BRITISH PATIENTS' CHOICE OF CARE FOR MINOR INJURY: A CASE-STUDY OF SEPARATE SAMPLE LOGISTIC DISCRIMINATION

I.T. Russell, Universities of North Carolina and Newcastle upon Tyne

Key words: Patient's choice of medical care; Separate sample logistic discrimintation; Stepwise procedure; Cross-validation.

Abstract

Under the British National Health Service, few types of patients are permitted a free choice between hospital care and family practice. However, the patient in need of treatment for minor injury is one of these types and this paper reports a study designed to identify factors which influence his choice.

Separate samples were drawn from patients presenting to hospital accident departments and from those consulting their family practitioner. Both samples were interviewed in their own homes as soon as possible after receiving treatment. These data have been analysed by stepwise application of separate sample logistic discrimination.

This analysis identifies only four 'objective' variables as conclusively affecting the patient's choice - his distance from his family doctor, his distance from the hospital, his diagnosis and his age. Cross-validation shows that the resulting discriminant function provides satisfactory estimates of the conditional probabilities associated with the patient's choice.

Introduction

British patients suffering from minor trauma enjoy the privilege, rare within the National Health Service (NHS), of making a free choice between two alternative systems of medical care; they are allowed to present for treatment either at a hospital Accident and Emergency Department (AED) or to their general practitioner (GP). Ever since the beginning of the NHS in 1948 (and even before that), there has been considerable debate, not only about the effects of such freedom, but also about its advisability.

In particular, the allocation of NHS resources to and within accident and emergency services, both in the hospital and in primary medical care, has been the subject of deliberation and of recommendation by a number of expert committees. The Platt Report (Central Health Services Council, 1962) recommended that the number of AEDs should be greatly reduced but that the level of staffing in the remaining units should be substantially raised. More recently, the Expenditure Committee of the House of Commons (1974) set on record its belief that the increasing use in general practice of appointment systems and deputising services had influenced patients' decisions to attend AEDs; it went on to propose a number of measures designed to counter these supposed determinants of the trend in patients' choices away from general practice and towards the AED.

The Newcastle Accident Survey was set up with a view to making an objective contribution to future decisions affecting the organisation of accident and emergency services. Specifically, it was designed to discriminate between two populations - patients in Greater Newcastle who consult general practitioners for the treatment of minor trauma and those who proceed direct to hospital.

Survey Method

As it is quite impractical to sample from the mixture of these two populations, we drew a separate sample from each. For the hospital population, it was quite easy to take a simple random sample from the register of each of the three AEDs within the survey area and then to exclude such 'foreign elements' as patients not suffering from trauma and those injured patients who were immediately admitted as inpatients and were thus, by definition, suffering from 'major' trauma.

However, there is no explicit sampling frame available for new patients consulting in general practice. Consequently, we drew a random sample, stratified by number of partners and geographical locality, of 58 GPs from the 290 doctors practising within the effective catchment area of the three AEDs. By observing each of these sampled GPs for one random week, we were able to define 'clusters' of patients, from which we excluded foreign elements much as before.

Both of these samples were interviewed in their own homes as soon as possible after receiving treatment. Unfortunately, practical constraints compelled us to handle the two samples as consecutive phases of the same study rather than concurrently. However, although the two phases had to be separated by a period of two years, we were able to carry them out over precisely the same quarter of the year. Furthermore, comparison of the numbers and characteristics of those patients in both samples who had been referred from general practice to one of the AEDs showed no significant differences.

Again, examination of routine NHS statistics and local demographic data produced no evidence of any appreciable secular trend either in the relative proportions of patients attending AEDs and general practice or in the distribution of the discriminating variables. (However, it is also worth recording that our method of analysis, yet to be described, is fairly robust to simple secular trends such as these; it requires a particularly perverse family of secular trends, those in which some discriminators become much stronger and other much weaker, to upset our analysis unduly.)

Our interviewers collected information on a wide variety of variables, of which 62 were common to both samples. However, this paper restricts attention to those variables which can be used in the future assessment of alternative policies by predicting likely responses. This requires, not only that there should be information available on the distribution of these variables in the community at large, but also that they should be objective, in two senses. We demand first that the responses should not, in all probability, have been

affected by anything occurring after the patient's choice of care (and, in particular, by the treatment he received) and secondly, that the same responses would probably have been obtained using a different method of data collection. Wherever possible, data on the 27 variables fulfilling these criteria were collected not from the patients themselves but from their medical records or by means of a postal survey of their GPs (Holohan et al., 1975).

Statistical Methods

Day and Kerridge (1967) have advocated the logistic form for posterior probabilities as a basis for discrimination between two populations, H_1 and H_2. Given that the values $x_1, x_2, \ldots, x_p$ of p potentially discriminating variables are known for an individual patient, they proposed that an appropriate formula for his resulting probability of belonging to (or in this case, opting for) population H_1 is $e^z/(1 + e^z)$ where z, usually known as the 'discriminant function' (DF), is given by:

$$z = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p$$

Replacing $\alpha_0, \alpha_1, \ldots, \alpha_p$ by their maximum likelihood estimates (MLEs) $\hat{\alpha}_0, \hat{\alpha}_1, \ldots, \hat{\alpha}_p$ leads to the discriminant rule 'Allocate to H_1 if $\hat{z}$ is positive, H_2 if $\hat{z}$ negative'. This rule is optimum in the sense that it chooses H_1 whenever $\text{Prob}(H_1/x_1,x_2,\ldots,x_p) > \text{Prob}(H_2/x_1,x_2,\ldots,x_p)$ and vice-versa.

But it was left to Anderson (1972) to consider the case, which arises here, when it is necessary, or preferable, to draw a separate sample from each population. He showed that the MLEs $\hat{\alpha}_1,\hat{\alpha}_2,\ldots,\hat{\alpha}_p$ are the same whether the sampling is carried out separately or from the mixture. However, the MLEs of the constant term, $\hat{\alpha}_0^*$ (separate samples) and $\hat{\alpha}_0$ (mixture sampling), are identical if and only if the proportion Π_1^* of H_1 in the separate samples taken together is the same as the proportion Π_1 of H_1 in the 'universe'.

In our study, we interviewed 155 patients who had opted for an AED (H_1) and 191 who had sought care in general practice (H_2); thus $\Pi_1^* = 0.448$. However, we have estimated[2] (Russell and Holohan, 1974) that $\Pi_1 = 0.482$ with an approximate confidence interval of (0.427, 0.538). Since we are concerned more with $\alpha_1,\alpha_2,\ldots,\alpha_p$ than with α_0, we are able to take $\Pi_1^* = \Pi_1$ and $\hat{\alpha}_0^* = \hat{\alpha}_0$.

The maximum likelihood equations for $\hat{\alpha}_0, \hat{\alpha}_1,\ldots,\hat{\alpha}_p$ are solved using a Newton-Raphson procedure and starting values of zero for all parameters. However, it is not computationally feasible to force all 27 variables under investigation into the discriminant function and then to test each coefficient for significance. We therefore build up the DF in a stepwise manner; at each step we identify that variable, not yet represented in the DF, which would generate the greatest improvement in the maximised log likelihood if incorporated in the DF. We then determine whether that improvement is statistically significant by taking into account not only the asymptotic χ^2 property of the maximised log likelihood (Cox, 1970) but also the combinatorial effect of choosing as test statistic the largest of (27-p) improvements in that maximised log likelihood. If this approximate test is significant, the variable is added to the DF; if not, the sequential procedure is terminated.

Findings

Our intention in applying logistic discriminant analysis to the 27 objective variables measured by the Newcastle Accident Survey is to identify that subset which *together* provide the best prediction (best, that is, in the sense that no statistically significant improvement is possible) of the patient's initial choice of care system. However, before analysing the data in this multivariate fashion, we examine the effects of certain variables *in isolation*.

TABLE 1

INITIAL CHOICE BY DISTANCE TO GP'S SURGERY

Distance from Site of Decision ('Source') to GP's Surgery	Initial Choice of Care			
	Hospital		GP	
	No.	%	No.	%
0.7 miles or less	47	30.3	112	58.6
0.8 to 1.7 miles	40	25.8	51	26.7
1.8 to 2.7 miles	34	21.9	20	10.5
2.8 miles or more	34	21.9	8	4.2
Total	155	99.9	191	100.0

Significance Test $\chi_3^2 = 44.3$ (Significant at 0.1% level)

It comes as no surprise to find in Table 1 that the farther the patient found himself from his GP, the less likely he was to consult him; similarly, the farther from the AED, the less likely he was to present there. Other attributes of the patient with effects significant at the 0.1% level are his diagnosis (fractures and wounds tend to be taken to the hospital, other conditions to the GP) and his age (patients between 15 and 44 are most likely to report to the AED, those over 65 least likely).

TABLE 2

INITIAL CHOICE BY PRACTICE APPOINTMENTS SYSTEM

Practice Use of Appointments System	Initial Choice of Care			
	Hospital		GP	
	No.	%	No.	%
Yes (All surgeries)	67	43.2	82	42.9
Yes (Some surgeries)	26	16.8	35	18.3
Yes (Not otherwise specified)	31	20.0	40	20.9
No	31	20.0	34	17.8
Total	155	100.0	191	99.9

Significance Test $\chi_3^2 = 0.38$ (Not significant)

Although variables describing the patient or the circumstances of his accident show marked univariate effects, the same is not true of those three variables which relate to the general practice with which he is registered. Indeed, the number of partners in that practice is the only one of these variables which is significant and, even then, only by testing for a linear trend in the proportion of patients choosing the hospital (a proportion which decreases with increasing partnership size). Furthermore, the two discriminators (implicitly) proposed by the Expenditure Committee of the House of Commons (1974) —whether the patient's GP makes use of an appointments system (Table 2) and of a deputising service —have no discernible one-way effects. In the case of deputising services, it is just possible that a genuine effect has been masked by the failure of GPs to respond to the postal questionnaire. However the patient's perception of whether his practitioner uses the deputising service seems to have no more effect on his actions than the objective version of that variable.

Two of the remaining 20 variables under consideration in this paper —the patient's occupational status and whether he had been treated at an AED in the preceding year —have univariate effects which are significant at the 1% level. A further five, including sex, martial status and 'external cause of injury' (International Classification of Diseases, 1967) were significant at the 5% level and the residual 13 had, in isolation, no significant effect on patients' choices.

TABLE 3

LOGISTIC DISCRIMINANT ANALYSIS:
DEFINITION OF VARIABLES
(RANKED BY ABILITY TO MAXIMISE LOG LIKELIHOOD)

Var. No. & Rank	Definition of Variables	No. of Cats.	Cat. Most Likely to Choose GP ($x_j = 0$)	Max. Log. L'hood
x_1	Distance to GPs surgery	4	≤ 0.7 miles	
x_2	Distance to hospital	4	≤ 2.8 miles	-192.90
x_3	Final diagnosis	2	All but fractures and wounds	-183.84
x_4	Age	4	≥ 65 years	-174.98
x_5	Has GP any partners?	2	Yes	-171.12

When we apply logistic discrimination to these data in the stepwise fashion already described, the two distances are the first to appear in the DF (Table 3). Since the corresponding parameter estimates, $\hat{\alpha}_1$ and $\hat{\alpha}_2$, are so similar (Table 4), it is clear that patients give equal weight to each of the two distances; all things being equal, it is the nearer of the two sources of emergency medical care which will be chosen. However, as is shown by the next two variables to enter the DF —final diagnosis and age —things are not always equal.

TABLE 4

LOGISTIC DISCRIMINANT ANALYSIS:
PARAMETER ESTIMATES AND THEIR STANDARD ERRORS

No. of Vars. in DF	Parameter Estimates (Standard Errors)				
	$\hat{\alpha}_0$	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\alpha}_4$
2	-1.971 (0.253)	0.837 (0.128)	0.799 (0.129)		
3	-2.492 (0.301)	0.865 (0.134)	0.850 (0.135)	1.106 (0.266)	
4	-3.911 (0.499)	0.830 (0.137)	0.898 (0.142)	1.180 (0.277)	0.980 (0.242)
4 (Hosp. A only)	-4.524 (0.802)	0.927 (0.197)	0.794 (0.234)	1.311 (0.422)	1.185 (0.399)
4 (Hosp. B only)	-3.121 (0.910)	0.865 (0.289)	0.854 (0.255)	1.356 (0.566)	0.656 (0.426)
4 (Hosp. C only)	-3.764 (1.071)	0.609 (0.264)	0.836 (0.328)	1.175 (0.556)	1.100 (0.473)

At this point, only one of the remaining 23 variables under consideration —the number of partners in the patient's general practice —is able to increase the maximised log likelihood by more than 1.92, the upper 5% point of $\frac{1}{2}\chi^2_1$ — the appropriate significance test in these circumstances (Cox, 1970). We deduce that none of the residual 22 variables has a significant effect on the patient's decision *over and above* that of the variables already selected —distance, age and final diagnosis. Since we have already mentioned that seven of the 22 have significant *univariate* effects however, it is helpful to consider one of these in a little more detail.

Of the patients who first sought care at an AED, 34% acknowledged that they had attended an AED at least once during the previous 12 months; the corresponding percentage among patients who reported to their GP was only 20%. That this tendency does not assist in the discrimination is explained by its positive correlation with all four variables already in the DF. In other words, of the factors which, according to our analysis, led patients to the AED on the first occasion, age is essentially immutable, and distance and diagnosis have a better than average probability of remaining unchanged.

It only remains to discuss whether partnership size should contribute to the DF. As Table 3 shows, this variable increases the maximised log likelihood by 3.86. Now although this value is approximately equal to the upper ½% point of $\frac{1}{2}\chi^2_1$, it must be remembered that 23 different variables are competing to become the fifth variable in the DF. Consequently, if all these variables were independent, the true significance level would be close to 10%. However, since our knowledge of the true correlation structure of these 23 variables is limited to this one survey, all we can say with any confidence is that the significance level to be attached to the proposition that partnership size has an intrinsic effect on the patient's initial choice of care system lies between ½% and 10%.

(Although computer simulation would enable us to be a little more precise about the size of this significance level, we doubt whether it would lead to a statement to the effect that it was less (or, for that matter, greater) than 5%.)

Thus our analysis has identified four variables which influence the patient's decision and one whose (independent) effect is is not proven. There is no evidence that the remaining 22 variables have any intrinsic effect. In particular, neither the use of an appointments system by the patient's GP nor that of a deputising service was able, at any of the five steps, to add more to the maximised log likelihood than the sixteenth (ranked by ability to maximise log likelihood) of the remaining 20 variables. However, before we can discuss the relevance of these findings for the National Health Service, we must assess how reliable they are and to what extent they may be regarded as representative of a wider population than that from which they have been derived.

Validation of Findings

Discriminant functions are traditionally appraised by examining the probabilities with which they misallocate to population H_2 patients who actually belong to (or, in this case, opt for) population H_1, and vice-versa (Hills, 1966). Although, this approach is less appropriate to logistic discrimination than it is to the classical method of linear discrimination, it provides a convenient starting point for our validation. However, there is no implied hierarchy among our populations, in the sense that misallocation from hospital to general practice is any more (or less) important than in the opposite direction. Further, both samples are of the same order of magnitude, as are the populations from which they were drawn. Consequently, there is no need for us to distinguish between the two types of misallocation.

TABLE 5

LOGISTIC DISCRIMINANT ANALYSIS:
MISALLOCATION RATES

No. of Vars. in DF	Crude Rates		Cross-Validatory	
	Within Hosps.	Overall	Random	Between Hosps.
2	0.286	0.301	0.301	0.328
3	0.251	0.272	0.275	0.282
4	0.240	0.234	0.254	0.253
5	0.231	0.234	0.237	0.231

The first two columns of Table 5 therefore present crude misallocation rates, 'crude' in the sense that they merely indicate what proportion of all 346 cases are misallocated by the discriminant functions specified in the first three rows of Table 4 and that based on all five variables included in Table 3. The second column is based on the estimation of a single DF for the entire data-set (the only case so far considered). The first column takes that analysis one stage farther by calculating a separate DF for each of the three hospitals involved; misallocation is then identified by comparing each patient's decision with that predicted by the DF appropriate to the AED in question (i.e. the one actually visited or which, according to the interview, would have been visited had the patient not elected to consult his GP).

To overcome the undesirability of testing a DF on the data which produced it, a number of authors, including Mosteller and Tukey (1968), have proposed the use of 'cross-validation', a technique in which the DF is estimated using all but one of the cases and the discarded case then used to assess that estimate. Eliminating each case in turn, thus repeating the procedure as many times as there are cases, leads to a less biased estimate of the misallocation rates.

However, even though it is an easy, if tedious, matter to carry out such an exercise, we are loath to commit ourselves to a further 346 computer runs whenever we have a DF to test. We therefore compromise by successively withdrawing each of 10 mutually exclusive and exhaustive random 10% samples. It is reassuring to find that the resulting cross-validatory estimates of the probabilities of misclassification, which appear in the third column of Table 5, are so close to the (more suspect) crude estimates in the first and second columns.

Since our 4-variable DF thus appears to be statistically valid within Greater Newcastle, we now enquire how relevant it is to other urban areas in the United Kingdom. Although a definitive answer to this question must wait for similar research to be carried out in other parts of the country, we make use of the fact that the three hospitals with which we are concerned are very different in character. Hospital A, which lies close to the centre of Newcastle, has for many years been the teaching hospital of the area; although hospital B has recently become a teaching hospital, its tradition is that of a municipal hospital serving one of the poorer parts of the city; hospital C, a smaller municipal hospital, is situated in the adjacent town of Gateshead.

This diversity suggests that, by validating across these hospitals (in much the same way as the validation across random sub-samples which we have already described), we can at least hint at what a validation across regions might eventually show. We first observe from Table 4 that the three hospital-specific 4-variable DFs show considerable similarities. Although hospital B has a non-significant age co-efficient $\hat{\alpha}_4$ (in other words, the sequential estimation procedure terminates after three steps rather than four), this is, arguably, attributable to the sample size of only 93. More important, the inter-hospital misallocation rates tabulated in the final column of Table 5 are remarkably close to the random cross-validatory rates, expecially when there are four variables in the DF.

Hence it may be suggested (and we put it no higher than that) that the logistic model which we have derived is applicable beyond the limits of Greater Newcastle. Furthermore, lest it should be thought that even this hint of wider applicability is compromised by our failure correctly to predict the decisions of as many as one-quarter of all patients who sustain minor injuries, it must be stressed that the advantage of the logistic method lies not in its power to make infallible forecasts

in the face of uncertainty but in its ability accurately to estimate the probabilities inherent in that uncertainty.

TABLE 6

LOGISTIC DISCRIMINANT ANALYSIS: GOODNESS OF FIT (4-VARIABLE DF)

Estimated DF	No. of Pats.	Predicted		Observed	
		Hosp.	GP	Hosp.	GP
$\hat{z} < -2$	59	4.9	54.1	8	51
$-2 < \hat{z} < -1$	53	10.7	42.3	11	42
$-1 < \hat{z} < -\frac{1}{2}$	40	11.4	28.6	9	31
$-\frac{1}{2} < \hat{z} < 0$	52	23.0	29.0	19	33
$0 < \hat{z} < \frac{1}{2}$	32	17.8	14.2	19	13
$\frac{1}{2} < \hat{z} < 1$	52	35.6	16.4	36	16
$1 < \hat{z} < 2$	30	25.0	5.0	27	3
$\hat{z} > 2$	28	26.6	1.4	26	2
Total	346	155.0	191.0	155	191

Significance Test $\chi^2_7 = 5.50$ (Not significant)

To illustrate this point, Table 6 compares the choices predicted by our 4-variable logistic model with those actually made by the survey patients. (If only for the sake of simplicity, the table presents a 'crude' goodness-of-fit test rather than the cross-validatory test which is the logical conclusion of the argument of this section.) Although our previous emphasis on 'misallocation' has served its purpose, it is worth stressing how misleading that term is in the context of logistic discrimination by pointing out that, until now, all the boxed figures on Table 6 have been so described. In view of the evidence of that table (and its cross-validatory equivalents), the DF which we have developed may fairly be described as a probabilistic model of the decision-making behaviour of patients suffering from minor trauma.

Discussion

The Newcastle Accident Survey has derived a statistical model which predicts minor accident patients' choices between AED and general practice with some accuracy in the face of the uncertainty evident in these decisions. Furthermore, the ability of this model to cope with three very different hospitals, admittedly all situated with Greater Newcastle, has led us to suggest that it may also be relevant to other urban areas.

Although this amounts to a claim that the original objective of our study has been achieved, it is not intended to suggest that there are no other objectives which could (or even should) have been tackled. Indeed, there are two particular ways of extending our research which have always seemed to us desirable but which, like all desirable things, are not without cost.

First, our survey has been designed, implemented and analysed under the assumption that those who sustain minor accidents first decide *whether* to seek medical care and only when they have so resolved do they choose *where* to seek it. This is, of course, an over-simplification because there are, for example, some injured patients who perceive their options as being limited to general practice or self-treatment. Any comprehensive discriminatory model of accident behaviour would have to acknowledge that there are (at least) three types of care for which the sufferer can opt. However, the identification of, and collection of *comparable* data from, the self-treaters is demonstrably much more costly and arguably less relevant to NHS decision-making than the exercise we have undertaken.

Secondly, we have made no attempt to compare the costs of treating the marginal minor trauma patient at an AED with those of caring for him in general practice, nor even to identify any of those costs. Consequently, the economic conclusions to be drawn from our study are limited to statements about the ordinal values which different minor trauma patients place on the two alternative forms of NHS treatment. We know nothing either of the way those values compare with that of self-treatment, or of the values of any of these treatments to society as a whole or even to the NHS as an institution.

Implicitly, it seems, patients accord the mile travelled to the surgery the same disutility as the mile travelled to the AED. The other two major discriminators tell us that hospital care is more highly valued by those suffering from fractures or wounds and by those between the ages of 15 and 44. (It is worth stressing here that although theses attributes are correlated, the multivariate nature of the statistical analysis ensures that the findings are independent; loosely speaking, the youth with a fracture is doubly likely to report to the AED.) It seems that the AED is held to be more proficient on the technical or instrumental side while the GP is seen as more supportive in the affective or emotional aspects of the care of accidents (Holohan, 1976). Hence the hospital is preferred both by those who need a technical service such as suturing and by those of an age-group which places a higher value on technical care than on affective care.

Our analysis is equivocal on the question of whether one further variable—whether the patient's GP has any partners—has any intrinsic effect on the patient's decision over and above that of the first four variables, and thus qualifies for inclusion on the model. If so, this would mean that the treatment of injuries by single-handed practitioners is valued less highly than when undertaken by partnerships. However, it is not yet clear whether this would, if true, reflect some inherent quality of one-man practices or whether it would be attributable, for example, to the lower proportion of such practices with attached and employed nurses, as reported by Reedy et al. (1976).

Much more certain, however, is the Newcastle Accident Survey's lack of support for the view taken by the Expenditure Committee of the House of Commons (1974) that 'the use of appointments systems and deputising services can be thought to have had some influence on patients' decisions to attend AEDs'. Not only are these two variables not even remotely associated with the choice of care system, as shown by a simple cross-tabulation, but neither was able to make any contribution to our

multivariate statistical analysis. Furthermore, similar negative findings with respect to deputising services have been reported by Williams et al. (1973), who analysed a year's deputising service consultations in Sheffield and Nottingham and secular trends in first attendances at AEDs.

Another argument which gets no support, either from the statistical analysis reported here or from the sociological analysis carried out by Holohan (1976), is that patients who present to AEDs for the treatment of minor injuries are either irrational or perverse, as suggested by some of the more outspoken writers in the medical press and even as hinted at by one or two of those who gave evidence to the Expenditure Committee. Indeed, the picture which emerges from our work is one of patients exercising considerable judgement in deciding which course of action is in their own better interest; so much so that their behaviour in aggregate conforms very closely to the mathematical model which we have proposed.

Acknowledgements

I acknowledge with gratitude the contribution of colleagues in the University of Newcastle upon Tyne, England — John Anderson, Ann Holohan, Bill Morgan and Peter Philips — and financial support from the Department of Health and Social Security, London.

References

Anderson, J.A. (1972). Separate Sample Logistic Discrimination. *Biometrika* 59, 19-35.

Central Health Services Council (1962). Accident and Emergency Services: Report of the Subcommittee of the Standing Medical Advisory Committee. H.M.S.O., London.

Cox, D.R. (1970). *The Analysis of Binary Data.* Methuen, London.

Day, N.E., Kerridge, D.F. (1967). A General Maximum Likelihood Discriminant. *Biometrics* 23, 313-323.

Hills, M. (1966). Allocation Rules and Their Error Rates. *Journal of the Royal Statistical Society, Series B,* 28, 1-31.

Holohan, A.M. (1976). Illness and Accident Behaviour. In *The Sociology of the N.H.S.* ed. Stacey, M., Sociological Review Monograph No. 22, 11-119.

Holohan, A.M., Newell, D.J., Walker, J.H. (1975). Practitioners, Patients and the Accident Department. *The Hospital and Health Services Review* 71, 80-84.

House of Commons (1974). Accident and Emergency Services: Fourth Report from the Expenditure Committee for the Session 1973-74. H.M.S.O., London.

International Classification of Diseases (1967). Manual of the International Statistical Classification of Diseases, Injuries and Causes of Death. World Health Organization, Geneva.

Mosteller, F., Tukey, J.W. (1968). Data Analysis including Statistics. *Handbook of Social Psychology* ed. Lindsay, G., Aronson, E. Addison-Wesley, Reading, Mass.

Reedy, B.L.E.C., Phillips, P.R., Newell, D.J. (1976). Nurses and Nursing in Primary Medical Care in England. *British Medical Journal* ii, 1304-1306.

Russell, I.T., Holohan, A.M. (1974). Newcastle Accident Survey: Report on Phases I and III — the Choice of Care for Minor Trauma. Medical Care Research Unit, University of Newcastle upon Tyne.

Russell, I.T., Philips, P.R., Anderson, J.A. (1976). Separate Sample Logistic Discrimination: A Case Study in Health Services Research. In *Compstat 76 — Proceedings in Computational Statistics* ed. Gordesch, J., Naeve, P. Physica-Verlag, Vienna.

ESTIMATING HOSPITAL MEAN LENGTH OF STAY FROM CENSORED DATA

R.L.W. Welch, Federal Reserve Board
T.L. Smith and M.W. Denker, University of South Florida

1. Introduction

This paper reports on a study of hospital census data undertaken at the University of South Florida Medical Center. It was conducted under the auspices of a joint project between the Departments of Psychiatry and Mathematics, the purpose of which was to bring together statistical and medical professionals, in order to investigate important and timely problems of interest to the latter. The primary area of concern was the evaluation of mental health care delivery. Two Tampa hospitals, St. Joseph's Hospital Mental Health Center and the Veterans Administration Hospital, generously made data available to the project.

The primary data base consists of several complete series dating from the opening of St. Joseph's Mental Health Center in late May 1971 to March 1975. These series include the daily census total (that is, number of beds occupied), number of admissions daily, number of discharges daily, and the length of stay and age of each discharged patient. In addition, the data for the first three series is extended through March 1977, for a total range of some 1,960 days. A time series analysis of this data, which should be read simultaneously with the present paper, can be found elsewhere in this volume (Smith, Welch, and Holland, 1977). We deal here with the analysis of the length of stay data, with the purpose of investigating procedures for estimating the mean length of stay.

The secondary data base consists of census information taken from a demographic survey on May 18, 1977, from the five psychiatric wards at the Tampa VA Hospital. Because most of the patients were not discharged on the day of the survey, the length of stay statistics are incomplete or "censored" by the time of observation. This situation is analogous to Type II censoring in reliability studies. We observe a sample of n patients until r, a fixed number, are discharged. The random lengths of stay of the remaining n-r patients are recorded as the total time each has spent in the ward, up to the discharge of the rth patient.

2. General Problem

The standard hospital procedure for compiling census statistics is to have a medical records clerk, once a month, collate all the relevant data pertaining to patients admitted or discharged in the previous month. The mean length of stay is computed for the latter. Now some, indeed many, of the patients may have been discharged near the beginning of the previous month. Thus, there is at least a month's time lag involved before the results become known. Furthermore, it is not customary to compute the variance or study any of the statistical properties of the distribution of lengths of stay.

On the other hand, such information would be useful in the administration of a ward or hospital. A medical director may wish to schedule admissions or discharges with a view toward maintaining the census near some equilibrium. If it is known in advance that there is likely to be a temporary decline in bed occupancy, personnel may be assigned to other duties. Conversely, during a period of high occupancy, certain non-vital tasks, like routine maintenance, may be postponed, or non-emergency admissions may be deferred until the beds are available.

In general, the random variable represented by length of stay is influenced by the type of ward, admission diagnosis, and age of the patient. In a medical or surgical ward, for example, the mean and standard deviation are commonly on the order of a few days to a week. At a voluntary admission, private care psychiatric facility like St. Joseph's, however, the typical adult stays two to three weeks. An adolescent or child, however, may stay an average of one or more months, with a corresponding increase in the standard deviation. The extreme case occurs in public, custodial institutions like state mental hospitals, where patients may be warehoused essentially for an indefinite period.

In the next section, we fit an exponential probability density function to the St. Joseph's data, in order to determine a theoretical model for further study. This is followed by a discussion of appropriate estimators for a censored sample, along with an example taken from the VA Hospital data. This latter type of sample is suggested as a means of computing a more timely estimate of the mean length of stay.

3. Exponential Model

Previous authors have considered the problem of fitting a density function to observed hospital lengths of stay. Cooper and Cocoran (1974) used an exponential distribution, and DuFour (1974) used a lognormal. In general, lengths of stay tend to be unimodal with a strong rightward skew. The mode may appear at or near 1 day, giving the histogram a J-shaped appearance. Thus, either distribution would appear as likely candidates, as well as might various forms of the gamma or Weibull distributions.

Table 1 reproduces the length of stay frequencies, with the data grouped into 25 classes, for 1,310 patients at St. Joseph's Hospital. (Note that we are not considering patients who did not occupy a bed for at least 1 night). Because of the difference in distributions between adolescents and adults, and because we wished to generalize the model to the VA Hospital data, only patients aged 20 or more were considered. The resulting sample of patients had a mean length of stay of $\bar{x} = 17.044$ days, with a

standard deviation of 22.241 days. Some age effect, however, still remains. Out of 21 patients with a recorded length of stay greater than 80 days, 14 were 20-24 years old. The distribution of ages of adult patients at St. Joseph's is more nearly uniform, with a range from 20 to 90 years.

A goodness-of-fit test for the exponential distribution

$$F(t) = 1 - \exp(-t/\beta) ,$$

where the mean β is estimated by $\bar{x}$, resulted in a value of $X^2 = 36.1$, $p = .042$ with 23 degrees of freedom. The "messiness" in the data mentioned above at least partially contributed to a too-heavy tail. Furthermore, the patients themselves have a broad heterogeneity of background and diagnosis in comparison to those at the VA Hospital. Thus, we are willing to accept this otherwise marginal value and assume that the data is exponential. Neither the lognormal, Weibull, or gamma distributions produced acceptable fits.

Table 1: Lengths of Stay for 1,310 Patients

Time in Days	Observed	Expected
< 3.5	244	243.19
3.5 - 6.5	181	172.18
6.5 - 9.5	163	144.39
9.5 - 12.5	119	121.09
12.5 - 15.5	108	101.54
15.5 - 18.5	82	85.155
18.5 - 21.5	82	71.411
21.5 - 24.5	76	59.886
24.5 - 27.5	35	50.221
27.5 - 30.5	42	42.115
30.5 - 33.5	33	35.318
33.5 - 36.5	16	29.618
36.5 - 39.5	15	24.838
39.5 - 42.5	16	20.829
42.5 - 45.5	17	17.467
45.5 - 48.5	12	14.648
48.5 - 51.5	6	12.284
51.5 - 54.5	8	10.301
54.5 - 57.5	8	8.6389
57.5 - 60.5	6	7.2446
60.5 - 63.5	3	6.0754
63.5 - 66.5	3	5.0948
66.5 - 69.5	4	4.2726
69.5 - 72.5	4	3.583
> 72.5	27	18.617

The probability density function of an exponential distribution is given by

$$f(t) = (1/\beta) \exp(-t/\beta) ,$$

with $E(t) = \beta$ and $Var(t) = \beta^2$. One interesting property for this distribution is that the hazard rate

$$h(t) = f(t)/(1 - F(t)) = 1/\beta$$

is a constant. The quantity $h(t)dt$ is the probability that a patient on the ward for t days will be discharged during the interval $(t, t+dt)$. For a long-term care, custodial institution it may be more realistic to fit a distribution that has a hazard rate which approaches 0 for large t. Either a lognormal or a Weibull density would fit this criterion.

4. Censored Sampling

For a complete sample, $\hat{\beta} = \bar{x}$ is an unbiased, maximum likelihood estimator for β. Let us assume a general case of censored sampling, where the sampling is progressive. We have a random sample of size n, where r patients are discharged at times $t_i (i = 1,\ldots,r)$, and n-r patients remain on ward at times t_{r_i} $(i = 1,\ldots,n-r)$. Then the likelihood of the r_i sample is given by

$$L = \frac{n!}{(n-r)!} \prod_{1}^{r} (1/\beta) \exp(-t_i/\beta)$$

$$\times \prod_{1}^{n-r} (1 - F(t_{r_i})) .$$

Thus,

$$\ell n\, L = -r\, \ell n\, \beta - (1/\beta) \sum_{1}^{r} t_i - (1/\beta) \sum_{1}^{n-r} t_{r_i} + \text{constant} ,$$

and

$$\frac{\partial\, \ell_n L}{\partial\, \beta} = -r/\beta + (1/\beta^2) \sum_{1}^{r} t_i + (1/\beta^2) \sum_{1}^{n-r} t_{r_i} .$$

The maximum likelihood estimator is given by

$$\hat{\beta} = \left(\sum_{1}^{r} t_i + \sum_{1}^{n-r} t_{r_i}\right)/r .$$

Mann, Schafer, and Singpurwalla (1974) also derive a best invariant estimator,

$$\tilde{\beta} = r\hat{\beta} / (r+1) .$$

The following numbers are the lengths of stay for 27 patients from Ward 2 of the Tampa VA Hospital, on May 18, 1977. The first 4 patients (reading left to right) were discharged that day; the remaining 23 observations are censored.

2	19	1	5	60	18	25	47	88
87	78	76	38	63	57	57	53	51
45	34	29	28	28	22	11	15	12

Thus,

$$\hat{\beta} = 27/4 + 1022/4 = 262.25 ,$$

and

$$\tilde{\beta} = (4/5)\hat{\beta} = 209.8 .$$

In practice, if there is a high mean length of stay and low daily turnover (admissions and discharges), one may find that this procedure leads to an unsatisfactorily great amount of censoring. An alternative scheme would be to follow the selected group of patients over a specific a amount of time (say, one or two weeks). The algebraic results would be identical to those shown above. This procedure will have the advantage of reducing the number of observations which are censored, though at the cost of some delay in obtaining the desired results. The study period might be varied relative to the type of ward and what is known a priori about the mean length of stay.

5. Acknowledgements

The authors wish particularly to thank Stephen White, Medical Director of St. Joseph's Mental Health Center, and Anthony Reading, Chairman of the University of South Florida Psychiatry Department and Acting Chief of the Tampa VA Hospital Psychiatric Service, for their assistance. This research was sponsored by a grant from the National Institute of Mental Health.

6. References

1. Cooper, J.K., and Cocoran, T.M. Estimating bed needs by means of queuing theory. New England Journal of Medicine, 1974, 291, 404-405.

2. DuFour, R.G. Predicting hospital bed needs. Health Services Research, 1974, 9, 62-68.

3. Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. Methods for statistical analysis of reliability and life data. New York, Wiley, 1974.

4. Smith T. L., Welch, R. L. W., and Holland, T. Time series analysis of mental health center census data. ASA Proceedings of the Social Statistics Section 1977

GRAPHIC MEASURES OF RESOURCE ABSORPTION IN ALCOHOLISM TREATMENT PROGRAMS

Alex Richman
Department of Psychiatry, Beth Israel Medical Center (307 Second Avenue, New York, N.Y. 10003) and the Mount Sinai School of Medicine of the City University of New York.

Quantification is essential in the planning, management and evaluation of health services; estimating cost effectiveness, and allocating resources. Alcoholism is a disorder for which quantification is particularly difficult.

> "...we have been employing rather nebulous variables to characterize a non-defined population of subjects treated by an ineffable process to produce a rather fuzzy outcome."
>
> Ludwig

Resource absorption refers to inequity in the use of clinical services wherein a minority of patients uses a disproportionately large volume of treatment. Such inequity, when recognized, is rarely quantified. This paper describes some graphic measures which will assist in quantifying and comparing resource absorption in alcoholism treatment programs.

CLINICAL PERSPECTIVES

The number of male mental hospital first admissions with alcoholic disorders increased 64% between 1962 and 1969. One-fifth of all male admissions to psychiatric facilities were alcoholic disorders in 1970. Nearly one-half of mental hospital male admissions aged 35-64 were diagnosed as alcoholic. (Redick) Since the significance of alcoholic disorders in hospital programs is increasing, we must review how hospitalization, the most costly form of treatment is being used.

What data do we have on initial outcome, relapse or recidivism in hospital programs for alcoholics? Baekeland, et al concluded that, despite the introduction of new treatment methods, the effectiveness of hospital treatment for alcoholism seemed no better from 1960 to 1973 than it was from 1953 to 1963, and no differences were found in the effectiveness of different kinds of treatment regimens. Detoxification programs, the most frequent type of treatment regimen, often care for persons who are drinking or drunk, but not in need of detoxification; intensive medical treatment is provided for some alcoholics who do not require intensive medical hospitalization; and some detoxification programs fail to provide alcoholism treatment. (Pattison) Nevertheless, established in-patient detoxification programs continue and acute-care hospitals are developing more in-patient detoxification programs resembling those which some already recognize as unsatisfactory.

> "The same alcoholics will repeatedly appear time after time in medical crises, the staff becomes demoralized and nothing effective is accomplished. It is a "revolving door".
>
> Pattison

SCARCITY OF DATA ON READMISSIONS

Patients with numerous readmissions are a major problem for alcohol treatment programs. New treatment programs rapidly accumulate readmissions (Richman and Smart); accomodation for new patients is reduced (Richman); and staff morale and therapeutic optimism is lowered (Richman and Dunham). However, little attention has been given to the biometrics of readmission and there are few reports of the readmission experience of specific programs against which detailed comparisons can be made.

Assessment of resource absorption is often resisted, as being in conflict with treatment philosophy; a problem that will not occur in well planned programs, or irrelevant to current models of alcoholism. Few statistical reports analyze the treatment events accumulated by a cohort over a period of time. Sophisticated statisticians are often reluctant to embark on descriptive studies which suffer from incomplete data, do not assess outcome in the community or treatment in other settings.

QUANTIFICATION OF READMISSIONS

Trends in readmissions are assessed by:

1) The percentage of readmissions among admissions,
2) The numerical distribution of previous hospitalizations for individuals; and
3) Actuarial rates of readmission, specific for number of previous admissions.

The percentage of readmissions among admissions is affected by changes in the absolute number of first admissions and, as well, the denominator does not include all those who are exposed to the risk of the occurence. (Moon and Patton)

An increasing percentage of readmissions among admissions is often alleged to be accounted for by the increasing number of former patients at risk of readmission from the community. It is sometimes rationalized that readmissions reflect the patients' satisfaction or confidence in the treatment program. Data analysis rarely substantiates these claims.

The numerical distribution of previous hospitalizations is sometimes tabulated. Usually such tabulations include persons with varying time-intervals of observation. In relatively new or expanding programs, the proportion of first admissions is particularly exaggerated.

Table 1 shows the distribution of events reported for alcoholics in two large-scale information systems; the Alcoholism Program Monitoring System operated by National Institute on Alcohol Abuse and Alcoholism (NIAAA), and that operated by the Missouri Department of Mental Health. The NIAAA data show the number of times detoxification services were received by 62,873 persons reported since 1973 by 42 NIAAA funded Alcoholism Treatment Centers. The Missouri data show the number of in-patient admissions between Jan. 1970-Nov. 1974 for 15,577 individuals who had received a diagnosis of alcoholic disorder on at least one discharge. In both systems, the majority of individuals had only one event reported. However, in NIAAA there were 1.5 detoxifications reported per patient, and in Missouri there were 1.9 hospitalizations per patient. The proportion of first admissions is exaggerated in both sources because the patients recently admitted for the first time have had less opportunity for readmission than those with longer periods of observation.

Lorenz-type curves can be constructed from these data to show the cumulative percentage of treatment events accounted for by various percentiles of the population ordered according to number of events. Disparities in the utilization of treatment by individuals thus become visually more apparent (Siegel and Goodman); in both treatment systems about 4 per cent of the alcoholics account for 24% of the events. (Fig. I)

Fig. I also shows inequity in the distribution of out-patient attendances by a group of alcoholics during 21-24 months following first admission. One quarter of the patients attended less than five times, half attended less than 12 times, one quarter attended more than 52 times and one-eighth attended 100 or more times; 13% of the patients accounted for 57% of the total attendances.

Actuarial analyses of readmission probabilities are needed to assess "recidivism". Recidivism, as defined by criminologists, is the progressive advance in readmission rates for persons with increasing numbers of previous admissions (Wilkins). Few such analyses of time-specific readmission rates have been reported for alcoholics with various numbers of previous hospitalizations.

Fig. II shows the time-specific readmission rates for alcoholic disorders discharged from one of the New York State Mental Hospital alcoholic units. These data, supplied by A. Weinstein, were part of a large scale analysis by the New York State Department of Mental Hygiene which collated treatment events reported for individuals. The time-specific probability of readmission progressively advances for those with increasing numbers of previous hospitalizations.

Fig. III shows the estimated rates of readmission for patients discharged from Canadian psychiatric institutions with the diagnosis of alcoholic disorder during April-June 1973. These estimates were derived by "inferential linking" of readmission events for a cohort of discharges on the basis of dates of previous discharge and the number of previous hospitalizations (Richman). This method of estimation does not require a unique, personal, life-time identifier, and thus avoids the difficulties of machine matching or the problems of maintaining confidentiality in large scale information systems.

RESOURCE ABSORPTION INDEX

Time-specific rates of readmission have been shown to increase for patients with progressive numbers of previous hospitalizations. How can these data be summarized and their impact on resource absorption in treatment programs emphasized? The time-specific, event-specific rates of readmission can be applied to a hypothetical program with constant admission capacity and stable duration of stay and the proportion of readmissions among admissions projected for successive time periods following opening of the program. (The algorithm was developed and programmed on a Wang 2200-B by David Ross Richman)

The Resource Absorption Index (RAI) is the proportion of resources used by readmissions in the hypothetical treatment program. This index stabilizes between one and two years. The proportion of readmissions is shown in Fig. IV for programs subjected to the readmission rates of Figs. II and III. At the readmission rates inferred for Canadian psychiatric institutions, 19% of the resour-

ces would have been used by readmissions; at the New York State unit readmission rates, 28% of the resources would have been used by readmissions at the end of two years. These values of resource absorption are minimized since the readmission rates are truncated at 9 months and limited to 5 readmissions.

The increase in readmissions and the progressive reduction in accomodation (or "silting up") for first admissions can be graphed and readily communicated to clinicians and program administrators.

DISCUSSION

It is clear that in-patient detoxification programs represent a form of treatment which is expensive; whose effectiveness is questionable; and whose potential benefits are markedly reduced by the small number of patients who frequently return. Resource absorption is critical for:

a) clinical information systems
b) cost-effectiveness estimates
c) program evaluation

It must be emphasized that resource absorption can escape detection from many clinical information system reports. Discussion of resource absorption frequently evokes a defensive response from clinicians or administrators. There are two types of questions: one is whether the observed level of resource absorption conforms to clinical expectations or program goals; the second is how the level of resource absorption compares to other programs. By analyzing the utilization of treatment, statisticians can provide specific data which clinicians and program administrators can relate to expectations or goals; and the means by which the inter-program comparison can be made. Statistical assessments of treatment programs must consider the impact of readmissions on treatment programs in terms of the analyses described earlier.

Cost effectiveness estimates are also affected by the problem of multiple treatment events for individuals not being brought together. Schwartz and Epps have emphasized the implications for cost-effectiveness assessments, of easy readmission and involvement of individuals in multiple programs. During the course of an individual's illness, numerous, short contacts in diverse treatment services and programs can exaggerate cost-effectiveness. The patient load reported by individual programs increased while cost per "illness-episode" decreased: when, in actual fact, there may have been no increase in program contacts with individuals during the year and no changes in costs per individual illness.

By itself, the Resource Absorption Index does not indicate the effectiveness of the treatment program. However, treatment may have a favorable outcome with a majority of patients while recidivism of a minority absorbs so much treatment resources that clinical attention is diverted from those who might benefit most.

PROGRAM EFFECTIVENESS: OUTCOME FOR THE MAJORITY OF INDIVIDUALS

PROGRAM RECIDIVISM (Repeated episodes of care for a minority)		*FAVORABLE*	*UNFAVORABLE*
	HIGH	A	C
	LOW	B	D

Programs can be effective for the majority while recidivism is high (Cell A) or programs can be ineffective and recidivism be low (Cell D). Program evaluation, in addition to considering outcome, must also assess the extent of recidivism in the use of treatment resources.

Various quantitative methods can be used to show the existence of resource absorption; to measure its extent and to monitor changes which might result from modification of admission and treatment procedures. Measures of resource absorption are an essential part of statistical reports of utilization; of assessments of cost-effectiveness and evaluation of the effectiveness of alcoholism treatment programs.

SUMMARY

The use of treatment services is unevenly distributed among patients. A small number of patients use a disproportionately large amount of treatment resources. Another group of patients have relatively little contact with the treatment program. Resource absorption is a critical problem in treatment programs for alcoholism. This paper describes the application of two graphic displays of resource absorption to alcoholism treatment programs, the Lorenz curve and the Resource Absorption Index.

The Resource Absorption Index (RAI), projects the use of program resources by readmissions. This new measure is derived from time-specific rates of readmission for discharges with specific numbers of previous hospitalizations. The RAI measures the number of readmissions generated in a hypothetical new program over successive time periods, and, shows

the progressive reduction in accomodation (or "silting up") for first admissions. This index has been calculated for alcoholic disorders in specific treatment programs, and from national data for Canadian psychiatric institutions.

These statistical measures are graphic, readily"grasped" and relevant for policy making, program planning and program management.

REFERENCES

Baekland F, Lundwall L and Kissen B: Methods for the treatment of chronic alcoholism: A critical appraisal. Chapter 7 in Research Advances in Alcohol and Drug Problems. Volume Two. (RJ Gibbins, et al eds.) New York: John Wiley and Sons, 1974.

Ludwig AM: quoted in Physician's Alcohol Newsletter (American Medical Society on Alcoholism) 6(3), 5, 1971.

Moon LE and Patton RE: First admissions and readmissions to New York State Mental hospitals - A statistical evaluation, Psychiatric Quarterly 39, 1-11, 1965.

Pattison EM: Rehabilitation of the chronic alcoholic, Chapter 17 in The Biology of Alcoholism. Vol. 3, (B Kissen and H Begleiter eds.) New York: Plenum Press, 1973.

Redick RW: Utilization of mental health facilities by persons diagnosed with alcohol disorders. DHEW Publication NO. HSM 73-9114, Washington, USGPO, 1972.

Richman A: Estimating bed needs for detoxification from alcohol, presented at the National Center for Health Statistics Second Annual Data Use Conference, Dallas Texas, March 28-30, 1977, in press Proceedings.

Richman A: Estimating readmission rates by inferential linking of reported hospitalizations - a surrogate for a case register, in preparation, September 1977.

Richman A and Dunham HW: Consciousness of kind, submitted for publication, April 1977.

Richman A and Smart RG: "Recidivism" in alcohol treatment programs. Abstract Dec. 1, 1976.

Schwartz DA and Epps LD: The numbers myth in community mental health. Psychiatric Quarterly, 48: 320-326, 1974.

Siegel C and Goodman AB: An evaluative paradigm for community mental health centers using an automated data system. Community Mental Health Journal 12: 215-227, 1976.

Wilkins LT: Recidivists and recidivism. Chapter 4 in Evaluation of Penal Measures. New York: Random House, 1969.

* * * *

The assistance of Tanya Dubrow, David Ross Richman and Drs. J.L. Hedlund, D.G. Patterson and Mr. A. Weinstein is acknowledged.

DISTRIBUTION OF TREATMENT EVENTS

FOR INDIVIDUAL ALCOHOLICS

TABLE 1

NUMBER OF ADMISSIONS	NUMBER OF PERSONS U.S. Number of people receiving detox services since 1973 in 42 NIAAA funded centers			NUMBER OF PERSONS MISSOURI In-patient discharges with diagnosis of alcoholism Jan. 1970-Nov. 1974		
1	50,457	=	80.3%	10,340	=	66.4%
2	7,005	=	11.1%	2,556	=	16.4%
3	1,793	=	2.8%	1,139	=	7.3%
4	968	=	1.5%	565	=	3.6%
5	624	=	1.0%	332	=	2.1%
6	537	=	0.8%	183	=	1.2%
7 - 10	1,076	=	1.7%	308	=	2.0%
11 - 15	255	=	0.4%	93	=	0.6%
16 - 20	81	=	0.1%	23	=	0.1%
Greater than 20	77	=	0.1%	37	=	0.2%
TOTAL:	62,873	=	99.8%	15,577	=	99.9%

SOURCE: NIAAA - Program Analysis and Evaluation Branch, Dec. 1976.
E.D. Bode, STD and J.L. Hedlund, Ph.D., Missouri Division of Alcoholism and Drug Abuse; Missouri Institute of Psychiatry .

FIG. I

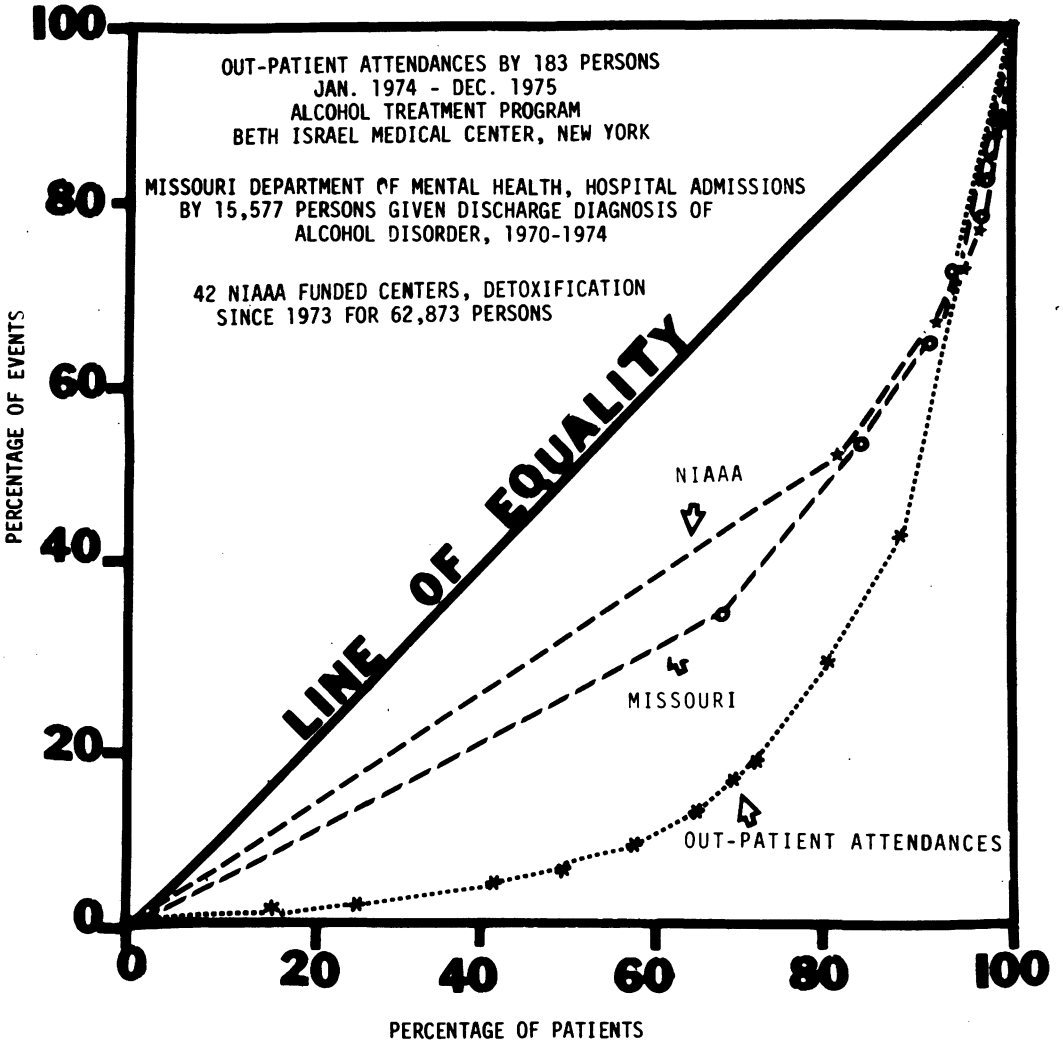

OUT-PATIENT ATTENDANCES BY 183 PERSONS
JAN. 1974 - DEC. 1975
ALCOHOL TREATMENT PROGRAM
BETH ISRAEL MEDICAL CENTER, NEW YORK
MISSOURI DEPARTMENT OF MENTAL HEALTH, HOSPITAL ADMISSIONS
BY 15,577 PERSONS GIVEN DISCHARGE DIAGNOSIS OF
ALCOHOL DISORDER, 1970-1974
42 NIAAA FUNDED CENTERS, DETOXIFICATION
SINCE 1973 FOR 62,873 PERSONS
LINE OF EQUALITY
NIAAA
MISSOURI
OUT-PATIENT ATTENDANCES
PERCENTAGE OF EVENTS
PERCENTAGE OF PATIENTS
0
20
40
60
80
100

FIG. II

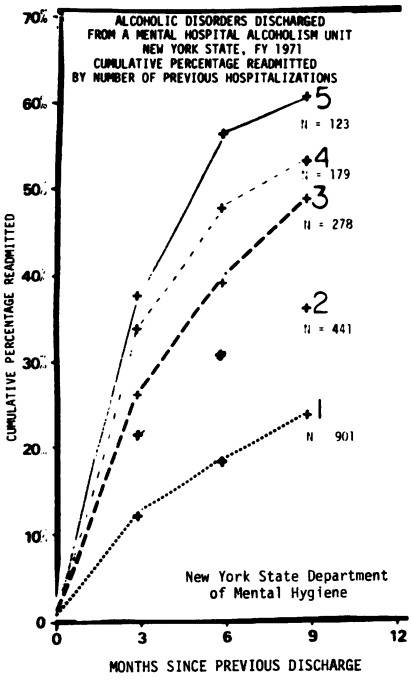


FIG. III

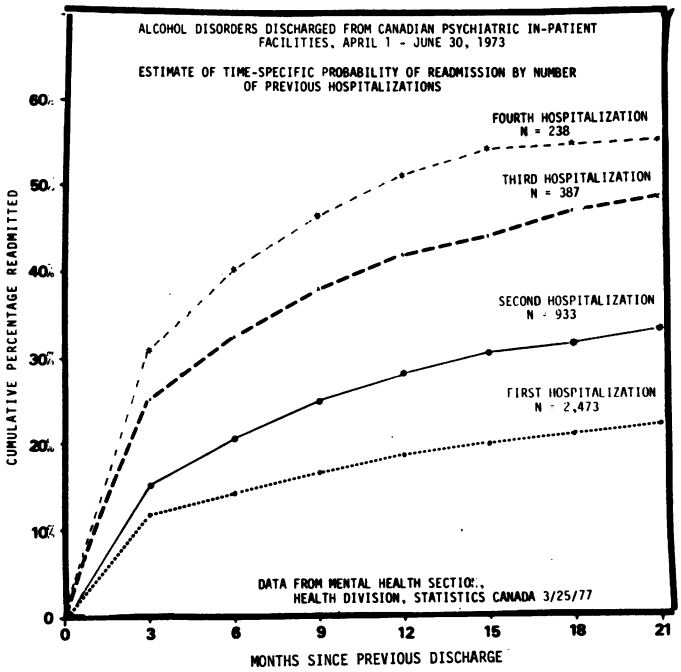
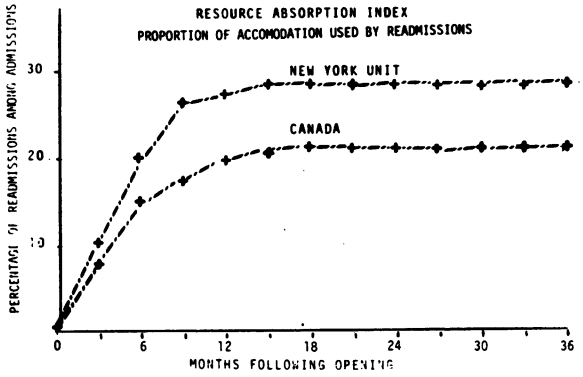


FIGURE IV



PSYCHOLOGICAL CORRELATES OF ATTITUDE TOWARD HEALTH AND PAIN: A VALIDATION STUDY

Phyllis R. Sweet Kathleen M. Johnson University of Lowell

METHOD

This study is part of an effort aimed at developing a reliable and valid paper and pencil test of attitude toward health and pain. The study is correlative in nature involving the administration, correlation and analysis of data from a series of survey instruments.

Four questionnaires were administered in two sessions to 120 students. Out of one hundred and twenty sets of questionnaires, only fifty-five were complete.

Sample. Fifty-five undergraduate students, enrolled in either psychology or sociology courses at the University of Lowell, completed all inventories used in this study. (One hundred twenty students completed some of the instruments used). Only 14.5% students were psychology or sociology majors. None of the students had undergone any psychiatric treatment. There are 23.7% females and 76.3% males. The average age is 20 years, 5 months. Foreign students made up 5.5% of the sample.

INVENTORIES USED

Personality Instruments. Two personality inventories were used: (1) The Edwards Personal Preference Schedule, (EPPS) and (2) selected scales of the Minnesota Multi-Phasic Personality Inventory, (MMPI). Both personality inventories are fixed-alternative questionnaires. The EPPS consists of 220 questions yielding 16 scales. The inventory derived from the MMPI consisted of 171 questions yielding 10 scales.

Pain Perception Inventory. A two page, 30 item index was used to measure the perception of pain and related health attitudes. This form of the 30 forced choice items inventory, yielded two derived scales described on the basis of apparent meaning, as follows:

(1) The Intellectualization-Suppression (I-S) Scale contrasts the respondent's belief in regard to the cause of pain with expected or experienced behavioral manifestations of the individual when faced with pain. High scores on this scale represent a tendency toward projection and intellectualization in the analysis by the respondent of the causality of pain. Low scores reveal a tendency to deny to act in a dependent manner and/or to isolate oneself when in a painful condition.

(2) The Anti-Professional Attitude (APA) Scale appears to reflect antagonism toward professional help. The high scoring individual exhibits chronic complaint behavior, sees no need to seek help for emotional pain and does not like doctors.

Personal Information Survey. Each subject was asked to fill out an optional personal information survey. From this sheet, information pertaining to age, sex, family income, extent of psychiatric care, and academic major was ascertained. Fifty-three respondents completed this form.

Hypotheses. Through analysis of prior studies and content analysis of the personality instruments used, a series of 151 hypotheses were developed. These hypotheses were intended as an exploratory test of the construct validity of the PPI-Form II as a paper and pencil measure of the perception of pain and of attitudes toward health.

These hypotheses were tested by correlational analysis. Pearson correlations, omitting missing data, were performed on all data from the 55 subjects. Sixty-eight null hypotheses were upheld.

TABLE I

REJECTED NULL HYPOTHESES RELATING PERSONALITY VARIABLES AND PPI ITEMS AND SCALES
($r \geq \pm 0.30$)

PERSONALITY VARIABLES	PPI ITEMS/SCALES NUMBER	PPI ITEMS/SCALES STATEMENT	PEARSONIAN CORRELATION(r=)	PROBABILITY LEVEL (p=)
EPPS-Intraception	P5	Pain may be a physical sensation.	-.377	.002
	P18	Certain types of pain are a physical sensation.	-.353	.004
EPPS-Abasement	P13	Pain is not only physical.	0.336	.006
	P25	I would never seek help for emotional pain.	-.306	.011
EPPS-Achievement	P28	Pain is only mental.	-.299	.013
EPPS-Nurturance	P16	Pain may be an emotional sensation.	0.334	.005
EPPS-Affiliation	P16	Pain may be an emotional sensation.	0.448	.001
EPPS-Aggression	P16	Pain may be an emotional sensation.	-.314	.010
	P25	I would never seek help for emotional pain.	0.410	.001
EPPS-Exhibition	P23	I tend to complain about even the smallest ache or pain.	0.297	.014
EPPS-Order	P3	People usually cause physical but not emotional pain.	0.408	.001
EPPS-Consistency	P12	I only seek help for check-ups.	-.333	.006
	P13	Pain is not only physical.	-.347	.005
	P24	Pain may be a physical and emotional sensation.	0.318	.009
MMPI - L Scale	P29	Pressure to succeed can cause emotional pain.	-.311	.010
MMPI - K Scale	P29	Pressure to succeed can cause emotional pain.	-.442	.001
MMPI - F Scale	P8	I have had no experience with physical pain.	0.310	.013
	P23	I tend to complain about even the smallest ache or pain.	0.298	.014
	P27	I would never seek help for physical pain.	0.301	.013
MMPI-Hysteria (Hy)	P11	Pain cannot be endured	0.334	.006
	PPI-1	PPI-Form II-Scale 1	0.334	.006
MMPI-HS (K) Scale	P18	Certain types of pain are a physical sensation.	-.303	.012
	P29	Pressure to succeed can cause emotional pain.	-.362	.003
MMPI-Admission of Symptoms	PPI 1	PPI-Form II-Scale 1	0.326	.007
MMPI-Denial of Symptoms	P29	Pressure to succeed can cause emotional pain.	-.314	.010

RESULTS

Of the 83 null hypotheses which were rejected, 25 showed correlations of $\pm$ 0.30, relationships of moderate or greater magnitude, (see Table 1). The items which showed moderate correlations may be viewed as clusters in relation to the personality scales with which they correlate. This study may be viewed as reflecting the personality structure of people, who, although not in pain, are responding to questionnaires within the confines of "what if I were ill, in pain, etc."

The EPPS Scale, Intraception, relates to observation, analysis and understanding of others motives,feelings and problems. The magnitude and direction of the correlation of this scale with items 5 and 18 indicate respondents expand the meaning of pain beyond the physical. These items may thus be seen as reflecting the individual's empathetic understanding of the meaning of pain.

The EPPS scale, Abasement, reflects the tendency of the individual to accept blame and to feel guilty when things go wrong. The implication behind the relationship of this scale and items 13 and 25 seems to be that the respondent is accepting an emotional component to pain. Further, if one accepts the correlation as evidence of related meaning, it appears that the emotional component brings with it a feeling of shame or guilt which is not present if one denies that pain expands beyond a physical sensation.

The EPPS scale, Achievement, relates to the effort to do one's best, to accomplish tasks requiring skill and effort. Once again, there is a relationship to item 28, which denies the unidimensionality of the pain experience. Here, the high achiever denies that pain is only mental.

The EPPS scale, Nurturance, relates to being supportive of others. Affiliation relates to loyal participation in groups. Aggression involves attacking contrary viewpoints. These three scales are all correlated with item 16. The apparent meaning is that supportive friendly individuals accept the likelihood that pain may have emotional component. To the more aggressive individual, the emotional component of pain is denied, both in item 16 and in the additionally correlated item 25.

The EPPS scale, Exhibition, involves the need to have others notice one. The PPI item 23 involves the same need, within the health context. Like the trait, exhibitionism, the complaint behavior admitted can be too extensive.

The EPPS scale, Order, involves being neat and orderly and making plans in advance. This scale is correlated with item 3 indicating that the orderly individual tends to blame others for inflicting physical pain but not for the emotional aspects of pain.

The EPPS scale, Consistency, is a measurement device intended to determine the seriousness with which an individual completes the questionnaire. If we can extend this attitude from one instrument in a set to another, then consistency may also be measured by items 12 and 13 being answered positively. Thus, content aside, we can assume that there is a serious attempt by the respondent to relate information if these 3 items are answered according to the formula above.

The three validity scales of the MMPI, F, L and K Scales, represent additional measurement devices. High F Scale scores represent confused or careless answers. High L Scale scores represent the attempt to avoid answering frankly and honestly. High K Scale scores represent defensiveness in answering questions. Under the same reasoning as given above for the EPPS Consistency Scale, agreement with items 8, 23 and 27 represent confused or careless answers (see especially item 8). Disagreement with item 29 would represent both high L and K Scale scores showing an attempt to maintain privacy.

The MMPI Scale, Hy (Hysteria), is described as one in which persons who score highly use physical symptoms as a means of solving difficult conflicts or of avoiding mature responsibility. This scale correlates positively with item 11 and with the PPI I-S Scale. The content of both the item and the scale do project the picture of acting out when faced with the pain experience, as would be consistent with the described personality scale.

The MMPI Scale, Hs (Hypochondriasis) with K correction, represents persons with the stereotyped pictures of hypochondriasis. These individuals show abnormal concern for their bodily functions. They are egocentric and immature. The K correction controls for a person covering up for his true response. The correlations between this scale and items 18 and 29 indicate a generalized, physically based view of pain.

The MMPI derived Scale, Ad (Admission of Symptoms) is a scale derived from the Hy Scale and composed of all the items related to somatic complaints. The higher the score, the greater the number of complaints. This scale correlates positively with the PPI I-S Scale. Since it is a derivation of the Hy Scale, it is likely that the correlation between Scale 1 and Hy accounts for the correlation between PPI Scales and the Ad Scales. However, it should be noted that the Dn Scale (described below) was also derived from Hy and there is no significant relationship between Dn and the PPI I-S Scale. In either case, the construct measured by the Ad scale is consistent with the tendency of the individual scoring high on the PPI toward intellectualization in the form of somatic symptoms.

The MMPI derived Scale, Dn (Denial of Symptoms) is also derived from the Hy Scale, It correlates negatively with the Ad Scale. The higher an individual's score, the more denial or problems relating to inadequacies, basic self control and empathy. This scale is negatively correlated with item 29 indicating a denial of the emotional effects of pressure.

In all, significant moderately high correlations appear to support the construct validation of 14 items of the 30 item PPI as well as of the first factor scale. However, the constructs against which the items and scales are judged themselves modify the meanings assumed under conditions of face validity. These modifications are discussed below:

DISCUSSION

This attempt to validate the PPI - Form II as a paper and pencil test of attitude toward health and pain, may be considered as relatively successful.

Hypothetical relationships between various selected personality variables and both the items and scales of the PPI were tested. Of the initial 151 hypotheses, no relationship was found for 68 in this sample, using the selected personality instruments. Fifty-eight correlations reached the level of significance, but the magnitude of these correlations (less than $\pm$ 0.30) did not reach the standard chosen for validation purposes. The remaining 25 hypotheses are herein accepted as construct validation of the relevent items and scales.

An internal validity scale of 7 items can be identified. These items, 8, 12, 13, 23, 24, 27 and 29, with appropriate responses can be taken as a measure of a serious, open and careful attempt to respond to the questionnaire. Since further tests of these items have not been undertaken, their use as a validity scale must presently be of "all or nothing" nature. A score of 7 would yield a valid questionnaire, a score of 0 would yield an invalid one, all other scores are more or less questionable.

Ten items, including three which are also part of the validity scale, may be said to be construct validated. Nine of these items make up part of the 26 item PPI I-S Scale. The remaining item falls on the 4 item PPI APA Scale. As indicated above, these items, in the process of construct validation, are also subject to some modification of meaning.

Eight of the validated items, 3, 5, 11, 13, 16, 18, 28 and 29 are, on the face of it, concerned with definitional aspects of pain. However, this study expands this meaning to include a component best described as one which addresses the question, "How do you know?" For each of these items, respondents indicate, through their responses in regard to the personality variables, that they have learned through observation of others than pain is multidimensional. Their view of the pain experience is mediated by their experiences within groups as loyal and supportive members. There is further an evident ability to put oneself in the place of one feeling pain which joins with a kind of egocentrism that relates all experiences to oneself, thus helping to define the experience.

The remaining two items, 23 and 25 relate to the seeking of care for painful experiences. On the one hand, these items indicate a reticent, intrapunitive and submissive attitude about pain. On the other hand, the individual is shown to be exhibitionistic, extrapunitive, impatient and a demanding subject.

Finally, the meaning of the PPI I-S Scale is, through this study, somewhat modified from that given above. The PPI I-S Scale may be viewed as one in which the high scoring individual uses pain as a way of avoiding responsibility for his own actions. Additionally, the scale may reflect somatic complaints which, although not part of the overt content of the scale, appears to underline the scale.

This attempt at validation should be viewed as exploratory, limited and subject to questions relating to sample generalizability and selections of validating standards. The latter questions can be answered in reference to the literature. The scales selected are consistent with those used in the studies reviewed. The former question can only be addressed by a series of replicatory studies using different age, educational and geographic background at the least. Even with these reservations, this study appears to be an appropriate beginning.

THE DESIGN AND METHODOLOGY FOR THE NATIONAL MEDICAL CARE EXPENDITURE SURVEY

Julia D. Oliver,[1] National Center for Health Statistics

1. INTRODUCTION

1.1 Background

The rapid escalation of medical care costs and their concomitant economic and social implications has received more national attention in recent years than any other health care issue. Since World War II the cost of medical care has risen dramatically: total personal health care expenditures have increased more than sevenfold; third-party payments have increased fiftenfold,[2] while the cost of living has more than doubled.[3] In just the four-year period of 1965-1969, the daily cost of a hospital room increased by 69.7 percent.[4] The effect of these expenditures on the health service delivery system and on the 75 million households has far reaching implications for our economy and all our citizenry.

At the present time, speculation about enactment of national health insurance legislation and its social and economic consequences is a major factor in generating the increased attention to the need for reliable information and research on the use of health care services, the expenditures for these, and the payment sources employed. Furthermore, if national health insurance legislation is passed, relevant and valid data are needed to evaluate the effect of the legislation.

The Health Interview Survey (HIS), conducted by the National Center for Health Statistics (NCHS) since 1957, has been a major vehicle to obtain operational, policy related, and evaluative data about the health system and its substantive aspects. However, even before the HIS, considerable interest had been directed to the problem of measuring and assessing medical and health expenditures. Studies were instituted as early as 1933 by I. S. Falk.[5] In 1953 the National Opinion Research Center (NORC) in collaboration with the Health Information Foundation (HIF) began a series of studies that were continued in 1958, 1963, and 1970. Subsequent studies such as the Federal Employees Health Benefits Program Utilization Study in 1972-1973 and the continuing studies by the HIS in 1971, 1975, and 1976 have monitored and documented individual and family health services utilization and expenditures.

However, previous surveys on health and medical expenditures have had trouble achieving desired levels of completeness and accuracy of reporting. Problems with recall periods and underreporting or incorrect reporting from the households have been major obstacles to obtaining valid health cost data in household interview surveys. The household interview process, by itself, has been thought to be an inadequate mechanism to obtain accurate data on the total costs of medical care. Collecting supplementary data from medical care providers, third-party payers, and employers can ameliorate the problems resulting from underreporting and incorrect reporting. Even where households report information on medically attended illness, however, most households cannot report diagnosis accurately.

In 1975, in the interest of improving medical expense data, NCHS conducted a pilot study to determine the most feasible survey design for a national study of medical care utilization, expenditures, and health insurance coverage.[6] During the same period, research on medical care utilization and expenditures was being conducted by the National Center for Health Services Research (NCHSR). In its program of intramural research, special emphasis was being given to the impact of possible changes in Federal income tax treatment of medical care expenses; the costs and benefits of various national health insurance proposals; the impact of Medicare and Medicaid programs on various subgroups of the population such as the poor and rural citizens; the impact of existing Federal health programs on access to care; the costs of illness for American families by diagnostic category and the impact of such costs on these families.

Given the shared interest in health expenditure data, NCHSR and NCHS decided to jointly sponsor the National Medical Care Expenditure Survey (NMCES). The principal funds for this research are provided through the Division of Intramural Research, NCHSR.

In June of 1976 a contract was awarded to the Research Triangle Institute (RTI), Research Triangle Park, North Carolina.[7] Dr. Daniel G. Horvitz, Vice President of RTI's Statistical Sciences Group, is project director. RTI has awarded subcontracts for components of the data collection to Abt Associates, Inc. (AAI), Cambridge, Massachusetts, and to NORC of the University of Chicago. RTI as the prime contractor has overall responsibility for the management of the contract. In conducting the survey, RTI and NORC shared responsibility for collecting the data. Each organization used their own probability samples and supervised their own field staff.

2. OVERALL DESIGN OF THE NMCES

NMCES is, in reality, three surveys: a survey of households, a survey of those medical providers and hospitals who provided services to the sampled population, and a survey of health insurers/employers who provided health insurance for them. Data on conditions, utilization and expenditures are being collected from about 11,500 households, oversampled for the uninsured population, during calendar year 1977.

The surveys of the medical providers and health insurers will begin in 1978. Questionnaires will be mailed to physicians, hospitals, and insurance companies or employers for each NMCES respondent who has signed a permission form authorizing the release of data.

2.1 Overall Design of the Household Survey

The original study design called for a national probability sample of the civilian noninstitutionalized population in the fifty States and the District of Columbia residing in 10,000 households. These households were to be interviewed 8 times--once every 9 weeks over a period of 16 months about medical care received during the 1977 calendar year. The first two interviews are to be in person, the next four by telephone, the seventh in person, and a final contact by telephone in April 1978. A diary was to be left with respondents and a computer generated summary was to be mailed to the households between interviewing rounds.

2.1.1 Sampling Protocol

Similar area probability designs were used by both RTI and NORC for the national survey. Both utilize a stratified multi-stage area probability sample. Subsets or half-samples of the full complement of Primary Sampling Units (PSU's) were selected from both organizations' national general purpose samples. The RTI and NORC half-samples contain 59 and 76 PSU's respectively and each represents the United States. Together the two independent samples overalp so that 106 separate locations are sampled. The second stage sampling units are large clusters of households (60 or more for RTI and 100 or more for NORC) from which 12 households were subsampled.

Given the analytical goals of the study, it was decided that the families and individuals identified for the sample in the initial round would be followed throughout the year even if they moved, unless they died, entered the armed forces, entered an institution, or left the country. Any new individuals who join the initial set of families or who join with a sampled individual to form a new family will be included for purposes of analyzing families but will not be included in any analysis of individuals.[8]

The sample was originally designed to provide eight rounds of data on 10,000 households. At the request of HEW's Assistant Secretary for Planning and Evaluation, this sample was supplemented with an oversample of 1,500 households which contained one or more persons without any health insurance. This was accomplished by selecting more households in the segments with a higher expected proportion of uninsured households. In order to produce a final sample of 11,500, an initial sample of approximately 16,000 households was selected to allow for an overall attrition of about 25 percent due to (1) unusable selected sampling units (e.g., vacant or demolished housing units), (2) unavailable respondents 19 years of age or older, (3) not-at-homes, (4) refusals and interview breakoffs, and (5) dropouts before end of all eight interviews.

2.1.2 Questionnaire Design

The questionnaire includes both items that are collected on special supplements and a repetitive core of questions asked on each round and on special one-time supplements. The core questions include the number of bed days, restricted activity days, hospital admissions, medical and dental out-patient visits, other medical care encounters, prescribed medicines, and coverage by private or public health insurance plans such as Medicaid. For each contact with the medical care system, data are obtained on health conditions, the characteristics of the provider, the services provided, the charges and methods of payment. Questions on the special one-time supplements include detailed data on health insurance coverage, access to medical care, limitation of activities, employment, income history, and socio-demographic characteristics.

2.1.3 Memory Aids

One of the ways NMCES addresses the problem of response error in the household survey is with the use of memory aids. A calendar/diary and a device called the summary have been developed for use in the household survey.

The calendar/diary is designed as a memory aid as opposed to a data collection instrument. It was developed and successfully used in the pilot study. It is left at the household at the end of the first personal interview. The field interviewer instructs the respondent in the use of the calender/diary and tells her that she will be asked to refer to it during the next interview. Instructions and examples are printed on the diary itself.

The diary consists of a monthly calendar with numbered date squares. Directly underneath the calendar is a ledger with boxes provided for each type of utilization reported: prescribed medicines, doctors, dentists, and other medical persons, hospitals, care in a nursing home or other similar place, and other medical items. It further identifies the reported utilization by household member, date obtained, name of provider or description of illness or injury and the cost. Beneath the ledger is a pocket with the printed instruction "Keep your medical bills in here."

The summary is a computer generated document that presents in a standard format the cumulative utilization and expenditures reported in each household for each round.

Data on utilization and expenditures collected during the household interview is keyed, processed through the computer and mailed to the household before the next interview.

The summary has separate computer printed pages for each household member. Each page contains a section for each type of utilization data collected: medical provider, hospital, prescription medicine, and other medical expenses. For each event of utilization the date obtained, the provider's name or source of care, the costs and source of payment are presented.

During the next interview, anywhere from 9 to 15 weeks later, the field interviewer systematically reviews and verifies the information with the

household. The interviewer enters any corrections, such as changes in dollar amounts or unreported visits and returns the corrected summary to RTI for rekeying and reprocessing to update the summary prior to the next interview.

Although, the primary purpose of the summary is to improve household reporting by collecting information that was forgotten or not available at the time of the survey, it has also served to legitimize the survey to respondents and to correct interviewer, respondent and processing errors.

The summary has been designed as an adjunct to the main questionnaire. Both flat fee and health insurance data from the summary are used in followup interviews if the reporting unit has reported either flat fees or health insurance plans/programs in previous interviews.

2.1.4 Achieving Desired Response Levels

In order to achieve the goal of a 95 percent response rate for the household survey and an 85 percent overall completion rate, it was decided to pay participating households a $20 incentive fee in addition to agressively pursuing traditional conversion procedures. Respondents are paid $5 after completing each of the first two rounds and $10 after the seventh. The conversion procedures include making calls at different times of the day for not-at-homes, making appointments at unusual hours for unavailables, and the use of supervisors as telephone converters for refusals.

2.2 Design of the Medical Provider Survey

The Medical Provider Survey is designed to obtain information on diagnosis, utilization, third-party payment, and total charges, and verification of payment from (1) hospitals and other institutional medical/health care facilities reported as utilized by sample individuals, and (2) physicians and doctors of osteopathy. It is necessary to ask providers of care for this information because in most instances respondents simply cannot provide charge and diagnostic information at the level of detail and accuracy necessary for our analysis plans. Data collection from providers of medical care such as dentists, nurses, chiropractors, etc., is not planned. Health insurance rarely directly covers the services of such providers and, hence, the costs of such care are usually paid directly and, therefore, known by patients. The data collection from doctors and hospitals will serve not only as a means for obtaining the required data but will also permit an assessment of the accuracy of household respondent reports on the utilization of medical care. Both household respondent over- and underreporting will be assessed using the data.

2.2.1 Physician Data Collection

The data element definitions and format for the physician data collection were developed in light of the pilot study and extensive consultation with the American Medical Association and knowledgeable individuals. In order to be responsive to the complexities of the medical environment, it was necessary to develop four separate data collection forms: the Medical Provider Questionnaire--the basic data collection instrument--and the Pregnancy Related Visit Form, the Inpatient Related Care Form, and the Repetitive Visit Form. These forms are designed to verify dates of visits reported by households and to gather detailed diganostic care information and the costs associated with that care. The Pregnancy Related Form, the Inpatient Related Care Form, and the Repetitive Visit Form are all designed to deal with the particular data collection complexities in each particular situation.

2.2.2 Hospital Data Collection

The design of these forms was influenced primarily by the experience of the pilot study and by the advice of outside consultants. Extensive consultation with the technical and research staff of the American Hospital Association and knowledgeable individuals in the hospital industry was undertakn on the format and data element definition.

Since in most hospitals there is not one central source for obtaining both cost and diagnostic data, the forms have been designed so that the cost data can be collected from a hospital accounting or billing department--the Financial Information Questionnaire--and the diagnostic data can be collected from the hospital medical records department--the Medical Information Questionnaire.

The Medical Information Questionnaire obtains the following basic information: (1) verification of the reported admission and discharge dates, (2) the respondent's chief complaint at the time of admission, (3) discharge diagnostic data including the H-ICDA or ICDA code, (4) operational and diagnostic procedural data, and (5) the names of physicians who provided care.

The Financial Information Questionnaire obtains information on the hospital charges by type for the admission, the amount of the bill already collected and the source of payment, the expected amounts of remaining payments and the sources of such payment. The questionnaire also includes a question that addresses the discounting of payments to hospitals.

2.3 The Health Insurer/Employer Survey

The insurance survey is designed to collect additional information on health insurance coverage, benefits, and premiums from two possible sources: (1) the employer, union, or group carrying the insurance; or (2) from the insurance company providing the coverage. Given the cost and complexity of collecting claims data in the 1970 Andersen survey and pilot study, it was decided not to collect claims data.

Detailed insurance coverage information is a crucial component of this study for several reasons. First, the analytical plans to simulate national health insurance proposals cannot be carried out without details of coverage, limits,

etc., which the household respondent cannot normally provide. Second, the information obtained from third-party payers will permit a comparison of how people perceive their coverage vs. their actual coverage and an assessment of the correlation of these measures with utilization. Third, premiums are a major component of medical costs each year. Both the pilot study and the pretest indicated that a majority of household respondents cannot provide accurately the amount of the premium paid by employers or other sources. The employer or insurance company is also the best source of information about premiums.

3. CURRENT STATUS OF THE SURVEY

3.1 Household

A scaled down version of the national study was pretested in Charlotte, North Carolina and Dayton, Ohio in the fall of 1976. Round 1 of the national household study began on January 17, 1977 and ended on April 1, 1977. The response rate for Round 1 was about 93 percent.

Round 2 of the national study was completed in June 1977. Preliminary results indicate that 96 percent of the households completing Round 1 completed Round 2.

During Round 1 it became obvious that processing and turning around the data for over 14,000 families within 9 weeks was next to impossible. This problem forced the delay of the start of Round 2 for three weeks and would consequently delay the start of every succeeding round, as well as start of the medical provider and health insurance surveys.

Given the necessity of keeping the survey on schedule and of collecting data for calendar year 1977, two telephone interviews were eliminated. These telephone interviews consisted only of the main questionnaire and contained no special supplements. This effectively extends the recall period between interviews, but the use of the summary and the use of the calendar/diary is expected to offset some of the memory problems.

3.1.1 Uninsured Sample

Of the 1,500 additional households added to the sample to increase the uninsured segment of the sample approximately 800 households turned out to actually have one or more household members who were uninsured. The 700 insured households were dropped from the survey after the first interview to reduce costs since they could not be used to offset losses due to attrition.

3.2 Current-Status--Medical Provider and Hospital Survey

The physician and hospital pretest began on September 17, 1977. Althogether 215 unique physicians and 5 hospitals were identified in the Charlotte pretest.

The physicians and hospitals that are being contacted have been identified from signed permission forms obtained during the household pretest interview carried out in Charlotte during the fall of 1976. Each adult and minor age 14 and over was asked to sign a permission form authorizing the NMCES contractor to obtain additional data from physicians and hospitals on medical care and cost. For minors under age 14, the parent or guardian was asked to sign.

An attempt was made to obtain signed permission forms for every physician who provided medical care during the pretest. In addition, signed forms were also sought for those physicians who were identified as a regular source of care in the "Access to Medical Care Supplement," even if no visits to the regular provider were mentioned during the pretest. This procedure was followed so that some measure of the extent of unreported visits could be obtained. Also, all respondents who reported a hospital stay of at least one night were asked to sign permission forms. Five of the 169 Charlotte household pretest families refused to sign permission forms: these five families reported eight doctor visits and no hospital admissions.

During the medical provider pretest, the forms are mailed to providers. Three to four weeks after mailing the forms, field interviewers are to telephone the medical providers. Depending on the situation, the field interviewer will then (1) arrange to pick up the forms, (2) remind the medical providers to pick up the forms and then arrange pickup, (3) offer to assist the provider in filling out the forms, or (4) in limited cases, where there is a small case load, offer to conduct the interview over the telephone.

The procedures for the Hospital Survey are identical to those developed for the medical provider surveys.

The medical provider/hospital surveys will end on October 21, 1977.

3.3 Health Insurer/Employer Survey

During the Charlotte household pretest, 41 unique insurers/employers were identified. For each health insurance plan reported in the pretest, the policyholder was asked, in Round 2, to sign a form authorizing the NMCES data collection contractor to obtain additional information from the employer/third-party payer. For all policies carried through employer, unions, or some other groups for which authorization was obtained, the employer/group was mailed a copy of the combined authorization form and questionnaire. If the employer or other groups was unable or unwilling to complete the questionnaire, the insurance company was contacted.

For individual policies for which signed authorization has been obtained, the insurance companies are being mailed a copy of the

combined authorization form and questionnaire directly. All providers of the insurance coverage are being asked to provide a copy of the policyholder's (or respondent's) insurance policy. These policies are being coded at AAI to test the codebook procedures. For contacts, whether with employers or other groups or with insurance companies, who did not respond to the first mailing, the study design calls for a follow-up effort consisting of two mailings, two telephone calls, and finally a personal visit from a field representative to conduct a face-to-face interview. However, preliminary results indicate that a high level of response was obtained without the use of the full regimen of follow-up procedures.

4. CONCLUDING REMARKS

Preliminary data from the NMCES will be published as soon as they become available in the form of joint publications from the National Center for Health Statistics and the National Center for Health Services Research. It is anticipated that some preliminary data will be available in 1978.

FOOTNOTES

1. The bulk of this report is drawn from internal working documents from the National Medical Care Expenditure Survey prepared by the staff of the National Center for Health Statistics and the National Center for Health Services Research, RTI, NORC, and Abt Associates, Inc. Dr. Daniel C. Walden, Mr. Robert A. Wright and Dr. Gordon S. Bonham reviewed the document and made many helpful comments.

2. Worthington, N. L. "National Health Expenditures, 1929-1974." Social Security Bulletin, 38, No. 2 (1975), p. 3.

3. U.S. Department of Commerce, Bureau of Economic Analysis. Survey of Current Business, May 1975, and U.S. Department of Commerce, Bureau of Economic Analysis. Business Statistics, 1975.

4. Jones, Sidney L. "Measuring National Wealth and Well Being." Included in Technical Paper No. 37 of the Bureau of the Census, June 1975.

5. Falk, I. S., et al. "The Incidence of Illness and the Receipt of Medical Care among Representative Families." Committee on the Cost of Medical Care, Publication No. 26 (Chicago: University of Chicago Press, 1933).

6. For a detailed description of the interviewing results of the pilot study, see Shapiro, S., Yaffe, R., Fuchsberg, R., and Corpeno, H. "Medical Economics Survey--Methods Study," a paper presented at the American Public Health Association, Chicago, Illinois, November 1975. This research was supported through Contract No. HRA 106-74-150 from the National Center for Health Statistics, Health Resources Administration, DHEW.

7. Contract No. HRA 230-76-0268 from the National Center for Health Services Research, Health Resources Administration, DHEW.

8. Additional information on the sampling design is available on request.

VALIDATION OF INFORMATION WITH THE HOUSEHOLD IN A PANEL STUDY OF MEDICAL CARE EXPENDITURES

Michele C. Gerzowski, National Center for Health Statistics

1. INTRODUCTION

A major problem in survey research is the effect of response errors on the accuracy and completeness of information obtained from household interviews. One component of response errors unique to studies on the cost of medical care is due to the complex structure of medical care expenses in this country. This kind of response error can occur in cases where even the most cooperative household respondents who keep meticulous records may not be able to answer questions on the cost of their medical care, because health insurance payments, reimbursements, and the like may lag months after a medical event. It might, therefore, take many household interviews to resolve the cost of a medical visit.

The National Medical Care Expenditure Survey (NMCES) is sponsored by the United States Public Health Service under the joint auspices of the National Center for Health Statistics and the National Center for Health Services Research. A probability sample of 13,500 households was selected for the NMCES so as to represent the civilian noninstitutionalized population of the United States.

The survey is being conducted by the Research Triangle Institute of North Carolina, in conjunction with its two subcontractors, the National Opinion Research Center of the University of Chicago and Abt Associates, Inc. of Cambridge, Massachusetts.

The NMCES and its pilot study, the Medical Economics Research Study (MERS), both addressed the problem of response error in panel studies with the use of a memory aid called the household summary. The household summary is a computer printout containing information on visits to medical providers from previous panel interviews and expenses associated with those visits. It also includes information on hospitalizations, prescribed medicines and miscellaneous medical expenses such as crutches and eyeglasses. The summary is used to verify and to check for completeness the information previously reported by the household.

This paper addresses the use of the household summary during the NMCES pretest conducted in 1976.

2. REVIEW OF RELATED LITERATURE

An example of a memory aid similar to the household summary is used in the Current Medicare Survey (CMS) administered by the Social Security Administration. The CMS provides information which can be used to produce national estimates on the kinds and costs of medical services used by Medicare subscribers.

The study consists of 16 interviews, some of which are personal and some of which are conducted by telephone, each of which is separated by an interval of one month. After every interview a document entitled the "Followup of Estimates and Omissions" is mailed to the interviewer prior to the next interview if necessary. The "Followup of Estimates and Omissions" is generated whenever estimates or "don't knows" were given as answers to questions on medical charges and/or whenever an item was omitted during a prior interview. During the next interview, the interviewer asks about the omitted and unknown information. If the household respondent does not know an answer after two more interviews, the questions are not asked again.

The only other example of an aid resembling the household summary found in the literature, was used in the MERS, the pilot for the NMCES. That household summary was very similar in content to the one used for the NMCES pretest. The final results of the pilot study have not, as of yet, been published.[1] However, preliminary results indicated that the use of a summary improved the quality of household data thus proving to be a feasible memory aid in studies on the cost of medical care.

3. METHODOLOGY OF THE NATIONAL MEDICAL CARE EXPENDITURE SURVEY PRETEST

The NMCES pretest was designed to test procedures for collecting information on medical care expenditures from household respondents and their medical providers. In the household portion of the pretest, all aspects of the national study were pretested, including the basic questionnaire, procedures for administering the household summary, the use of a diary for recording medical events, and the collection of signed permission forms for the Medical Provider Survey (MPS).

The household pretest was a two round panel study administered in two sites--Dayton, Ohio and Charlotte, North Carolina in the fall of 1976. All household interviews were conducted in person.

The MPS pretest began in August 1977 and is scheduled to end in October 1977. Those physicians and hospitals for whom a signed permission form has been obtained from household respondents will be asked questions about medical care which they provided. The medical providers will be contacted initially by mail, with telephone and personal followup procedures used whenever necessary.

The NMCES began on a national level in January 1977. A description of all aspects of the entire NMCES and pretest is given by Julia D. Oliver at these proceedings.[2]

In the NMCES pretest, approximately 160 households in each pretest site of Dayton, Ohio and Charlotte, North Carolina participated.

During the first interview, questions were asked about disability days, days lost from work due to illness, the costs and services provided during hospitalizations and visits to physicians, dentists, and all other medical providers, the

costs of prescribed medicines and other medical expenses (e.g., crutches and eyeglasses) and health insurance coverage.

After this first household interview, the respondents were informed that they would be contacted again for another interview in 8 weeks. Before the next interview, they were told that they would receive in the mail a computerized summary containing information on medical costs and visits that they had just reported in the first interview. In addition, they were instructed on the use of a diary to record medical expenses incurred before the next interview.

Data from the first interview were coded, keyed and edited before being generated into household summaries. Two copies of each summary were produced--one copy for the household respondent and one copy for the interviewer. Also included in the packet mailed to the respondent was a letter explaining how to read the computerized format. An example of the household summary is shown in Illustration 1.

After the administration of the questionnaire during the second interview, the summary was reviewed jointly by both the respondent and the interviewer. In most cases, the same person was the respondent for both interviews.

All interviewers were instructed to specifically ask about error codes appearing on the summary. Error codes (i.e., "not known," "not available") were programmed onto the summary whenever missing, illegible, or outrageous information, such as a $50,000 dental bill, were reported, during the first interview. In the course of the summary review, the interviewer would ask the respondent if that information was now available. Some interviewers reviewed the summary line by line, asking if each item was correct. Others reviewed only the error codes specifically and then asked respondents if the rest of the summary was correct.

Corrections were made on the interviewers' copy of the summary, which was returned to data processing. Previously unreported information, questions which were originally misunderstood or omitted, incorrect answers given by proxy respondents, keying, coding, and interviewer errors were reconciled on the summary. In this way, information from the first interview was validated and checked for completeness with the household respondent.

The summary review only took a few minutes on the average and never exceeded ten minutes. Respondents reacted favorably to the summary, in general, commenting that seeing their own data made the survey seem more legitimate and worthwhile.

Many respondents had difficulty in reading and understanding the computerized format, but most understood the summary after it was explained to them by the interviewer. The summary was read to those respondents who were illiterate, had poor eyesight, or who were, for some other reason, unable to review the summary by themselves.

During the pretest, the summary was only reviewed once. The summary is being reviewed several times during the national study, during every interview after the initial one. The summary in the national study will be cumulative, with the last household summary containing information from all prior interviews.

4. FINDINGS

The analysis is primarily concerned with the number and types of changes made regarding information from the first interview and how these changes affect out-of-pocket expenses for medical care.

The variables which were analyzed for changes on the summary were: (1) name, address, source of payment, amount paid by each source of payment, and total charge for visits to physicians, dentist and other medical providers (not M.D.'s) and hospitalizations; (2) name, source of payment, amount paid by each source of payment and total charge for prescribed medicines and other medical expenses (i.e., eye glasses, crutches).

The variables which were not analyzed include date of visit and demographic information. Changes in the date of a medical event were far too few to warrant any analysis. If the respondent knew a date during the first interview, it was rarely changed during the summary review 8 weeks later. Similarly, if a respondent did not know the date during the earlier interview, it was unlikely to be remembered later.

Demographic information such as education, income and race was collected during the second interview, but was not available for this analysis However, the name, address, age and sex of each respondent was printed onto the summary and corrected if necessary. The number of these changes was insignificant, but the effect of seeing an incorrect name, age or sex, appearing on the summary, disturbed the respondents more than anything else.

The results of the summary review were separated into two main groups: those entries which were changed and those entries which remained the same. The changes were then regrouped into reports of new visits (additions), deletions of previously reported visits, updating of information (corrections) and changes caused by previous omissions of questions. The latter referred to a built-in feature of the questionnaire where if a bill was to be expected, questions about the source or amount of payment were not asked.

Those items which remained the same were classified into two categories, those items which the respondent knew in the first interview and verified 8 weeks later and those items which the respondent did not know during the first interview and still did not know at the time of the summary review. The number of changes reflect the impact of the summary review.

Table 1 gives the frequency of different medical events which occurred during the first interview of the pretest and combines results for both Dayton and Charlotte.

Table 2 shows the number and percentage of changes made on the summary for certain key variables in both pretest sites. Changes were highest for those variables involving dollar amounts, "amount of payment for each source" (19.7 percent) and "total charge" (21.0 percent). The fewest number of changes were made for names (7.5 percent) of medical providers or items, such as the name of a prescribed medicine and the address of medical providers (6.0 percent). Most of these changes were corrections in spelling.

Another way of looking at changes on the summary is to view the number of charges per line of data. A line of data on the summary closely approximates a medical event and contains an average of 4 key variables (name, address, source of payment, amount paid by each source, and total charge). The average number of changes per line of data is 0.6 (826 ÷ 1,398), with 2.6 changes per household (826 ÷ 317) and 1.0 changes per person (826 ÷ 819).

Table 3 shows the number and type of new and deleted visits and items picked up by the summary review. The net change of all medical events was shown to be 24, with 34 additional and 10 deleted items and visits resulting after the summary review. Nine of the ten deletions resulted because the visit or event appeared on the wrong family member's page of the summary. The net change (24) represented 2.2 percent of all 1,116 reported items and visits.

Table 4 shows the changes made in out-of-pocket expenses for dental visits, physician visits and prescribed medicines as a result of the summary review. Changes in the amount paid by the family were made in 15.0 percent of physician visits, 34.3 percent of dental visits and 13.4 percent of prescribed medicines.

Table 5 shows the average out-of-pocket expense for different medical events for the first interview compared with the subsequent summary review, and the number of out-of-pocket expenses which were known and not known at the two times. At the time of the first interview, 174 (23.7 percent) of all out-of-pocket expenses were unknown. After the summary review, only 61 (8.3 percent) of these expenses still remained unknown.

The ratios and net differences of the average out-of-pocket expense for different items as reported in the two time periods is given in Table 6. Average out-of-pocket costs per household dropped $2.82 for dental visits, $0.05 for prescribed medicines, $0.91 for other medical provider visits and $1.56 for other medical expenses. Some costs increased, e.g., $5.48 for physician visits, $58.50 for hospitalizations and $4.66 (a 31 percent increase) for all medical events.

Table 7 is similar to Table 5, except that it shows the average expense paid by private insurance for different medical events as given during the first interview and after the summary review. In this case, 59.0 percent of all amounts paid by insurance were unknown during the first interview and 38.8 percent remained unknown after the summary review.

Table 8 shows that the net ratio (1.04) of the average amount paid by insurance as reported in the two time periods is much less than that reported for out-of-pocket expenses (1.31)

The net differences are also given in Table 8 for insurance payments reported before and after the summary review. There was a decrease in the amount paid by insurance of $1.67 for dental visits, $0.80 for physician visits and $1.21 for prescribed medicines. There was an increase of $149.12 for hospitalizations and $2.26 for the total of all medical events. However, after the review of the summary, there are still many insurance payments which remain unknown.

Information on the names and addresses of physicians after the summary review was compliled in Table 9. While the data are not directly related to medical expenditures, it is crucial for the MPS when it will be necessary to contact all the physicians. Only in a few Instances (2.8 percent) will both the name and address of a physician be unknown.

5. CONCLUSIONS

The household summary was a valuable tool for obtaining missed visits and for correcting and completing information on the cost of medical care. A question which remains unanswered is the length of an optimum period of time for summary review, which would be both long enough to allow for unknown variables to be resolved and short enough so that recall would not be a problem. An issue which can be resolved after the entire pretest is the relative effectiveness of the household summary versus going directly to the providers in obtaining accurate cost data.

REFERENCES

1. "Medical Economics Survey Field Operations and Costs, January through October, 1975." Westat Research, Rockville, Md., 1212175.

2. Oliver, Julia D. "The Design and Methodology of the National Medical Care Expenditure Survey." American Statistical Association, Contributed Paper Session, Chicago, Illinois (August 1977).

Illustration 1

MEDICAL CARE AND EXPENSES SUMMARY

FOR: HEALTHY, JACOB J. MALE AGE 66 FROM: 01/15/75 TO: 05/21/76

PROVIDER NAME / ADDRESS / CITY, STATE	DATE OF CARE	TYPE OF SERVICE / SPECIALITY OR ITEM PROVIDED	AMOUNT	BILL WAS OR WILL BE PAID BY SOURCE OF PAYMENT	AMOUNT OF PAYMENT
*** I. DENTAL CARE EXPENSES					
PULLEM X. Y. DDS 123 MAIN STREET CHICAGO, ILL.	03 04 76	EXTRACTION	$ 32.00	FAMILY AETNA	$ 14.00 $ 18.00
*** II. HOSPITAL VISIT EXPENSES					
NONE					
*** III. DOCTOR VISIT EXPENSES					
SMITH, JOHN D. MD 22 MAIN STREET CHICAGO, ILL.	05 12 76	GEN PRACT	NOT AVAIL	MEDICARE	NOT AVAIL
*** III. OTHER HEALTH CARE EXPENSES					
NONE					

TABLE 1. Number of Medical Events, by Type

	Number of Events	Average Number of Events/Household
Dental Visits	174	0.55
Hospitalizations	15	0.05
Physician Visits	443	1.40
Prescribed Medicine	399	1.26
Other Medical Providers' Visits	58	0.18
Other Medical Expenses	27	0.08
Total	1,116	3.52

TABLE 2. Total Changes Made on the Summary

Variable	Number of Items	Number of Items Not Changed			Number of Changes				
		Item Known	Item Unknown	Total	Corrections	Additions	Deletions	Previous Omissions of Ques.	Total
Name of Provider/Item	1,116 (100%)	933 (83.6%)	99 (8.9%)	1,032 (92.5%)	40 (3.6%)	34 (3.0%)	10 (0.9%)	0	84 (7.5%)
Address of Provider	696 (100%)	625 (89.8%)	29 (4.2%)	654 (94.0%)	15 (2.2%)	20 (2.9%)	7 (1.0%)	0	42 (6.0%)
Source of Payment	1,272 (100%)	973 (76.5%)	83 (6.5%)	1,056 (83.0%)	91 (7.2%)	34 (2.7%)	12 (0.9%)	69 (5.4%)	216 (17.0%)
Amt. of Payment by Source	1,272 (100%)	733 (57.6%)	289 (22.7%)	1,022 (80.3%)	133 (10.4%)	34 (2.7%)	12 (0.9%)	58 (4.6%)	250 (19.7%)
Total Charge	1,116 (100%)	661 (59.2%)	221 (19.8%)	882 (79.0%)	123 (11.0%)	34 (3.0%)	10 (0.9%)	57 (5.1%)	234 (21.0%)
Totals	5,472			4,646 (84.9%)					826 (15.1%)

TABLE 3. Number of New and Deleted Medical Events

Type of Event	Number of New Events	Number of Deleted Events	Net Change
Dental Visits	12	4	8
Physician Visits	5	2	3
Prescribed Medicines	13	4	9
Visits to Other Medical Providers	4	0	4
Other Medical Expenses	0	0	0
Hospitalizations	0	0	0
Totals	34	10	24

TABLE 4. Changes in Out-of-Pocket Expenses for Dental Visits, Physician Visits and Prescribed Medicine

	Number of Items=n	Number of Items Not Changed			Number of Changes				
		Item Known	Item Unknown	Total	Corrections	Additions	Deletions	Previous Omissions of Ques.	Total
Dental Visits	131 (100%)	80 (61.1%)	6 (4.6%)	86 (65.7%)	31 (23.7%)	10 (7.6%)	2 (1.5%)	2 (1.5%)	45 (34.3%)
Physician Visits	232 (100%)	176 (75.9%)	21 (9.1%)	197 (85.0%)	27 (11.6%)	3 (1.3%)	2 (0.9%)	3 (1.3%)	35 (15.0%)
Prescribed Medicine	312 (100%)	245 (78.3%)	26 (8.3%)	271 (86.6%)	22 (7.0%)	13 (4.2%)	4 (1.3%)	2 (0.6%)	41 (13.4%)

TABLE 5. Average Out-of-Pocket Expenses by Medical Event

Medical Event	As Reported in the First Interview			After Summary Review		
	n(K)	n(DK)	Average Cost per Household	n(K)	n(DK)	Average Cost per Household
Dental Visits	86	45	$36.16	125	6	$33.34
Physician Visits	180	57	$16.97	214	23	$22.45
Prescribed Medicine	246	59	$ 5.76	277	28	$ 5.71
Other Medical Providers Visits	25	5	$17.04	30	0	$16.13
Other Medical Expenses	22	4	$22.42	24	2	$20.86
Hospitalizations	2	4	$21.75	4	2	$80.25
All Medical Events	561	174	$15.21	674	61	$19.87

n(K) = Number of out-of-pocket expenses known

n(DK) = Number of out-of-pocket expenses unknown

TABLE 9. Name and Address of Medical Doctor After Summary Review

	Number	Percent
Both Name and Address Known	378	86.7
Only Name Known	3	0.7
Only Address Known	43	9.9
Neither Name nor Address Known	12	2.8
Totals	436	100.0%

TABLE 6. Comparisons of Average Out-of-Pocket Expenses Between the First Interview and the Summary Review

	(Average Out-of-Pocket Expense After the Summary Review) ÷ (Average Out of Pocket Expense Reported in the First Interview)	Net Change in Out-of-Pocket Expenses After the Summary Review
Dental Visits	.93	- $ 2.82
Physician Visits	1.32	+ $ 5.48
Prescribed Medicines	0.99	- $ 0.05
Other Medical Providers Visits	.94	- $ 0.91
Other Medical Expenses	0.93	- $ 1.56
Hospitalizations	3.69 *	+ $58.50 *
All Medical Events	1.31	+ $ 4.66

* n-very small

TABLE 7. Average Amount Paid by Insurance by Medical Event

Medical Event	As Reported in the First Interview			After Summary Review		
	n(K)	n(DK)	Average Cost per Household	n(K)	n(DK)	Average Cost per Household
Dental Visits	11	24	$42.50	24	11	$40.83
Physician Visits	41	44	$32.50	53	32	$31.70
Prescribed Medicines	18	34	$ 5.22	28	24	$ 4.01
Other Medical Provider Visits	0	3	-	2	1	$26.25
Other Medical Expenses	2	1	$16.88	2	1	$16.88
Hospitalizations	5	5	$416.81	6	4	$565.93
All Medical Events	77	111	$52.12	115	73	$54.38

n(k) = Number of out-of-pocket expenses known

n(DK) = Number of out-of-pocket expenses unknown

TABLE 8. Comparisons of Average Insurance Payments Between the First Interview and the Summary Review

	(Average Insurance Payment After the Summary Review) ÷ (Average Insurance Payment Reported in the First Interview	Net Change in Out-of-Pocket Expenses After the Summary Review
Dental Visits	0.96	- $ 1.67
Physician Visits	0.98	- $ 0.80
Prescribed Medicines	0.77	- $ 1.21
Other Medical Providers Visits	**	**
Other Medical Expenses	1.00 *	0 *
Hospitalizations	1.36 *	+$149.12
All Medical Events	1.04	+ $ 2.26

* n-very small

** not available

ESTIMATION PROCEDURES USED TO PRODUCE WEEKLY FLU STATISTICS FROM THE HEALTH INTERVIEW SURVEY

James T. Massey, Gail S. Poe, Walt R. Simmons
National Center for Health Statistics

1. INTRODUCTION

In April 1976, the United States Congress appropriated $135 million for a national immunization program against the A/New Jersey or "Swine Flu." The Center for Disease Control (CDC) was charged with the responsibility of developing a comprehensive immunization delivery system and with the assessment of the coverage of the vaccination program, as well as the surveillance of flu cases. Missing from CDC's surveillance systems was a system through which national estimates could be made from a national probability sample or a full census. Although CDC's basic systems could provide partial information for the entire country, they could not provide sufficient data for production of estimates that could be assessed for precision. CDC therefore, requested the National Center for Health Statistics (NCHS) to collect influenza activity data in the Health Interview Survey (HIS).

In the Health Interview Survey a probability sample of households representing the civilian, noninstitutionalized U.S. population is interviewed each week. Interviewing is done continuously on a weekly sample of about 800 households. In response to CDC's need a supplemental set of questions on influenza and influenza vaccinations was added to the regular HIS interview questionnaire in the last week of September 1976.

In the regular HIS processing procedures, the time between the data collection and publication of the results is generally at least one year. Because of the demand for timely data on influenza cases and vaccinations the HIS implemented a rapid reporting system in which estimates of influenza-like illnesses; bed days due to such illnesses; and all types of influenza, including swine flu, vaccinations were published weekly three weeks after the week for which the estimates were made and only one week after the data were collected.

The HIS sample is designed so that tabulations can be provided for each of four major geographic regions, for large metropolitan areas, and for urban and rural sectors of the United States. The sample is also designed so that households interviewed each week represent those in the target population and that the weekly samples are additive over time. A rapid reporting system was used one other time in the history of the HIS, and that was during the 1957-58 influenza epidemic. At that time, weekly reports also were issued.

The weekly reports for 1977 were continued through April, and the estimates presented were provisional. Final estimates will be published after several months of extensive data processing in which medical coding is completed and many error and consistency checks are made on the data. The HIS weekly estimates were not part of CDC's systems for detecting early outbreaks of influenza. Because of the national scope of the data, local outbreaks of influenza-like illness possibly were undetected. However, when used in conjunction with other sources of information within the CDC surveillance system, the HIS data could confirm or deny early inferences regarding the spread of this disease and its effect.

2. STATISTICAL METHODS

Several estimation procedures were considered by NCHS for estimating the weekly number of flu cases, the number of bed-days due to flu, and the number of all types of flu shots and swine flu shots. The two most prominent estimators are described below along with some of their properties.

Since the HIS uses a two-week reference period to collect data on the incidence of acute conditions a two-week reference period was also chosen for the influenza supplement. That is, during each week of interviewing a flu case is enumerated if its onset occurred during the two weeks preceding the interview week, a bed-day is enumerated if it occurred during the two weeks prior to the interview week, and a flu shot is enumerated if it were received in the two weeks prior to the interview week. Thus, for each week i of interest the following two independent estimates can be made for the number, say X_i, of flu cases, bed-days, or flu shots.

α_i' - the estimate for the "last week" obtained from interview week (i+1) and,

β_i' - the estimate for "week before" obtained from interview week (i+2).

The first estimator of X_i considered was used during the 1957-58 flu epidemic to estimate the incidence of acute upper respiratory conditions and is given by

$$X_i'' = \frac{1}{2}(\alpha_i' + \beta_i').$$

The estimator, X_i'', is unbiased and, if one assumes that the variance of X_i'' is constant from week to week, then the variance of X_i'' is given by

$$\sigma^2_{X_i''} = \frac{1}{2}\sigma^2_{\alpha_i'}.$$

The second estimator to be considered is given by

$$X_i' = \frac{1}{4}\left(\beta_{i-1}' + \alpha_i' + \beta_i' + \alpha_{i+1}'\right)$$

$$= \frac{1}{2}\left(U_i' + U_{i+1}'\right)$$

where $U_i' = \frac{1}{2}\left(\beta_{i-1}' + \alpha_i'\right)$ is the average weekly estimate obtained from interview week (i+1).

The estimator X_i' is a weighted average of four weekly estimates obtained from interview weeks (i+1) and (i+2). Since the estimator contains information from the week on either side of the week of interest a smoothing effect results. The expected value of X_i' is given by

$$E(X_i') = X_i + \frac{1}{4}\left(X_{i-1} + X_{i+1} - 2X_i\right).$$

The bias of X_i' is given by the second term on the right hand side of the above equation. In most situations (especially if a linear trend is present) this bias will be small. The only time when this bias might be more than a few percentage points is at the maxima or minima points of a trend.

Again, assuming that the variance of the weekly statistics remains constant from week to week, the variance of X_i' can be expressed as

$$\sigma_{X_i'}^2 = \frac{1}{16}\left[4\sigma_{\alpha_i'}^2 + 4\ \mathrm{COV}\left(\beta_{i-1}', \alpha_i'\right)\right]$$

$$= \frac{1}{4}\sigma_{\alpha_i'}^2\left[1 + r_{\beta_{i-1}', \alpha_i'}\right]$$

where $r_{\beta_{i-1}', \alpha_i'}$ is the correlation between the incidence of "last week" and the "week before" from a single week's sample. For most acute conditions the correlation is assumed to be small, although the correlation will be higher for very contagious diseases. The assumption of equal weekly variances should hold unless X_i changes considerably from week to week. It should also be noted that the weekly statistics U_i' and U_{i+1}' are independent since they are obtained from independent weekly samples.

Comparing the variances of X_i'' and X_i',

$$\sigma_{X_i'}^2 < \sigma_{X_i''}^2 \quad \text{for } r_{\beta_{i-1}', \alpha_i'} < 1.$$

Using the data on the number of flu cases for the 1975-76 flu season (the last quarter of 1975 plus the first quarter of 1976) the correlation coefficient, $r_{\beta_{i-1}', \alpha_i'}$, was estimated to be approximately 0.75. Thus, for flu cases the variance of X_i' is approximately 13 percent smaller than the variance of X_i''. Based on a mean squared error criteria there is little to choose between X_i' and X_i''.

Another important feature of the estimator X_i', however, is that it can be formed using the two-week average estimates U_i' and U_{i+1}' and doesn't require the formation of two separate weekly estimates for each week of interviewing. Operationally, this feature reduces the number of weekly tabulations in half. For this reason the estimator X_i' was selected for making our weekly estimates.

The difference between incidence for adjacent weeks is estimated by

$$d_i' = X_i' - X_{i-1}'$$

and the variance of d_i' can be shown to be

$$\sigma_{d_i'}^2 = \frac{1}{4}\sigma_{\alpha_i'}^2\left[1 + r_{\beta_{i-1}', \alpha_i'}\right]$$

$$= \sigma_{X_i'}^2.$$

The variance of $\sigma_{d_i'}^2$ can also be shown to be equal to $2\sigma_{X_i'}^2\left(1 - r_{X_i', X_{i-1}'}\right)$ and, thus, $r_{X_i', X_{i-1}'} = \frac{1}{2}$. This correlation is intuitively obvious since U_i' is used to form one half of both the estimators X_i' and X_{i-1}'.

The weekly estimates can be summed to form aggregates such that for N weeks

$$X_s' = X_1' + X_2' + \ldots\ldots + X_N'.$$

The variance of X_s', assuming equal weekly variances, is given by

$$\sigma_{X_s'}^2 = N\sigma_{X_i'}^2 + 2r_{X_1'X_2'}\sigma_{X_i'}^2 + 2r_{X_2'X_3'}\sigma_{X_i'}^2 + \ldots. + 2r_{X_{N-1}'X_N'}\sigma_{X_i'}^2$$

$$= (2N-1)\sigma_{X_i'}^2.$$

The relative standard error of X_s' can be written as

$$V_{X_s'} = \frac{\sqrt{(2N-1)}\sigma_{X_i'}}{X_s'}$$

$$\doteq \sqrt{\frac{2N-1}{N^2}}\ V_{X_i'}.$$

For NCHS's weekly publications $V_{X_s'}$ is further approximated by

$$\sqrt{2/N}\ V_{X_i'}$$

where $V_{X_i'}$ is the relative standard error of X_i'

and is defined as $\sigma_{X_i'}/X_i'$.

2.1 Estimating Sampling Variance

There are several alternative methods for approximating the sampling variance of X_i' and three such methods which are described below were compared.

If the weekly statistics are reasonably stable from week to week with no apparent trend, then a simple estimate of $\sigma_{X_i'}^2$ can be made using any two consecutive weekly estimates of X_i.

For a single week i, the estimated sampling variance is given by

$$s_{X_i'}^2 = \frac{1}{4}(U_i' - U_{i+1}')^2$$

and the relative standard error is given by

$$v_{X_i'} = \frac{U_i' - U_{i+1}'}{U_i' + U_{i+1}'}.$$

By summing over k weeks a more stable estimate of $\sigma_{X_i'}^2$ can be obtained such that

$$s_{X_i'}^2 = \frac{1}{4(k-1)} \sum_{j=1}^{k-1} (U_j' - U_{j+1}')^2$$

and

$$v_{X_i'}^2 = \frac{s_{X_i'}^2}{\overline{U}}$$

where

$$\overline{U} = \frac{1}{k} \sum_{j=1}^{k} U_j'.$$

A second method of approximating $\sigma_{X_i'}^2$ uses least squares regression to fit three consequtive values of U_i' and looks at the deviation about the regression line to estimate $\sigma_{X_i'}^2$. A more stable estimate is again obtained by averaging a series of estimates. This method is satisfactory for linear trends but tends to over estimate the sampling variance when there are changes in direction in the trend. The approximation is given by

$$s_{U_i'}^2 = \frac{1}{6(k-2)} \sum_{j=2}^{k-1} (U_{j-1} - 2U_j + U_{j+1})^2$$

and

$$s_{X_i'}^2 = \frac{1}{2} s_{U_i'}^2.$$

Yet another method of estimating the variance of the weekly statistics is to compute a simple random sample variance estimate and inflate by a design factor. The design factor represents the increase or decrease in precision due to deviations from a simple random sample design such as stratification and clustering. For the HIS the design factor has been shown to be around two. Using this assumption estimates of variances were calculated for the number of flu cases and compared to the estimates obtained using the second method of approximation presented above. The results are presented below in terms of percent relative standard error (PRSE) which is the standard error of an estimate divided by the estimate itself multiplied by 100.

Size of Estimate (In thousands)	Method 3 (Simple Random Sample)	Method 2 (Regression)
1500	19	21
2000	16	18
2500	14	15
3000	13	13
4000	11	10
5000	10	9
6000	9	8

The overall average values of the percent relative standard errors for the number of flu cases and bed days shown in NCHS's weekly flu publication were obtained using data from the 1975 flu season (September to March) and then reverified using the first 14 weeks of the 1976 flu season. The PRSE for the weekly estimate of flu shots was not available in 1975 and was approximated using the 1976 data. The first two methods presented in this section were used to estimate the PRSE's and the methods were found to be quite comparable. The average weekly PRSE is about 16 for flu cases and 20 for bed days and flu shots.

2.2 Weighting and Post-Stratification

Up to this point it has been assumed that the average weekly estimates, U_i', have already been calculated. The first step in the weekly estimation procedure is to calculate U_i'. This is done by weighting the weekly sample data. Except for minor adjustments due to nonresponse and subsampling the HIS sample is self-weighting (each sample person has the same probability of selection in the national sample). For weekly estimation each sample person is assigned the same probability of selection. One final post-stratification adjustment is required, however, to adjust each week's sample to the same national population. Since each week's sample is a random sample, the distribution of sample persons will vary from week to week by age and race and an adjustment to the population distribution will improve the precision of the weekly estimates. The population distribution is obtained from the Bureau of the Census and adjustments are made each week for ten age-race groups. If

y_{jk} = total number of sample persons in the jk^{th} age-race cell reported during interview week (i+1),

z_{jk} = total number of flu cases, bed days, or flu shots in the jk^{th} age-race cell

reported during interview week (i+1), and

Y_{jk} = population control (Census value) for jk^{th} age-race cell for week i,

then the average weekly estimate U_i' obtained from interview week (i+1) is given by

$$U_i' = \frac{1}{2} \sum_{jk} z_{jk} Y_{jk}/y_{jk}.$$

The U_i' are then used to calculate the X_i's and their sampling errors.

3. RESULTS

Figure 1 below shows the weekly estimates of flu-like illness for September 20, 1976 through April 17, 1977 and Figure 2 contains the estimates of bed days due to flu for the same time period. The curves are similar to ones for the last several years. The relative smoothness of the curves further indicates the stability of the estimators which were employed in the rapid reporting system. The table below gives the actual weekly estimates for the variables of interest.

4. CONCLUSION

Although there was, fortunately, no epidemic of Swine Flu this year we felt that the HIS rapid reporting system was a success. Reliable weekly estimates were published only one week after the data were collected and only three weeks after the reference week. It would have been impossible to implement such a system after the detection of an epidemic because of the rapidity with which such a virus spreads throughout the entire country.

The influenza supplement will provide health data analysts and planners with extensive information on the correlates of influenza. Among the many areas that may be examined are the relationships among other health and demographic characteristics (obtained in the main interview) and influenza, as well as the effect of influenza symptoms on limitation of usual activity (such as work loss). Additionally, the characteristics of persons who obtained flu vaccinations as opposed to those who did not and the timing of the vaccination in relation to the contraction of an upper respiratory illness may be studied.

REFERENCES

Poe, Gail S. and Massey, James T., "Estimating Influenza Cases and Vaccinations by Means of Weekly Rapid Reporting System," Public Health Reports, 92 No. 4 (1977), 299-306.

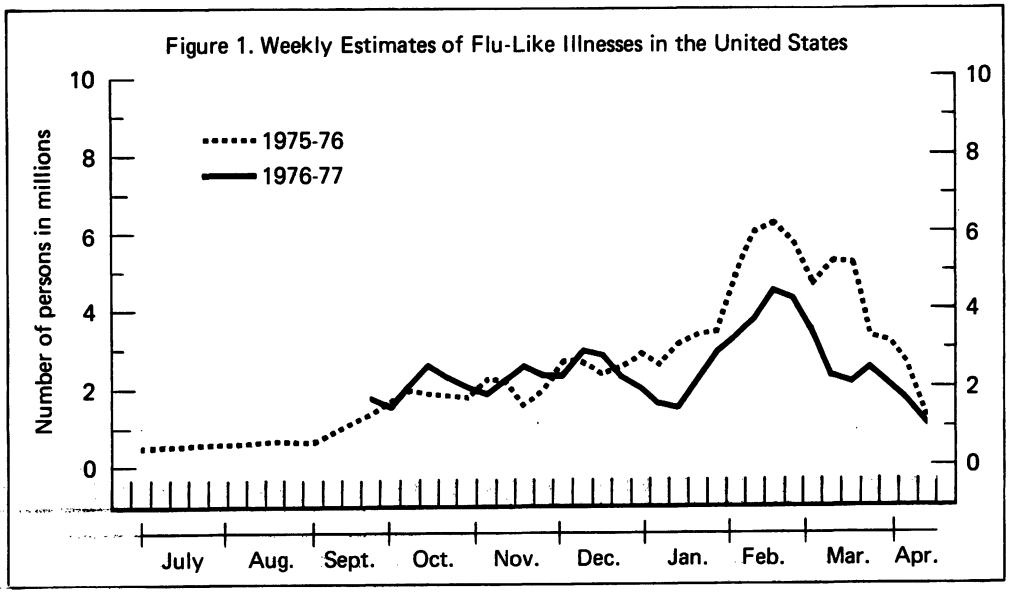
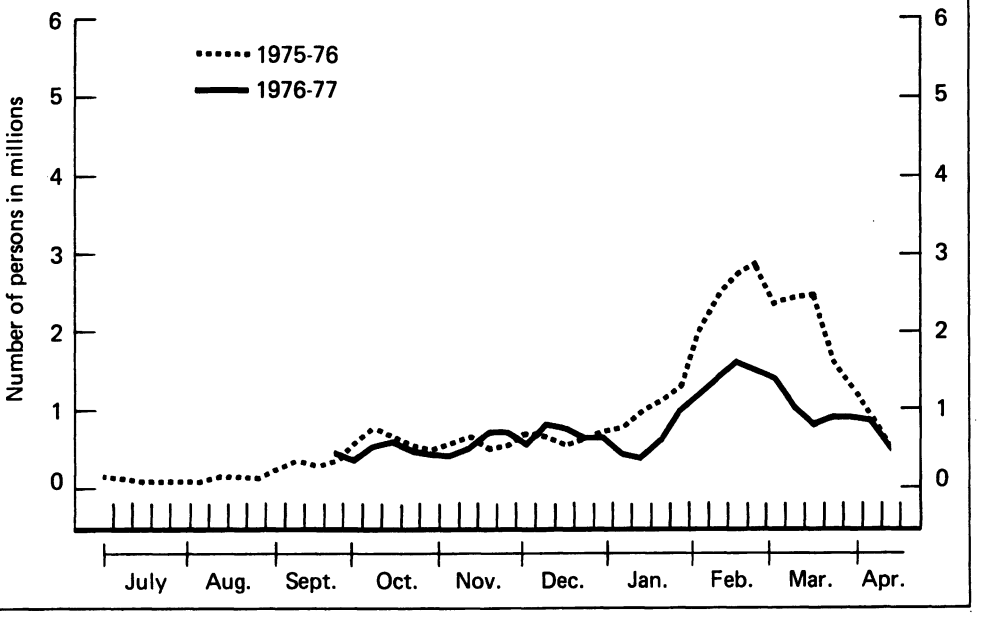


Figure 2. Weekly Estimates of Average Number of Persons in Bed Each Day Because of Flu-Like Illness in the United States



Weekly Estimates of Flu-Like Illnesses, Average Number of Persons in Bed Each Day, and Flu Shots: United States, 1976-77

Week	Flu-like illness	Average number of persons in bed each day because of flu-like illness	All types of flu shots		Swine flu shots	
			Each week	Cumulative since September 20	Each week	Cumulative since September 20
	Number in thousands					
September 20-26, 1976	1,710	419	638	638	*	*
September 27-October 3, 1976	1,490	369	1,130	1,768	*	*
October 4-10, 1976	2,115	517	1,846	3,614	1,200	1,668
October 11-17, 1976	2,567	591	3,196	6,810	2,378	4,046
October 18-24, 1976	2,239	489	4,737	11,547	4,014	8,060
October 25-31, 1976	2,062	413	5,122	16,669	4,634	12,694
November 1-7, 1976	1,880	410	5,580	22,249	5,019	17,713
November 8-14, 1976	2,319	511	6,749	28,998	6,391	24,104
November 15-21, 1976	2,493	704	5,379	34,377	5,154	29,258
November 22-28, 1976	2,277	700	4,101	38,478	3,921	33,179
November 29-December 5, 1976	2,276	573	4,128	42,606	3,972	37,151
December 6-12, 1976	2,857	789	2,616	45,222	2,361	39,512
December 13-19, 1976	2,831	753	1,235	46,457	1,063	40,575
December 20-26, 1976	2,207	631				
December 27,1976-January 2,1977	2,023	636				
January 3-9, 1977	1,646	465				
January 10-16, 1977	1,457	411				
January 17-23, 1977	2,127	562				
January 24-30, 1977	2,938	945				
January 31-February 6, 1977	3,242	1,205				
February 7-13, 1977	3,766	1,381				
February 14-20, 1977	4,540	1,597				
February 21-27, 1977	4,333	1,517				
February 28-March 6, 1977	3,290	1,299				
March 7-13, 1977	2,308	969				
March 14-20, 1977	2,190	778				
March 21-27, 1977	2,525	882				
March 28-April 3, 1977	2,123	896				
April 4-10, 1977	1,614	772				
April 11-17, 1977	1,138	516				

*Figure does not meet standards of precision.

NOTE: Even though the suspension of the Public Health Service immunization program was lifted on February 7, 1977, estimates of flu shots are not shown after the week ending December 19, 1976.

LENGTH OF INTERVIEW: THE NATIONAL SURVEY OF FAMILY GROWTH

Gordon Scott Bonham, National Center for Health Statistics

1. INTRODUCTION

Gathering information takes time. Time represents both burden on the respondent and cost to the organization collecting the data. Any reduction in the amount of time for interview would be desirable, as long as it did not decrease the amount of information or its quality. This paper presents data from the first cycle of the National Survey of Family Growth (NSFG). It analyzes the time required to complete interviews as related to characteristics of the respondent, of the interviewer and of the interviewer training. Characteristics of the respondent that affect the length of interview are not subject to modification without modifying the design of the survey itself. The effect of characteristics of the interviewer on the length of interview might be modified by differential selection of interviewers, but it is interviewer training that is most subject to modification by the survey organization.

The author has found little published data relating to the length of interview. One study showed that the time spent in actual interviewing is a fraction of the total time spent by the survey organization for each case (Sudman 1965). Another study found that younger and higher educated interviewers had higher production ratios than did older or less educated interviewers. Production ratios also increased with length of experience with the Bureau and with progression through the assigned work load (U.S. Bureau of the Census 1972).

Changes in length of interview do not necessarily imply changes in interview quality. A decrease in the length of interview could be at the expense of quality if the interviewer is taking shortcuts, not probing completely, or failing to record all relevant verbatim information given by the respondent. The National Center for Health Statistics (1977) found a decrease in the amount of information obtained with increased experience of the interviewer. However, characteristics of interviews that shorten the time required for interview might also increase the quality of the interview. The quality or completeness of the information obtained in the NSFG is beyond the scope of this analysis. However, a brief investigation showed that the longer the interview, the greater the number of errors discovered in the preliminary office edit ($r=0.14$) and the greater the percentage of items that were not ascertained ($r=0.07$).

2. DATA AND METHODS

The NSFG is a periodic survey conducted by the National Center for Health Statistics. It was designed to provide information about fertility, family planning, and aspects of maternal and child health that are closely related to childbearing. Data on these topics were collected in the first cycle by personal interviews with 9,797 women aged 15-44 years who had ever been married or who had children of their own living in the household. Field work was conducted between July 1973 and February 1974 by the National Opinion Research Center of the University of Chicago. Background data were obtained for the 335 female interviewers working on the project.

The time at the beginning and end of interview was recorded on the questionnaire for 9,676 interviews. The length of these interviews ranged from 15 minutes to 4 hours and 40 minutes, with 60 minutes being the mode. The average length of interview was 67 minutes with a standard deviation of 22 minutes.

Multiple Classification Analysis (MCA) is used in the analysis (Andrews et al. 1973). MCA is basically a multiple regression analysis using dummy variables that shows deviations from the grand mean. The unadjusted deviation in the tables shows the actual number of minutes more or less that interviews in the category took compared with the grand mean. The adjusted deviation is interpreted as what the deviation of the category would be if the cases in the category had the same distribution as all cases with respect to all other characteristics entered into the analysis. The "eta" statistics shows the strength of the simple bivariate relationship between each of the predictor variables and the length of interview; eta squared is the proportion of the variance in length of interview explained by the predictor characteristic. The "beta" statistic is a measure of the relationship between the length of interview and the predictor characteristic holding constant the effect of the other predictor characteristics. As with a multiple regression "beta," the MCA beta can be used to show the relative importance of the different predictor characteristics. Because of the large number of interviews, almost all differences are statistically significant. Therefore, the analysis will focus on the relative importance of the characteristics and on their practical significance.

3. FINDINGS

3.1 Respondent Characteristics

The design of the NSFG meant that women with more pregnancies were asked more questions than women with fewer pregnancies. Likewise, currently married women were asked more questions than other women. These characteristics should definitely affect the length of interview, while other characteristics like race, age, work status and education might affect the length. Although neither the survey design nor respondent characteristics can be modified to reduce the length of survey without changing the nature of the survey or the surveyed population, the effect of these characteristics should be understood and should be controlled in subsequent analysis.

Table 1 shows that each pregnancy increased the interview length by an average of 3.5 minutes.

1. Percent of interviews and unadjusted and adjusted[a] deviations in minutes from the mean length of interview by characteristics of the respondent: National Survey of Family Growth, 1973.

Characteristics of the respondent	eta	beta	Percent of Interviews	Unadjusted deviations	Adjusted deviations
All interviews			100.0	0.0	0.0
Number of pregnancies	0.35	0.35			
No pregnancies			10.6	-14.2	-13.7
1 pregnancy			19.5	- 4.5	- 4.8
2 pregnancies			23.3	- 0.7	- 0.8
3 pregnancies			18.2	0.7	0.9
4 pregnancies			11.0	2.9	3.0
5 pregnancies			7.1	6.7	7.0
6 pregnancies			4.0	10.4	10.4
7 pregnancies			2.5	13.9	13.7
8 pregnancies			1.6	14.5	14.3
9 pregnancies			0.9	25.0	23.9
10 pregnancies			0.6	20.9	23.9
11 pregnancies			0.4	28.9	27.4
12 pregnancies			0.2	41.5	40.6
13 pregnancies			0.1	31.6	31.7
14 pregnancies			0.0	63.3	61.3
17 pregnancies			0.0	113.3	111.3
19 pregnancies			0.0	78.3	76.5
26 pregnancies			0.0	33.3	33.4
Race	0.14	0.12			
White and other			61.0	- 2.5	- 2.1
Negro			39.0	3.9	3.3
Marital status	0.00	0.08			
Currently married			77.4	- 0.0	1.0
Widowed, divorced, separated, single with own children			22.6	0.2	- 3.3
Age	0.10	0.06			
15 to 19 years			5.4	- 4.9	- 0.0
20 to 24 years			18.9	- 3.3	0.9
25 to 29 years			21.6	- 0.2	1.6
30 to 34 years			20.1	1.2	0.7
35 to 39 years			17.9	2.1	- 1.1
40 to 44 years			16.1	1.8	- 2.1
Working status	0.10	0.05			
35 or more hours per week			32.9	- 2.8	- 1.3
1 to 34 hours per week			9.1	- 1.4	- 0.8
Not working			58.0	1.8	0.9
Education	0.14	0.04			
Not a high school graduate			34.3	4.2	1.2
High school graduate			45.2	- 1.9	- 0.9
Some college or more			20.5	- 2.7	0.8
Number of interviews	9696				
Mean length of interview	66.7				
Coefficient of determination	0.14				

[a]Adjusted for the other characteristics in the table by use of Multiple Classification Analysis.

The number of pregnancies by itself explains 12 percent (eta^2) of the variation in the length of interview.

Interviews with black women took 6.4 minutes longer, on the average, than did interviews with white and other race women. Adjusting for the other characteristics of the respondent shown in the table reduced the differential to 5.4 minutes. However, there was an attempt to match interviewers and respondents by race, and so this characteristic of the respondent may be picking up effects

2. Percent of interviews and unadjusted and adjusted[a] deviations in minutes from the mean length of interview by characteristics of the interviewer: National Survey of Family Growth, 1973.

Characteristics of the interviewer	eta	beta	Percent of interviews	Unadjusted deviations	Adjusted deviations
All interviews			100.0	0.0	0.0
Age	0.18	0.16			
Under 30 years			14.0	- 3.9	- 3.2
30 to 39 years			27.5	2.7	1.6
40 to 49 years			31.0	- 3.8	- 3.4
50 to 59 years			22.4	2.7	3.4
60 to 69 years			2.7	- 1.0	- 0.7
Not ascertained			2.4	16.3	13.3
Race	0.20	0.15			
White and other			62.2	- 2.4	0.0
Black			37.5	3.5	- 0.5
Not ascertained			0.3	60.1	58.4
Religion	0.06	0.09			
Protestant			68.2	0.0	- 0.8
Catholic			11.2	0.2	- 0.2
Jewish			10.3	0.5	4.1
None			5.5	0.3	3.8
Other			4.3	- 4.0	- 1.9
Not ascertained			0.4	16.4	15.2
Education	0.05	0.07			
Not a high school graduate			3.1	- 1.8	- 2.5
High school graduate			21.0	- 0.2	2.2
Some college			40.1	- 0.7	- 1.3
College graduate or more			35.7	1.1	0.5
Not ascertained			0.1	-19.0	-19.9
Marital status	0.09	0.06			
Currently married			77.5	- 0.7	- 0.2
Widowed, divorced, separated			18.2	2.9	0.8
Never married			2.9	4.7	5.0
Not ascertained			1.4	- 9.2	- 7.3
Children ever born	0.07	0.06			
No children			14.2	1.6	0.2
1 or 2 children			37.0	- 1.0	- 1.1
3 or 4 children			34.7	- 1.0	- 0.2
5 or more children			13.2	3.3	3.2
Not ascertained			0.2	5.7	2.6
Number of iterviews	9696				
Mean length of interview	66.7				
Coefficient of determination (adjusted):					
interviewer characteristcs			0.08		
Interviewer and respondent characteristics			0.20		

[a]Adjusted for other characteristics of the interviewer in the table as well as characteristics of the respondent in Table 1.

of race of the interviewer (to be discussed below).

The NSFG questionnaire consisted of separate forms for currently married and currently unmarried women. Table 1 shows no differences in the unadjusted length of interview by marital status. However, once the number of pregnancies and other characteristics of the respondent are controlled, currently married women, as expected, took an average of 4.3 minutes longer per interview than did previously married or single women.

The unadjusted deviations in Table 1 show an increasing length of interview with increasing age of the respondent. However, the adjusted deviations suggest that the relationship is curvilinear. Women 25-29 years old took longer to interview than did women who were either older or younger.

Interviews with women working full time were shorter on the average than were interviews with women working part time; interviews with women working part time were shorter than were interviews

with women not working at all. The amount of the differences is over twice as great in the unadjusted figures than it is when other characteristics of the respondent are controlled.

Education considered separately has a strong relationship with the length of interview. Women with at least some college education took almost 7 minutes less to interview than women who had not graduated from high school. Once other characteristics are controlled, however, the effect of education is reduced.

In summary, the NSFG was designed to ask currently married women and women with greater numbers of pregnancies more questions than previously married or single women and women with fewer numbers of pregnancies, respectively. These aspects of the survey design are reflected in the length of interview. In addition, interviews were slightly longer for black women, women in their twenties, women who were not working, and women who were not high school graduates. Altogether, characteristics of the respondent explain 14 percent of the variance in the length of interview.

3.2 Characteristics of the Interviewer

The length of the interview may be affected by the interviewer as well as the respondent. Table 2 shows the deviations from the overall average length of interview for selected characteristics of the interviewers. The adjusted deviations control for all respondent characteristics of Table 1 as well as all of the other interviewer characteristics in Table 2.

Although the beta coefficient indicates a fairly important relationship between age of interviewer and length of the interview, the lack of pattern may indicate that only random variations are occurring.

Interviews conducted by black women took about 6 minutes longer than interviews conducted by white women. The differential is reduced to 0.5 minutes once other characteristics of the interviewer and characteristics of the respondents are controlled.

The NSFG matched interviewers and respondents on race to the extent possible. In only 3.3 percent of the interviews were the respondent and the interviewer not of the same race, making it difficult to differentiate how much of the effect on length is due to race of the respondent and how much is due to the race of the interviewer. Table 3 shows the adjusted deviations from the mean length of interview by the cross classification of race of respondent and race of interviewer. Interviews with white or other race respondents conducted by white or other race interviewers were 4.3 minutes shorter (net of other factors) than interviews with black respondents conducted by black interviewers. Controlling for the race of respondent, race of interviewer affects the length of interview by 2.7-3.2 minutes. Controlling for the race of interviewer, race of respondent affects interview length by 1.3-1.6 minutes. Therefore, it appears that race of interviewer has a slightly greater effect than the race of the respondent.

3. Adjusted[a] deviations in minutes from the mean length of interview by race of respondent and race of interviewer: National Survey of Family Growth, 1973 (numbers of interviews in parantheses)

Race of Interviewer	Race of Respondent	
	White/Other	Black
White or other	-1.8 (N=5801)	-0.2 (N= 217)
Black	1.4 (N= 104)	2.5 (N=3524)
Not ascertained	-2.8 (N= 28)	65.5 (N= 2)

[a]Adjusted for other characteristics of the respondent in Table 1 and other characteristics of the interviewer in Table 2 (excluding the race variables).

The religion, education, marital status and children ever born by the interviewer were collected to determine if they had an effect on the interview. They do have some effect on the variation in interview length, but less so than the interviewer's race and age.

The six interviewers' characteristics in Table 2 explain 8 percent of the variation in length of interview by themselves. The combination of the respondents' and the interviewers' characteristics explains 20 percent of the variation in the length of interview.

3.3 Interviewer Training and Experience

Interviewer training and experience, whether on this particular survey or surveys in general, is related to the length of interview. Length decreases rapidly with the first few interviews, as the interviewer gains experience with the NSFG questionnaire. It then decreases more slowly with additional experience. A log model (Length = 78.6-9.9164 Log Order) fits better than a quadratic or a cubic model, but it does not fit the unadjusted group means very well between the 50th and the 150th interview. The fit of the log model to the adjusted group means is better for the first 100 interviews than for more than 100 interviews. Adjusted for other characteristics and training, interviews which were among the first 10 conducted by an interviewer took an average of 25 minutes longer than did those which had been preceeded by at least 149 other NSFG interviews. It could be that interviewers, who take a long time completing their first few interviews, drop out earlier than interviewers who take less time. Investigation did not show this to be the case.

The number of years the woman has been interviewing is related to the average length of interview (Table 4). However, this relationship

4. Percent of interviews and unadjusted and adjusted[a] deviations in minutes from the mean length of interview by experience and training of the interviewer: National Survey of Family Growth, 1973.

Experience and training of the interviewer	eta	beta	Percent of interviews	Unadjusted deviations	Adjusted deviations
All interviews			100.0	0.0	0.0
Order number of interview	0.20	0.23			
1st to 9th interview			28.0	6.8	7.0
10th to 19th interview			25.0	- 0.5	0.3
20th to 29th interview			18.2	- 3.5	- 2.4
30th to 39th interview			11.0	- 3.9	- 3.3
40th to 49th interview			5.8	- 4.4	- 4.8
50th to 59th interview			3.3	- 3.3	- 4.0
60th to 69th interview			2.2	- 4.2	- 7.0
70th to 79th interview			1.7	- 3.0	- 7.0
80th to 89th interview			1.2	- 2.7	- 5.9
90th to 99th interview			0.8	0.5	- 6.0
100th to 119th interview			1.3	- 2.3	-10.2
120th to 149th interview			1.0	- 6.2	-16.3
150th or higher interview			0.7	- 9.6	-17.7
Years with organization	0.16	0.10			
Less than 1 year			56.9	0.9	- 0.3
1 or 2 years			8.0	5.6	3.0
3 or 4 years			14.9	- 6.1	- 3.9
5 to 9 years			13.4	- 2.1	2.3
10 or more years			4.3	7.6	5.4
Not ascertained			2.4	- 4.2	- 2.0
Years interviewing	0.12	0.10			
Less than 1 year			37.8	1.3	1.7
1 or 2 years			12.5	3.8	3.0
3 or 4 years			18.6	- 3.2	- 2.1
5 to 9 years			18.4	- 2.9	- 2.8
10 or more years			11.4	2.5	0.3
Not ascertained			1.3	- 8.8	- 6.4
Training team	0.10	0.10			
Team A			30.1	- 1.2	- 1.4
Team B			32.5	0.7	1.3
Team C			35.7	- 0.4	- 0.7
Locally trained			1.6	16.9	15.6
Training session	0.14	0.07			
June 24-30, 1973			37.6	- 3.9	- 1.9
July 8-14, 1973			32.1	2.8	1.8
July 22-28 or later			30.2	1.9	0.5
Supervisory position	0.08	0.02			
Supervisor or coordinator			18.9	- 3.4	- 0.3
Not supervisor or coordinator			79.9	0.8	0.1
Not ascertained			1.2	0.8	- 4.1
Number of interviews	9696				
Mean length of interview	66.7				
Coefficient of determination (adjusted):					
interviewer experience and training			0.08		
interviewer experience and training and respondent characteristics			0.23		
interviewer experience, training and characteristics					
interviewer experience and training and interviewer and respondent characteristics			0.29		

[a]Adjusted for the other interviewer experience and training factors in the table as well as characteristics of the respondent in Table 1 and characteristics of the interviewer in Table 2.

is curvilinear with interviewers of 3-9 years experience conducting shorter interviews than interviewers with fewer or more years experience.

Most of the interviewers in the NSFG had not worked with the contracting organization prior to the NSFG, or had worked for it less than one year. However, this lack of experience with the contracting organization did not increase their average length of interview. There is also no indication that increased length of service with the contracting organization reduced the length of interviews.

The majority of interviewers for the NSFG were trained during one of three week-long training sessions. During each training session, interviewers were trained by one of three teams (denoted A, B, and C), each team comprised of 2-4 trainers. Due to interviewer loss, a fourth training session of 3½ days was held part way through the field work conducted by a single trainer. In addition a few interviewers were trained by local supervisors. The training team of the regular sessions had some effect on the length of interviews. However, interviewers trained in the mid-fieldwork session or by local supervisors took substantially longer to complete interviews than those trained by the regular trainers prior to field work, independent of the fewer number of interviews they were able to complete.

Analysis of the adjusted deviations from the overall mean length of interview shows a curvilinear relationship between length and the time of the three regular training sessions (mid-fieldwork training is included with the last session). The first session produced the shortest interviews and the second session the longest. It does not appear, therefore, that the experience of the trainers in training on the NSFG had any substantial effect on reducing the length of interviews.

Interviewers who were also supervisors or coordinators took less time interviewing respondents than interviewers with no administrative responsibilities, at least in respect to the unadjusted deviations. When other factors are controlled, this difference is reduced. Those selected for supervisory positions were probably those interviewers who had proved themselves most efficient in the past and had those characteristics or training background that enabled them to conduct interviews in shorter periods of time.

The six interviewer training factors by themselves explain 8 percent of the variation in the length of interview, and add more independent explanation to the respondent characteristics than do the characteristics of the interviewer. When combined with both respondent and interviewer characteristics, interviewer training increases the amount of the explained variance to 29 percent. The most important part of interviewer training is experience with the NSFG interview.

4. CONCLUSION

The first cycle of the NSFG took over an hour, on the average, per interview. There was considerable variation in the length of interview, however, and 29 percent of the variance in the length of interview can be explained by characteristics of the person interviewed, characteristics of the interviewer and the training of the interviewer. The two single most important variables were the number of pregnancies the respondent had had and the order number of the interview (the experience the interviewer had on the NSFG questionnaire).

Would reducing the variation in the length of interview actually save much time in terms of overall respondent burden or time costs? Most of the length of interview is determined by the content and design of the survey. Some of the variation in length is due to characteristics of the respondent and of the interviewers which might be difficult, or impossible, to modify without changing the nature of the survey. However, there are interviewer experience and training items that could be modified and reduce the length of interview.

Suppose that 100 interviewers, each with 5-9 years of previous interviewing experience, had been selected for the NSFG. If they had all been trained during the first training session and each had completed 97-98 interviews, the average savings per interview would have been: (taken from adjusted deviations of Table 4):

Interviewers with 5-9 years experience	- save 2.8 minutes
Training during first training session	- save 1.9 minutes
Conducting 97-98 interviews each	- save 3.4 minutes
Total	save 8.1 minutes

The mean interview length would have been reduced from 66.7 minutes to 58.6 minutes. This savings would have resulted in 1,323 fewer hours spent in interviewing. At $3-$4 dollars per hour for interviewers, $4,000-$5,000 would have been saved. This is a small amount in relationship to the overall cost of the survey, but is nevertheless a savings. However, the savings would have resulted in a 12 percent reduction in respondent burden.

REFERENCES

Andrews, Frank., Morgan, James N., Sonquist, John A. and Klem, Laura (1973), Multiple Classification Analysis, Ann Arbor: Institute for Social Research, University of Michigan.

National Center for Health Statistics (1977), "A Summary of Studies of Interviewing Methodology," by Charles F. Cannell, Kent H. Marquis and André Laurent, Vital and Health Statistics, Series 2-No. 69. DHEW Pub. No. (HRA) 77-1343.

Sudman, Seymour (1965), "Time Allocation in Survey Interviewing and in Other Field Occupations," Public Opinion Quarterly, 29, 638-648.

U.S. Bureau of the Census (1972), Investigation of Census Bureau Interviewer Characteristics, Performance, and Attitudes: a Summary, by Gail Poe Inderfurth, Working Paper No. 34, Washington, D.C., U.S. Government Printing Office.

LESSONS LEARNED ON PROCEDURES FOR LONGITUDINAL STUDY OF NURSING HOME RESIDENTS

Iris M. Shimizu, National Center for Health Statistics

1. INTRODUCTION

During the fall of 1975, a longitudinal study of nursing home residents was included in a pilot test for the National Nursing Home Survey (NNHS) in order to measure changes that occur in the health status and the activities of residents. Since changes are bound to occur if enough time is allowed to elaspe, the primary object of the test was to determine whether changes occur rapidly enough to be detected within an 8 week period of time. Thus, data for the test was collected at 8 week intervals by repeating the same questions verbatim about a sample of residents. Since it is possible that changes implied by the data collected could be due to error, a reconciliation study was conducted during the second survey. A consistency study was also done on the data collected.

This paper deals with lessons learned from the pilot test about the conduct of longitudinal surveys. While some of the observations in this study have possible implications on the quality of data from the NNHS itself, those implications are ignored, here, due to space limitations for the present paper.

The basic survey design for the pilot test is described in sections 2 and 3. The methodology for the reconcilliation study and results of the pilot test are discussed in section 4.

2. BASIC DESIGN OF THE PILOT STUDY

2.1 Sample of Facilities

Since the project was a pilot study, the sample of facilities was restricted for convenience to 4 cities, one in each of the Census Regions to allow geographical differences, if any, to surface during the study.

The sampling frame used for the first stage consisted of facilities listed in the 1973 Master Facility Inventory (MFI) of nursing homes. To reduce respondent burden, homes known to be in other surveys just prior to our pilot study were eliminated. These were homes in the 1975 pretest of the MFI, the 1975 Pretest Study of Institutionalized Persons done by the Social Security Administration, and the pilot study for the Survey on Head and Spinal Cord Injuries sponsored by the National Institute of Neurological Diseases and Stroke. Also deleted from the frame were homes with 300 or more beds since these are included in the NNHS with certainty or near certainty.

In order to have a variety of facilities represented in the test, 24 strata were defined and at least one facility was selected from each non-empty stratum. The variables used for defining the strata were:

a. Certification status as listed in 1975 by the Social Security Administration: (1) Certified for Medicare, with or without Medicaid (2) Certified for Medicaid only, or (3) Not certified.

b. Bed Size: (1) Less than 25 beds, (2) 25 - 49 beds, (3) 50 - 99 beds and (4) 100 - 299 beds.

c. Ownership: (1) proprietary or (2) non-proprietary.

Some of the strata described were empty. Hence, more than one home was selected from some strata with the restrictions that 8 homes had to come from each certification class and 6 homes had to be located in each of the 4 cities. (Health status of the resident was deemed to be more closely related to the certification of the home than any of the other stratifying variables.)

Of the 24 homes selected for the longitudinal study, 20 participated in both the initial survey and the resurvey. Of these 20, nine were chosen at random, with at least two per city for a reconciliation study.

2.2 Procedures Used Within Facilities

The homes selected for the study were visited by survey interviewers twice with about 8 weeks between visits. (Eight weeks appeared to be the maximum length of time feasible for a longitudinal study in the 2 to 3 months of data collection planned for the full NNHS.) On the first visit to each home, the interviewer selected a sample of residents by using a systematic random sampling procedure. This yielded a total of 197 sample residents for the study.

The resurvey in each facility was done by an interviewer other than the one who conducted the initial survey in the home. This was done to prevent the possibility that an interviewer might be biased due to memory of answers given on the prior visit to the facility.

Data were collected on sample residents on both visits. This meant that in addition to the usual practice of keeping the identity of sampled residents confidential, the residents sampled in the initial visit had to be identified during the revisit. Where permitted, residents' names were used as the link between the initial survey and the resurvey. If the administrator of a home objected to the use of residents' names, a code was used which permitted facility personnel to uniquely identify sampled residents during the resurvey but which prevented meaningful identification by any one not connected with either the home or the survey.

During the first survey, the staff person present who was most responsible in the facility for a sample resident was asked a series of questions about the health and activities of that resident. When possible the same staff member was interviewed concerning the sampled resident

during both visits to the home. Otherwise during the second survey that staff member present who was most responsible for the particular resident was interviewed. For both interviews, the respondent was asked to consult the resident's records for answers to the questions.

Among the questions asked about the sampled residents, about 30 questions were identical on the two surveys. The concept assumed, here, was that a change in response to any of these particular questions about a resident would indicate a change in the resident's status.

3. QUALITY IN DATA PROCESSING AND COLLECTION

A major concern in the resurvey study was the presence of errors in the data which might result in changes being indicated for an individual resident when indeed no change occurred. The errors could be due to the respondents, the interviewers, or the data processing. The errors due to respondents are dealt with only in the next section. This section deals with the quality of the data as it is affected by data processing and collection. Since, the study was only a pilot test, all the quality control procedures usually established for a full fledged NNHS were not instituted for the study and, hence, the data quality is not expected to be the same as that of a full survey. However, efforts were made to minimize the possibility of errors which could affect the number of changes in residents that would be detected in the pilot.

For the pilot test, keypunching was verified 100% and then the data was subjected to simple computer edits for such things as illegal codes and improper skip patterns. All errors detected in the computer edit were corrected manually after a review of both the error and the original questionnaire. The editing of the record was repeated until all edits were passed. Hence, it is expected the data processing has little effect on the counts of changes that resulted from the test.

Another factor in the data quality is the interviewer. It is conceivable that pilot test data can have proportionally more interviewer errors over all data collection than a full survey since interviewers are not as familiar with the survey or the data collection forms as they would become over a full survey. The interviewers did receive as much training as is usually given in the full survey. That is, they were asked to read the interviewer's manual during the two weeks prior to their participation in a training session of several days length.

Interviewers could possibly have transmitted their biases to the data recorded or to the cues which they gave the respondent. This type of error is not easy to detect and, indeed, no effort was made to measure it in the pilot. However, since the respondents for residents were for the most part nurses, it is assumed that the respondents had the ability to choose the correct answers from the options offered for each question with a minimum of influential cues from the interviewers. Furthermore, the respondent had to be in agreement with the answers recorded except for accidental recording errors. It is assumed for purposes of analysis that any accidential recording errors are random and not due to consistent bias on the part of the interviewer.

Interviewer error did affect the count of residents for which usable data could be tabulated for each question included in the pilot test questionnaire. When the interviewer failed to mark any answer for a question and when the interviewer marked an illegal number of options for a question during either the first survey or the resurvey, a zero was coded for the question and data for that question about the resident was then omitted from all tabulations made for the pilot study. Thus the resident was not counted among those for which data was available for the question. This error probably occurred most often because several answer options applied to the resident and the respondent, having difficulty choosing only one answer for the question, changed the answer originally given and then the interviewer forgot to draw a line through the original answer. Subsequently it was impossible to determine which answer was intended for the question since only one was allowed. The result is that the counts of residents included in tabulations ranged from 197 (the total sampled) on one question to 134 on another. For the majority of the questions, data were available for at least 170 residents. Analysis on any question must, hence, be restricted to those residents for which data could be tabulated.

4. QUALITY OF RECORDED DATA

Since it was desired to know whether changes indicated by the recorded data were real rather than due to error, two studies were conducted. A consistency check was done of the responses given on the two surveys for individual sample residents and a reconciliation study was conducted in a subsample of the facilities.

4.1. Results of Consistency Check

Among the questions that were repeated verbatim on the two surveys there were four questions for which the second answer should have been implicated by the first answer. Analysis of the consistency in responses to each of these questions was restricted to only those residents for which answers were available from both surveys, as mentioned above.

The questions used in the consistency study are:

1. What was this resident's primary diagnosis at admission?

2. Has this resident lived in this facility one full month or longer?

3. What was the primary source of payment when he/she was admitted to the home?

4. Does he/she have any of the following conditions or impairments?

The correct answers to the first and third of the above questions for any one resident must be the same on the two surveys. The correct answer to the second question was "yes" on the resurvey except for those residents who were discharged during the 8 week interval between the surveys and who had been in the home less than a full month at the time of the discharge.

On the fourth question the interviewer was to mark all answer options which applied to the resident. Among conditions listed as options were the following eight which are generally considered incurable once a person acquires them:

a. Mental retardation
b. Heart trouble
c. Arthritis or rheumatism
d. Parkinson's disease
e. Chronic respiratory disease
f. Diabetes
g. Permanent stiffness or deformity of back or extremities
h. missing extremities.

A resident having one of these particular conditions during the first survey would still have it during the resurvey. The reverse is not necessarily true since a resident could just begin to show the symptoms for some of the conditions for the first time during the 8 week interval between surveys.

To determine a lower limit on the frequency of errors which occurred for these questions, an error was counted each time the data recorded in the resurvey was inconsistent with that recorded for the same resident during the first survey. That is, errors were counted if changes other than the possible ones described above occurred to the resident according to the data recorded on the two surveys. The errors that did not yield inconsistencies in the data could not be detected and, hence, were not counted here. Changes that resulted because the response on one survey was "Don't know" were also not counted as due to errors. The error counts obtained are given in Table 1 together with the percent of residents for which errors were found.

Since it was thought that the number of residents for which change occurred may be affected by changes in respondents between the two surveys or the resident's discharge from the facility during the 8 week interval, tabulations were also made according to the respondent during the two surveys and according to the discharge status of the resident. These counts are also shown in Table 1.

It can be seen that errors occurred for at least 37% of the sample residents for which data on primary diagnosis at admission was available from both surveys. The error percentages shown for the other items range from 0 to 20 percent. If one defines 5 percent as the maximum amount of error that is permitted before data are labeled as being unreliable, then data for 6 of the 12 question items considered here would be labeled as unreliable on the basis of inconsistencies alone.

The error percentages for the sampled residents are two or more percentage points higher on 10 out of 12 items for those discharged than for those not discharged. Likewise, it is noted that errors occurred relatively more often on 9 of 11 items when the respondent was different between the two surveys. On the first 3 questions above, it is possible that the differences between the groups of residents could be affected by the lack of data for some of the 197 sample residents. However, in view of the relationships between error percentages shown for the condition items, where data are available for all the sampled residents, it appears that differences would probably still exist, and the tendency toward higher error percentages would likely continue for the residents who were discharged between surveys and the residents with different respondents.

These observations from the pilot study suggest that answers to questions about residents can vary with the date of interview. Part of the variation can result from a difference in the respondent that would be interviewed on different dates. But inconsistencies such as those found in the pilot test for the primary diagnosis question studied here suggest that variation in answers are possible even though the respondent were to remain the same.

During field observations made in the pilot test it was noted in regard to the primary diagnosis question, that several diagnoses could be recorded in a resident's file with no indication about which is the primary one. Hence, the respondent used judgement to pick the diagnosis most likely to be primary for the resident. On the basis of the pilot test data, it is evident that one's judgement of what should be primary can vary with time even though the correct answer remains the same.

4.2 Reconciliation Study

As indicated above, error could occur in the data reported during the two surveys for a resident without any indication of impossible changes such as those described above. A reconciliation study was conducted in a subsample of 9 facilities which contained 84 of the sampled residents. In these homes a copy of the questionnaire completed for each resident during the first survey was given to the interviewer for the second survey. After the second interview for each resident was completed, the two forms were reviewed and the respondent was asked to explain any differences in answers. Since there was a concern that the questionnaire design or some other item in the survey could be responsible for erroneous changes in responses, the respondent was also asked to identify the source of the error when an error was reported. The errors are tabulated in Table 2 by sources of errors.

In the reconciliation homes, a total of 1077 changes were noted over all the approximately 30 questions asked. Of these 527 or 48.9 percent were reportedly due to errors. Seventy of the erroneous changes were blamed on the questionnaire. A review of these cases revealed that at most 7 erroneous changes were blamed on any one questionnaire item. This implies that the questionnaire itself did not appear to be much of a problem to the respondents.

Forty percent of the erroneous changes were blamed on the "unknown". These changes probably include many of those that resulted because the respondent could not find the requested data in the resident's file or if the data were found, the information was not adequate. That is the respondent(s) had to rely on memory or judge which of the data that were in the file best described the resident's status. An example of such a question is the diagnosis question discussed earlier. Certainly, when "concrete" data is not available for the record, it is possible for a change of respondents or other circumstances present at the time of the survey to influence the responses given and, thus, any changes in responses that would be detected between two points in time.

About half (260) of the erroneous changes identified were blamed on the respondents and/or the interviewers.) This suggests that in a longitudinal study where questions are reasked verbatim about nursing home residents, if it were possible to design the survey and questionnaire in such a way that the only possible sources of error would be the participants in the interview, then it could be that as much as a third ($260 \div [1077 - 267] = .32$) of the changes detected would be due to error.

5. SUMMARY AND CONCLUSIONS

On the basis of the pilot test for a longitudinal study of residents in nursing homes, it is evident that when questions are simply repeated verbatim at two points in time, the result is that the changes detected in individual sample units are as likely to be due to error as not.

The experience indicates that the attention of the respondent for the second survey should be focused more specifically on change. One could simply ask whether a change had occurred since the last interview. In the NNHS such a question would require that the respondent know the resident's status at the time of the prior interview. Based on observations in the pilot test, the respondent may not always have that information.

It appears at least for the NNHS that a better procedure for measuring change would be to first tell the respondent what was recorded in response to the question during the first survey and then to ask what answer would apply at the time of the resurvey. A bias in responses may be introduced by informing the respondent about the past answer in that it may encourage some resurvey respondents to repeat the response that was recorded during the first survey. This would especially be true in the event that the respondent, who for some reason, is not absolutely sure of the correct answer.

The proposed procedure would at least force the respondent to think about whether a change has occurred since the earlier survey. That is, if no change has occurred, then the answer in the resurvey would have to be the same and if a change has occurred then the resurvey answer must be consistent with the answer given during the first survey.

Admittedly, the response recorded during the first survey may not be correct. In this case, results from the reconciliation study indicate that resurvey respondents may identify errors in the data recorded during the first survey so that changes will not be erroneously implied by the resurvey answer which they supply.

In any event it is expected that the proposed procedure would yield fewer erroneous changes in the resulting data.

TABLE 1. Error Counts for Each Question According to Whether Resident was Discharged and Whether Respondent for the Resident was the Same

(Numerators of ratios are error counts and denominators are number of residents for which useable data was available from both surveys for the question.)

Abreviated Question	Total	Resident Was Not Discharged	Resident Was Discharged	Respondent Was Same	Respondent Was Different
1. Primary Diagnosis at Admission	56/152= 37%	50/133= 38%	6/19= 32%	34/103= 33%	22/49= 45%
2. In Home One Month	9/180= 5%	5/163= 3%	4/17= 24%	6/133= 5%	3/47= 6%
3. Primary Payment Source at Admission	25/127= 20%	19/115= 17%	6/12= 50%	*	
4. Impairments or Condition					
a. Mental Retardation	6/197= 3%	3/172= 2%	3/25= 12%	6/142= 4%	0/55= 0%
b. Heart Trouble	13/197= 7%	10/172= 6%	3/25= 12%	6/142= 4%	7/55= 13%
c. Arthritis	26/197= 13%	22/172= 13%	4/25= 16%	15/142= 11%	11/55= 20%
d. Parkinson's Disease	4/197= 2%	3/173= 2%	1/25= 4%	2/142= 1%	2/55= 4%
e. Chronic Respiratory Disease	5/197= 3%	4/172= 2%	1/25= 4%	2/142= 1%	3/55= 5%
f. Diabetes	5/197= 3%	3/172= 5%	2/25= 8%	3/142= 3%	2/55= 4%
g. Permanent Stiffness	19/197= 10%	17/172= 10%	2/25= 8%	11/142= 8%	8/55= 15%
h. Missing Extremities	0/197= 0%	-	-	-	

*The financial questions about many residents were answered by someone other than the respondent for the remainder of the questionnaire but no records were made on whether the respondent to financial questions was the same or different between the two surveys.)

TABLE 2. Counts of Changes Due to Errors Identified in Reconciliation Study by Source of Error and Survey in Which Error Occurred

Source of Error	Total	Survey in which Error Occurred: Initial Survey	Resurvey	Both
Total Changes Due to Error	527	191	69	267
Respondent only	145	92	52	1
Interviewer only	111	94	17	0
Both Respondent and Interviewer	4	4	0	0
Questionnaire	70	1	0	69
Unknown	197	0	0	197

EFFECT OF CHANGING AGE COMPOSITION ON MEASURES OF MORTALITY FROM MALIGNANT NEOPLASMS, FOR WHITE AND ALL OTHER RACES: UNITED STATES, 1950-75

A. Joan Klebba, *National Center for Health Statistics*

First I shall review a few statements made long ago by some illustrious statisticians about the selection of a standard population for computing age-adjusted death rates. Then I shall examine with you the effect of the selection of the total population in the United States in 1940 as the standard population for almost all age-adjusted death rates (based on the "direct method") published by the National Center for Health Statistics. We shall base our examination on the effect on the age-adjusted death rate for malignant neoplasms.

Mr. G. Udny Yule began his paper on the use of indices for measuring occupational mortality, read before the Royal Statistical Society, November, 1933, with the question, "What do we want to do by standardization?"[1] His answer to his question was in part as follows: "The problem is simply to obtain some satisfactory form of average . . . an average which will measure in summary form the general fall in mortality between two epochs, just as an index-number measures the general fall or rise in prices."

Professor M. Greenwood in his laudatory discussion of Mr. Yule's paper said: "In the first section of his paper Mr. Yule defines the limitations of processes of so-called standardization, and I think it is important to bear those limitations in mind, because some of the mistakes and ambiguities to which he has drawn attention do arise from neglect of the fact that a pint pot cannot contain more than a pint."

Professor Greenwood reminded his audience that: "It was not until the seventeenth century that it was realized that in seeking to grasp everything, one tended to grasp nothing. Then we had the introduction to a statistical tabulation in which our predecessors a hundred years ago were engaged, and we merged individuals into groups, deliberately sacrificing some valuable information for the sake of retaining a clearer view of other important facts. What has happened is that the groups themselves and the information relating to those groups have become more and more detailed, and so we reached the stage when it was necessary to summarize the summaries and that is at the base of these various methods."

Dr. Percy Stocks, who, at this same meeting of the Royal Statistical Society, followed Professor Greenwood with a discussion of Mr. Yule's paper, cautioned the audience that in using either the direct or indirect method of age-adjustment with the 1901 population of England as the standard ". . . we are now over-weighting our standard death-rates with the mortality of childhood, for whereas the 1931 Census population contained 24 percent of children under 15, we weight their mortality as though they formed 32 percent, as they did in 1901. Since the greatest decline in mortality has occurred at these ages, this fact is over-represented in the fall of standardized death-rates, and the lack of improvement at the older ages is not sufficiently represented."

Before any attempt to determine whether Dr. Stocks' cautionary remarks about over-weighting England's age-adjusted death rates with childhood mortality are also applicable to the age-adjusted death rates published by the National Center for Health Statistics and its predecessor agencies, let us look at the summary of standard populations used by England and Wales and the United States, as given to us by Linder and Grove[2] in their momentous work entitled *Vital Statistics Rates in the United States 1900-1940.*

These authors state: "When age-adjusted rates were first published regularly in the English official reports of vital statistics (in 1883), the standard population used was that of the entire country at the previous census. The changing of the standard at the end of each decade was later felt to be unsatisfactory, so the age and sex distribution of the population of England and Wales in 1901 was taken as a 'permanent' standard near the beginning of the century. Despite the fact that it is no longer at all typical of the existing population, this distribution is still used in the Registrar-General's reports for age-adjustment of death rates by the direct method."

In this same work Linder and Grove report that for the United States the direct method of age adjustment was adopted by the Bureau of the Census in 1911, and that at that time "the rates were based on the same standard million of England and Wales in 1901, since it was felt that the results 'would be of greater value for comparison'."

The study of Linder and Grove appeared in 1973. In this work the authors chose for their standard population that of the total United States in 1940. Since then NCHS and its predecessors have clung as tightly to this standard as England and Wales have clung to the population of 1901.

As the charts I shall now show you indicate, the effect of this choice of the 1940 population on the level of the curves for age-adjusted death rates is quite different for the white male and female populations than for the male and female populations of Negro and other minority races (hereinafter denoted by "all other" or "all other races").

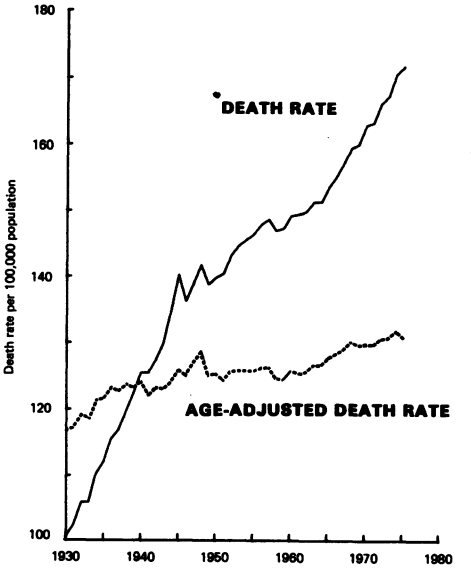
To better assess what information is summarized by the age-adjusted death rate for malignant neoplasms, let us first review what is summarized by the total or unadjusted death rate. For any given year this rate may be defined as the total number of deaths from malignant neoplasms in that year per 100,000 exposed to the risk of death throughout the year. The number of persons thus exposed is taken to be the enumerated population for census years and the estimated mid-year population for inter-censal years. The total death rate is clearly upward (figure 1).

These total or unadjusted death rates are undoubtedly valuable rates. They tell us the total probability of death from malignant neoplasms from the combined influence of all factors affecting death from this cause.

The following two well-known trends inform us why for malignant neoplasms the most important of these factors is the age distribution of the population under study: (1) Death rates for malignant neoplasms rise steeply with advance in age until

FIGURE 1

MALIGNANT NEOPLASMS: DEATH RATES AND AGE-ADJUSTED DEATH RATES: 1930-75



the end of the life span; and (2) since 1940 the proportion of persons at ages 45 years and over in this country has increased appreciably.

The first question we must answer, therefore, is whether the long upturn in the total death rate for malignant neoplasms is attributable for the most part to the increased proportion of older people in the population, or to increases in age-specific death rates for this cause. The measure we will use to answer this question is, of course, the age-adjusted death rate. Inasmuch as the standard population is the population of 1940, for that year the age-adjusted rate is the same as the total or unadjusted death rate for malignant neoplasms.

A comparison of age-adjusted death rates for malignant neoplasms with the corresponding total or unadjusted death rates for the period 1930-75 shows only a moderate rise in the age-adjusted death rates. It would be wrong, however, to conclude from this, as one might easily be misled to do, that the steep rise in the total or unadjusted rate is attributable almost entirely to the increasing proportion of older persons in the population. As will be shown below, this relative stability of the age-adjusted rate for the total population results for the most part from the offsetting of the steep rise in the age-specific death rates for the male population with the fall in the age-specific death rates for the female population.

Before 1940 the curve for the age-adjusted death rate is higher than that for the total death rate; and after 1940, progressively lower than that for the total death rate. This change in position results from the fact that the proportion of persons at the high risk ages of 45 years and over was lower before 1940 and progressively higher after 1940 than the proportion at these ages in the population of 1940. It follows, therefore, that since 1940 we are to an increasing extent over-representing in our age-adjusted death rate for the total population the low mortality from malignant neoplasms for children and young adults and under-representing the rising mortality for older age groups.

For the male population the total or unadjusted death rate for malignant neoplasms increased 34.6 percent during 1950-75, an average annual increase of 1.4 percent (figure 2). The corresponding unadjusted death rate for the female population also increased, but only 11.2 percent, an average annual increase of about 0.4 percent.

A different pattern emerges when this mortality is measured by age-adjusted rates. For the male population the trend is still upward; but for the female population the trend for the age-adjusted death rate is downward.

The trends of the age-specific death rates for malignant neoplasms for the male population make clear that the steep rise in mortality from this cause as measured by both the total or unadjusted death rates and by the age-adjusted rates is attributable for the most part, not to the ageing of the population, but to an increase in the force of mortality from this cause. For every age group in the span 45 years and over the male death rates for malignant neoplasms are substantially higher for 1975 than for 1950.

The gap between the total or unadjusted death rate and the age-adjusted rate for the male population widened during 1950-75—from 12.1 to 30.9 deaths per 100,000. This gradual enlargement of the gap reflects the slowly increasing percentage during 1950-75 of men in the total male population at ages 45 and over—from 28.2 to 28.9 percent.

In contrast the trends of the age-specific death rates for malignant neoplasms for the female population indicate that the moderate rise in mortality from this cause as measured by the total or unadjusted death rate is attributable in great part, if not entirely, to the ageing of this population, and not to an increase in the force of mortality from this cause. For every age group in the entire life span the female death rates for malignant neoplasms are substantially lower for 1975 than for 1950.

Also in contrast to the pattern for the male population, as a result of the decline in the age-adjusted rate, the gap between the total or unadjusted death rate and the age-adjusted death rate widened rapidly during 1950-75—from 16.0 to 43.8 deaths per 100,000. This enlargement reflects the increasing percentage during 1950-75 of women at ages 45 years and over—from 28.7 to 32.9 percent.

The total death rate and the age-adjusted death rate for malignant neoplasms rose during 1950-75 for both white males and for all other males (figure 3). But the increases were much greater for the latter group.

For white males the position of the curve for the age-adjusted death rate below that for the total rate beginning in 1940 reflects the fact that in their age-adjusted death rates we are under-representing mortality from malignant neoplasms at older ages. For example, whereas in 1975 white men at ages 45 years and over formed 29.8 percent of their population, we weight their mortality as though they formed only 26.7 percent as did persons at these ages in our standard population.

In contrast, for all other males the position of the age-adjusted death rate above the unadjusted death rate throughout 1914-75 (the longest period for which we have these rates by race) results from the fact that in their age-adjusted death rates we are over-representing mortality from malignant neoplasms at older ages. In 1975 men of races other than white at ages 45 years and over formed only 22.1 percent of their population. Yet we weight their mortality as though they formed 26.7 percent, as did persons at these ages in our standard population.

Before taking up our examination of age-adjusted death rates for the female population, let us pause briefly to recall the absolute number of men in the United States who lost their

FIGURE 2
MALIGNANT NEOPLASMS: DEATH RATES AND AGE-ADJUSTED DEATH RATES, BY SEX: 1950-75

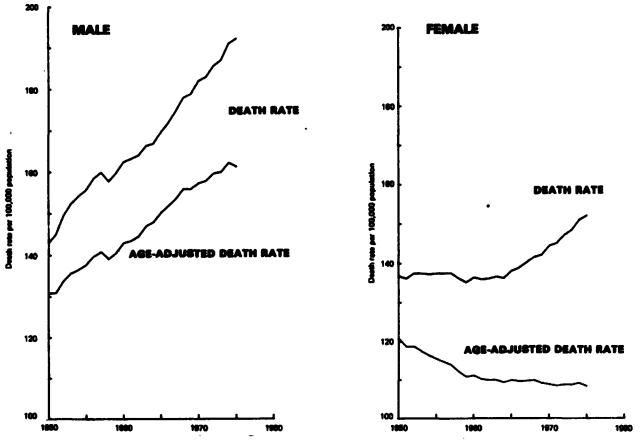


FIGURE 3
MALIGNANT NEOPLASMS: DEATH RATES AND AGE-ADJUSTED DEATH RATES FOR THE MALE POPULATION, BY COLOR: 1950-75

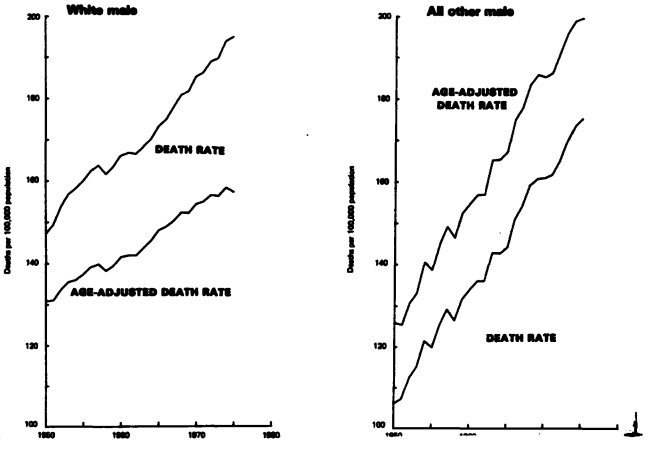
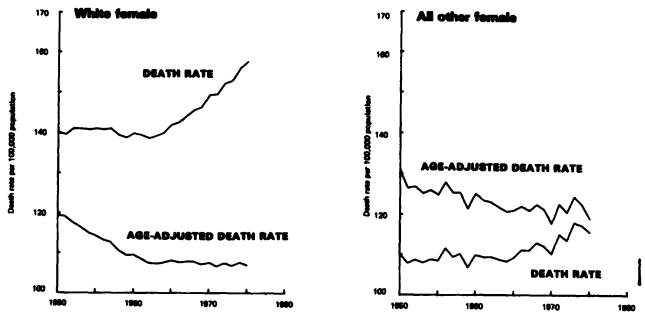


FIGURE 4
MALIGNANT NEOPLASMS: DEATH RATES AND AGE-ADJUSTED DEATH RATES FOR THE FEMALE POPULATION, BY COLOR: 1950-75



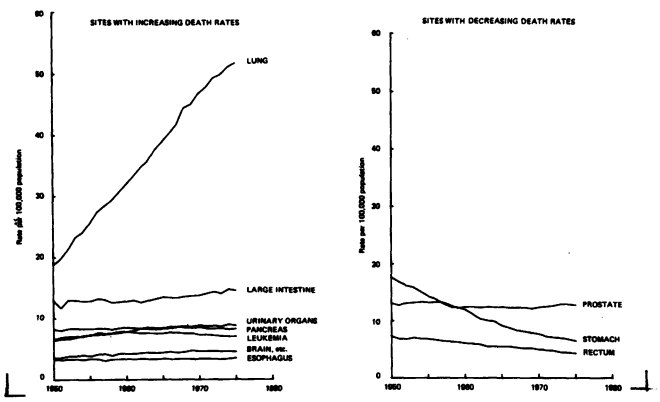
lives at ages 45-64 and 65-74 years—as a result of increases in the death rates for malignant neoplasms between 1950 and 1975 (figure 4).

Also for white females the total death rate for malignant neoplasms rose during 1950-75, especially during 1965-75 (figure 5). But as indicated by the negligible decline (1.1 percent) in the age-adjusted death rate between 1965 and 1975 (from 108.1 to 106.9 deaths per 100,000 population) most of this upturn in the total death rate is attributable to the ageing of the population. It is likely that their curve for the age-adjusted death rate will soon turn upward—pushed up by the rapidly increasing death rate among women for malignant neoplasms of trachea, bronchus, and lung.

Again, as for white males, the position of the curve for the age-adjusted death rate for white females below that for their total rate beginning in 1940, reflects the fact that in their age-adjusted death rates, we are under-representing mortality from malignant neoplasms at older ages. In 1975 white women at ages 45 years and over formed 34.2 percent of their population. Nevertheless we weight their mortality as though they formed only 26.7 percent as did persons at these ages in 1940.

For all other females the total death rate rose slightly, but the age-adjusted death rate declined between 1950 and 1975. The position of their age-adjusted death rate above the unadjusted death rate throughout 1914-75 results from the fact that in their age-adjusted death rate we are over-representing mortality from malignant neoplasms at older ages. Illustrating again with data for 1975, whereas in that year women of races other than white at ages 45 years and over formed only 24.4 percent of their population, we weight their mortality as though they formed 26.7 percent, as did persons at these ages in our standard population.

FIGURE 5
AGE-ADJUSTED DEATH RATES FOR WHITE MALES FOR LEADING SITES OF MALIGNANT NEOPLASMS: UNITED STATES, 1950-75



Space does not allow us to show mortality trends for both total and age-adjusted death rates for all body sites in which malignant neoplasms occur. Therefore we shall limit our remaining charts to age-adjusted rates for some of these sites. As may be expected, however, from the discussion above for mortality from all malignant neoplasms combined, for most of these body sites, for the white population the position of the curve for the total death rate is above that for the age-adjusted rate; and for the population other than white, the position of the curve for the total death rate is below that for the age-adjusted rate. This is true for both males and females of the white and other races (figures 6-9).

FIGURE 6
AGE-ADJUSTED DEATH RATES FOR ALL OTHER MALES FOR LEADING SITES OF MALIGNANT NEOPLASMS: UNITED STATES, 1950-75

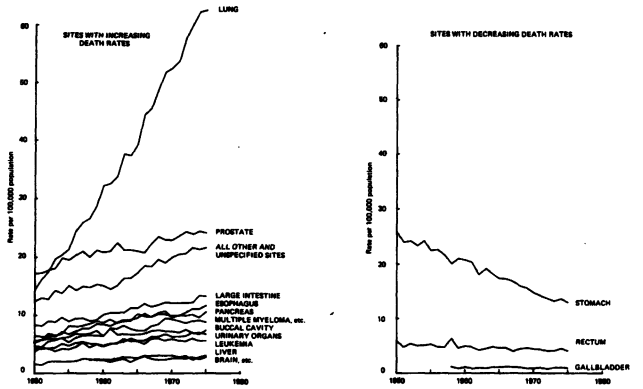


FIGURE 7
NUMBER OF MEN LOST AT AGES 45-64 AND 65-74 YEARS AS A RESULT OF INCREASES IN DEATH RATES FOR MALIGNANT NEOPLASMS BETWEEN 1950 AND 1975 BY COLOR: UNITED STATES

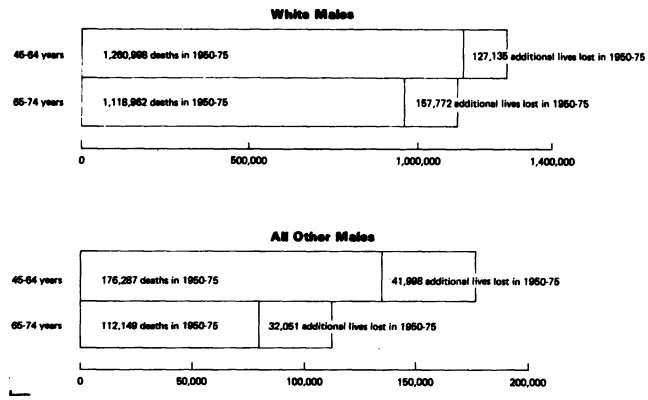


FIGURE 8
AGE-ADJUSTED DEATH RATES FOR WHITE FEMALES FOR LEADING SITES OF MALIGNANT NEOPLASMS: UNITED STATES, 1950-75

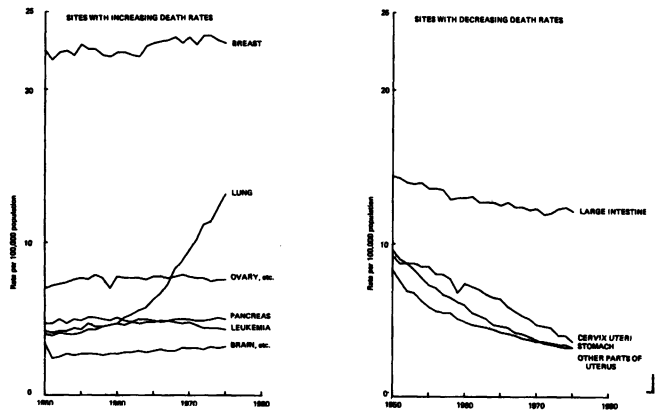
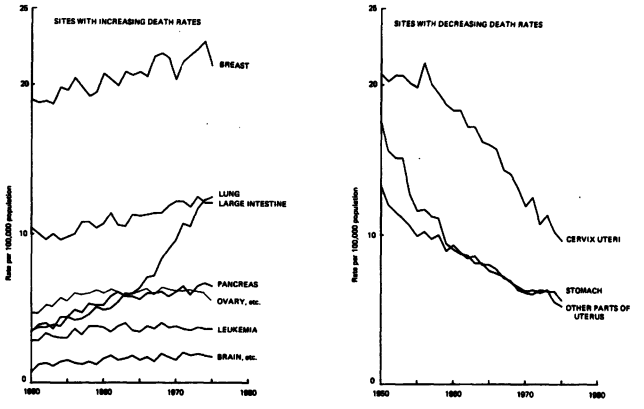


FIGURE 9
AGE-ADJUSTED DEATH RATES FOR ALL OTHER FEMALES FOR LEADING SITES OF MALIGNANT NEOPLASMS: UNITED STATES, 1950-75



Conclusion

Despite the fact that our standard population of 1940 is no longer typical of the existing total population, should we continue to use it? We have seen that the choice of the standard definitely influences the magnitude of the index. But why should we be concerned if the curve for the age-adjusted death rate is above or below the curve for the total or unadjusted-rate?

This writer agrees with Greenwood who had this to say "... If we could only persuade people that the whole process of standardization was merely to facilitate comparison, the psychological difficulty created by a choice of a wholly fictitious standard population would vanish."

References

[1] Yule G. U. *On some Points relating to Vital Statistics, more especially Statistics of Occupational Mortality*. Journal of the Royal Statistical Society, vol. 97, pp. 1-84. 1934.

[2] Linder, Forrest E. and Grove, Robert D., *Vital Statistics Rates in the United States, 1900-1940* (Washington, D.C.: Bureau of the Census, U.S. Government Printing Office, 1943).

ETHNIC DIFFERENCES IN MARITAL INSTABILITY

Dennis P. Hogan, Community and Family Study Center, The University of Chicago

Introduction

Ethnic differences in rates of marital instability have been observed by virtually every demographic and sociological study of the topic. Higher rates of marital disruption among blacks have not been accounted for by reference to a variety of social and demographic background variables with conventional methods of crosstabular analysis and ordinary least squares regression techniques. Sources of differential rates of disruption between Anglos and Spanish origin men have received only limited attention.

The research reported here answers two questions:

1. Can ethnic differentials in marital instability be accounted for by reference to additional characteristics of family background and early career?
2. Do log-linear models of analysis do better in accounting for ethnic differences in marital instability than have previously applied analytic methods?

Data

The data for this study are drawn from the 1973 "Occupational Changes in a Generation" Survey (OCG-II), which was carried out in conjunction with the March demographic supplement to the Current Population Survey (Featherman and Hauser, 1975). In 1973, the eight-page OCG questionnaire was mailed out six months after the March CPS and was followed by mail, telephone, and personal callbacks. The respondents, comprising 88 percent of the target sample, included more than 33,500 men aged 20 to 65 in the civilian noninstitutional population. Also, blacks and persons of Spanish origin were sampled at about twice the rate of whites, and almost half of the black men were interviewed personally. In this paper we shall examine factors that determine rates of marital disruption; therefore, we restrict our analysis to the ever-married men who compose the population at risk.

The OCG Survey is unique in providing an extensive account of demographic and family background data for large numbers of men so that separate analyses by ethnic groups are possible even though the event studied is relatively rare. The OCG-II data produce a pattern of racial differences of the sort expected—14.5 percent of the Anglos and 12.7 percent of the Spanish men experiencing a marital disruption because of separation, divorce, or widowhood as compared with 23.6 percent of the ever-married black men 20 to 65 years of age.

The CPS-OCG data do not include information about the cause of termination of first marriage for those men who have subsequently remarried. Men in their second or later marriage are counted along with those who are currently separated, divorced, or widowed as having experienced a disruption. Since date of termination of first marriage is not available, the dependent variable measures the prevalence (i.e., ever-occurrence) of a marital disruption. All models include marriage cohort (years since first marriage) as a control for differing length of exposure to the risk of a disruption.

Acknowledgements

Paper read at the annual meetings of the American Statistical Association (August, 1977) Chicago, Illinois. The research reported here is an outgrowth of a dissertation written at the University of Wisconsin-Madison, under the supervision of Professor David L. Featherman. I gratefully acknowledge his suggestions and assistance, as well as the comments provided by James Sweet, Larry Bumpass, Amy Tsui, and the anonymous reviewers. This research has been supported by National Science Foundation Grant GI-31604, "Occupational Changes in a Generation-II," and by NIH Training Grant GM01190 from the National Institute of General Medical Services (NIGMS). Computing facilities at the Center for Demography and Ecology were supported by the National Institute of Child Health and Human Development (NICHHD) grant IPOI-HD05876. Any opinions, findings, conclusions, or recommendations are solely those of the author and do not necessarily reflect the views of any of the agencies supporting this work.

Ordinary Least Squares Regression Models of Marital Disruption

Ordinary least square multiple regression models of marital disruption, by ethnic group, for ever-married men aged 20 to 65 as of March, 1973, are shown in Table 2. The models indicate that socioeconomic characteristics of family of origin have only trivial direct effects on marital disruption. Farm origin is the only exception—men from a home in which the father was primarily employed as a farmer or farm laborer experience rates of disruption about two and a half points lower among whites and over six points lower among blacks.

These trivial effects of socioeconomic origins on marital stability are not the result of collinearity among the independent variables. When measures of socioeconomic origin are entered into the equation in stepwise fashion the regression coefficients are no larger than in a model incorporating the entire set of family background factors. Even the zero-order regression coefficients are trivial.

The structure of the family of origin is of more relevance to the permanence of a man's own marriage. Net of socioeconomic standing of family of origin, growing up in a nonintact home reduces the chances of an intact marriage by about five points among whites and about two points among blacks.

Net of family background (including farm origin) and early socioeconomic attainments, region of birth is a factor of substantial relevance for later marital stability. Generally, men born in the Southern, Central, or Western United States experience higher rates of marital disruption than those born in the Northeastern United States or born abroad. The differences are significant among Anglos, with men from the Northeast experiencing rates of disruption about five points below other native-born men. The foreign-born enjoy rates about three points lower than men from the South, West, or Central States.

An additional year of schooling reduces the likelihood of a marital disruption about one point among Anglos. While the relationship of education to marital instability among Anglos is similar in magnitude to what has previously been estimated, it can hardly be termed massive.

First job status bears no relationship to the intactness of a man's first marriage among any of the ethnic groups. These same results characterize even the youngest men among whom the effects of first job on marital stability should be most apparent (Hogan, 1976: Chapter 5).

While years of schooling is of limited importance and first job of no importance for the stability of a man's first marriage, these models indicate rather large differences in levels of marital disruption between men who served in the Armed Forces and those who did not. Among Anglos, men who are veterans experience rates of disruption 3.7 points higher than other men. The size of the coefficients are smaller for blacks and Spanish ancestry men but there is no clear evidence that the consequences of military service differ by ancestry. Speculatively, we believe that the negative consequences of military service are specific to those men who were married either prior to or during their military service. Such couples would be more subject to frequent moves, prolonged absences of the husband from home, and any strains induced by on-base living. Higher rates of separation and divorce result among servicemen and among veterans to the extent that the initially heightened levels of marital discord persist after military service.

Finally, these models indicate that among Anglos the ordering in which a man finishes school, begins work, and marries is of considerable consequence for the stability of his marriage. A man who is still in school at the time of marriage or who returns to school after marriage experiences rates of disruption four points higher than men who have finished school and have begun a first job by the time of marriage. Men who have finished school by the time they are married but who either began work at their first jobs prior to completing schooling or after marriage experience rates two points higher than men who follow school with the beginning of work and then marry. The coefficients among the blacks and Spanish fluctuate somewhat but are not significantly different from those of the Anglos.

Since the ethnic groups differ in family background and early attainment compositions, the demographic and socioeconomic differentials in marital instability may at least partly account for ethnic differences in rates of disruption. Following the procedures of Sweet and Bumpass

(1974) the black and Spanish rates can be standardized by inserting the black means into the least squares regression estimates for Anglos. The linear regression models of Table 2 account for only 15 percent of the excess of black disruptions over Anglos, but all (105 percent) of the difference between Spanish and Anglos under this procedure. The differing compositions of the racial groups as regards family background and early career achievements are thus insufficient to explain the higher rates of marital instability among blacks.

Log-Linear Models of Marital Disruption

In a series of papers introducing log-linear modelling techniques to the discipline of sociology, Goodman (1970; 1971; 1972; 1976) has demonstrated that conventional techniques of cross-tabular analysis can produce fallacious interpretations of one's data. Linear multiple regression models are plagued with problems of possible bias and unreliability in situations where the dependent variable is dichotomous and the proportion in each category is outside the 25 to 75 percent range (Goodman, 1976; Knoke, 1975). The study of differentials in marital disruption is, of course, exactly such a case. How well do the findings reported here hold up under a more rigorous log-linear analysis? This is the question to which we now turn.

The log-linear model estimated (Table 3) includes as independent variables those factors most relevant in determining rates of marital disruption, as identified by the least squares regression model. The baseline model allows for the associations among the independent variables (marriage cohort, age at marriage, military service, parental structure, ancestry, region of birth, and education) and for the associations due to the proportion of the total population experiencing a marital disruption. The association unexplained by the baseline model (and measured by the chi-square statistic) is entirely due to the associations of the independent variables, either alone or jointly, with the dependent variable. The first model fit is a full additive structural model which includes a parameter for the direct association of each independent variable with marital disruption. The amount and proportion of the baseline chi-square statistic attributable to the direct effect of each variable, net of the direct effects of each other variable, is shown in Panel C. By far the largest component of association of these variables with marital disruption is accounted for by the direct association of rate of disruption with marriage cohort. This is followed in order by age at marriage, military service, parental structure, ancestry, region of birth, and education each of which has a significant direct association with disruption, net of the direct effects of all other variables.

The nature of the effect of each variable on rate of marital disruption is shown in Table 4. The gross effects are similar to zero-order regression coefficients. The full additive structural model is analogous to a logit model which incorporates the direct net effect of each independent variable on the dependent variable. The direct net effects are generally of the sort expected on the basis of the earlier linear models. The longer ago a man was married, the higher his chances of having experienced a disruption of his first marriage. The gross rates of disruption are 57 percent higher among men growing up in a nonintact home but this is reduced to 38 percent once compositional differences in regard to origin and early attainments are controlled.

The blacks are 65 percent more likely to experience a disruption than Anglos, while the Anglos have net rates only 6 percent higher than the Spanish. Net of controls for education, age at marriage, and ancestry, men who are foreign-born or born in the South have rates of disruption that differ rather little from the average. But the gross differentials for men born in the Northeastern United States persist net of controls, men from the Central and Western States 46 percent more likely to experience a disruption than men born in the Northeast. With the exception of birth in the Northeastern States, the region is which a man is born has considerably less effect on his chances of marital stability than is ordinarily believed (c.f., Carter and Glick, 1970).

Gross education differentials in rates of disruption are quite pronounced. However, men with the lowest education are especially likely to be in the earliest marriage cohort and/or are more likely to have married at an early age. When these sources of spurious and indirect effects are removed, the net effects of educational attainment on disruption are greatly reduced. The pattern is generally monotonic with a lower education inducing disruption. The single exception to this pattern is that men with some college who failed to receive their degrees have higher rates of disruption than either high school or college graduates.

The deleterious effects of service in the Armed Forces for stability of first marriage are enlarged when other variables are controlled. The gross differential of 25 percent is increased to a 39 percent higher rate of marital disruption among veterans when the effects of all other variables are statistically controlled.

Age at marriage has the expected inverse relation to disruption, but the big difference is for men married prior to age 21 as compared with all other men. Men married at this very young age have a net rate of disruption 71 percent higher than men marrying between age 21 and 24. These latter men, in turn, are only six percent higher in rates of disruption than men married at an even later age.

Tests for joint associations (interactions) among the independent variables in their effects on the dependent variables were performed using a forward stepwise procedure. Two interactions were observed to be significant net of the direct effect of each independent variable, as well as net of each other. The military service-age at marriage interaction is the larger of the two (Table 3). The older the age at marriage of a man, the less the importance of his service in the military for the stability of his marriage. Generally, the older a man's age at marriage, the less likely marriage is to have occurred prior to discharge from the service. This suggests that it is spending some of the years of early married life in the Armed Forces that strains the marital bond and increases the likelihood of a disruption, rather than any psychological concomitant of service in the military.

The second interaction observed is in the joint effects of marriage cohort and ethnic ancestry on marital disruption. The patterns are complex, but the major difference is between black men married 1920–47 and all other men. While the racial difference is particularly pronounced for earlier marriage cohorts, it is less evident for men first married after 1962. While this relationship may, in part, result from higher black rates of mortality (which produce progressively larger racial differentials in widowhood in the older ages), it seems more likely to be a result of a failure to locate separated and divorced young blacks for CPS-OCG interview.

Summary and Conclusions

The substantive results of the analysis in this paper have largely served to verify previous research findings about differentials in marital disruption. Thus, we have confirmed that there is a tendency to inherit a pattern of marital instability from the parental generation net of socioeconomic background factors. Men born in the Northeastern United States do experience lower rates of disruption than men born elsewhere and this differential cannot be explained by socioeconomic, demographic, or ethnic national origin factors. While higher levels of schooling generally decrease the risk of a marital dissolution, men who drop out of college suffer higher rates of disruption than men who complete either high school or college. Age at marriage likewise displayed its traditional inverse relationship with rates of separation and divorce.

New findings indicate that men who have finished school and are in the labor force at the time of marriage enjoy more stable marriages than men who spend a part of their married years in school or military service. The appropriate sequencing of events in the life cycle and especially the disruptive effects of military service on the timing and achievement of job status and marital stability are subjects especially worthy of further pursuit. These findings hold true with both traditional linear regression models of the determinants of marital instability, as well as with the new and more appropriate log-linear modified regression models.

Conventional linear regression models indicate no major differences among the ethnic groups in the impact of other socioeconomic and demographic variables on marital stability. Such models indicated that only 15 percent of the racial difference in rates of marital instability can be attributed to a wide variety of social and demographic characteristics. The relatively similar gross rates of marital instability characterizing the Spanish and Anglo ancestry men persisted with controls for other variables. These findings are unchanged when the mode of analysis is switched to a statistically more appropriate log-linear modified multiple regression models. These findings substantively suggest that the traditional characterization of Spanish origin people as having an especially strong family structure is essentially incorrect.

References

Bumpass, Larry L. and James A. Sweet

1972 "Differentials in Marital Instability: 1970." *American Sociological Review* 37 (December): 754–66.

1975 "Background and Early Marital Factors in Marital Disruption." Presented at the annual meeting of the American Sociological Association, San Francisco (August).

Carter, Hugh and Paul C. Glick

1970 *Marriage and Divorce: A Social and Economic Study.* Cambridge, Mass.: Harvard University Press.

Featherman, David L. and Robert M. Hauser

1975 "Design for a Replicate Study of Social Mobility in the United States." In K. C. Land and S. Spilerman (eds.), *Social Indicator Models.* New York: Russell Sage Foundation.

Goodman, Leo A.

1970 "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications." *Journal of the American Statistical Association* 65: 226–36.

1971 "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications." *Technometrics* 13: 36–61.

1972 "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables." *American Sociological Review* 37 (February): 28–46.

1976 "The Relationship Between Modified and Usual Multiple-Regression Approaches to the Analysis of Dichotomous Variables." Pp. 83–110 in D. R. Heise (ed.), *Sociological Methodology* 1976. San Francisco: Jossey-Bass Publishers.

Hogan, Dennis P.

1976 "The Passage of American Men From Family of Orientation to Family of Procreation: Patterns, Timing, and Determinants." Unpublished Ph.D. dissertation, Department of Sociology. The University of Wisconsin-Madison.

forthcoming "The Variable Order of Events in the Life Course." *American Sociological Review.*

Knoke, David.

1975 "A Comparison of Log-Linear and Regression Models for Systems of Dichotomous Variables." *Sociological Methods and Research* 3 (May): 416–34.

Sweet, James A.

1972 "Measurement of Trends and Socioeconomic Differentials in Marital Instability." Center for Demography and Ecology Working Paper 72–20, The University of Wisconsin-Madison.

1973 "Differentials in Marital Disruption." Center for Demography and Ecology Working Paper 73–28, The University of Wisconsin-Madison.

Sweet, James A. and Larry L. Bumpass

1974 "Differentials in Marital Instability of the Black Population: 1970." *Phylon* 35: 323–31.

U.S. Bureau of the Census

1971 *People of the United States in the 20th Century,* by Irene B. Taeuber and Conrad Taeuber (a Census Monograph). U.S. Government Printing Office, Washington, D.C.

1975a *Current Population Reports,* Series P-20, No. 287, "Marital Status and Living Arrangements: March 1975." U.S. Government Printing Office, Washington, D.C.

1975b *Current Population Reports,* Series P-20, No. 290, "Persons of Spanish Origin in the United States: March 1975." U.S. Government Printing Office, Washington, D.C.

Table 1: MEANS AND STANDARD DEVIATIONS OF DEMOGRAPHIC, FAMILY BACKGROUND, AND EARLY ACHIEVEMENT VARIABLES, BY ANCESTRY, EVER-MARRIED U.S. MALES BORN 1907–1952.

Variables[a]	Anglos		Spanish		Blacks	
	X	S.D.	X	S.D.	X	S.D.
Place of Birth						
Central, West	.403	.491	.230	.421	.911	.285
Northeast	.252	.434	.032	.176	.062	.240
South	.296	.457	.251	.434	.824	.381
Foreign	.049	.216	.487	.500	.025	.157
Parental Family Intact	.857	.350	.771	.420	.666	.472
Siblings	3.670	2.612	5.257	2.875	5.138	2.949
Father NILF	.056	.230	.106	.308	.078	.269
Father's Education	8.632	3.927	5.007	4.596	6.371	3.953
Father's Occupation, SEI	30.274	22.524	20.742	18.885	16.238	13.935
Farm Origin	.229	.421	.409	.492	.401	.490
Mother's Education	9.156	3.614	4.815	4.305	7.422	3.883
Family Income–100s	99.558	76.541	60.687	67.692	51.094	49.363
Education	11.985	3.044	9.043	4.193	9.901	3.602
Military Service	.580	.494	.289	.454	.415	.493
Ever Worked at First Job	.968	.175	.936	.245	.971	.169
First Job, SEI	33.542	24.676	24.193	20.381	20.149	18.029
Temporal Ordering						
Typical, schooling, job, marriage	.657	.475	.698	.459	.709	.454
Atypical, marriage follows school	.176	.381	.197	.398	.184	.387
Atypical, school follows marriage	.167	.373	.105	.306	.107	.310
Age at Marriage	23.579	4.977	23.854	5.745	23.634	5.831
Years Since First Marriage	18.598	11.727	15.311	10.650	17.666	12.073
Disrupted Marriage	.145	.352	.127	.333	.236	.425

[a]An intact parental family is one in which both parents are reported as having lived with the respondent most of the time up to his sixteenth birthday. Siblings refers to number of brothers and sisters. Father NILF is a dummy variable scored one if the respondent's head of family was not usually in the labor force. Education variables are scored in years of regular schooling completed (ranging from 0 for those with no schooling to 17 for those with one or more years of graduate or professional schooling). Occupations are scored using Duncan's index of socioeconomic status. Respondent's report of family income when he was age sixteen is inflated to 1972 dollars using the consumer price index for 1972 and the year of his sixteenth birthday. Military service is scored one if a man served six months or more on active duty in the regular armed forces, and zero otherwise.

Table 2: REGRESSION ANALYSIS[a] OF DISRUPTED FIRST MARRIAGE ON DEMOGRAPHIC, FAMILY BACKGROUND AND EARLY ACHIEVEMENT VARIABLES, BY COLOR, EVER-MARRIED U.S. MALES BORN 1907–1952.

Independent Variables	Anglos		Spanish		Blacks	
	b	Se(b)	b	Se(b)	b	Se(b)
Place of Birth						
Central, West	.....	.....	.....	.....	.....	.....
Northeast	-4.97	.79	-9.33	6.84	-7.12	5.24
South	.49	.75	-.23	3.39	-2.62	3.74
Foreign	-3.27	1.51	-.86	3.21	6.13	7.33
Parental Family Intact	-4.73	.91	-6.13	2.90	-2.10	2.27
Siblings	-.34	.13	-.24	.43	-.02	.36
Father NILF	-.77	1.38	-3.15	3.91	1.17	3.87
Father's Education	-.12	.12	.20	.40	-.23	.36
Father's Occupation, SEI	.00	.02	-.09	.08	-.07	.08
Farm Origin	-2.66	.84	-1.81	2.89	-6.23	2.44
Mother's Education	.25	.13	.10	.42	.22	.37
Family Income–100s	.01	.01	.03	.02	.01	.03
Education	-1.00	.15	-.30	.40	.33	.40
Military Service	3.74	.64	.75	2.88	1.89	2.18
Ever Worked at First Job	.43	1.83	4.25	5.29	-1.97	6.20
First Job, SEI	-.03	.02	.06	.07	-.05	.07
Temporal Ordering						
Typical, schooling, job, marriage	.....	.....	.....	.....	.....	.....
Atypical, marriage follows school	1.90	.86	4.80	3.34	4.80	2.75
Atypical, school follows marriage	4.21	.96	3.74	4.03	-.04	3.65
Age at Marriage	-.21	.06	-.39	.21	.16	.18
Years Since First Marriage	.38	.03	.50	.12	1.03	.10
R^2	.035		.054		.091	
Constant	26.28		16.35		13.14	

[a]Men who have experienced a disruption of first marriage are scored 100; all others are scored 0. Unstandardized (metric) coefficients are shown. See Table 1 for definitions of the independent variables. The sample cases have been weighted to reflect true population proportions. The estimated standard errors are based on sample frequencies that are adjusted to reflect departures from a simple random sample.

Table 3: MODELS OF SELECTED FAMILY BACKGROUND AND EARLY ACHIEVEMENT DETERMINANTS OF MARITAL DISRUPTION, EVER-MARRIED U.S. MALES BORN 1907–1952[a].

Model[b]	χ^2_{LR}	df	p	Δ	χ^2_H/χ^2_T
A. Baseline Model [D] [ERAPVMC]	1830.51	2159	> .5	8.02	100.00
B. Full Additive Structural Model [DE] [DR] [DA] [DP] [DV] [DM] [DC] [ERAPVMC]	1038.52	2144	> .5	5.16	56.74
C. Direct Effect Net of All Other Direct Effects					
1. [DC] (marriage cohort)	304.54	2	.000	1.38	16.64
2. [DM] (age at marriage)	142.25	2	.000	0.68	7.77
3. [DV] (military service)	51.33	1	.000	0.23	2.80
4. [DP] (parental structure)	35.90	1	.000	0.12	1.96
5. [DA] (ancestry)	45.47	2	.000	0.15	2.48
6. [DR] (region of birth)	43.63	3	.000	0.19	2.38
7. [DE] (education)	45.77	4	.000	0.25	2.50
D. Gross Effect of Each Three-way Parameter[c]					
1. [DAC] (ancestry-marriage cohort)	19.65	4	.000	0.08	1.07
2. [DVM] (military service-age at marriage)	20.16	2	.000	0.07	1.10
E. Net Effect of Each Three-way Parameter[d]					
1. [DAC] (ancestry-marriage cohort)	20.46	4	.000	0.10	1.12
2. [DVM] (military service-age at marriage)	20.97	2	.000	0.09	1.15
F. Full Structural Model [DE] [DR] [DP] [DAC] [DVM] [ERAPVMC]	997.89	2138	> .5	4.99	54.51

[a]The sample cases have been weighted to reflect true population proportions. The estimated sample frequencies have been adjusted to reflect departures from a simple random sample.

[b]D=Marital disruption (yes/no); E=Education (0–8/9–11/12/13–15/16–17+); R=Region of birth (South/Northeast/Central, West/Foreign); A=Ancestry (Anglo/Spanish/Black); P=Parental structure (intact/nonintact); V=Military service (non-veteran/veteran); M=Age at first marriage (less than 21/21–24/25 or older); C=Years since first marriage (i.e., marriage cohort) (1920–47/1948–61/1962–73). The notation indicates those marginal tables that are fit (i.e., used to predict cell frequencies) under that model. [D] indicates that the marital disruption margin for the entire table is fit. [DC] indicates that the marriage cohort by marital disruption marginal table is fit.

χ^2_{LR} is the likelihood ratio chi-square statistic.

df are the degrees of freedom.

p is the probability level that the chi-square statistic is due to chance.

Δ is the index of dissimilarity between the observed sample frequencies and the expected frequencies obtained with that model.

χ^2_H/χ^2_T is the percent of the baseline (total) chi-square accounted for by the chi-square statistic of that model.

[c]Only interactions significant ($p < .001$) net of the full additive structural model are shown.

[d]The effect of each interaction net of the full additive structural model and the other three-way parameter from Panel D.

Table 4: STRUCTURAL MODELS OF SELECTED FAMILY BACKGROUND AND EARLY ACHIEVEMENT VARIABLES ON MARITAL DISRUPTION, EVER-MARRIED U.S. MALES BORN 1907–1952[a].

Independent Variables	Gross Effects		Full Additive Structural Model[c]	
	$\beta=\ln\gamma$	γ	$\beta=\ln\gamma$	γ
(Intercept)	NA[b]	NA	-1.677	.187
Marriage Cohort				
1962–73	-.599	.549	-.579	.560
1948–61	.176	1.192	.172	1.188
1920–47	.423	1.526	.407	1.503
Ancestry				
Anglo	-.159	.853	-.148	.862
Spanish	-.244	.784	-.206	.814
Black	.403	1.496	.354	1.425
Region of Birth				
South	.287	1.333	.092	1.096
Northeast	-.225	.799	-.215	.806
Central, West	.168	1.183	.165	1.179
Foreign	-.231	.794	-.042	.959
Parental Structure				
Nonintact	.225	1.252	.163	1.177
Intact	-.225	.799	-.163	.850
Education				
0–8	.338	1.403	.162	1.176
9–11	.211	1.235	.094	1.098
12	-.085	.919	-.084	.919
13–15	.027	1.028	.120	1.127
16–17+	-.493	.611	-.292	.747
Military Service				
No	-.111	.895	-.163	.849
Yes	.111	1.117	.163	1.177
Age at Marriage				
0–20	.362	1.437	.377	1.458
21–24	-.173	.841	-.158	.854
25–65	-.189	.828	-.220	.803

[a]These estimated effects are net of the associations among the independent variables. The parameters shown refer to the estimated odds of having experienced a disruption of first marriage vs. having an intact first marriage.

[b]Not shown due to different intercepts for each set of coefficients shown below.

[c]This model results in a 5.1 percent reduction in the conditional uncertainty of marital disruption. The maximum reduction obtainable with this set of independent variables is 11.7 percent.

AN APPROACH TO QUANTIFYING BREAKTHROUGH BLEEDING IN ORAL CONTRACEPTIVE STUDIES

Prem P. Talwar, International Fertility Research Program

INTRODUCTION

Breakthrough bleeding has been reported as one of the important causes of dissatisfaction with fertility control methods. Its importance was amply indicated when the WHO Task Force on Acceptability of Fertility Regulating Methods recommended at its 1973 meeting that multinational social science research be initiated on patterns and perceptions of menstrual bleeding (1). Although the methodology and approaches adopted differ from one study to another, every study of contraceptive methods, especially those concerning IUDs and steroidal contraceptives, has described the phenomenon of bleeding associated with the contraception. In some studies, the bleeding is described quantitatively by indices such as the percentage of women reporting the symptoms, or the number of days or number of episodes of bleeding, or by more than one of these indices. In other studies, women are asked to report subjectively whether "increase", "decrease" or "no change" occurs in menstrual bleeding when the period of contraceptive use is compared to the period of precontraception (2,3,4,5). These different approaches in analysis, in addition to differences in the design of the study and conceptual definitions, have made it difficult, if not impossible, to compare results of different studies. In 1976, Rodriguez attempted to standardize the definitions and the method of analysis of the menstrual patterns (6). In a later workshop, problems related to the study of menstrual patterns were discussed more thoroughly, and most of the recommendations made by Rodriguez were accepted with or without modifications (7).

The framework of standardized methodology (6, 7) is more appropriate when the contraceptive method under study tends to disrupt the menstrual cycle completely. In that situation, the use of indices showing only bleeding episodes (which includes all types of bleeding--menstrual or intermenstrual) is valid. But with oral contraceptives (OCs), there are two distinct types of bleeding which occur cyclically; combining the two and ignoring their distinctiveness can in no way be justified. Equating the effects of one episode of breakthrough bleeding with withdrawal bleeding on acceptability and continuation is not only undesirable, but also objectionable. Moreover, there is a definite need for treating each contraceptive cycle, especially the earlier ones, separately because of (a) evidence of a stabilization trend in breakthrough bleeding during the use of the first few contraceptive cycles, and (b) higher differences in the incidence of breakthrough bleeding among different OCs during the first and second cycles of use (2). Therefore, it is recommended that for the study of oral and other contraceptives which have cyclical patterns of bleeding, a reference period should consist of one contraceptive cycle rather than a period of 60 or 90 days (6,7).

This paper presents an approach, differing from that of Rodriguez, for quantifying breakthrough bleeding in oral contraceptive users. A method of computing an index of breakthrough bleeding is discussed which not only can be used to compare bleeding patterns of different OCs but is also useful to relate this symptom to a woman's medical, physiological, and biological profile. It is assumed that there are two main dimensions of this symptom, namely, persistence and severity (Snowden, reference 7, has described them by duration of bleeding and volume of blood loss), and each dimension has more than one level. Thus, a multivariate observation consisting of two levels of persistence (number of days of bleeding and number of episodes of bleeding in a contraceptive cycle) and three levels of severity (spotting, and light and heavy bleeding) describes the breakthrough bleeding pattern of a woman. Three severity levels of bleeding have been considered instead of two (considered in references 6 and 7) because of the different effects they may have on acceptability and continuation rates. These three levels can be suitably defined or modified to reflect the degree of dissatisfaction in the cultural context of the country. This six-variate information was converted to a composite index because any index based on one of the variates was not found to be comprehensive enough to include information on all aspects of bleeding. The composite index thus derived was used to study differential patterns of breakthrough bleeding for three groups of women using Ovral (ethinyl estradiol 0.05 mg and dl-norgestrel 0.5 mg), Norinyl (mestranol 0.05 mg and norethindrone 1.0 mg), and Norlestrin (ethinyl estradiol 0.05 mg and norethindrone acetate 1.0 mg).

MATERIALS AND METHOD

The data were obtained in a double-blind comparative study on Ovral, Norinyl and Norlestrin carried out in 1974 at the Planned Parenthood Clinic of Seattle, Washington. A sample of 480 women who had no contraindications for OC use, including no irregularities in their menstrual cycles, and who had not used steroidal contraceptives in the preceding three months were randomly assigned to one of these study OCs and given three cycles of supply, each containing 21 hormonal and seven placebo tablets. Information on 27 symptoms associated with OC use was collected by a public health nurse who telephoned each subject every two weeks to ask whether she had experienced any symptom (which women associate with OC use) since her last contact. Inquiry was made about each such symptom in order to ensure the reporting of all pertinent events, and the information was recorded on a symptom grid by the day on which symptoms appeared (2).

Only the information on intermenstrual vaginal bleeding during the period hormonal pills were taken in the first cycle of OC use was

utilized in the analysis. Questions were asked about the day when breakthrough spotting or bleeding occurred, whether protection by tampons or pads was required, and in what quantity. Events of breakthrough bleeding were defined in this study as spotting, when no protection was needed; light bleeding, requiring one pad a day; and heavy bleeding, when two or more pads were used. Future studies may alter these definitions, when appropriate, to reflect the degree of dissatisfaction in the cultural context of the population under study.

The choice of a composite index based on both the dimensions of breakthrough bleeding was made because of its comprehensiveness in covering information on all aspects of bleeding; this is not possible if one or more levels of one dimension is used as an index (Table I). The first component of the Principal Component Analysis was used as an index. Because more than one such index can be obtained by adopting variations of the Principal Component technique on the same set of variables or by using different subsets of the variables, a choice of the "best" index was made based on the degree of information on different aspects of breakthrough bleeding contained in the index. (Eigenvectors vary by the choice of the matrix used to extract eigenvalues. Variance-Covariance matrix gives a set of eigenvectors different from the ones obtained from the correlation matrix.)

DEVELOPING AN INDEX TO QUANTIFY BREAKTHROUGH BLEEDING

Choice of variables

The combinations of two levels of persistence and three levels of severity resulted in the following six variables (X_1, $X_2 \ldots X_6$) which cover most of the information on breakthrough bleeding on an individual woman:

$X_{1(2,3)}$ = Number of days of spotting (light bleeding, heavy bleeding)

$X_{4(5,6)}$ = Number of episodes of spotting (light bleeding, heavy bleeding)

Because the first set of three variables are different in measurement units from the second, it was desirable either to convert them into comparable units or make them unitless (dimensionless) so that they might be combined into one index. For this purpose, the observed measurements (X's) were converted to percentages (P's) by relating them to the overall value for the population. (Because the purpose was to obtain dimensionless variables, any average value could have served.) That is, the new variables P_i's were obtained in the following fashion:

$$P_i = [X_i / X_i(A)] \times 100$$

Where:

X_i = observed value of the ith variable

$X_i(A)$ = average value of X_i $\left(\frac{\sum_{j=1}^{n} x_{ij}}{n}\right)$,

where n is the number of women on whom the information on the X_i variable was available

Thus, the redefined six-variate information for women used in the development of the index is:

$P_{1(2,3)}$ = Number of days of spotting (light or heavy bleeding) for a woman as the percentage of the average number of days of spotting (light or heavy bleeding) in the population under study

$P_{4(5,6)}$ = Number of episodes of spotting (light or heavy bleeding) for a woman as the percentage of the average number of episodes of spotting (light or heavy bleeding) in the population

Choice of several indices

Once these variables were made dimensionless, the next step was to look for a comprehensive index which includes the most information contained in the six variables. The first consideration was whether one of the variables could serve as the index. To test it, a correlation matrix consisting of correlation coefficients between different P's was computed. It may be seen (Table I) that a single variable is not sufficient to be an index because it does not contain information on the other variables. Thus, a need for a composite index was indicated. The first component of the Principal Component Technique on the six-dimensional vector ($P_1, P_2 \ldots P_6$) is a valid index, often used by statisticians to represent a multivariate observation more parsimoniously (8, 9, 10). Geometrically, this index is a linear combination of the variables which cover the maximum variance in the sample scatter configuration. For computations, BMD computer programs prepared by the University of California, Los Angeles, were used (10). The first principal component was obtained by deriving the weights from the eigenvector corresponding to the largest eigenvalue of the variance-covariance or correlation matrix of the multivariate observations. The breakthrough bleeding score for an individual was then obtained from the product of the vector of the standardized variables and the eigenvector. That is, if $(b_1, b_2 \ldots b_k \ldots)$ is an eigenvector corresponding to the largest eigenvalue, the breakthrough bleeding score for the individual is obtained as

$$\frac{b_1(P_{i1}-\bar{P}_1)}{\sigma_{P_1}} + \frac{b_2(P_{i2}-\bar{P}_2)}{\sigma_{P_2}} + \ldots + \frac{b_k(P_{ik}-\bar{P}_k)}{\sigma_{P_k}} + \ldots$$

where σ_{P_k} is the standard deviation for variable P_k, P_{ik} is the observed measurement (as derived for this study), and $\bar{P}_k$ is the mean measurement of the variable P_k. Two modifications were done to derive scores for this analysis.

Table I

CORRELATION COEFFICIENTS BETWEEN DIFFERENT VARIABLES AND INDICES

Variables & Indices	Variables P_1	P_2	P_3	P_4	P_5	P_6
P_1	1.000					
P_2	0.073	1.000				
P_3	-0.055	0.141	1.000			
P_4	0.744	0.125	-0.045	1.000		
P_5	0.104	0.707	0.123	0.197	1.000	
P_6	0.0002	0.184	0.820	-0.009	0.227	1.000
I_1	0.6850	0.656	0.198	0.718	0.681	0.305
I_2	0.317	0.693	0.595	0.369	0.733	0.688
I_3	0.125	0.695	0.778	0.062	0.513	0.635
I_4	0.303	0.576	0.430	0.445	0.811	0.659

(1) The scores were approximated as

$$\frac{b_1}{\sigma_{P_1}}(P_{i1}) + \frac{b_2}{\sigma_{P_2}}(P_{i2}) +\ldots+ \frac{b_k}{\sigma_{P_k}}(P_{ik})+\ldots$$

and

(2) $(\frac{b_1}{\sigma_{P_1}}, \frac{b_2}{\sigma_{P_2}}, \ldots)$ were so chosen that

$$\Sigma_k(\frac{b_k}{\sigma_{P_k}}) = 1$$

These two modifications were advantageous in that a woman whose breakthrough bleeding measurements for all six variables were equal to the average in the sample, would score 100.

Using this method, several indices could be developed from the six-variate information on breakthrough bleeding. This was done in two stages. In the first stage, two indices were developed by introducing slight variation in the Principal Component Technique--one of them used a variance-covariance matrix of the six variates to extract eigenvalues and eigenvector and the other used the correlation matrix. As we will see in the next section, the index corresponding to the correlation matrix was found to be better. Thus, in the next stage, correlation matrix was used to extract the eigenvectors. In this stage an attempt was made to determine whether all six variables were needed or whether a subset of them, either P_1 to P_3 or P_4 to P_6, could by itself lead to an index as good as the one based on P_1 to P_6. Thus, two more indices were computed, one based on (P_1, P_2, and P_3) and the other on (P_4, P_5, and P_6).

Choice of the "best" index of breakthrough bleeding

The "best" index is the one which contains most of the information on breakthrough bleeding available in each of the six variates (P_1, $P_2 \ldots P_6$). The operational meaning of this definition is that the index which has the maximum correlation with the individual variable $P_1, P_2 \ldots P_6$ will be the "best". Four indexes considered for the selection of the "best" were the following:

I_1: Index based on (P_1, $P_2 \ldots P_6$), eigenvector obtained from the variance-covariance matrix of the six variables

I_2: Index based on ($P_1, P_2 \ldots P_6$), eigenvector obtained from the correlation matrix of the six variables

I_3: Index based on (P_1, P_2, P_3), eigenvector obtained from the correlation matrix of the three variables

I_4: Index based on (P_4, P_5, P_6), eigenvector obtained from the correlation matrix of the three variables

The correlation coefficients between these indices and the individual variables (P_1, P_2, $\ldots P_6$) are shown in Table I. I_3 has very low correlation with P_1 and P_4 and thus can be omitted when compared to others. Between I_2 and I_4, I_2 is preferred for higher correlation with most of the variables P_1 to P_6, though I_4 has higher correlation coefficients with P_4 and P_5. I_2 is to be preferred over I_1 because of (1) overall higher correlation with all variables P_1 to P_6 and (2) higher correlation with the heavy breakthrough bleeding (variables P_3 and P_6), a more serious side effect. Thus I_2, the "best" index of breakthrough bleeding, is given by

$$I = 0.0676P_1 + 0.163P_2 + 0.200P_3 + 0.104P_4 + 0.205P_5 + 0.255P_6$$

The order of the magnitude of weights may be noted--variables corresponding to heavy bleeding get more weight than those corresponding to

Table II

DISTRIBUTION OF WOMEN USING OVRAL, NORINYL AND NORLESTRIN BY THE BREAKTHROUGH BLEEDING SCORE

Breakthrough Bleeding Score	Types of Oral Contraceptives		
	Ovral (%)	Norinyl (%)	Norlestrin (%)
0	128 (90.1)	72 (51.4)	66 (47.8)
0-10.0	2 (1.4)	3 (2.1)	3 (2.2)
10.0-20.0	5 (3.5)	8 (5.7)	7 (5.1)
20.0-30.0	4 (2.8)	20 (14.3)	20 (14.5)
30.0-40.0	1 (0.7)	9 (6.4)	9 (6.5)
40.0-50.0	0 (0.0)	6 (4.3)	7 (5.1)
50.0-60.0	0 (0.0)	8 (5.7)	8 (5.8)
60.0-70.0	0 (0.0)	4 (2.9)	8 (5.8)
70.0-80.0	1 (0.7)	1 (0.7)	3 (2.2)
80.0-100.0	1 (0.7)	3 (2.1)	4 (2.9)
Over 100.0	0 (0.0)	6 (4.3)	3 (2.2)
Mean Score (all users)	2.7	20.9	22.6
Mean Score (breakthrough bleeders)	27.5	43.0	43.4

light bleeding which, in turn, get more weight than those for spotting. Also, those corresponding to number of episodes are relatively more important than those related to total number of days.

APPLICATION

This index was used to obtain breakthrough bleeding scores for 480 women using Ovral, Norinyl, and Norlestrin in the Seattle study. The distribution of women by their scores and mean scores is given in Table II. The distribution of breakthrough bleeding scores for Ovral users was significantly different ($P<0.01$) from the similar distributions exhibited by Norinyl and Norlestrin users.

SUMMARY AND DISCUSSION

The usual indices of breakthrough bleeding, such as the percent of women reporting breakthrough bleeding and the duration or the number of episodes of bleeding, are not adequate to quantify breakthrough bleeding associated with OC use because they do not contain information on all aspects of bleeding. A composite index was therefore developed by using information on two levels of persistence and three levels of severity of breakthrough bleeding. It was found that more severe bleeding and more frequent episodes of bleeding (rather than days of bleeding) had higher weights in the index. This index was used to quantify breakthrough bleeding scores for women using Ovral, Norinyl, and Norlestrin; Ovral users had significantly different patterns, compared to Norinyl and Norlestrin users whose patterns were similar to each other.

Because breakthrough bleeding is an important factor contributing to dissatisfaction among OC users, it is recommended that more detailed data be collected to develop the composite index. The Principal Component Technique may be used to determine the coefficients of a linear combination of the variables because the coefficients may differ from one population to the other. Such an index can assign scores to individual women; the next step is to identify some medical, physiological, and biological variables which are positively associated with these scores. Such investigation will be helpful in better education and management of the OC acceptors.

Although the index developed here has been discussed in the context of oral contraceptive use, the technique is quite general and can be used for measuring breakthrough bleeding in any setting. It is generally applicable in any situation where multivariate data are to be presented in fewer dimensions to make them more comprehensible.

REFERENCES

1. Snowden, R: Report on a Planning Meeting of the Task Force on Acceptability of Fertility Regulating Methods. WHO, Geneva, 1973.
2. Ravenholt, R T, Kessel E, Speidel, J J, Talwar P P & Levinski M J: A comparison of symptoms associated with the use of three oral contraceptives: a double blind cross-over study of Ovral, Norinyl, and Norlestrin. Adv Plann Parent (in press).
3. Hines, D C & Goldzieher, J W: Large-scale study of an oral contraceptive. Fertil Steril 19: 841, 1968.
4. Mishell, D R & Freid, N D: Life table analysis of a clinical study of a once-a-month oral steroid contraceptive. Contraception 8:37, 1973.
5. Tatum, H J, Coutinho E M, Filho, J A & Sant'anna, A R: Acceptability of long-term contraceptive steroid administration in humans by subcutaneous Silastic capsules. Am J Obstet Gynecol 105:1139, 1969.
6. Rodriguez, G, Faundes-Latham A & Atkinson, L E: An approach to the analysis of menstrual patterns in the critical evaluation of contraceptives. Stud Fam Plann 7(2):42, 1976.
7. Snowden, R: The statistical analysis of menstrual bleeding patterns. Biosoc Sci 9:107, 1977.
8. Morrison, D F: Multivariate Statistical Methods, p. 221. McGraw-Hill, New York, 1967.
9. Talwar, P P: Developing indices of nutritional level from anthropometric measurements on women and young children. Am J Public Health 65:1170, 1975.
10. Dixon, W J (ed.): BMD: Biomedical Computer Programs, p. 193. University of California Press, Los Angeles, 1973.

ACKNOWLEDGMENT

This study was supported in part by the International Fertility Research Program, Research Triangle Park, N. C. (AID/csd-2979).

TYPE OF DELIVERY ASSOCIATED WITH SOCIAL AND DEMOGRAPHIC, MATERNAL HEALTH, INFANT HEALTH, AND HEALTH INSURANCE FACTORS: FINDINGS FROM THE 1972 U.S. NATIONAL NATALITY SURVEY

Paul J. Placek, National Center for Health Statistics

Five types of delivery (spontaneous, forceps, Cesarean section, breech, and other) are examined according to a wide variety of social and demographic, maternal health, infant health, and health insurance characteristics. Data are from the 1972 National Natality Survey, a 1 in 500 survey of legitimate live births linked with a mail followback survey of the mothers, physicians, and hospitals associated with those births. Maternal health factors include previous fetal loss, underlying medical conditions, complications of pregnancy, earliness and amount of prenatal care, complications or unusual conditions noted during each trimester, complications of labor, anesthetics used, duration of labor, pre- and post-delivery hospital stay, and postpartum care information. Infant health factors include period of gestation, birthweight, Apgar score, sex, multiple births, congenital malformations and anomalies, birth injuries, unusual resuscitative efforts required, discharge examination information, and whether the infant was discharged from the hospital alive. Health insurance characteristics of the mother include the amount of coverage, if any, for prenatal care, hospital bill, and doctor bill.

Of 2,818,000 legitimate live hospital births occurring in the United States in 1972, 52.7% were spontaneous, 36.8% were forceps, 7.3% were Cesarean section, 2.3% were breech, and 0.9% were other deliveries. Since national data on type of delivery has not been heretofore available, this study both provides baseline data and examines relationships not previously studied.

The 1972 National Natality Survey was designed by the National Center for Health Statistics to extend the scope of data which are collected through the registration system, to gather information comparable to that collected in previous U.S. national natality surveys to assess trend changes in birth related matters, and to evaluate the accuracy and completeness of selected items on the live birth certificate as reported through the vital registration system. Birth certificate records for 1972 from all 54 birth registration areas of the U.S. were divided into 6505 groups of 500 certificates each, and one certificate was selected from each of these 6505 primary sampling units. Births which were reported to be illegitimate (N=555) or inferred to be illegitimate by comparisons of names of father, mother and baby (N=261) were eliminated from this study. The remaining 5689 certificates comprised the sample. Additional information was secured as follows: (1) all mothers named on those certificates were mailed a questionnaire to obtain additional social and demographic information, a complete pregnancy history, detailed information about prenatal care received during the last pregnancy, and the mother's expectations of additional births; (2) if the attending physician and the hospital where the birth occurred had different addresses on the birth certificate, the physician was mailed a questionnaire to obtain medical information about the mother and her infant. Also, the hospital was mailed a questionnaire to assess additional medical information about the mother and her infant; (3) if the attending physician and hospital of birth had the same address, the hospital was sent one longer questionnaire which gathered all the same information as the physician and hospital questionnaire just discussed in section (2); (4) if the place of delivery was not a hospital but a physician was the attendant at birth, only the physician questionnaire discussed (2) was mailed. NNS data are weighted by means of a post-stratified ratio estimation procedure for age, race, and parity to reflect national estimates of 2,839,000 legitimate live births in the United States.

Since type of delivery data is from the hospital, only hospital births are included here. Thus, the data presented here refer to 2,818,000 legitimate live hospital births in the United States in 1972.*

* Additional tables, standard errors, etc. can be obtained from the author:

3700 East-West Highway
Center Bldg.
Room 1-44
Nat. Ctr. for Health Statistics
HRA
Hyattsville, MD 20852

Table 1. Percent distribution of type of delivery for mothers of legitimate live hospital births according to social and demographic characteristics: United States, 1972 National Natality Survey

Social and Demographic Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
All births	2,818	100.0	52.7	36.8	7.3	2.3	0.9
Color of mother[bc]							
White	2,490	100.0	51.9	37.9	7.0	2.4	0.9
All other	328	100.0	59.0	29.1	9.4	1.9	0.6
Metropolitan/non-met. county of residence[bc]							
Metropolitan	1,874	100.0	50.6	37.9	8.3	2.2	0.9
Non-metropolitan	944	100.0	56.9	34.7	5.2	2.4	0.8
Region of residence[bc]							
Northeast	603	100.0	54.0	35.8	7.0	2.5	0.6
North Central	775	100.0	55.3	34.9	7.2	1.9	0.8
South	940	100.0	51.2	38.4	7.2	2.3	0.8
West	500	100.0	49.8	38.1	7.8	2.6	1.7
Total family income[m]							
Less than $4,000	296	100.0	57.0	31.5	7.2	3.8	0.5
$4,000 to $6,999	537	100.0	54.1	36.7	6.8	1.6	0.7
$7,000 to $9,999	681	100.0	55.5	35.3	6.1	2.5	0.7
$10,000 to $14,999	818	100.0	50.2	39.0	7.4	2.1	1.3
$15,000 or more	487	100.0	48.8	38.7	9.2	2.2	1.1
Mother's education[m]							
None or elementary	121	100.0	63.4	27.9	5.3	3.0	0.4
1-3 years high school	475	100.0	56.0	34.3	6.5	1.8	1.4
High school graduate	1,348	100.0	52.5	37.4	7.4	1.9	0.9
1-3 years college	541	100.0	49.7	39.3	7.2	3.0	0.8
College graduate	334	100.0	49.7	37.5	8.8	3.2	0.7
Father's education[m]							
None or elementary	179	100.0	62.7	26.4	7.0	2.3	1.7
1-3 years high school	400	100.0	55.1	36.4	6.0	1.9	0.6
High school graduate	1,180	100.0	54.0	35.9	7.1	2.2	0.8
1-3 years college	467	100.0	48.0	39.3	8.7	2.7	1.3
College graduate	592	100.0	49.0	40.2	7.5	2.6	0.7

Table 1. Percent distribution of type of delivery for mothers of legitimate live hospital births according to social and demographic characteristics: United States, 1972 National Natality Survey (Cont'd.)

Social and Demographic Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
Age of mother[bc]							
Under 18 years	124	100.0	46.7	45.9	6.2	0.8	0.4
18-19 years	291	100.0	48.9	40.7	6.8	2.0	1.7
20-24 years	1,031	100.0	51.3	38.6	6.7	2.7	0.8
25-29 years	850	100.0	54.6	36.2	6.5	2.0	0.8
30-34 years	356	100.0	54.4	32.2	9.3	3.2	1.0
35 and over	166	100.0	59.1	25.9	12.1	1.7	1.2
Live-birth order[bc]							
First	1,072	100.0	39.1	49.3	8.7	2.0	0.7
Second	869	100.0	56.0	34.8	6.3	2.1	0.8
Third	428	100.0	63.0	27.0	6.4	2.5	1.1
Fourth	214	100.0	67.2	22.4	5.7	3.5	1.2
Fifth or higher	236	100.0	70.3	18.6	7.1	2.8	1.2
Interval between 1972 birth and previous live birth[m]							
1972 birth was one of a multiple birth	26	100.0	41.9	34.1	8.1	15.9	-
12 months or less	213	100.0	58.6	30.7	6.3	3.0	1.4
13-24 months	502	100.0	60.4	30.9	5.6	2.2	1.0
25 months or more	1,047	100.0	58.6	31.8	6.9	2.0	0.8
No previous live births	1,030	100.0	42.0	46.2	8.6	2.2	0.9

Sources of each variable:

bc = birth certificate
m = mother questionnaire
hp = hospital and/or physician questionnaire

* Figure does not meet standards of reliability or precision
\- Quantity zero

Table 2. Percent distribution of type of delivery for mothers of legitimate live hospital births according to maternal health characteristics: United States, 1972 National Natality Survey

Maternal Health Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
All births	2,818	100.0	52.7	36.8	7.3	2.3	0.9
Previous fetal losses[m,1]							
None	2,434	100.0	52.3	37.7	6.8	2.3	0.9
One	266	100.0	55.1	32.5	9.5	2.1	0.7
Two+	119	100.0	54.9	29.3	11.6	2.5	1.7
Underlying medical conditions[hp,2]							
None	2,427	100.0	53.1	37.3	6.4	2.3	0.9
One+	391	100.0	50.1	34.1	12.8	2.1	0.9
Complications of pregnancy[hp,3]							
None	2,359	100.0	54.2	36.9	5.9	2.1	0.8
One+	459	100.0	45.0	36.3	14.0	3.1	1.5
Trimester of pregnancy that prenatal care began[hp,4]							
First trimester	1,870	100.0	51.3	38.3	7.3	2.2	0.9
Second trimester	585	100.0	55.5	34.6	6.6	2.1	1.2
Third trimester	168	100.0	58.9	30.7	6.5	3.0	0.9
No prenatal care	195	100.0	51.9	35.4	9.2	3.2	0.3
Number of prenatal visits reported by medical sources[hp,4]							
No visits	195	100.0	51.9	35.4	9.2	3.2	0.3
1-4 visits	204	100.0	60.3	27.8	6.3	4.2	1.4
5-9 visits	809	100.0	56.7	33.6	6.6	2.0	1.1
10-14 visits	1,306	100.0	50.8	39.4	6.8	2.0	0.9
15-19 visits	285	100.0	45.5	41.3	9.9	2.8	0.5
20+ visits	18	100.0	42.8	40.6	16.6	-	-
Complications or unusual conditions noted during first trimester[hp,4]							
No prenatal care	195	100.0	51.9	35.4	9.2	3.2	0.3
No complications	2,393	100.0	53.6	36.9	6.5	2.1	0.9
One complication	200	100.0	43.9	38.1	13.8	3.5	0.8
Two+ complications	30	100.0	44.7	34.9	15.1	1.9	3.4
Complications or unusual conditions noted during second trimester[hp,4]							
No prenatal care	195	100.0	51.9	35.4	9.2	3.2	0.3
No complications	2,266	100.0	53.7	36.8	6.4	2.1	1.0
One complication	287	100.0	45.8	39.4	11.1	2.8	0.9
Two+complications	70	100.0	49.6	31.9	14.3	2.9	1.3

Table 2. Percent distribution of type of delivery for mothers of legitimate live hospital births according to maternal health characteristics: United States, 1972 National Natality Survey (Cont'd.)

Maternal Health Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
Complications or unusual conditions noted during third trimester[hp,4]							
No prenatal care	195	100.0	51.9	35.4	9.2	3.2	0.3
No complications	2,151	100.0	54.0	37.3	5.8	2.1	0.8
One complication	398	100.0	48.0	35.6	12.1	2.6	1.6
Two+ complications	74	100.0	42.8	33.8	18.8	2.7	2.0
Number of prenatal trimesters mother experienced complications[hp,4]							
No prenatal care	195	100.0	51.9	35.4	9.2	3.2	0.3
No trimesters	1,830	100.0	55.1	37.1	5.0	2.0	0.9
One trimester	589	100.0	48.7	36.2	11.2	3.0	0.9
Two trimesters	142	100.0	45.1	40.9	10.9	1.7	1.4
Three trimesters	62	100.0	40.2	32.1	21.1	4.0	2.5
Complications of labor[hp,5]							
None	2,248	100.0	58.1	37.3	2.3	1.6	0.8
One+	570	100.0	31.5	35.0	27.0	5.2	1.3
Type of anesthetic used[hp]							
Inhalation - Yes	970	100.0	55.3	33.9	7.8	2.1	0.9
No	1,848	100.0	51.3	38.4	7.0	2.4	0.9
Spinal and epidural - Yes	661	100.0	23.8	57.8	15.5	2.2	0.7
No	2,157	100.0	61.5	30.4	4.7	2.3	1.0
Local - Yes	653	100.0	76.7	20.2	0.4	2.2	0.6
No	2,166	100.0	45.4	41.9	9.3	2.4	1.0
Other - Yes	660	100.0	47.1	42.2	6.8	2.3	1.6
No	2,158	100.0	54.4	35.2	7.4	2.3	0.7
Number of anesthetics used for delivery[hp]							
None	203	100.0	85.3	9.3	1.0	3.8	0.7
One	2,296	100.0	48.9	40.1	7.9	2.2	0.9
Two+	320	100.0	58.9	30.8	7.0	2.2	1.1
Total duration of labor[hp]							
0-3 hours	647	100.0	55.9	27.7	12.2	3.2	1.1
4-7 hours	1,109	100.0	56.2	36.6	4.1	2.1	1.0
8-11 hours	565	100.0	47.2	45.0	5.4	1.8	0.5
12+ hours	497	100.0	46.8	40.0	10.1	2.1	0.9

Table 2. Percent distribution of type of delivery for mothers of legitimate live hospital births according to maternal health characteristics: United States, 1972 National Natality Survey (Cont'd.)

Maternal Health Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
Pre-delivery hospital stay[hp]							
Less than one day	2,158	100.0	55.0	37.0	4.4	2.6	0.9
One day	582	100.0	44.9	36.6	16.4	1.2	0.9
Two or more days	78	100.0	45.3	33.9	17.6	2.5	0.6
Post-delivery hospital stay[hp]							
0-2 days	554	100.0	65.2	30.6	0.8	2.7	0.8
3 days	1,016	100.0	56.0	39.9	0.6	2.5	1.1
4 days	693	100.0	50.4	42.9	3.6	2.2	0.8
5 or more days	556	100.0	37.0	30.0	30.5	1.8	0.7
Total hospital stay of mother[hp]							
0-2 days	438	100.0	66.5	29.2	0.6	2.8	1.0
3 days	915	100.0	57.7	38.0	0.5	2.6	1.1
4 days	744	100.0	51.6	43.6	1.7	2.4	0.6
5 or more days	722	100.0	39.0	33.0	25.6	1.6	0.9
Complications to mother's health noted after delivery[hp]							
Yes	187	100.0	38.3	36.0	20.1	4.6	1.0
No	2,632	100.0	53.7	36.9	6.4	2.1	0.9
Postpartum sterilization of mother[hp]							
Yes	220	100.0	50.9	22.9	23.5	1.6	1.1
No	2,598	100.0	52.8	38.0	5.9	2.4	0.9
Interval between delivery and first postpartum visit[hp]							
Under 30 days	592	100.0	47.9	34.2	14.5	2.3	1.1
30-35 days	1,710	100.0	52.2	38.9	5.8	2.3	0.8
60-90 days	154	100.0	61.0	33.9	2.6	2.0	0.6
No postpartum visits reported	362	100.0	59.2	32.7	4.3	2.7	1.2
Reason for and number of postpartum visits[hp]							
One routine visit	887	100.0	53.6	38.6	4.9	2.1	0.8
One non-routine visit	724	100.0	52.4	37.8	6.7	2.6	0.5
Two routine visits	152	100.0	47.8	34.2	13.7	2.3	2.0
Two non-routine visits	90	100.0	50.6	38.8	8.9	0.5	1.1
Two visits, one non-routine	357	100.0	52.3	34.6	9.6	2.0	1.5
Three+ routine visits	30	100.0	44.8	38.2	13.5	1.7	1.8
Three+ visits, incl. one+ non-routine	217	100.0	45.0	37.6	14.1	3.0	0.2
No postpartum visits reported	362	100.0	59.2	32.7	4.3	2.7	1.2

Table 3. Percent distribution of type of delivery for mothers of legitimate live hospital births according to infant health characteristics: United States, 1972 National Natality Survey

Infant Health Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
All births	2,818	100.0	52.7	36.8	7.3	2.3	0.9
Birth weight in grams (lbs. & oz.)[bc]							
2500 grams (5 lb. 8 oz.) or less	197	100.0	52.1	26.0	11.6	9.4	0.9
2501-3000 grams (5 lb. 9 oz. to 6 lb. 9 oz.)	493	100.0	55.2	34.1	7.1	2.6	1.0
3001-3500 grams (6 lb. 10 oz. to 7 lb. 11 oz.)	1,093	100.0	53.1	37.1	7.0	2.0	0.9
3501-4000 grams (7 lb. 12 oz. to 8 lb. 13 oz.)	742	100.0	50.2	41.3	6.4	1.0	1.0
4001 grams (8 lb. 14 oz.) or more	294	100.0	53.4	36.9	7.7	1.5	0.5
Period of gestation[hp]							
36 weeks or less	273	100.0	56.9	25.8	9.7	6.3	1.2
37-39 weeks	1,093	100.0	54.4	34.2	8.4	2.0	1.0
40 weeks	642	100.0	48.9	43.4	5.3	1.7	0.8
41 weeks or more	810	100.0	51.9	38.9	6.5	1.8	0.8
Sex of child[bc]							
Male	1,456	100.0	51.7	37.3	8.0	2.0	1.0
Female	1,363	100.0	53.7	36.3	6.5	2.7	0.8
Number at birth[hp]							
Single	2,761	100.0	53.0	36.9	7.2	1.9	0.9
Plural birth	57	100.0	37.0	33.1	9.3	19.8	0.8
Order of presentation at birth[hp]							
Single birth	2,761	100.0	53.0	36.9	7.2	1.9	0.9
First	28	100.0	46.8	36.9	11.3	5.0	-
Second or higher	29	100.0	27.3	29.4	7.4	34.3	1.7
Congenital malformations or anomalies reported on birth certificate[bc]							
No stated condition	2,647	100.0	52.4	37.0	7.4	2.2	0.9
Any stated condition	20	100.0	53.7	32.0	9.5	4.7	-
Not on State's certificate	151	100.0	56.6	34.8	4.2	3.0	1.3
Congenital malformations or anomalies noted at delivery by hospital[hp]							
No	2,686	100.0	52.9	37.0	7.0	2.2	0.9
Yes	133	100.0	49.0	33.5	12.4	4.6	0.4

Table 3. Percent distribution of type of delivery for mothers of legitimate live hospital births according to infant health characteristics: United States, 1972 National Natality Survey (Cont'd.)

Infant Health Characteristics	Number in Thousands	Type of Delivery[hp]					
		Total	Spontaneous	Forceps	Cesarean section	Breech	Other
Birth injuries noted at delivery[hp]							
No	2,757	100.0	53.3	36.2	7.3	2.3	0.9
Yes	62	100.0	24.7	66.2	4.9	3.3	0.8
Unusual resuscitative efforts required[hp]							
No	2,609	100.0	53.5	36.8	6.9	2.0	0.9
Yes	209	100.0	42.9	37.5	11.6	6.8	1.2
Apgar score - one minute[hp]							
Not done	467	100.0	57.7	32.1	7.1	2.2	1.0
0-3	41	100.0	43.1	17.2	25.0	13.6	1.1
4-7	310	100.0	44.2	39.0	8.6	6.9	1.3
8-10	2,000	100.0	53.0	38.0	6.7	1.4	0.8
Apgar score - five minutes[hp]							
Not done	1,146	100.0	56.2	34.9	6.3	1.8	0.9
0-3	16	100.0	42.0	12.8	22.9	15.8	6.5
4-7	52	100.0	39.5	33.6	16.5	8.6	1.8
8-10	1,605	100.0	50.7	38.6	7.5	2.3	0.9
Age when baby was first examined outside the delivery room[hp]							
One hour or less	911	100.0	52.9	32.4	10.7	3.0	1.0
2-6 hours	584	100.0	54.3	35.5	6.5	3.1	0.7
7-23 hours	655	100.0	51.3	41.3	4.7	1.5	1.1
24 hours or more	669	100.0	52.3	39.7	5.8	1.5	0.7
Birth injuries noted before discharge from hospital[hp]							
No	2,810	100.0	52.7	36.8	7.3	2.3	0.9
Yes	*8	100.0	57.4	42.6	-	-	-
Congenital malformations or anomalies noted before discharge[hp]							
No	2,803	100.0	52.7	36.8	7.3	2.3	0.9
Yes	15	100.0	45.3	45.0	3.3	6.3	-
Any other illnesses noted before discharge[hp]							
No	2,806	100.0	52.7	36.9	7.3	2.3	0.9
Yes	13	100.0	58.7	33.9	7.4	-	-
Infant discharged from hospital alive[hp]							
Yes	2,791	100.0	52.7	37.0	7.3	2.1	0.9
No	27	100.0	49.6	23.7	7.9	18.7	-

Sources of each variable:

bc = birth certificate
m = mother questionnaire
hp = hospital and/or physician questionnaire

* Figure does not meet standards of reliability
- Quantity zero

ESTIMATION OF FERTILITY RATE WITH OPEN INTERVAL DATA

P. T. Liu, A. W. Kimball & L. P. Chow
The Johns Hopkins University

1. Introduction

Measurement of fertility changes before and after implementation of family planning is essential for better program planning, management, and evaluation. However, in most of the developing countries in which large scale family planning programs are in operation, poor or non-existent vital statistics and registration systems are the rule rather than the exception.

Various methods and techniques for determining the fertility rates using data of poor quality or from sources other than vital registration and census have therefore been developed by demographers. Such approaches include child-ever-born ratios, the reverse survival method, pregnancy history analysis, and own children living with mother. Manuals prepared by the United Nations present a number of methods for estimation of fertility from incomplete data.(1) Another manual prepared by Bogue and his associate described many of these measures. (2)

Various fertility indicators have also been developed to detect changes in fertility level. These indicators do not measure the fertility of the population; rather, their changes reflect changes in fertility. Examples of such indicators include the age-parity distribution of annual births, age-parenthood status distribution, proportional fertility ratios, cumulative fertility for women over 30; proportion of women who are currently pregnant, live birth pregnancy rate, mean birth intervals, and mean open intervals.

All of these measures or indicator of fertility are useful but their utility depends on the type of populations. Most of these measures need accurate age of mothers and children - data difficult to collect from an illiterate population in the rural areas of developing countries. Moreover, a long recall period is frequently required as, for example, in the pregnancy history analysis technique.

There is pressing need for the development of a simple technique for estimating the fertility of a population. Such a technique would require relatively little information and the information would be of the type that most respondents are able and willing to report. Open interval appears to come close to such requirements.

2. Review of Literature

A number of researchers have discussed the utility of the mean open interval as a fertility indicator. Mohapatra (1966) investigated the relative importance of wife's age at marriage, length of the completed birth intervals, and length of open intervals in explaining the fertility differentials by socio-economic status. He found that among women over 30 years of age, modernization is likely to be more strongly associated with the length of open intervals.(3)

Srinivasan (1966) used the open interval as an index to detect fertility change.(4) In 1967, he further investigated the distribution of open intervals for women under three sets of assumptions concerning the rate of occurrence of births of a specified order and the parity progression ratio, and estimated the first and second moments of the distribution under each of the assumptions.(5) In the same year, he published another article proposing a method for study of interval between live births. Such a study would be applicable to the cases for which data collected in a survey are limited to information about the last two live births. This method of observation yields two kinds of intervals: the birth interval and the open interval. He assumed that the open interval is part of a complete interval from the last live birth to the time immediately proceeding the next birth. Therefore, the open interval is a random segment, which may be assumed to be uniformly distributed within the birth interval. From the rectangular (or uniform) distribution, the first and second moments of the open interval distribution can be obtained.(6)

Leridon (1970) made some comments on the Srinivasan's article, pointing out that the Srinivasan's estimation is based on the assumption that the distribution of open intervals from a survey has a mean equal to one-half of the mean birth interval. However, he proved that the longer the interval from the last live birth to the next, the more likely it is to be included in the survey. Therefore, the mean open interval including the survey point must be greater. In other words, Srinivasan's method under-estimates the mean open interval.(7)

Sheps (1970,1973) et al investigated the truncation effect and problems of interval analysis through computer simulation. They found that the mean open interval does not properly reflect the fertility change, and doubt that the current emphasis on securing such data is justified. (8,9)

Pathak (1971) developed a stochastic model for the study of open interval and reported that by taking account of parity progression variation, the open interval can be shown to predict the current fecundability, and thus, fertility of the women.(10)

Venkatachaya (1972) pointed out the weakness of using the mean open interval as a fertility indicator. His criticism was that the mean open interval does not properly reflect the effect of long-term and continuous use of a less than perfect contraceptive; it will only show the effect of a contraceptive method used since

the last live birth. When the mean open interval is used without adjustment it is not sensitive to changes in fertility. He indicated that the mean open interval standardized by age-parity distribution, might provide a more useful measure of fertility changes.(11)

More recently, Hastings and Robinson replicating and expanding an earlier study of Srinivasan on the open interval reported that "the open interval is more sensitive as an index of marital fertility when marital duration and parity are controlled than when mother's age and parity are controlled."(12)

In spite of some drawbacks, mean open interval is a fertility indicator frequently used in evaluating family planning programs impact. No attempt, however, has been made to convert change in the length of mean open interval into change in fertility rate. Potter (1968) mentioned that if acceptors of programmed contraception exhibit a consistently longer open interval than a matched sample of couples outside the program, then there is little question but that these participants are lowering their fertility. However, he said, "... there is no way to translate a change in mean open interval into an estimate of births averted."(13)

Venkatachaya (1972) also mentioned that the data on open intervals have been collected on a longitudinal basis in some standard fertility surveys in India, but they do not appear to have been used for an analysis of fertility.(11)

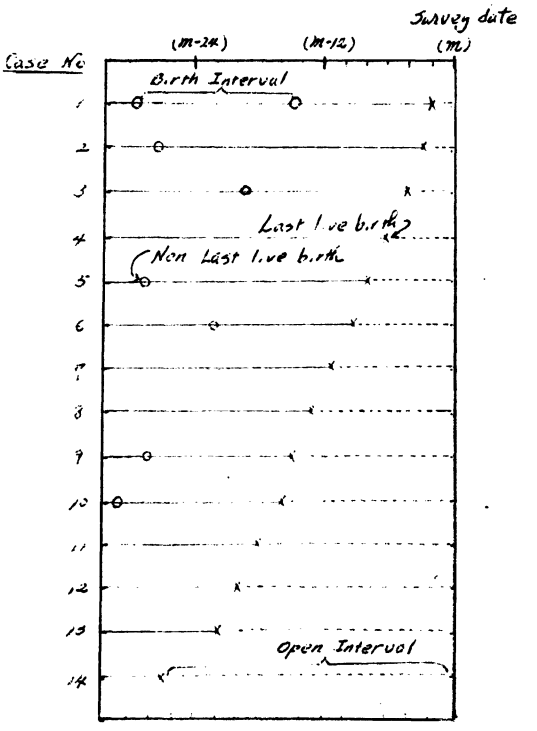
A considerable amount of work has been done on the own children method of estimating fertility. One study by Rindfuss (1976) compared the annual fertility rates obtained from census data on own children and the corresponding rates obtained from the vital statistics data for the United States during 1964 and 1970. He reported that the agreement between the rates obtained from these two sources was greater when own children rates were not adjusted for under-enumeration of women by the census. He also found that the estimated adjusted age-specific rates for the younger age groups were consistently lower than the recorded rates, and the estimated rates for the older age groups were consistently higher than the recorded rates.(14)

3. Rationales for the Current Study

Assume that a random sample of women of childbearing age of size m is drawn, and that a survey is conducted at the end of month m, which includes the following question:

"When was your last child born?" (or,
When did your last live birth occur?)

Assume further that the respondent is requested to answer the question by telling the interviewer the month and year of her last live birth, the following diagram depicts the live births and birth intervals schematically, showing the information obtained from the survey.



The characteristics of the data shown in the diagram may be expressed in different ways. One may compute the number of live births within a specified time interval, which is an essential element in calculating fertility rates. Alternatively, we may compute the duration between two successive live births (i.e., birth interval) or the interval between the date of the last live birth and the date of the survey (i.e., open interval). Since fertility rate and birth or open intervals are derived from a same set of information concerning live births, these three parameters are mutually related to one another. The fundamental rationale of this study is based on the above relationship, which provides a basic converting information on fertility measures.

In other countries where national family planning programs are being implemented, nationwide family planning surveys (or KAP surveys) are usually conducted at two to three year intervals. Questions concerning the respondents' reproductive history are usually asked, and at least one question about the date of the last live birth (or age in months of the last child) will be asked. Open interval information, therefore, is usually available from this type of study.

The advantage of use of open interval data for the estimation of fertility is self-evident; the data are more easily obtainable and with relatively higher accuracy because:

(1) The recall period is shorter, extending only to the last live birth.

(2) The question is asking for a clearly identifiable event, namely a live birth. There may be some ambiguity between stillbirth and

live birth when a baby dies immediately after delivery, but a few supporting questions should minimize the errors.

(3) The information can be obtained by asking a simple and short question taking very little time for the respondents to answer.

(4) The question is essentially non-sensitive, and there is little reason for the respondents to refuse to answer.

(5) The event of last live birth can always be related to a major event which is common to partically all cultures, i.e., a new-year celebration. In a community where most people are illiterate, the question may be modified: e.g., "Was your last child born before or after the last new year festival?" or "Was your youngest son or daughter born before the last new year festival, or the one before the last?".

4. Methods and Procedures

All the live births occurring in any one calendar year may be classified into two mutually exclusive categories: "last live births" and "non-last live births." The number and distribution of last live births occurring in a year is known from the open interval data. The problem of estimation of fertility rate, therefore, is simplified to the estimation of the distribution of "non-last live births" in each calendar year, which is unknown. For this purpose, some assumptions are needed:

(1) First, it is assumed that no two consecutive live births will occur within nine months (we ignore multiple births at this point). In other words, the birth interval must be greater than nine months, or the probability of getting another live birth within nine months after delivery is assumed to be "zero."

(2) Secondly, birth intervals are distributed as a certain function which depends on the fertility at the end of a birth interval. (Retrospective or backward approach rather than perspective or forward approach in estimating fertility.)

Let n_i be the number of last live births month i;

$\bar{n}_i$ be the number of non-last live births at month i;

N_i be the total live births at month i;

T_i be the corresponding number of women at month i; and

f_i be the fertility rate at month i;

then, $N_i = n_i + \bar{n}_i$,

and $f_i = \frac{N_i}{T_i}$

For simplicity, we further assume that within a same birth interval (excluding the duration of gestation), the probability of becoming pregnant in each month is the same. (The distribution of birth intervals is not necessarily restricted to an exponential function. It is also possible to assume an unequal probability of conception during each month.)

Since it is impossible for two consecutive live births to occur within nine-month period, hence,

$$\bar{n}_i = 0 \qquad \text{for } i = m-9, m-8, \ldots m$$

or $N_i = n_i$ for $i = m-9, m-8, \ldots m$, and

and $N_i = n_i + \bar{n}_i$ for $i = 0, 1, 2 \ldots, m-10$

$$f_i = \frac{N_i}{T_i} = \frac{n_i}{T_i} \qquad \text{for } i = m-9, m-8, \ldots m$$

$$\bar{n}_{m-10} = N_m \cdot e^{-f_m}$$

$$N_{m-10} = n_{m-10} + \bar{n}_{m-10}$$

$$f_{m-10} = \frac{N_{m-10}}{T_{m-10}}$$

$$\bar{n}_{m-11} = N_m \cdot e^{-2f_m} + N_{m-1} e^{-f_{m-1}}$$

$$N_{m-11} = n_{m-11} + \bar{n}_{m-11}$$

$$f_{m-11} = \frac{N_{m-11}}{T_{m-11}}$$

.....

In general,

$$\bar{n}_{m-k} = N_m \cdot e^{-(m-k+9)f_m} + N_{m-1} e^{-(m-k+8)f_{m-1}}$$

$$+ N_{m-k+9} e^{-(2)f_{m-k+9}} + N_{m-k+10} e^{-f_{m-k+10}}$$

$$= \sum_{j=o}^{k-10} N_{m-j} e^{-(m-k+9-j)f_{m-j}} \qquad \text{for } m > k \geq 10$$

$$N_{m-k} = n_{m-k} + \bar{n}_{m-k}$$

and $f_{m-k} = \frac{N_{m-k}}{T_{m-k}}$

5. References

(1) Methods of Estimating Basic Demographic Measures from Incomplete Data, United Nations. Manual IV, 1967.

(2) Bogue, D. J., Rapid Feedback of Family Planning Improvement, Manual 1-14, Community and Family Study Center, University of Chicago, 1970-75.

(3) Mohapatra, P. S., "The Effect of Age at Marriage and Birth Control Practices on Fertility Differentials in Taiwan," Ph.D. dissertation,

The University of Michigan, 1966.

(4) Srinivasan, K., "Open Interval as an Index of Fertility" Journal of Family Welfare (India) 13(2):40-44, 1966.

(5) Srinivasan, K., "A Probability Model Applicable to the Study of Inter-live Birth Intervals and Random Segments of the Same," Population Studies 23(1):63-70, 1967.

(6) Srinivasan, K., "A Set of Analytical Models for the Study of Open Intervals," Demography 5(1):34-44, 1968.

(7) Leridon, H., "Some Comments on Article by K. Srinivasan: A Probability Model Applicable to the Study of Inter-live Birth Intervals and Random Segments of the Same," Population Studies 23(1):101-104, 1969.

(8) Shep, M. C., J. K. Menken, J. C. Ridley, and J. W. Lingner, "Truncation Effect in Birth Interval Data," J.A.S.A. 65(330):678-693, 1970.

(9) Shep, M. C., and J. A. Menken, Mathematical Models of Contraception and Birth, The University of Chicago Press, Chicago and London, 1973.

(10) Pathak, K. B., "A Stochastic Model for the Study of Open Interval - A Cohort Approach," Sankhya (Series B) 33(3):305-314, 1971.

(11) Venkatachaya, K., "Some Problems in the Use of Open Intervals as Indicators of Fertility Change," Population Studies, 26(3):494-505, 1972.

(12) Hastings, D. W., and R. W. Robinson, "Open and Closed Birth Intervals for Once-Married Spouse-Present White Women," Demography 12(3):455-466, 1975.

(13) Potter, R. G., "Effect of Programme on Future Fertility and Birth Rates" (Unpublished Mimeographed), May 1968.

(14) Rindfuss, R. R., "Annual Fertility Rates from Census Data on Own Children: Comparisons with Vital Statistics Data for the United States" Demography 13(2), 1976.

A COMPOUND POISSON PROCESS FORMULATION OF THE PARTIY DISTRIBUTION

Lester R. Curtin and Chirayath M. Suchindran, University of North Carolina

1. Introduction

Various attempts have been made to describe the parity distribution as the realization of some type of Poisson process. Dandekar (1955) developes a modified Poisson distribution which is applied to data on the number of children born in a fixed time period. Brass (1958) and Singh (1968) assume that the number of live birth conceptions follow an underlying Poisson process with modifications for non-susceptible periods following a live birth. Further modifications for heterogeneity among women and for conceptions which end in fetal loss (pregnancy wastage) must also be considered. The probability distributions which result from these models are somewhat cumbersome and difficult to apply.

An assumption of an underlying non-homogeneous Poisson process leads to more theoretical models of the parity distribution such as those by Hoem (1969) and Nour (1972). This paper derives a model of the parity distribution which incorporates Nour's concept of conditional fecundability. The resulting model is a realization of a compound Poisson process and is a particular case of Hoem's model. Estimation of the model parameters from U.S. cohort fertility data will be briefly examined.

2. The H_1 and H_2 distributions

Suppose we observe a cohort of women of current age x. Assume that there has been no mortality, that each woman has been susceptible to the risk of a live birth conception for a fixed number n of time units, and that the probability of a live birth conception in a unit time is a constant p, $0 < p < 1$. Under these assumptions, the number of births to a woman aged x is a random variable having a Binomial distribution with parameters n and p.

Actually, the number of time units that a woman is susceptible to the risk of conception can be considered a random variable. That is, n will vary among women due to the influence of such variables as age at first marriage, non-susceptible periods following a live birth conception (the nine months of gestation plus a period of post-partum amenorrhea), and non-susceptible periods associated with pregnancy wastage. We consider two cases. For the first case, we assume that n is a random variable having a Poisson distribution with parameter λ. This gives the compound distribution for the number of births to a woman aged x as a Poisson distribution with mean λp For the second case, we assume that n has a Negative Binomial distribution with parameters K and p'. The resulting compound distribution is then a Negative Binomial distribution with parameters K and pp'.

Heterogeneity amoung women is introduced by considering the parameter p of the Binomial distribution as a random variable having a Beta distribution with parameters a and b. Specifically, let

$$f(p) = \frac{\Gamma(a)\Gamma(b-a)}{\Gamma(b)} p^{a-1}(1-p)^{b-a-1} \qquad (2.1)$$

with $b > a > 0$ and $0 > p > 1$.

This gives the parity distribution conditional on age as either Katti's (1968) H_1-distribution (for case 1) or the H_2-distribution (for case 2). Using Gurland's (1957) notation for compound distribution we have

$$H_1(\lambda,a,b) \sim \text{Bin}(n,p)^{\wedge}_{n}\ \text{Poisson}(\lambda)^{\wedge}_{p}\ \text{Beta}(a,b)$$

$$H_2(k,a,b,p') \sim \text{Bin}(n,p)^{\wedge}_{n}\ \text{Neg.Bin}(k,p')^{\wedge}_{p}\ \text{Beta}(a,b)\ .$$

The probability generating functions are given by

$$g_{H_1}(x) = {}_1F_1[a;b;\lambda(s-1)] \qquad (2.2)$$

and

$$g_{H_2}(s) = {}_2F_1[k,a;b;p'(s-1)] \qquad (2.3)$$

where ${}_1F_1[a;b;\lambda(s-1)]$ is the confluent hypergeometric function and ${}_2F_1[k,a;b;p'(s-1)]$ is the hypergeometric function. These are defined (Erdéyli, 1953) as follows.

$${}_1F_1[a;b;\lambda(s-1)] = \sum_{n=1}^{\infty} \frac{(a)n}{(b)n} \frac{\lambda^n(s-1)^n}{n!}$$

$${}_2F_1[k,a;b;p;(s-1)] = \sum_{n=0}^{\infty} \frac{(k)n(a)n}{(b)n} \frac{[p'(s-1)]^n}{n!}$$

and

$$(a)n = \frac{\Gamma(a+n)}{\Gamma(a)} = \begin{cases} 1 & \text{if } n = 0,-1,\ldots \\ \prod_{k=0}^{n-1}(a+k) & \text{if } n = 1,2,\ldots \end{cases}.$$

Differentiating the probabiltiy generating function gives the probability density functions for the H_1 and H_2 distributions as

$$p_{H_1}(x) = \frac{\lambda^x}{x!} \frac{(a)x}{(b)x} {}_1F_1[a+x;\ b+x;\ -\lambda] \qquad (2.4)$$

and

$$p_{H_2}(x) = \frac{(p')^x}{x!} \frac{(k)x(a)x}{(b)x} {}_2F_1[k+x;a+x;b+x;-p'] \qquad (2.5)$$

The factorial moments of these distributions are given by simple recurrence relations. For the H_1-distribution

$$\mu'_{(r+1)} = \frac{\lambda(a+r)}{(b+r)} \mu'_{(r)} \quad \text{for} \quad r = 0,1,2,\ldots \qquad (2.6)$$

For the H_2-distribution the relation is

$$\mu'_{(r+1)} = \frac{p'(k+r)(a+r)}{(b+r)}\mu'_{(r)} \quad \text{for} \quad r = 0,1,2,\ldots$$

The mean and variance of each distribution are easily desired to be:

$$\mu'_{H_1} = \frac{\lambda a}{b} \quad \text{and} \quad \sigma^2_{H_1} = \frac{\lambda a}{b}\left[1 + \frac{\lambda(b-a)}{b(b+1)}\right]$$

while

$$\mu'_{H_2} = \frac{kap'}{b} \quad \text{and} \quad \sigma^2_{H_2} = \frac{kap'}{b}\left[1 + \frac{kp'(b-a)}{b(b+1)} + \frac{p'(a+1)}{(b+1)}\right] .$$

Setting $\mu_{H_1} = \mu_{H_2}$ it is seen that $\sigma^2_{H_1} < \sigma^2_{H_2}$. Also, both distributions are over-dispersed in the sense that

$$\frac{\sigma^2_{H_1}}{\mu_{H_1}} > 1 \quad \text{and} \quad \frac{\sigma^2_{H_2}}{\mu_{H_2}} > 1 .$$

3. The Compound Poisson Process

A compound Poisson process is defined by Parzen (1962) in the following manner. Consider the stochastic process $\{x(t),\ t > 0\}$. Let

$$x(t) = \sum_{n=1}^{N(t)} Y_n \qquad (3.1)$$

such that $\{Y_n;\ n = 1,2,\ldots\}$ are independently identically distributed random variables and $\{N(t),\ t > 0\}$ is a Poisson process with intensity function $\nu(t)$. Then $x(t)$ is said to be a compound Poisson process. Also, we can define

$$m(t) = \int_0^t \nu(\tau)d\tau \qquad (3.2)$$

as the mean value function of the Poisson process $N(t)$.

We now define $x(t)$ to be the number of live births in the interval $(0,t)$ and $N(t)$ is the number of time units that a woman is susceptible to the risk of a live birth conception. As before we can define p as the probability of a live birth conception in a unit time given that the woman is susceptible to the risk of a live birth conception. This corresponds to Nour's definition of conditional fecundability (Nour, 1972). In the context of the compound Poisson process we now have

$$Y_n = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } (1-p) \end{cases} \qquad (3.3)$$

for $n = 1,2,\ldots$.

The unconditional fecundability (the probability of a live birth conception) can now be defined as $p\nu(\tau)\Delta\tau + o(\Delta\tau)$ where the probability of a woman being susceptible to the risk of conception in the interval $(\tau,\tau+\Delta\tau)$ is given by $\gamma(\tau)\Delta\tau + o(\Delta\tau)$. Thus the "force of fertility" is simply

$$\phi(t) = p\nu(t) . \qquad (3.4)$$

Given the above formulation, we can derive the distribution of $x(t)$ for an arbitrary, but fixed, value of t. Let $P_k(t|p)$ be the probability of k births $(k = 0,1,2,\ldots)$ in the interval $(0,t)$ given the value of p. We have

$$P_k(t|p) = \exp\left[-\int_0^t \phi(\tau)d\tau\right] \frac{\left[\int_0^t \phi(\tau)d\tau\right]^k}{k!}; \quad k = 0,1,2,\ldots$$

This can also be written as

$$P_k(t\ p) = \exp[-pm(t)] \frac{[pm(t)]^k}{k!}; \quad k = 0,1,2,\ldots$$

Assume that p has a denisty function $f(p)$, we have that $P_k(t)$, the unconditional probability of k births in $(0,t)$, is given by

$$P_k(t) = \int_0^1 P_k(t|p)f(p)dp \qquad (3.5)$$

since $0 < p < 1$. An obvious choice of $f(p)$ is the Beta distribution (equation 2.1). Substitution yields:

$$P_k(t) = \frac{[m(t)]^k}{k!}\frac{\Gamma(b-a)\Gamma(a)}{\Gamma(b)}\int_0^1 e^{-pm(t)}p^{k+a-1}(1-p)^{b-a-1}dp$$

or

$$P_k(t) = \frac{[m(t)]^k}{k!}\frac{(a)k}{(b)k}\ {}_1F_1[a+k;b+k;-m(t)] \qquad (3.6)$$

for $k = 0,1,2,\ldots$. This is simply the H_1-distribution.

If $N(t)$ is considered to be a homogeneous Poisson process, then $\nu(t) = \nu$ and $m(t) = \nu t$. This gives

$$P_k(t) = \frac{[\nu t]^k}{k!}\frac{(a)k}{(b)k}\ {}_1F_1[a+k;b+1;-\nu t] \qquad (3.7)$$

and the probability generating function for $X(t)$ is then $g(s) = {}_1F_1[a;b;\nu t]$.

Further heterogeneity amoung women can be introduced by assuming that ν is a random variable having a Gamma distribution with parameters k and β. That is,

$$f(\nu) = \frac{1}{\beta^k\Gamma(k)}\nu^{k-1}\exp[-\nu/\beta] . \qquad (3.8)$$

Letting $p' = 1/(B+1)$, the probability generating function for $X(t)$ now becomes

$$g(s) = {}_2F_1[k,a;b;p't] \qquad (3.9)$$

which is the H_2-distribtuion.

4. Estimation of Parameters

The maximum likelihood equations for the estimators of the parameters of the H_1 and H_2 distributions involve finite series and can not be solved explicitly. Iterative procedures, such as those outlined in Kaplan and Elston (1972) can be used to obtain the maximum likelihood estimates. However, for these distributions, the iterative

procedures either fail to converge or converge to an arbitrary upper or lower bound. This could be a result of a very flat likelihood surface. Table I shows how similar H_1-distributions can be obtained for quite different values of the parameters a and b.

Minimum chi-square estimates have the same problems as the maximum likelihood estimates. The minimization procedures used so far have been inadequate to give valid estimates. Moment type estimates are also inadequate.

In an attempt to get preliminary estimates, the function

$$\sum_i |O_i - E_i| \qquad (4.1)$$

was minimized. Here O_i refers to the observed frequency and E_i refers to the expected frequency for the desired distribution. These estimates can only be considered as very rough estimates and are used only as the initial step in the examination of the goodness of fit of the H_1 or H_2 distributions.

5. Modified Models

The data used is from Heuser (1976) and consists of the parity distribution by single years for the 1920 birth cohort (while women only). Initial estimates of the parameters of the H_1-distribution, conditional on age, indicate a possible lack of fit of the model. A modified H_1-distribution can be developed by adjusting the zero-parity class.

We let $(1-\alpha)$ be the proportion of the cohort at age x which can be considered as *never* having been susceptible to the risk of conception due to natural sterility or due to never being married. The modified H_1-distribution is then given by

$$P_0^* = (1-\alpha) + \alpha P_0$$

and (5.1)

$$P_i^* = \alpha P_i \ , \quad i = 1,2,\ldots$$

where the $P_0, P_1, P_2, \ldots,$ are the probability under the H_1-distribution.

The modified H_1-distribution does not provide an adequate fit to the data at the older ages. The major discrepancy lies in the class where parity equals two. The H_2-distribution fails to correct this problem. In an attempt to correct the problem with the second parity class, two mixtures were considered. Namely, a mixture of two H_1-distributions with different values of the parameter λ:

$$(1-\alpha)H_1(a,b,\lambda_1) + \alpha H_1(a,b,\lambda_2) \qquad (5.2)$$

and a mixture of the H_1 and H_2 distributions as

$$(1-\alpha)H_1(a,b,\lambda) + \alpha H_2(k,a,b,p') \qquad (5.3)$$

Examples of estimates of the parameters of the four distributions considered are presented for various ages. The distributions used are the modified H_1-distribution (Table II), the H_2-distribution (Table III), the mixture of two H_1-distributions (Table IV) and the mixture of an H_1 and an H_2-distribution (Table V).

6. Conclusions

Of the four distributions considered, the modified H_1-distribution appears to provide the best approximation to the observed parity distribution except at the oder ages. After age 40, the mixture of two H_1-distributions is a better approximation. This is indicated by Table VI. Again, the difference between the observed distribution and the fitted, or expected, distributions is most apparent at parity two. It seems that a certain proportion of the population terminate their reproduction after their second birth. Perhaps the model can be improved by treating the study population as a mixture of two populations with one group consisting of people who wish to terminate their fertility at two and the other group who does not terminate at two.

It is also obvious that the estimation procedures must be improved. It may yet be possible to obtain maximum likelihood estimates for the above distribution. These problems will be examined in subsequent reports.

Acknowledgments

This research was in part supported by the National Institute of Child Health and Human Development (Grant HD-00371). The authors would also like to thank Jackie O'Neal for her conscientious typing of this manuscript.

References

1. Brass, W. (1958). "The distribution of births in human populations," *Population Studies* 12, 51-72.

2. Dandekar, V.M. (1955). "Certain modified forms of Binomial and Poisson distributions," *Sankhyā A* 15, 237-250.

3. Erdélyi, A. (1953). *Higher Transcendental Functions, Vol. I.* McGraw-Hill Book Company, Inc., New York.

4. Gurland, J. (1957). "Some interrelations amoung compound and generalized distributions," *Biometrika* 44, 265-268.

5. Heuser, R.L. (1976). Fertility tables for birth cohorts, by color: United States 1917-1973. DHEW Pub. no. (HRA) 76-1152, National Center for Health Statistics, Rockville, MD

6. Hoem, J.M. (1969). "Fertility rates and reproduction rates in a probabilitic setting," *Biométric-Praximetre* 10, 38-66.

7. Kaplan, E. and Elston, R. (1972). "A subroutine package for maximum likelihood estimation (MAXLIK)," Institute of Statistics Mimeo Series No. 823, University of North Carolina.

8. Katti, S.K. (1966). "Interrelations amoung generalized distributions and their components," *Biometrics* 22, 44-52.

9. Nour, El-Sayed (1972). "A stochastic model for the study of human fertiltity," Institute of Statistics Mimeo Series No. 879, University of North Carolina.

10. Parzen, E. (1962). *Stochastic Processes.* Holden-Day, Inc., San Francisco.

11. Philipson, C. (1960). "The theory of confluent hypergeometric functions and its application to compound poisson processes," *Skandinarisk Aktuarietidskrift* 43, 136-162.

12. Singh, S.N. (1968). "Chance mechanisms of the variation in the number of births per couple," *Journal of the American Statistical Association* 63, 209-213.

TABLE I: The H_1-distribution with $\lambda = 2.5$

PARAMETER		PARITY							
A	B	0	1	2	3	4	5	6	≥7
4	5	.1489	.2674	.2582	.1743	.0913	.0392	.0143	.0064
8	10	.1423	.2670	.2641	.1768	.0906	.0378	.0133	.0081
24	30	.1376	.2706	.2684	.1790	.0903	.0367	.0125	.0049
40	50	.1367	.2706	.2693	.1796	.0902	.0365	.0123	.0048
56	70	.1363	.2706	.2697	.1798	.0902	.0363	.0122	.0049

TABLE II: Estimates of the Parameters of the Modified H_1-distribution

	PARAMETER			
Age	A	B	λ	α
20	1.708	3.660	0.943	0.526
25	2.385	3.256	1.553	0.839
30	4.011	4.643	2.058	0.964
35	4.531	5.312	2.695	0.971
40	5.456	6.283	2.934	0.968
45	4.953	5.803	2.935	0.978

TABLE III: Estimates of the Parameters of the H_2-distribution

	PARAMETER			
Age	k	A	B	p'
20	1.525	1.198	2.346	0.275
25	2.857	3.174	3.998	0.422
30	4.766	5.188	6.151	0.475
35	5.137	33.156	38.950	0.560
40	6.159	4.130	4.861	0.504
45	5.153	17.332	20.286	0.630

TABLE IV: Estimates of the Parameters of the Mixture $(1-\alpha)H_1(A,B,\lambda_1) + \alpha H_1(A,B,\lambda_2)$

	PARAMETERS				
Age	α	A	B	λ_1	λ_2
25	0.159	1,246	2.426	2.031	0.893
30	0.484	2.758	4.015	2.605	2.402
35	0.209	3.374	4.440	2.812	2.952
40	0.527	3.390	4.473	3.201	3.038
45	0.969	3.437	4.475	3.291	3.109

TABLE V: Estimates of the Parameters of the Mixture $(1-\alpha)H_1(a,b,\lambda) + \alpha H_2(k,a,b,p')$

	PARAMETER					
Age	α	λ	A	B	k	p'
25	0.478	2.360	2.367	4.700	2.031	0.389
30	0.166	5.343	4.612	6.427	2.311	0.491
35	0.090	2.158	4.713	6.061	2.809	0.323
40	0.088	5.831	4.165	4.924	2.688	0.557
45	0.004	5.476	3.480	4.623	3.201	0.601

TABLE VI: The Observed and Expected Parity Distributions

		PARITY								
Age	Parity Distribution	0	1	2	3	4	5	6	≥7	$\sum\lvert O_i-E_i\rvert$
35	Observed	.1351	.1929	.2950	.1908	.0962	.0443	.0220	.0238	—
	Modified H_1	.1352	.2243	.2493	.1912	.1128	.0542	.0220	.0110	.1170
	H_2	.1362	.2243	.2212	.1695	.1113	.0657	.0360	.0358	.1802
	Mixture H_1, H_1	.1350	.2466	.2512	.1828	.1046	.0495	.0200	.0103	.1352
	Mixture H_1, H_2	.1300	.2454	.2548	.1859	.1055	.0492	.0196	.0096	.1336
	Observed	.1161	.1685	.2722	.1965	.1148	.0599	.0328	.0452	—
	Modified H_1	.1101	.2015	.2432	.2023	.1293	.0673	.0296	.0167	.1220
	H_2	.1106	.1968	.2101	.1746	.1244	.0806	.0476	.0553	.1680
	Mixture H_1, H_1	.1100	.2197	.2449	.1951	.1223	.0635	.0281	.0164	.0987
	Mixture H_1, H_2	.1095	.2179	.2432	.1949	.1233	.0647	.0291	.0174	.0993

THE EFFECT OF RESTRICTED UNIVERSES ON COHORT BIRTH EXPECTATIONS

Maurice J. Moore and Martin O'Connell, U.S. Bureau of the Census

ABSTRACT

Past national surveys regarding birth expectations have usually been restricted to currently married women, a fact which has led demographers to question the usefulness of these data. Because the June 1976 Current Population Survey includes the expectations of all women in a cohort regardless of marital status, it provides the data needed to evaluate biases due to restricted survey universes. At older ages, where there are substantial differences in lifetime expectations between currently married and single women, there are relatively few single women; at younger ages, however, where the proportion of single women in a cohort is relatively large, the differences in expectations are small. This counterbalancing effect makes the lifetime birth expectations of currently married women a close approximation of all women in a cohort. The analysis also indicates that the observed intracohort declines in lifetime birth expectations since 1967 were due largely to the addition at subsequent survey dates of previously unmarried women; nevertheless, some "true" cohort declines also seem to have occurred.

INTRODUCTION

Since the 1955 Growth of American Families (GAF) study (Freedman et al. 1959) the hope has been that data on birth expectations could be used successfully to estimate completed marital fertility for cohorts of women still in their childbearing years. After the 1960 GAF study Whelpton, Campbell, and Patterson (1966) assessed the accuracy of birth expectations for the period 1955 to 1960. Births expected in the next 5 years by women surveyed in the 1955 GAF study were compared with those actually born in the previous 5 years to women surveyed in 1960. The result was that close agreement was found between <u>expected</u> and <u>actual</u> average numbers of births for the 5-year period. However, the fact that surveys regarding birth expectations have previously been limited to currently married or to ever-married women has led certain demographers, notably Ryder and Westoff (1967), to question the usefulness of these data for population projections or for intracohort fertility comparisons.

Siegel and Akers (1969) have summarized two principal drawbacks of expectations data due to limiting the sample to currently married women:

1. "Because women have most of their babies shortly after marriage and because the surveys covered married women only, in a very short time the majority of births occur to women who were not represented in the survey. (Our rough calculations show that in five years about 50 percent of the births, and in ten years nearly 80 percent of the births, will occur to women not now married.)
2. "The limitation of the surveys to married women means that the proportion of women covered by the surveys varies between ages in the same cohort and for the same age over time. Because of this, comparisons are difficult to interpret."

The June 1976 Current Population Survey was the first nationwide survey to ask number of children born to date and additional births expected for all women, regardless of marital status, in a broad age range of the childbearing years (Moore 1976). The age of the woman surveyed was 14 to 39 years for women currently married and living with their husbands and 18 to 34 years for all other woman. Thus, expectations of completed fertility are available for total cohorts of women in the age range 18 to 34 years.

The fact that all women in a cohort were surveyed allows us to examine differences by marital status. An additional survey question on date of first marriage made it possible to select out women who were already married at the time of previous surveys. This enables us to make some assessment of whether previously observed intracohort changes over time in the birth expectations of currently married women are "true" cohort changes or whether they are artifacts of adding to the survey universe women first marrying at later ages.

DIFFERENTIALS IN LIFETIME BIRTH EXPECTATIONS BY MARITAL STATUS

An indication of the magnitude and direction of the bias in cohort lifetime birth expectations resulting from the exclusion of single (never-married), widowed, divorced, and separated women is shown in table 1. The data show that, for individual age groups, the lifetime birth expectations of currently married women exceed the expectations of all women in the cohort by about one-tenth of a child per woman. For example, <u>all</u> women 22 to 24 years old in 1976 expected an average of 2,022 children per 1,000 women, whereas <u>currently</u> married women (excluding separated) of the same age expected an average of 2,145.

The second block of data in table 1 shows that the expectations of currently married women are from 4 to 7 percent higher than those of all

Table 1. Lifetime Births Expected per 1,000 Women by Marital Status: June 1976

(Data limited to women reporting on birth expectations.)

Age (1)	All women (2)	Ever married women (3)	Currently married (exc. separated) (4)	Widowed, divorced, and separated (5)	Single (6)
			Lifetime births expected per 1,000 women		
Total, 18-34	2,160	2,286	2,291	2,256	1,794
18-19	2,087	2,123	2,163	(B)	2,072
20-21	1,989	2,111	2,122	2,030	1,859
22-24	2,022	2,121	2,145	1,940	1,781
25-29	2,098	2,197	2,202	2,161	1,424
30-34	2,445	2,541	2,536	2,574	939
			Ratio to rates for all women		
Total, 18-34	1.000	1.058	1.061	1.044	0.831
18-19	1.000	1.017	1.036	(B)	0.993
20-21	1.000	1.061	1.067	1.021	0.935
22-24	1.000	1.049	1.061	0.959	0.881
25-29	1.000	1.047	1.050	1.030	0.679
30-34	1.000	1.039	1.037	1.053	0.384
			Number of women (thousands)		
Total, 18-34	23,125	17,174	14,880	2,296	5,952
18-19	2,768	733	670	63	2,036
20-21	2,847	1,463	1,296	167	1,384
22-24	4,350	3,084	2,714	371	1,266
25-29	7,153	6,246	5,394	853	907
30-34	6,007	5,648	4,806	842	359

B: Base less than 75,000.

women in the cohort. On the other hand, the expectations of single women fall short of those of all women in the cohort by a range of 1 percent for 18 and 19 year olds to 62 percent for 30 to 34 year olds.

Although the expectations of older single women differ markedly from the average for all women in the same cohort, nevertheless, the effect of single women's expectations on the average for all women depends not only on the difference of the level of expectations between single and ever-married women but also on the proportion of single women in each cohort. Among 18 and 19 year olds, where single women constitute about 74 percent of the cohort, the difference in expectations between single and ever-married women is less than one-tenth of a child per woman. Among 30 to 34 year olds, however, where the expectations of ever-married women exceed those of single women by an average of 1.6 children, single women constitute only 6 percent of the cohort. These two examples illustrate the counterbalancing effects on cohort averages of proportion single and of differences in levels of expectations by marital status. Thus, the ratios of lifetime births expected by ever-married women to those of all women in the cohort (table 1, column 3) show that the overall effect in 1976 of excluding single women from a cohort did not exceed 6 percent (for the 20 to 21 year olds) and was as low as 2 percent (for the 18 to 19 year olds).

The June 1976 CPS also included lifetime birth expectations for widowed, divorced, and separated women. Comparing lifetime expectations of currently married women with those of all ever-married women shows the effect of excluding women who are widowed, divorced, and separated. Rates shown for currently married women are only marginally higher, with the exception of women 30 to 34 years old, than rates for all ever-married women. Thus, the exclusion of widowed, divorced, and separated women from many previous surveys of birth expectations may very well be

Table 2. Lifetime Births Expected per 1,000 Wives Reporting on Birth Expectations in 1971 and 1976 by Number of Years Since First Marriage: June 1976

1971 Current Population Survey		1976 Current Population Survey			
				Wives first married:	
Age in 1971 (1)	All wives (2)	Age in 1976 (3)	All wives (4)	On or before June 1971 (5)	After June 1971 (6)
		Lifetime births expected per 1,000 wives			
14-17	2,497	19-22	2,159	2,166	2,158
18-19	2,256	23-24	2,128	2,193	2,094
20-21	2,375	25-26	2,111	2,224	1,944
22-24	2,404	27-29	2,258	2,317	1,952
25-29	2,620	30-34	2,536	2,571	1,865
30-34	2,991	35-39	2,994	3,017	(B)
		Percentage difference from 1971 rate for "All wives"			
14-17	(X)	19-22	-13.5	-13.3	-13.6
18-19	(X)	23-24	- 5.7	- 2.8	- 7.2
20-21	(X)	25-26	-11.1	- 6.4	-18.1
22-24	(X)	27-29	- 6.1	- 3.6	-18.8
25-29	(X)	30-34	- 3.2	- 1.9	-28.8
30-34	(X)	35-39	+ 0.1	+ 0.9	(B)
		Number of wives (thousands)			
14-17	165	19-22	2,491	199	2,292
18-19	687	23-24	1,932	669	1,263
20-21	1,342	25-26	2,054	1,230	824
22-24	2,957	27-29	3,339	2,792	547
25-29	4,514	30-34	4,806	4,561	245
30-34	3,982	35-39	4,206	4,145	61

B: Base less than 75,000. X: Not applicable.

discounted as having much of a biasing effect on expectations data.

Although the expectations data shown in table 1 indicate the likelihood of biases in previous surveys that have omitted unmarried women from their sample universe, nevertheless, lifetime birth expectations of currently married women closely approximate expectations of all women in a cohort. Thus, the omission of women not currently married in previous surveys does not seem to diminish significantly the value of these statistics in examining intercohort differences in birth expectations.

DIFFERENTIALS IN BIRTH EXPECTATIONS BY INTERVAL SINCE FIRST MARRIAGE

The June 1976 CPS also indicates the extent of biases in intracohort comparisons of the birth expectations of currently married women which are due to the increasing proportions of women first married as a cohort ages. When making intracohort comparisons of data from two different survey dates, we would ideally like to select at the later data only those women who were in the survey universe at the earlier date. Ryder and Westoff (1967) have suggested that, if the data are not analyzed in this manner, observed intracohort differences in expectations may be due to including women at later dates who were not currently married at a prior survey date.

The inclusion of more recently married women in a cohort tends to lower the birth expectations of all currently married women in a cohort because of the negative relationship between age at first marriage and birth expectations. This is shown in columns 5 and 6 of table 2 where women first married since June 1971 have lower expectations than the women first married on or before June 1971. However, where large differences exist in lifetime expectations between these two groups of women (for example, among women age 25 to 29 in 1971), the number of women married within the last five years is so small as to produce almost no difference between the rates expected by all wives (2,536 per 1,000 wives) and those married on or before June 1971

Table 3. Lifetime Births Expected per 1,000 Wives Reporting on Birth Expectations in 1967 and 1976 by Number of Years Since First Marriage: June 1976

1967 Survey of Economic Opportunity		1976 Current Population Survey			
				Wives first married:	
Age in 1967[a] (1)	All wives (2)	Age in 1976[b] (3)	All wives (4)	On or before March 1967 (5)	After March 1967 (6)
		Lifetime births expected per 1,000 wives			
18-19	2,720	27-28	2,232	2,593	2,098
20-21	2,916	29-30	2,367	2,569	2,123
22-24	2,856	31-33	2,541	2,669	2,053
25-29	3,037	34-38	2,940	3,014	1,885
		Percentage difference from 1967 rate			
18-19	(X)	27-28	-17.9	- 4.7	-22.9
20-21	(X)	29-30	-18.8	-11.9	-27.2
22-24	(X)	31-33	-11.0	- 6.5	-28.1
25-29	(X)	34-38	- 3.2	-0.8	-37.9
		Number of wives (thousands)			
18-19	588	27-28	2,107	570	1,537
20-21	1,087	29-30	2,084	1,139	945
22-24	2,486	31-33	2,979	2,361	619
25-29	3,773	34-38	4,328	4,043	285

a - Age in February-March 1967 SEO study

b - Age in 1976 is 9 years and 3 months older than 1967 ages since midpoint of 1967 survey is taken as March 1.

X - Not applicable.

(2,571 per 1,000 wives). In fact, none of the differences shown in the 1976 birth expectations of all wives and those of wives first married 5 or more years ago exceed one-tenth of a child for any age group in table 2.

Table 2 shows two types of intracohort changes between June 1971 and June 1976, a "gross" change where there is no control for interval since first marriage (column 4) and a "net" change where the interval since first marriage is used as a controlling variable (column 5). The "gross" declines over the 5-year period appear to be greater than the "net" declines. However, no statistically significant ($p < .05$) declines in lifetime expectations occurred between 1971 and 1976 for wives who were first married by June 1971. In fact, with the exception of wives 20 to 21 years old in 1971, no differences were statistically significant between the 1971 rates in column 2 and the rates for all wives in column 4.

Differences over the longer time period of 1967 to 1976 are shown in table 3. Lifetime birth expectations for all wives in 1976 (column 4) show a large gross decline for women who were 18 to 24 years old in 1967. However, the net cohort rates for 1976 (column 5), based on women who had been married in 1967, are much closer to the 1967 rates, although these too show evidence of a decline over the 9-year period. The magnitude of the differences of the 1976 rates relative to the 1967 rates is especially striking for the youngest age group. For all wives who were 18 and 19 years old in 1967, a gross decline of 17.9 percent in lifetime expectations is recorded over the 9-year period, whereas a net cohort decline of only 4.7 percent occurs for those women who had married by the time of the 1967 survey. The difference between gross and net changes for women 25 to 29 years old in 1967 is relatively small, since this group had already completed the majority of its lifetime fertility by 1967.

In addition to showing differences in gross and net cohort changes, tables 2 and 3 also indicate the level of consistency in the expectations of comparable samples of women at two survey dates. Over the 5-year period expectations appear to have changed by no more than 6 or 7 percent for women 18 to 34 years old in 1971, with the average being around 2 to 3 percent (table 2). The level of change over the 9-year period is slightly greater with an average change of about 4 percent for women 18 to 29 years old in 1967 (table 3).

The decomposition of birth expectations by interval since first marriage suggests that, for the 5-year period of time examined in this paper, lifetime expectations for all currently married women in a cohort can reasonably be utilized to

measure short term intracohort changes in lifetime birth expectations. The longer 9-year period, however, substantiates the Ryder-Westoff statement that the confounding effect of the addition of recently married women to a cohort as it ages produces gross changes in cohort expectations that significantly overestimate the true cohort change. Unfortunately, the younger age groups, which will contribute substantially more future births than older age groups, are more subject to such a bias since they incur the greatest number of future additions from subsequent first marriages.

SUMMARY

Data from the June 1976 Current Population Survey indicate some substantial differences in lifetime birth expectations between currently married and single women. Where the differences in the expectations are the largest at the older ages, there are relatively few single women; at the younger ages, however, where the proportion of single women in a cohort is relatively large, the differences in expectations between single and currently married women are small. Thus, the lifetime birth expectations of either currently married or ever-married women, to which previous surveys have been limited, may reasonably serve as a proxy for the expectations of all women in a cohort, regardless of marital status.

The data were also examined to ascertain whether recently observed intracohort declines in lifetime births expected by currently married women were true declines or whether they were artifacts of the changing composition of the cohorts due to the subsequent addition of women first marrying at later ages. Although the observed intracohort declines in birth expectations were shown to be due, in a large part, to the subsequent addition to the sample universe of previously unmarried women, nonetheless some "true" cohort declines seem to have occurred since 1967.

This paper is substantially the same as our paper which appeared in Demography, August 1977, Vol. 14, No. 3.

REFERENCES

Freedman, Ronald P., Pascal K. Whelpton, and Arthur A. Campbell. 1959. Family Planning, Sterility, and Population Growth. New York: McGraw-Hill.

Moore, Maurice J. 1976. Asking Single Women About Their Children: The Census Bureau's Experience. Proceedings of the Social Statistics Section, American Statistical Association annual meeting, Boston, pp. 618-623.

Ryder, Norman B., and Charles F. Westoff. 1967. The Trend of Expected Parity in the United States: 1955, 1960, 1965. Population Index 33: 153-168

Siegel, Jacob S., and Donald S. Akers. 1969. Some Aspects of the Use of Birth Expectations Data from Sample Surveys for Population Projections. Demography 6: 101-115.

Whelpton, Pascal K., Arthur A. Campbell, and John E. Patterson. 1966. Fertility and Family Planning in the United States. Princeton: Princeton University Press.

AN EXAMINATION OF THE INTERRELATIONSHIPS BETWEEN NOMINAL AND DEMOGRAPHIC DIMENSIONS WITHIN THE AMERICAN PROFESSORIATE: A CASE STUDY IN MANOVA

Jerrold P. Katz, Simmons College
Andrei S. Markovits, Wesleyan University

Most modern societies embody important structures of stratification which affect the existence of their citizens, both publicly and privately. The major agents of stratification differ geographically as well as longitudinally with, however, certain indicators prevailing over time. Hence, class, ethnicity (race), religion, age and sex have divided all societies, providing unearned advantages to a select few while causing undue hardships to a great number.

The very essence of the existence of the United States is the result of stratification and its inequities in other parts of the world. Unfortunately, however, the new country perpetuated its own structure of stratification and hiearchical differentiation which, albeit different--thus, for example, most political sociologists agree that class is a weaker discriminating factor in the United States than in Europe--is no less an empirical reality and a moral bane. Still, a sincere meritocratic ethic, an ideological characteristic of a burgeoning capitalist development, has pervaded various social structures, notably the realm of knowledge. Science ostensibly rewarded only meritocratically attained achievements which followed strict requirements of intellectual rigor and objectively defined criteria. The American belief in the positive values of meritocracy became institutionally epitomized in the structure of academia.

Universities and their constituents have replicated--and initiated--many changes of contemporary American life. Most importantly, they have stood in the forefront of the battle against discrimination thereby representing a major protagonist for equality and justice. Without denigrating the sincerity of this noble effort in the least we would like to shed some light on certain structural inequities which, despite a meritocratic ideology, have remained inherent to American higher education with all its organizational manifestations.

Using the data from the extensive Carnegie Faculty Survey of 1969 furnished to us by the courtesy of Professors E.C. Ladd and S.M. Lipset we have attempted an in-depth analysis of the extent to which meritocratic criteria determine the institutional existence and rewards of the American professoriate. The present paper embodies a preliminary report of a partial segment of this larger project.

We have selected a number of measures related to personal background such as sex, race, religion and parental education on the one hand, and variables measuring professional achievement such as salary, quality of institution and research funding on the other. Although we have been working extensively with numerous other variables such as for example "parental occupation" and "regional origins" on the independent dimension and "number of publications" in the dependent cluster, our principal concern in this particular endeavor is to explore some suitable analytic procedures to be implemented in our larger project.

We proceeded to use "analysis of variance" to determine whether the background variables affected the achievement variables. We considered running an ANOVA by using each dependent variable and all six of its independent counterparts. This would have entailed six different analyses each being a six-way ANOVA. A major constraint presented itself in the fact that each analysis would have required the excess of 20 million bytes of core storage. Thus we subdivided the analysis into several one, two, and three-way ANOVAs thereby exploring the effects of different combinations of the independent variables on each of the dependent ones. For example, one analysis encompassed "quality of institution" by "religion", "race" and "father's education". The main effects were significant at a level of less than .001. The overall two-way interaction was significant at .01, while the individual two-way interactions were not all significant. The entire set of results will be found in Table 2.

Similarly we performed another three-way ANOVA using the same dependent variable with the independent variables being "religion", "race" and "sex". The results were again quite similar. Details can be found in Table 3.

Numerous other ANOVAs were performed in the same manner employing each of the dependent achievement variables and different combinations of the independent background variables. In every case, each of the main effects showed an F-statistic which was significant at a level of less than .001 thus indicating the strong effect of the independent background variables on the dependent achievement variables.

The main problem with the above is that we used several different measures of achievement. Ideally, we would have liked to determine whether overall achievement is affected by the set of independent background variables. A conventional procedure would have been to construct a single composite index of achievement derived from its individual components. Regardless of the construction procedure, one faces conceptual problems such as those of proper weighting and scaling in addition to incurring an inevitable loss of information. In order to overcome this obstacle, we decided to try the technique of multivariate analysis of variance. MANOVA permits the use of a set (more than one) of interval variables as the dependent variable in addition to treating independent variables in a manner analogous to ANOVA. The benefit of this process lies in the fact that there is no necessity of building a composite index; rather, the entire set of achievement variables can be incorporated. Subsequently, we can observe whether a difference exists as to the overall achievement determined by the independent

background variables. Furthermore, MANOVA is also suitable for handling several nominal dependent dariables. Thus, for example, in a subsequent stage of our research, we would like to use "teaching discipline" as a dependent variable. (This variable has four categories.) This step would require the creation of dummy variables for each category. Under MANOVA we will treat the entire set as a dependent variable, thus avoiding a number of separate and tedious procedures.

We performed several MANOVAs. Each MANOVA consisted of the entire set of achievement variables by three of the background variables. The output was much more complex than that of ANOVA. Summaries of the rather interesting results will be found in Table 4-space does not permit the inclusion of all the tables.

As can be gathered from the table, the independent background variables have a significant effect on the achievement variables. Although more complex than other analytic procedures, the results of this analysis were very informative conceptually and methodologically. While this brief paper has tried to show that achievement is affected decisively by the selected set of background variables, it would be an interesting addition to measure the degree to which <u>each</u> background variable affects the level of achievement. Thus, for instance, it would be important to investigate whether religion has a more substantial effect on academic achievement towards elucidating this problem. However, most of the background variables are nominal in nature. Hence, the use of regression analysis would necessitate the creation of a vast array of dummy variables.

Unquestionable, there is more work to be done in this direction. As pointed out earlier, the present preliminary report represents but a fraction of our eventual analyses concerning this rich data. One can nevertheless state, even at this juncture of our research, that the evidence is overwhelming as to the fact that ascriptive variables such as sex, religion and race play a crucial role in determining an individual's success or failure in a nominally meritocratic environment.

Table 1

VARIABLE SET

Background or Ascriptive Variables--Independent	Achievement Variables--Dependent
Religion in which person was raised	Total annual salary
Sex	Quality of institution
Race	S.A.T. scores of students at institution
Father's political beliefs	Research funded by outside sources
Father's education	Research dollars per student
Mother's education	Revenue per student

Table 2

ANALYSIS OF VARIANCE

QUALITY OF INSTITUTION BY

RELIGION, RACE AND FATHER'S EDUCATION

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	7286.488	13	560.499	133.072	0.000
religion	3518.414	4	879.604	216.680	0.000
race	498.894	3	166.298	40.965	0.000
father's education	2728.738	6	454.790	112.032	0.000
2-WAY INTERACTIONS	320.316	52	6.160	1.517	0.010
religion race	59.271	10	5.927	1.460	0.148
religion father's education	194.507	24	8.104	1.996	0.003
race father's education	71.655	18	3.981	0.981	0.479
3-WAY INTERACTIONS	172.766	49	3.526	0.869	0.730
religion race father's education	172.768	49	3.526	0.869	0.730
EXPLAINED	7779.625	114	68.242	16.811	0.000
RESIDUAL	231994.500	57149	4.059		
TOTAL	239774.125	57263	4.187		

Table 3

ANALYSIS OF VARIANCE

QUALITY OF INSTITUTION BY

RELIGION, RACE, AND SEX

* *

SOURCE OF VARIATION	SUM OF SQUARES	DF	MEAN SQUARE	F	SIGNIF OF F
MAIN EFFECTS	6488.738	8	811.092	199.153	0.000
Religion	3567.358	4	891.839	218.979	0.000
Race	504.696	3	168.232	41.307	0.000
Sex	1799.183	1	1799.183	441.765	0.000
2-WAY INTERACTIONS	198.324	17	11.666	2.864	0.000
religion race	71.489	10	7.149	1.755	0.063
religion sex	110.159	4	27.540	6.762	0.000
race sex	13.992	3	4.664	1.145	0.329
3-WAY INTERACTIONS	39.660	10	3.966	0.974	0.464
religion race sex	39.661	10	3.966	0.974	0.464
EXPLAINED	6726.750	35	192.193	47.190	0.000
RESIDUAL	236111.563	57974	4.073		
TOTAL	242838.313	58009	4.186		

Table 4

Multivariate Analysis of Variance Salary,
Quality of Institution, Sponsored Research;
Research Per Student and Revenue Per Student
by Father's Education and Mother's Education

OVERALL SALARY	QUALITY OF INSTITUTION	MEAN SAT SCORE FOR STUDENTS AT INSTITUTION	AMOUNT OF SPONSORED RESEARCH	RESEARCH DOLLARS PER STUDENT	REVENUE PER STUDENT
Overall Salary	0.636E 03				
Quality of Institution	0.133E 03	0.179E 03			
Mean SAT score for students at Institution	0.806E 02	0.918E 02	0.135E 03		
Amount of sponsored research	0.163E 03	0.889E 02	0.544E 02	0.234E 03	
Research dollars per student	0.210E 03	0.152E 03	0.112E 03	0.309E 03	0.420E 03
Revenue per student	0.176E 02	0.861E 02	0.594E 02	0.138E 03	0.181E 03	0.116E 03

TEST OF FM

TESTS OF SIGNIFICANCE USING WILKS LAMBDA CRITERION
AND CANONICAL CORRELATIONS.

TEST OF ROOTS	F	DFHYP	DFERR	PROB.	R
1 THROUGH 6	1.442	216.000	311418.688	0.001	0.041
2 THROUGH 6	1.269	175.000	311224.688	0.013	0.036
3 THROUGH 6	1.130	136.000	311011.375	0.073	0.032
4 THROUGH 6	0.997	99.000	310776.375	0.001	0.030
5 THROUGH 6	0.801	64.000	310517.250	0.895	0.026
6 THROUGH 6	0.529	31.000	310230.938	0.990	0.018

UNIVARIATE F TESTS

VARIABLE	F(36,52526)	MEAN SQ	PROB.
Overall Salary	1.942	17.653	0.001
Quality of Institution	1.210	4.973	0.172
Mean SAT score for students	2.036	3.757	0.001
at Institution	1.117	7.893	0.261
Amount of sponsored research	1.556	11.675	0.016
Research dollars per student	1.375	3.224	0.062
Revenue per student			

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

VARIABLE	1	2	3	4
Overall Salary	0.456	-0.695	-0.451	0.364
Quality of Institution	-0.253	-0.010	0.712	0.883
Mean SAT score for students	-0.695	-0.748	-0.246	-0.593
at Institution	0.928	-0.031	-0.041	-1.007
Amount of sponsored research	-0.731	0.549	-1.007	0.511
Research dollars per student	0.214	0.130	-0.092	0.636
Revenue per student				

TEST OF F

SUMS OF PRODUCTS FOR HYPOTHESIS ADJUSTED FOR 0 COVARIATES

OVERALL SALARY	QUALITY OF INSTITUTION	MEAN SAT SCORE FOR STUDENT AT INSTITUTION	AMOUNT OF SPONSORED RESEARCH	RESEARCH DOLLAR PER STUDENT	REVENUE PER STUDENT
Overall Salary	0.156E 04				
Quality of Institution	0.560E 03	0.246E 04			
Mean SAT score for students	0.349E 03	0.197E 04	0.159E 04		
at Institution	0.374E 03	0.204E 04	0.161E 04	0.173E 04	
Amount of sponsored research	0.406E 03	0.269E 04	0.216E 04	0.225E 04	0.298E 04	
Research dollars per student	0.332E 03	0.169E 04	0.136E 04	0.139E 04	0.185E 04	0.117E 04
Revenue per student						

TESTS OF SIGNIFICANCE USING WILKS LAMBDA CRITERION
AND CANONICAL CORRELATIONS.

TEST OF ROOTS	F	DFHYP	DFERR	PROB.	R
1 THROUGH 6	37.564	36.000	230638.500	0.001	0.148
2 THROUGH 6	7.667	25.000	210088.500	0.001	0.056
3 THROUGH 6	1.550	16.000	183229.063	0.070	0.018
4 THROUGH 6	0.801	9.000	148556.813	0.607	0.010
5 THROUGH 6	0.614	4.000	105046.000	0.651	0.007
6 THROUGH 6	0.0	1.000	52523.500	1.000	0.000

UNIVARIATE F TESTS

VARIABLE	F(6,52526)	MEAN SQ	PROB.
Overall salary			
Quality of Institution	28.672	260.638	0.001
Mean SAT score for students	99.809	410.151	0.001
At Institution	143.909	265.512	0.001
Amount of sponsored research	40.733	287.861	0.001
Research dollars per student	66.113	496.096	0.001
Revenue per student	82.848	194.227	0.001

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

VARIABLE	1	2	3
	0.359	0.950	-0.089
Overall Salary	0.386	0.106	0.224
Quality of Institution	0.605	-0.236	-0.353
Mean SAT score for students	0.228	0.534	1.437
At Institution	0.167	-0.452	-0.482
Amount of sponsored Research	0.260	0.018	-0.324
Research dollars per student			
Revenue per student			

	OVERALL SALARY	QUALITY OF INSTITUTION	MEAN SAT SCORE FOR STUDENTS AT INSTITUTION	AMOUNT OF SPONSORED RESEARCH	RESEARCH DOLLAR PER STUDENT	REVENUE PER STUDENT
Overall Salary	0.287E 04					
Quality of Institution	0.173E 03	0.989E 02				
Mean SAT score for students	0.916E 02	0.877E 02	0.983E 02			
at Institution	0.122E 03	0.754E 02	0.685E 02	0.105E 03		
Amount of sponsored research	0.396E 02	0.832E 02	0.919E 02	0.107E 03	0.134E 03	
Research dollars per student	0.954E 02	0.544E 02	0.567E 02	0.527E 02	0.710E 02	0.426E 02
Revenue per student						

TESTS OF SIGNIFICANCE USING WILKS LAMBDA CRITERION
AND CANONICAL CORRELATIONS.

TEST OF ROOTS	F	DFHYP	DFERR	PROB.	R
1 THROUGH 6	11.939	36.000	230638.500	0.001	0.082
2 THROUGH 6	2.938	25.000	210088.500	0.001	0.031
3 THROUGH 6	1.481	16.000	183229.063	0.098	0.017
4 THROUGH 6	0.924	9.000	148556.813	0.473	0.012
5 THROUGH 6	0.150	4.000	105046.000	0.962	0.004
6 THROUGH 6	0.0	1.000	52523.500	1.000	0.001

UNIVARIATE F TESTS

VARIABLE	F(6,52526)	MEAN SQ	PROB.
Overall Salary	2.641	478.523	0.001
Quality of Institution	4.012	16.489	0.001
Mean SAT score for student	3.879	16.382	0.001
at Institution	2.487	17.579	0.018
Amount of sponsored research	2.986	22.405	0.006
Research dollars per student	3.031	7.107	0.005
Revenue per student			

STANDARDIZED DISCRIMINANT FUNCTION COEFFICIENTS

VARIABLE	1	2	3
Overall Salary	1.024	0.122	-0.007
Quality of Institution	0.174	-0.089	0.701
Mean SAT score for students	0.141	-0.943	-0.030
at Institution	0.020	0.057	1.458
Amount of sponsored research	0.016	-0.030	-1.764
Research dollars per student	0.108	0.023	-0.187
Revenue per student			

BIBLIOGRAPHY

Buhler Roland and Shirrell Buhler, P-Stat-A Computing System for File Manipulation and Statistical Analysis, Princeton, N.J.: Princeton University Computing Center, 1976

Cooley and Lohnes, Introduction to Statistical Procedures, Wiley, NY: 1968

Morrison, Donald F., Multivariate Statistical Methods, New York, New York: McGraw-Hill Book Company, 1967

Nice, Norman, H., C. Hadlan Hull, Jean G. Jenkins, Karin Steinbrenner, Dale H. Bert, Statistical Package for tr Social Sciences;

Press, S. James, Applied Multivariate Analysis, New York, New York: Holt, Rinehart and Winston, Inc., 1972

Tatsuoka, Maurice, M., Multivariate Analysis, New York, New York: John Wiley & Sons, Inc., 1971

Winer, B. J. Statistical Principles in Social Design, New York, New York: McGraw-Hill Book Company, 1971

The Women and Mathematics Program: A Preliminary Statistical Evaluation

L. Denby
S. J. Devlin
Bell Laboratories
Murray Hill, New Jersey 07974

E. L. Poiani
Saint Peter's College
Jersey City, New Jersey 07306

This paper describes preliminary analyses of a pilot study designed to evaluate the effectiveness of the Women and Mathematics program. WAM, an acronym for Women and Mathematics, is a secondary school lectureship program sponsored by the Mathematical Association of America under a grant from IBM. Since math is a "critical filter" to many careers, the purpose of WAM is to interest high school 10th graders, women in particular, in studying more math by providing role models, and to acquaint counselors and teachers with career opportunities open to students with good math backgrounds (MAA, 1976 and Ernest, 1976).

The evaluation that will be described is a first attempt at determining if WAM lectures had any short-term effect on attitudes toward math and sex roles in math-related fields.

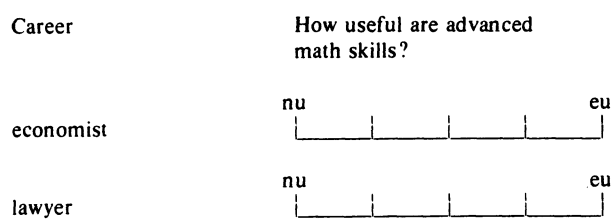
The study was conducted in two New Jersey cities, chosen not only because of their different demographic characteristics but also because of our own connections within the school systems. The first city is a large urban center with 260,000 people according to the 1970 census and has five public high schools. The other, a suburban city, has 24,000 people with only one public high school. The urban city has a median income of $9,000; the suburban city's median income is $18,000. The median education level of an adult in the urban city is 10th grade — less than a high school degree. In the suburban city the median is "some college."

The 10th graders (males and females) in each public high school were divided into two groups. One group heard a WAM lecture; the other did not. This division had to be done within the confines of the school schedule, so as to cause minimal disruption to the existing classes. An attempt was made to get the same range of math ability and background in both groups.

A questionnaire was given to all 10th graders about two weeks after the experimental group heard the lecture. The student questionnaire contained three sections:

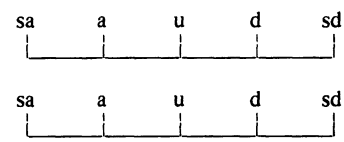
(1) Demographics: This included what courses each student had taken, the student's grade point average, parents' occupations and schooling, and the student's future plans.

(2) Career awareness questions: The students evaluated the usefulness of math to eight careers. The following depicts a part of the questionnaire exemplifying these questions:



where nu = not useful
eu = extremely useful

(3) 24 attitude statements: The student indicated his/her agreement-disagreement to statements concerning confidence towards math, usefulness of math, teachers'-parents' influence, and perception of math as a male domain. These statements were extracted mainly from a study by Fennema and Sherman (1976). The following are two examples:



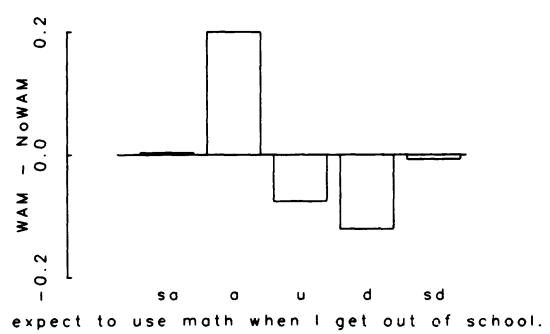
where sa = strongly agree
a = agree
u = uncertain
d = disagree
sd = strongly disagree

Also, a representative of the school was asked to fill out a form about the make-up and characteristics of the school population, the history of enrollment by sex in each of the upper level math courses for the current and previous three years, and the standardized testing that is given to the students and the availability of these scores for our use.

The analyses that are reported here are preliminary because only two of the tested schools have been analyzed and because of the pilot nature of the study. Results are from two of the schools — one from each city. The sample sizes were 236 in the urban school and 339 in the suburban school. The urban school is larger but absenteeism was a problem there.

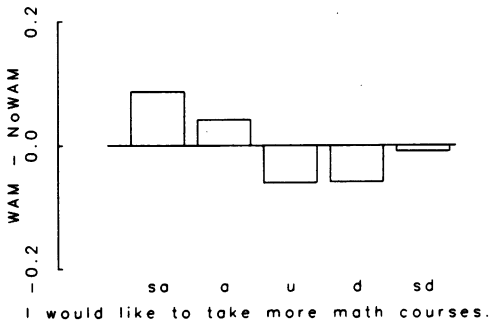
The first step of the analysis involved looking at each of the attitude and career awareness responses individually at each school to see if there was any effect from hearing a WAM lecture. For each statement the χ^2 statistic to test for independence in two-way tables was calculated to check if there was a statistically significant difference between WAM and no-WAM responses. In this case we have a 2×5 table: WAM or no-WAM vs. five possible responses to each statement or question.

Figure 1
Urban School



For example, consider the statement: "I expect to use math when I get out of school." For the urban school there is a significant difference between WAM and no-WAM response at the 2% level. Figure 1 gives a way of looking at the responses to see how the WAM answers differ from the no-WAM answers. The figure depicts a bar graph of the difference between the proportion of WAM and the proportion of no-WAM students giving a certain response. Each bar represents one of the five possible responses — sa, a, u, d, sd. For example, seven-tenths of the WAM group responded "agree" to the statement. Only one-half of the no-WAM group responded "agree." The difference between the two proportions is .2 as seen in Figure 1. This shift towards agreement and away from disagreement (the negative bars at "u" and "d") suggests that the WAM talks may be adjusting student attitude in the urban school towards usefulness of math. Interpretation of this effect requires the raw bar charts of the no-WAM responses, which measures the attitude of the students before a WAM visit, in conjunction with Figure 1.

Figure 2
Urban School



For the statement, "I would like to take more math courses," the χ^2 test says that the two groups at the urban school are not significantly different. However, Figure 2 shows a shift which is systematic — proportionately more WAM respondents agree or strongly agree with this statement. Unfortunately the χ^2 test ignores the order of the categories which is important here; hence these difference bar graphs are necessary. After inspecting these χ^2 results, it is evident that in future analyses a test which looks for systematic trends would be more appropriate.

Figure 3
Suburban School

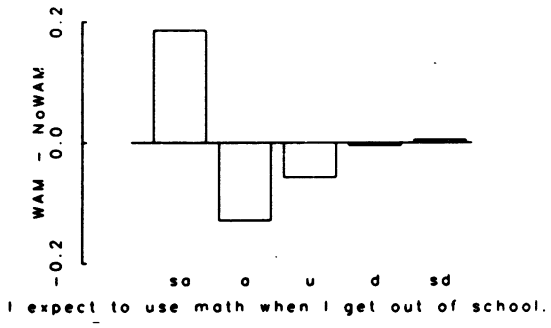
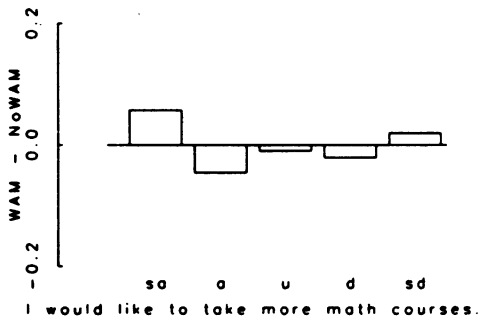


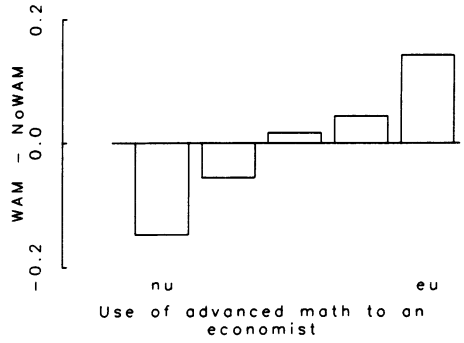
Figure 4
Suburban School



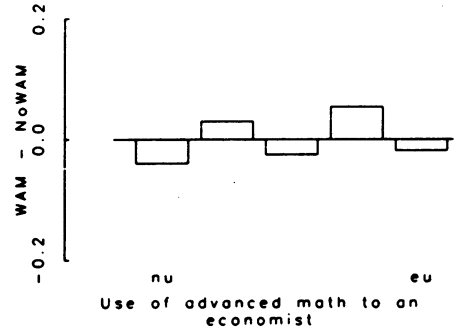
Inspecting the same questions for the suburban, more affluent school district, we find a significant WAM effect at the 1% level for the first statement (Figure 3). Here the shift is more from "uncertain" and "agree" for the no-WAM group to strongly agree for WAM. Looking at the second statement (Figure 4), not only is there no significant difference but also there is no distinct pattern seen. It could be that no WAM effect is evident because most of the suburban students already had intended to take more math. However, this was not the case. 54% of the no-WAM group responded uncertain to strongly disagree.

Summarizing the results of all 24 attitude statements, a WAM effect was found for about one-half of the statements in the urban city school. This was not true in the suburban school where little WAM effect was seen.

Figure 5
Urban School



Suburban School



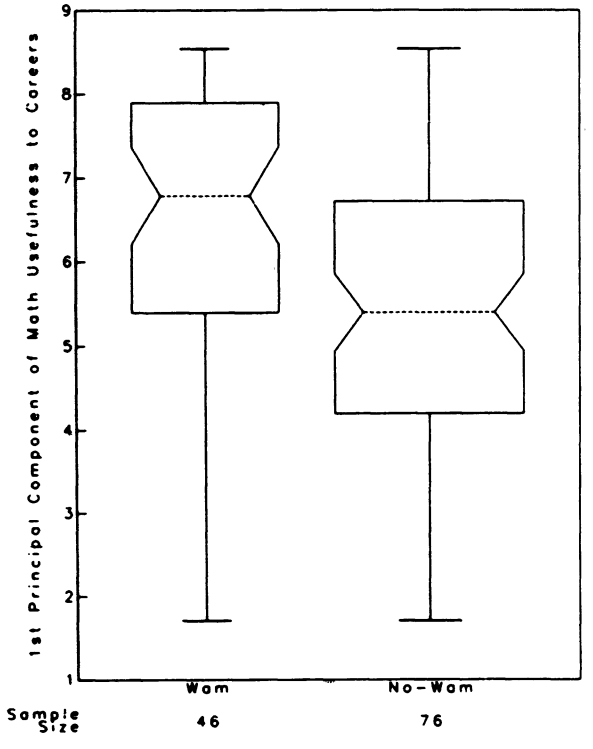
The second section of the questionnaire contained a list of eight careers. The students were asked to evaluate the usefulness of math to each career. Figure 5 exemplifies the responses at both schools regarding usefulness of math to an economist — the scale now ranges from not useful to extremely useful. In both schools the χ^2 test shows a significant WAM effect at an

11% level. After eliminating the nonrespondents, the suburban school had a much larger sample answering this statement; hence, smaller differences were judged significant. However, inspection of the difference bar graph shows that the suburban school's large χ^2 value is due to randomly ordered differences and thus is not impressive for our purposes. However, the top difference bar graph (the urban school) does show a distinctive pattern — a larger proportion of those hearing the WAM lecture thought that math was useful to an economist. For the urban school, patterns somewhat similar to this one were seen for all eight careers with varying levels of significance. In the suburban school only two careers showed a pattern in favor of a positive WAM effect.

We have just described some of the univariate methods used to inspect this data. It is very voluminous to summarize results in this fashion and these methods do not take account of interresponse correlations. Perhaps it might be better to summarize each student's impression of the importance of mathematics to various careers as a linear combination of their responses to the eight careers. The advantage of such a measure is that it is more continuous and thus more types of analyses are applicable. Also, it focuses more generally on career awareness than on a specific career.

To detect a WAM effect the choice of linear combination should be that which best discriminates between the WAM and no-WAM mean vectors — the discriminant axis. However, in doing some general exploratory analysis of the urban school data, the direction accounting for the greatest variability of all eight career responses — the 1st principal axis — was used to inspect for WAM and the many other possible differences (e.g., sex, attitude towards math, parents' occupations, etc.) simultaneously. It proved to be of interest from the WAM and no-WAM viewpoint. The data for students in one particular school who answered all eight questions are projected onto this axis. Next, the students were grouped first by WAM and no-WAM, and then boxplots were used to compare how the distribution of the measures varied between WAM and no-WAM.

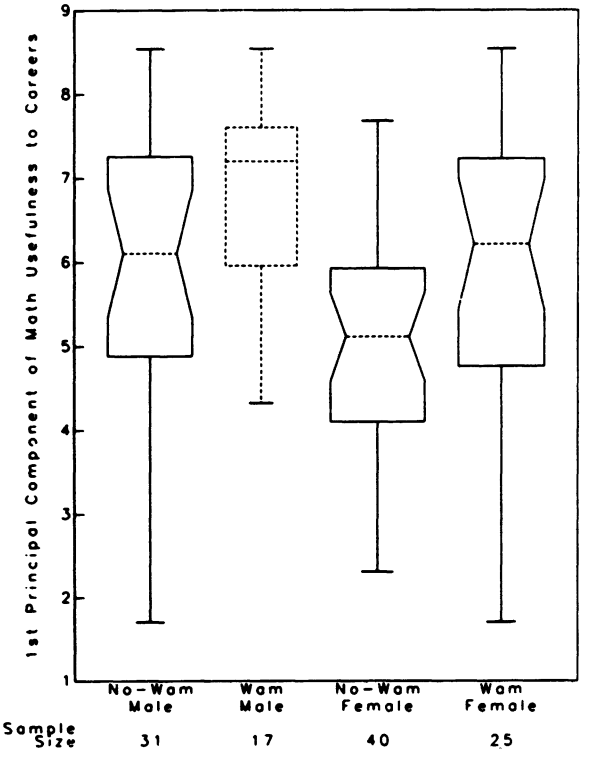
Figure 6
Urban School



A boxplot (McGill et al., 1977) is a summary of the distribution of a sample displayed graphically as seen in Figure 6. The quartiles define the upper and lower edges of the box. The lines out of the box extend to the upper and lower extremes of the data set. The dashed line through the middle of the box indicates the median. The width of the box is proportional to the square root of the sample size of the data set. When comparing two or more independent data sets, upper and lower notches are added to each box which are defined so that if the notches of two boxes overlap, the medians are insignificantly different at approximately the 5% level.

Figure 6 depicts the boxplots for the urban students' usefulness responses to the eight careers projected onto the first principal axis. Note that the WAM median is above the no-WAM median and the notches do not overlap. This suggests that the two groups are different. Since the direction is data determined from the combined groups, the notches may be too small for a test of significance at the 5% level. The coefficients of the eigenvector which defines the linear combination are all positive supporting that the WAM group generally perceives math as more useful.

Figure 7
Urban School

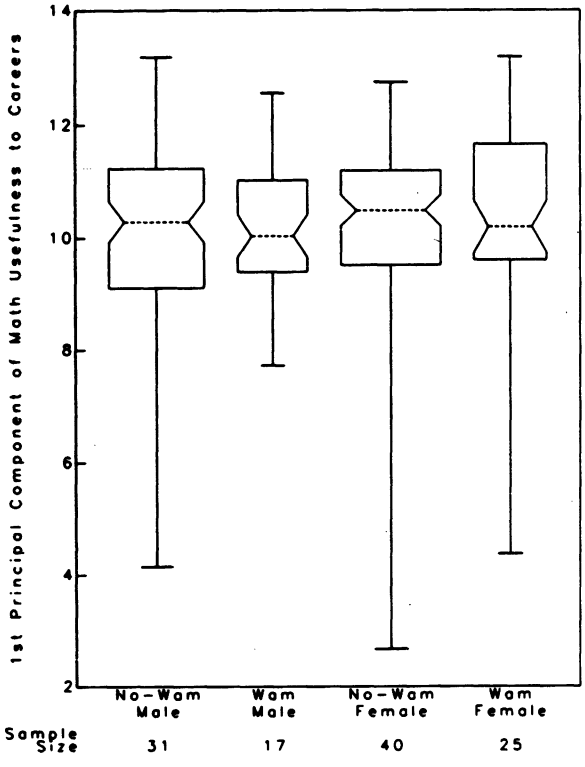


The WAM audience was composed of both sexes. So perhaps this significant difference was due solely to effects on the male respondents, with no effect on the females. Hence the students were regrouped, separating males and females. Students who did not indicate their sex were eliminated. Now there are four sample distributions — corresponding to no-WAM males, WAM males, no-WAM females, and WAM females — as shown in Figure 7.

The WAM female median is higher than the no-WAM female median. The sample sizes, which are indicated on the figure, are quite small and the enlarged notches now overlap slightly.

The same conclusion can be drawn from the two male groups. The dashed box indicates that the notches for the WAM males are so wide, due to the small sample size, that they extend beyond at least one quartile.

Figure 8
Suburban School

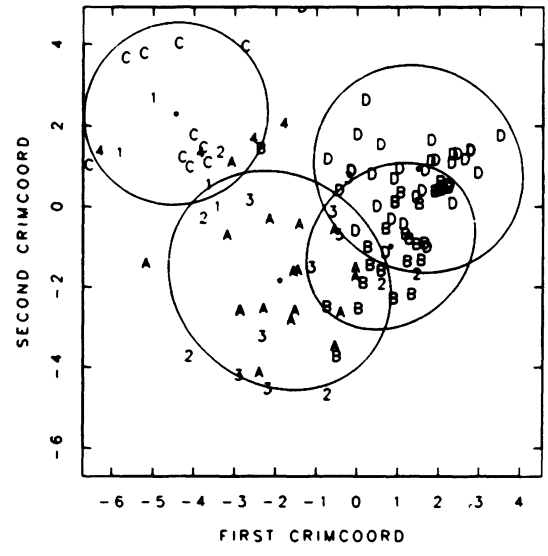


Figures 6 and 7 suggest that a WAM talk increases awareness of the importance of math to careers for both sexes in the urban school. However, Figure 8 shows that when using the same procedure in the suburban school no such effect was found. All four medians are about the same. Since the coefficients of the suburban school's principal axis differ from those of the urban school, direct comparison between Figures 7 and 8 cannot be made. Though no WAM effect is seen here, projection along the first discriminant axis did uncover an effect.

Finally, we demonstrate another multivariate approach for analyzing this data. Perhaps there is a natural grouping of students suggested by their responses which may have some interesting interpretation. Again, focus is on the eight responses regarding the importance of math to careers using the urban school as it has proved to show the greatest WAM effect thus far. To find this natural grouping hierarchical clustering (Johnson, 1967) was used on all urban students in the eight dimensional space without regard to WAM or sex. The focus of the discussion will be on the display and interpretation of the clusters.

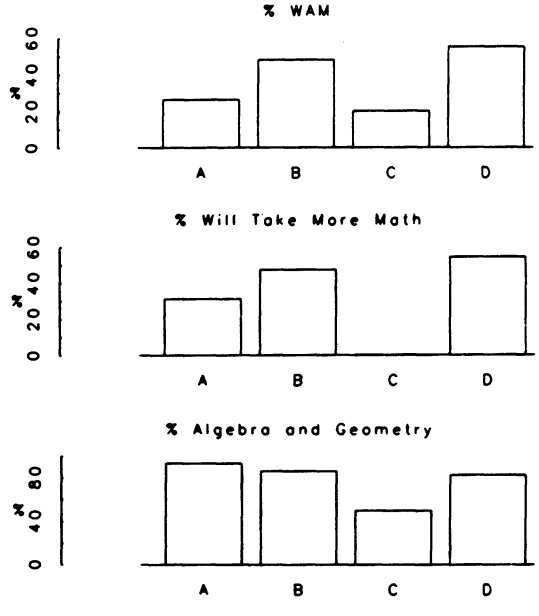
Four main clusters (A,B,C,D) and five smaller clusters (1,2,3,4,5) with fewer than eight students in each were found. In order to display and evaluate how well separated the four larger clusters are, a graphical technique suggested by Fowlkes and McRae (1977) was used. Figure 9 shows a scatter plot of all the students responding to all eight careers as plotted in the two-dimensional space which best shows the separation between the clusters — the first two discriminant coordinates. That is, the X-axis is the direction which most greatly separates the mean vectors of the clusters relative to the within group separation. The Y-axis is the direction which gives the next greatest separation of cluster means relative to the within group separation such that the projected data on the two axes are uncorrelated. The cluster sizes and individual cluster covariance structures are used to determine these directions. The ellipses, which correspond to the four large clusters, are centered at the cluster means and are scaled so that about 90% of the corresponding cluster is expected to be inside the ellipse.

Figure 9
1st Two Discriminant Coordinants
for Math Usefulness Perceptions



In this two dimensional space clusters B and D do not seem very different, whereas A and especially C are separated from the others and from each other.

Figure 10
Urban School



Next we attempted to see if there was a relationship between these clusters and any of the demographic variables. Figure 10 indicates the three demographic variables which did seem to discriminate among clusters. In Figure 10a each bar represents the percentage of WAM respondents in the corresponding cluster. Clusters B and D have a high WAM precentage — 1/2 are WAM where only 1/3 of the total population is WAM. The other two clusters, A and C, have fewer WAM respondents than the average. Figure 10b shows that the two high WAM clusters (B and D) also have a higher percentage of students who want to take more math than in the high no-WAM clusters. This is very encouraging, though not conclusive. Finally, from Figure 10c, it is seen that cluster C, which is located at the top of Figure 9, can be differentiated from the other three clusters in that fewer of its members were taking

algebra and geometry. There is no noticeable difference between clusters in terms of any other demographic variables.

Summarizing Figure 9, clusters B and D are predominantly WAM and also contain more students who want to take more math. It is encouraging that this interest in math is confounded with the WAM effect. Clusters A, B and D have a larger percentage of algebra and geometry students than cluster C.

In summary,

(1) Indications of a WAM effect were seen in the urban school.

(2) Only slight evidence of a WAM effect was found in the suburban school. In analyses not discussed above, male and female responses to the math-as-a-male-domain statements were significantly different. Further study is warranted to determine if this is masking a slight WAM effect.

(3) Only two of the schools visited have been analyzed. Consequently caution must be exercised in considering these preliminary results.

Clearly, as many questions have been raised as answered by this analysis. Additional work is in progress.

REFERENCES

Ernst, J. (1976). Mathematics and sex. *American Mathematical Monthly. 83,* 595-613.

Fennema, E. and Sherman, J. A. (1976). Fennema-Sherman mathematics attitudes scales: instruments designed to measure attitudes toward the learning of mathematics by females and males. *Catalog of Selected Documents in Psychology 6.*

Fowlkes, E. B. and McRae, J. (1977). Graphical techniques for displaying multidimensional clusters. Unpublished document.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika 32,* 241-54.

Mathematics Association of America (MAA). Women and Mathematics. A pamphlet describing the program available from MAA, 1225 Connecticut Ave., N.W., Washington, D.C. 20036.

McGill, R. Tukey, J. W. and Larsen, W. A. (1977). Variations of the Box Plots. To appear in *American Statistician.*

AN INDIVIDUAL GROWTH MODEL PERSPECTIVE FOR EVALUATING EDUCATIONAL PROGRAMS

Judith F. Strenio, Harvard University & The Huron Institute
Anthony S. Bryk, Harvard University & The Huron Institute
Herbert I. Weisberg - The Huron Institute

1. Introduction

In evaluating educational programs it is often not possible to conduct a rigorous randomized experiment. Estimates of program effects must be based on uncontrolled observational studies or partially controlled quasi-experiments. These studies generally involve comparisons of treatment group performance with that of a nonequivalent control group. Because the groups being compared are not completely equivalent prior to the intervention, observed outcome differences may reflect these prior differences in addition to the treatment effect. That is, estimates of the effect based on a direct comparison of post-treatment measures will be biased.

Traditional analysis methods use an adjustment approach in attempting to reduce this bias. Pre-treatment differences between a treatment and control group are modelled. Statistical techniques based on the model are used to compensate, or adjust for these initial differences when comparing outcome data for the two groups.

One of the major potential sources of bias in such studies derives from the fact that individuals grow at different rates in the absence of a treatment. Thus the effects of a program may be confounded with natural growth, or maturation. In a previous paper (Bryk and Weisberg, 1977) we have detailed some of the problems encountered by traditional statistical methods for analyzing quasi-experiments when individuals are growing.

In this paper we discuss an alternative analysis strategy based on a projection approach. Utilizing information in the data set on individual growth, the strategy involves explicitly projecting the growth the program group would have achieved without any intervention. Actual growth can then be compared with projected growth; the difference is termed the value-added by the program.

The value-added technique was originally presented (Bryk and Weisberg, 1976) in terms of a very restricted model: all individuals were assumed to have identical growth rates. In this paper we extend the model to consider variable individual growth rates.

Note that, unlike adjustment techniques, the value-added approach does not necessarily assume the availability of data on an untreated control group. It is essentially a single-group design. On the other hand, it does require a sufficient combination of theory and empirical data to estimate natural growth. In this paper we assume that subjects are tested twice: once prior to the program, at a pretest time that we denote by t_1, and once at the end, at a posttest time t_2. Our objective is to estimate the average increment at the posttest time which is attributable to program experience.

2. Model and Rationale

We assumed that each individual's growth is a linear function of age. Let us denote by $a_i(t)$ the age of individual i at time t. Individuals are assumed to vary in terms of growth rate π_i and onset age δ_i (the age when non-negligible growth begins). Moreover, they are assumed to be sampled from a population in which π and δ are independently distributed with means μ_π, μ_δ and variances σ^2_π, σ^2_δ.

This model represents the simplest situation that incorporates varying individual growth. While too simple to represent realistically many educational processes, linear growth may be a reasonable approximation over a short term even if long-term growth has a more complex form.

For the present, we will also assume that δ and π are distributed to children independently of their age at pretest. That is, the older children are not more likely to have started growth at a particular age than the younger, nor are they growing slower or faster. We examine this assumption, and some consequences of its violation, in a later section.

Finally, we assume that observed growth $Y_i(t)$ is made up of two components: systematic growth $G_i(t)$ and a random noise component $R_i(t)$ determined by the particular circumstances at time t. Our basic model can be represented as

$$(1)\quad G_i(t) = \begin{cases} \pi_i[a_i(t) - \delta_i] & \text{for } a_i(t) \geq \delta_i \\ 0 & a_i(t) < \delta_i \end{cases}$$

$$(2)\quad Y_i(t) = G_i(t) + R_i(t)$$

where for all t, t'

$$E[R_i(t)] = 0$$

$$\mathrm{Cov}[R_i(t), R_i(t')] = 0$$

We also assume that the pretest time t_1 is set so that all subjects have begun to grow by that time. Combining equations (1) and (2) we can write

$$(3)\quad Y_i(t_1) = \pi_i[a_i(t_1) - \delta_i] + R_i(t_1)$$

Let us for convenience define

$$(4)\quad \Delta = t_2 - t_1$$

Then if no treatment were introduced, we would have

$$(5)\quad Y_i(t_2) = \pi_i[a_i(t_2) - \delta_i] + R_i(t_2)$$
$$= G_i(t_1) + \pi_i\Delta + R_i(t_2)$$

In order to model a treatment effect, we assume that over the time interval t_1 to t_2 the treatment increases each subject's growth by an amount v_i (the value-added). The mean and variance of v_i are μ_v and σ^2_v and v is assumed to be uncorrelated with any other variable in the model. Since v_i is a random variable, this model in principle allows for individual effects. Finally, then, we can represent the measured growth that subject i in the program group achieves by time t_2 as

$$(6)\quad Y_i(t_2) = G_i(t_1) + \pi_i\Delta + v_i + R_i(t_2)$$

We take the estimation of μ_v as the object of our analysis.

Before proceeding further with the examination of this model, we present the rationale underlying the method. During the period between pretest and

posttest, the average growth for the treatment group is $\bar{Y}(t_2) - \bar{Y}(t_1)$. The expected growth under the model is $\mu_\pi\Delta$. If we knew the value of μ_π, a natural estimator of the value-added would be $\bar{Y}(t_2) - \bar{Y}(t_1) - \mu_\pi\Delta$. So if we have an estimator $\hat{\mu}_\pi$ of μ_π we might use

$$(7) \quad V = \bar{Y}(t_2) - \bar{Y}(t_1) - \hat{\mu}_\pi\Delta$$

From equations (5) and (6) it is clear that any unbiased $\hat{\mu}_\pi$ will yield an unbiased estimator $\hat{\mu}_V$ of μ_V. In this paper we propose to use the ordinary least-squares regression coefficient of $Y(t_1)$ on age. This estimator is simple to compute nd intuitively appealing. In the next section we show that it is unbiased.

3. Examining the Value-Added Method: Properties of $\hat{\mu}_\pi$.

In this section we consider some properties of the least-squares regression coefficient we are proposing as an estimate of μ_π.

Lemma 1: Taking expectation over the distributions of π and δ,

$$E(\hat{\mu}_\pi) = \mu_\pi.$$

Proof: Our model is given by equation (3) with π_i, δ_i, and R_i mutually independent. We can rewrite this equation as

$$(8) \quad Y_i(t_1) = \mu_\pi a_i(t_1) - \mu_\pi\mu_\delta + \{(\pi_i - \mu_\pi)a_i(t_1) - (\pi_i\delta_i - \mu_\pi\mu_\delta) + R_i(t_1)\}$$

This equation is now in the form

$$(9) \quad Y_i = \alpha + \beta a_i + e_i$$

with

$$\alpha = -\mu_\pi\mu_\delta$$

$$\beta = \mu_\pi$$

$$e_i = \{(\pi_i - \mu_\pi)a_i(t_1) - (\pi_i\delta_i - \mu_\pi\mu_\delta) + R_i(t)\}$$

Under our assumptions it is straightforward to obtain

$$(10) \quad E(e_i|a_i) = 0.$$

Thus our model satisfies the usual conditions under which ordinary least squares yields unbiased estimates of α and β. Q.E.D.

We note, however, that the variance of the error term works out to be

$$(11) \quad \mathrm{Var}(\mathrm{error}_i) = a_i^2(t_1)\sigma_\pi^2 - 2a_i(t_1)\sigma_\pi^2\mu_\delta + \sigma_\pi^2\sigma_\delta^2 + \sigma_\pi^2\mu_\delta^2 + \mu_\pi^2\sigma_\delta^2 + \sigma_R^2.$$

Thus the error variance is a quadratic function of $a_i(t_1)$ and the OLS estimate, though unbiased, will be inefficient. In practice, we might wish to use a generalized least squares procedure to estimate μ_π. Implementing this idea involves some complex problems which we are currently investigating.

We next derive the variance of $\hat{\mu}_\pi$:

Lemma 2:

$$(12) \quad \mathrm{Var}(\hat{\mu}_\pi) = \frac{\sigma_\pi^2 \Sigma A_i^2 a_i^2(t_1) - 2K_2 \Sigma A_i^2 a_i(t_1)}{(\Sigma A_i^2)^2} + \frac{\sigma_R^2 + K_1}{\Sigma A_i^2}$$

Where $K_1 = \mathrm{Var}(\pi\delta) = \mu_\delta^2\sigma_\pi^2 + \mu_\pi^2\sigma_\delta^2 + \sigma_\pi^2\sigma_\delta^2$

and $K_2 = \mathrm{Cov}(\pi,\pi\delta) = \sigma_\pi^2\mu_\delta$.

Proof: Let $\bar{a}(t)$ be the average age of the program group at time t, and $A_i = a_i(t) - \bar{a}(t)$, noting that $\Sigma A_i = 0$. Then the usual least-squares estimate is given by:

$$(13) \quad \hat{\mu}_\pi = \frac{\Sigma A_i[Y_i(t_1) - \bar{Y}(t_1)]}{\Sigma A_i^2},$$

which simplifies to

$$(14) \quad \hat{\mu}_\pi = \frac{\Sigma A_i Y_i(t_1)}{\Sigma A_i^2} \quad \text{because } \Sigma A_i = 0. \quad \text{Now}$$

$$(15) \quad \mathrm{Var}(\hat{\mu}_\pi) = \frac{\Sigma A_i^2 \mathrm{Var}[Y_i(t_1)]}{(\Sigma A_i^2)^2}$$

There are no covariance terms, since $\mathrm{Cov}(Y_i,Y_j) = 0$.

Thus we require $\mathrm{Var}[Y_i(t_1)]$ (recalling $Y_i(t_1)$ from equation (3)):

$$(16) \quad \mathrm{Var}[Y_i(t_1)] = a_i^2(t_1)\sigma_\pi^2 + \mathrm{Var}(\pi_i\delta_i) + \sigma_R^2 - 2a_i(t_1)\,\mathrm{Cov}(\pi_i, \pi_i\delta_i)$$

Because we assume π_i and δ_i are independent, we find $\mathrm{Var}(\pi\delta) = K_1$, and $\mathrm{Cov}(\pi,\pi\delta) = K_2$ as given in the statement of the Theorem. Thus equation (11) is indeed the variance of $\hat{\mu}_\pi$. Q.E.D.

This gives the variance of $\hat{\mu}_\pi$ in terms of the parameters of the model. Note that the usual variance of $\hat{\mu}_\pi$ is simply $\frac{\sigma_R^2}{\Sigma A_i^2}$, one term of our variance.

4. Examining the Value-Added Technique: Properties of V.

We now consider some statistical properties of the value-added estimator itself. From equations (5) through (7) we have

$$(17) \quad V = \frac{\Delta}{n}\sum_{i=1}^{n}\pi_i + \frac{1}{n}\sum_{i=1}^{n}v_i + \frac{1}{n}\sum_{i=1}^{n}R_i(t_2) - \frac{1}{n}\sum_{i=1}^{n}R_i(t_1) - \hat{\mu}_\pi\Delta$$

Theorem:

(18) (a) $E(V) = \mu_V$

(b) $\text{Var}(V) = \dfrac{\sigma_v^2 + 2\sigma_R^2 - \Delta^2\sigma_\pi^2 + n\Delta^2 \text{Var}(\hat{\mu}_\pi)}{n}$

Proof: (a) Apply expectation to both sides of equation (17);

(19) $E(V) = \dfrac{\Delta}{n}(n\mu_\pi) + \dfrac{1}{n}(n\mu_V) + 0 - 0 - E(\hat{\mu}_\pi)\Delta$

$= \Delta\mu_\pi + \mu_V - \mu_\pi\Delta$

$= \mu_V.$

(b) Take variances of both sides of equation (17);

(20) $\text{Var}(V) = \dfrac{\Delta^2\sigma_\pi^2}{n} + \dfrac{\sigma_v^2}{n} + \dfrac{2\sigma_R^2}{n} + \Delta^2 \text{Var}(\hat{\mu}_\pi)$

$- \dfrac{2\Delta^2}{n} \sum_{j=1}^{n} \text{Cov}(\pi_i, \hat{\mu}_\pi).$

We already have $\text{Var}(\hat{\mu}_\pi)$ from Lemma 2.

We require $\sum_{i=1}^{n} \text{Cov}(\pi_i, \hat{\mu}_\pi)$:

First, using equation (13),

(21) $\text{Cov}(\pi_i, \hat{\mu}_\pi) = \dfrac{\sum_{j=1}^{n} A_j \text{Cov}[\pi_i, Y_j(t_1)]}{\sum_{j=1}^{n} A_j^2}.$

We have

(22) $\text{Cov}[\pi_i, Y_j(t_1)] = \begin{cases} 0 & \text{if } i \neq j \\ -K_2 + \sigma_\pi^2 a_i(t_1) & \text{if } i = j. \end{cases}$

So

(23) $\sum_{i=1}^{n} \text{Cov}(\pi_i, \hat{\mu}_\pi) = \dfrac{-K_2 \sum_{i=1}^{n} A_i + \sigma_\pi^2 \sum_{i=1}^{n} A_i a_i(t_1)}{\sum_{i=1}^{n} A_i^2}$

$= \sigma_\pi^2$

Substituting this into equation (20) above and collecting terms, we get the expression in (b) of the Theorem. Q.E.D.

Comments On This Theorem:

1) V is unbiased, because $\hat{\mu}_\pi$ is. It may look as though we are using independent variables with error when we write the model--that is, we want $[a_i(t_1) - \delta_i]$ and we know only $a_i(t_1)$--but the proof of $\hat{\mu}_\pi$'s unbiasedness shows that age alone is valid for estimating μ_π, and we usually know age accurately.

2) If everyone had the same growth rate, and we knew what it was, $\text{Var}(V)$ would be $\frac{2\sigma_R^2}{n} + \frac{\sigma_v^2}{n}$. The $\frac{\Delta^2\sigma_R^2}{n}$ arises from the differences in growth rate, and the other terms from the estimation of μ_π from data.

5. Testing Significance of V

In practice, we generally wish not only to estimate the treatment effect, but also to test its significance and/or to state a confidence interval. To derive such tests and intervals requires derivation of the distribution of V under various assumptions about the distribution of π, δ, R and V. In the previous section we derived an expression for the variance of V. It is not obvious how to use it in developing the necessary statistical procedures.

While the development of procedures based on the distribution of V is worth pursuing, another general purpose approach may prove useful. The jackknife technique (described in Chapter 8 of Mosteller and Tukey, 1977) can provide both a test statistic and standard error for use in forming confidence intervals. To apply the jackknife to our situation is fairly straightforward. Let $\hat{\mu}_\pi(\text{all})$ be the least squares coefficient computed from the whole data set, and let $\hat{\mu}_\pi(i)$ be the coefficient computed with only observation i removed from the data. Then for each individual i a pseudo-value V_{*i} is calculated as

(24) $V_{*i} = Y_i(t_2) - Y_i(t_1) - \hat{\mu}_{\pi *i}\Delta$

where $\hat{\mu}_{\pi *i} = n\hat{\mu}_\pi(\text{all}) - (n-1)\hat{\mu}_\pi(i)$

The V_{*i} are then treated as data points. Their mean $\bar{V}_*$ provides an unbiased estimate of μ_V, and the standard error allows calculation of a t-statistic with (n - 1) degrees of freedom for testing or interval estimation.

6. Illustrative Example

We take as an example a subset of the data collected to evaluate the Head Start Planned Variation program. We will consider the data on one curricular model for one outcome, the Preschool Inventory (described in Walker, Bane and Bryk, 1973). All children were pretested at ages between 50 and 63 months, with mean age 56.80 months.

The mean pretest score is 14.116 and the mean posttest score is 20.454, out of a possible 32. The mean time between tests is 7.40 months. The least-squares regression coefficient of pretest on age is 0.484. Thus the estimated value-added is given by

$V = 20.454 - 14.116 - (0.484)(7.40) = 2.756.$

To test this value for statistical significance, the jackknife procedure was carried out as described above. This resulted in a mean $\bar{V}_*$ of 2.764 which has a standard error of 1.192. The resulting t-value of 2.319 with 96 degrees of freedom is significant at the .05 level.

7. Independence of Age and Individual Growth Characteristics.

The value-added method as applied in this paper uses the cross-sectional relationship between score and age at a particular point in time, t_1, to estimate the mean growth rate for

the program group. This approach assumes that individual growth characteristics (reflected by π_i and δ_i in our model) are independent of age. If there exists a systematic relationship between these characteristics and age, then the pretest/age relationship reflects not only individual growth but also the age gradient of δ_i and π_i.

Non-independence can occur in at least two different ways. First, in the population from which individuals are sampled, there may be historical trends causing children born at different times to differ. For example, during the period when Sesame Street was first being introduced, younger children exposed to the program may have had different characteristics from older children not exposed.

Second, even if this stable universe assumption (Kodlin and Thompson, 1958) is true for the population being studied, selection of the experimental sample may introduce an age by characteristic relationship. Criteria of selection may have operated so that younger children tend to have different characteristics from older ones. For example, the youngest children in a Head Start program may be there because they are unusually mature for their ages, possibly entering a bit below the age threshold. The oldest children may be particularly slow, possibly even old enough to enter kindergarten but not really ready.

To understand the effects of these phenomena, we develop a simple model. Let A_i represent the deviation of a subject from the group mean (as before),

$$A_i = a_i(t_1) - \bar{a}(t_1) .$$

Let us assume further that the expected values of π and δ are functions of A_i:

$$\begin{aligned} E(\pi_i|A_i) &= f(A_i) \\ E(\delta_i|A_i) &= g(A_i) \end{aligned} \qquad (25)$$

To see how this would affect our value-added technique, we look first at $E[Y_i(t_1)|A_i]$, to see what the age versus pretest score graph will look like; that is, what the cross-sectional data will become. We have equation (3) for $Y_i(t_1)$. If we take expectations, substitute for $E(\pi_i|A_i)$ and $E(\delta_i|A_i)$ from equation (25) and rewrite $a_i(t_1)$ as $[A_i + \bar{a}(t_1)]$, we arrive at this result:

$$(26) \quad E[Y_i(t_1)|A_i] = f(A_i)\ [A_i + \bar{a}(t_1) - g(A_i)]$$

We can see that unless we choose some special f and g, or they have some special parameterization, $Y_i(t_1)$ will become a nonlinear function of age. Thus the age versus pretest score graph will show curvature, and we can test for age selection by testing the age by pretest score graph for nonlinearity.

8. Linear Individual Growth Assumption

Another possible problem is that individual growth may be non-linear. With extreme nonlinearity, the linear approximation will not be trust-worthy even in the short term. For example, on a particular test as soon as a subject has thoroughly mastered all items, Y_i flattens out at the perfect score (although the type of skill that had been measured may continue to improve).

If we wish to retain the idea that each subject has different parameters of the growth curve, then this problem becomes very complex. In Bryk (1977) an individual negative exponential growth curve is examined. This is a very appealing model for growth, which has been widely used in biological growth studies. Bryk derives the expected value of $\bar{Y}(t)$ and shows that it is not a negative exponential function of time. More generally, the average of non-linear growth curves, even when taken over subjects the same age, will not trace out the same shaped curve as the individuals are following. This will make model identification difficult when only cross-sectional data are available. But, we with the age-growth dependence problem, at least we can see that the age versus pretest plot will not be linear. So, again, a test for linearity can be used as an indicator of failure to meet the model's assumptions.

9. Directions for Further Research

The use of the ordinary least-squares regression coefficient to estimate μ_π was chosen for simplicity and intuitive appeal. We have shown that it leads to an unbiased estimate of μ_v. In large samples, this estimator should be quite adequate. With smaller samples, however, it is not clear whether this approach yields estimates that are efficient enough for practical purposes. This question needs to be investigated. It may well be necessary to develop alternative estimation procedures with greater efficiency.

Secondly, the model we have assumed here is the simplest model which incorporated diffential growth rates across individuals. Investigation of more complex models and development of corresponding analysis strategies is needed. For example, models could reflect various kinds of dependence between π and δ, various forms of nonlinear growth, and various kinds of age-selection effects.

Finally, a very important research area lies in the attempt to assess individual values of v_i. If we could do this, we would be able to estimate σ_v^2 and the distribution of v. We could also estimate interactions between the v_i and measured covariates. Particularly in this educational context, we are often interested in more than the simple average effect. Rather, we wish to discover which programs help which students, by how much.

In order to achieve this objective, the estimation of the individual v_i seems necessary. To accomplish this, however, more information will be needed. We have gone quite far with only two cross-sections, one as proxy for longitudinal data and the other to gauge progress. The next logical step is to gather more data on the same group, so that we really have, say, four or five data points on each subject. Through the combination of cross-sectional and longitudinal perspectives on the same data set we should be able to estimate more precisely both the mean effect μ_v and other aspects of the distribution of individual v_i's.

References

Bryk, A.S. An Investigation of the Effectiveness of Alternative Statistical Adjustment Strategies in the Analysis of Quasi-Experimental Growth Data. Thesis for Doctor of Education,

Harvard University (1977).

Bryk, A.S. and Weisberg, H.I. Value-Added Analysis: A Dynamic Approach to the Estimation of Treatment Effects. Journal of Educational Statistics, Vol. 1, No. 2 (1976).

Bryk, A.S. and Weisberg, H.I. Use of the Nonequivalent Control Group Design When Subjects Are Growing. Psychological Bulletin, Vol. 85, No. 4 (1977).

Kodlin, D. and Thompson, D.J. An Appraisal of the Longitudinal Approach to Studies of Growth and Development. Monographs of the Society for Research in Child Development, Inc., Vol. 23, Serial No. 67, No.1 (1958).

Mosteller, F. and Tukey, J.W. Data Analysis and Regression: A Second Course in Statistics. Reading, MA: Addison-Wesley (1977).

Walker, D.K., Bane, M.J. and Bryk, A.S. The Quality of the Head Start Planned Variation Data. Cambridge, MA: Huron Institute (1973).

AN EXPLORATION IN EDUCATION AND ATTAINMENT: THE SHEEPSKIN EFFECT

James E. Katz, Massachusetts Institute of Technology

PROBLEM

Is there a "sheepskin effect"? Do additional benefits accrue to individuals who pass "certification points" (generally understood to be either high school or college completion) which go above and beyond the regular increments for each year of schooling completed?

It has been demonstrated that education serves as a screen blocking those with low education from, and facilitating the entry of those with high education to, desirable and prestigious jobs. Yet it is unclear whether or not the attainment of a certification point (narrowly defined as the passage of a particular year) in and of itself makes a significant difference in the socio-economic level a person attains. This paper demonstrates whether or not the certification effect exists, and reviews some consequences for social policy.

PERSPECTIVE

The human capital model developed by Schultz (1964), Becker (1964), and Mincer (1975) view education as an actual investment of finite resources, subsuming the educative process under an economic model. However, the human capital model does not distinguish specific years as being more economically significant than another contiguous year.

A parallel, but distinct concern has been the argument that the role of education is to screen individuals. The idea here is that the goal of education is to attain a certificate which then assigns one a "niche" in society. Thus, the idea of screening has both a socio-psychological and an economic rationale.

It is also argued that screening is the expensive result of an imperfect market.

Taubman and Wales (1974) argue that education is both an investment and a screening device.

Ivar Berg, in his book, Education and and Jobs: The Great Training Robbery (1970), exhaustively examines the certification uses of education in industry. Berg demonstrates that job skill requirements have changed little between 1940 and 1965, but educational requirements have risen tremendously.

Ironically, despite the importance of these economic and sociological studies, research interpretation of the relationship between education and occupational status (or achievement) is usually restricted to only a narrow segment of the diverse American population. In addition, most studies approach education as a more or less continuous process and pay relatively little attention to specific certification points.

This study incorporates white males and females of all working ages and all occupations (except farming) and its implications are subsequently broader.

It is important to point out that it is not the intention of this study to explain occupational prestige or earnings on the basis of various background variables. The goals of the present study are much more modest; it seeks to learn whether the certification point serves as a screen, not to job entry, but to higher occupational prestige and earnings.

METHODOLOGY

To test for a certification or sheepskin effect, I seek to see if the functional form of the regression equation is piece wise linear. This is tested by estimating the equation for working white males and females with 9-11 or 13-15 years of schooling separately for three broad work experience categories. Certification occurs if the actual mean is significantly above the predicted mean. The 1970 Public Use Sample files, 15 percent 1:1000 files of the U.S. census provide the data for this study. Every case was drawn in which the individual met the criterion. For this study all white males and females who had 9-16 years of schooling, were between 22 and 65 years of age, and worked full-time (more than 35 hours a week) were selected. The full sample size was 36, 304. The sample was broken down by three career groups: early career, those with 5-14; medium career, 15-29 years; and thirty and above years since leaving school.

Dollar earnings and occupational prestige (based on the Duncan scale, 1-1000) were the primary variables used to examine the presence or absence of a certification effect.

A 95% confidence interval was computed based on the data points for the three years before the certification points, the twelfth and sixteenth years. The data for the certification points was then tested to see whether or not it fell outside the confidence interval. The Walter-Lev (1953) test was used to test for significance at the .025 level.

FINDINGS

If career stage is controlled, there is a significant certification effect at the earliest career stage solely job prestige for white women who have graduated from college. At the mid-career stage, college graduation has a significant impact on job prestige for white males (.025 level). At the later career stage, both men and women demonstrate certification effects and at both college and high school graduation levels. For women both earnings and job prestige are significantly related to obtaining high school certification but there is no certification effect for college graduation. For white males job prestige is significantly related to higher earnings.

Clearly the certification effect is more prevalent for the late career group. Because of the cross-sectional nature of this study, it cannot be determined whether these differences can be explained by developmental causes (maturation, lag time before certification effects take hold, and so on) or generational (the certificate meant more for the older generation and its employers, the historical differences in the labor market and so on).

The findings indicate that for late career individuals, there is a strong certifica-

tion effect, and while there is some indication of a certification effect for those in earlier career groups it is much less pronounced.

IMPLICATIONS WITH SPECIAL REFERENCE TO ADULT EDUCATION

The findings reflect at least the partial existence of a certification or "sheepskin" effect. Rather than simply an increased transfer of cognitive material, the certification point also represents a socially symbolic achievement; it is a "rite of passage" denoting the crossing of an important although rather arbitrarily designated point in the education system. If the findings have validity, any theory which tries to explain the social functions of education must account for a "certification effect" ascribed to the completion of specific diploma-conferring years of education.

There is an abiding faith in America in what Ralph Turner (1960) has called "contest mobility", the idea that "elite status should only be given to those who earn it." Because society at large establishes the criteria of elite status, the possession of visible credentials becomes a vital component of success. Of all such credentials, the high school diploma would seem to be elemental and indispensible.

As Turner points, out, the "contest" idea defines the accepted mode of upward mobility, and in judging a contest there is always the fear of premature closure. Hence, in the educational sphere, options are provided so that adults who failed to attain credentials the first time around may try again. At the secondary level the most wide-spread of these options are high school credit or equivalency programs for adults. It is believed that many people have been denied life's rewards simply because they have failed to attain a credential -- regardless of their other inherent capabilities.

High school completion programs for adults then comprise a large and still growing field.

More and more adults are being converted to the idea that the mobility"contest" continues well past adolescence and so are earning larger and larger numbers of diplomas. At the same time, a quiet revolution is taking place in the courts which threatens to undermine the whole endeavor. The center of the controversy is a 1970 U.S. Supreme Court decision, Griggs vs Duke Power Company. In this decision the Court held that unless it could be demonstrated that a credential (in this case a high school diploma) or standardized examination was related to job performance it could not be used in personnel decisions related to job entrance or promotion.

The implication of Griggs, if broadly interpreted, could seriously undermine the usability of a high school diploma or even college degrees as an arbiter of job entrance or promotion. The Griggs decision may serve to accelerate the movement toward competency-based certification as the way out of a thorny predicament; how does a credential demonstrate anything more than the attainment of a credential? Only if the relationship of the credential to actual job skills is verified can this dilemma be resolved. The findings of this study, like the Griggs decision may be upsetting to those who accept on faith the inherent meaningfulness of a diploma. If a high school diploma is a poor indicator of skill attainment, it now appears that its posession does little to insure one of a higher income or greater job prestige.

RESOURCE ALLOCATION WITHIN SECONDARY SCHOOLS: A GOAL PROGRAMMING APPROACH

J. Michael Morgan, Western Kentucky University
Elchanan Cohn, University of South Carolina

The objective of this paper is to develop and implement a static educational resource-allocation model so that estimates of the resources necessary to satisfy a set of pre-specified conflicting educational outputs can be obtained. The outputs are ranked by secondary school administrators in order of their importance. The optimal resource mix is that which meets, as closely as possible, the output target values given by the school administrators. If the exact attainment is not possible, the output solution vector will be that which minimizes both the positive and negative deviations from the pre-specified targets. Since the determination of a price vector for the outputs of a state's educational system is virtually impossible, (and hence the determination of marginal values necessary for optimization in the traditional sense is unavailable), a model which computes efficient output vectors in terms of the physically necessary resource requirements will allow the school administrator to alter the input mix based on the subjective rankings of the output target values.

This study presents a goal programming/input-output model for the Pennsylvania secondary school system. The goals (output targets) of the model represent the Goals of Quality Education as defined by the Pennsylvania Educational Quality Assessment Program (E.Q.A.), and a brief description is presented in Table I. The data employed in the model consist of an aggregation of the individual rankings of the goals as expressed by twenty-eight school administrators in Pennsylvania; a primal objective function reflecting the priorities of the goals; a set of technical production constraints which represent the influence of input factors which can often be controlled by the school administrator; and a set of factor-availability constraints. The data reflecting administrator preferences and resource availability are drawn from a questionnaire submitted to selected school principals who have been participants in the E.Q.A. program. The technical production relationship has been estimated by Cohn [2].

This paper is divided into four major sections: (1) the presentation of a theoretical model; (2) a discussion of the data; (3) the empirical results; and (4) conclusions. The product of this study is twofold. First, it presents a workable model which can be applied directly to public school systems where a constrained efficient input mix is desirable. Second, the empirical results for Pennsylvania suggest that it is possible to increase the level of attainment of school outputs by altering the input mix available during the short period and to attain that resource mix which, over the long term, produces the most efficient output vector, given the subjective preferences and the state of the technological arts.

THEORETICAL MODEL

The goal programming approach to creating effective decision models has restrictive assumptions and requirements (Lee [2], pp. 32-35). One assumption is that the environment contains goals which are incompatible and incommensurable. A conflict area for the decision maker is therefore established, and, given a set of realistic constraining relationships, it is impossible to completely satisfy all of the goals simultaneously. With a set of incompatible and incommensurable goals, it must be assumed that the decision maker can correctly and meaningfully specify and ordinally rank his goals. The ranking assumption permits goal j to be revealed preferred to goal j+1 (assuming that each goal can be met only at the expense of the other). The establishment of priority factors is based on the ranking assumption and hence reflects the decision maker's subjective preference map. In addition to the ranking property, it must be assumed that the decision maker can specify deviational variables to be associated with each goal. It is necessary to be able to determine whether or not it is preferable to underachieve (d^-), overachieve (d^+), or exactly attain ($d^- - d^+$) each goal.[1] It is necessary also that goal attainment and the level of resource use measurements be proportional to the magnitudes which would be encountered if the model consisted of individual activities. The assumption and requirement that both the objective function and constraints are additive will insure that no joint interaction exists between any activities of either the goal attainment function or the constraining functions. In a goal programming model, non-integer solutions must be acceptable. The requirement of fractional solutions has the disadvantage that what may be optimal in terms of the model may be totally unrealistic in the real world. It must also be assumed in the model that the technical coefficients are constant, which invokes the requirement that the model must be evaluated from a static-analytic approach. Finally, it must be assumed that the number of constraints in the model exceeds the number of variables in order to prevent a trivial solution.

By properly specifying and examining the decision environment relevant to a particular situation, it is possible to formulate the constraints, choice space and objective function of the decision model. Once these three components have been established, it is possible to specify a goal programming model.

Suppose there exists an (M X N) simultaneous input-output model representing a school system where the outputs of the system are the desired goals of the production process, with M outputs and N inputs. Suppose, also, that the school administrator is able to assign priority weights to the outputs in such a manner that P_i is strictly preferred to P_{i+1}. Also, suppose that the estimated reduced form coefficients of the input-output-model and the level of resource availability are acceptable as constraining the system, and that some target level of goal attainment is desirable. The goal programming model might then take

the form:

(1) MINIMIZE:

$$Z = \sigma_1 \pm P_1 d_1 \pm + \sigma_2 \pm P_2 d_2 \pm + \ldots + \sigma_m \pm P_m d_m \pm +$$

$$\sum_{j=1}^{n} P_{m+1} (\sigma_{m+i}^{+} d_{m+i}^{+} + \sigma_{m+i}^{-} d_{m+i}^{-})$$

(2) SUBJECT TO:

$$b_{11}X_1 + b_{12}X_2 + \ldots + b_{1n}X_n - d_1^{+} + d_1^{-} = T_1 - S_1$$

$$b_{21}X_1 + b_{22}X_2 + \ldots + b_{2n}X_n - d_2^{+} + d_2^{-} = T_2 - S_2$$

. . . . .

. . . . .

. . . . .

$$b_{m1}X_1 + b_{m2}X_2 + \ldots + b_{mn}X_n - d_m^{+} + d_m^{-} = T_m - S_m$$

(2')

$$X_1 + d_{m+1}^{-} - d_{m+1}^{+} = X_1^*$$

$$X_2 + d_{m+2}^{-} - d_{m+2}^{+} = X_2^*$$

. . . .

. . . .

. . . .

$$X_n + d_{m+n}^{-} - d_{m+n}^{+} = X_n^*$$

(3) $X_j \leq h_j$ and $h_j \geq K_j$ $\quad i = 1, \ldots, m$

$(T_i - S_i)$, d_i^{+}, $d_i^{-} \geq 0$ $\quad j = 1, \ldots, n$

where:

Z = the objective function of the model with the priority factors, determined by the administrators preference function, associated with each goal.

b_{ij} = the estimated reduced form input coefficient from the simultaneous system. These coefficients represent those inputs over which the decision-maker has control.

X_j = the inputs over which the administrator has control. These inputs may be altered by the decision-maker when he attempts to optimize his objective function.

d_i^{+} = deviational variable representing the overachievement of goal i with its value determined ex post in solution.

d_i^{-} = deviational variable representing the underachievement of goal i (also determined ex post).

P_i = the preemptive priority factor for the ith goal.

T_i = the predetermined target level for each goal.

S_i = the contribution to the ith goal attributable to the socio-demographic variables and the variables over which the educational administrator has no control. The expression for S_i is additive and linear.

K_j = the level of resource utilization.

h_j = the level of resource availability.

$\sigma_i^{\pm}$ = an ex-ante determined coefficient of regret (weighting factor) associated with goals which occupy the same priority level in the objective function. The coefficient of regret gives the relative importance of goal i to goal j when each occupies priority level k. Also, it is required that $\sigma_i^{\pm} \geq 0$.

m = the number of goals, including subgoals.

n = the number of inputs over which the administrator has control.

X^*_j = the desired value of the subgoal associated with each manipulable variable.

TABLE I
GOALS OF QUALITY EDUCATION

Goal	Short Name	Target Output Number
I	Self Concept	1
II	Understanding Others	2
III-V	Verbal Basic Skills	3
III-M	Math Basic Skills	4
IV	Learning Attitudes	5
V	Citizenship	6
VI	Health Habits	7
VII-P	Creativity Potential	8
VII-O	Creativity Output	9
VIII	Vocational Development	10
IX	Knowledge of Past	11
X	Readiness for Change	12

Source: Cohn and Millman [2], p. 58. A more detailed description is contained in Beers [1] and Cohn and Millman [2], Table A-1.

In the above model, note that the objective function (1) incorporates the preemptive priority factors of the decision maker. The priority factors indicate which goal should be met first and continue through to the last goal. The preemptive priority factors, however, do not indicate how much goal i is preferred over goal j. The objective function also expresses the deviational variable (d_i) in terms of either + or -. In the actual model either one or both deviations will be assigned to each goal (priority level), depending on the decision maker's preferences. The case where both signs appear associated with a single priority level indicates that the decision maker seeks to exactly attain his goals and thus wishes to minimize both under and overachievement.

The expression $\sum_{i=1}^{m} P_{m+1} (\sigma_{m+i}^{+} d_{m+i}^{+} + \sigma_{m+i}^{-} d_{m+i}^{-})$ allows for the set of factor constraints, as given in (2'), to enter into the objective function as a subgoal. The factor availability subgoal must also be assigned a priority factor (P_{m+1}). It is necessary in this model that the factor constraints be incorporated directly into the objective function since they will determine the boundaries of the choice space and hence determine the feasible region. When no boundaries are explicitly expressed in the model, then $-\infty \leq K_j \leq +\infty$ is the boundary. Also, in the set of factor constraints, the positive and negative deviational variables indicate that the attainment of a target level of factor utilization, X^*_j, is desired. The assignment of a priority level to the factor constraints

depends upon the decision maker's particular goal structure and hence may range from the highest to the lowest point in the ordering.

The constraints (2) reflect the input-output technical coefficients of production. The deviational variables associated with each production constraint reflect a particular goal of the system. It should be noted that $-d_i^+ + d_i^-$ incorporated into the production constraints suggest that only the exact achievement of the goal is desirable, and therefore both positive and negative deviations are to be minimized. This is only one particular case, and the decision maker could indicate that either over or underachievement is desirable.

The objective function of the general model thus relates the priorities (P_j) of a goal to the production function associated with that goal. That is, $P_1 d_1 \pm$ indicates that the highest priority of the model is to be assigned to the exact achievement of goal one. Goal one ($d_1 \pm = d_1^- - d_1^+$) is reflected by the first production constraint with its right hand side value assigned as a target for that goal. If the statement $P_1 d_m \pm$ appeared in the objective function, then top priority is assigned to the mth goal which is reflected by the mth production constraint. The objective function also reflects the desired level of resource utilization and availability by its inclusion of the factor subgoal. Each deviational variable within the subgoal priority expression relates some indicated level of resource usage. The expression $+d_{m+1}^- - d_{m+1}^+$ suggests that only some specific level of resources should be used and hence implies a very restricted boundary; however, this need not necessarily be the case.

The constraint set (2') also reflects the boundary constraints. It states that X_j is constrained by resource availability and legal or institutional constraining factors. And constraint set (3) imposes non-negativity restrictions on the deviational variables, the target values, and the X_j desired values.

The solution to a goal programming model using input-output-information and the ranked goals of the administrator provides an empirical identification of the input requirements, in terms of manipulable factors, necessary to attain all of the specified goals. Even though these resource requirements are identified, no assurance can be given that all goals are attained because the school system may not be able to purchase or secure the necessary inputs.

AN EMPIRICAL MODEL

The Data

The data employed in the goal programming/ input-output model presented in this study can be divided into two categories: (1) objective data designed to estimate the technical production relationships of the school system, and (2) subjective data designed to establish an ordered set of priorities with priority weights for a prespecified set of goals for the school system. The subjective data are also designed to establish the relative importance of various decision variables in a school's production process.

Input-Output data: The data describing the technical production relationship for the Pennsylvania secondary school system consist of a set of simultaneous production functions estimated by Cohn [2]. That study is based on output measures and input variables for fifty-three public secondary schools in Pennsylvania for the 1971-72 school year. Output data are based upon performance in basic skills and replies to various instruments measuring both cognitive and affective traits. The ten initial goals of quality education, presented in Table I, were modified by the Pennsylvania Department of Education to consist of a set of twelve measurable outputs of an educational program, by separating Goal III into verbal and mathematical skills, and by separating Goal VII into creativity potential and output. The manipulable input variables are presented in Table 2.

Two-stage least squares regression methods were applied to the data from which the reduced-form coefficients of the educational inputs were estimated. Since we are concerned here with a management model, we must distinguish between manipulable and non-manipulable inputs. The non-manipulable variables included in the study were composed of 14 different socio-cultural and demographic characteristics of the students. These were reduced, by means of factor analysis, to a set of four socio-economic factors (SEFAC). Although initially it was believed that the SEFAC variables would be an important explanatory element in the regression equations, test results indicate that they exert a minimal contribution to the predicted outputs of the system.[2]

Subjective data: To obtain information concerning the preference rankings and the availability of resources, a survey was conducted of the fifty-three school systems for which input-output data were already available. Of the fifty-three principals surveyed, twenty-eight acceptable responses were obtained and used in this study.

The twelve goals were ranked in order of their importance from 1 to 12, inclusive. A ranking of 1 designated the highest priority and 12 the lowest. The principals were also asked to indicate whether or not he or she would be willing to overachieve (+), underachieve (-), or exactly achieve (0), a particular goal, given budgetary limitations and resource availability. The priority rankings for each questionnaire do not permit any two goals to occupy the same priority level; however, when the objective function of the model is specified, two or more goals may occupy the same priority level. If it is the case that the same priority level is assigned to two or more goals in the objective function, then each must be appropriately weighted by its coefficient of regret.[3]

Resource use data: Since the manipulative inputs represent elements over which the administrator exercises some control, each principal surveyed was asked to assign maximum, desired, and minimum values to the specified set of input factors. In addition, the principal was asked to indicate whether or not he or she would prefer to overachieve (d_i^+), underachieve (d_i^-), or exactly achieve (d_i^0) the indicated desired level for each goal.

Although the full set of inputs contains eighteen manipulable factors, it was necessary to present only twelve variables to the principals. The justification for not listing all of the con-

TABLE 2
MANIPULATIVE INPUT VARIABLES USED BY COHN

Label	Description	Goal Program Symbol
TEDUC	Teacher's education	X_1
GUIDANCE	Counselors/pupil	X_2
TLOD	Teacher load	X_3
CSIZ	Class size	X_4
AEE	Average extracurricular expenditure/pupil	X_5
TSALARY	Teacher's salary	X_6
PSUP	Paraprofessional support	X_7
CUG	Curriculum units/grade	X_8
PRCO	Preparation coefficient (teacher specialization)	X_9
SFRAT	Student/academic faculty ratio	X_{10}
BOOKSP	Library books/pupil	X_{11}
TEXPER	Teacher's teaching experience	X_{12}
LIBRARY	Accessibility of library	X_{13}
CLPRACT	Teacher classroom practices	X_{14}
INNOVATE	School usage of innovations	X_{15}
BRAT	Ratio of actual enrollment to building capacity	X_{16}
AMAN	Administrative man hours/pupil	X_{17}
AXMAN	Auxiliary man hours/pupil	X_{18}

Source: Cohn and Millman [2], p. 59.

trollable factors and requesting the principals' responses rests primarily on the fact that certain of the variables do not lend themselves to the necessary quantification by school administrators. Also, some of the variables were based on the student or teacher's response along with that of the principal's. All of the eighteen manipulative variables are, however, included in the goal programming model.[4] Table 3 presents the descriptive statistics for the resource factors.

Target Values: The computation of the target values for the goals is based on the assumption that the student observations used by the Pennsylvania Department of Education during the E.Q.A. program were normally distributed. Thus, based on the Tchebysheff theorem, three standard deviation units above or below the initial target mean should capture the true population mean.[5] It is assumed, however, that the principals would prefer to have a value greater than the computed mean of the goal. As a result, the initial target value for the ith goal is computed as:

(4) $T_i^* = Y_i + 3\,\gamma_i$

where:

$\hat{\gamma}$ = the ith estimated standard deviation.

The contribution of the socio-economic variables (SEFAC) to the educational output targets should be removed since the administrator exercises no control over their input into the productive process. Since the SEFAC variables exert a very negligible influence on the target level of each goal, they were assigned a value of zero in the goal programming model.[6]

The initial target values, with the exclusion of the SEFAC variables, however, reflect the influence of both manipulative and non-manipulative variables. It is, therefore, necessary to remove this influence of the non-manipulative factors from the output targets since they cannot be controlled. In order to net the non-manipulative factors, we use the relation:[7]

(5) $\tilde{T}_i = T_i^* - [\alpha_i + (B_{nmi})(\bar{X}_{nmi})]$,

where:

$\tilde{T}_i$ = the target value of the ith goal reflecting only the influence of the manipulative factors.

T_i^* = the initial target value of the ith goal as expressed by (4).

α_i = the estimated intercept of the ith production relationship.

b_{nmi} = the estimated reduced form coefficient of the ith non-manipulative variable.

$\bar{X}_{nmi}$ = the mean of the ith non-manipulative factor.

The Objective Function

The objective function of the goal programming model is based on the concept of a value restricted transitive ranking and the simple majority rule decision criterion.[8] Based on the twenty-eight acceptable responses from our survey, we examined the binary choices of each principal for each possible pair of goals. Aggregation was based on the rule that for goal i to be preferred to goal j, at least fifteen principals (simple majority) must prefer i to j. Also, in order to determine the position of goals i, j, and k, in the ranking, we examined the frequency of binary comparisons between goals i and j, j and k, and i and k, respectively. Recalling that P_i represents the ith priority level of the jth output target ($d_i \pm$), the objective function takes the form:[9]

$$(6)\quad Z = P_1(\sigma_3^+ d_3^+ + \sigma_5^+ d_5^+) + P_2 d_4^+ + P_3[\sigma_1^+ d_1^+ + (\sigma_2^+ d_2^+ + \sigma_2^- d_2^-)] + P_4 d_4^+ + P_5(\sigma_8^+ d_8^+ + \sigma_8^- d_8^-) + P_6(\sigma_{12}^+ d_{12}^+ + \sigma_{12}^- d_{12}^-) + P_7 d_7^+ + P_8[(\sigma_9^+ d_9^+ + \sigma_9^- d_9^-) + \sigma_{10}^+ d_{10}^+] + P_9 (\sigma_{11}^+ d_{11}^+ + \sigma_{11}^- d_{11}^-) + P_{10}\left[\sum_{i=13}^{66} (\sigma_i^+ d_i^+ + \sigma_i^- d_i^-)\right].$$

GOAL PROGRAMMING RESULTS

The results of the goal programming model are presented in Table 4. The column labled RHS: Target Value provides the values estimated in expression (5) above. The priority column reflects the value-restricted ranking of (6). The overachievement (d_i^+) and underachievement (d_i^-) columns provide the ex post values of the deviational variables associated with each goal in the objective function. The sign column indicates the ex ante deviational variables assigned in the

TABLE 3
RESOURCE (FACTORS) STATISTICS

	Minimum Values			Desired Values			Maximum Values		
	Min-Min Value	Mean	Max-Min Value	Min Value	Mean	Max Value	Min-Max Value	Mean	Max-Max Value
TEDUC (X_1)	2	3.9	5	4	4.93	6	5	6.4	7
GUIDANCE (X_2)	100	202.7	300	200	267.9	400	250	392	600
TLOD (X_3)	5	15.0	30	17	26.1	30	25	33	40
CSIZ (X_4)	8	18.6	25	20	25	32	25	33.3	40
AEE(X_5)	0	18.6	75	2	44	150	5	60.3	200
TSALARY (X_6)	8500	10614	12000	10000	12642	16000	13500	17394	20000
PSUP (X_7)	0	16.5	40	0	31.25	48	0	37.4	50
CUG (X_8)	5	18.5	50	10	34.7	60	12	46.7	80
PRCO (X_9)	1	2.2	5	2	3.5	10	3	5.3	8
SFRAT (X_{10})	10	16.2	20	18	22.55	35	22	30.6	40
BOOKSP (X_{11})	3	9.7	20	8	20.85	100	10	31.4	50
TEXPER (X_{12})	0	3.2	8	2	8.9	19	10	17.8	37
LIBRARY (X_{13})*		1.0			4.37			5.0	
CLPRACT (X_{14})*		11.0			38.09			55.0	
INNOVATE (X_{15})*		12.0			33.55			60.0	
BRAT (X_{16})*		0.75			1.08			2.0	
AMAN (X_{17})*		1.0			3.95			10.0	
AXMAN (X_{18})*		1.0			8.02			16.0	

* Designated variables excluded from Questionnaire.

NOTE: The variables X_2, X_3, and X_6 are defined in this table somewhat differently than in the model. The results are based, however, upon consistent definitions of all variables.

TABLE 4
VALUE RESTRICTED GOAL PROGRAMMING RESULTS

RHS: Target Value	Goal	Priority	d_i^+ (Overachievement)	d_i^- (Underachievement)	Sign	$\sigma_i^+ d_i^+$ or $\sigma_i^- d_i^-$: Value or Over- or Underachievement
5.574	1	3	13.256	0.0	$-d_1^+$	17.259
6.194	2	3	0.0	6.903	$-d_2^+ + d_2^-$	20.363
1.620	3	1	0.685	0.0	$-d_3^+$	0.723
1.023	4	2	0.0	0.0	$-d_4^+$	0.0
22.802	5	1	0.0	0.0	$-d_5^+$	0.0
9.715	6	4	45.056	0.0	$-d_6^+$	45.056
0.0	7	7	11.338	0.0	$-d_7^+$	11.338
6.492	8	5	20.592	0.0	$-d_8^+ + d_8^-$	36.037
6.305	9	8	0.0	15.353	$-d_9^+ + d_9^-$	33.085
4.295	10	8	18.516	0.0	$-d_{10}^+$	23.552
13.644	11	9	17.821	0.0	$-d_{11}^+ + d_{11}^-$	33.255
0.0	12	6	6.205	0.0	$-d_{12}^+ + d_{12}^-$	15.791

objective function. Note that on levels one, three, and eight, two goals occupy the same priority level in the ranking. Also, for goals two, eight, nine, eleven, and twelve exact achievement is desired. As a result, their coefficients of regret are assigned a value greater than one. The column $\sigma_i^+ d_i^+$ or $\sigma_i^- d_i^-$: Value of Over- or Underachievement gives the magnitude of non-attainment of each goal. The minimized Z-value is 236.45.

From Table 4 it is clear that goals four and five (priority level one and two) have been exactly met. Also, goals one, three, six, seven, eight, ten, eleven, and twelve have been exceeded; only goals two and nine have not been achieved. Although the target values for goals two and nine have been underattained by an amount exceeding their initial target values, the target value, in solution, is zero.

The resource requirements necessary for solution are presented in Table 5. The impact of the restriction that in goal programming models non-

TABLE 5
VALUE RESTRICTED ORDERING RESOURCE REQUIREMENTS: INTERPRETATION

Variable		Required Usage	Interpretation
(X_1)	TEDUC	6.3	Teachers should possess a Master's degree plus two years.
(X_2)	GUIDANCE	.005	The pupil/counselor ratio in solution is 200 to one.
(X_3)	TLOD	3.0	Optimal teaching loads are established at three classes per day or fifteen classes per week.
(X_4)	CSIZ	18.6	The average number of students per class.
(X_5)	AEE	20.3	The number of dollars spent in the school district per student for extracurricular activities.

(X_6)	TSALARY	173.40	**Scales** back to an average annual salary of $17,340 per teacher.
(X_7)	PSUP	16.5	Paraprofessional support per week, in hours.
(X_8)	CUG	18.6	The number of different subject matter courses available for student registration per grade.
(X_9)	PRCO	4.05	Number of different subject matter preparations per teacher per week.
(X_{10})	SFRAT	30.6	The ratio of students to academic (teaching and non-teaching) faculty.
(X_{11})	BOOKSP	9.7	The number of library books available for check out per pupil.
(X_{12})	TEXPER	3.2	Total years of teacher service in education.
(X_{13})	LIBRARY	1.0	Library accessibility index. Solution values may range from 1 (minimum accessibility) to 5 (maximum accessibility).
(X_{14})	CLPRACT	11.0	Teacher classroom practices. Solution values may range from 11 to 55.
(X_{15})	INNOVATE	59.8	School usage of twelve or more relatively new educational practices. Solution values may range from 11 to 60.
(X_{16})	BRAT	0.75	An index of crowding of physical plant.
(X_{17})	AMAN	1.0	Administrative man-hours per student. The solution value can range between 1.0 and 10.0 man-hours per student.
(X_{18})	AXMAN	1.0	Auxilliary man-hours per student. The solution values may range between 1.0 and 16.0.

integer solutions must be acceptable is readily apparent. For instance, the optimal level of teacher education is seen to be 6.3 academic years, which would provide certification at least at the level of Master's plus two years. Two quite interesting results are concerned with the BRAT variable and the AMAN variable. Since BRAT reflects the building occupancy ratio and a value of one indicates that actual occupancy equals state rated capacity, the solution value of .75 indicates that overcrowding of the physical plant should be avoided when possible. Building programs currently are emphasizing the modular and open classroom concepts, and thus are attempting to remove classroom crowding conditions. The AMAN and AXMAN variables reflecting the level of administrative man-hours per student and auxiliary man-hours per student, respectively, have a solution of 1.0. This result is interesting because it indicates that in the actual production of education outputs, the administrative and auxiliary support functions are rather secondary. Instead of purchasing more administrative and auxiliary services, these resources could possibly be allocated more effectively along other channels.

CONCLUSION

Probably the most immediate and obvious conclusion is that, properly specified, the goal programming approach to decision making within educational systems appears to be useful. Thus, the present approach is a step forward in the development of educational decision models.

No attempt was made here to determine the financial feasibility of securing the resource mix necessary for the level of goal attainment presented above. Once financial information is incorporated into the constraints, an even closer approximation of the real world can be made. The concern here has been, however, to determine the physical level of resources required to meet, as closely as possible, the school principal's priorities.

The sample size employed here is very small. Only one specification of the input-output model used for the technical constraints has been tried, and different specifications could yield different goal programming results. Since it has been demonstrated here that the methodology is operable and applicable to public education, we feel that efforts should be intensified to thoroughly define and specify the educational environment.[10]

Footnotes

[1]The term "exactly attain" reflects a situation where deviations in both directions are minimized, but does not guarantee that both deviational variables would be reduced to zero (at least one deviational variable, however, must be reduced to zero).

[2]For a discussion of the manipulable and non-manipulable variables, the estimated reduced form coefficients, and the contribution of the SEFAC variables to the system's output, see Morgan [5], pp. 135-137. See also Cohn and Millman [2], p. 63. In single equation educational production functions, the socio-economic factors generally do exert a very strong explanatory influence. However, in a simultaneous input-output system as developed by Cohn, the influence of socio-cultural and demographic factors has not been proven.

[3]See Morgan [4], pp. 145-146 for a discussion of the priority frequency matrix for the goals. The value of the implicit weights ($\sigma_i^{\pm}$) used in the objective function can be computed from the goal deviation frequency matrix. The computation takes the form: $\sigma_i^{\pm} = [\frac{1}{2}n^{o}+n^{\pm}/N]^{-1}$, where

$\sigma_i^{\pm}$ = the weight associated with both positive and negative deviations from the ith goal;
n^{o} = the frequency of responses where exact attainment was indicated for the ith goal;
$n^{\pm}$ = the frequency of responses indicating that over (d^{+}) or under (d^{-}) achievement would be desirable;
N = the total number of responses.

[4]For a discussion of the values for the variables excluded from the questionnaire, see Morgan [4], pp. 154-157.

[5]For a description of the initial output means and standard deviations, see Cohn [2], p. 58. The level of confidence is at least 89 percent and could be 99 percent if the normality assumption is appropriate.

[6]See Morgan [4], p. 136.

[7]See Morgan [4], p. 185 for a discussion of the values for $\hat{T}_i$; T_i^*; α_i; b_{nmi}; and $\bar{X}_{nmi}$.

[8]See Sen and Pattaniak [8] for a discussion of the value restricted social rankings.

[9]For a discussion of the compilation, significance, and implication of the value restricted preference ranking among school administrators, see Morgan, McMeekin, and Cohn [6].

[10]A more detailed analysis is contained in Morgan and Cohn [5], which will be made available upon request.

References

1. Beers, J.S. The Ten Goals of Quality Education: Rationale and Measurement. Harrisburg, PA: Pennsylvania Department of Education, 1970.

2. Cohn, E. with Millman, S.D. Input-Output Analysis in Public Education. Cambridge: Ballinger Publishing Co., 1975.

3. Lee, Sang M. Goal Programming for Decision Analysis. Philadelphia: Auerback Publishers, Inc., 1972.

4. Morgan, J.M. Goal Programming and Resource Allocation Within the Pennsylvania Secondary School System. Unpublished doctoral dissertation. Columbia: University of South Carolina, 1977.

5. Morgan, J.M. and Cohn, E. Goal Programming and Resource Allocation Within Educational Systems. Unpublished Working Paper, Western Kentucky University and the University of South Carolina, 1977.

6. Morgan, J.M., McMeekin, G.C., and Cohn, E. "Value Restricted Preferences and Educational Planning." Unpublished Working Paper, Department of Economics, Western Kentucky University, 1977.

7. Russel, N.F. Public School and Community Conditions: Definition and Measurement. Harrisburg, PA: Pennsylvania Department of Education, 1971.

8. Sen, A. and Pattaniak, P.K. "Necessary and Sufficient Conditions for Rational Choice Under Majority Decision." Journal of Economic Theory. I (1969): 178-202.

DISCUSSION

Abbott L. Ferriss, Emory University

1.

The Strenio-Bryk-Weisberg paper presents an individual growth model for evaluating programs, designed particularly for the Head Start program for children 4 to 6 years of age. The paper develops and assesses the traditional analysis of covariance approach, which compares change in treatment and nontreatment groups, and a value-added approach wherein the individual's growth is projected by regression, obtained from the initial cross-sectional data, and this projection of growth then is compared with the obtained growth at time t_2. The paper deals with four problems in the development of models for this problem: errors of measurement, the assignment of subjects to groups, the problem of individual growth, and the treatment effects. My comment concerns only one aspect of the problem of individual growth which the authors recognize in their paper.

The authors suggest that the assumption that "the cross-sectional data mirrors the longitudinal data may be wrong." This is critical to the use of the value-added procedure.

There are two aspects of this assumption that may not hold.

The Head Start program attempts to compensate for the variation in the early learning experiences of the child. Even among low-income families considerable variation exists in the attention and stimulation children receive, resulting in different developmental rates. The problem is whether a growth curve based upon such heterogeneity is an adequate basis for predicting the expected development of the individual child. It would seem desirable to seek some basis for controlling on prior learning environments and experiences.

A second problem is whether growth is linear with age. Studies by Gesell, Breckenridge and Vincent and others have shown that, while growth is continuous, it is not observably smooth and uniform over time in its many facets. "...what happens at one stage carries over into and influences the next and ensuing stages." All aspects of growth do not "develop at the same rate at the same time. . ."[1] Gesell singled out two-and-one-half years and three-and-one-half years for special consideration because they were particularly significant in the growth of the third and fourth years.[2]

The authors' future plans to develop individual growth curves by obtaining longitudinal observations on each child would appear to be a satisfactory approach to these problems.

With biological, social and cultural influences affecting the rate of growth of an individual, it is not surprising that a complicated design is required to tease out the effects of a Head Start program. The authors have approached this difficult problem on a sound basis.

2.

Professor Katz's analysis of the "sheepskin effect" also uses a regression technique. He tests the hypothesis that high school or college graduation (with the sheepskin) produces significantly higher income (and prestige) in later career than does the all-but-diploma earner. The technique predicts earnings (or prestige) by regression of earnings (or prestige) for the three years prior to the normal graduation year. If the predicted is less than the earnings actually obtained by those who attained the extra year of schooling, the difference is attributed to the "sheepskin effect." His analysis is by sex for three career groups, using educational attainment at both the 12th and 16th educational years to represent graduation.

It is a study of the marginal, incremental value of an additional year of schooling. The assumption that the sheepskin made the difference is questionable, because the data actually do not answer to the question, "Did you graduate?" Having attained 12 years or 16 years of schooling is not precisely synonomous with graduation. Indeed, in the past, some school systems have granted diplomas after 11 years of schooling. During World War II, a graduation date likely to affect high school graduation among Katz's group 15-29 years since leaving school, accelerated programs enabled early high school graduation, that is, with less than 12 years of schooling. During that period there were cases of college graduation at ages 18 or 19. Finally, the recent study of the High School class of 1972, while not falling within the time-frame of the Katz study, shows that only 75.4% of the graduating class were 18 years of age in spring of graduation year.

Suppose Katz had predicted the earnings for those with 11 years of schooling, or those with 15 years of schooling, upon the basis of the previous three years experience, would the results have demonstrated a "11th grade effect" or a "college junior effect"? In short, I would feel more confident of these results if the actual determination of graduation or non-graduation were the basis for the classification.

For the college-level data for women the results were contrary to the hypothesized result for the early and middle career women. I suggest that the reason for this inconsistent result is that the basis for classifying career level for women is less reliable than for men, since women typically have less continuous work histories than men, the years since leaving school containing fewer working years among women than men.

Small increments in education may make larger differences in earnings early in one's career but the advantage of the sheepskin might be expected to decrease as additional years of experience become more influential in determining earning power. Katz's data on earnings generally show an increasing value of the sheepskin effect with increasing experience, rather than less

effect. This is another of the "anomalous results" for Professor Katz to worry about.

3.

The Morgan-Cohn paper presents a model for allocating resources within secondary schools that uses specifically defined goals. They give an overview of a much more extensive Pennsylvania study. My comment concerns only one small aspect of their work.

Morgan and Cohn reduced 14 socio-economic and demographic in-put variables to four socio-economic factors and discovered, after regression equations were computed, that these non-manipulable variables "exert a minimal contribution to the predicted outputs of the" school system. The measurable outputs are the goals of quality education, listed in Table 1 of their paper. The authors do not describe in this paper their method of measuring these characteristics, but this result is contrary to many studies. Verbal and math skills, commonly measured with some uniformity in different studies, are found to be highly associated with demographic and socio-economic factors, e.g., sex, socio-economic status of the family as measured by income, education of head of household, and occupation of head of household. That the Morgan-Cohn study did not find verbal and math skills to be associated with socio-economic factors requires further exploration or explanation. In a recently reported follow-up Longitudinal Study of the High School Class of 1972, socio-economic status is associated with each of the items entering the measurement of self-concept.[3] However, other studies have found a low association between self esteem and SES among low income families, but the relationship usually is found when SES covers a wide range.[4] Could their sample of schools have come from a strata with low SES variance?

References

1. Breckenridge, Marian E. and E. Lee Vincent, "What are Some of the Laws Which Govern Growth?" in Morris L. Haimowitz and Matalie Reader Haimowitz (eds.) Human Development: Selected Readings, (3rd edition), New York: Thomas Y. Crowell Company, 1973, pp. 116-117.

2. Gesell, Arnold, Frances L. ilg, Louis Bates Ames, and Janet Learned Rodell, Infant and Child in the Culture of Today, (Rev. Ed.), New York: Harper & Row, Pub., 1974, p. 19.

3. Jay Levinson, Louis Lewis, John A. Riccobono, R. Paul Moore, National Longitudinal Study of the High School Class of 1972: Base Year, First and Second Follow-up Data File Users Manual, Research Triangle Park, N.C.: Center for Educational Research and Evaluation, 1976.

4. Morris Rosenberg and Roberta G. Simmons, Black and White Self-Esteem: the Urban School Child, (Arnold and Caroline Rose Monograph Series) Washington: American Sociological Society, 1971, p. 71.

COMPARATIVE EFFECTS OF COAL AND NUCLEAR FUEL ON MORTALITY

S. C. Morris, Brookhaven National Laboratory

While information upon which to base risk assessment is often scanty, assembling the data available, organizing them in a way to facilitate choices among energy policy options, including evaluation of uncertainties, is a useful aid to decision-making. Since decisions usually involve choosing among different technologies, standardized comparisons are essential to avoid misleading results. For technologies producing the same form of energy (e.g., electricity) a standardized unit of production can be used, for instance, a 1000 mWe power-plant year. When comparisons must stretch across technologies producing different energy forms, (e.g., coal electric versus coal gasification versus coal liquefaction) the proper basis of comparison is not always obvious. Indeed, there may not be a totally satisfactory basis. Streams of electricity, gas, and oil with the same energy content are not really equal; they are used by the consumer for different purposes and with different efficiencies. This difficulty can largely be overcome by examining the impacts of complete energy systems made up of different technological mixes. Risk assessment must attribute risk to each component of the energy system. Valid comparisons can be made only between entire fuel cycles or between alternative energy systems. While we are not yet able to completely analyze environmental and health impacts from quantitative data for the entire energy system, current economic- and technology-oriented models use this integrated framework.[1]

A key part of risk assessment is estimation of population exposure. This might ideally be a compilation of the number and characteristics of people exposed to given kinds, levels, and combinations of risk. The compilation ideally would be sufficiently disaggregated to allow calculation of the joint frequency of various combinations of risk to which a single population might be exposed. The number of people exposed at each level of risk is important since the true health damage function (or dose-response function) is likely to be nonlinear. Joint frequencies of risk are important since combined exposures from multiple agents may have synergistic effects. Information on pertinent population characteristics would allow differences in susceptability within the total population to be considered.

This ideal compilation would be very complex. Knelson[2] has suggested the framework of such a compilation which remains complex although synergisms among pollutants were not considered. Even were we to establish such a framework for analysis, however, current knowledge of dose-response relationships is insufficient to calculate effects of specific mixes of exposure levels to specific population subgroups except in rare situations. Available data are inadequate, for example, to adequately allocate the observed effect of air pollution to specific pollutants.

At this point, in our models, we are not considering synergisms. We attempt to define the population exposed and the degree of exposure, but treat the population at each exposure level as a single class. We also use linear damage functions. While these probably do not adequately represent the true effects over a wide range of exposure, we believe they are adequate to predict the effect of small changes of exposure within the general range of previous observation. Moreover, in our air pollution models we are generally allocating part of the total effect of air pollution to a specific source. A linear model seems completely appropriate for this use.

We measure mortality in "excess" or attributable deaths per power-plant year. "Excess deaths" is a convenient way to express changes in mortality rates. Although one expects only one death per lifetime, there can be more than the expected number of deaths in a population during a given time period. The time period we take is a year. Thus, an excess death represents at least one person-year of life lost, although for the most part we have only poor estimates of how much more than a year has been lost. If 130 coal miners are killed in accidents in the process of mining 300 million tons of coal, then there are 130/300 = 0.4 deaths per million tons of coal mined that would not have occurred had the coal not been mined. We can then apportion the attributable deaths based on the annual coal consumption of a power plant. In a strict sense this is not quite correct since there are competing risks. The miners would face other risks were they not in the mines. In this case the correction would not seem to be a major one and the years lost might be approximated by the expected remaining lifetime of the rest of the population in the age group. Other classes of effects are not as simple. Excess deaths due to air pollution are derived from linear regression models relating mortality rates in Standard Metropolitan Statistical Areas with air pollution and socio-economic variables.[3]

We are not very happy with excess deaths as a measure of health impact. It

withholds much important information about the impact being measured. One cannot distinguish among an accidental death, a heart attack and a cancer, or among the death of a child, a young adult or a senior citizen. It is not exactly clear how one should weigh these factors or what other factors should be included, but we would like to include more information of this type. There is a problem in the data as well as in the conceptual formulation. In many instances we do not have sufficient knowledge to estimate years of life lost per death very well for example.

The difficulty in using excess deaths as our measure is compounded by the confusion over the goal at which we are aiming. It has become general practice to total up the number of deaths that can be attributed to nuclear power or coal or to auto accidents, smoking, etc. Since the analysis is done to affect decisions, the implicit notion is that we should act to reduce the total numbers of deaths. I believe a major philosophical question arises over whether one should treat well-defined deaths such as accidental fatalities among coal miners the same as deaths that can only be calculated by extrapolation, such as deaths in the general population caused by air pollution or radiation exposure. To some degree, this can be handled by taking the level of uncertainty associated with the estimated number of deaths into account. The level of individual risk can have importance as well as the total number of deaths. A high risk of accidental death among a few coal miners may be perceived differently than an infinitesimally small additional risk assumed by a large population, even though the absolute number of annual excess deaths may be the same. In some cases this may be a function of the state of knowledge. Coal miners are a well-defined group and the 100 or so that die annually in mine accidents are easily counted and attributable to coal mining. There may be a group within the general population exposed to air pollution from coal combustion that has a particular, but undetected, constitutional susceptability to air pollution health damage. People in that group may face an individual risk as high as a coal miner. Until such a population can be defined and its level of risk determined, however, we perveive the effects of air pollution as spread over the entire population at a very low individual risk level. There have been some attempts to derive damage functions for specific groups believed to be at high risk to air pollution, to determine the exposure level of these groups and calculate the impact in that manner.[4] One might also hypothesize, however, that everyone is affected at least to some degree.

Table 1 provides two measures. The total risk, in excess deaths per power-plant year, and the individual risk, in excess deaths per power-plant year per person. The latter might be taken as the increased probability that an individual in the exposed population will die in a manner attributable to the operation of a power plant or part of its supporting fuel cycle for a year. To the extent that attributable deaths from different causes in different populations are considered equal, the total effects are additive. The individual risks are additive only when the same population is involved in each case. In addition to a goal of decreasing total attributable deaths, disproportionately high levels of individual risk should indicate areas of concern.

The level of individual risk in Table 1 provides only a crude estimate of the range of effect on the individual. The number of people exposed to the risks of an activity, particularly among the public, and the distribution of exposures among that population varies greatly according to the location and the specific design of the facility. Average population density alone differs by more than two orders of magnitude between the Middle Atlantic and the West North Central regions of the country. Differences among individuals in the exposed population are not considered in the tabulation. These differences could include individual activity patterns that enhance exposure, concurrent exposure from other sources (e.g., smoking or occupational exposures), or individual variations in susceptability to a given environmental stress. Thus the individual risk levels given must be taken as merely crude guidelines subject to much more uncertainty and variability than the total effects.

Most of the kinds of health impacts quantified in Table 1 are either occupational effects that occur frequently enough in well-defined populations so that sufficient data are available from which to make reasonable estimates of risk or, particularly in the case of radiation exposures, exposure situations for which established methods of estimating health impacts are available.

Underground mining is a dangerous occupation as can clearly be seen from the risk levels for underground coal mining in Table 1. Underground and surface mining are combined for uranium mining in the table since the fuel from both are combined in the cycle well before the power plant. A coal-fired power plant, on the other hand, is usually served from one or from a well-identifiable group of mines. The major difference in total deaths between coal and uranium miners stems from the higher energy content of the nuclear fuel. It

requires only about one-tenth the man-days of effort in the mines to fuel a 1000 mWe nuclear plant compared to a similar coal plant. The wide range in estimates of disease-related deaths in coal miners stems from a wide range of disease rates among different coal mining regions, difficulty in attributing an appropriate share of observed disease deaths among miners to their occupation, and uncertainty in the efficacy of recently mandated improvements in the mines.

Transport accidents in the coal fuel cycle range from mine-mouth plants with essentially no transport to fairly long distance transport by rail involving the risk of railroad associated accidents. The largest share of these are train-auto collisions at grade crossings. Although the individual risk level in the table is calculated as if the entire population of the country were at risk, the true exposed population is probably limited to people living near major coal train routes. The individual risk might then be an order of magnitude or more higher. The routine impact of transporting nuclear fuel is very small relative to coal because of the much smaller mass of material to be handled.

An exception to the notion that the risk estimates are fairly well defined is the health impact of air pollution from coal combustion. We have spent considerable effort attempting to estimate this impact and to define the uncertainty associated with these estimates.[5,6] Our models are based on the currently held theory that the principle agents of health damage are sulfate compounds mainly resulting from SO_2 emitted from tall stacks undergoing chemical reaction in the atmosphere.[7] Uncertainties in both the toxicological and epidemiological studies linking sulfate compounds with health effects are such that the possibility of no effect is not foreclosed. The bulk of the evidence, however, suggests that there is an effect. Local estimates are based on stocastic models developed at Brookhaven by Morgan, et al.[6] They are based on typical 1000 mWe power plants with tall stacks in the western Pennsylvania area. (These plants have an average of 2-4 million people within the 80 km radius.) Emission rates have been adjusted to match current New Source Performance Standards. Although the range is from 0-24 excess deaths annually, the expected value is about 4. Due to limitations in the meteorological model, rather than any physical break-point, the exposed population is limited to an 80 km radius. Current work with long-range transport models being developed in Brookhaven's Atmospheric Sciences Division suggests that the effect on more remote populations may exceed the local effects by as much as an order of magnitude.[8]

Sulfates are not the only pollutant of health concern from coal power plants. Lundy has estimated the impact of polycyclic hydrocarbon emissions to be in the range of 0-4 excess deaths per power-plant year.[9] Various toxic trace metal emissions could be of concern, but probably have a much smaller direct impact than the sulfur and polycyclic hydrocarbon compounds.[10]

An additional impact which has considerable uncertainty and controversy associated with it is the possibility of major radiation releases associated with catastrophic events, particularly from nuclear power plants. These are not shown explicitly in the table, but the annual expected value of these highly unlikely events is so low that it does not significantly affect the totals. The major work in this area has been the Atomic Energy Commission sponsored Reactor Safety Study (Rasmussen study) which estimated the expected annualized loss of life from nuclear power plant accidents as 0.02.[11] One can argue that the population is strongly a risk avoider for very large accidents. One way to take this into account is to multiply the annualized impact by a weighting factor before comparing it with effects which happen routinely. A weighting factor of 100 (which seems very high) is needed to even put accidents into the range of routine effects.

It has been suggested that the Rasmussen estimates may be too low. Most suggestions are by a factor of 2 to 10. The recent report of the Nuclear Energy Policy Study Group states that "...the WASH-1400 estimate could be low by a factor of as much at 500."[12] With this estimate, fatalities due to nuclear fall within the range of estimated effects of coal, but a direct comparison is not a fair one. This was not put forth as a best estimate as the Rasmussen number was, but as an upper limit. It is based on the very pessimistic assumptions that (1) the probability of a core meltdown is 5×10^{-3} per reactor year, 100 times more likely than estimated by Rasmussen and high enough that were it the true value we have been quite lucky not to have had a core meltdown yet; (2) the probability of emergency core cooling system (ECCS) failure of 1.0; (3) probability of breach of containment of 0.2 (twice the Rasmussen estimate) and (4) three to four times the average fatalities predicted by Rasmussen given a major accident. The fact that an estimate very far out on the tail of the nuclear effects distribution intersects the coal effects distribution does not negate the clearly significant difference between the estimated health effects of the two energy forms.

There is a fair possibility that coal electric has a relatively much

greater impact on mortality than nuclear. The reverse does not seem to be true. Some degree of perspective is necessary, however. Neither coal nor nuclear has a very big impact on mortality relative to other factors. Were the high end of the coal range to prove correct, a large increase in coal fired electric power might bring the impact up to 5 to 7 percent of total mortality--a big effect. This could be reduced considerably by stricter controls on sulfur emissions, a step already being considered by the Environmental Protection Agency. It is more likely that the effects are considerably lower, around 1 percent of total mortality attributable to coal and nuclear power. This must be compared to 2 to 3 percent from automobile accidents and 17 percent attributable to smoking. It is my personal conclusion that while we must continue to do our best to reduce the total effects from both coal and nuclear electric generation, the primary emphasis should be placed on areas such as coal and uranium mining where the highest individual levels of risk are faced.

Acknowledgement

Work supported by Division of Technology Overview, AES/ERDA. I have drawn heavily on work by my colleagues at Brookhaven, particulary G. Morgan, S. Finch, K. Novak, R. Meyers, R. Cederwall, and L. D. Hamilton. I am particularly grateful for discussions of drafts of this paper by S. Finch, M. Rowe, and L. D. Hamilton.

References

1. M. Beller, ed., Sourcebook for Energy Assessment, Brookhaven National Laboratory, Upton, New York, 1975 (BNL No. 50483).

2. J. H. Knelson, Designing the Exposure/Response Matrix in Environmental Health Studies, Proceedings, Internation Symposium on Recent Advances in the Assessment of Health Effects of Environmental Pollution (EUR 5360), Commission of the European Communities, Luxembourg, 1975, pp. 181-88.

3. L. B. Lave and E. P. Seskin, An Analysis of the Association Between U.S. Mortality and Air Pollution, Journal American Statistical Association 68, pp. 284-90, 1973.

4. J. F. Finklea, et al, The Role of Environmental Health Assessment in the Control of Air Pollution, in Advances in Environmental Science and Technology, J. N. Pitts and R. L. Metcalf, eds., vol. 7, John Wiley and Sons, New York.

5. S. J. Finch and S. C. Morris, Consistency of Reported Health Effects of Air Pollution, in Advances in Environmental Science and Engineering, in press, J. R. Pfafflin and E. N. Ziegler, eds.

6. M. Granger Morgan, S. C. Morris, A. K. Meier, D. L. Shenk, A Probabilistic Methodology for Estimating Air Pollution Health Effects From Coal-Fired Power Plants, Energy Systems and Policy, in press.

7. U.S. Environmental Protection Agency, Position Paper on Regulation of Atmospheric Sulfates (EPA-450/2-75-007), Research Triangle Park, 1975.

8. R. Meyers and R. Cederwall, Brookhaven National Laboratory, work in progress.

9. R. T. Lundy and D. Grahn, Predictions of the Effects of Energy Production on Human Health, American Statistical Association annual meeting, Chicago, August 1977.

10. M. R. Baser and S. C. Morris, Assessment of the Potential Role of Trace Metal Health Effects in Limiting the Use of Coal Fired Electric Power, draft informal report, Brookhaven National Laboratory, Upton, New York, 1977.

11. U.S. Nuclear Regulatory Commission, Reactor Safety Study (WASH-1400), Washington, D.C., 1975.

12. Nuclear Energy Policy Study Group, Nuclear Power Issues and Choices, Ballinger Publishing Co., Cambridge, Massachusetts, 1977.

Table 1

MORTALITY RISKS IN COAL AND NUCLEAR FUEL CYCLES
(TOTAL RISKS ARE PER 1000 mWe PLANT YEAR)

ACTIVITY	COAL		NUCLEAR	
	INDIVIDUAL RISK	TOTAL RISK	INDIVIDUAL RISK	TOTAL RISK
Mining				
Underground - Accident	1×10^{-3}	0.5 - 1.1		
Surface - Accident	7×10^{-4}	0.2	3×10^{-3}	0.09 - 0.2
Underground - Disease	4×10^{-4} - 3×10^{-2}	0.2 - 13	8×10^{-4}	0.04
Processing				
Occupational Accidents	6×10^{-4}	0.04	~1×10^{-5}	0.004
Occupational Disease	-	-	0-1 $\times 10^{-4}$	0 - 0.03
Transport				
Accidents	0-4 $\times 10^{-8}$	0 - 4	~10^{-10}	0.01
Electric Generation				
Occupational Accidents	1×10^{-4}	0.01	1×10^{-4}	0.01
Local (80 km) Disease	0-8 $\times 10^{-6}$	0 - 24	0-3 $\times 10^{-9}$	0-6 $\times 10^{-2}$
Global Disease	(0-3 $\times 10^{-6}$)	(0 - 240)	0-2 $\times 10^{-11}$	0 - 0.1
Waste Management				
Disease	-	-	-	0 - 0.04
Totals		0.3 - 300		0.1 - 0.5

URBAN AIR POLLUTION AND MORTALITY: TEN YEARS OF PHILADELPHIA DATA

Ronald E. Wyzga,* Electric Power Research Institute

I. INTRODUCTION

In spite of the large amount of research that has been undertaken to learn about the relationship between human health and air pollution, we still remain relatively ignorant. We cannot enumerate all of the health impacts; we are not at all certain about the identity of those pollutants, singly or in combination, which may be responsible for health effects; and we are ignorant about the dose-response relationship between these two elements. Further information is badly needed about this relationship to guide us in the development of future energy technologies.

All believable energy scenarios for the U. S. indicate an important role for coal. Uncontrolled coal combustion produces significant air pollution. To date, we have regulated the combustion of coal and adopted technologies to reduce the emissions of particulates and SO_2. There are those who question whether present control technologies are sufficient or whether or not we should concentrate our pollution control efforts on other substances. This question is also posed by those developing new technologies in which some trade-offs may be necessary. For example, there are new technologies which could allow us to reduce our NO_x emissions further, but at the expense of increased polycyclic organic emissions. There is also concern that some currently suggested methods of SO_2 control in coal-fired power plants could lead to the increased formation of sulfate and other oxidized sulfur compounds.

This paper describes a model used to estimate the association between air pollution and health as measured by mortality and then tries to identify those pollutants which are more closely associated with mortality.

II. THE DATA AND VARIABLES

This study uses Philadelphia data for the years 1957-1966. Daily mortality data by cause of death are available for those residents of Philadelphia who died in that city. Two daily pollution measures were available for the ten-year period: coefficients of haze (smoke shade) and total suspended particulate (HI-VOL) measures. These two measurements were taken at two or three sampling stations in Philadelphia. Three mutually exclusive time periods (1957-1960, 1961-1963, and 1964-1966) are defined to accommodate changes in sampling sites over the ten-year period. Data from the same stations are then generally available for each day of a particular time period. This partition into three time periods allows three replications of each subsequent model examined and aids in the model and variable development. The first and third time periods of the study use data from two sampling stations. Coefficients of haze (COH) measurements and total suspended particulate (TSP) data from the two stations are generally available for each day of these time periods. Those days for which one or more station measures are missing are excluded from this study. For the second study period, data are available from three sampling stations. When daily measures were missing from one station, they were estimated through use of an iterative regression procedure (1). The COH variable for such a day would then be the average of the observed COH values and the estimated value. When measurements were missing for two or more stations on a given day in the second time period, that day was eliminated from the investigation.

For the third time period, measures of six additional pollutants are available from one monitoring station in Philadelphia. These pollutants are NO, NO_2, SO_2, hydrocarbon, CO and oxidants. The means and standard deviations of the pollution variables for the winter months of the 1964-1966 period are given in Table 1. Table 2 gives the estimated correlation coefficients for each pair of pollution variables during that period.

Several seasonality variables were compared and a weighted 30-day moving average of past temperatures, which gave twice the weight to the most recent 15 days, was chosen because it correlated more highly with total mortality than several other moving-averages of temperature and because it was significantly more highly correlated with mortality than Fourier functions of time.

The performance of the seasonal adjustment variable for the winter months differed significantly from its performance for the summer months. Accordingly, the year is divided into halves in subsequent analyses. An epidemic variable for the winter months and a heat-wave variable for the summer months were also found to be important contributors to the variation in mortality data. These variables are included in the following analyses. The epidemic variable is defined from the residuals obtained from regressing monthly New Jersey mortality data appropriately detrended on current and preceding Philadelphia temperature averages.

The heat-wave variable is the weighted product of lagged and unlagged values from a one-to-six corrected effective temperature scale. If the corrected effective temperature scale value for the day of mortality is represented by $E(D)$, then the variable used can be written $E(D)^3 E(D-1)^2 E(D-2)$. For each season, a two-day moving-average of temperature which represents recent weather is also added to the set of variables included in the analysis.

The means and standard deviations of each variable analyzed for all time periods are given in Table 3.

III. DEVELOPMENT OF A MODEL

Regression models are considered. Total mortality was first regressed upon the group of adjustment variables and COH and TSP for the summers and winters of each of the three time periods. Given the high correlation between TSP and COH, it was felt that any further model development would best consider only one of the two variables, and since the regression coefficients of COH were associated with larger t-statistics than those of TSP, the COH variable was used in further model development.

Linear and non-linear models were considered, and linear models performed better than non-linear models and were therefore considered in the subsequent analyses. The residuals of the linear regression appeared to be normally distributed although they were serially correlated in some

time periods.

IV. RESULTS

The results of the regressions on total mortality, using the COH (coefficient of haze) measure as a pollution variable, are given in Table 4. The COH coefficients are all positive, and those for all of the winter periods are for the 1957-1960 summer period are significant. The mean pollution levels for the two summer periods in which the COH coefficients are not significant are noticeably smaller than the pollution variables for the other time periods. This fact might explain the non-significance of the COH coefficients for these time periods.

There could be two reasons for detecting a weaker relationship between the COH values and mortality for those periods with smaller COH values. First of all, the response of mortality could be non-linear with a proportionately stronger response to higher pollution levels than a linear model suggests in spite of the fact that a linear model performed better than non-linear models tested. Functions using the COH values only above a certain threshold and exponential functions of the COH values were introduced into the regressions, but they gave no higher association with mortality than the initial COH variables. The second reason could explain a smaller association between the COH measure and mortality when the COH measures are small, even if the relationship were linear. The measurement error of smaller COH values is far greater relative to their size than the measurement error of the larger COH values. As measurement error would bias the regression coefficients of the COH variable downward (2,3), the coefficients of smaller COH variables would be subject to a greater bias than the coefficient of larger COH variables.

The significant Durbin-Watson statistics indicate the presence of serial correlation, which can lead to overestimates of the (absolute values of the) t-values used to test the significance of the coefficients (2,3). To adjust for this problem, a non-linear regression model incorporating serial correlation was fitted. The results showed no changes in the significance levels for any of the COH coefficients.

The beta coefficients (β coeff.) presented in Table 4 indicate the predicted number of standard deviations the mortality variable will change for each increase of one standard deviation in that variable, if one assumes that the other variables remain constant. Thus if the linear regression model is correct for the 1964-1966 winter data, the estimates predict that an increase of one standard deviation in the COH variable will lead to an increase of mortality on that day of 0.1349 times the standard deviation of total mortality (9.22) or to an increase of about one death.

The results indicate how important the epidemic and "heat wave" variables are in explaining daily mortality. The 2-day temperature variable in the summer months is also an important predictor of mortality, and it probably complements the "heat wave variable" as an index of hot weather.

Data for the other pollutants (NO, NO_2, SO_2, hydrocarbons, CO, and oxidant) were available only for the 1964-66 time period. Given the lack of significance of the COH variable coefficient in the 1964-66 summer period, the data for this period were not analyzed with the additional pollutants. The winter 1964-66 data were analyzed using a series of regression models similar to that described above, but with a different pollution variable in each regression. The series of regressions permitted a comparison of regression coefficients and avoided a multicollinearity problem which would have arisen given the degree of correlation between several pairs of pollutants. (See Table 2). Serial correlation was not statistically significant, and no adjustment was undertaken.

Table 5 summarizes the results of regressing the various pollution variables upon daily total mortality. The seasonality variable, two-day temperature variable and epidemic variable were also included as independent variables in these regressions. All of the pollution variables have positive coefficients, but only the COH, NO and hydrocarbon variables have significantly positive coefficients.

V. DISCUSSION

From the results it is difficult to generalize about which pollutant is best as an index, or which may affect health most. Differences in the estimated coefficients could be due to differences in measurement error among the pollutants. The greater the random measurement error of a variable, the larger the downward bias in the coefficient of that variable (2,3). As different measurement methods are involved in measuring the various pollutants, the measurement errors cannot be expected to be the same for each variable. Another source of error leading to the same type of downward bias is the local influence upon a variable. Local influence is the influence of nearby sources upon a pollution measure. These local influences can be thought to be a kind of measurement error imposed upon an overall urban index. The variables other than COH and TSP are particularly susceptible to this type of error, as measures from only one station are available.

The estimated increase in the number of deaths associated with an increase of one standard deviation in the pollution variable ranges from 1.24 when COH or NO are the pollutants in the regression to 0.25 when oxidant is the pollutant examined. These estimates only consider deaths on the day of pollution; lagged or delayed effects are not included here.

A model to examine lagged effects was developed (4) using the COH variable. The large number of missing observations for the other pollution variables made it difficult to apply lagged models with these variables. The model developed was a geometrically-distributed lag model which adjusted for serial correlation. This model was applied to the four time periods in which the COH variable was statistically significant and yielded similar estimates of the COH impact for each period. Table 6 presents the results of this model for the 1964-66 winter period. The total increase in the estimated impact of the pollution variable on mortality is about one third, with almost no impact of pollution occurring beyond two days after the pollution occurred.

Chronic or greatly delayed effects cannot be estimated with time series models of daily data.

VI. SUMMARY AND CONCLUSIONS

Environmental air pollution is associated with increased mortality. Although this association is significant, the other environmental phenomena, such as heat waves, may be responsible for a larger number of deaths.

The use of different pollution variables was investigated. One would expect the different pollution measures to perform quite similarly as meteorological conditions largely determine the concentrations of pollutants in the atmosphere. All of the pollutants were positively associated with mortality, but only variables derived from COH, NO and hydrocarbon measurements were significantly associated with mortality. Until further information is obtained about the effects of measurement error and local influence upon the various pollution measures, it is impossible to associate mortality more closely with one type of pollution than with another. It should also be added that it will be necessary to consider additional pollutants or combinations of pollutants. Certainly one hears the names of additional emitted compounds as one investigates new and existing energy technologies.

REFERENCES

1. Wyzga, R. E. (1973): "Note on a Method to Estimate Missing Air Pollution Data in A Statistical Analysis of the Relationship Between Daily Mortality and Air Pollution Levels". Journal of the Air Pollution Control Association, 23(3), 207-208.

2. Johnston, J. (1963): Econometric Methods. McGraw-Hill, New York.

3. Malinvaud, E. (1966): Statistical Methods of Econometrics. Rand-McNally and Co., Chicago.

4. Wyzga, R. E.: The Effect of Air Pollution upon Mortality: A Consideration of Distributed Lag Models. Submitted for publication.

TABLE 1

Means and standard deviations of pollution variables, 1964-1966

Variable			Winter Periods
COH[a]	(per 1000 ft.)	Mean	131.00
		S.D.	55.14
TSP	($\mu g/m^3$)	Mean	161.59
		S.D.	66.95
NO	(parts per hundred million)	Mean	6.36
		S.D.	5.59
NO_2	(parts per hundred million)	Mean	3.41
		S.D.	1.26
SO_2	(parts per hundred million)	Mean	9.57
		S.D.	6.90
Hydrocarbon	(parts per ten million)	Mean	22.65
		S.D.	7.02
CO	(parts per million)	Mean	7.54
		S.D.	3.11
Oxidant	(parts per hundred million)	Mean	2.02
		S.D.	1.24

[a]The COH variable has been multiplied by 100.

TABLE 2

Correlations between pollution variables, 1964-1966 winters

Variable	COH	TSP	NO	NO_2	SO_2	HC	CO	OX
COH	1.000	.795	.811	.600	.667	.688	.329	.403
TSP		1.000	.657	.629	.654	.570	.306	.302
NO			1.000	.596	.520	.625	.325	.435
NO_2				1.000	.544	.508	.203	.378
SO_2					1.000	.562	.074	.163
HC (Hydrocarbon)						1.000	.147	.458
CO							1.000	.233
OX (Oxidant)								1.000

TABLE 3

Means and standard deviations of variables

		Winters			Summers		
Variable		1957-60	1961-63	1964-66	1957-60	1961-63	1964-66
Total daily mortality	Mean	66.90	66.89	64.62	58.55	58.85	60.36
	S.D.	10.19	10.36	9.22	10.74	10.21	10.27
2-day moving-average temperature	Mean	90.16	87.18	89.56	149.79	149.66	151.15
	S.D.	25.94	27.25	22.32	19.51	18.81	19.98
30-day moving-average temperature	Mean	1008.49	978.03	1008.63	1692.83	1683.30	1680.62
	S.D.	200.15	241.08	181.40	148.81	148.33	191.23
COH variable[a]	Mean	189.42	160.84	131.00	122.02	92.39	87.13
	S.D.	76.28	69.23	55.14	44.11	37.62	41.33
Epidemic variable	Mean	21.09	19.95	19.20			
	S.D.	7.08	7.29	6.22			
Effective temperature function	Mean				502.75	531.33	533.47
	S.D.				1676.49	1691.38	1467.94

[a]The COH variable have been multiplied by 100

TABLE 4

Comparison of results from multiple regressions - total mortality

Variable		1957-1960	1961-1963	1964-1966
A. Winter Periods				
COH variable	Coeff	0.0098	0.0126	0.0226
	β Coeff.	0.0735	0.0840	0.1349
	t-value	2.00*	2.02*	2.85**
Seasonality variable	β Coeff.	0.3077	-0.4292	-0.2812
	t-value	-5.72**	-7.09**	-4.44**
2-day temperature variable	β Coeff.	-0.0165	0.0626	0.0742
	t-value	-0.31	1.12	1.18
Epidemic variable	β Coeff.	0.2000	0.3468	0.0820
	t-value	5.57**	9.40**	1.75
Multiple correlation coefficient squared (R^2)		0.1550	0.2879	0.0926
Durbin-Watson statistic		1.7622**	1.7652**	1.8809
Number of observations		660	532	421
B. Summer Periods				
COH variable	Coeff.	0.0286	0.0199	0.0052
	β Coeff.	0.1174	0.0734	0.0208
	t-value	3.53**	1.82	0.41
Seasonality variable	β Coeff.	0.4663	0.2407	0.4426
	t-value	-9.89**	-4.48**	-6.53**
2-day temperature variable	β Coeff.	0.2048	0.0794	0.2273
	t-value	4.11**	1.42	3.06**
Heat wave variable	β Coeff.	0.4269	0.4671	0.2730
	t-value	12.15**	10.83**	5.01**
Multiple correlation coefficient squared (R^2)		0.3194	0.2360	0.1522
Durbin-Watson statistic		1.6503**	1.7597**	1.3458**
Number of observations		688	540	386

*Significant at the 0.05 level.
**Significant at the 0.01 level.

TABLE 5

Comparison of results from multiple regressions, 1964-1966, total mortality upon various pollutants

Winter 1964-1966

Pollution Variable	β coefficient	t-value
COH	0.1349	2.85**
TSP	0.0808	1.85
NO	0.1347	3.11**
NO_2	0.0565	1.28
SO_2	0.0443	0.94
Hydrocarbon	0.0999	2.17*
CO	0.0620	1.35
Oxidant	0.0269	0.53

*Significant at the 0.05 level.
**Significant at the 0.01 level.

TABLE 6

Geometrically decreasing lag model with serial correlation 1964-1966 Winter Data

Parameter	Estimate	Std. Error of Estimate	t-value
30-day seasonality variable parameter	-0.0166	0.0038	-4.30**
2-day temperature variable coefficient	0.0349	0.0296	1.18
Epidemic variable coeff.	0.1637	0.0848	1.93
COH variable coeff. b	0.0204	0.0072	2.83**
Lag parameter λ	0.3251	0.0847	3.84**
Serial correlation ρ	-0.2671	0.0872	-3.06**
Total effect $b/(1-\lambda)$	0.0302	0.0106	2.85**

Regression Constant: 60.78

Source	Sum of Squares	Degrees of Freedom	Mean Square
Regression	4258.102	6	709.684
Residual	30035.984	390	77.015
TOTAL	34294.086	396	86.601

Multiple Correlation Coefficient R: 0.3524
R^2: 0.1242

* Significant at the 0.05 level.
** Significant at the 0.01 level.

MORTALITY TRENDS IN COUNTIES SURROUNDING THE OAK RIDGE NUCLEAR FACILITIES

Clifford H. Patrick, U.S.ERDA

With the anticipated growth of nuclear facilities in the coming decade, it is imperative that the public health effects of nuclear power-plant operations be ascertained. In this study, changes in selected measures of ill health in the population surrounding the Oak Ridge plants are compared to changes in Tennessee as a whole for the period from 1929, prior to Oak Ridge's existence, through 1971. Tennessee is used as a control population against which to measure changes induced by strictly localized factors, such as the nuclear facilities, as opposed to statewide or national epidemics or trends. Because of the myriad potential causes of the measured effects and the paucity of actual measurements for these competing factors in the general public, a quantification of a dose term is not included in this analysis.

Potential Health Effects of Low-Level Exposure

The somatic and genetic effects associated with radiation exposure are briefly enumerated herein to indicate the types of public health changes which might be induced by increased radiation exposure in the population at risk (6, 16). The possible somatic effects of low-level radiation include cancers which have relatively long latent periods. The specific cancers most often cited in relation to radiation exposure are leukemia, thyroid, bone, breast, lung, and gastrointestinal tract. The noncarcinogenic diseases associated with radiation (based on studies at high levels of exposure) include cataracts, central nervous system disorders, premature aging ("life shortening"), fertility impairment, congenital malformation, and increased incidences of cardiovascular-renal diseases. The possible genetic effects of radiation exposure may be seen in the population as increased rates of spontaneous abortion or fetal wastage, neonatal and infant mortality, infertility, and congenital malformations, including rare syndromes such as Mongolism. In this study, four measures relating to possible radiation effects (cancer, infant mortality, congenital malformations, and fetal deaths) are examined.

Past Studies of Public Health and Radiation

The potential health effects of Oak Ridge's nuclear operations have been examined in three previous studies. One study attempted to determine if there is a relationship between cancer morbidity in the public and potential radiation exposure (5), while two later studies examined the relative mortality of Oak Ridge nuclear facilities' employees (2, 12).

The study by Moshman and Holland only examined the Oak Ridge resident population for a single year, 1948; the incidence of cancer morbidity in the Oak Ridge population was compared to expected rates to determine if Oak Ridge residents were more susceptible to cancer than the U.S. population. Computed age-adjusted cancer incidence in Oak Ridge was only 123 per 100,000 compared to the national average of 230, reflecting the highly selected Oak Ridge population.

Incidence rates of cancer for both males and females in Oak Ridge were lower than the national norms. On a relative basis, the proportions by primary sites of cancer occurrence in white females in Oak Ridge were not significantly different from nationally based expected values. Only one significant difference was found in males; a higher proportion of respiratory cancer was found in white males than would have been expected, using 1938 cancer data to compute expected values. The authors felt the continuous upward trend in respiratory system cancer among males since 1938 would account for this higher incidence in 1948 in Oak Ridge. The study was rather limited because it covered only one year and used prewar bases for computing expected values. However, it was perhaps the first study to test the hypothesis that the nuclear facilities at Oak Ridge might be a potential source of ill health.

The 1966 study by Larson et al. compared the number of actual deaths in the three Oak Ridge nuclear plants from 1950 through 1965 with the number expected by applying 1962 U.S. age-specific mortality rates to the age distribution of workers. Based on 207,204 man-years of employment, 692 deaths occurred compared to the 992 expected using the 1962 U.S. rates. Thus, workers exposed to the environment of the Oak Ridge facilities appear to live longer than their cohorts in the general population.

Such a result seems to indicate a low dose of radiation exposure is healthy, but such an interpretation of the results may be erroneous. The result only shows the workers to be less likely to die at a given age than the control population (in this case, the 1962 U.S. population of the same age distribution). This control population includes the disabled and institutionalized segments who are in a much lower state of health than any normal work force, and especially workers at the Oak Ridge facilities who have on-site medical care and periodic plant physicals. While it is valid to conclude that these workers have better health than the control group, further analysis is required to test whether potential exposure to low-level radiation is related to the better state of health.

Such an analysis, based on age-adjusted data, has been attempted by Scott et al. Workers from two Oak Ridge facilities involved with uranium processing were separated into two groups based on their work areas at the plants. The uranium workers were predominantly technicians and craftsmen, while the nonuranium workers covered a broader spectrum of job classifications. The study covered employees from 1951 through 1969 and applied the 1960 U.S. mortality tables to

each of the two distributions to determine expected deaths in each group. As in the Larson study, one would have expected to find the actual number of deaths to be less than for the U.S. average; but the critical question, which the earlier study did not consider, is whether the uranium workers are relatively more healthy than the nonuranium workers.

Scott et al. found the uranium workers had a mortality experience 59% as high as the general population, while the nonuranium workers had a mortality rate 76% as high. Thus the uranium workers appear relatively less subject to the risk of dying at a given age than the nonuranium workers. This result could be even more significant, because the average age of the uranium workers was about five years greater than that of the nonuranium workers, potentially giving radiation workers a longer period of exposure.

Though these studies uncovered no adverse health effects, the evidence is not overwhelming and indicates the need for in-depth epidemiological studies. While research of this type still appears to be in its infancy, research in the areas of occupational and medical exposures and by the Radiation Effects Research Foundation (RERF) suggests a well-trod path to follow (6, 16).

Methodology for Examining Mortality Trends in the Public

Methodology is important, particularly in studies of public health, because of the paucity of both reliable exposure data and knowledge of dose-response at low level chronic exposures (3). Data available for examining health effects in the public include time series of vital statistics for both the local area in which the facility is located and a comparable nonimpacted area to act as a control. Included under the rubric of vital statistics are data on population size, births, deaths, illness, and migration. These data ideally should be categorized by demographic variables such as age, race, sex, and socioeconomic characteristics. Vital statistics data are generally published annually by each state for counties and larger cities in a Vital Statistics series, usually by race, but seldom by other traits (15). Annual vital statistics and related data also are available from the National Center of Health Statistics of HEW and the Bureau of the Census (7, 17). Vital Statistics usually contain very limited information on morbidity, but seldom contain migration data. The Census Bureau, however, publishes estimates of population change and migration; some morbidity data is available through the U.S. Public Health Service (8, 10, 11).

In this study, time series data from 1929 to 1971 for four types of mortality and for total population in each area examined are taken from Tennessee Vital Statistics for the given year. Stillbirths (fetal deaths), infant mortality, cancer deaths, and congenital malformations were chosen because they are representative of effects of radiation found in the literature.

Age, sex, and race breakdowns of the data are preferred because of the differential mortality among demographic groups but were not available in necessary detail in the published sources provided by the State (1).

Given the data in hand, separated by race for all years except 1959 and 1970, we restricted the analysis to the white population for two reasons: (1) The nonwhite population is quite small, generally younger and subject to much larger errors in reporting than the white population, especially prior to the 1950's; and (2) Rates of age-specific mortality by cause are probably based on more reliable data for whites. An effect in the largest, most statistically reliable group should be present in other demographic groups unless a race-specific selection of radiation-induced ill health exists.

The absence of age structure for the local populations is a severe shortcoming, because differences in age structure should be taken into account in comparing death rates for different areas. Older populations tend to have higher cancer rates, while younger populations tend to have greater incidences of death due to congenital malformation, fetal deaths, and infant deaths. Although these age factors may be offsetting, we do not know this for sure. For the purposes of this analysis, we are assuming (1) the age effects are sufficiently reflected using alternative measures with old age and young age biases and (2) the statistics will show offsetting trends where age structure per se creates an effect. The age structure problem may be more acute for analysis of cancer trends than for analysis of natality related measures. The lack of age structure is probably the largest drawback to the use of this annual data, because neither direct nor indirect standardization can be applied to develop measures unless decennial Census figures are used as estimators.

Given these limitations, we examine the yearly statistics for population, deaths, and death rates for both the local plant area and the control area--here the State (national statistics are often used as controls). The death rates for smaller areas almost always appear much more variable than the rates for larger areas because of the smaller number of deaths and the smaller base populations. Nonetheless, a steady rise in death rates could indicate that the local area is either getting older or that the relative risk is increasing and needs to be analyzed more closely to determine what factors have induced this comparative rise. Wide fluctuations around a generally constant trend should occur under normal circumstances. Given these factors, let us now examine the trends in the Oak Ridge area to determine the direction of the mortality trends.

Trends in Selected Mortalities in the Oak Ridge Area

Trends in mortality from 1929 to 1971 include a 14-year period prior to the existence of Oak Ridge and its three nuclear facilities and a 29-year period after its founding in 1943. Fetal

deaths, infant deaths, and deaths from congenital malformation have been declining slowly over time in the white populations of the Oak Ridge area and Tennessee (14). Trends in cancer among whites in the Oak Ridge area and in Tennessee have been increasing over the period; this reflects the conquest of competing causes of death resulting in rising rates for chronic diseases such as cancer as longevity increases (3).

The first vital statistics for Oak Ridge became available in 1949. If the data for the period 1949 to 1971 are examined, the trends in deaths for the four causes reflect no particular sequence which would suggest that the Oak Ridge area has been or is becoming a relatively hazardous locale. Since the number of deaths is small, the variability is large; but the trends are fairly consistent. The city of Oak Ridge, which is closest to the nuclear facilities, does not show any consistent increasing trend, nor do Anderson and Roane Counties in which the facilities are located. All three areas reflect the same general trends as the State of Tennessee. However, rates (the ratio of deaths to population) are more appropriate for comparative purposes in relation to ascertaining a radiation effect or any other type of health effect gradient. Note well that cancer rates are total cancer deaths per 10^5 total population, while the rates for infant and fetal deaths are per 10^3 live births (18).

An examination of trends of cancer mortality rates among whites reveals nothing that would indicate the presence of a radiation effect. The rates for the Oak Ridge white population which would be the closest to the releases of radioactive materials--hence, most exposed--are the lowest rates depicted. They also have undergone wide fluctuations, not the consistent upward trend expected if cumulative radiation exposure were a primary etiologic agent. In only two periods, 1964-65 and 1968-69, are the cancer rates higher than in the previous year. At least since 1949, the trends in cancer mortality, though rising in the Oak Ridge area (and in the State), have not shown a consistent pattern that might suggest a radiation problem. In fact, the 1929-43 trends in Anderson and Roane Counties would appear to naturally extend into the 1949-71 trend in the same general upward flow as demonstrated by the State trend.

The trends in rates of fetal deaths (stillbirths), infant deaths, and congenital malformations appear to be equally consistent as cancer in not revealing a trend which would suggest an effect subsequent to Oak Ridge operations. Oak Ridge has consistently had lower mortality rates than either Anderson or Roane Counties, suggesting an inverse distance gradient. Trends for all three causes, though showing wide fluctuations, have been downward which is not suggestive of an adverse or cumulative radiation effect. All three areas tend to reflect the experience shown by the trend in the State rates.

In addition, the "relative risks" of death in the Oak Ridge area have also been computed using the local population rates and Tennessee rates. Oak Ridge appears to have consistently had a lower relative risk from each cause. Anderson County appears to have shifted in the 1940's from an area of relatively higher risks to an area of relatively lower risks for all causes except cancer, while actual cancer risks have been lower than expected since 1929. Roane County appears to have had higher risks due to infant deaths and congenital malformations and lower risks due to cancer since 1929; while risks of fetal deaths were, more often than not, lower in the pre-Oak Ridge years and higher in the post-Oak Ridge years. In the post-Oak Ridge years, Oak Ridge has consistently had the lowest relative risk for each cause, no doubt a reflection of age and socioeconomic factors.

The upward convergence of the Oak Ridge and Anderson County crude cancer death rates toward the State rate is consistent with several hypotheses, including an effect due to the nuclear facilities, though such an effect is not shown in any of the other mortalities or in Roane County. Since the most obvious reason for such an increase in cancer rates is the increasing age of the local population, the age-adjusted cancer mortality data by county for the 1950-69 period produced by the National Cancer Institute were analyzed statistically (Chi square) to compare Anderson County, Roane County, and Tennessee (4). These age-adjusted data (Table 1) indicated that there are no significantly greater rates in Anderson and Roane Counties and suggest that the nonage-adjusted temporal trends seen in the convergence of the cancer rates in Oak Ridge and Anderson County toward the State rate are probably due to the increasing proportion of older ages over time in Oak Ridge and Anderson County.

Conclusion

Thus, the statistical evidence, through preliminary and certainly not conclusive, suggests that Oak Ridge's nuclear facilities have not adversely affected mortality from selected causes often associated with high doses of radiation. Further analysis is needed to insure that instances in which rates of mortality were greater than those of the State were either (1) in response to relative changes in age structure of the local population over time, as seems to be indicated by independent verification using NCI age-adjusted figures, or (2) due to convergence as socioeconomic differentials between the State and Oak Ridge area have narrowed over time, rather than to environmental agents such as radiation.

In essence, the mortality trends do not show a pattern in time or space which would suggest that the presence of the Oak Ridge nuclear facilities has resulted in adverse impacts on the health of the local population. While the statistical results seem to imply the local environment is relatively safe, as in the studies previously cited, there remain potentially serious limitations in the data which are being more fully assessed, including the roles of migration, age structure, and socioeconomic factors (1, 9). Although high levels of radiation are a proven threat to man's health, no evidence of harm to

the general public has yet been found to be due to low levels of radioactivity such as might result from Oak Ridge's nuclear facilities (6, 9, 13, 16). Nevertheless detailed epidemiological analysis is still needed in this area because existing studies have been unable to detect consistent changes in measures of health in the area from preoperational years. Using measures of both potential somatic and genetic effects, this study of the Oak Ridge nuclear facilities has found no adverse impact on public health that can be attributed to the operation of the facilities. The low-level radiation effects from nuclear facilities remain relatively unknown but appear to be less a hazard than the fossil fuel pollutants. Nonetheless, further indepth epidemiological research is needed before this issue is settled and risks of alternative energy technologies are preceived fully by the public and policy makers.

References

1. Kitagawa, E. M., and Hauser, P. M., 1974, Social and Economic Differentials in Mortality (Harvard Univ. Press: Cambridge).

2. Larson, C. E., Lincoln, T. A., and Bahler, K. W., June 9, 1966, Comparison of Mortality of Union Carbide Employees in Oak Ridge Atomic Energy Facilities with U.S. Bureau of Vital Statistics Mortality, Report No. K-A-708 (Oak Ridge).

3. Lilienfeld, A. M., Pedersen, E., and Dows, J. E., 1967, Cancer Epidemiology: Methods of Study (Johns Hopkins Press: Baltimore).

4. Mason, T. J., and McKay, F. W., 1974, U.S. Cancer Mortality by County: 1950-1969, DHEW Pub. No. (NIH) 74-615 (Public Health Service Institute: Bethesda).

5. Moshman, J., and Holland, A. H., Jr., 1949, "On the Incidence of Cancer in Oak Ridge, Tennessee," Cancer 2(4): 567-575.

6. National Academy of Sciences Advisory Committee on the Biological Effects of Ionizing Radiations (BEIR Committee), November 1972, The Effects on Populations of Exposure to Low Levels of Ionizing Radiation (NAS: Washington).

7. National Center for Health Statistics, November 1968, Migration, Vital, and Health Statistics, PHS Pub. No. 1000, Series 4, No. 9 (USGPO: Washington).

8. National Center for Health Statistics, Standardized Micro-Data Tape Transcripts, DHEW Pub. No. (HRA) 74-1213 (PHS: Rockville).

9. Patrick, C. H. 1977, "Trends in Public Health in the Population Near Nuclear Facilities: A Critical Assessment," Nuclear Safety, 18(5), 647-662.

10. Public Health Service National Cancer Institute, 1975, Third National Cancer Survey: Incidence Data (NCI Monograph 41), DHEW Pub. No. (NIH) 75-787 (PHS: Bethesda).

11. Public Health Service Center for Disease Control, 1975, Congenital Malformations Surveillance (Annual Summary 1974) (PHS-CDC: Atlanta).

12. Scott, L. M., Bahler, K. W., de la Garza, A., and Lincoln, T. A., 1972, "Mortality Experience of Uranium and Nonuranium Workers," Health Phys. 23(4): 555-557.

13. Sonnenblick, B. P., 1972, Low and Very Low Dose Influences of Ionizing Radiations on Cells and Organisms, Including Man: A Bibliography, DHEW Pub. No. (FDA) 72-8029, BRH/DBE 72-1 (Public Health Service: Rockville).

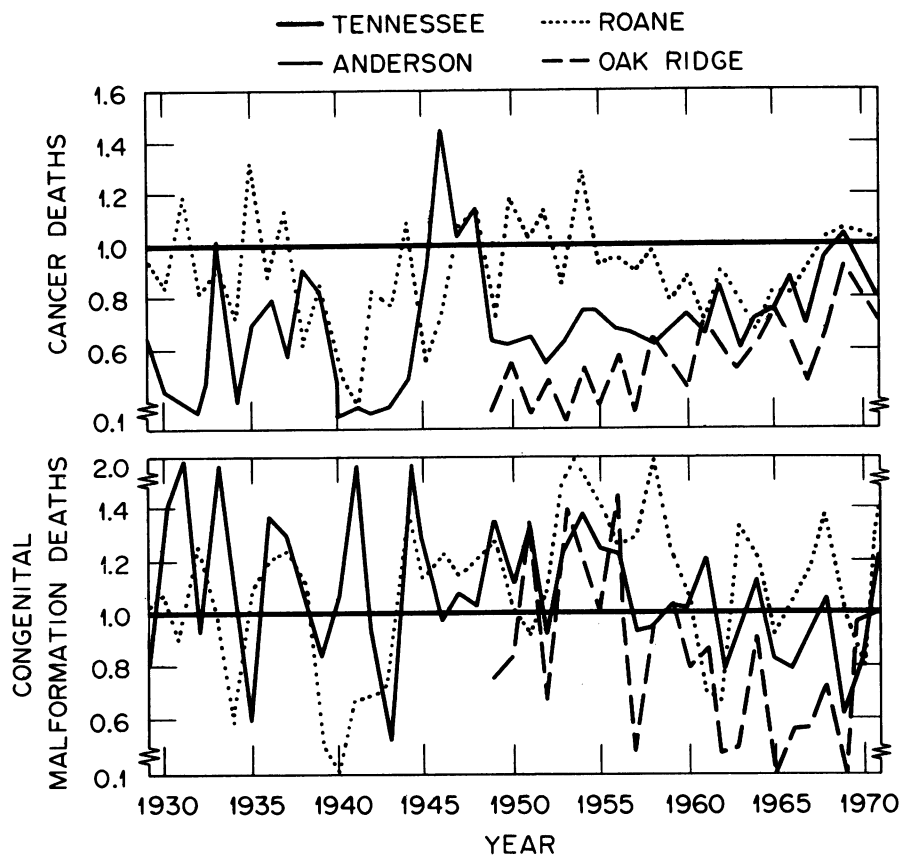
14. Tennessee Department of Public Health, issued annually, Annual Bulletin of Vital Statistics (TDPH: Nashville).

15. Tennessee Department of Public Health, 1971, Annual Bulletin of Vital Statistics (TDPH: Nashville).

16. United Nations Scientific Committee on the Effects of Atomic Radiation, 1972, Ionizing Radiation Levels and Effects, Vols. I and II (United Nations: New York).

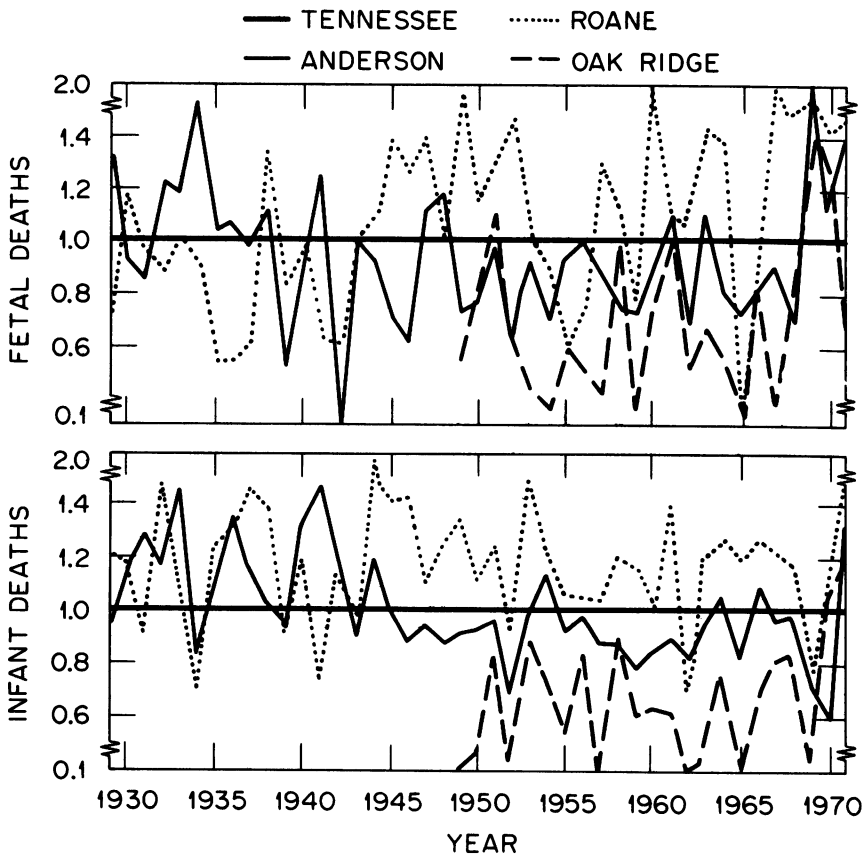
17. U.S. Bureau of the Census, Various issues, Current Population Reports, Series P-25 (USGPO: Washington).

18. Data and results available upon request.



SOURCE : TABLE 4

Ratio of Actual to Expected Cancer and Congenital Malformation Deaths in the Oak Ridge Area and the State White Populations, 1929–1971.



SOURCE : TABLE 4

Ratio of Actual to Expected Fetal and Infant Deaths in the Oak Ridge Area and the State White Population, 1929–1971.

Table 1. Age-Adjusted Mortality Rates and Actual and Expected Deaths for Selected Cancer Types by Sex and Rate in Tennessee, Anderson County and Roane County -- 1950-1969

Type	Tennessee		Anderson County				Roane County			
	No.	Rate	No.	Rate	Expected	x^2	No.	Rate	Expected	x^2
ALL CANCERS										
WM	38,356	146.3	544	143.8	553.5	0.16	432	154.4	409.3	1.25
WF	35,763	116.0	510	117.4	503.9	0.07	384	120.2	370.6	0.49
NM	7,874	163.8	22	236.3	15.3	2.99	22	145.3	24.8	0.32
NF	2,268	142.5	22	213.3	14.7	3.63	26	165.8	22.4	0.60
LEUKEMIA										
WM	2,268	8.4	39	8.7	37.7	0.05	20	6.4	26.3	1.49
WF	1,700	5.6	32	6.2	28.9	0.33	20	6.0	18.7	0.10
NM	301	6.0	1	9.6	0.6	0.23	2	12.5	1.0	1.13
NF	222	3.9	2	16.7	0.5	5.03[a]	0	--	--	--
LUNG										
WM	8,885	33.5	151	38.6	131.1	3.04	111	38.7	96.1	2.32
WF	1,673	5.5	23	6.0	21.1	0.17	15	4.7	17.6	0.37
NM	1,387	28.8	6	59.0	2.9	3.22	2	11.7	4.9	1.74
NF	300	5.5	2	20.8	0.5	4.09[a]	1	6.0	0.9	0.01
BONE										
WM	371	1.4	8	1.7	6.6	0.30	5	1.4	5.0	0.0
WF	366	1.2	4	0.7	6.9	1.19	6	1.8	4.0	1.00
NM	59	1.2	0	--	--	--	0	--	--	--
NF	43	0.8	0	--	--	--	0	--	--	--
THYROID										
WM	91	0.3	2	0.6	1.0	1.00	0	--	--	--
WF	195	0.6	2	0.6	2.0	0.0	2	0.6	2.0	0.0
NM	12	0.3	0	--	--	--	-	--	--	--
NF	28	0.5	0	--	--	--	0	--	--	--

*$x^2 = \frac{(\text{Observed} - \text{Expected})}{\text{Expected}}$, with one degree of freedom; expected number based on the state rate applied to the local population.

**WM (white male), WF(white female), NM (nonwhite male), NF (nonwhite female).

[a]Significant at the 0.05 level ($x^2 \geq 3.84$).

PREDICTIONS OF THE EFFECTS OF ENERGY PRODUCTION ON HUMAN HEALTH

R. T. Lundy and D. Grahn, Argonne National Laboratory

In recent years, it has been discovered that there are certain risks to human health associated with various facets of an affluent industrialized society. As a a consequence regulations have been promulgated with the intent of protecting our health, and numerous studies have been done to determine whether additional protection is needed. As many of the effluents from energy systems are among those identified as hazardous, those who must plan and analyze various energy options must perforce take into account the health effects anticipated in any situation being considered.

Some effluents are now regulated; others may be in the future. Pressures are occasionally brought to bear from industry to relax or eliminate regulations once instituted. The person trying to foresee the effects of a given policy needs to have some way to guess the likely course of future regulations, which are among the major economic and engineering constraints that must be considered. Future regulations may in part be projected on the basis of health effects. Also, there are economic tradeoffs to be made, whatever the constraints, and these, too, require a realistic estimate of the health consequences. All of this analysis requires an appropriate quantitative model.

For purposes of analysis, it is desirable to be able to project what will happen; how much of it will happen; when it will happen; and to whom it will happen. This information can be expressed from several perspectives. The most important of these are the "personal" and the "real population" perspectives. The "personal" perspective expresses risks as seen by an individual -- "what will happen to my personal chances of survival?" It is from this point view that insurance premiums are (ideally) calculated. This is in many cases a useful point of departure, but it carries with it some important assumptions which are not always apparent: For example, it assumes that , if your expected days in the hospital are raised by 20%, that the hospital facilities will be available for your use and the doctors will be there to treat you. Such an assumption may be reasonable if we are talking about a relatively small occupational group within the larger society, in which case the situation would fall within the normal variation in the usage of the facilities available to society as a whole. The same assumption may not be reasonable when the group at risk is essentially the whole society. To deal effectively with that case, we must look from the perspective of the "real population" to determine from society's point of view what the potential demand for health-related facilities might be. This determination is made by integrating the "personal" risks over the distribution of persons at various types of risk. At this level, estimates are often made by defining a single does-response coefficient and applying it to an estimate of the total population at risk. This, however, requires that the distribution of persons within the population at risk remain constant. This is not a safe assumption. Also, the use of an independently derived population estimate can lead to the theoretical death of the same person more than once in the course of a projection.

Frequently a dose-response coefficient derived from one population is applied to another population whose composition and characteristics are so different that the results become unreliable. For example, such an error would involve projecting health effects in a general population by using dose-response coefficients from a study of asthmatics. Although such a blatant error has never to our knowledge been made, a more subtle form of this error occurs whenever dose-response coefficients derived from one population are applied to a population whose age profile differs significantly from the one from which the coefficient was derived. This error can occur even if the population appears at first glance to be the same. For example, a study examining hypothetical health effects expected in the population of the U. S. in 1970 would make such an age distribution error if the dose-response coefficients used had been generated from the U. S. population in 1960, since the age distribution shifted markedly in that decade as a result of changing fertility levels over the previous 40 years. Consequently, a model must carry out two functions:

1. It must project the response to an exposure as a function of level and duration of exposure, and of the age, sex, and any other predisposing factor associated with a definable class of person.

2. It must project the distribution of such people during the period of time to be covered by the analysis.

Projecting Distributions of Persons at Various Levels of Risk: The Demographic Module.

Most major risk factors are associated

with age and sex. The susceptibility of most people to the ill effects of exposure to a toxic material tends to increase exponentially with age though there is an additional peak in susceptibility in the first year of life. Thus, one can go a long way towards projecting the risk level distribution simply by projecting the age and sex distribution.

The problem of projecting the future population has concerned demographers for over a century. A number of procedures, some of them quite sophisticated, have been devised to deal with it. The most appropriate procedure for any particular population will depend on its particular characteristics. However, for illustrative purposes, the component projection model developed by Whelpton, generalized by Leslie, and described by Keyfitz (1), will be presented here. The procedure by which the model is extended to project deaths as well as living population may be applied to any projection scheme.

Let

x = exact age index.

n = the length of an age interval or projection interval.

i = the age group definition index.

$= \text{trunc}\left(\frac{x}{n}\right) + 1$

${}_s k_i^t$ = population in age group i and sex group s.

${}_s S_i$ = probability of survival from age group i to age group i+1 during an n-year interval.

${}_s F_i$ = expected number of children that will be born to a woman starting in age group i during n years.

We can then assemble the ks into a column vector of population K and the F and S terms into a square matrix M such that the population vector at time t+n is related to the population vector at time t by

$$\underset{\sim}{K}^{t+n} = \underset{\sim}{L} \times \underset{\sim}{K}^{t} \qquad (1)$$

The cells in the projection matrix are customarily estimated by assuming that the age distribution within each age group is similar to that in a stationary population in which case the subdiagonal survival terms are given by

$$S_i = \frac{{}_n L_{x+n}}{{}_n L_x}$$

where ${}_n L_x$ = numbers in life table age age distribution aged x to x+n.

The reproductive terms are given by

$$F_i = 2.5\ ({}_n b_x + S_i\ {}_n b_{x+n}) \qquad (2)$$

where ${}_n b_x$ is the yearly birth rate for women age x to x+n.

By interdicting the computations at the appropriate point, it is possible to estimate the distribution of deaths by age group as well. The total deaths in a cohort starting in age group i and surviving to age group i+1 is given by

$$d^*_i = k_i - k_{i+1} \qquad (3)$$

We can define a factor Z_i such that

$Z_i \cdot d^*_i$ people die in age group i

$(1.0 - Z_i) \cdot d^*_i$ people die in age group i+1

If we assume that the age distribution of the deaths as well as the population within each 5-year age group during the passage between one age group and the next is the same as in the life table, then

$$Z_i = \frac{{}_n d_x}{({}_n d_x + {}_n d_{x+n})} \qquad (4)$$

where ${}_n d_x$ is the number of life table deaths in age group x to x+n.

The number of deaths in age group i during the projection interval is then

$$D_i = (1 - Z_{i-1})\ d^*_{i-1} + Z_i\ d^*_i. \qquad (5)$$

Projecting Changes in Health: The Dose/Response Module.

It is in the area of dose/response relationships that most other modeling efforts are concentrated. A dose/response function is a relationship between the degree of exposure to a toxic substance and the degree of excess risk that can be observed as a consequence of that exposure. In its simplest form, the function states that:

$$du = Bdp,$$

where

u = risk of death

p = exposure index

B = proportionality factor

It is apparent, however, that the change in the risk of death as a consequence of any given change in

exposure will not be the same for all persons exposed. Also, the consequences of the pollutants with which we will be dealing tend to show a prolonged latent period before the full effects can be seen. Consequently, the function should be disaggregated to whatever degree is necessary to assure reasonable homogeneity within groups, and it should be made duration-specific as well. At the current stage of development , this disaggregation is limited, as is the demographic module, to age (in 5-year groups) and sex.

Let us now focus our attention on the response function, and the effluents to which it refers.

In the area of energy production and public health, one class of effluents is of particular importance: airborne combustion products and the by-products which they give rise to in the course of their travels through the atmosphere. There is a bewildering array of them, and almost all can be found in any given sample of polluted air. For purposes of analysis, however, most investigators have chosen to index air pollution levels on one or two of the more prominent, easily measured, or otherwise interesting components. The most commonly used of these are total suspended particulates (TSP), sulfur dioxide (SO_2), or suspended sulfates (SO_4). The next thing traditionally done in such studies is to focus attention on a carefully selected subgroup of the population, usually chosen on the basis of a compromise between high a priori susceptibility and large numbers.

Most existing models had their origins in studies in which the major interest was in the derivation of qualitative estimates of relationships (e.g., is SO_2 bad or isn't it?), or in estimating in retrospect what the cumulative quantitative effects had been. Epidemiological models especially tend not to consider explicitly that the composition of the population being studied can (and usually will) change markedly with time. They generally refer to populations defined so broadly that their internal structure can change drastically with respect to many factors often confounded with pollution-related health effects (i.e., age, socioeconomic status, % suffering from morbid conditions, etc.), while still remaining within the original definition of the study population (e.g., "total," "whites 35 years of age and over," "employees hired in 1950-55," etc.)

There is, however, one major source of air pollution associated with combustion products that has been studied very throughly indeed: the cigarette. It is not, of course, usually considered in the context of fossil energy sources, although the number of BTU's of cigarettes burned each hour in the United States is the approximate equivalent of 12-15 tons of coal, an amount great enough to operate a 26 MWe power plant.

Unlike the epidemiological studies of air pollution, in which neither duration nor magnitude of exposure are easily measured, the investigators of smokers have been able to do reasonably well-controlled prospective studies in which age at onset, degree of exposure, and outcome are all defined with reasonable accuracy. Assuming cigarette smoke, then, to be just another air pollutant, let us look at the function relating increments in age-specific death rates to exposure measured in number of cigarettes smoked per day, as shown in Table 1. The same data are graphed against age in figure 1.

We note that above age 50, the semilog plot of the response curve constitutes for both sexes (males particularly) a reasonably straight line indicating a constant exponential increase in damage, while below that age the curve drops away from this line and presumably would, if extended properly, hit 0 at around the mean age at which each sex begins to smoke. This is close to 15 for males and 20 for females. Why should this be so?

If we assume any of several models indicating that the ability of mammalian organisms to withstand the ravages of their environment declines in inverse proportion to their age, the familiar Gompertz law of exponentially increased risk would be expected

$$g(x) = a\, e^{bx} \qquad (6)$$

On the other hand, it can be shown that whenever some increment of damage occurs to an organism, various repair mechanisms are brought into play. In a situation of constant exposure to a toxic agent, the amount of repair taking place tends to rise in direct proportion to the damage accrued. Under this assumption, one would expect the damage function to rise asymtotically to some constant value as the incremental damage and repair effects reached equilibrium over a period of time. Under this assumption, furthermore, the change in the damage function would follow the logistic function

$$v(x) = \frac{1}{1 + ce^{-d(x-x_o)}} \qquad (7)$$

where x_o is the age at onset of exposure.

Both effects would appear to be operating simultaneously in the present case. The constant exposure to the toxic agent would be initiating a process

whereby the damage function would attempt to rise over a period of several years to an equilibrium value; at the same time, however, this equilibrium value would be changing with the advancing age of the exposed organisms according to the Gompertz law. Hence, the damage function, which describes the data of Table 1, ought to have the form

$$B(x,x_o) = \frac{a\, e^{bx}}{1 + c\, e^{-d(x-x_o)}} \qquad (8)$$

This function has been fitted to the cigarette data, as shown in Table 2, and is shown superimposed on the data points in figure 1.

Applicability of the Cigarette Model to Other Forms of Air Pollution

Cigarette smoke and coal smoke differ markedly in some respects. In particular, carbon monoxide is found in far higher concentrations in cigarette smoke than in coal smoke. (Its presence indicates inefficient combustion, anathema to engineers.) The cigarette model can be justified, however, on two grounds: First, the kind of damage done by air pollutants is not specific by causative agent; sulfuric acid droplets, fly ash particles, NO_2, O_3, and SO_2 all cause the same kinds of damage to the lung in appropriate concentrations, as indeed do most of the aldehydes, ketones, and other noxious organics likely to be encountered under similar circumstances. Second, the response curves for smoking seem to fit those for air pollution data reasonably well.

It will be noted in the fit of the cigarette data given in Table 2 that the age at onset of exposure is around age 15 for males and 20 for females. When one is dealing with other airborne pollutants, however, it is immediately clear that no decision on the part of the person involved, other than migration, will prevent exposure from commencing at birth.

During the time when the data for the most recent studies of air pollution and health were collected, it is probably fair to assume that the then current level had prevailed long enough for the latency effects to have worked themselves out some time previously. Therefore, a response function was calculated from the cigarette model of equation 13 assuming constant exposure to the effluent of interest from birth. For comparative purposes, the response function for the case where SO_2 and TSP are incremented equally for whites on the basis of the regressions derived by Lave and Seskin and presented in Finch and Morris (4) were plotted against the mean of the age groups considered in their analysis as derived from the 1960 U. S. population. Figure 2 compares the cigarette function thus adapted with a plot of the Lave-Seskin points.

It can be seen that while the response pattern seems to fit very well for males, the fit for females is not as close. Three possible explanations suggest themselves : First, the exposure data used in the Lave analysis may not be as well matched to the female population in his sample as it is to the male population. There is reason to believe that the males in the SMSAs (Standard Metropolitan Statistical Areas) treated were more likely than the females to spend a significant part of their day in areas close to the locations of the sampling stations from which the the exposure data were derived. These stations were for the most part in the central urban areas, and males are more likely to commute into these areas for work than females, whose lower labor force participation rates and and differing array of employment opportunities would tend to keep them out of the relatively more polluted areas. Second, it might well be the case that there is a strong interaction between the effects of air pollution and smoking history, in which case the later onset of smoking among females might contribute significantly towards the pattern seen here. Third, one of the points in the cigarette data, specifically the one for females in the 70-79 age group may be a spurious point. If this point is eliminated from the calculations, the resulting function fits the air pollution data far more closely, although it seems not to fit the cigarette data quite as well at the lower ages. The curve derived when this point is eliminated is shown as the dashed line in figures 1 and 2.

Fitting the Cigarette-Derived Model to Air Pollution Data

A number of studies have been done that give response coefficients for various population subgroups exposed to various pollutants. We would now like to fit these data into the framework of our model. Fitting is most conveniently done by simulating the particular study and determining the cigarette-equivalent dose needed in the current model to reproduce the effect of a given pollutant dose. For example, Finch & Morris (4) have determined that the response function implied in Winkelstein's study of air pollution in Buffalo, NY, as indexed by Total Suspended Particulated (TSP) for white males 50-69 years of age is about 14 deaths per 10^5 population per ug/m^3/day incremental long-term exposure. Starting with the life table and age distribution

of U.S. white males in 1960, one finds that an assumed increment of 0.35 cigarettes per day will have the same effect in that age group. Consequently, to convert the cigarette model into a TSP model one needs only multiply the a coefficient in Table 2 by .35. Similar fits for data from other studies are given in table 3.

This fitting procedure yields a further dividend in that it gives us a method for projecting the results of some studies beyond their original age boundaries.

Merging the two components

The complete model, then, operates as follows: The exposure level of the index pollutant and the initial population are both defined. Then the population is projected forward in time, with the projection matrix being modified at each cycle according to the dose-response function.

Example

The question might be raised, what advantage do we derive from using such an elaborate model? How will its results differ from those obtained with one of the simpler methods, e.g., OBERS or other projections of the total population size, and using a simple response coefficient? How important are these distributional factors? Table 4 compares the results obtained with a single-coefficient procedure and with the one proposed here. Estimates were calculated on the assumption that fertility levels in the 30-state region would be the same as the 1971 level through the year 2020, and that the 8.95 $\mu g/m^3$ increment in suspended SO_4 was instituted in 1970. The pattern of deviations between the two systems is striking. The simple model grossly overestimates the number of excess deaths in 1985 due to the latency factor. In 2000, the latency effect has passed, and, by coincidence, the age distribution estimated for that year leads to a reasonably close concordance between the results of the two models. By 2020, however, the simple coefficient suggest 23% fewer excess deaths than does the model because the population at that time will have a decidedly older average age profile than in either 2000 or in 1960, the time at which the current age distribution was used to fit the simple coefficient to the response function.

Discussion

We have defined here a model system for projecting the excess mortality that might be observed in a population exposed to an increment of environmental insult. It avoids many of the pitfalls found in most current approaches.

The system here presented is not meant to be the last word on the subject. Among the features not considered here, but which deserve attention, are:

- The effects of constantly changing exposure levels on the response function.
- The effects of migration into and out of a polluted area.
- The effects of reductions, as opposed to increases, in exposure levels. Cigarette data would suggest that for some phenomena, particularly cardiovascular disease, the recovery rate once exposure has ceased is far faster than would be anticipated on the basis of the current equations. This phenomena could have a strong impact when investigating the policy implications of tightening air quality standards.

References

1. Keyfitz, Nathan, Introduction to the Mathematics of Population, Addison-Wesley, New York, 1968.
2. Lave, Lester, and E. P. Seskin, table cited in Finch & Morris below.
3. Carnow, B. W., and P. Meier, "Air Pollution and Pulmonary Cancer," Archives of Environmental Health, vol. 27, Sept. 1973, pp. 207-218.
4. Finch, S. J., and S. C. Morris, Consistency of Reported Health Effects of Air Pollution, Brookhaven National Laboratory, BNL-21808.
5. Hammond, E. C. "Smoking in Relation to the Death Rates of One Million Men and Women," in W. Haenzel, ed., Epidemiological Study of Cancer and Other Chronic Diseases, NCI Monograph 19, Washington, D.C., 1966.

Table 1
Effects of smoking on death rates for both sexes by age. Data derived from Hammond(5) especially appendix tables 2 and 3.

Age Group	Mean Cigs/day All Smokers Female	Male	Increase in Death Rate/Cig. Female	Male
35-39	20.6	28.5	0.34	3.9
40-44	20.3	28.8	1.3	7.1
45-49	20.0	28.9	3.1	14.6
50-54	19.5	28.6	6.0	19.2
55-59	18.7	27.3	9.5	33.3
60-64	17.6	25.4	12.2	46.1
65-69	16.4	23.4	28.8	72.3
70-74	14.9	21.0	51.1	94.6
75-79	14.2	18.0	19.8	139.2
80-84	12.0	17.4	172.4	188.3

Table 2
Fitted coefficients of equation using data of table 1.

Coef.	Females	Males
a	6.24×10^{-7}	9.14×10^{-6}
b	8.84×10^{-2}	6.44×10^{-2}
c	100.0	100.0
d	0.2	0.2
x_o	20	15

Table 3
Coefficients to convert $\mu g/m^3$ pollution exposures into cigarette/day equivalents.

Study	Index Pollutant	Crude Response	Conv. Coef.
Winkelstein	TSP	.00014	.35
Morris & Novak	SO_4	.000033	.21
Lave & Seskin	TSP	.835	.09
	SO_2	.715	

Table 4

Projected premature deaths in a hypothetical population with a resemblance to that of the North Central and Northeastern regions in 1985, 2000, and 2020 assuming that the mean suspended sulfate exposure rises to 8.95 $\mu g/m^3$ of air starting in 1970.

Year	Est. Pop. $\times 10^6$	Simple Estimate (Morris & Novak)	Model Estimate
1985	201	48,000	12,000
2000	225	66,400	62,500
2020	256	75,600	92,800

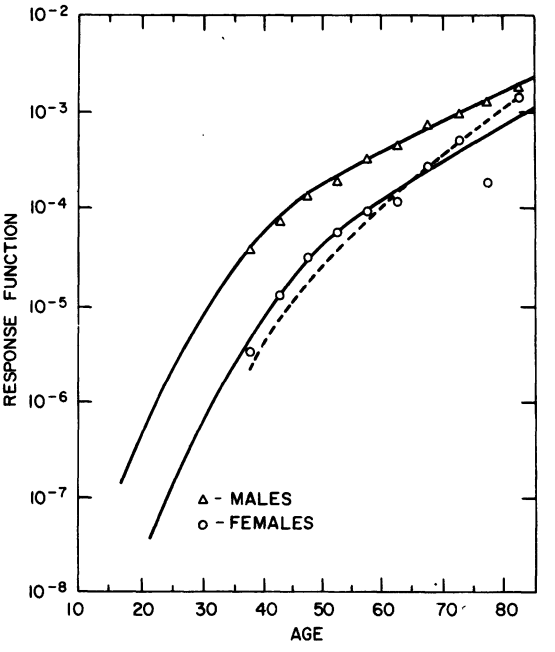


Figure 1

Increment in death rates per cigarette plotted with fitted response function.

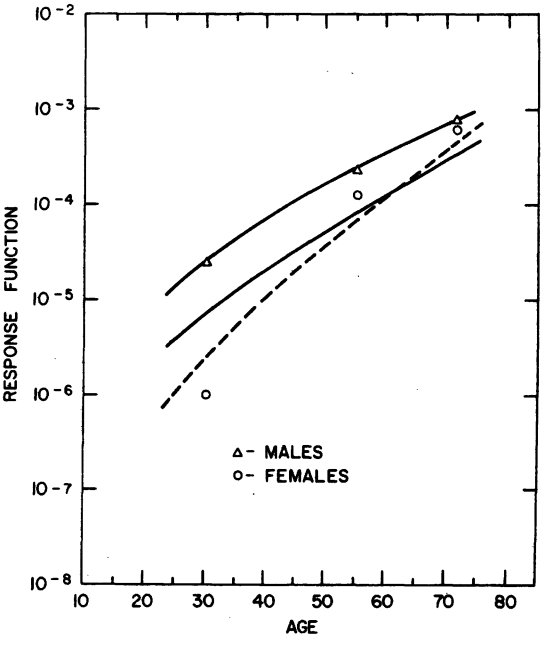


Figure 2

Increment in death rates per unit of polluted air plotted with adapted cigarette response function.

A GENERAL INDEX OF HEALTH: SOME PROBLEMS AND DESIRABLE CHARACTERISTICS

Martin K. Chen

National Center for Health Services Research
Rockville, Maryland

The need for quantitative indices of health status has been recognized by health workers for almost half a century. As early as the 1930's, the Health Organization of the League of Nations charged two scientists to develop expressions of health in numerical terms. The results of the scientists' efforts were three indices: vitality and health, environment, and public health activity (Stouman et. al, 1939). Since that time a variety of health status indices applicable to individuals and populations have come into being. Paradoxically, however, the burgeoning proliferation of health status indices of various descriptions and orientations in the literature has not made the task easier for health planners who must use these indices to evaluate the effectiveness of current and proposed health programs. In point of fact, the search for a usable general index of health applicable in various health planning settings is becoming ever more frenzied, as witness the spate of mail requesting information each time the Clearing House on Health Status Indexes, a quarterly publication of the National Center for Health Statistics, U.S. Public Health Service, prints a new item in its bibliography.

Nothing in the preceding paragraph should be construed to mean that there are no useful of usable health status indices on the market. Several indices, including the Activities of Daily Living or ADL (Katz et. al, 1963) and the G-index (Chen, 1973) have been applied successfully in health program evaluation, but these are special-purpose indices that do not have general applicability. There are, however, no known general-purpose health indices that have been tested and are ready for application. The stochastic models of population health status developed by Chiang (1965) and by Chiang and Cohen (1973) are mathematically elegant and logically sound and straight forward, but they have not been tested with real data nor are they testable until the problem of determining the values of various functional or dysfunctional states of health is resolved. The values are the weights for the discrete segments of a health status continuum from death to optimum health called for in the models.

The Definitional Problem

One of the basic problems of designing a general health status index applicable to individuals or populations is the problem of defining health to the satisfaction of the scientific community, and if possible, the lay public. Scientifically, health must be defined in concrete terms that are both quantifiable and consistent with the available body of medical knowledge about human health. Further, the definition must be comprehensive and inclusive of all known aspects of health and their dynamics. Such a definition would not be in the spirit of the usual "operational definition" that arbitrarily limits the scope of the definition only to parameters or aspects that are concrete and measurable. "Operational definitions," while necessary for research and scientific progress, usually reflect the orientations of the researchers who use them, and as such may not be acceptable to the majority of the scientific community.

Many attempts at defining health have been made by various scientists and organizations in the past decades. Stocks (1955) believes that the assessment of the "healthiness" of a community in terms of a numerical index useful for cross-community comparisons "poses a problem of the greatest difficulty" because it is impossible to have a clear definition of the concept of positive health as expressed in the WHO definition, "Health is a state of complete physical, mental, and social well-being and not merely the absence of disease and illness." He suggests that designers of health indices use measures based on "either freedom from illness or ability to continue living," but he makes no attempt to define illness or ability to continue living.

Wylie (1970), deploring the fact that it is circular reasoning to attempt to define health in terms of the absence of disease without trying also to define disease, offers his own definition of health as "the perfect continuing adjustment of an organism to its environment." However, he neither defines adjustment not suggests any way of measuring it.

Klarman (1965), aware of the vagueness of the WHO definition, believes that it is hopeless to attempt to have a standard definition that is universally accepted. An economist, he offers a pragmatic solution to the problem by suggesting that the definition of health be left to the health care industry and health care administrators in terms of the costs of services, personnel and facilities. This suggestion, of course, is no help to authors of health status indices because costs of health services, personnel and facilities are not legitimate proxy measures of health status.

The theory of homeostasis, both biological and social, is apparently the basis of Sigerist's (1941) definition of health as "something positive, a joyful attitude toward life, and a cheerful acceptance of the responsibilities that life puts on the individual. The imprint of this theory is even more pronounced in his later attempt to define health as "undisturbed rhythm and harmony with nature, culture and habit," (Sigerist, 1960). The vagueness of the terms used, such as "undisturbed rhythm" and "harmony,"

makes his definitions of dubious value to workers in the area of health status indices.

The American Public Health Association (1961) differentiates four stages of health as the discrete steps of an ordinal scale that comprise mortality, serious morbidity, minor morbidity, and positive health. Until the terms "serious morbidity," "minor morbidity" and "positive health" are given concrete definitions, it is unlikely that this definition of health can ever be of anything more than theoretical interest to health researchers.

This sampling of the definitions of health makes it abundantly clear that health is an elusive concept that is difficult to pin down neatly in a concise definition. Practically all the definitions employ terms that themselves require definition. Some of the definitions are oriented toward certain aspects of health. For instance, Sigerist's definition (1941) pertains to mental and perhaps social health, but has nothing to do with physical health. A Pollyanna philospher who is dying of cancer of the lung would be considered healthy by this definition. Other definitions, such as Stock's definition, are mere tautologies. Needless to say, without a satisfactory definition of health, there cannot be a satisfactory health status index.

Methodological Difficulties

The concept of health status as a continuum is intuitively appealing because individuals can be neatly represented as points moving along this continuum toward a more or less healthy state. This is the concept used, for instance, by Chiang and Cohen (1973) in deriving their health index. This concept, however, is not a definition of health; it provides no information about factors or forces that are responsible for movements of the points along the continuum in either direction at varying speeds. In other words, the concept is merely a unidimensional representation of a phenomenon called health that is not only multidimensional, but whose multi-dimensions are most probably not orthogonal.

In terms of indices applicable to individuals, the problem then becomes the location of an individual in hyperspace and representing this location by a scalar that is some function of the various dimensions. While the statistical methodologies in multi-variate analysis are currently available for performing this task, the dimensionality of health is unknown and even if it were valid and reliable measures of these dimensions would have to be developed first. Further, a dynamic model of individual health must also take into account the dynamics of health, genetics and environment, and knowledge about this dynamics is sketchy and fragmentary at this time and will probably remain so for years to come.

As for health indices applicable to populations, the problems affecting individual health indices are further compounded by the fact that somehow values must be assigned to the graduations along the health continuum, so that the summary scalar representing population health reflects not only the distribution of people in the graduated states, but also the degree of desirability of that particular distribution. Without the assigned values, which ideally should be derived through social concensus, the scalar would be meaningless as an index because it would lack the properties of an ordinal scale along the desirability dimension. Without the ordinal properties an index cannot be used to evaluate the health status of populations or individuals.

As a dynamic model, the population index must consider, not only the distribution of people in the graduated states at a given point in time, but also shifts in the distribution within a stated time span. Information about the shifts is derived from the transition probabilities involving Markov chain processes. A formidable problem in the estimation of transition probabilities is the appropriate classification of people into the graduated states. If the graduated states are too gross, then many people may be in the same state due to vastly different underlying causes. For instance, if one graduated state were categorized as "bedridden," it would include people who sprained their ankles, people who had active pulmonary tuberculosis, and people with bad colds. The transition probabilities of these three types of people would not be the same. Yet the transition probabilities estimated from this state would be based on all types of people. These transition probabilities would be different, perhaps drastically, if a different combination of types of people were in it. Thus no stable transition probabilities could be estimated. On the other hand, a too fine graduation would reduce the numbers of people in some of the states to the degree where no reliable estimates of transition probabilities would be feasible.

The Validation Problem

Although a variety of health status indices are available, very few of them have been validated to generate evidence that they truly measure health or at least some aspects of health. As is evident from the definitions of health previously cited, the concept of health is not a discrete entity that can be directly observed. What is observable is the totality of physiological, biological, and behavioral manifestations of the underlying health status. Both inductive and deductive logic is required to establish evidence of causality between health and its manifestations. Thus establishing the validity of health status indices, whether the indices apply to individuals or to populations, is a time-consuming process.

One of the main reasons authors of health

status indices generally fail to validate their products is that a well-conceived health status index usually encompasses most of the salient aspects of health, and once these aspects are incorporated into the index, they cannot be used as criteria for validation because the relationship between the criteria and the index would be spurious. Another reason, already alluded to previously, is the lack of adequate knowledge about the interrelationships of various manifest health-related behaviors, including physiological behavior, as well as the relationships of the behaviors to the underlying health level.

This lack is particularly vexing to authors of health indices that include the mental health component. So-called aberrant behaviors in one culture are perfectly normal in another. Even within one culture the distinction between normal and deviant behavior is not all that clear and some distinctions disappear with the changes in social mores, as is the case with homosexuality in the United States. Thus a new dimension comes into the picture: cultural factors, along with the underlying health level, may influence the manifest health-related behaviors. This new dimension compounds the problems of attempts to validate health status indices.

Some Desirable Characteristics

The difficulties relative to the definition of health may appear--indeed, may actually be, insuperable. Nonetheless, general indices of health are needed by health services researchers and health planners. As a matter of fact, the National Health Planning Act (1975) specifically directs that Health Systems Agencies study the impact of health care delivery systems on the health of residents under their jurisdictions. Unless the law is satisfied with the individual health indicators such as mortality rate and/or hospitalization data, some kind of general index of health will have to be developed in spite of the difficulties discussed. What characteristics should such an index have to be useful?

At a minimum, the index must possess the properties of the ordinal scale. That is to say, the values of the index can be used to rank order communities or individuals in terms of their underlying health status, but not to determine the extent of differences among the communities or individuals. In other words, ordinal scale satisfies the following two postulates and no other: (1) if $a > b$, then $b \not> a$, and (2) if $a > b$ and $b > c$, then $a > c$.

As previously stated, the desirability of the states of health should reflect the values of a society comprising the individuals whose health is measured. While in general it is true that life is preferred to death (with the exception of suicide cases in which death is obviously preferred to life), it may be extremely difficult to attach preference values to different health conditions that are acceptable to all members of society. For instance, in terms of physical health, it would not be easy to rank order the health status of two individuals, of whom one has frequent and severe colds and the other suffers an occasional, but paralyzing, arthritic pain, assuming that they are comparable in other aspects. Nonetheless, such preference values must be derived and incorporated into the index as weights for it to have ordinality.

Another basic requirement or characteristic is that the index values should reflect the underlying health status independently of the biological, physiological and behavioral manifestations of the normal aging process. Unless this requirement is met, the index may measure health status largely as a function of age: the older one gets, the less healthy one is. An index so formulated would preclude statements such as '"Some young people are sickly whereas some older folks are "hale and hearty."' Certainly there are people in their seventies or even eighties who enjoy the best of health possible among their age groups.

The practical implication of this requirement in terms of designing a general health status index is that norms must be used, since, in the words of Dubos, (1959) "health (and happiness) cannot be absolute and permanent values, however careful the social and medical planning." In fact, the World Health Organization (1957), after a lengthy discussion of the meanings and definitions of health, concluded that health would be best expressed as "a degree of conformity to accepted standards of given criteria in terms of basic conditions of age, sex, community and region, within normal limits of variation." Thus an index that fails to take into consideration these factors may indeed be a measure of demographic and geographic artifacts rather than true underlying health status of an individual or community.

Even a norm-oriented index of health may be difficult to interpret unless the range of index values, which usually are abstract or pure numbers, is known or pre-determined. Many health status indices could be cited that, because of their employment of arbitrary measurement scales, have arbitrary values that have no lower or upper bounds and that in themselves have no meaning although they can be used to rank order individuals or communities with respect to health status. Notable exceptions are the index of Chiang and Cohen (1973) and that of Chen (1976). These indices range in value in the closed range between zero and one, which enables the reader to know the relative health status of a community by its index value without reference to other communities.

If, however, the index is not a pure number and employs known units of measurement, a closed range of values is still desirable, but not crucial. For instance, Chen's G-index (1973) is

in unit of years unnecessarily lost by a population group through poor health and/or living conditions, and this is meaningful, although it would be more informative to know the numbers of years lost by other population groups.

Other desirable characteristics or attributes of a useful general index of health pertain to the feasibility of application, its validity and reliability, and its sensitivity to changes in the underlying health status. These will not be discussed here since they have been adequately treated elsewhere (Chen, 1975). Suffice it to say here that an index without such desirable attributes has rather limited utility either as a tool in health services research or for health planning purposes.

REFERENCES

American Public Health Association
1961 "A broadened spectrum of health and morbidity." American Journal of Public Health 51: 287-294.

Chen, M.K.
1973 "The G-index for program priority." in Robert Berg (ed.), Health Status Indexes, Chicago American Hospital Research and Educational Trust.

Chen, M.K.
1975 "Health status indicators: from here to where?" Invited paper presented at Data Use and Analysis Laboratory of the National Center for Health Statistics, Orlando, Florida, April 28-29.

Chen, M.K.
1976 "The K-index as a proxy measure of health care quality." Health Services Research, in press.

Chiang, C.L.
1965 "An index of health, mathematical models." United States Public Health Service publication No. 1000-series-2-No. 5. Washington: U.S. Government Printing Office.

Chiang, C.L. and R.D. Cohen.
1973 "A stochastic model for an index of health" International Journal of Epidemiology 2:7.

Dubos, R.J.
1959 Mirage of health: utopias, progress and biological change. New York: Harper.

Katz, S. et al.
1963 "Studies of illness in the aged: the index of ADL, a standardized measure of biological and phychosocial function." Journal of American Medical Association 185:914-919.

Klarman, H.E.
1965 The economics of health. New York: Columbia University Press.

Sigerist, H.E.
1941 Medicine and human welfare. New Haven: Yale University Press.

Sigerist, H.E.
1960 "The special position of the sick." In Henry E. Sigerist on the sociology of medicine. Roemer, M.E. (ed.). New York: M.D. Publications.

Stocks, P.
1955 An index of vitality for assessment of national health levels. Unpublished report, World Health Organization, Geneva.

Stouman, K. and I.S. Falk.
1939 "Health Indices--a study of objective indices of health in relation environment and sanitation." Bulletin of Health of the League of Nations 8:63.

U.S. Congress
1975 Public law 93-651, part B, health systems agencies.

World Health Organization
1957 Measurement of health levels. Geneva: World Health Organization.

Wylie, C.M.
1970 "The definitions and measurement of health or disease." Public Health Reports 85:100-104.

COMPARISON OF METHODS FOR MEASURING SOCIAL PREFERENCES FOR A HEALTH STATUS INDEX

James W. Bush*, Robert M. Kaplan**, and Charles C. Berry*

University of California, San Diego*, and San Diego State University**

INTRODUCTION

THE HEALTH STATUS INDEX

In previous publications, a health status index has been described which could be an effective tool in health planning, health program evaluation, and population monitoring [Patrick et al. 1973a, 1973b; Bush et al. 1973; Chen et al. 1973, 1975; Blischke et al. 1975; Chen and Bush 1976; Kaplan et al. 1976].

The index separates two distinct components: Levels of Well-Being, the weights, social values, or utilities that members of society associate with a person's level of functioning at some point in time, and prognoses -- the probabilities of transition to other levels of function and Well-Being on future occasions. Treating these components as analytically distinct allows the quantitative expression of the two variables.

Since the quantities vary independently, joint functions of the two variables are necessary to fully describe health status. Thus, no precise statement of health status can be made for an individual or a group without knowledge of the expected transitions among the function levels over time. We shall, therefore, reserve the term "health" for a composite expression of prognosis and function level as well as Level of Well-Being.

The present report concerns the utility dimension of health. This is the social preference or "Level of Well-Being" for states of function on a continuum from optimum function (1.0) to death (0.0). When these weights have been measured, health status can be expressed precisely as the expected value (product) of the preferences associated with the states of function at a point in time and the probabilities of transition to other states over the remainder of the life expectancy [Kaplan et al. 1976].

Steps from three scales -- Mobility, Physical Activity, and Social Activity -- can be combined into sets called Function Levels.* Any individual can be classified into one of the mutually exclusive and collectively exhaustive Function Levels. Subjective, symptomatic disturbances are incorporated in an independent set of symptom/problem complexes whose presence or absence can be noted in surveys and follow-up studies.

Levels of Well-Being are the weights, social preferences, or measures of relative importance that members of society associate with each of the Function Levels. These preferences may be measured by having consumers rate sets of standardized but realistic case descriptions. The case descriptions consist of the items of information describing a Function Level and a Symptom/Problem Complex, and describe how a person would be classified according to the items in the Index. Thus, unlike weights obtained from arbitrary, disease specific scenarios, the weights obtained can be assigned with little error to all actual persons.

Since utilities are an important component of the Index of Well-Being, accurate, reliable ratings on an interval or ratio scale of measurement are highly desirable. This study compares results obtained via magnitude estimation, a method purported to yield ratio scales, with data obtained by a simpler, more widely accepted method known as category rating.

PREFERENCE MEASUREMENT

In a number of his publications, S.S. Stevens refers to two classes of psychological continua: prothetic describes intensity dimensions such as light or sound, and metathetic describes qualitative dimensions, such as pitch or visual position. The functional form of the responses among the scaling procedures determines the type of continuum.

Category rating is a simple partition method in which subjects are requested to assign each stimulus to a set of numbered categories representing equal intervals. This method, exemplified by the familiar 10-point rating scale, is efficient, easy to use, and applicable in a large number of laboratory and survey settings. Stevens [1966, 1971, 1974] questioned the assumption that the subjective impressions of a stimulus can be discriminated equally at each level of the scale. With Galanter [1957] he claimed that the category method is biased because subjects attempt to use each category equally often -- spreading out the ratings when the stimuli are actually close together, and pushing them together when the true values are far apart.

In a long series of studies, the same authors [1957] purportedly demonstrated that the results of magnitude estimation accurately represent sensory and nonsensory perceptions. With this procedure, a subject is given a standard stimulus and asked to provide a subjective ratio by assigning numbers to other stimuli "in proportion to" the number assigned to the standard case. Except in rare cases, the mean category ratings are linearly related to the logarithms of the arithmetic or geometric mean magnitude estimation judgments.

The present analysis extends a previous study [Patrick et al. 1973b] which described a linear relationship between magnitude estimation and category rating. That study could be criticized, however, because a standard (Well-Day) for magnitude estimation was assigned the value 1000 to represent the top extreme of the scale. The bounding of the scale, which is not standard in magnitude estimation, might have forced the linear relationship because it effectively made the procedure a form of category rating. The present study examines the relationship between category scaling and an unbounded form of magnitude estimation.

* See Appendix I.

METHOD

SUBJECTS AND CASE DESCRIPTIONS

The subjects were 65 volunteers from introductory psychology courses at San Diego State University with roughly equal proportions of males and females.

The items or case descriptions were drawn from a sample frame which includes all possible combinations of Function Levels and Symptom/Problem Complexes. Since age is necessary to provide a meaningful case description but contributes little to the variance of the ratings [Chen et al. 1973], one of four age groups was also identified with each item.

Thirty items were chosen to represent the full range of dysfunctions imposed on all types of patients by multiple symptoms and problems, including near well states. Each step in the scales of Mobility (MOB), Physical Activity (PAC), and Social Activity (SAC) was included at least once in the set of case descriptions. The first five items included a description of a completely well person and a person in a comatose state. These items familiarized the subjects with scale extremes. In sum, each item is a combination of an age group, one step from each of the three scales, and one symptom/problem complex (CPX), as follows:

School age (6-17),	(AGE)
Used car, bus or train as usual for age,	(MOB)
Walked with physical limitations,	(PAC)
Limited in amount or kind of school work,	(SAC)
Had pain, bleeding, itching or discharge from sexual organs.	(CPX)

The stimuli were presented as single pages in thirty item booklets. The content of the items within each booklet was identical and the order of the first five (warm-up) items was constant. The study items, however, were in a computer generated random order. Half the subjects were assigned to do category first, and the other half to do magnitude first, using different booklets for the two procedures. The subjects were run in groups of three to five students. Detailed instructions are available from the authors.

DATA CLEANING

A set of rules was created to eliminate judges who had apparently not paid close attention or did not understand the instructions. The rules eliminated subjects who rated two or more items above the well case (Item 1), or who assigned the well case a number less than 9 on the category scale, since the instructions specifically noted that 10 is for a well day. This process eliminated 11 subjects, leaving 54 subjects who produced a total of 3,240 usable observations.

RESULTS

As Stevens and Galanter initially demonstrated [1957], the arithmetic means of category rating (on the ordinate) exhibit a concave downward relation to the geometric means of magnitude estimation (on the abscissa). Figure 1 reveals this well known concave downward relation in our data. Thus the dimension of Well-Being behaves as a prothetic continuum. The product moment correlation for this relationship is .76.

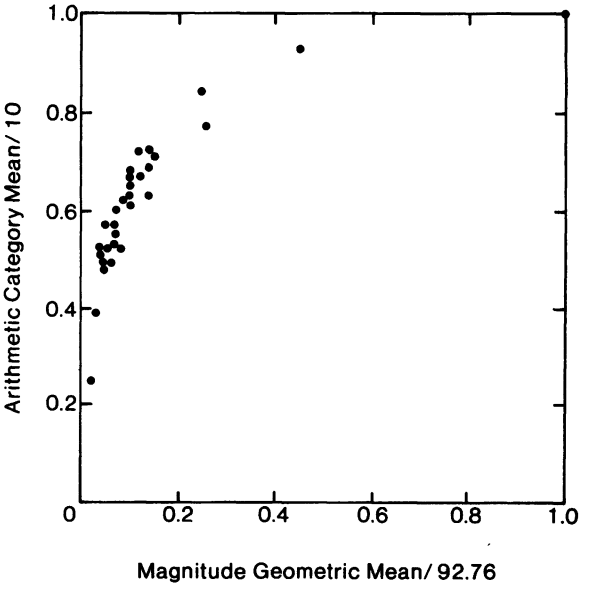


Fig. 1. Social preference ratings for 30 items representing states of dysfunction showing the classical concave downward relation between category rating and magnitude estimation.

Note that category and magnitude means have been transformed to a 0-1 scale. For category, all means were divided by 10, the top step of the scale. For magnitude, all geometric means were divided by the geometric mean weight assigned by the subjects to the Well-Day on the open-ended scale (92.76). This transforms the otherwise arbitrary numbers of the scaling procedures to a meaningful, comparable unit. The dramatic result is that all the magnitude measures of central tendency (median, arithmetic, and geometric means) compress the social preferences for almost all the items near the death state below 0.2. An item with a mean value of .72 using category rating, for example, receives a value of only .12 using magnitude estimation. If the relationship between the scaling methods is logarithmic, then a plot of category means against the logarithms of the magnitude geometric means should be approximately linear. Figure 2 demonstrates that the relationship, which has a product moment correlation of .96, is indeed approximately linear. The equation for this relation is:

$$C = .22 + .18 (\log M)$$

where

C is the arithmetic mean for the category rating for an item on a 0-1 scale, and

log M is the mean of logs (log of the geometric mean) for an item rated by magnitude estimation.

A similar comparison of the arithmetic category means versus the arithmetic magnitude means (and their logarithms) is not shown but was almost identical. This relation was apparent even when the confused and uncooperative subjects were not eliminated from the data set.

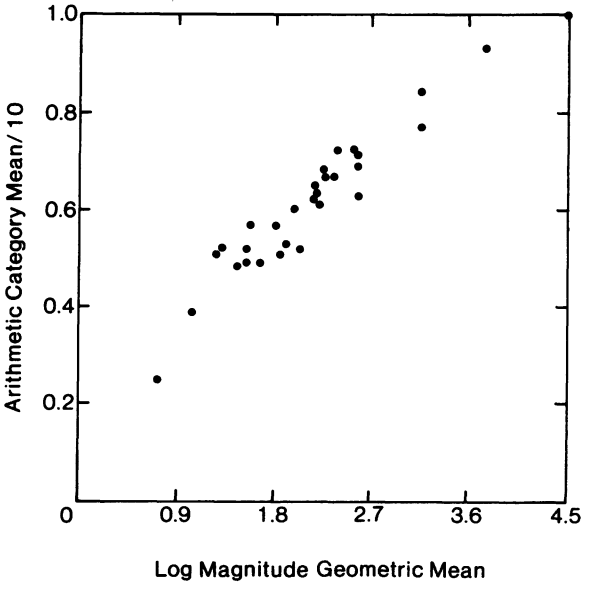


Fig. 2. Approximately linear relation of category arithmetic means to logarithms of the magnitude geometric means for items (data points) in Figure 1.

DISCUSSION

The results of this analysis confirm that social preferences or ratings of Well-Being behave as a prothetic continuum. If the continuum were metathetic, and the two methods had yielded identical results, scaling method would not be a concern for health index construction. We had originally used magnitude estimation because a fairly extensive literature held that it produced scale values with optimal properties [Stevens 1968]. The current results indicate that the scaling technique is now somewhat problematic and criteria must be established to select the best weights from those produced by different methods.

The needs of a Health Index per se are neutral in any disagreement between advocates of different scaling techniques. If magnitude estimation or a more complex technique were established as more valid, then any category data from a field survey could be transformed to yield the equivalent of the more desirable score by using a functional relation established in a careful laboratory study.

Our previous finding that the two methods agreed [Patrick et al. 1973b] was unexpected but gratifying. On logical grounds, it could be argued, either of the methods could produce an equal interval response scale. In closing the methodological loophole of the previous study, however, the non-linear relationship between the two sets of responses is now apparent. Both methods cannot be producing an equal-interval measure of preference. The results of this and subsequent research, on the other hand, do not support transforming field data from category rating to its magnitude counterpart. Figure 1 reveals that when the "ratios" from magnitude are transformed to a scale whose meaning can be interpreted directly and intuitively, the weights appear unreasonable.

Stevens was disappointed that most social scientists continued using category scales despite his repeated and vociferous objections. The major support for his magnitude estimation technique was the face validity argument that the subjects were instructed to assign their numbers "in proportion to" subjective ratios. This instruction is insufficient to establish the properties of the scale in theory [Krantz et al, 1971, p. 11], and several authors have noted Stevens' failure to provide empirical criteria for the properties that he claimed [Garner 1954; Torgerson 1960; Junge 1965; Anderson 1976].

Anderson has recently [1974, 1976] proposed a test for the equal interval property based on a simple analysis of variance. According to his functional measurement technique, the absence of a significant interaction effect in the analysis of variance establishes the equal interval property. Differences between preferences for two items which differ on only one attribute should be equal to the difference between two other items which have the same difference on that attribute. Experiments using functional measurement have demonstrated that category ratings meet this empirical criterion for the interval property while magnitude estimation does not [Anderson 1974, 1976; Weiss 1972, 1975].

Previous studies using our own case attributes have also demonstrated this absence of interaction [Patrick et al. 1973b]. One concern with the functional measurement test, which involves accepting the null hypothesis, is a possible false negative because of lack of power. In data from a probability sample of 900 San Diego households, however, this property was reconfirmed with approximately 100 subjects rating each item. Figure 3, showing data from four items, clearly demonstrates the parallelism exhibited by equal interval scales. For this analysis, both main effects were highly significant, while the F-ratio for the interaction was less than 1.0. This illustration is one from twelve similar analyses (to be reported) from balanced designs in the household survey, in which all possible interactions were non-significant.

An equal if not more important criterion for choosing between methods is whether the weights are consistent with ethical preferences [Harsanyi 1955] -- not the preferences that respondents would theoretically use for themselves, but the stated weights that they favor implementing for public policy. Our previous study [Patrick et al. 1973b] reported the only results in Health Index research (and, as far as we are aware, in social indicators research) to date using an equivalence technique which forces the trade-off among target population beneficiaries that are implied in the weighting scheme. Each of 12 comparisons among multiple groups, many composed of statewide health leaders and decisionmakers in health services, revealed non-significant differences between category rating and equivalence. The equivalence technique uses the natural social metric of the numbers of similar persons affected to provide a precisely adjustable response scale that is not biased by income, non-linearities in

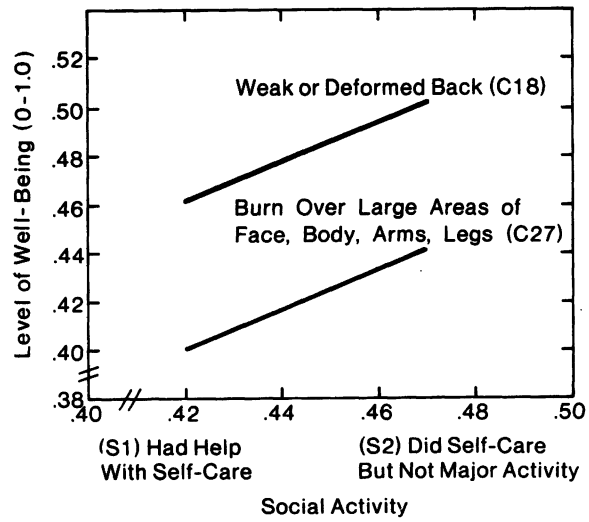


Fig. 3. Functional measurement test showing lack of interaction among items that differ by the same levels on each of two attributes (SAC and CPX) characteristic of equal interval scale.

the utility of money, prognostic or personal factors in time-tradeoffs, or aversions to gambling, which render suspect many other techniques used to measure utilities for health states. This consistency with the trade-offs implied in social choices is of major importance for the preference scale for which an equal interval measure is derived. So far as we are aware, this property has never been tested (much less demonstrated) for weights derived using magnitude estimation or any other technique in health index research.

The equal interval property may derive from the ease of administration of category scales, which means that single global ratings can be given to total case descriptions, thus considering the multiple dimensions of health states (including Symptom/Problem Complexes) jointly and simultaneously. This completely bypasses the need to rate separate attributes individually and later combine the ratings by arbitrary rules. Using such methods, the equal interval property of the total case score cannot be tested. The variance in our global ratings can be disaggregated and related to the case attributes using a simple linear model, which provides separate main-effect weights for Function Levels and Symptom/Problem Complexes and explains 96% of the variance [Chen et al. 1973].

In rejecting magnitude estimation, we do not necessarily reject all of Stevens' reservations about other attitude and preference measurement methods. In particular, we would agree that methods that unitize the dispersion in subject's responses -- just noticeable differences (jnd's) -- are not a desirable measurement unit. In our magnitude data, standard deviations increase with increasing desirability of the stimulus, but were roughly the same using the category method. The present evidence for the metric property of category responses is based, not on assumptions about stimulus or response dispersion, but on empirical tests and congruence with social choice metrics. Thus, with adequate warm-up and proper administration, category ratings apparently quantify subjective preferences directly, making later adjustments of category widths unnecessary [Blischke et al. 1975].

Torrance [1976] found that results from a time trade-off technique (of his invention) conformed to results from a version of the von Neumann-Morgenstern standard gamble better than results from category ratings. In view of its wide previous use in many circumstances, the difficulty that Torrance' subjects experienced with category rating is puzzling. This may have been because category rating was always administered at the very beginning of the interview as the first technique with very complex items.

Measures of internal reliability were not performed for category rating. In addition, the correspondence of the category rating and the time trade-off technique to the standard gamble were tested on only six items clustered near the middle of the scale. Such a study does not seem to justify Torrance's conclusion that category rating is inadequate for health index construction.

Unfortunately, even magnitude estimation does not offer the opportunity to incorporate the unbounded concept of "positive" mental health states in a health index. That limitation is due to the lack of an operational (observable or reportable) definition of the "positive" attribute to which utilities can be assigned. If it were possible to say that some persons had "positive" health attributes, while others did not, then the presence of the attribute(s) could be incorporated in the state of optimum function weighted 1.0, and the absence of the attribute(s) would simply be scored lower.

Although this would depress all values on the 0-1 scale, the scale would have been altered by incorporating a higher standard into the state of optimum function. The terms "positive" and "negative", in which much health and mental health jargon is couched, are totally arbitrary from an algebraic perspective. If a superior state of "positive" health were operationalized, it could be easily incorporated in the strategy of assigning consensus preferences to predefined states, regardless of the rating technique used. To the extent that such "positive" attributes affect current symptoms, problems and functioning, or prognoses, they are, of course, already reflected in the existing Index.

The demonstration of method differences should not lead to the conclusion that preference measures in health indexes are any more biased or unreliable than much health data that is currently published. All existing morbidity and mortality statistics have an implicit value component that is incompletely specified. In addition, all such specific statistics are upwardly biased as comprehensive health indicators because of the multiple other factors that they omit. The current life expectancy, for example, greatly overestimates the health status of a population because it includes no indication at all of the decreased quality of life.

Previous efforts to compensate for this lack has led to the publication of frankly subjective

data on scales such as "excellent/good/fair/poor" whose metric properties (despite high correlations with utilization, number of chronic conditions, etc.) have hardly been examined [USD/HEW, 1976, pp. 242-243]. Serious question can be raised, in fact, about even the ordinal properties of the scale [Kaplan et al. 1976], and yet its levels have frequently been treated as interval numbers in statistical models.

Almost any reasonable or approximate set of weights, applied to objectively verifiable states of function, would give a far more valid, reliable, and mathematically manipulable health indicator than aggregation of such crudely expressed individual opinions, for which the word "validity" has little if any meaning. As the science of function state classification and preference measurement progresses, actual values can be better approximated allowing consumer preferences to prevail over implicit, investigator assigned, or other ad hoc weighting procedures. Although arbitrarily weighted indexes can be shown to correlate highly with simplified versions of the IWB that omit variations at high levels of Well-Being -- the major source of IWB variance -- such numbers cannot be used to compute a meaningful weighted life expectancy which depends on precise 0-1 scale locations for the levels [Miles 1977]. Such an interpretation is essential to use a health index as a social indicator, as a tool for resource allocation, and even to quantify the health status impact of programs in evaluation research.

Anderson and his colleagues have demonstrated that the interval properties of the attribute ratings are preserved when the items include probabilities (prognoses) so the category ratings are consistent with the multiplicative properties required to treat them as expected values in decision models [Shanteau 1974, 1975; Anderson 1976]. These are precisely the properties required to compute the Weighted Life Expectancy and to estimate the output of a health program [Chen et al. 1975, Chen and Bush 1976].

In addition, all the preference distributions for the items rated were unimodal ("single-peaked"), which Black [1958] has demonstrated provides a sufficient condition to insure the transitivity of the resulting social preference function.

With the addition of the present results, our psychometric studies may be summarized as follows:

1. Preferences can be measured reliably ($r = 0.91$) from cross-validation studies using randomly created parallel forms of the procedure;
2. The values on the 0-1 scale possess equal-interval properties;
3. The category ratings are stable across different orders of testing and modes of test administration;
4. Linear statistical models accurately represent and predict ($R^2 \geq 0.96$) the mean and median global consumer ratings for individual case descriptions;
5. Age groups representing different phases of the life cycle in the case descriptions account for only about 1 percent of the variance in the preference ratings;
6. The preferences are generalizable across different social groups and their leaders, all of whom seem to share a consensus on the terminal values associated with the Function Levels; and
7. The category ratings are consistent with results from procedures designed to test for the ethical preferences implied in social choices, and have unimodal distributions which insure social transitivity.

With data now available, we will soon be able to examine the stability of the mean and median preferences over time.

This accumulation of evidence supports the notion that category ratings give social preference weights that are as nearly valid and with as desirable properties as any other techniques tried to date. Contrary to previous suggestions [Arrow 1963, Stevens 1966], magnitude estimation does not appear appropriate as a measurement method for a health status index and is probably inappropriate also for social indicators [Sellin and Wolfgang 1964] and other criteria of social choice.

REFERENCES

Anderson NH, Algebraic Models in Perception. In EC Carterette and MP Friedman, eds., HANDBOOK OF PERCEPTION, V. 2. NY: Academic Press, 1974, 215-291.

Anderson NH, How Functional Measurement Can Yield Validated Interval Scales of Mental Quantities. J APPL PSYCHOL 61:677-692, 1976.

Arrow KJ, SOCIAL CHOICE AND INDIVIDUAL VALUES. New Haven: Yale Univ. Press, 1963.

Black D, THE THEORY OF COMMITTEES AND ELECTIONS. Cambridge: University Press, 1958.

Blischke WR, Bush JW and Kaplan RM, A Successive Intervals Analysis of Social Preference Measures for a Health Status Index. HEALTH SERV RES 10(2):181-198, 1975.

Bush JW, Chen Milton and Patrick DL, Cost-Effectiveness Using a Health Status Index: Analysis of the New York State PKU Screening Program. In R Berg, ed., HEALTH STATUS INDEXES. Chicago: Hospital Research and Educational Trust, 1973, 172-208.

Chen Milton and Bush JW, Maximizing Health System Output with Political and Administrative Constraints Using Mathematical Programming. INQUIRY 13(3):215-227, Sept 1976.

Chen Milton, Bush JW and Patrick DL, Social Indicators for Health Planning and Policy Analysis. POLICY SCIENCES 6(1):71-89, 1975.

Chen Milton, Bush JW, Patrick DL and Blischke WR, Statistical Models of Social Preferences for Constructing a Health Status Index. Springfield, VA: National Technical Information Service, Pub. No. PB 236 155/8ST, 1973.

Garner WR, Context Effects and the Validity of Loudness Scales. J EXP PSYCHOL 48:218-224, 1954.

Harsanyi JC, Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. J POLIT ECON 63(3):309-321, 1955.

Junge K, SOME PROBLEMS OF MEASUREMENT IN PSYCHOPHYSICS: A THEORETICAL STUDY. Oslo, Norway: Scandinavian University Books, 1966.

Kaplan RM, Bush JW and Berry CC, Health Status: Types of Validity for an Index of Well-Being. HEALTH SERV RES 11(4):478-507, 1976.
Krantz DH, Luce RD, Suppes P and Tversky A, FOUNDATIONS OF MEASUREMENT. NY: Academic Press, 1971.
Miles DL, Health Care Evaluation Project Terminal Progress report, National Center for Health Services Research Grant 5 R01 HS 01568, July 1977.
Patrick DL, Bush JW and Chen Milton, Toward an Operational Definition of Health. J HEALTH SOC BEHAV 14(1):6-23, 1973a.
Patrick DL, Bush JW and Chen Milton, Methods for Measuring Levels of Well-Being for a Health Status Index. HEALTH SERV RES 8(3):228-245, 1973b.
Sellin T and Wolfgang ME, THE MEASUREMENT OF DELINQUENCY. NY: Wiley, 1964.
Shanteau JC, Component Processes in Risky Decision Making. J EXP PSYCHOL 103:680-691, 1974.
Shanteau JC, An Information Integration Analysis of Risky Decision Making. IN MF Kaplan and S Schwartz, eds., HUMAN JUDGMENT AND DECISION PROCESSES. NY: Academic Press, 1975.
Stevens SS, Issues in Psychophysical Measurement. PSYCHOL REV 78:426-450, 1971.
Stevens SS, A Metric for the Social Consensus. SCIENCE 151:530-541, Feb 1966.
Stevens SS, Perceptual Magnitude and Its Measurement. In EC Carterette and MP Friedman, eds., HANDBOOK OF PERCEPTION, V. 2. NY: Academic Press, 1974, 361-387.
Stevens SS, Ratio Scales of Opinion. In DK Whitla, ed., HANDBOOK OF MEASUREMENT AND ASSESSMENT IN BEHAVIORAL SCIENCES. Reading, Mass.: Addison-Wesley, 1968, 171-199.
Stevens SS and Galanter E, Ratio Scales and Category Scales for a Dozen Perceptual Continua. J EXP PSYCHOL 54:377-411, 1957.
Torgerson WS, Quantitative Judgment Scales. In Gulliksen and Messick, eds., PSYCHOLOGICAL SCALING: THEORY AND APPLICATIONS. NY: Wiley, 1960.
Torrance GW, Social Preferences for Health States: An Empirical Evaluation of Three Measurement Techniques. SOCIO-ECON PLANN SCI 10:129-136, 1976.
USDHEW, Public Health Service, HEALTH, UNITED STATES, 1975. Rockville, MD: USDHEW Pub. No. (HRA) 76-1232, 1976.
Weiss DJ, Averaging: An Empirical Validity Criterion for Magnitude Estimation. PERCEPTION AND PSYCHOPHYSICS 12:385-388, 1972.
Weiss DJ, Quantifying Private Events: A Functional Measurement Analysis of Equisection. PERCEPTION AND PSYCHOPHYSICS 17:351-357, 1975.

APPENDIX I: SCALES AND DEFINITIONS FOR CLASSIFICATION OF FUNCTION LEVELS*

MOBILITY

5 Drove car and used bus or train without help
4 Did not drive, or had help to use bus or train
3 In house
2 In hospital
1 In special care unit

PHYSICAL ACTIVITY

4 Walked without physical problems
3 Walked with physical limitations
2 Moved own wheelchair without help
1 In bed or chair

SOCIAL ACTIVITY

5 Did work, school or housework, and other activities
4 Did work, school, or housework, but other activities limited
3 Limited in amount or kind of work, school, or housework
2 Performed self-care, but not work, school, or housework
1 Had help with self-care activities

* Instruments for classification of persons into one and only one Function Level for multiple days available from the authors

INDICATORS OF COMMUNITY HEALTH STATUS FOR HEALTH PLANNING

Anne Mooney and N.W. Rives, Jr., University of Delaware

The National Health Planning and Resources Development Act of 1974 (P.L. 93-641) mandates Health Systems Agencies (HSAs) to assess the health status of residents of their planning areas. To satisfy this mandate, HSAs must be able to measure health status and to give such measures empirical content. There are barriers to pursuing effectively both activities. First, it is not clear what dimensions of health status planners should assess and monitor, although numerous measures have been developed. Second, given the capacity to measure health status, limitations on the statistical activities of HSAs suggest that most empirical analysis in support of health planning will be conducted with aggregate information routinely available.

This paper addresses four questions that raise fundamental considerations in the design and estimation of health status measures suitable for local health planning. What considerations are central to the design of health status measures? What segment, if any, of the technology of measuring health status can be adapted to local areas? How can HSAs reconcile the need to measure health status with the available resources and the restrictions placed on their use? Finally, what longer run developments in health status measurement are desirable, and what can be done to facilitate their achievement?

Dimensions of Health Status Measures

The purpose of measuring community health status is to summarize the health of human populations. Measures can be developed theoretically not only for physical health, but also for mental health and social functioning. The manner in which health status is measured clearly depends upon the manner in which health is defined. As Lerner [9] has pointed out, this is a particularly difficult problem because health is a multidimensional characteristic.

Three general considerations are fundamental to the design of health status measures for health planning. These considerations are: (1) measurement of health-related conditions, including both conceptual and statistical issues; (2) determination of the size and structure of the population at risk in the geographical area for which estimates are being made; and (3) specification of intervention approaches with respect to prevention and treatment of conditions that adversely affect health status.

Conditions

In developing health status measures for health planning, particular attention must be given to the identification and classification of health-related conditions, including functional conditions not derivative of any unique disorder or illness in clinical terms. Certain conditions are interesting simply because it is known that medical care could prevent their occurrence or control their effects, including death. In particular communities, recognition of high prevalence rates for preventable or controllable conditions gives these conditions the visibility to encourage efforts to reduce their prevalence, possibly by reducing incidence. Health status measures that relate to these conditions would be especially valuable tools for the identification of problems and the evaluation of intervention programs.

Populations at Risk

Determination of populations at risk is frequently easier from a conceptual standpoint than an empirical one, especially at lower geographical levels. Delineation of populations at risk in demographic terms facilitates the estimation process, at least to some extent, owing to the generally wider availability of demographic statistics. The problem can become quite complex when the population has been conceived in non-demographic terms. Persons at risk of contracting an illness against which a preventive inoculation is available, such as tetanus or diptheria, are generally those who have not been exposed to the inoculation. Depending upon the availability of information about levels of immunization in the population, health planners may or may not have a useful screening device for determining populations at risk.

Intervention

Health status measures will be particularly useful for health planning when they relate to planning, implementation, and evaluation of health services. Activities of planners with respect to health status can usefully be categorized in terms from preventive medicine. Thus, they may be oriented to primary prevention, the reduction of the incidence of a condition. They may also be oriented toward secondary prevention, reducing the incidence of complications of an illness or injury, or tertiary prevention, reducing the levels of residual disability or other long-term effects, given that an illness or injury has occurred. It follows that health status measures are needed that reflect different effects, ranging from changes in the incidence of a condition through changes in rates of residual disability.

Approaches to Design of Health Status Measures

Siegmann [16] suggests that current strategies for health status measurement are divided between the development of disciplinary research indicators and policy research indicators. The former consist of measures designed to summarize population health levels, possibly to provide a basis for allocating health resources. The latter consist of measures designed to identify the effect on health status of particular health

programs; measures in this category are meant to support evaluation research. Siegmann concludes her assessment of the technology of health status measurement by suggesting that disciplinary research indicators and policy research indicators might serve as points of departure for the development of an epidemiology of health.

The measurement strategy of greater interest to HSAs, following the Siegmann typology, is the development of policy research indicators. If an important function of health status measures in health planning is to facilitate an improved understanding of the relation of health status to the allocation of resources to health programs, then policy research indicators at the local level must be thoroughly investigated and fully exploited. But the technology of health status measurement, and specifically the technology adaptable to local areas, simply has not reached the point where health planners can expect to have access to a wide range of useful measures. Furthermore, until the data systems to support newer measurement schemes have been fully implemented, and until the feasibility and validity of these schemes have been demonstrated, local health planners will simply have to resort to less precise, and possibly less useful, measures of health status. [4]

Current Measurement Possibilities

This section of the paper identifies possibilities for measuring health status which do not seem to exceed the legal, institutional, and technical constraints under which local health planning must operate. An examination of approaches one might reasonably expect HSAs to employ in the measurement of community health status suggests the following classification scheme: (1) the use of mortality data to summarize the risk of dying in an existing population; (2) the use of mortality data to infer conditions of morbidity; (3) the use of morbidity data to measure the incidence and prevalence of specific conditions; (4) the use of utilization and treatment data to measure the absolute frequency of specific conditions in a selected segment of an existing population; (5) the use of indicator measures which are known or suspected covariates of health status; and (6) the use of synthetic measures of health status. There are precedents for each approach, and each approach has specific advantages and disadvantages.

Risk of Death

The use of mortality data to summarize the risk of dying is the traditional approach of demography to measurement of health status. The most refined mode of analysis is the life table, an analytical technique for expressing mortality in terms of probabilities. Concepts and methods surrounding the construction of life tables are well developed, and the life table provides measures of mortality which are easily interpreted. Unfortunately, even abridged life tables for relatively short time intervals cannot be constructed for all levels of characteristic and geographical detail. Statistical standards may preclude the estimation of life table functions for many health planning areas. Furthermore, the speed with which local health statistics are processed for planning applications may present problems.

Morbidity Inferred from Mortality

Since the risk of dying represents only a single dimension of health status, it is necessary to consider other approaches. The use of mortality data to infer morbidity recognizes another element of the classic relation of vital events and health status. Several recent studies illustrate the potential of this approach [3,6,15]. All three studies are concerned with the relation of morbidity and mortality, and all three have local adaptability due to their primary reliance on vital statistics, but their use at the HSA level cannot go unquestioned, because inferences concerning morbidity from mortality data are not as direct as one would wish.

Incidence and Prevalence of Morbidity

The use of morbidity data to measure the incidence and prevalence of specific conditions follows the classic tradition of epidemiology. Data sufficient to construct incidence and prevalence measures can be obtained from either surveys or reporting systems. Measures based on morbidity data represent an improvement over inferential mortality methods with respect to the directness of measurement, but the approach is not without limitation. Survey-based measures are subject to all of the usual problems of sample estimators, including random and systematic errors. Reporting-system-based measures are subject to problems of completeness of coverage and of response error.

Furthermore, given the comparative advantage of surveys as a method of data collection at higher geographical levels, one is not likely to find much in the way of subnational sample data on morbidity. Greater geographical detail will almost certainly accompany measures derived from reporting systems, but the larger bureaucracy required to operate a reporting system, particularly a national system, will invariably restrict the flow of information, thus reducing the processing speed and the timeliness of the information. Both surveys and reporting systems must operate under the present uncertainties surrounding privacy and confidentiality, and these factors add to the general problem of data access for organizations like HSAs.

A final limitation concerns the conceptualization of morbidity. If health status implies something beyond a simple assessment of the presence of illness, then technically astute HSAs may become disenchanted with the performance of such common measures as "disability days" and "work-loss days [19]." A recent review of newer sociomedical indicators identifies at least four that give promise of meeting criteria of reliability, sensitivity, and applicability to community populations [17].

Morbidity Inferred from Utilization and Treatment

The fourth measurement approach is based on the notion that populations in treatment and populations utilizing medical services may resemble populations at risk of various conditions. Data on medical services utilization from both consumer and provider perspectives are highly standardized, owing to the increasing use by the National Center for Health Statistics (NCHS) and other agencies of abstracting systems and "minimum basic data sets."

Much of the information on utilization of health services is derived from household interviews conducted at the national level. A prime example is the current estimates program of the Health Interview Survey (HIS), which provides national data on such variables as health care expenditures and physician visits. The NCHS Hospital Discharge Survey produces annual data on the utilization of inpatient services at short-stay hospitals, and the recently implemented National Ambulatory Medical Care Survey (NAMCS) produces both medical and nonmedical statistics on physician episodes obtained from a national sample of cooperating physicians.

The problems of using utilization and treatment data to measure health status are considerable [25]. The expected number of persons in treatment is theoretically the product of the population at risk of a specific condition, the unconditional risk of this condition, and the probability of medical service utilization given the presence of the condition. The expected number of persons with a specific condition in the larger population can then be computed either by multiplying the population at risk by the unconditional prevalence rate or by dividing the number of persons in treatment by the propensity of persons who get sick to seek professional assistance. Unfortunately, although both approaches are theoretically correct, neither is easily given empirical content. The former approach is impractical because neither the prevalence rate nor the population at risk are known with precision, and the latter approach is impractical because reliable estimates are never available on the completeness of coverage of treatment programs. Furthermore, since persons entering treatment are not drawn at random from the population with the condition of interest, one cannot expect treatment statistics to portray accurately the true nature of the problem.

A second limitation of utilization data involves both the amount of geographical detail one can expect to obtain and the speed with which the data are processed. Most of the better known utilization data sets are produced by NCHS at the national level and cannot be disaggregated to provide even divisional and state detail without some loss in either precision or characteristic detail; it should be noted that the survey design does permit publication of regional and selected SMSA estimates from the HIS. Altogether, these limitations would seriously restrict the use of such data by HSAs, even if processing time were not a problem.

Treatment data do not suffer as much from problems of geographical detail, although the speed with which these data are processed varies significantly from place to place; some reporting systems are still not automated, while others change too rapidly to become efficient under any one system design. It is in this light that the advantage of data produced at the national level can now be seen. Despite problems of geographical detail and processing time, national data systems are able to capitalize on the greater technical competence of system personnel, which implies better design procedures and more rigidly enforced standards of quality control. Few states can produce utilization or treatment data sets of comparable stature, although a number of states have developed statistical systems in selected areas that produce reliable information on a regular basis, with appropriate characteristic and geographical detail. The comparative advantage of national utilization data sets from the standpoint of technical refinement makes them prime candidates for synthetic estimation in health planning areas.

Social Indicators of Health Status

The use of indicator measures which are known or suspected covariates of health status represents the conceptual point of departure for the construction of ecological models of health status. There are several precedents for this approach. First, the National Institute of Mental Health has developed the Mental Health Demographic Profile System (MHDPS) [23]. This system is designed to facilitate the estimation of indicator measures and the construction of indicator models for local mental health needs. The demographic data items (social indicators) were selected from 1970 census data to permit delineation of meaningful social areas and subsequent inferences concerning community health status. A second precedent is the Social and Health Indicators System piloted by the Bureau of the Census, Census Use Study, in Atlanta and Los Angeles [24]. The objectives of this system are quite similar to those of the MHDPS, and both systems are census-based, although the Atlanta and Los Angeles programs were designed to accept local data.

There are several important limitations to the use of indicator measures to describe health status. First, almost all of the measures now available are derivative of the decennial census. The utility of census-based measures declines rapidly as the census becomes less recent in time, despite efforts to update census information; the problem is almost certainly greater at lower geographical levels. The mid-decade census, to begin in 1985, should reduce considerably the problems associated with intercensal estimation, even for small areas.

A second limitation is the compatability of census data and data locally produced. The failure of the decennial census to include even basic information on health subjects means that many HSAs will find a need to supplement census data with data locally produced on subject items not found on a census schedule. The need to combine

local data with census data raises questions of the compatability of geographical areas and the comparability of definitions for common terms. One approach to this problem is to manipulate local data to fit the census framework, but this may not always be the best solution. Census regions are designed primarily to facilitate data collection. Whether areas like census tracts and minor civil divisions can be considered meaningful ecological units of analysis, or whether one can build meaningful units by aggregating these areas, is quite another matter.

A final limitation of most indicator measures is their essentially static nature. Aigner and Simon [1] have shown that cross-section estimators behave quite differently from their time-series counterparts, and this cautions dynamic inferences from static indicator models. Since many statistical systems supporting health services research have only recently come into existence, there is not a wealth of time-series information with which to study the behavior of indicator measures over time.

Synthetic Estimates

Synthetic estimation methods have considerable appeal for HSAs because synthetic estimates are conceptually and mechanically simple and potentially useful. Synthetic estimates are indirect measures, not in the sense that they are covariates of some condition, but rather in the sense that they measure the condition without direct observation. A recent report by NCHS [21] on state estimates of disability and utilization of medical services provides an example.

Estimation through direct observation at the state level is not permitted by the current design of the National Health Survey. The HIS presently generates national data on disability and medical services utilization, however, and if one were willing to assume that both factors are related to such variables as age, sex, race, and family income, then differences among states with respect to disability and medical services utilization might simply reflect differences with respect to the control variables. This implies, of course, that the age-sex-race-income specific disability and utilization rates, observed in the national survey, are constant among states. Whether this is a strong assumption, or even a defensible one, remains an empirical question, dependent largely upon the situation under study. If one is willing to make the assumption, then state estimates are simply a matter of weighting the schedules of survey rates by the age-sex-race-income specific state populations.

Although synthetic estimates are potentially useful in areas where circumstances preclude other forms of measurement, one must be extremely careful in the interpretation of such estimates, because their statistical properties are not well known [5,11]. Only further empirical research can establish the extent to which the "homogeneity-of-risk" assumption underlying synthetic estimation has any practical validity.

Directions for Further Research

Implied in the previous sections is the conclusion that current efforts at estimating health status of populations for local health planning are constrained by limitations of: (1) conceptual and statistical aspects of health status measurement; (2) quantity and quality of data available with the appropriate characteristic and geographical detail, and the requisite frequency; and (3) dynamic models of health status. The remainder of this paper is devoted to consideration of a series of issues and recommendations with respect to measures, data, and models.

Multiple Measures

We believe that global measures of health status are of limited use to local health planners; even sets of comparative summary measures for subareas within an HSA may be incomplete or misleading. Rather, a variety of measures focused on components of health status needs to be devised and used to estimate levels of health in sectors of the population that are at risk of particular conditions. Some of these measures are easy to estimate from data currently available on a regular basis; others will have to be invented and the data collected to use them. Imagination and experimentation will be required to determine the dimensions and combinations of population, condition, and intervention outcomes that are of greatest interest to measure.

Some of the problems associated with summary measures of mortality for small populations can be overcome, as Kleinman [22] suggests, by computing separate parallel measures for segments of the population divided by age: infants, and ages 1-34, 35-64, and 65 and over. A possible refinement in some larger communities would be to take infants, children aged 1-14, then 15-year age groups from 15 to 74, with an open category 75 and over. These seven groups seem to capture some important differences by age in risk of conditions important to planning. Indeed, a set of preventive measures has recently been proposed for each of seven age groups, starting with the fetus and mother and extending through the life cycle to older adults [2]. If some of the services recommended for each of these groups were to be designated for special programming, appropriate health status measures for the specific conditions of interest and age-graded populations at risk would be in order.

In addition to age, sex and race are standard factors by which rates are often adjusted or specified. Given an age-sex-race breakdown of the population, what measures should health planners try to estimate for these groups? For mortality in populations of 25,000 or more, Kleinman [7] argues for years-of-life-lost measures, at least for the population under age 65 or 70, as displaying the best combination of statistical properties, including stability, availability of data, and sensitivity to conditions along the age span.

For any given level of statistical precision, the requirements of characteristic, geographical, and temporal detail will tend to conflict. This fundamental constraint must be confronted, although it will no doubt be dealt with differently, depending on the use for which estimates are intended. As we see it, there is no escape for health planners from these dilemmas, only a set of solutions of varying utility according to the situation.

Comparability of Estimates Among HSAs

HSAs will probably develop unique sets of health status estimates based on the most relevant combinations of conditions, populations and interventions. For other purposes, however, such as comparisons among areas with respect to particular dimensions of health status or for allocation of program funds with respect to need, some minimum basic set of measures should be estimated for all HSAs. It would be helpful, therefore, if national standards were promulgated in the near future, recommending appropriate definitions, data sources, and computations for a variety of measures, some of which would be calculated for all HSAs and some of which would not. But each would be comparable from one area to another within the set of areas for which it was available. The standards and estimates presently provided for infant mortality and total mortality in the Statistical Notes for Health Planners [20,22] provide an excellent start in the direction we are suggesting. At the same time as we advocate national standards and comparability, we would also encourage HSAs to be the sources of innovations in small-area measurement and estimation that could be adapted for use throughout the country.

Synthetic Estimates

Many measures of health status ultimately depend on observations drawn from a survey. Because of the strong comparative advantage of the survey as a scheme for data collection in large populations, methods of inferring from larger to smaller areas are of utmost importance. In addition, the costs associated with data collection and analysis, in general, and with surveys, in particular, will make it necessary to limit the scope of direct observation and to rely as much as possible on indirect methods of estimation. Therefore, synthetic estimation deserves particular attention.

Some important questions for consideration, as these techniques are developed and applied, are:

1. What are the appropriate characteristics on which to base synthetic estimates of various measures? To what degree do geographical units vary with respect to these characteristics? What is the effect of this variation on the synthetic estimates compared with estimates based on the assumption that a general population rate at one level of aggregation applies to all lower levels?
2. In the case of synthetic estimates, fixing the level of characteristic detail and allowing the geographical level of aggregation to vary assumes that the risk of any condition for any population group does not vary geographically. How geographically homogeneous is the risk of health-related conditions of concern to planners?
3. What are the optimal geographical units for development of synthetic estimates? At what level should surveys or other direct observations be made and to what level can they validly be projected, using synthesizing techniques?

Some evaluations of synthetic estimation procedures at the national level have already been made [5,14,21]. Additional work needs to be done at the local level.

Ecologically Homogeneous Areas

One response to the problem of balancing needs for geographical and characteristic detail is to work with areas relatively homogeneous in population characteristics. In general, designation of such areas makes the tasks of measurement and planning easier, since one dimension of variation is eliminated. Therefore, the development of procedures in social area analysis and factorial ecology for the designation of ecologically homogeneous areas would be an asset to local health planning [12].

Several studies have pointed out the importance of looking at small-area variations in population characteristics, prevalence of conditions, and availability and utilization of services [26]. Others are taking advantage of mortality data available by census tract to investigate the extent to which mortality risk varies by tract, given variation in tract population characteristics [10].

Social Indicator Models

Since health planning is ultimately concerned with resource allocation, presumably with the objective of improving health status, more attention will need to be paid to covariates of health status and to the conditions under which the status of populations changes. Further exploration is needed of the extent to which covariates--especially those for which estimates are regularly made at appropriate levels of geographical detail--can be used to make estimates of health status in the absence of surveys, registration, or reporting systems. In cases where it is expensive and/or difficult to obtain direct observations, then a good indicator, taken from a social indicator model, may actually provide as much information as a survey, but at a much lower expenditure of funds and effort. To the extent that this approach proves to be appropriate and practical, it can be related to periodic surveys for benchmarking purposes. An indicator scheme for states and local areas is a desirable long-run objective of this enterprise.

Only a bare beginning has been made, however, in establishing the characteristics of communities associated with different levels of health and in understanding the factors associated with changes in health levels. Social indicator models, both cross-sectional and time-series, need to be developed to provide a context for interpreting estimates of health status and a basis for appreciating the interrelation of health and non-health variables. Recent attempts to develop dynamic social indicator models for selected health status measures include the work of Land and Felson [8] and of Brenner [18].

The approaches discussed above commend themselves on various grounds as productive directions for developing and implementing health status measurement schemes for local health planning. Other steps that we believe will facilitate the research and development process are (1) the re-orientation of existing data collection efforts toward providing statistics for local areas and (2) the reinforcement of interagency organizational linkages, both vertically from the federal to the state and local levels and horizontally among federal agencies [13].

REFERENCES

[1] Aigner, D.J., and Simon, J.L. "A specification bias interpretation of cross-section vs. time-series parameter estimates." West. Econ. J. 8: 144, 1970.

[2] Breslow, L. "Prevention in personal health services."In U.S.P.H.S., Health Resources Admin., Papers on the National Health Guidelines: The Priorities of Section 1502, pp. 102-119, 1977.

[3] Donahue, C.L. et al. "The use of vital events as a data source in the planning of maternity and newborn services." Presented at the annual meeting, Amer. Public Health Assn., 1976.

[4] Elinson, J. "Insensitive health statistics and the dilemma of HSAs." Amer. J. Public Health 67: 417, 1977.

[5] Gonzalez, M.E. "Use and evaluation of synthetic estimates." In U.S. Bureau of the Census, Census Tract Papers, Series GE-40, No. 10, Statistical Methodology of Revenue Sharing and Related Estimate Studies, pp. 46-50, 1974.

[6] Haberman, P.W. and Baden, M.M. "Drinking, drugs, and death." Internat. J. Addictions 9: 761, 1974.

[7] Kleinman, J.C. "A new look at mortality indexes with emphasis on small area estimation." Proceedings, Amer. Statistical Assn., Social Statistics Section, 1976, pp. 485-490.

[8] Land, K.C. and Felson, M. "A dynamic macro social indicator model of changes in marriage, family, and population in the United States: 1947-1974." Social Science Res., Dec., 1977 (forthcoming).

[9] Lerner, M. "Conceptualization of health and social well-being." In Health Status Indexes, R.L. Berg, ed., pp. 1-6. Chicago: Hospital Research and Educational Trust, 1973.

[10] Lerner, M. and Stutz, R.N. "Mortality differentials among socioeconomic strata in Baltimore, 1960 and 1973." Proceedings, Amer. Stat. Assn., Soc. Stat. Sec., 1975, pp. 517-22.

[11] Mooney, A. Estimation of the Prevalence of Alcohol Abuse in Pennsylvania Counties. Newark: Univ. of Delaware, College of Urban Affairs and Public Policy, 1977.

[12] Plessas, D.J. and Carpenter, E.S. "Empirical designation of health service areas." Health Services Res. 10: 333, 1975.

[13] Rice, D.P. "The role of statistics in the development of health care policy." Amer. Statistician 31: 101, 1977.

[14] Rives, N.W., Jr. "Estimating 1970 census coverage for metropolitan areas." South Atlantic Urban Stud. 2: 55, 1977.

[15] Rutstein, D.D. et al. "Measuring the quality of medical care: A clinical method." N. E. J. Med. 294: 582, 1976.

[16] Siegmann, A.E. "Readiness of sociomedical sciences to measure health status." In Health Goals and Health Indicators, J. Elinson, A. Mooney, and A.E. Siegmann, eds. Washington: AAAS, 1977 (in press).

[17] Siegmann, A.E. and Elinson, J. "Newer sociomedical health indicators." Med. Care 15 (May, supp.): 84, 1977.

[18] U.S. Congress, Jt. Econ. Comm., Achieving the Goals of the Employment Act of 1946-Thirtieth Anniversary Review. Vol. I-Employment. Paper No. 5, by M.H. Brenner.

[19] U.S.N.C.H.S. "Disability components for an index of health," by D.F. Sullivan." Vital and Health Statistics, PHS Pub. no. 1000, Series 2, No. 42, 1971.

[20] U.S.N.C.H.S. "Infant mortality," by J.C. Kleinman. Statistical Notes for Health Planners, no. 2, 1976.

[21] U.S.N.C.H.S. State Estimates of Disability and Utilization of Medical Services: United States, 1969-71, 1977.

[22] U.S.N.C.H.S. "Mortality," by J.C. Kleinman. Statistical Notes for Health Planners, no.3, 1977.

[23] U.S.N.C.H.S. "Mental health demographic profile for health services planning," by E.S. Pollack. Statistical Notes for Health Planners, no. 4, 1977.

[24] U.S. Office of Economic Opportunity and U.S. Bureau of the Census. Social and Health Indicators System: Los Angeles. Census Use Study, 1973.

[25] Wennberg, J.E. "Using localized population based data in evaluating planning programs." In U.S.P.H.S., Health Resources Admin., Papers on the National Health Guidelines: The Priorities of Section 1502, pp. 78-91, 1977.

[26] Wennberg, J. and A. Gittelsohn. "Small area variations in health care delivery." Science 182: 1102, 1973.

A MICROSIMULATION MODEL FOR HEALTH STATUS PROJECTIONS

George C. Myers
Alfred M. Pitts
Roger Hillson
Eric Stallard

Duke University

INTRODUCTION

The development of a microsimulation model that captures the sickness-death process has been a central focus of several research projects undertaken in the past few years by the Center for Demographic Studies at Duke University. Implicit in this development has been the intent that such a model provides a suitable framework for producing national population projections that contain not only age, sex and race specificity, but also estimates of the health status of these population sub-groups. Moreover, the model affords the researcher an experimental tool for assessing the changes that may be experienced in the incidence of specific diseases and the probabilities of dying (or surviving) from the diseases. In this paper, the main features of the model are described and two applications are discussed.

The projection of the health status of future national populations is clearly of great importance in anticipating the demands that will arise in medical manpower, facilities and fiscal support systems. Moreover, it is likely that existing differentials in health status by age, sex, race, socioeconomic and other social characteristics will persist to varying degrees in the future. Thus, these health status projections must be disaggregated for important segments of the population if they are to be responsive to the growing concerns with health policy formulations and health service program planning.

Demographic specialists concerned with national population projections have largely ignored considerations of health status change and disease processes in examining the differentiation of mortality risk in human populations. Although the compositional variables that are typically used to explain changes in mortality might be regarded as major, implicit correlates of morbidity, these projection models tend to view changes in aggregate survivorship as arising *pari passu* with changes in sociodemographic composition alone. The processes of disease onsets, virulence, and recovery, through which the effects of compositional change are transmitted, have not been a traditional projection concern.

MODEL FOR PROJECTING HEALTH STATUS

The health status model that has been developed is quite straightforward in its conceptualization. The general approach is similar to that of the POPSIM simulation model developed by Horvitz, *et. al.*(1) although there are a number of important differences that cannot be discussed here. The model constructs for each member of a random sample of the United States population, a health status history by stochastically exposing the individual to a set of health status change probabilities. Although these probabilities are allowed to vary cross-sectionally by age, race, and current health status, in effect, we are projecting the future health status of the population to the year 2000 on the assumption that the probabilities observed in 1970 remain unchanged. After repeating these simulations for every member of the sample sufficiently often to assure that the relative age distributions of deaths and population are free of random experimental error, the results for the sample are extrapolated to the national population.

Health statuses are divided into three acute and eight chronic conditions; these conditions are indicated in the tables that follow. In addition, individuals also can die from external causes (e.g., accidents). However, the model does not capture the temporary or permanent disability that might result from a non-fatal external incident, though nothing in the structure of the model precludes such a refinement.

Paper presented at the Annual Meetings of the American Statistical Association held August 15-18, 1977, in Chicago, Illinois. Support for this research was from a contract from the Adult Development and Aging Branch of the National Institute of Child Health and Human Development, No. 1-HD-2-267, and a contract from the National Academy of Sciences' Committee on Health Care Resources in the Veterans Administration.

The model projects a life history for each member of the simulation sample by establishing the time of disease onsets and deaths, on the basis of a number of simplifying assumptions.

First, the model assumes that the onset or presence of any one health status condition is not correlated with the onset or presence of any other condition or group of conditions (except via an indirect path through mortality). Therefore, health status changes occur as independent events.

Second, the onset of any acute condition has a fixed initial duration of three months, during which time the afflicted individual experiences a risk of dying from that condition. Persons who experience an acute onset are allowed to "recontract" that condition prior to the termination of the three-month onset period. The effect of recontracting the disease is to extend the recovery date of the condition by another three months from the month of recontract; during this second onset interval, the condition-specific mortality risk remains at the same level at which it was during the initial onset period.

Third, although acute illnesses constitute transient health statuses, the onset of a chronic condition results in a permanent independent increase in the risk of dying, although the amount of the increase is allowed to vary as the person ages forward from the time of onset. In other words, persons who experience a chronic onset never recover, in the sense that they never experience a remission of the rise in mortality risk that results from the onset of a chronic condition at any time during their lifetimes.

Finally, the experience of an "external incident" has no effect on the individual except to bring about an instantaneous, momentary rise in one's risk of dying. That is to say, that the model explicitly concerns itself only with those external events that are immediately lethal.

Mortality, except from an external cause, is handled by the model as an age and health status contingent process. Individuals die according to a set of independent probabilities of death from each of the health conditions that they are experiencing at a given moment. Persons will experience a continuous rise in their death risk as they accumulate more and more conditions.

Perhaps the most problematic aspect of the above formulation is that it ignores considerations of disease latency and recovery. We would argue, however, that the model's lack of any explicit representation of latency or recovery is less relevant to the task at hand--namely, that of assuring consistency between the model's health status projections and its projections of overall mortality--if one uses a broad definition of "recovery" or latency that includes any elements of the disease process that do not directly influence mortality risk. If the remission of a condition does not carry with it a reduction in the risk of dying, the individual cannot be considered as having "recovered" from the standpoint of the model. The mere resumption of normal activity--if, for example, this is what one means by "recovery"--has no relevance in the present context.

Thus the model embodies an "ever experienced" notion of morbidity, unlike the "currently manifest" notion implicit in most point-prevalence measures of health status. In a sense, it accordingly captures the conceptualization of disease latency that characterizes the standard multiple decrement, cause-elimination life table, in which the survivorship column (l_x) of the life table is segregated into subpopulations of individuals who are ultimately and inevitably destined to die from a specific cause--"marked for life", so to speak, by the onset of a specific condition.

THE MICROSIMULATION FRAMEWORK OF THE MODEL

In its present application, this formulation is operationalized by drawing upon its implications for the timing of changes in health status and death. Simply stated, each individual's life span is segmented into a series of age intervals over each of which the individual is assumed to be at constant risk of experiencing a given event. A stochastic procedure then is applied to determine whether the individual survives through each consecutive interval without experiencing the event. When the interval is finally reached in which the event is projected to occur, the model assigns that event to a precise time point within the interval.

Derivation of the appropriate timing functions is relatively straightforward. Define,

- t' as the time at which an event is projected to occur
- $t(i)$ as the number of years in the i-th age interval

$p(i)$ as the probability that an event will not occur in the i-th age interval

r as a number that is randomly selected from a rectangular distribution of numbers between 0 and 1

If one assumes that the hazard of an event remains constant over all age categories, it can be shown that,

$$(a) \qquad p(i)^{t'} = r$$

Appropriately transposing t' in (a) yields the expression,

$$(b) \qquad t' = \frac{\ln r}{\ln p(i)}$$

Expression (b) specifies the time at which the event is expected (projected) to occur, if the hazard of its occurrence remains constant over and across all age intervals.

To derive an estimated failure time from hazards that vary across age intervals, consider first the case in which

$$t' > t(1) \text{ and } t' < [t(1) + t(2)]$$

that is, the case in which an event is projected to occur within the second age interval. It can be demonstrated that expression (a) and the randomness of r together imply that,

$$(c) \qquad r = P(1)^{t(1)} \cdot p_{(2)}^{[t'-t(1)]}$$

Appropriately transposing (c) yields the expression,

$$(d) \qquad t' = t(1) + \frac{\ln r}{\ln p(2)} - \frac{\ln \left[p(1)^{t(1)}\right]}{\ln p(2)}$$

which is the precise time within the second age interval at which the event is projected to occur. This result can be further expanded to yield the general timing expression,

$$(e) \quad t' = \frac{\ln r}{\ln p(j)} + \sum_{i=1}^{j} \left[t(i-1) - \frac{\ln p(i-1)^{t(i-1)}}{\ln p(i)} \right]$$

where,

$$t' < \sum_{i=1}^{j} t(i) \text{ and } t' > \sum_{i=1}^{j-1} t(i)$$

and j denotes the age interval within which the event is projected to occur.

Expression (e) is thus used in the model to determine the timing of disease onsets and deaths. During risk intervals having constant occurrence hazards, it in effect assumes a Poisson process for the event.

Because the model assumes the onset of any given condition to occur independently of the presence or onset of all and any other conditions, one can use (e) to separately project the expected onset times of each of the eleven conditions. Later, when age at death is determined, onsets that are projected to occur after death are "erased". With regard to chronic onsets, recall that the model postulates non-recovery. Hence, only one random number needs to be generated to determine for the entire lifespan the projected period during which the individual will have a given condition not already being experienced at the start of the projection. The projection of acute onsets varies in detail, though not in principle, from that of chronic onsets.

Once the health history is available, it is then possible to determine the precise age at which the individual will die, using a single cast against the general timing function (e). The relative event probability, $p(i)$, is the joint probability of surviving all of the conditions extant in the i-th risk-homogeneous interval. Since disease onsets can occur at precise time points within age intervals over which the hazard of disease onset is constant, the mortality risk homogeneous intervals will be bounded not only by the age boundaries across which changes in the condition-specific probabilities of dying occur, but also by the time points when the individual experiences the onset of additional conditions.

APPLICATIONS

Two main applications of the model have been made to date, each intended for specific projection purposes that made the final output somewhat different, but each involving the general strategy as previously outlined. The first entailed a national projection of the elderly population (65 years of age and older) to the year 2000 by age, sex, race and marital status categories and twelve health condition states, as noted previously. In addition, of course, deaths could arise from accidents in the model as well as deaths from any of the eleven disease conditions. The elderly population segment is therefore decremented by deaths arising from the sickness to death process, and it is incremented by persons turning 65 years of age.

The second application involved projection to 2000 of a national population; in this case the male veterans entitled

to benefits from the Veterans Administration. These projections were specific by age and race. These projections took into account that the veterans population is continuously being incremented by discharges from the military by means of estimates provided by the Department of Defense.

Parameter Estimation

The most serious obstacle to the success of any model is the degree to which adequate data can be derived to apply and test it on actual populations. The numerous decisions regarding specification of the parameters made in these two applications relate fundamentally to data considerations, but a full discussion of these matters is clearly not appropriate here.

In brief, disease incidence and prevalence rates for the projections were estimated using data from the 1970 Health Interview Survey, using procedures similar to those employed by the National Center for Health Statistics in preparing its national morbidity estimates. Estimates of disease prevalence were obtained by pooling the point-prevalence data of the HIS over the entire data year and then averaging. A gross adjustment for prevalence underreporting was attempted by adding the total number of deaths from a given cause to the estimated prevalence of that condition; in effect, the prevalence estimates assume that all of the deaths from a given cause occurred among individuals that were not covered by the Health Interview Survey. Data on deaths by underlying cause were tabulated from NCHS 1969 complete file of United States death certificates, adjusted to reflect 1970 levels of total death rates.

For the elderly projection a sample of 49,000 individuals, age 34 years and over in 1970, comprised the "start population". The statistical theory upon which the model depends requires that the projected sample be genuinely random in nature. As the 49,000 HIS cases that comprised our projected sample were, in fact, differentially weighted, an adjustment of the file was required before the actual simulation was carried out. Consequently, each case in the sample was duplicated by a factor equal to its case weight divided by the lowest case weight found in the sample. The result of this adjustment procedure was to expand the original sample to approximately 110,000 projected cases.

For the projection of the veteran population, the sample that was actually used for the microsimulation consisted of the 16,000 United States male veterans surveyed in the 1970 Health Interview Survey. The result of the sample replication procedure was to expand the veteran sample to approximately 31,000 projected cases.

Experimental Error

The stochastic nature of the event-timing functions implies that the projection contains an element of random variance, "experimental error". A simulation will lead to a "correct" projection only if the simulation of each sampled life history is carried out "sufficiently often", as on the familiar coin toss experiment. One way to assure that the projection is relatively free from experimental error is to repeat the simulation of all sampled cases repeatedly until the relative frequencies of each life history characteristic, specific by whatever demographic categories are of interest, cease to change. In our applications of the model, there were good reasons to suspect that such stability had been attained after only a single simulation was run, thanks to our large sample size. Indeed, stability could have been achieved with a much smaller sample than the one that was used. Stability was tested in the following manner.

After the simulation was carried out, the resultant sample of projected life histories was randomly divided in half. The hypothesis was then tested that the relative distributions of selected characteristics in the half samples could be reflective of two different sampling universes. As the primary interest in devising the model was to prepare joint projections of population size and health status in which survivorship patterns were consistent with patterns of disease prevalence, we selected for the test the distribution of projected deaths jointly tabulated on age of the decedent in 1970, age at death, and conditions present at death, with age specified in terms of 5-year age intervals. For each condition, the age-specific death rates in one-half of the file were regressed against those in the other half. The results for the veterans are shown in Table 1.

Projection Results

Illustrative results for the elderly white male population are presented in Tables 2 and 3. These tables reveal that the results for this exercise are generally in accord with our expectations -- that is the model behaves correctly. In the actual results obtained in this test, the total numbers of the elderly exceed those estimated in other national projections. At the same time, the preva-

lence of chronic disease (in both numbers and rates) is sharply increased.

These results indicate that the structure of the model is basically sound and does capture the interaction of disease prevalence and age structure. A major deficiency would seem to lie in the assumption regarding non-recovery since a disease is acquired that creates overestimates of prevalence as the projections proceeds. What is clearly called for are refinements in the estimated levels of prevalence in the start population and the incidence rates of disease conditions that drive the model. These considerations have received additional treatment in further elaborations of the model.

The veterans application offered a possibility of assessing the population results against alternative procedures. Table 4 provides the results of this exercise. The alternatives are Method I which involved projecting the 1970 veteran population forward on a cohort-component basis, with future discharges from the military taking the place of births and using survivorship ratios derived from the life table for all United States males including non-veterans.

A second approach -- Method II -- was to prepare a cohort-component projection similar to that of Method I, using veteran-specific survivorship ratios. Although it is true that the mortality data that would enable one to directly estimate a veteran-specific life table are not readily available, it is at least conceivable that such estimates might be obtained on an indirect basis, as follows.

For any cohort of veterans alive in a given year, the number of veterans alive after a certain period of time has passed is equal to the number alive initially plus the number of military discharges entering the cohort in the interval less the number of deaths occurring among both the initial cohort and those who were discharged from the military into the cohort during the interval. Suppose that the veteran population is closed to migration. Because such surveys as the Current Population Survey provide data on the size of veteran cohorts at successive intervals in time, and also is available, it was possible to estimate the rate of decrement over given time intervals.

Table 4 indicates that all three projections show the same general trend--a veteran population of 27.2 million in 1970 that increases to somewhat more than 28.5 million in the late 1970's, but then begins an uninterrupted decline through the year 2000. The main reason for the decline is that the large World War II veteran cohort will be entering old age, and the projected number of new entrants into the veteran population is simply not sufficient to offset the increased number of veteran deaths that are expected to occur as a result.

The populations projected under Methods I and II both peak in 1980, while the peak for the microsimulation is reached in 1975. Under Method II, with its veteran-specific survivorship, however, the population reaches both a higher peak size--28.8 million--and declines to a lower level in the year 2000--25.4 million, than occurs in the Method I projection with its total male survivorship schedule (somewhat less than 28.7 million, and 25.5 million, respectively). Method III, on the other hand, generates lower projected numbers at all times than are projected by the other two methods. Method I generates the highest proportion of projected elderly--27.9 per cent of the veteran population, as compared to the 26.4 per cent projected under Method II and the 23.2 per cent generated by the microsimulation, for the year 2000.

The survivorship dynamics hold the key for explaining these differences. Methods I and II assume that constant levels of survivorship prevail throughout the period 1970 through 2000. Method I subjects veterans to the schedule of survivorship rates that was experienced by all United States males, including non-veterans, in 1974. Method II, on the other hand, uses a schedule of survivorship experience in which veterans experience lower levels of mortality during the younger ages, but in which the favorable differentials fades with age. Indeed, examination of the Method II life expectancies reveal a slight mortality crossover in the older age categories; mortality rates derived from the 1970-73 CPS for ages past 55 years (not shown here) are actually slightly higher than those observed for all males in 1974. As a result, Method II population grows somewhat faster than the Method I population, as long as it is a relatively younger population; when the projected age distribution under Method II becomes a relatively older one than that of Method I, the Method II population declines much more rapidly because of its lower survivorship levels at the older ages.

In the microsimulation projection, however, survivorship levels vary over time as the relative prevalence of disease varies. An examination of the Method III life expectancies, derived from the age-specific death rates generated by the projection, reveal that the

implied survivorship function for veterans, on the assumption that they experience fixed rates of disease onsets and condition-specific mortality characteristic of the total United States male population, endows them with substantially lower mortality at the younger ages (initially) than that which characterizes either Method I or Method II, but that the rates rise more rapidly with age and lead to much higher levels of mortality at the older ages than is characteristic of either of the other two survivorship regimes. The result is a considerably smaller population projected for the year 2000, with a smaller proportion of aged individuals.

Thus, the Method II projection presents the population size implications of a regime of veteran mortality in which veterans are somewhat favored over non-veterans initially, with gradual convergence to the mortality experience for all males (assuming that age is a relatively good proxy for time since discharge), and with a slight mortality cross-over at the older ages. Method III shows the implications of a regime with initially much lower veteran mortality, rapid convergence and a much deeper cross-over.

(1) Horvitz, D. G., F. G. Giesbrecht, B. V. Shah, and P. A. Lachenbruch "POPSIM, a Demographic Microsimulation Model". Monograph 12. Chapel Hill, North Carolina: Carolina Population Center, University of North Carolina at Chapel Hill. 1971.

Table 1

Condition	Regression Slope	Correlation Coefficient
Acute infectious disease.....	1.077	0.992
Acute respiratory disease.....	0.811	0.986
Miscellaneous acute diseases.....	1.017	0.998
Chronic respiratory disease.....	0.983	1.000
Malignant neoplasms.....	0.998	0.998
Endocrine and metabolic disorders.	0.990	0.998
Cardiovascular disease.....	1.006	1.000
Cerebrovascular disease.....	0.922	0.999
Arterioschlerosis.....	1.043	0.999
Chronic digestive, liver disease..	0.970	0.995
Miscellaneous chronic disease.....	1.002	1.000
External events.....	0.943	0.978

Table 2

Projected Population of the Elderly, 1980-2000

White Males Age 65 Years and Over

Population By Age:	1980	1990	2000
All Ages, 65+ Years	10931462	14325538	15303999
65-69 Years of Age...	3808479	4388216	3861230
70-74 Years.....	2802199	3623962	3650442
75-79 Years.....	2028967	2816234	3263160
80-84 Years.....	1230640	1835174	2320493
85-89 Years.....	770708	1112436	1444209
90-94 Years.....	241681	417990	605811
95 Years and Over....	48788	131526	158654
Percentage Age Distribution:	**1980**	**1990**	**2000**
All Ages, 65+ Years	100.0	100.0	100.0
65-69 Years of Age...	34.8	30.6	25.2
70-74 Years.....	25.6	25.3	23.9
75-79 Years.....	18.6	19.7	21.3
80-84 Years.....	11.3	12.8	15.2
85-89 Years.....	7.1	7.8	9.4
90-94 Years.....	2.2	2.9	4.0
95 Years and Over....	0.4	0.9	1.0
75 Years and Over....	39.5	44.1	50.9
85 Years and Over....	9.7	11.6	14.4
Selected Characteristics:	**1980**	**1990**	**2000**
Mean Age (Years)	74.25	75.06	76.17
Percent of all Persons Age 65	38.98	39.24	38.86
Est. Annual Growth Rate (Percent)	3.70	1.76	0.18
Total Percentage Change Since 1970	47.61	93.45	106.66
Annual Deaths Per 1000 Persons	41.28	46.50	51.51

Table 3

Projected Total Prevalence
Of Selected Conditions Among The Elderly, 1980-2000

White Males Age 65 Years and Over

Persons With Selected Conditions:	1980	1990	2000
Total	10931462	14325538	15303999
No ill health	1997641	1383519	881611
Acute infectious disease	250296	338111	317779
Acute respiratory disease	682174	865988	868661
All other acute conditions	732784	913415	914767
Chronic respiratory disease	2654727	4541182	5827826
Malignant neoplasms	918501	1669400	2022090
Chronic endocrine and metabolic diseases	809535	1463881	1993063
Chronic cardiovascular disease	3125858	5595072	7010214
Chronic cerebrovascular disease	892016	1564239	1924073
Arterioschlerosis	349738	687425	836493
Chronic digestive and liver disease	764422	1513946	1985268
All other chronic conditions	4914559	7818069	9692446
Crude Total Prevalence Rates:	**1980**	**1990**	**2000**
No ill health	182.74	96.58	57.61
Acute infectious disease	22.90	23.60	20.76
Acute respiratory disease	62.40	60.45	56.76
All other acute conditions	67.03	63.76	59.77
Chronic respiratory disease	242.85	317.00	380.80
Malignant neoplasms	84.02	116.53	132.13
Chronic endocrine and metabolic diseases	74.06	102.19	130.23
Chronic cardiovascular disease	285.95	390.57	458.06
Chronic cerebrovascular disease	81.60	109.19	125.72
Arterioschlerosis	31.99	47.99	54.66
Chronic digestive and liver disease	69.93	105.68	129.72
All other chronic conditions	449.58	545.74	633.33
Mean Number of Conditions Per Person:			
All persons, age 65+ years	1.47	1.88	2.18
Persons, age 65+ years, w. 1+ conditions	1.80	2.08	2.32

*Crude total prevalence rates are expressed in terms of prevalent conditions per 1000 persons, age 65 years and over, of specified age, race and sex.

Table 4

Alternative Projections
Of The U.S. Male Veteran Population:
Selected Characteristics, 1970-2000

Population Size (1000's):

Year	Method I	Method II	Method III
1970	27,203	27,203	27,203
1975	28,614	28,647	28,390
1980	28,670	28,801	27,931
1985	28,398	28,595	27,061
1990	27,766	27,866	25,757
1995	26,772	26,711	24,224
2000	25,485	25,353	22,559

Method I: Cohort-component projection, assumes observed 1974 survivorship for total U.S. male population.

Method II: Cohort-component projection, assumes CPS estimate of veteran-specific survivorship for 1970-1973.

Method III: Sickness-death microsimulation, assumes HIS estimates of disease onset and virulence rates for total U.S. male population, 1970.

ECONOMIC AND ENVIRONMENTAL DETERMINANTS OF URBAN MORTALITY

Timothy D. Hogan and Lee R. McPheters, Arizona State University

The urban environment has become the dominant setting for contemporary life in the United States, with more than 73 percent of the population residing within metropolitan areas. The benefits of urbanization include numerous economies of scale in production and consumption. However, the process of urbanization has entailed mounting negative side effects or externalities, including crime, congestion, and environmental deterioration. In addition, urban mortality rates exceed those of non-metropolitan areas (7).

This paper examines factors associated with the level of urban morality in the United States. We seek evidence on the existence of a systematic relationship between urban mortality and a variety of socio-economic, environmental and health care characteristics of the urban setting. In particular, several alternative measures of the supply of health care services and of environmental quality have been included as variables in our analysis. This has been done so that we can explicitly investigate the relative effects of differences in the provision of health services and of variations in pollution levels upon urban mortality levels, since these factors are two potentially important determinants that can be affected by social policy. The U. S. society currently allocates huge sums of money to health care expenditures, and there is growing pressure for the establishment of an even more expensive national health insurance system. At the same time, we have in recent years seen a social and governmental response, involving both increased expenditures and regulation, to the growing awareness of the adverse effects of pollution and other environmental contaminants. Many argue that, in an affluent society such as the United States, other factors such as personal habits, diet, pollution, etc., have more impacts on health than the availability of more and better medical services (6,9). We hope to gain some insight into this controversy within the context of the urban environment with this empirical analysis.

II. Review of Previous Studies

Although the determinants of urban mortality are undoubtedly complex, initial studies attempted to link variation in mortality with differences in a single or small number of suspected influencing variables. For example, in one of the earliest reported studies, Altenderfer (1) examined the relation between per capita income and mortality in 1940 for 92 cities with population greater than 100,000 persons. Altenderfer concluded that overall mortality and the death rate for ten broad diagnosis groups were inversely related to income (1, p. 1688).

In a later study, Patno (14) related mortality to "economic level," using 1940 and 1950 census tract data for the white population of Pittsburgh, Pennsylvania. The measures employed as indicators of economic level were median value of owner occupied housing and median monthly rental, and median family income. Based upon this evidence, Patno determined "In general, the highest mortality occurred among persons within the areas designated as being of low economic level, and the most favorable experience was found among the residents of the areas of higher economic status" (14, pp. 845-846).

The particular problem of racial differences in infant mortality in urban areas was examined by Jiobu (11) who found post-natal infant mortality related to socio-economic measures. He suggested that ghettoization may affect infant mortality due to influences of factors such as overcrowding and quality of medical care.

In a more qualitative investigation considering a wider range of influences of mortality, Biraben linked urban mortality to certain aspects of urban living such as increased personal contacts, traffic, pollution, and the general pace of city life (4).

A major collection of studies on the economics of health and medical care was published by the National Bureau of Economic Research in 1972 (8). One of these essays (2) attempts to test the impact of medical care variables on health, as measured by white mortality rates. While this study analyzed state, rather than urban data, the authors, Auster, Leveson, and Sarachek, took a more sophisticated approach than previous studies by including a number of socio-economic and environmental variables to control for differences among geographic areas. They discovered the influence of medical care on mortality to be small, while the association between mortality and education was strong and negative. Surprisingly, an income measure was found to be positively related to mortality in this study, contrary to much of the previous evidence.

Another study included in that collection, which is perhaps the most comprehensive investigation of the determinants of urban mortality to date, is that of Silver (15) who examined both Standard Metropolitan Statistical Area and state data to explain spatial variation in black and white mortality rates. Silver applied regression analysis to some forty explanatory variables. Again, Silver found the relation between education and mortality to be negative and usually significant. The excess of black over white mortality was attributed to differences in income and educational levels for the two groups. Generally, Silver found a negative relation between income and mortality, with education excluded from the model. One exception was the case of white male mortality, using state data. Re-estimating the white male mortality equations with income broken into labor and non-labor components, the sign of non-labor income was strongly negative, while labor income was not usually significant. This

suggested to Silver that "pure" increases in income may have a positive effect on health, but incomes earned by more strenuous or dangerous work may be unfavorable to health.

While the more recent attempts to estimate the relative contributions of various determinants of urban mortality have included larger numbers of variables than earlier approaches, the increased number of variables has introduced the possibility of multi-collinearity and thus mis-interpretation of the empirical results. Below, we follow an approach which allows for consideration of a large number of variables, but minimizes the problems associated with multicollinearity.

III. Design of the Study

Our fundamental hypothesis is that variation in urban mortality is dependent upon a complex milieu of determinants, including economic variables, environmental variables, population characteristics, and various measures of health care availability and utilization. We test this hypothesis in a least squares regression analysis, using data for the 64 largest SMSA's within the continental U. S. for 1970.[1] Since the inclusion of all the potential determinant variables in one regression equation would prove unwieldy and statistically suspect, we reduce the information contained in the original large data matrix through extraction of its principal components.

Briefly, principal component analysis takes observations on a large number of correlated variables and finds a smaller set of orthogonal or uncorrelated variables which capture as much of the variability of the original data set as possible (5, pp. 53-65). The resulting components may be used to construct index variables which can be employed as independent variables in ordinary least squares regression analysis.

The original data used in our analysis are shown in Table 1. Included are measures of economic variables, population characteristics, environmental measures, and medical care variables. Many of these variables have appeared independently as explanatory variables in previous analyses of mortality.

Extracting components until the resulting Eigen-values fell to 1.0, we derived nine orthogonal components or factors which accounted for over 75 percent of the variance in the original data matrix. To facilitate interpretation of each of the derived factors, we performed varimax rotation, which preserves orthogonality while simplifying the columns of the factor matrix.[2] The loadings shown in Table 2 measure the correlation between the original variables and each of the respective components after varimax rotation.

The first factor extracted in principal component analysis is usually a general factor expressing a summary of the linear relationships present in the data. After rotation, the first factor (F1) seemed to measure the general character of urban areas especially with respect to motor vehicle dominance. This factor is positively associated with motor vehicle registrations, days of sunshine, per capita income and education, and negatively related to hospital occupancy rates and pollution variables. We found SMSA's with high scores on this factor are likely to be newer "sun belt" cities.

The second factor (F2) tends to be most highly associated with variables measuring economic level, including percent of the population above the poverty level and retirement benefits. The third factor (F3) reflects influence of suburbanization on the middle class; it is negatively related to population density and housing dilapidation, but positively related to education and proportion of housing owner occupied.

Factor four (F4) is most strongly associated with property or non-labor income, while factor five (F5) is linked to measures of health care services, including doctors and dentists per 100,000 persons and per capita health care expenditures by local governments. Factor six (F6) is associated with air quality and water pollution variables, while factor seven (F7) is interpreted as a medical facilities factor, loading highly on hospital beds per 100,000 persons. Factor eight (F8) seems closely associated with both savings and black white income differentials, and factor nine (F9) loads most highly on unemployment.

These factors were used as weights in constructing indices corresponding to each component. Each of the 64 SMSA's thus received a weighted value or "score" for each component. The SMSA's with the highest and lowest index values for each factor are shown in Table 3.

The constructed indices were utilized as orthogonal independent variables in an ordinary least squares regression equation of the form

$$M_i = a + b_1F1 + b_2F2 + b_3F3 + b_4F4 + b_5F5 + b_6F6 + b_7F7 + b_8F8 + b_9F9 + u$$

where M_i is the age-adjusted mortality rate for the ith population category a is an intercept term and u is the stochastic error term.

The mortality variables used in the study are age-adjusted mortality rates for whites (MW) and blacks (MB) for 1970 expressed in index form. Data for construction of the mortality variables were obtained from (17) and (21). The computation procedures followed those set out in (16, p. 242). The mortality index measures the ratio of SMSA mortality to expected deaths based upon national age-specific mortality rates.

IV. Empirical Results

Regression analysis yielded the following regression results for white and black mortality (t-values are in parentheses, with those significant at the 5% level marked with an asterisk):

(1) MW = 100.1 - 1.99 F1* + .746 F2 - 3.09 F3*
(2.90) (1.08) (-4.49)

- 1.71 F4* - .448 F5 + 1.46 F6* + 1.85 F7* -
(-2.489) (-.652) (2.13) (2.69)

1.69 F8* - .086 F9 R^2 = .501
(-2.47) (-.126)

(2) MB = 106.4 - 2.68 F1 - 6.45 F2* - 3.03 F3 -
(-.911) (-2.19) (-1.03)

891 F4 - 5.55 F5* + 1.21 F6 + 3.24 F7 +
(-.303) (-1.89) (.410) (1.10)

6.03 F8* - 2.96 F9 R^2 = .239
(2.05) (-1.02)

Since the indices are based on standardized variables, the importance of each factor to the equation can be measured by the size of the regression coefficient. For white mortality, the most important factor is F3, the measure of suburbanization, which includes low population density, high median education, and high proportion of owner occupied housing. The negative sign on this factor suggests that the process of suburbanization has a definite favorable impact upon mortality.

These results are consistent with a number of previous studies which found an inverse relationship between education and health and mortality. For example, Kitagawa and Hauser found a "consistent decline in mortality as years of schooling increased" (12, p. 38). They interpreted the education variable as a proxy for all the various socioeconomic variables which may be linked to education, including income, level of occupation, style of life, diet, quality of housing, and others. Grossman also found a similar relationship, within the framework of a more sophisticated human-capital model (10).

Our approach suggests that higher levels of education in urban areas are closely tied to other important socioeconomic variables, especially those associated with suburbanization. Previous studies utilizing single variables in regression equations may have failed to capture the composite nature of the relationship involved.

The second most important determinant of white mortality in this equation is F1, the measure of general urban character. This index enters with a negative sign, which suggests newer, rapidly growing SMSA's may have a more favorable mortality experience, in spite of high motor vehicle density.

Also important is F7, the medical facility factor. The sign is positive, which may be an indication of simultaneity in the underlying structure of the relationships of the equation. That is, areas with higher mortality rates may require a large stock of medical facilities.

Factor F4, non-labor or property income, is significant and negative in sign. This confirms earlier findings of Silver (15). That is, increases in income may be beneficial to health and longevity if they are not directly associated with additional emotional or physical stress. Similarly, factor F8, a factor which was highly associated with savings, has a negative effect on mortality.

The final significant determinant of white mortality is the pollution index (F6). The positive sign here supports the view that higher pollution levels are detrimental to health.

Three of the factors were not significant at the 5% level. These were the economic level index (F2), the health services index (F5) and the unemployment measure (F9). The insignificance of the health services factor supports a growing body of literature which suggests that in advanced societies there may not be a strong link between supply of health services and health status (6, 9).

The results for black mortality are substantially weaker than for white mortality, with less than 25 percent of the variation explained by three significant indices. The most important of the significant variables is factor two which is highly correlated with the proportion of the population above the poverty level. This variable was insignificant for whites, suggesting mortality gains to whites from increasing affluence have possibly been exhausted, but such benefits to blacks are still forthcoming.

This interpretation is given weight by the results for factor eight, which is highly associated with savings and the black-white income differential. While this factor had a negative influence on white mortality, the sign for black mortality is positive.

The third variable of interest is the health care services index (F5). As with the poverty variable, health care services was not significant for white mortality, but is significant and of negative sign for black mortality. While incremental physicians or public expenditures on medical services have no apparent significant influence on white mortality, black mortality seems to be influenced in a beneficial way by the availability of such services.

IV. Conclusions

While there is substantial variation in mortality rates among urban areas, there appear to be certain urban characteristics which are systematically related to mortality levels. These characteristics have a differential impact on white and black mortality rates.

The empirical results of our study demonstrate the mortality experience of whites in newer, automobile oriented, suburbanized areas is more favorable than that found in older, higher density metropolitan areas. While early studies found a persistent negative relationship between economic level and urban mortality, studies based on more recent data have found a positive relationship between economic factors and mortality. Our results confirm the explanation offered by Silver for this anomaly. The negative relationship between our property income factor and white mortality and the insignificance of our index of economic level (F2) replicate Silver's findings for non-labor income and aggregate income measures.

The results for mortality experience of urban blacks seem the virtual inverse of that for whites. Factors related to suburbanization and character of the SMSA are found to be insignificant determinants of black mortality. On the other hand, two factors found to be insignificant for whites (economic level and medical services) were significant and of the theoretically expected sign for blacks. This evidence is consistent with the hypothesis that improvements in economic level and medical services continue to offer potential mortality gains to blacks, which may be no longer true for whites.

The empirical results also provide some interesting evidence for the controversy dealing with the proper mix of public policies to improve the health of the U. S. population. The findings relating to white mortality tend to support the contention that efforts should be concentrated away from the traditional medical approach toward a broader approach of lifestyle modification. At the same time, however, results from the analysis of black mortality imply that continued emphasis on improved medical care and increased availability of such services to the black and minority populations would have significant impact upon the health status of these disadvantaged populations.

FOOTNOTES

[1]Honolulu was excluded due to its unusual socio-economic and demographic characteristics.

[2]Such a simplification is equivalent to maximizing the variance of the squared loadings of each column. Hence, the name "varimax."

REFERENCES

1. Altenderfer, M. E. "Relationship Between Per Capita Income and Mortality," Public Health Reports, Vol. 62, pp. 1681-1691, Nov. 1947.

2. Auster, Richard, Irving Leveson, and Deborah Saracheck, "The Production of Health, An Exploratory Study," in Victor R. Fuchs, (ed.), Essays in the Economics of Health and Medical Care, National Bureau of Economic Research, Columbia University Press, New York, 1972.

3. Berry, Brian, Land Use, Urban Form and Environmental Quality, University of Chicago, Department of Geography Research Paper No. 155, 1974.

4. Biraben, Jean-Noel, "Quelques Aspects de la Mortalite en Milleu Urbain," Population, Vol. 3, pp, 509-520, 1975.

5. Dhrymes, P. J., Econometrics: Statistical Foundations and Applications, Harper and Row, New York, 1970.

6. Dubos, Rene, The Mirage of Health, Harper, New York, 1959.

7. Erhard, Carl L. and Joyce E. Berlin (eds.), Mortality and Morbidity in the United States, Harvard University Press, Cambridge, 1974.

8. Fuchs, Victor, Essays in the Economics of Health and Medical Care, National Bureau of Economic Research, Columbia University Press, New York, 1972.

9. Fuchs, Victor, Who Shall Live, Basic Books, New York, 1974.

10. Grossman, Michael, "The Correlation Between Health and Schooling," in Household Production and Consumption, National Bureau of Economic Research, 1976.

11. Jiobu, Robert M. "Urban Determinants of Racial Differentiation in Infant Mortality," Demography, Vol. 9, pp. 603-615, November, 1972.

12. Kitagawa, E. M. and P. M. Hauser, "Education Differentials in Mortality by Cause of Death: The United States, 1960," Demography, Vol. 5, pp. 318-353, 1968.

13. Liu, Ben-Chieh, Quality of Life Indicators in United States Metropolitan Areas, 1970, U. S. Environmental Protection Agency, Washington, D. C. 1975.

14. Patno, M. E., "Mortality and Economic Level in an Urban Area," Public Health Reports, Vol. 75, pp. 841-851, September, 1960.

15. Silver, Morris, "An Econometric Analysis of Spatial Variations in Mortality Rates by Age and Sex," in Victor Fuchs, ed., Essays in the Economics of Health and Medical Care, National Bureau of Economic Research, Columbia University Press, New York, 1972.

16. Stockwell, Edward G. (ed.), The Methods and Materials of Demography, San Francisco, Academic Press, 1976 (condensed edition).

17. United States Department of Commerce, Bureau of the Census, 1970 Census of Population and Housing, United States Government Printing Office, Washington, D. C., 1972.

18. United States Department of Commerce, City and County Data Book, United States Government Printing Office, Washington, D. C., 1972.

19. United States Department of Commerce, Survey of Current Business, United States Government Printing Office, Washington, D. C., May, 1974.

20. United States Department of Commerce, Statistical Abstract of the United States, United States Government Printing Office, Washington, D. C. Various Issues.

21. United States Department of Health, Education and Welfare, Public Health Service, Vital Statistics of the United States, 1970, U. S. Government Printing Office, Washington, 1971.

22. United States Department of Justice, Federal Bureau of Investigation, Uniform Crime Reports of the United States, 1970, U. S. Government Printing Office, Washington, D. C., 1971

23. United States Department of Transportation, Federal Highway Administration, Motor Vehicle Registrations by SMSA, United States Government Printing Office, Washington, D. C., 1971.

Table 1

ORIGINAL VARIABLES USED IN PRINCIPAL COMPONENTS ANALYSIS

Symbol		VARIABLES
V1	(IVFG)	Annual Inversion Frequency
V2	(SUN)	Annual Sunshine Days
V3	(T32)	Number of Days Temperature Above 32°
V4	(HBDS)	Hospital Beds Per 100,000 Persons
V5	(TCRM)	Total Crime Rate Per 100,000 Persons
V6	(COST)	Cost of Living Index
V7	(SEG)	Housing Segregation Index
V8	(APOV)	Percent Families Above Poverty Level
V9	(MCYC)	Motorcycle Registrations per 100,000 Persons
V10	(PDNS)	Population Density
V11	(BWYA)	Ratio of Black to Total Median Family Income
V12	(MV)	Motor Vehicle Registrations per 100,000 Persons
V13	(DDS)	Dentists per 100,000 Persons
V14	(HSOC)	Hospital Occupancy Rates
V15	(MD)	Physicians per 100,000 Persons
V16	(HCEX)	Per Capita Local Government Expenditures on Health
V17	(MDED)	Median School Years Completed
V18	(PART)	Mean Level for Total Suspended Particulates
V19	(SLDX)	Mean Level for Sulfur Dioxide
V20	(DLPD)	Percent Housing Units Dilapidated
V21	(WSTE)	Tons of Solid Waste from Manufacturing
V22	(WTPL)	Water Pollution Index
V23	(RTBF)	Average Monthly Retiree Benefits
V24	(UNPLY)	Unemployment Rate
V25	(PRSY)	Per Capita Personal Income
V26	(SVG)	Per Capita Savings
V27	(PYPSY)	Property Income as a Percent of Personal Income
V28	(OWOC)	Percent Owner Occupied Housing
V29	(MVAL)	Median Value of Owner Occupied Housing

Sources: V1, V2, V3, V6, V7, V14, V18, V19, V22 are from (13);
V8, V10, V17, V20, V25, V28, V29, are from (17);
V11, V23, V24, V26 are from (18); V4, V13, V15, V16 are from (20);
V9, V12 from (23); V5 is from (22); V21 is from (3);
V27 is from (19).

Table 2

LOADINGS USED TO CONSTRUCT INDICES

VARIABLES	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7	FACTOR 8	FACTOR 9
IVFG	.418	-.043	.314	-.582	-.013	.021	.127	-.240	.214
SUN	.748	-.364	.024	.120	.131	-.037	-.108	.039	-.100
T32	-.530	.439	.168	-.283	-.147	.167	.249	-.067	-.328
HBDS	-.036	-.100	-.031	.013	.081	.156	.895	-.022	-.114
TCRM	.548	-.397	-.015	.014	.351	.171	.180	.231	.158
COST	-.089	.455	-.432	.232	.457	-.102	.004	-.012	-.119
SEG	-.154	.772	-.025	-.038	.145	.144	-.121	-.198	-.054
APOV	.032	.892	.107	-.005	.147	-.029	-.029	.062	-.024
MCYC	.772	-.095	.334	-.114	-.045	-.146	-.150	-.020	.321
PDNS	-.036	.113	-.849	-.033	.153	.132	-.069	.091	.070
BWYA	-.424	-.095	.173	.164	-.282	.036	-.252	.627	.114
MV	.728	-.113	.301	.180	-.288	-.207	-.029	.183	-.059
DDS	.027	.255	.026	.304	.744	-.006	.253	.049	.284
HSOC	-.755	-.162	.020	.051	-.099	.052	-.056	.088	-.105
MD	.079	-.030	.012	.162	.797	.040	.293	.023	-.032
HCEX	-.116	.005	-.024	-.428	.724	-.104	-.103	.072	.130
MDED	.220	.286	.681	.114	.415	-.015	-.148	-.071	.122
PART	.037	.006	.043	-.275	-.065	.765	.172	.035	-.219
SLDX	-.351	.365	-.316	.106	.085	.628	-.073	.011	-.026
DLPD	-.130	.079	-.698	.035	-.028	-.175	.099	-.392	.058
WSTE	.097	-.458	.136	.589	-.033	-.004	-.429	.005	-.192
WTPL	-.258	.123	.064	.393	.001	.645	.112	-.048	.233
RTBF	-.215	.802	-.103	.176	.060	.217	.038	.207	.190
UNPLY	.218	.030	-.006	-.150	.133	-.060	-.109	.030	.861
PRSY	.139	.376	.071	.172	.548	-.033	-.371	.062	.114
SVG	.181	.092	-.016	.187	.298	-.031	.079	.771	-.009
PYPSY	.159	.120	.109	.789	.173	-.022	.087	.191	-.020
OWOC	-.024	.196	.646	.089	-.605	-.079	.009	.023	.072
MVAL	.142	.414	-.208	.014	.724	.051	-.274	.088	-.037
Cumulative Proportion of Variance	19.2	35.6	45.8	54.6	61.3	66.3	70.8	74.5	78.0

Table 3

URBAN AREAS WITH HIGHEST AND LOWEST SCORES ON CONSTRUCTED INDICES

FACTORS	HIGHEST	LOWEST
F1	Anaheim	Milwaukee
	Phoenix	Buffalo
	San Bernadino-Riverside-Ontario	Albany
F2	Patterson-Clifton-Passiac	Memphis
	Hartford	San Antonio
	Minneapolis-St. Paul	Norfolk-Portsmouth
F3	Denver	Jersey City
	Portland	Newark
	Grand Rapids	New York
F4	Ft. Lauderdale	Gary-Hammond
	Miami	Sacramento
	Tampa-St. Petersburg	San Jose
F5	New York	Youngstown-Warren
	Washington, D. C.	Ft. Worth
	San Francisco-Oakland	Gary-Hammond
F6	Cleveland	Rochester
	Detroit	Seattle-Everett
	Pittsburgh	Allentown-Bethlehem-Easton
F7	Oklahoma City	Norfolk-Portsmouth
	Minneapolis-St. Paul	San Diego
	New Orleans	Washington, D. C
F8	Ft. Lauderdale	Boston
	Miami	Springfield-Chicopee-Holyoke
	Tampa-St. Petersburg	Providence-Pawtucket-Warwich
F9	Seattle	Washington, D. C
	Portland	Dayton
	Pittsburgh	Denver

DISCUSSION

Monroe Lerner, The Johns Hopkins University

My assignment for this discussion was the two papers, by Mooney and Rives 1/ and by Hogan and McPheters 2/. Both are excellent and hard to criticize negatively, and I shall not attempt to do so. Rather, I will summarize their presentations, selecting the points most salient to me, and propose some rather simple-minded suggestions for the authors' consideration.

Mooney and Rives

The National Health Planning and Resources Development Act of 1974 (P.L. 93-641) mandates that Health Systems' Agencies (HSAs) assess the health status of residents of their planning areas; to satisfy this mandate, they must be able to measure health status empirically. However, at least two major barriers to empirical measurement exist: 1) The dimensions of health status are not specified, and 2) Since data-collection activities of HSAs are necessarily restricted, measurements will have to be made with routinely available data. This paper examines six possibilities or approaches to identify the dimensions of community health status characterized as feasible and/or practical for HSAs.

These are:

1. Use mortality data to summarize the risk of dying.
2. Use mortality data to infer morbidity.
3. Use morbidity data to measure incidence and prevalence.
4. Use utilization and treatment data to infer morbidity from specific conditions and/or in selected population segments.
5. Use uni-dimensional indicators to represent the multi-dimensional concept of health status.
6. Use synthetic measures of health status.

The pros and cons of each approach are reviewed very competently by these authors and their conclusion is that none by itself does the entire job. Therefore Mooney and Rives suggest the promulgation, in the near future, of a minimum basic set of health status measures, i.e., a national standard involving appropriate definitions, uniform data sources, and standard methods of computation for a variety of measures. Presumably each would be derived from available data. Some would be calculated for all HSAs, some not, but all would be comparable across areas.

They suggest also paying particular attention to synthetic estimation, a technique for using national data to make sub-national estimates, at the same time that they caution about the hazards inherent in its use. Finally, they suggest that the construction of social indicator models at the county or HSA level for the measurement of health status is a desirable long-run objective.

Overall, this paper very competently illuminates the framework within which HSAs will necessarily have to operate, and in so doing it makes a real contribution by synthesizing knowledge in this currently very salient area.

If one single, most important, suggestion may be made, it is this: I would have liked some greater recognition that health status today, in response to changing health problems, is best considered as multi-dimensional, including both the quantity and quality of life and, under quality, physical, mental or emotional, and social well-being. The implication is that the authors' review of problems facing HSAs in developing indicators of community health status provides too little discussion of the dilemma that, even as social well-being increases in importance relative to other dimensions of health status, HSAs are nevertheless unlikely to measure it. This is in part because the technology does not as yet exist, in part because the health field lags widely behind some social science investigators in recognizing that social well-being is today properly a component of health, but also because the organizational structures of the health and social welfare fields remain largely separate and distinct, reflecting the development of separate professions to deal with the major problems of man—physical, mental, social (and moral)—rather than one profession treating him as a whole human being.

By thus omitting measurement of social well-being, HSAs will omit measuring an important dimension of health status, and one not necessarily highly correlated with other dimensions. As a consequence, if this continues long enough, their partial definitions may become "frozen", i.e., locked into the process of health status measurement, and the field may be set back substantially; alternatively, measurements under the limited definition may subsequently require substantial revision, thus making them less than optimally useful. Planning meanwhile on the basis of this partial definition is sure to be correspondingly inadequate.

Perhaps one other point may be made, underscoring what the authors have said. One of the most frustrating experiences is to be told that some aspect of life—health, intelligence, socio-economic status, etc.—is so complex that no single definition or measurement is adequate in capturing it. Yet our investment to improve or at least maintain it, e.g., health, is enormous, and cost-benefit analysis and/or planning cannot proceed optimally without some measurement. The pressure to aggregate into a single, summary measure will properly be enormous, but it isn't clear how HSAs will derive this measure. Imperfect measures will be employed, perhaps differing among HSAs, and these may even be empirically quite useful. At the same time, work on the theoretical aspects of the measure will continue. The parallel to developments in measurement of intelligence are obvious.

Hogan and McPheters

The second paper examines the influence of

1/ Anne Mooney and Norfleet W. Rives, Jr., Indicators of Community Health Status for Health Planning.

2/ Timothy D. Hogan and Lee R. McPheters, Economic and Environmental Determinants of Urban Mortality.

urbanization on age-adjusted mortality rates for the 64 largest U.S. SMSAs for 1970. Using an econometric model, the authors test also for the influence of population density, income level, housing conditions, air quality, and health services' expenditures on mortality for blacks and whites separately. A principal components' analysis prior to the use of two regression equations reduces 29 original variables to nine factors.

The authors find that for whites the most important determinant of mortality is their measure of suburbanization, a construct including low population density, low housing dilapidation, high median education, and a high proportion of owner-occupied housing. This construct is negatively related to mortality.

Their second most important determinant of white mortality is their construct measuring general urban character, with a negative sign, suggesting that newer, rapidly growing SMSAs may have a more favorable mortality in spite of the high motor-vehicle density characteristic of these cities. The following three factors were also found to be significant: medical facilities, with a positive sign, perhaps indicating only that high mortality areas require a large stock of medical facilities; non-labor or property income, negative sign, indicating that increases in income may be beneficial to longevity if not directly leading to additional stress; and pollution, positive sign, detrimental to health. Not significant, at the .05 level, were economic level, health services, and unemployment.

Results for black mortality are substantially weaker than for white. The most significant variable here is the factor highly correlated with proportion of population above the poverty level, suggesting that mortality gains from increasing affluence are still accruing to blacks, not so for whites. Also for blacks, the health care services' index was significant, and again not so for whites.

Intuitively the results make considerable sense. Nevertheless, as in so many analyses of this type, and especially here where principal components' analysis and econometric models are used, the results often seem forced. For example, some of the factors, combining quite unlike items, represent "artificial" constructs, possibly a consequence of the aggregate nature and crudeness of the original data on which they are based. The logic of their combination seems strained. As a consequence, even though the results make intuitive sense, policy implications should be drawn only with considerable care. This is particularly true of the conclusion that, for the reduction of white mortality, "effects should be concentrated away from the traditional medical approach toward broader life-style modification", while for blacks "continued emphasis on improved medical care and increased availability....would have significant impact on health status".

Data analyzed by me suggest that advances in medical technology in recent years have had a significant impact on heart disease and cancer. We see this in the reduction in recent years in overall mortality rates, but especially in the mortality rates from these two major diseases, while the end is, hopefully, not yet in sight. Much further work is clearly indicated, especially on socio-economic differentials in mortality within the major metropolitan areas and on differentials among central cities, suburbs, and non-metropolitan areas especially by cause-of-death, age, race, and sex. Planning will proceed by identifying these differentials and their causes, and by locating the pockets of excess mortality and their causes.

A COMPLETE FACTOR ANALYSIS BY AN INDIRECT METHOD

Harry R. Barker Barbara M. Barker

University of Alabama

A novel, indirect method of factoring data matrices was developed by Horst (1965). The method involves first reducing the variable matrix to a few subsets of variables and then deriving a score (such as a total) from each subset for each subject. (Rules for forming subsets are unspecified). The matrix of intercorrelations between subset totals is factored; and, then, taking into account intercorrelation of subset totals with variables, an estimate is made of the factor structure of the variables. The advantage of this method is that it avoids the direct factoring of the larger data matrix. This permits very rapid computer solution and enormously increases the number of variables which can be factored on a computer. Conventional factor methods permit about 80 to 100 variables which can be factored simultaneously on a computer.

During the past several years the investigators have extensively explored the indirect factor method. A summary report of this research by Barker and Barker (1975) indicated two vital requirements for accuracy of the method:

(1) Subsets of variables must be of homogeneous factor composition.

(2) Variables must enter subset totals with appropriate sign (+ or -).

These findings suggested the total impracticality of the indirect solution. In essence, one had to know the factor structure of the set of variables in order to assign the variables properly into subsets and to assign the correct sign for addition. Subsequent research revealed that the indirect method could be useful in testing theories of the factor structure of large data sets.

Three studies demonstrated the theory testing value of the indirect method. Hamlett (1976) evaluated theories of the personal orientation inventory. Barker and Barker (1976A, 1976B) evaluated competing theories of the factor structure of the MMPI on the original normative sample. Interest centered on the degree of association between subset specification (according to theory) and clusters of items identified by factor analysis. The information measure D was used to quantify the degree of agreement between theory and empirical results. Although the theories of the MMPI were successfully rank ordered according to the D measures, none of the D measures was high thereby indicating none of the theories was adequate.

Further attempts were made to extend the usefulness of the indirect factor method, by using results of the indirect method to refine the variable subsets. Starting with an initial clustering of the variables into subsets, an indirect factor solution was obtained and used for the purpose of reassigning the variables to subsets. A second indirect factor solution was obtained, and further refinement of the variable subsets was attempted. This iterative process was continued until computer time was exhausted or convergence on a stable set of variable subsets was reached. Barker and Barker (1977) tested this refined procedure on several computer generated data sets, which varied in strength of factor structure (strong, moderate and mixed) and obtained excellent results. After a few iterations, indirect solutions were virtually identical to those of conventional factor solutions.

The purpose of this study was to replicate the earlier indirect factor analysis of the MMPI normative data (male and female separately) using the iterative factor method in order to arrive at a definitive factor structure.

Method

The data consisted of item answers on the MMPI of 225 males and 325 females. These subjects were originally used as normative groups for the conventional MMPI clinical scales. The data resided on computer tape and were analyzed separately for male and female.

In an attempt to provide an objective and hopefully satisfactory starting place for the indirect factor method, the following steps were taken:

(1) Twenty subjects were selected randomly and then were used in an obverse factor analysis (CORR98, Barker and Barker, 1977). Eigenvalues were examined by a scree test in order to roughly identify the number of factors to retain. Subsequent varimax rotations were performed on successively fewer principal axes factors until the correct number of factors was identified. A varimax factor load equal to or greater than .3 on only one factor was the criterion used for clustering a

---

*Paper presented at American Statistical Association, August 17, 1977 (Chicago, Ill.). The research was made possible by funds from the Research Grants Committee of the University of Alabama.

variable into a subset. The sign of the factor load was used to identify the manner in which the variable entered into the totalling operation.

(2) The initial subsets of variables assigned by CORR98 were used for the first indirect factor solution (CORR99, Barker and Barker, 1977). The computer program (CORR99) utilized the outcome of the factor solution to reassign items to subsets and continued the iterative process for the specified number of interations or until convergence was reached. Convergence was defined as two consecutive identical factor solutions.

The computer program (CORR99) was modified to compute the information measure D between location of items in subsets and in factor clusters for each iteration.

The similarity of factor solutions for male and female was evaluated by (CORR22, Barker and Barker, 1977). This computer program rotates one factor solution to another, and determines degree of contiguity of variables attained in factor space.

Results

The obverse factor solutions, (CORR98) for the male and female samples suggested seven factors in each data set. Clustering of variables on the varimax rotated factors was used to identify the initial subsets of variables and sign for totaling operation for the indirect factor analysis.

Twenty-five iterations of the indirect factor solution were performed on both the male and female data sets. In neither case was convergence between variable subsets and factor clusters attained. Failure of convergence suggests that the factor structure of the data sets is quite weak. Earlier research reported by Barker and Barker (1977) noted that the weaker the factor structure of the data, the greater the number of iterations required to attain convergence.

The twenty-fifth iteration produced a D measure between variable subsets and obtained variable clusters of .60 for males and .70 for females. Although the D measures were lower than expected, the obtained D measures exceed those earlier obtained when evaluating competing theories of the factor structure of the MMPI. Therefore, this solution appears superior to those earlier attained.

The original estimate of seven factors for the males was supported whereas only four factors were retained for the females. Tables 1 and 2 identify items associated with each of the factors for male and female. In order for an item to be identified with a factor, a varimax factor load of .3 or greater on only one factor was required. For males, a preliminary labeling of factors is as follows:

(1) General mental health
(2) Religious
(3) Adventurous, independent
(4) Moralistic
(5) Unclear-items range from fear of catching disease to awareness of ears ringing and dreaming
(6) Neurotic
(7) Phobias, Anxiety

Factor 3 and 5 might well be dropped because they contain so few items.

Suggested labels for the female factors are:

(1) General mental health
(2) Moralistic
(3) Neurotic
(4) Phobias, anxiety

Factor 4 contains only 6 items and could be dropped.

Rotation of the varimax factor structure for females to maximum contiguity with the varimax factor structure of the males resulted in relatively good factor alignment; however item locations in factor space were quite separated. This further supports the apparant lack of similarity in male and female factor structure.

Data were processed on a UNIVAC 1110 system with 128K core allocation. The required computer time and costs for the indirect solutions were as follows:

(1) The male data required a computer run time of 25 min. 50 sec. and cost $176.42.

(2) The female data required about 34 min. at a cost in excess of $210.00.

Discussion

It appears that a weak factor structure characterizes the original MMPI normative data for both male and female. In view of the decided slant of MMPI items towards measuring pathology and the alleged normalcy of the subject samples, this is to be expected. A finding of considerable interest is the difference in factor structure for male and female.

In interpreting the findings of the study several cautions should be observed. The

data were obtained in 1957, on principally Mid-Western rural subjects. Cultural biases are very likely reflected in the data. The ratio of subjects to variables is grossly inadequate according to several criteria for multivariate work. For example, Cattel's rule which suggests that the number of subjects equal or exceed the number of variables by 100 is far from met. The data set would be of less interest except that the standard norms for the MMPI were obtained on these two samples. These norms have remained unchanged throughout the test's history.

Considering the large number of items on the MMPI which are not scored on the regular clinical scales, it was anticipated that additional useful factors might emerge which could be used to measure dimensions among normals. The obtained results support this view.

Summary

Application of a refined version of Horst's indirect factor method to the original normative data of the MMPI disclosed weak factor structure for both male and female. Seven factors for male and 4 factors for females were extracted and rotated to a varimax criterion. Factors extracted for male did not closely resemble factors extracted for female.

References

Barker, H. R. and Barker, B. M. Behavioral Sciences Statistics Program Library. University of Alabama, 2nd. rev. edition, 1977.

Barker, B. M. and Barker, H. R. Evaluation of theorized factor structure of the MMPI for male and female populations. Proceedings of the Social Statistics Section, American Statistical Association, 1976A, 174-178.

Barker, H. R. and Barker, B. M. An indirect method for testing the dimensionality of large data sets. Proceedings of the Social Statistics Section, American Statistical Association, 1976B, 179-183.

Barker, H. R., and Barker, B. M. Summary of research on a novel, indirect factor analytic solution. Proceedings of the Social Statistics Section, American Statistical Association, 1975, 298-301.

Barker, H. R. and Barker, B. M. Further refinement of an indirect method for factoring large data sets. Paper presented at Southeastern Psychological Association, Hollywood, Fla., May 6, 1977.

Hamlett, C. C. Validation of the theorized factor structure of the personal orientation inventory. Proceedings of the Social Statistics Section, American Statistical Association, 1976, 366-370.

Hathaway, S. R. and Briggs, P. F. Some normative data on new MMPI scales. Journal of Clinical Psychology 1957, 13, 364-368.

Horst, P. Factor analysis of data matrices. New York: Holt, 1965.

Table 1
Identification of Item Numbers with
Seven MMPI Factor Scales (Male)
(MMPI Normative Data)

I						II	III	IV	V	VI	VII
13	123	244	328	375	506	46	257	6	37	2	3
15	129	248	331	381	507	50	289	30	131	20	7
24	136	252	332	382	509	58	497	45	281	60	8
27	138	259	333	383	511	98	501	90	302	114	9
28	139	266	337	384	517	115	502	95	329	119	36
32	146	269	338	386	518	249		111	432	130	79
33	147	275	339	388	526	287		118	524	152	91
34	158	278	341	389	530	483		135		161	153
35	162	280	343	390	531	558		198		174	160
40	171	282	345	392	543			215		190	163
41	172	284	346	395	551			225		214	170
42	182	290	348	397	553			231		242	176
44	184	291	349	398				255		310	178
48	191	292	350	404				427		330	188
61	194	293	351	406				430		533	230
66	197	296	352	418				446		540	243
82	200	299	354	426				488			318
84	202	301	355	438				490			353
85	205	303	357	442				548			367
97	210	305	358	448							379
100	224	307	359	469							399
104	226	312	365	472							401
109	234	314	366	473							412
117	239	323	368	499							479
121	241	325	374	505							521
											522

Table 2
Identification of Item Numbers with
Four MMPI Factor Scales (Female)
(MMPI Normative Data)

I						II	III	IV
3	129	247	335	385	506	2	37	170
5	136	251	337	388	509	21	55	348
13	138	252	338	389	511	30	60	429
15	142	259	340	390	525	45	68	521
16	147	265	343	392	526	80	103	534
22	148	266	344	396	530	99	130	546
24	157	267	345	397	531	111	133	
29	158	273	350	398	535	135	153	
32	163	278	351	407	543	181	154	
40	166	284	352	411	551	208	175	
41	171	290	354	414	553	231	187	
43	172	292	356	416	560	285	193	
62	179	301	357	418	564	308	214	
67	186	303	359	421		378	281	
72	189	305	360	439		391	294	
76	190	307	361	442		427	302	
82	191	314	362	443		446	330	
84	201	315	366	448		452	460	
86	214	316	368	465		457	464	
94	217	317	374	468		481	478	
106	224	321	375	475		490	486	
108	234	322	377	487		527	496	
109	241	326	382	489		548	540	
117	244	328	383	492				
120	245	333	384	499				

AN INDIRECT FACTOR ANALYSIS OF A 300 ITEM ACHIEVEMENT TEST

William H. Resha, University of Alabama

Introduction

The use of factor analytic techniques has been known since the latter half of the nineteenth century. However, due to the mathematical complexities of this technique, factor analysis on large data matrices have been limited to factoring subscales or using Q-methodology when test items exceed more than 80 to 100 variables. Such substitute methods raise many questions concerning potential inaccuracies.

Due primarily to the work of Barker and Barker in the 1970's, computer programs have been developed to handle large data matrices, using an indirect factor analysis model presented by Horst (1965). Horst's (1965) model is based on three steps:

1. The sets of variables is reduced to a limited number of totals by grouping individual variables in some unspecified order.
2. Factor analysis is applied to the matrix of totals.
3. Matrix operations are used to estimate the factor loads for individual variables.

However, until recently there has been no way to measure how practical his model was. The problems in developing a computer program appeared insurmountable, but Barker succeeded in implementing a computer program which has been refined and updated over a period of years. An article by Barker and Barker (1975) gives a complete summary of the problems and successes they have encountered.

The General Board of Examining Chaplains of the Episcopal Church has developed a 300 item instrument which is administered to seminary students during the last year of their academic training prior to ordination. The instrument, General Ordination Examination (GOE), has been used for the past three years (1975, 1976, 1977). Thus far there has been no validation nor reliability studies done on this instrument.

The purpose of this study is to provide some initial data on this instrument by factor analyzing it using Barker's indirect factor analysis programs based on Horst's (1965) model.

The study hypothesis is that the six subtests of the GOE measures six distinct areas of Bible related content and knowledge the factor analysis of this instrument will also function as a theory generator for the possible future refinement of the GOE.

Methodology

The Population:

Subjects of this study were students who were trying to become ordained Episcopal priests. The majority of the students were third year seminary students in attendance at the 11 Episcopal theological seminaries in the United States.

Population represented the vast majority of all of the students who had taken the GOE since its inception in 1975. The population consisted of 786 students. Cattell (1966) has set a criterion for number of subjects required in factor analysis at 100 subjects plus the number of variables. If using this criterion, factor analysis of the GOE would require a minimum of 400 subjects.

The GOE was administered in a standardized form to all of the students in group settings. Security of test questions was maintained at a maximum.

General Ordination Examination:

The GOE is a 300 item objective multiple choice examination. In its present form, it has six subtests covering the following areas: Old Testament (60 questions); New Testament (60 questions); Church History (60 questions); Theology (60 questions); Ethics (30 questions); and Liturgics (30 questions). There are four alternatives to each question. Subtest questions are intermixed with one another. However, there is a pattern in the way the questions are listed. The scale is set up in the following pattern, 1 2 3 4 5 6 1 2 3 4 1 2 3 4 5 6 1 2 3 4, etc.

The GOE has not been normed nor has it been examined for reliability and validity. The Educational Testing Service (ETS) served as consultants during the development of the instrument.

The GOE is basically an achievement test since its primary function is to estimate the person's present knowledge of previous course work (Anastasi, 1976). Further, the test is considered norm-referenced since its results are interpreted in relation to other students performance who took the test (Anastasi, 1976; Aiken, 1971). There was no deliberate weighting of test items in reference to item difficulty. Test items are treated as quantitative variables with the data being considered as dichotomized normal. The test items will be interpreted as discontinuous dichotomous data (Ghiselli, 1964; Edwards, 1972).

The General Board of Examining Chaplains of the Episcopal Church has not formally stated a theory for the way this test is set up. It will be assumed that the theory that is implied is that the GOE tests the student's knowledge in the area of church history, literature and vocabulary of the christian tradition. This theory is encompassed in the six subtests that have been previously mentioned.

Methodology:

The 786 subject data base was factor analyzed by an indirect method to arrive at an estimate of a conventional principle axis solution, and this solution was rotated to a varimax criterion.

The raw data were punched on IBM cards and were then transposed to dichotomous data using a scoring key developed by the church. The dichotomized scored data was in the form where 1's equaled an incorrect or omitted item and 2's represented a correct response.

A SPEC01 program (Barker and Barker, 1977) listed and numbered all of the students. This allowed for a visual check on all data cards to insure they were in proper order, had the proper number of cards per student (five cards per student) and were aligned in the proper columns. In an effort to arrive at homogeneous items for clustering into totals, a Q-analysis program (CORR98) was used on 20 subjects across the 300 items. The factor loads for the items were then estimated as in R-methodology. This 20 subjects factorization provided a rough grouping rationale for the totals used to start the iterative processes involved in the indirect factor analytic approach (CORR99).

A SPEC50 program (Barker and Barker, 1977) which generates random numbers was run to obtain the random sample of 20 subjects which was used as the seed data to be initially run with CORR98, (Barker and Barker, 1977).

Using the Eigenroots from the CORR98 a Scree Test was drawn to indicate the starting point from which programming could begin. CORR98 was then programmed to rotate six factors.

CORR99 (Barker and Barker, 1977) was then used to factor analyse the GOE's 300 items, using the six rotated factors as a starting point for clustering into totals. An item had to load ±.30 or better on only one factor before it was identified with that factor. These results were compared with the theorized factor structure as presented in the GOE.

The information measure D (Relative Uncertainty Reduction) was selected to provide an objective measure of the degree to which theorized dimensions reflect estimated factor structure of the data set. In an ideal solution, all entries in the matrix of factors and item subsets would appear in the diagonal. Such a solution would indicate complete agreement of items subsets (totals) with actual factor structure. Frequently, however, certain items are found to load inappropriately. Items which fail to load as expected on factors appear as false negatives. Those which load into factors contrary to expectations appear as false positives. The D measure expresses the relationship between rows (subsets of items) and columns (factors). Use of this statistics also permits comparison between theories of degree of agreement between a priori item subsets and actual factor structure.

Results

CORR98 identified six factors as the most parsimonious initial set to use. The decision was made that a factor had to have at least three or more items loading on it before it would be considered for clustering into a total.

CORR99 was programmed for 20 iterations using six factors. The computer reached convergence at 12 iterations for maximum factor alignment.

The indirect factor analysis method identified three general factors which make up the greater part of the GOE and which contribute the most to the D value (.87). These three general factors are labeled as follows: Factor I = General Bible Content; Factor V = Historical Theology; Factor VI = Contemporary Theology.

In addition, there were three other factors that had some item loadings. The number of items loading on these factors appear to be inconsequential when the total number of items (300) on the GOE is taken into consideration.

The data indicates most impressive results, obtaining a D = .87. This D value indicates that item subsets identified by the Horst's (1965) indirect factor analysis method reflect actual factor structure. The higher D value is reflected in the lower number of false negatives (98).

There were two items which loaded at ±.30 or better on two factors (item 205 loaded on factors V and VI; item 232 loaded on factors I and V).

CORR99 was also run using scoring on subtests as hypothesized by the Episcopal

Church. Six totals consisting of 60 items each for the first 4 totals and 30 items each for the other two were used, according to the scales. The theory hypothesized by the Episcopal Church does not appear to have been substantiated by the data, which indicates a D measure of .32 which represents a weak agreement between the theorized item clusters and actual factor structure. The partial nature of the theory is reflected by the large number (206) of items, expressed as false negatives which failed to load appropriately on factors.

Conclusion

The 300 item GOE as developed by the Episcopal Church appears to have three factors which measures general knowledge in Historical Theology, Contemporary Theology and Bible Content. A total of 135 items loaded on one of six new factor structures. Of these 135 items, 84% of them (114) loaded on one of the three main factors.

The results from Horst's (1965) indirect method of factor analysis appears to be quite impressive and persuasive in reference to the possibility of either reducing the GOE down to a smaller test or revising the examination, improving the questions so more of the items load on the identified factors. Out of the 300 items, 45% loaded on one of six new factors and 38% loaded on one of three major factors. This indicates that well over 50% of the present items are not loading on any factors (subtests) with which they were initially identified. It is suggested, therefore, that until further revision is made of the GOE, caution needs to be stressed if it is used in a decision making capacity.

Results of this study support earlier studies previously cited by Barker and Barker and Hamlett (1976) in demonstrating usefulness of the indirect method as an appropriate technique in evaluating theories regarding factor structure of large data sets.

It is further noted that the overall computer cost of using the Barker and Barker programs remain at a minimum when the number of variables is considered. For example, using six factors, 300 variables, 786 subjects (five cards per subject) and 12 iterations, it took a total of in/out time of 8 1/2 minutes at a cost of $110.45. This further substantiates previous Barker articles cited in reference to the economy of Horst's (1965) model as programmed by CORR98 and CORR99 (Barker and Barker, 1977).

The present study demonstrates the usefulness of performing factor analysis on psychological tests as a first step in the overall process of test validation.

Two further conclusions appear appropriate from this study. First, the indirect factor analytic method as presented has proven its flexibility. It is noted that all of the computer runs were made during normal operating hours and it was not necessary to shut down other projects or handle the Barker program in any special manner by the computer center.

Finally, regardless of the significance of the reported findings, it is significant that someone with relatively limited training in statistics could be successful in completing such a project. The relative simplicity of the Barker programs will allow future researchers in the behavioral sciences this additional tool in their empirical investigations.

References

Aiken, L. R., Jr. Psychological and Educational Testing. Boston, Mass.: Allyn and Bacon, Inc., 1971.

Anastasi, A. Psychological Testing (4th Ed.). New York: Macmillan, 1976.

Barker, H. R., Barker, B. M. and Carlton, B. B. Accuracy Of An Indirect Factor Analytic Solution As A Function Of The Method Of Assigning Variables To Totals. Paper presented at American Statistical Association, Atlanta, 1975A.

Barker, H. R. and Barker, B. M. Summary of research on a novel, indirect factor analytic solution. Paper presented at American Statistical Association, Atlanta, 1975B.

Barker, H. R. and Barker, B. M. Behavioral Sciences Statistics Program Library. (2nd Ed.). University of Alabama, 1977.

Barker, H. R. and Barker, B. M. An indirect factor method for testing the dimensionality of large data sets. Paper presented at Southeastern Psychological Association, New.Orleans, 1976.

Barker, H. R., Fowler, R. D., and Peterson L. P. Factor Analytic Structure of the Short-Form MMPI Items. Journal of Clinical Psychology, 1971, 27, 228-233.

Cattell, R. B. Extracting the correct number of factors in factor analysis. Educational and Psychological Measurement, 1958, 18, 791-838.

Edwards, A. L. Experimental Design in Psychological Research (4th Ed.).

New York: Holt, Rinehart and Winston, Inc., 1972.

Ghiselli, E. E. Theory of Psychological Measurement. New York: McGraw-Hill, 1964.

Horst, P. Factor Analysis of Data Matricies. New York: Holt, 1965.

Hunnicutt, B. M. and Barker, H. R. Application and evaluation of a novel and indirect factor method to a very large data matrix. Paper presented at Southeastern Psychological Association, Hollywood, May, 1974.

Sloan, H. C. A simplified procedure for estimating the factor structure of large data matrices. Unpublished doctoral dissertation. University of Alabama, 1973.

Stallings, N. A. Evaluation of an indirect method of estimating the factor structure of large data matrices. Unpublished doctoral dissertation. University of Alabama, 1973.

Table 1

Sex and Population Distribution			
Year	Male	Female	Total
1975	189	30	219
1976	227	37	264
1977	224	59	303
			786

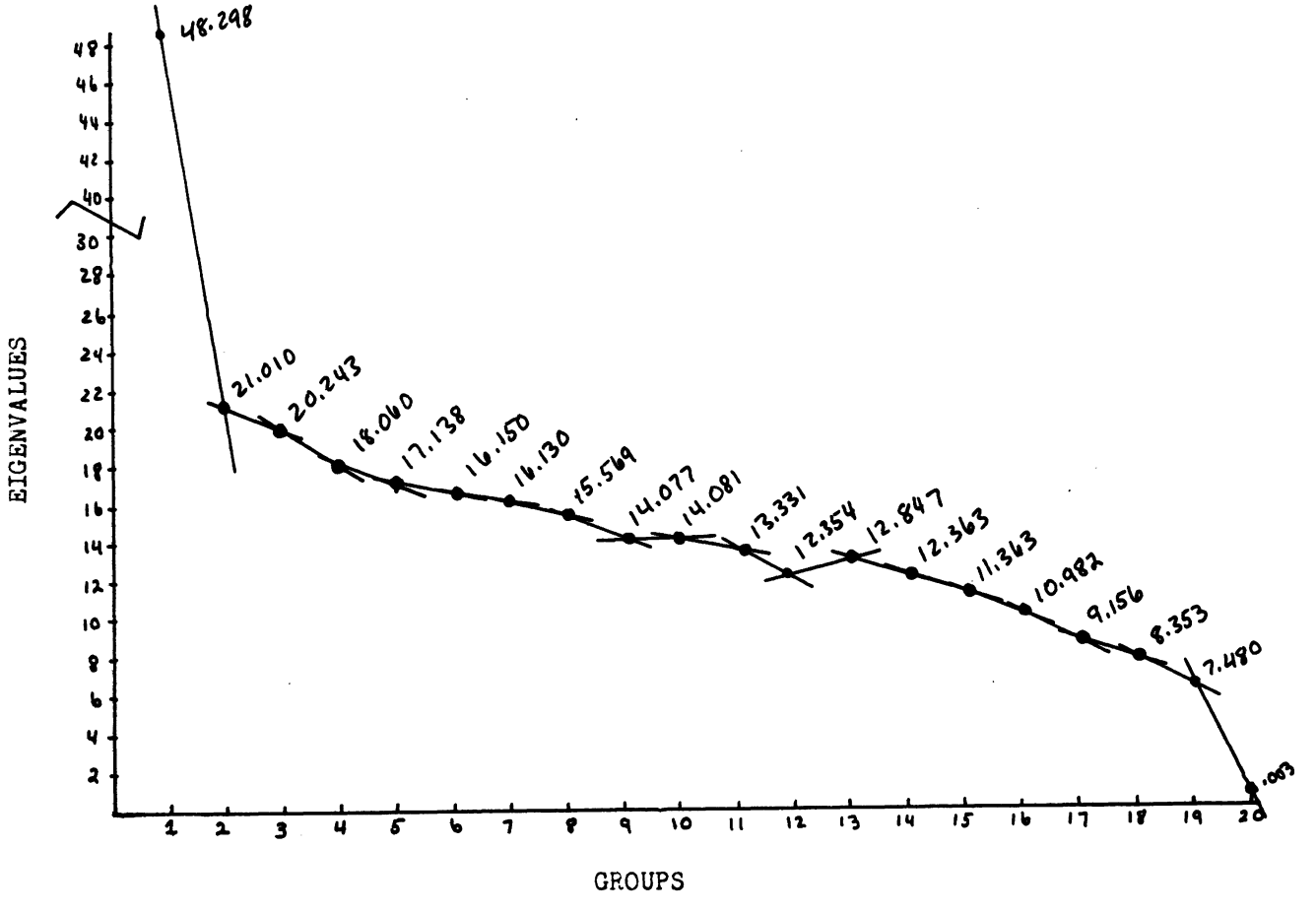


Figure 1. Scree Test on Eigenvalues of General Ordination Examination

Table 2

Year	Subject #	Total Sample
1975	51 107 121 185 203	5
1976	256 362 382 443 465	5
1977	501 508 512 517 548 551 664 680 694 724	10
		20

Table 3

Items Loading ±.3 Or Better Onto One Of Six Factors

Factor I (Gen. Bible Content)	Factor II (Specific Identification)	Factor III (Biblical Identification)	Factor IV (Church Definitions)	Factor V (Historical Theology)	Factor VI (Contemporary Theology)
1	187	12	53	9	56	284	5
8	188	21	57	105	129	287	10
11	192	44	103	133	180	289	35
17	197	72	119	139	186	293	14
27	198	256	135	144	196	294	38
31	212		141	175	204	296	45
32	217		286	223	206	298	50
37	218			255	211	299	75
48	237			275	216	300	83
51	288				229		89
61	242				230		96
97	246				231		99
98	257				235		100
102	288				248		110
107					249		120
108					253		140
111					258		170
117					263		179
118					264		184
122					265		185
127					266		189
131					268		200
138					270		210
147					271		220
148					272		221
152					273		227
157					276		240
161					277		245
172					278		260
181					281		285

Table 4

Association Between Item Subsets and Varimax Factors
(Actual Data Results)

	I	II	III	IV	V	VI	False -	Sum
1	41	0	0	0	0	0	0	41
2	0	5	0	0	0	0	0	5
3	0	0	7	0	0	0	0	7
4	0	0	0	7	0	0	0	7
5	0	0	0	0	29	0	0	29
6	0	0	0	0	0	27	0	27
False +								
	2	0	0	0	8	1	98	109
Sum	43	5	7	7	37	28	98	225

HX = 2.135704 HY = 2.214495 HXY = 2.417882 D = .87258

Table 5

Association Between Variable Subsets and Varimax Factors
(Church Hypothesized Data Results)

	I	II	III	IV	V	VI	False -	Sum
1	26	0	0	0	0	0	34	60
2	0	21	0	0	0	0	39	60
3	0	0	17	0	0	0	43	60
4	0	0	0	12	0	0	48	60
5	0	0	0	0	8	0	22	30
6	0	0	0	0	0	10	20	30
False +								
	0	0	0	0	0	0	0	0
Sum	26	21	17	12	8	10	206	300

HX = 2.521930 HY = 1.670171 HXY = 3.398042 D= .31486

A FACTOR ANALYSIS OF THE ITEMS OF THE EPPS

Barbara M. Barker and Nancy G. Williams
University of Alabama

Murray and others (1938) proposed 40 needs of which Edwards (1954) chose to develop 15 into a measure of personality, the Edwards Personal Preference Schedule (EPPS). To control for social desirability, Edwards arranged the items of the inventory into pairs matched in terms of their social desirability ratings. The results are ipsative, i.e., the 15 needs scores sum to a certain fixed constant. This format has produced different opinions as to how factor analysis should be applied (Horst & Wright, 1959; Tatsuoka, 1971). Sherman and Poe (1959) used a normative EPPS based on 135 distinct items rated on a nine-point Likert-type scale. Others (Dixon & Ahern, 1973; Heilizer, 1963; Levonian, et al, 1959) have been content to factor only the scale scores.

The 225 items in the test actually represent 450 separate statement choices. This large a set of items or variables has posed a problem for factor analysis because of computer limitations. Normally, 80 to 100 variables are all that can be factor analyzed. Barker and Barker (1977) developed a computer program (CORR99) based on a mathematical model by Horst (1965) that will factor very large numbers of variables. This program has been revised and tested over a number of years (Barker & Barker, 1975). The indirect method has been used for theory testing and has proved to be quite useful (Barker & Barker, 1976(a) and (b); Hamlett, 1976). Recent refinement of the indirect factor method makes possible use of the method without the need of a priori theory.

This paper proposes to analyze subject preference on the 225-item EPPS by means of the indirect factor method. The items will be treated as though the subject either agreed with or did not agree with a fixed alternative.

METHODOLOGY

The 225-item EPPS was administered to two groups. One group consisted of randomly selected female teachers in Alabama who responded to a mailed questionnaire. The other half of the 315 subjects were graduate level students taking a course in statistics.

Responses to the 450 test items were punched onto computer cards. If a subject chose an A or a B alternative, it was represented by a 1. If the corresponding A or B were not chosen, or if neither alternative were chosen, it was represented by a zero. A Univac 1110 computer was used for all analyses.

The first step in factor analysis by the indirect method (CORR99) required the items to be grouped for totaling. The more homogeneous the items in factor structure, the better. To identify clusters for totaling, an obverse factor analysis (CORR98) was applied to the data for 20 randomly selected individuals. A SCREE test (Cattell, 1966) was used to determine the number of factors to retain. The items were then pooled according to the results of the varimax rotation and these items were used to begin the iterative indirect factor analysis. A criterion of $\pm.3$ or greater on only one factor was used to identify items with factors.

RESULTS

The indirect factor analysis program (CORR99) was begun with eight totals. After the 14th iteration, one total was dropped due to a lack of any items loading $\pm.3$ or greater on the eighth factor. Of the seven factors remaining, one had too few items for interpretation (2). Table 1 shows the item numbers associated with the factors.

The factor analysis required 14 minutes using 128K core storage for a total cost of $72.69. Twenty-five iterations were performed, and the measure of association D increased at each iteration. The D measure was used to determine the degree of agreement between the items contained in totals and the items which loaded $\pm.3$ or greater on a particular factor. The D measure for the 25th iteration was .70.

Factor I contained 43 items which seemed to represent an interpersonal orientation. Twenty of the items came from the Need for Heterosexuality subscale and others came from the Succorance, Dominance, and Affiliation subscales.

Factor II contains twenty items which appear to be measuring assertive aggressiveness. Items relating to aggression, autonomy, exhibitionism and dominance were included on this factor.

Factor III appears to measure an anti-social attitude (six items). It is the smallest interpretable factor. Items relating to telling others how to do their jobs and avoiding responsibilities and obligations define this dimension.

Factor IV, with items such as, "I like to do new and different things.",

appears to measure the need for change with 16 items.

Factor V contains only two items (10, 44) and was not interpreted. Nine items loaded on Factor VI. These items suggest submissiveness based on items relating to feeling inferior and needing encouragement.

On Factor VII, twenty-one items from the subscales of achievement, endurance, and order appear to represent a need for personal responsibility. Of the total of 225 items, approximately half (107) of the items did not appear on any factor.

DISCUSSION

Edwards designed the PPS to measure 15 different needs. Factor analysis of the test failed to support the theorized structure. Instead, six factors appear to measure personality traits suggested by items from different Edwards subscales.

Sherman and Poe (1969) factor analyzed items in a normative version of the EPPS and found four main factors. Three of these factors were replicated by factoring the need scale scores.

Factors I and II appear to be virtually identical to factors labeled Interpersonal Orientation and Assertive Aggressiveness by Sherman and Poe. One of our factors, although resembling their Persistence-Dependence factor, also contained many achievement-oriented items. We suggest that a better name for this factor might be Personal Responsibility.

A factor measuring preference for change and one measuring submissiveness were identified. The last factor, considered minor because it contained only six items, appears to be measuring an antisocial attitude.

Horst and Wright (1959) found no essential difference in using forced-choice over a rating scale type of instrument. Tatsuoka (1971) suggests an initial factorization of ipsative instruments and subsequent use of factor scores. In comparing the results of the factoring of the 225-item ipsative version of the EPPS with the 135-item normative version, support was found for Horst and Wright's stand.

Levonian and associates (1959) evaluated the EPPS items (by scales) statistically and found "...an unexpectedly large discrepancy between what the PPS is designed to measure and the actual item factorial content." They objected to the repetitive nature of the forced-choice format. Because factoring subscales scores or items within single subsets cannot estimate complete factor structure, past factor analyses of the EPPS have been inconclusive. This study, which utilized all items of the test, clearly reveals a factorial structure different from what was intended by the test maker.

SUMMARY

A factor analysis of the 225 items of the EPPS identified six factors. These factors appear to be measuring: (1) Interpersonal Orientation, (2) Assertive Aggressiveness, (3) Personal Responsibility, (4) Change, (5) Submissiveness, and (6) Antisocial Attitude.

REFERENCES

Barker, B. M., and Barker, H. R. Evaluation of theorized factor structure of the MMPI for male and female populations. Proceedings of the Social Statistics Section, American Statistical Association, 1976, 174-178. (a)

Barker, H. R., and Barker, B. M. An indirect method for testing the dimensionality of large data sets. Proceedings of the Social Statistics Section, American Statistical Association, 1976, 179-183. (b)

Barker, H. R., and Barker, B. M. Behavioral Sciences Statistics Program Library. Tuscaloosa, Al.: University of Alabama Reproduction Services, 1977.

Barker, H. R., and Barker, B. M. Summary of research on a novel, indirect factor analytic solution. Proceedings of the Social Statistics Section, American Statistical Association, 1975, 298-301.

Cattell, R. B. (Ed.) Handbook of Multivariate Experimental Psychology. Chicago: Rand-McNally, 1966.

Dixon, P. W., and Ahern, E. H. Factor pattern comparisons of EPPS scales of high school, college, and innovative college program students. Journal of Experimental Education, 1973, 42, 23-35.

Hamlett, C. C. Validation of the theorized factor structure of the Personal Orientation Inventory. Proceedings of the Social Statistics Section, American Statistical Association, 1976, 366-370.

Heilizer, F. An ipsative factor analysis of the ipsative EPPS. Psychological Reports, 1963, 12, 285-286.

Horst, P. Factor Analysis of Data Matrices. New York: Holt, 1965.

Horst, P., and Wright, C. E. The comparative reliability of two techniques of personality appraisal. Journal of Clinical Psychology, 1959, 15, 388-391.

Levonian, E., Comrey, A., Levy, W., and Proctor, D. A statistical evaluation of Edwards Personal Preference Schedule. Journal of Applied Psychology, 1959, 43, 355-359.

Murray, H. A., and others. Explorations in Personality. New York: Oxford University Press, 1938.

Sherman, R. C., and Poe, C. A. Factor analytic scales of a normative form of the EPPS. Measurement and Evaluation in Guidance, 1969, 2, 243-248.

Tatsuoka, M. M. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: Wiley & Sons, 1971.

Table 1

Items loading ±.3 on only one factor.

I		II	III	IV	V	VI	VII
3	142	72	25	51	10	36	1
9N	143	75	44	56	77	47	4
18	153	91	74	57		49	29
27	154N	93	119	58		50	53N
28	159N	96	175	59		52	76
38	164N	97	195	60		114	81
43	168	98		132		125	83
66	173	106		133		177	86
67	184N	116		134		188	87
68	186	146		135			88
69	193	147		158			136
70	197N	166		183			137
73	199N	167		201			138
84N	204N	171		203			140
89N	209N	180N		206			151
105	214N	181		207			155
112	216	190N					156
113	217	191					161
115	218	192					196
117	219	223					211
130	220	225					215
141							

AN ALTERNATIVE METHOD FOR EVALUATING SUBJECTIVELY INTERPRETED TEST DATA

Carolyn Minder, Northeast Louisiana University
Betty Carlton, The University of Alabama
Charles Minder, Northwestern State University

Introduction

Many standardized tests in use today involve subjective interpretation of results. In clinical settings or when decisions are to be made about a patient, individual interpretation of results is necessary. On some occasions, however, subjective interpretation of findings may be inefficient. A method of test interpretation which utilizes statistical procedures rather than subjective clinical judgement would be useful in dealing with large numbers of subjects and in studies in which the researchers are interested in groups of people rather than in individuals.

Sources of Data

The sample consisted of 500 subjects, students enrolled in graduate and undergraduate programs of several colleges in a university in the southeastern United States. These people were asked to respond on a voluntary basis to the Personal Orientation Inventory (POI), an instrument designed by Everett L. Shostrom to measure characteristics of self-actualization. The POI consists of 150 items which yield 12 scores purported to reflect various dimensions of self-actualization.

The POI was selected as the instrument for this study because of the manner in which results are interpreted. Test results are scored objectively in that totals are obtained for items in each scale. Profiles are interpreted by comparing them to sample profiles described in the POI manual. A number of profiles are presented including those for college students, Peace Corp Volunteers, psychopathic felons, hospitalized persons, and others. These profiles show typical scores obtained by poorly functioning, normally adjusted, and self-actualized persons. The clinician compares the subject's profile with those in the manual and makes a subjective judgement as to the person's adjustment.

Methodology

Test results were scored using a computer program written by Dr. Harry Barker of the University of Alabama. This program obtained totals of items for each of the twelve scales of the POI. Other computer programs used were also written by Dr. Barker (1973). Ward's Hierarchical Grouping Technique was applied to test score variables. As described by Ward and Hook (1961), this technique is used to group test profiles so as "to maximize the homogeneity of profiles within the same clusters, taking into account of all profile variables and all clusters at the same time" (p. iii). Ward's Hierarchical Grouping Technique is used appropriately with measures of profile similarity and does not require prior formation of nucleus groups.

The computer program (CORR23) used has a subject limitation of 350 subjects. Due to the fact that the total number of subjects in the present study exceeded this number, two applications of Ward's Hierarchical Grouping Technique were required. Two groups were formed by combining results of these applications.

To test the appropriateness of each subject's placement within the designated groups, a discriminant analysis program (CORR06) was used to classify subjects. CORR06 reports a probability term associated with the largest discriminant function. The higher this term, the more likely the subject belongs to the designated group. Application of this type of discriminant analysis was required as a check on subject placement since groups were formed on the basis of combining results of two different applications of Ward's Hierarchical Grouping Technique. Finally, a second discriminant analysis program (CORR20) was used to test the discriminating power of the variables (scale scores) between the two groups.

Two POI scores, Time Competence and Inner-Other Support, are reported in terms of ratios. Shostrom believes that response on these dimensions is best represented as position on a continuum. The other ten scores represent totals of items within each of the ten profile scores. In this study the two Ratio scales were treated as totals, rather than ratios. Results of discriminant analysis with the ratio scales as variables are reported separately from discriminant analysis for which the ten scale scores were variables.

Results

Results of applications of this technique to data of 350 subjects resulted in two groups composed of 173 and 177 subjects, respectively, accumulated error = 96.0099. Ward's Hierarchical Grouping Technique applied to the remaining data of 150 subjects resulted in two groups of 69 and 81 subjects, accumulated error = 46.5457. Results of the discriminant analysis run to test appropriateness of each subject's group placement indicated a very small percentage (10%) had been grouped inappropriately by Ward's.

Results of the conventional discriminant analysis indicate that groups formed on the basis of Ward's Hierarchical Grouping Technique were significantly separated by the profile score variables. F test on Wilks Lambda was found to be significant at the .01 level of confidence.

Insert Table 1 about here

Results of the univariate F tests indicate that each of the variables differentiated (P .01) between the groups.

Insert Table 2 about here

Visual examination of data for the two groups indicates that average scores of individuals within the two groups are roughly comparable to scores obtained by poorly functioning and normally functioning subjects as described

by Shostrom. The POI manual states, "self-actualized groups are significantly higher on all scales and nonself-actualized groups tend to be lower on all scales. Normal groups tend to score in between" (Shostrom, 1972, p. 21).

Insert Table 3 about here

Discussion

Application of the procedures described in this paper provides an alternative method for interpretation of test data. Applications of this technique are not limited to POI data and could be made to similar types of tests when scores are not to be interpreted on an individual basis. This efficient procedure would be most useful in dealing with large numbers of subjects in which groups are defined in liberal terms, rather than in cases in which each member of a group must be precisely described. It is recognized that in many cases application of these procedures would be inappropriate and, in such cases, individual clinical judgement of the psychologist would be the appropriate method used to evaluate test data.

Results of discriminant analysis in classifying subjects indicate impressive accuracy (90%) of subject placement in groups by Wards, while results of the second discriminant analysis which examines relationships between variables in groups indicate that the procedures applied result in groups significantly separated by profile variables. Since profile scores were used as grouping variables, it is not surprising that discriminant analysis reveals that groups were significantly separated when scores from the POI are used as independent variables. However, when results of conventional discriminant analysis are treated as a statistical test of the success with which Wards and classifacatory discriminant analysis form groups, this statistic is found to contribute in an important way to data analysis.

While results of applying statistical procedures to POI data appear to be impressive, these findings would be meaningless if the groups formed based on these applications bore little resemblence in terms of average subject scores to Shostrom's profile descriptions of similar groups. For this reason examination of Table 3 is particularly relevant. The higher scoring group bears remarkable resemblance to Shostrom's descriptions of typical college students while the scores for the lower group appear similar to those of less self-actualized persons, represented by POI profiles for entering college freshmen and alcoholic males. The POI profile for alcoholic males was selected to represent those of poorly functioning persons as was that for POI entering college freshmen (male and female). These profiles are very similar to the lower scoring group's profile with this study data. According to self-actualization theory younger people as a group are less fully functioning that are mature adults. Therefore, scores obtained by older college students which bear resemblance to those of entering freshmen indicate that these older people appear to be relatively poorly adjusted.

In conclusion, examination of results of this study indicate that statistical procedures can be used as an alternative to subjective interpretation of test data in certain circumstances. Groups were formed on the basis of applying statistical procedures to test data rather than by relying on clinical judgement to form groups. Scores obtained by subjects in the two groups appear to be similar to scores described by the POI manual as being typical of poorly adjusted and normally adjusted people.

References

Barker, H. R. Behavioral sciences statistics program library. University Al.: University of Alabama Reproduction Services, 1973.

Shostrom, E. L. Manual for the Personal Orientation Inventory. San Diego: Educational and Industrial Testing Service, 1972.

Ward, J. H., & Hook, M. E. A hierarchical grouping procedure applied to a problem of grouping profiles (ASD-TN-61-55). Lackland Air Force Base, Tex.: Personnel Laboratory, Aeronautical Systems Division Air Force Systems Command, October 1961.

Table 1

Results of Conventional Discriminant Analysis
for Two Groups on POI Data

Variables	Wilks Lambda	DF-B	DF-W	F*	X^2	DF
Two Ratio Scales----	.430	2	0	.000	419.771	2
Ten Profile Scales--	.537	10	489	42.195	307.328	10

*For 2 groups and 2 variables the F ratio is not correct. Chi square for the discriminant root is interpreted.

Table 2

Univariate Analysis of POI Variables
$df_b = 1$, $df_w = 498$

Variable	MS-B	MS-W	F
Time Competence (Tc)	2169.5166	5.0584	428.8937
Inner-Other Support (I)	28884.7812	74.4766	387.8370
Self-Actualizing Value (SAV)	1145.7891	8.4041	136.3377
Existentiality (Ex)	2602.4941	13.5842	191.5828
Feeling Reactivity (Fr)	874.4678	7.9401	110.1332
Spontaneity (S)	922.5137	5.3047	173.9036
Self Regard (Sr)	860.5361	4.5992	187.1070
Self Acceptance (Sa)	828.1611	7.6614	108.0959
Nature of Man (Nc)	262.3179	4.3450	60.6519
Synergy (Sy)	148.1802	1.7141	86.4479
Acceptance of Aggression (A)	976.9150	8.0493	121.3665
Capacity for Intimate Contact (C)	2099.8223	10.3878	202.1427

Table 3

Comparison of Average Scores and Standard Deviation
for Study Data and for POI Manual Data

POI Scales:	Tc	I	SAV	Ex	Fr	S	Sr	Sa	Nc	Sy	A	C
Average Scores for Study Sample, Poorly Functioning People												
Mean:	13.1	73.0	17.1	16.6	14.0	10.2	10.3	13.0	10.5	6.4	14.4	15.2
S.D.:	2.6	9.0	3.3	3.5	3.0	2.4	2.4	2.9	2.4	1.5	3.1	3.4
Average Scores for Study Sample, Normally Adjusted People												
Mean:	17.3	88.3	20.4	21.2	16.7	13.0	12.9	15.6	11.9	7.5	17.1	19.3
S.D.:	1.9	8.2	2.5	3.9	2.6	2.2	1.8	2.6	1.7	1.1	2.5	3.0
Average Scores for College Sample, POI Manual, Males												
Mean:	15.1	75.6	18.8	16.7	13.8	9.7	11.5	13.7	11.6	6.3	15.1	15.6
S.D.:	2.9	8.9	2.6	4.4	2.9	2.2	2.2	3.1	2.0	1.4	3.0	3.4
Average Scores for College Sample, POI Manual, Females												
Mean:	16.2	76.0	19.1	17.2	13.7	9.6	11.5	14.3	11.9	6.6	15.0	15.6
S.D.:	2.7	9.7	3.4	4.2	2.8	2.4	2.3	2.8	1.9	1.3	2.9	3.3
Entering College Freshmen (Male and Female), POI Data												
Mean:	15.1	75.6	18.8	16.7	13.8	9.7	11.5	13.7	11.6	6.3	15.1	15.6
S.D.:	2.9	8.9	2.6	4.4	2.9	2.2	2.2	3.1	2.0	1.4	3.0	3.4
Alcoholic Males, POI Data												
Mean:	13.0	73.6	18.4	16.6	14.2	8.7	9.9	13.8	11.2	5.6	13.8	15.6
S.D.:	3.2	9.9	2.4	4.1	2.2	2.4	2.7	2.5	2.0	1.7	2.5	4.2

DETERMINATION OF LIKELIHOOD OF BELONGINGNESS OF AN INDIVIDUAL TO A NORM GROUP BY MEANS OF STANDARD SCORES ON INDEPENDENT PSYCHOLOGICAL MEASURES

Ajit Kumar Mukherjee, Corpus Christi State School
Texas Department of Mental Health and Mental Retardation

In psychology it is always a problem to determine with some certainty whether an individual belongs to a group or not. An example may make the issue clear. Suppose that Mr. A takes a test composed of various subtests to determine A's suitability for a training program. The test is designed to measure various independent abilities needed to complete the training program successfully. The test A has taken has been standardized using a sample of people who completed the training program successfully (norm group). There are three possibilities:

1. A's abilities may be above the norm group.
2. A's abilities may be like the other members of the norm group.
3. A's abilities may be below the norm group.

The question is often asked --- Does A belong to the norm group? Cronbach and Glesser (1953) recommended a model for assessing similarity between profiles designed to handle the question "How similar is Person 1 to Group Y?" Cronbach and Glesser - D^2 and Mahalanobis - D^2 are identical to each other when variates are standardized and uncorrelated (Cronbach and Glesser, 1953). However, for studying groups of persons, Cronbach and Glesser - D^2 did not prove to be very effective. In 1928, Pearson recommended "coefficient of racial likeness" which was designed to measure the similarity between two groups or the similarity of an individual to a group. Unfortunately, Pearson's index proved unsatisfactory. Since Pearson, several techniques have been recommended by several individuals. Cattell (1949) introduced the concept of r_p as a coefficient of pattern similarity. One of the assumptions of Cattell was that for computing r_p, variates were needed to be uncorrelated. Williams (1969) in his study found that moderately correlated variates could be used effectively for computing r_p.

When statistics only deal with probability, all statistical models deal with 'chance'. And, no statistical model is designed to predict certainty.

In this article, an attempt has been made to deal with the issue of chance of an individual to belong to a group.

PROCEDURE

The following theorems have been used to develop the technique:

1. If the random variable X is $N(\mu,\sigma^2)$, $\sigma^2>0$ then the random variable $V = (X-\mu)^2/\sigma^2$ is χ^2 with df = 1.

2. Let $X_1, X_2 \ldots X_n$ denote a random sample of size n from a distribution which is $N(\mu,\sigma^2)$. The random variable $Y = \sum_{i=1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$ has a chi-square distribution with n degree of freedom.

3. $F = \frac{U/r_1}{V/r_2}$ has F distribution when U and V are independent chi-square variables with r_1, and r_2 degrees of freedom respectively.

MODEL

We have a test with K subtest such that each subtest measures independent psychological trait of an individual. A sample of n individuals are randomly selected to standardize the test. On each independent subtest standard score (Z-Score) of each n individual is calculated. Then we have a matrix,

	1	2	3	K
1	Z_{11}	Z_{12}	Z_{13}	Z_{1k}
2	Z_{21}	Z_{22}	Z_{23}	Z_{2k}
3	"	"	"	"
"	"	"	"	"
"	"	"	"	"
"	"	"	"	"
"	"	"	"	"
n	Z_n	Z_n	Z_n	Z_{nk}

squaring each Z - Score we have,

Z^2_{11}	Z^2_{12}	Z^2_{13}	Z^2_{1k}
Z^2_{21}	Z^2_{22}	Z^2_{23}	Z^2_{2k}
"	"	"	"
Z^2_{n1}	Z^2_{n2}	Z^2_{n3}	Z^2_{nk}

Now, Z^2_{ij} is chi-square with one degree of freedom. Since, each subtest is independent, $\sum_{i=1}^{k} Z^2_{ij}$ is chi-square with k degrees of freedom. Let that be written as χ^2_k.

Since, performance of each individual is independent of each other, $\sum_{i=1}^{n} \chi^2_k$ is chi-square with n·k degrees of freedom.

Now, let us consider a case who is not a member of the sample used for standardization and whose Z-scores are $Z_{m_1}, Z_{m_2} \ldots . Z_{m_k}$

$\sum_{i=1}^{n} {}_i\chi_k^2$ and ${}_m\chi_k^2$ are independent chi-squares.

A F-ratio can be obtained and F ratio can be defined as:

$$F = \frac{\sum_{i=1}^{n} \chi^2 \; i^k / n \cdot k}{\chi^2 \; m^k / k}$$

A F-table can be utilized to determine the significance of F with degrees of freedom as nk and k at .05 or .01 level. The obtain F ratio indicates to which extent between the subjects variability is larger than that of within subject variability. When F is significant, one may state that a significant positive correlation exists. Therefore, it may be concluded that there is a likelihood that the subject outside the "norm group" belongs to the norm group.

ILLUSTRATION

Let us consider a sample where n=4 and a test with three subtests (k=3). The following matrix represents standard score of each subject in each subtest.

S_1	-1.1	1.0	0.8
S_2	1.2	-1.3	0.9
S_3	1.4	2.0	1.3
S_4	2.0	1.7	-0.5

Square of Standard Score Matrix:

1.21	1	0.64	${}_1\chi_3^2 = 2.85$
1.44	1.69	0.81	${}_2\chi_3^2 = 3.94$
1.96	4.0	1.69	${}_3\chi_3^2 = 7.65$
4.00	2.89	0.25	${}_4\chi_3^2 = 7.14$

$\chi_{12}^2 = 2.85 + 3.94 + 7.65 + 7.14 = 21.58$

$\chi_{12/12}^2 = \frac{21.58}{12} = 1.79$

Let us consider two individuals with standard scores in the same test but not members of the "norm group".

S_5	1	0.5	2	
	1	0.25	4	$\therefore \chi_3^2 = 5.25$

$\chi_{3/3}^2 = \frac{5.25}{3} = 1.75$

What is the likelihood that S_5 belongs to the norm group?

$F = \frac{1.79}{1.75} = 1.02$ F with degrees of freedom 12 and 3, to be significant at .05 level needs to be 8.74 F is not significant. Likelihood of S_5's belongingness is not significant.

Let us consider:

S_6	0.3	0.5	0.2
	0.09	0.25	0.04

$\chi_3^2 = 0.38$ $\chi_{3/3}^2 = 0.13$

$F = \frac{1.79}{0.13} = 13.84$

Here, F is significant at $P < .05$ level.

Likelihood of S_6's belongingness to the norm group is significant.

DISCUSSION

Data from a research project (Title IVC, funded by Texas Education Agency) was used to determine the limitation of the model. A sample of N=600 with measures on seven independent variates was used. In some cases, variates were moderately correlated (Table 1).

	1	2	3	4	5	6	7
1	1	.33	.32	0.06	.36	.28	.13
2		1	.35	0.02	.37	.34	.14
3			1	0.07	.33	.28	.13
4				1	.10	.11	.07
5					1	.32	.16
6						1	0.10
7							1

Various random samples of different sizes and different number of variates were drawn from the original sample (N=600) to examine the limitation of the model. The criterion for belongingness to the norm group was the agreement between mental age of an individual and the mean mental age of the norm group. That means:

(a) mean mental age for each random sample was computed from the raw scores (refer to the report of the research project), and

(b) an individual was selected at random and the model discussed in this article was used to compute the F-ratio. When F-ratio was significant, mental age of the individual in question was computed. If the mental age of the individual and the mean mental age of the group were in agreement(range ± 0.5 years), prediction of the likelihood of belongingness to the norm group was

considered to be accurate. The following are the inferences:

(1) At least measures on five independent variates and a sample size of 30 were needed for predicting the belongingness of an individual to the "norm group".

(2) N = 50 and above tend to predict the belongingness of individuals with very flat profiles ($_{m}\chi^2_{k} \leq 3.52$, # variates = 7, N = 50)

(3) The model was most suitable for samples when N = 40 and # of variates = 7.

REFERENCES

Cattell, R.B., r_p and Other Coefficients of Pattern Similarity, Psychometrika, 1949, 14, 279-298.

Cronbach, L.J. and Glesser, G.C., Assessing Similarity Between Profiles, Psychological Bulletin, 1953, 50, 456-472.

Hogg, R.V. and Craig, A.T., Introduction to Mathematical Statistics (3rd). New York, Macmillan Company, 1971.

McNemar, Q., Psychological Statistics (4th). New York, John Wiley and Sons, Inc., 1969.

Pearson, K., On the Coefficient of Racial Likeness, Biometrika, 1928, 18, 105-117.

Williams, J.M., An Empirical Study of Cattell's Coefficient of Profile Similarity, Unpublished doctoral dissertation, University of Alabama, 1969.

SPECIAL REPORT

Mukherjee, A.K., Ander, S., Sluyter, G.V., and Leal, A., Measurement of Intellectual Potential in Mexican-American School Age Children. Corpus Christi State School, Texas Department of Mental Health and Mental Retardation, June, 1976.

ELIMINATING RATER BIASES IN PERFORMANCE APPRAISAL

Purnell H. Benson
Rutgers University

Sources of Error in Evaluating Performance with Rating Scales

The use of rating scales to measure the performance of individuals raises questions of sources of errors and methods for detecting and controlling the errors. In using a rating scale where a numbered step or numerical point is selected to judge performance, various types of error appear.

(1) The constant bias of the individual rater who habitually overrates or underrates all ratees is a bias which shifts the origin or zero point of the scale used. If the correct mean for a set of individuals rated by rater $\underline{k}$ is $\bar{X}_k'$, and the mean of the ratings reported by rater $\underline{k}$ is $\bar{X}_k$, the two values are related by $\bar{X}_k = \bar{X}_k' + Z_k$, where Z_k is the bias or shift in the zero point of the scale resulting from $\underline{k}$'s constant error in judgment.

(2) The habitual contraction or expansion in the dispersion of ratings by rater $\underline{k}$ is a distortion introduced by those who are either reluctant to give extreme ratings or who go to extremes in choosing ratings. If the standard deviation of ratings by $\underline{k}$ is SD_k, and the correct standard deviation for those ratings is SD_k', the two are related by a stretch correction factor F_k such that $SD_k' = F_k SD_k$.

(3) Also involved is the interpersonal error which rater $\underline{k}$ makes with regard to ratee $\underline{i}$. This is an error unique to ratee $\underline{i}$ and rater $\underline{k}$, designated as P_{ik}.

(4) Remaining is a residual random error which depends upon the precision of judgment of which rater $\underline{k}$ is capable, E_k. In practice, P_{ik} may not be separable from E_k, and the two may be considered together as a residual error R_k characteristic of rater $\underline{k}$.

Combining the three types of error into a single measurement equation, the correct rating of individual $\underline{i}$ is related to the rating given of individual $\underline{i}$ by rater $\underline{k}$ by

$$X_i' = (X_{ik} - \bar{X}_k) \cdot F_k + \bar{X}_k - Z_k - R_k, \qquad (1)$$

where $\bar{X}_k$ is the mean of the ratings made by rater $\underline{k}$ of $\underline{k}$'s ratees, F_k is the stretch correction for rater $\underline{k}$, Z_k is the average bias in the ratings by $\underline{k}$, and R_k is the residual error for rater $\underline{k}$.

Computation to Obtain Scale Values for Ratee Performance Which Eliminate Zero-Point Biases of Raters

With an array of ratings X_{ik} of ratees $\underline{i}$ by raters $\underline{k}$, we seek to use the information contained in this matrix to learn the correct ratings X_i' of the performance of the individual ratees. This involves removing the zero-point and stretch errors. We first consider eliminating the zero-point biases of the raters. Reduction of the residual error to a minimum random error of judgment will be considered later.

The matrix of ratees rated by raters yields intervals between the performance ratings of all pairs of ratees from the sets of ratees with common raters. Each rating interval is correct in the sense that the constant zero-point bias has been subtracted in defining the rating interval. If these intervals are correct except for a random residual error, they can be averaged by the arithmetic of paired comparisons to obtain more accurate estimates of the rating intervals between all pairs of ratees from the entire group of ratees.

We proceed as follows. For each pair of ratees $\underline{i}$ and $\underline{j}$ for whom rating intervals are given by one or more raters $\underline{k}$, we average the $\underline{q}$ intervals to obtain for the pair of ratees $\underline{i}$ and $\underline{j}$:

$$Y_{ij} = \sum_{k=1,q} (X_{ik} - X_{jk})/q. \qquad (2)$$

The average interval Y_{ij} is posted in the cell for the $\underline{i}$th column and the $\underline{j}$th row of a paired matrix, and again with the sign reversed in the cell for the $\underline{j}$th column and the $\underline{i}$th row. If the data yield pair differences for all possible pairs of ratees in the matrix, the numerical average of the pair differences down the $\underline{i}$th column gives the average interval between ratee $\underline{i}$ and all of the ratees included by the matrix.

If, as is often the situation, the matrix is incompletely paired, the regression procedure reported by F. Mosteller (1951) is used to find scale values whose differences provide the best fit in the least squares sense to the pair differences which are included in the incomplete matrix. The input for the regression calculation consists of 1 and -1 in the $\underline{i}$th and $\underline{j}$th columns of the row with Y_{ij} as the entry for the dependent variable. Entries elsewhere are 0's, except for the last row which contains 1's to establish the origin for the system of scale values. The entry for the dependent variable in this added row is 0.

The sums of cross-products and squares are calculated about an origin of 0, rather than the mean. This reflects the circumstance that only the entries from one side of the diagonal of the paired matrix need be used in the calculation. The matrix of squares and cross-products whose solution yields the scale values for performance contains entries as follows, using W_{ij} as the weight for the number of raters who define the interval between ratees $\underline{i}$ and $\underline{j}$.

The diagonal cells of the $\underline{i}$th row and column contain $\sum_{j=1,n} W_{ij} + 1$. The off-diagonal cells for the $\underline{i}$th row and $\underline{j}$th column have $1 - W_{ij}$. The $\underline{i}$th row of the n+1 column for dependent variable is $\sum_{j=1,n} W_{ij} Y_{ij}$. The sum of squares for the dependent variable is $\sum_{i=1,n-1} \sum_{j=1,n} W_{ij} Y_{ij}^2$.

The scale values S_i found from solving the equations in this matrix are performance ratings about a mean of zero. While they define rating intervals between ratees, they are not performance ratings in an absolute sense. The norm for the group of ratees must be known, so that the scale values can be transformed to this norm.

The norm may be defined according to some external behavior criterion, or it may be fixed by expert judgment, or it may be taken as the simple average of all of the ratings of ratees by raters. The proper performance norm is the one which is meaningful to those using the rating scale for which a norm is needed.

Calculation to Eliminate the Stretch Bias of Raters

We now consider removal of the stretch bias evident in the contraction or expansion of ratings by each rater. The stretch differences of individual raters can be made uniform by imposing the same rating dispersion upon all raters. Of course it is necessary to consider that each rater may rate a somewhat different set of ratees. First, the standard deviation of performance ratings calculated for all ratees is fixed. Then, the spread of ratings by each rater is altered to agree with the spread of ratings calculated for that rater's ratees. This provides an adjusted set of ratings by each rater for a next iteration of computation. Iteration continues until no further adjustment in the spread of any rater's ratings takes place.

The rating X_{ik}' for ratee $\underline{i}$ and rater $\underline{k}$ corrected for the stretch in the rater's scale is related to the unadjusted rating X_{ik} by

$$X_{ik}' = F_k(X_{ik} - \overline{X}_k) + \overline{X}_k, \qquad (3)$$

where $\overline{X}_k$ is the mean of the original ratings made by rater $\underline{k}$ of rater $\underline{k}$'s ratees, and F_k is a stretch correction factor defined by

$$F_k = SD_k'/SD_k, \qquad (4)$$

with SD_k the standard deviation of rater $\underline{k}$'s ratings, and SD_k' is the standard deviation of $\underline{k}$'s ratees obtained from the ratings calculated for these ratees.

Like the mean imposed as the correct norm upon the system of performance ratings, the standard deviation designated can be defined by an external behavior criterion, or by expert judgment, or simply taken as the standard deviation of all ratings by all raters.

If the standard deviation of the performance scale values X_i' calculated from the inputted ratings is SD', and these scale values are about a mean of zero, then the scale values X_i'' transformed to a designated mean $\overline{X}_o$ and standard deviation SD_o are given by

$$X_i'' = \overline{X}_o + (SD_o/SD')X_i' \qquad (5)$$

Control of Interpersonal and Random Errors in Rating

No simple computational procedure is available to eliminate the interpersonal error peculiar to a particular rater and ratee. This type of error arises from favoritism and misjudgment of the unique achievements of the ratee. As for the random error remaining, this is an error unrelated to systematic analysis.

Both of these types of residual error depend upon the ability and motivation of the rater to control them. Instruction of raters in the criteria for making ratings is important in reducing residual errors, as well as zero-point and stretch errors. Improvement in precision of judgment requires measurement of rating accuracy to grant recognition to those who are efficient raters. If those who play favorites or who fail to take the rating effort seriously are detected by having their rating efficiency measured, this affords means for improving rating efficiency or avoiding those whose rating activity is of poor quality.

Several components of rating efficiency can be isolated and measured by comparing ratings made by raters with the ratings calculated from input by competent raters. We will call these calculated ratings "adjusted group ratings" or AGR. Various comparisons of original ratings with the calculated ratings yield scores.

(1) The zero-point score of rater $\underline{k}$, referred to as score ${}_1T_k$, can be defined as follows for $\underline{m}$ items of performance rated:

$$ {}_1T_k = \left[{}_1\bar{Z}' - \left[\sum_{h=1,m} \frac{(\bar{X}_{h.k} - \bar{X}_{h.k}')^2}{m} \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_1SD_o}{{}_1SD} \right] + {}_1\bar{X}_o, \qquad (6) $$

where ${}_1\bar{X}_o$ is the mean imposed upon performance scores, ${}_1SD_o$ is the standard deviation imposed, ${}_1\bar{Z}'$ is the mean zero-point bias of raters. (Rater bias is a standard deviation of item biases of each rater). ${}_1SD$ is the standard deviation of the zero-point biases of raters (calculated as standard deviations of item biases), $\bar{X}_{h.k}$ is the mean rating by rater $\underline{k}$ of ratees for performance item $\underline{h}$, and $\bar{X}_{h.k}'$ is the mean AGR calculated for rater $\underline{k}$ on item $\underline{h}$.

(2) The score ${}_2T_k$ for stretch bias of rater $\underline{k}$ as a standard deviation of item differences from the AGR spread for $\underline{k}$'s ratees is

$$ {}_2T_k = \left[{}_2\bar{Z}' - \left[\sum_{h=1,m} \frac{(SD_{h.k} - SD_{h.k}')^2}{m} \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_2SD_o}{{}_1SD} \right] + {}_2\bar{T}_o, \qquad (7) $$

with the same identification of variables as before, except for the prescript 2 reference to stretch bias.

(3) The score ${}_3T_k$ for residual error of rater $\underline{k}$ after adjusting $\underline{k}$'s ratings for zero-point and stretch biases depends first upon the calculation for each item of the residual standard error. This is calculated with $\underline{n}$ - 2 degrees of freedom (or $\underline{n}$ - 1 if $\underline{k}$ has only 2 ratees, not permitting a valid adjustment for stretch). Then the root mean square of the residual standard errors is obtained with $\underline{m}$ degrees of freedom for $\underline{m}$ items.

$$ {}_3T_k = \left[{}_3\bar{R}' - \left[\sum_{h=1,m} \frac{(R_{h.k})^2}{m} \right]^{\frac{1}{2}} \right] \left[\frac{{}_3SD_o}{{}_1SD} \right] + {}_3\bar{T}_o, \qquad (8) $$

where $R_{h.k}$ is the residual error of rater $\underline{k}$ rating item $\underline{h}$.

(4) If provision is made for raters to make self-ratings, the discrepancy between the self-rating and AGR can be made the basis for a score for accuracy in self-rating, ${}_4T_k$. It seems more meaningful to those whose rating is evaluated to make this score reflect the total discrepancy between the self-rating and AGR, rather than the residual error after adjusting the self-rating for zero-point bias and stretch bias.

$$ {}_4T_k = \left[{}_4\bar{Z}' - \left[\sum_{h=1,m} \frac{(X_{h.kk} - X_{h.k}')^2}{m} \right]^{\frac{1}{2}} \right] \cdot \left[\frac{{}_4SD_o}{{}_4SD} \right] + {}_4\bar{T}_o, \qquad (9) $$

where $X_{h.kk}$ is the self-rating by rater $\underline{k}$ on item $\underline{h}$, $X_{h.k}'$ is the AGR for ratee $\underline{k}$, ${}_4\bar{Z}'$ is the mean self-rating error on all items (expressed as a standard deviation), and ${}_4SD_o$ and ${}_4T_o$ are the standard deviation and mean imposed for self-rating scores.

In practice, the conversion of error quantities into standard scores is more simply accomplished if the four separate error quantities are averaged into a single score for rating efficiency. Then the mean and standard deviation are imposed upon the overall rater score. Each separate rater score has subtracted from it the group mean for that type of score and then is divided by the group standard deviation for that type of score. This converts all four error scores to the same standard deviation. Then the average of the four scores for each rater is calculated, and a group standard deviation for the overall rater scores is calculated. With the ratio of this to the imposed standard deviation used as a multiplier of the divergence of the overall score from the group mean, in the same manner as the separate formulas already given, the overall scores are converted to those with the required mean and standard deviation. The formula for the combined score adjusted to the imposed standard deviation and mean for the group is applied to

$$ T_k = {}_1T_k + {}_2T_k + {}_3T_k + {}_4T_k, \qquad (10) $$

if equal weights are assigned to each of the four error components with unit standard deviations, and the final formula for adjustment is

$$ T_k' = (T_k - \bar{T}) \left[\frac{SD_o}{SD} \right] + \bar{T}_o, \qquad (11) $$

with T_k' the final rater score, SD the standard deviation of T_k for the group, and $\bar{T}$ the mean of the T_k before adjustment, and $\bar{T}_o$ the mean rater score imposed for the group.

Comparison of ratings by each rater with those made by the leader for that rater's ratees permit four more error scores to be defined. The overall score can be made a weighted combination of the two sets of four scores, if all are available.

Since these measures of rating efficiency depend upon the difficulty of the task of rating, some adjustment is needed when poor performers are

rated who cannot be rated with the same absolute precision as good performers who are near the top of the rating scale.

In the PEERRATE system, diminution of the measure of rater error is accomplished by one of the following two formulas.

$$R_{i.k}' = R_{i.k}\left[\frac{a_3}{a_0 - a_1 X_i'}\right], \qquad (12)$$

$$R_{i.k}' = R_{i.k}\left[a_4 . 10^{a_2(a_1 X_i')}\right]. \qquad (13)$$

$R_{i.k}'$ is the error after adjusting the residual error $R_{i.k}$ in the rating of ratee $\underline{i}$ by rater $\underline{k}$, and a_0, a_1, a_2, a_3 and a_4 are parameters found suitable for the error adjustment. Such parameters can be selected to maximize the correlation between the score for rater efficiency and some criterion, such as the rating received for performance on items.

The rater's score for rating efficiency and the same rater's performance score as a ratee can be used to calculate a suitable weight in the calculation of the adjusted group rating AGR. Commencing with equal weights for raters, these can be progressively improved through iteration, using fresh weights at each stage of iteration obtained from the rater scores from the previous stage of iteration. In the PEERRATE system, the performance score and rater score are combined by parameters for linear, square and cross-product terms.

Operation of the PEERRATE Computer Program

The PEERRATE rating system described here has been implemented by a computer program prepared by the author of the system. An early version of the program was reported by Benson (1976).

The computer program permits a variety of computations to be made to meet various rating situations, such as use or non-use of leader ratings, use of team or department ratings, and combination of ratings into rating scores by either addition or multiplication of ratings together. The program also calculates a matrix of intercorrelations between the performance and rating scores, item by item or overall scores. These intercorrelations help guide the operator towards the selection of proper parameters for the calculations made. All of the results and intermediate steps of calculation can be outputted on cards, tape or disk, at the option of the user of the program, to facilitate further research.

The PEERRATE program, consisting of a deck of approximately 3,000 cards, is available on application to: Dean Horace J. De Podwin, Graduate School of Business Administration, Rutgers University, Newark, N. J. 07102. The program is free of cost to educational and non-profit users except for cost of transcribing the program on cards or tape.

Tables 1, 2, & 3 contain inputted ratings and calculated scores for item performance and rating efficiency.

References

Benson, P.H. A computerized system of student grading of student assignments. Third Annual New Jersey Conference on the Use of Computers in Higher Education, Rutgers University, New Brunswick, N.J., March 22, 1976, pp. 21-30.

Mosteller, Frederick. Remarks on the method of paired comparisons: the least squares solution assuming equal standard deviations and equal correlations. Psychometrika, 16 (March 1951, pp. 3-9.

Table 1

Rater-Ratee Matrix of Ratings for Items

Rater	Item	1	2	3	4	5	6	7	8	9	10	11	12
1	1	***	80	80	87	***	73	87	67	67	93	73	80
	2	***	67	100	67	***	53	80	47	100	73	53	80
2	1	***	87	93	93	***	80	80	73	73	87	93	93
	2	***	87	87	93	***	80	80	73	80	80	100	93
3	1	***	87	100	80	***	67	93	80	67	87	100	87
	2	***	80	93	93	***	67	93	67	80	87	80	87
4	1	***	73	73	80	***	27	67	53	80	73	67	67
	2	***	67	47	87	***	33	87	67	87	73	67	80
5	1	***	93	***	***	***	***	***	***	67	87	***	***
	2	***	73	***	***	***	***	***	***	93	87	***	***
6	1	***	100	93	100	***	100	100	87	87	93	93	93
	2	***	87	87	93	***	100	100	93	87	100	93	93
7	1	***	80	80	87	***	60	87	73	73	87	93	73
	2	***	67	93	73	***	67	87	73	80	73	87	73
8	1	***	***	87	87	***	87	80	87	***	***	93	87
	2	***	***	87	87	***	93	87	87	***	***	93	87
9	1	***	80	73	87	***	93	87	93	87	73	93	93
	2	***	87	93	93	***	93	93	100	93	87	100	93
10	1	***	87	87	87	***	87	93	73	80	100	87	87
	2	***	87	87	87	***	80	93	80	93	93	93	93
11	1	***	***	53	67	***	***	80	***	***	***	73	93
	2	***	***	93	93	***	***	53	***	***	***	100	80
12	1	***	100	100	93	***	93	93	93	87	93	100	93
	2	***	87	100	93	***	93	93	93	87	93	100	93

Table 2

Ratee Performance Scores

	Rating by Group				Rating by Leader			
Id. No.	Rater Quality	Overall Rating	Item 1 Rating	Item 2 Rating	Rater Quality	Overall Rating	Item 1 Rating	Item 2 Rating
1	***	***	***	***	***	***	***	***
2	91	82	89	76	100	74	80	67
3	82	90	88	92	100	90	80	100
4	92	91	92	90	100	77	87	67
5	***	***	***	***	***	***	***	***
6	70	70	71	68	100	63	73	53
7	85	91	91	92	100	84	87	80
8	83	73	73	73	100	57	67	47
9	78	82	73	92	100	84	67	100
10	87	88	91	86	100	83	93	73
11	88	92	94	91	100	63	73	53
12	96	89	89	90	100	80	80	80

Table 3

Rater Performance Scores

		Comparison With Group Calculations				Comparison With Leader Rating			
Id. No.	Overall Rater Score	Average Deviation	Difference In Range	Residual Error	Self Deviation	Average Deviation	Difference In Range	Residual Error	Self Deviation
1	100	100	100	***	***	***	***	***	***
2	90	99	96	85	90	87	88	83	90
3	93	96	93	92	88	91	92	99	90
4	86	70	74	84	88	90	100	89	90
5	98	96	90	100	***	100	100	99	***
6	75	79	87	73	70	70	76	75	70
7	95	84	97	90	98	100	91	96	100
8	76	93	80	70	78	76	70	70	71
9	80	86	89	70	86	75	74	71	90
10	89	93	90	89	89	81	77	99	90
11	81	70	70	100	75	93	95	76	70
12	82	79	84	79	100	70	72	80	92

AVOIDING DISCLOSURE IN TABULATIONS

Richard Bell, Social Security Administration

The various types of disclosure in tabulations are discussed. Appropriate examples, taken from tables appearing at the end of this paper, are presented.

The classification which follows represents an effort to develop a comprehensive and logical description of different types of disclosure. Suggestions for improvement will be welcomed.

Disclosure will be studied both for tabulations involving count (frequency) data and for those containing quantity (magnitude) data.

1. Exact disclosure
   a. Count data: A marginal total equals one of its detail cells; this detail cell is defined as narrowly as possible from the records upon which the tabulation is based.
   Table 1: All beneficiaries in County B are black.

   b. Magnitude data:
   (i) A quantity corresponds to a cell with only one member.
   Table 2: Total sales for the single establishment in Industry B is $125,000,000.

   (ii) A quantity assumes its maximum or minimum possible value.
   Table 3: If the maximum possible payment under the program is $190, then each person in State B receives precisely $190.

2. Approximate disclosure
   a. Count data: A detail cell is zero; the disclosure is not exact.
   Table 1: No beneficiary in County C is white.
   Table 4: The age of each beneficiary in County B is restricted to the interval (65, 69).

   b. Magnitude data:
   (i) The value of a quantity corresponding to an individual cell member is restricted to an interval (L, U) where the lower and upper limits are determined by such relationships as the following from published data and logical operations:

$$U = M;$$
$$U = m + N(A - m);$$
$$U = T - m(N - 1);$$
$$L = m;$$
$$L = M - N(M - A);$$
$$L = T - M(N - 1),$$

   where

   N is the number of members in the cell,

   A is the published, average value of the quantity for the cell,

   T is the published, total value of the quantity for the cell (note T = AN),

   M and m are the maximum and minimum possible values, respectively, for any member in the cell.

   (If two or more distinct values are available for either U or L, select the largest of the possible lower limits for L and the smallest of the possible upper limits for U, respectively.)

   Table 2: Total sales for each establishment in Industry C of table 2 is between 0 and $125,000,000.

   Table 3: Monthly benefit for each of the four beneficiaries in State A cannot exceed $632.

$$L = m = 0$$
$$U = m + N(A - m) = 4(158).$$

   Table 3: If the maximum possible payment under the program is $192, then each person in State B receives at least $120.

$$U_1 = m + N(A - m) = 0 + 36(190) = 6840$$
$$U_2 = M = 192$$
$$U = \text{minimum}(6840, 192) = 192$$
$$L_1 = m = 0$$
$$L_2 = M - N(M - A) = 192 - 72 = 120$$
$$L = \text{maximum}(0, 120) = 120.$$

   (ii) Information about other characteristics associated with the same cell is used to estimate the value of a quantity corresponding to an individual cell member.
   Table 2: If it is known from another source that all five establishments in Industry C have about the same number of employees, then total sales of $25,000,000 can be estimated for each member of the industry. Similarly, if one of the five has 80% of the employment, then the estimate for this "largest" establishment of $100,000,000 in total sales is reasonable.

3. Probability-based disclosure
   The probability that a given member of a total cell with T members belongs to a particular detail cell with D members is D/T.

   Table 1: Assign a probability of 28/30 to the event that a person known to be a

beneficiary in County C is black.

4. Internal disclosure
A member (or a coalition of members) of a group uses his own (their own), as well as published data, to obtain confidential information about others in the group.

a. Count data:
(i) Exact disclosure: The difference between the values of a marginal total and one of its detail cells is equal to the number of members of a coalition not belonging to the detail cell; the detail cell is as narrowly defined as possible.

(ii) Approximate disclosure: The difference between the values of a marginal total and the sum of a proper subset of detail cells is equal to the number of members of a coalition not belonging to the proper subset (equivalently, all members of a detail cell also belong to a coalition); but the disclosure is not exact.

(iii) Probabilistic disclosure: Define the following

S = the published number of members in the total cell,

D = the published number of members in the detail cell,

C = the coalition size,

X = the number of coalition members also belonging to the detail cell.

The probability is $\frac{D-X}{S-C}$ that another member of the marginal total, but outside the coalition, lies in the detail cell.

Table 5: The black worker in County A knows that all the other workers in his county are white. A black worker in County B deduces that the probability is 65/66 that another worker, unknown to him, is white; the coalition of two black workers in County B knows that all other workers in County B are white. The black worker in County C knows he is the only black worker in his county.

b. Magnitude data:
(i) Exact disclosure: After a coalition of size C adjusts a published figure by means of its own data, the revised value of the characteristic involves either type of disclosure described in 1b. (Equivalently, a quantity is published for a cell of size C + R where one of the following conditions hold:
(1) R = 1
(2) The revised value of the published figure, obtained by adjusting for the contribution of the coalition, is a maximum or a minimum.)

(ii) Approximate disclosure: After a coalition adjusts a published figure by means of its own data, the value of a quantity corresponding to an individual cell member is restricted to an interval as described in 2b.

Table 2: If one of the establishments in Industry C has total sales = S, then total sales for each of its competitors must be less than \$125,000,000 - S.

Table 6: If the maximum possible benefit for each of the beneficiaries is \$140, then a person receiving \$40 in County A can deduce that each of the other two members of his cell must receive between \$120 and \$140. Although it would be impossible for a user, not belonging to the cell for County B, to restrict the payment amount to either person in that county to any interval smaller than (0, 140), either beneficiary can readily compute the payment to the other person by use of the published cell.

5. Indirect disclosure
Any of the above types of disclosure is derived by algebraic manipulation and/or logical operations.

a. Count data: Neither table 7 nor table 8 provides sensitive information directly. However, by combining information from both tables, it is seen that all men over 75 with medical coverage have hospital coverage; all women with medical coverage but without hospital coverage are under 65.

Table 9: By subtraction, it follows that there are no workers of race other than white or black in Industry A and that all workers in Industry C are white.

b. Magnitude data: Suppose Industry A consists of two disjoint sub-industries A1 and A2 and that the following information is available from various tables:

Industry	Number of companies	Total sales
A	5	\$200,000,000
A1	4	\$150,000,000

By subtraction, the one company belonging to Industry A2 has total sales of \$50,000,000.

Table 1: Number of beneficiaries by county and race

County	Race			
	Total	White	Black	Other
A	40	15	20	5
B	30	0	30	0
C	30	0	28	2

Table 2: Total sales, by industry

Industry	Number of establishments	Total sales
A	18	$450,000,000
B	1	125,000,000
C	5	125,000,000

Table 3: Average monthly benefits, by State

State	Number of beneficiaries	Average monthly benefit
A	4	$158
B	36	190

Table 4: Number of beneficiaries, by county and age

County	Age class				
	Total	Under 65	65-69	70-74	75 and over
A	37	3	15	11	8
B	4	0	4	0	0

Table 5: Race of workers by county

County	Total	White	Black	Other
A	94	93	1	0
B	67	65	2	0
C	103	101	1	1

Table 6: Number of beneficiaries and average payment amount

County	Number of beneficiaries	Average payment amount
A	3	$100
B	2	70

Table 7: Number of persons with hospital and medical coverage, by age and sex

Age	Hospital and medical coverage		
	Total	Male	Female
Total.....	9,593	4,633	4,960
Under 65.....	3,534	1,714	1,820
65-74.....	3,147	1,517	1,630
Over 75.....	2,912	1,402	1,510

Table 8: Number of persons with medical coverage, by age and sex

Age	Medical coverage		
	Total	Male	Female
Total.....	9,609	4,640	4,969
Under 65.....	3,548	1,719	1,829
65-74.....	3,149	1,519	1,630
Over 75.....	2,912	1,402	1,510

Table 9: Race of workers by industry

Industry	Total	Male	Female
Total.....	400	328	62
A.....	60	30	30
B.....	236	194	32
C.....	104	--	--

AVOIDING DISCLOSURE IN THE RELEASE OF MICRODATA

Paul Zeisset, U.S. Bureau of the Census

For the purpose of this discussion, we will use the term microdata to refer to files in which each record provides data about an individual person, household, establishment, or other unit. Microdata thus include an agency's own confidential files of questionnaires or basic records from a survey or other data collection. Normally we think of these data as being summarized or aggregated to produce statistics for the reports and publications discussed in the previous paper. Nonetheless, release of information in microdata form to a data user outside the originating agency can serve legitimate and important public purposes--in that the data may be useful for many more tabulations or other analyses than the originating agency is prepared to provide. Further, certain statistical applications (for example, similation models) require the user to have input in microdata form.

Release of records about individuals raises the issue of disclosure. Some files are by law not confidential, for example, from the Census of Governments, where detailed data are released identified to the specific governmental unit. On the other hand, most statistical data bases are covered by statutes which prohibit the release of data from which information may be gained about particular individual entities, be they persons, households, establishments, corporations, or other reporting units. In the latter situation, microdata are releasable only if the information is not specific enough to allow identification of the individual. Invariably names and addresses, social security numbers, and other positive identifiers must be removed. Further, certain other information, such as residential location, is generally abbreviated or withheld.

Federal Agency Examples of Microdata Release

For those of you not familiar with what types of microdata files are being released by Federal agencies, let me give you a few examples.

Probably the best known of all Federal microdata bases are the public use samples of basic records from the 1960 and 1970 censuses of population and housing. From the first release in 1963, these samples have provided nearly the full richness of detail about households derivable from the decennial censuses: age, education, income, occupation, etc., of each family member along with characteristics of the family's housing. The sample originally released in 1963 had little geographic information and the sampling fraction was only 0.1 percent of all American households. But, based on the public acceptance and demonstrated utility of that microdata product, public use samples from the 1970 census were created with a larger sampling fraction (one-percent) and more specific geographic information (that is, areas as small as 250,000 population were identified).

The Census Bureau also releases survey data files on a similar basis, with certain added qualifications regarding the smallest areas that can be identified. Microdata are available from the Current Population Survey, the Annual Housing Survey, and the National Travel Survey, to name just a few. Other agencies frequently contract with the Census Bureau to conduct surveys for them, and these surveys also result in microdata files released by either Census or the sponsoring agency: for example, the National Crime Survey sponsored by LEAA, the Consumer Expenditures Survey sponsored by BLS, and the Survey of Income and Education by HEW. In general, all of these files become available for unrestricted public use after identifiers, detailed geography, and some subject information are removed.

Several agencies also release microdata based on administrative records. The Social Security Administration makes several files available from its Continuous Work History Sample derived from payroll tax records and from records of each applicant for a social security number. The Longitudinal Employee-Employer Data (LEED) file is a one-percent sample of employees covered by Social Security. For every individual in the file there is age, race, sex and a record for each place of employment since 1957, indicating the industry, State, county, taxable wages, and estimated total wages for each year. In view of the disclosure potential of the county and industry identification, purchasers must enter into a written agreement with SSA specifying the purpose for which the file may be used, prohibiting further dissemination without SSA authorization, and specifically precluding any attempt to identify specific individuals or establishments or to match individual records with information in other files on specific individuals.

The National Center for Health Statistics also releases a number of microdata files. In this context the most interesting of these is the file on natality which provides a 50-percent sample of records in its birth registration system. No other federal microdata file allows so large a sampling fraction. Records include the age, race, and education of the father and mother, the State and county of residence of the mother, the birth date, legitimacy (if recorded) and several characteristics of the mother's previous childbearing history. Purchasers of a NCHS microdata file must sign a statement that the microdata file will be used solely for statistical research purposes.

Factors Bearing on the Likelihood of Disclosure

While we are confining our consideration to microdata files with no positive identifiers, it should be recognized that a combination of data elements, such as geographic location, age, race, and occupation, if sufficiently detailed, could be used to identify an individual if the investigator knew those characteristics of his subject in advance. Other information on a microdata

record so identified would then be disclosed about the individual, for instance, his income, marital history, educational attainment, and so forth.

Let me discuss three factors bearing on the likelihood that such disclosure might occur: (1) sample size, (2) geographic and subject detail--or the degree to which records in the file are unique, and (3) recognizability of the sample record.

(1) Sample size or fraction of the universe

If an investigator were searching for a particular individual in a microdata file his probability of success can be no greater than the chances that the individual's record is present in the file. In a one-percent sample the chances are 99 to 1 against a particular individual having a record in the file, assuming one has no external way of knowing that the individual was included in the sample. A larger sample size would create greater disclosure potential; a smaller sampling fraction would yield less.

(2) Uniqueness

I use the term uniqueness to refer to whether an individual can be distinguished from all other members in a population in terms of information available on the microdata record. That uniqueness is determined by the size of the population and the degree to which it is segmented by geographic information, and the number and detail of characteristics provided for the sample unit.

The smaller the population, the more easily an individual can be unique; the larger the population, the more likely that his or her set of characteristics is duplicated by somebody else's. Size of the population, or of the smallest segment that can be readily identified, can be varied quite directly by varying the amount of geographic information supplied on a microdata file.

It can also be said that the greater the number and detail of characteristics reported about an individual the more likely it is that the individual's representation on the file would be different from that of any other individual in the population. Just 10 characteristics with four categories each create over a million possibilities (4^{10}), and when one considers that some data items may have 100 or more potential categories (e.g., age, occupation, industry, income, place of birth) the number of possibilities becomes astronomical in a file with a large number of characteristics. Many characteristics are, however, likely to be correlated with one another, thus reducing the degree to which an additional item creates additional unique records.

Assuming that we need to control the degree of differentiation available, it might then seem reasonable to designate a minimum category population, for instance, to collapse country-of-birth categories with less than 50 cases in the file. The technique appears inadequate, however, since for instance, while there may be many Russian-born persons sampled, only one may be black, or only one may live in a particular identified area. More important, uniqueness in the sample is not the critical factor, for there may be a hundred such individuals in the population with no possibility of discriminating among them. Uniqueness in the population is the real question, and this can not be determined without a census or administrative file exhausting the population or at least an identifiable subset thereof (such as a file of all doctors). Precluding uniqueness in the sample would be a very conservative approach to avoiding disclosure.

Some public-use microdata files provide characteristics for all or at least multiple members of a household. The association of the characteristics of household members greatly increases the potential for unique combinations (for example, a 66-year-old judge married to a 23-year-old actress would be a rather unusual combination.)

(3) Recognizability

Suppose we determine that a given record is unique. The next question is whether that record can be linked to a specific person, without which disclosure does not occur. I will refer to this property as a record's recognizability, and I'll discuss three factors affecting it: (a) the existence of a population register, (b) inaccuracy or "noise" in the microdata file, and (c) time lag or the degree to which the microdata information becomes out-of-date for an individual.

(a) Population Registers

Suppose there were a list of everyone in the population, including each person's age, place of birth, and a few other items which were also on a public-use microdata file. Such a list, or population register, could make it not too difficult to find the identity of any one with a unique set of those characteristics.

In some countries, Sweden to name one, such registers are publicly available. In this country the best lists would be in the hands of the Internal Revenue and the Social Security Administration, but these are not available to the public. But neither nationwide coverage nor coverage of all segments of the population is required. Reasonable coverage of a defined subpopulation, along with a number of reliable matching characteristics may suffice. A register of some groups like black architects, American Indians, high public officials, or birth records is not improbable. The existence of rather extensive registers of business establishments, in the hands of government agencies, trade associations or firms like Dun and Bradstreet, has virtually ruled out the possibility of releasing microdata files about businesses for statistical purposes.

One needn't associate the idea of a population register with the dossiers of an investigative agency. If Who's Who in America or the Congressional Directory were in computerized form they could be quite useable for the restricted populations they cover. Welfare agencies and credit bureaus might have information useable for matching in computerized form although access to these files is assumed to be restricted. Those lists which are public--city directories, voter registration lists, or the records of motor vehicle agencies, tax assessors or real estate agencies--probably don't contain a broad enough set of characteristics for matching, at least with the microdata files we have examined. There should be no doubt, however, that any new file considered for availability in microdata form should be reviewed for its correspondence to various existing population registers.

(b) "Noise" in the Data

Another factor which affects recognizability is inaccuracy or "noise" (random error) in the microdata. Usually we think of noise in data as undesirable--respondent mistakes, intentional misrepresentation, coding or processing errors--but that noise also reduces disclosure potential in that unreliability in the microrecord degrades its matchability to a referent in the population. The effect is more severe to attempted identification through matching than it is to the more appropriate statistical uses because there is no chance for compensating errors to average out or to appear small in perspective.

If unintended error or unreliability helps reduce disclosure potential, then intentional noise added to a microdata file could be still more effective, particularly in touching all records rather than just some. Doing so without damaging the usefulness of the file for statistical purposes is the problem.

(c) Time Lag

Time lag is a third factor affecting recognizability. There is inevitably some lag between the date of data collection or reference date and the date the microdata become available, usually at least several months and sometimes several years. As the data become less current they become less useful for many statistical purposes, but they may also become less potentially dangerous to confidentiality.

First, the user will have greater difficulty in reconstructing a given individual's characteristics as of the reference date. Secondly, whatever possible gain the user might expect from the match will presumably be less. Welfare agencies and credit bureaus might have the best files for matching purposes, but the fact that the linked microdata may be one or more years out of date should reduce the utility of the match substantially. A microdata file could be withheld from public use for a number of months or years to reduce its disclosure potential, or "old" files could be released with less stringent protection than contemporary files.

Hypothesized Relationships Among the Various Factors

Now, in examining the relative impact on disclosure potential of the various factors we have discussed, it is useful to hypothesize how an investigator might go about identifying microdata records. There appear to be two different broad types of potential disclosure situations, and they are affected by the various factors in differing degrees. The first scenario is where the investigator searches the file for a specific individual, using certain characteristics of which he is already aware. The second is where the investigator is just "fishing" for a set of characteristics he recognizes.

The first type is quite volatile. If a public-use microdata file were to be useful for investigatory purposes, the breach of confidentiality would be extremely serious. The most obvious factor working against misuse of this type is the sample size. Even considering the largest of the existing public-use microdata files, the six 1970 census one-percent public use samples, and under hypothetically perfect matching conditions, the investigator would have a 94-percent probability of failure with regard to a particular individual.

Only where there is an extremely large number of subjects for whom excellent matching data are available, and under conditions where success in only a few cases will suffice, could the file seem to be of any use. The existence of some sort of population register would be almost a necessity for investigatory use. It is also true that any substantial noise or inaccuracy in the data would preclude an exact match rather effectively.

By contrast, in the second type of disclosure situation the investigator is not searching for a particular individual, but is just "fishing" for a set of characteristics he or she recognizes. Such an occurrence does not immediately seem to be very serious, since there is presumably no profitable purpose to be served by such an investigation. Such an effort might, however, be undertaken in an attempt to discredit the issuing agency or the practice of releasing microdata.

Since one is not starting with a specific set of target individuals, the low probability that any one individual is in the sample is not a problem to the investigator. The investigator selects certain unusual and highly noticeable characteristics, then extracts corresponding records from the sample. The task then is to recognize well-known households or individuals among the extracted records. A population register would be useful but not mandatory here. In the absence of a population register, geographic information on a file is very important since it may be the most specific matching characteristic known to the investigator. Number of characteristics reported is important since the matching will depend on some sort of pattern recognition. Minor aberrations introduced into the data may not inhibit the match if they do not disturb the general pattern, quite unlike the situation with a population register where a minor discrepancy might defeat the match. Compared to searching for a specific individual, the technical requirements for a fishing expedition are relatively modest.

Techniques for Avoiding Disclosure

(1) General Tradeoffs

From the foregoing it should be apparent that a number of factors impact on disclosure potential, and also that no one of them alone can be so restricted as to prevent disclosure by itself. A file which exhausts a universe, or comes close, presents considerable disclosure potential if it contains any unique records. Geographic information must be restricted beyond the point where an individual user could be familiar with a significant proportion of the universe, but whether that point comes at 25,000, 250,000 or 1 million will depend on the detail in the file and other restrictions imposed. The Census Bureau has imposed a 250,000 minimum population criterion across the board, but that is in the context that the Bureau normally provides data files with highly detailed subject matter (for instance, single years of age, detailed occupation). No formula has been worked out adequately representing the tradeoffs between level of geographic identification, detail of individual subject items, and sample size.

(2) Elimination of Categories Identifying Small Salient Groups

Another technique is to avoid categories so detailed that they define a small and easily identifiable group. Providing income groupings so that persons with very high incomes cannot be separately identified is a common technique and may be seen as a more generalized approach to insuring that corporate executives and other highly recognizable individuals not be so easily identified from the rest of the population. A common upper limit for detailed income categories is $50,000 per year, although inflation may soon make a somewhat higher cutoff appropriate.

(3) Allowing No Unique Cases

It has also been proposed (Fellegi, 1972) that microdata files can be made disclosure free by making sure that there are no unique records in the file, which is to say that every set of characteristics is replicated at least once. There is little doubt that this standard would prevent disclosure since any match attempt would never result in only a single qualifying individual. This is, however, an unrealistic standard for a file with many data items, since the number of possible combinations would be astronomically high when in fact relatively few of those data items would be involved in any conceivable match attempt.

That procedure does have some relevance when a particular population register is recognized as threatening the confidentiality of a microdata file, for example, a drivers license file with date of birth, state of birth, sex, and marital status. If a four-dimensional cross tabulation of the microdata within the area to be identified had any cells with only one case, categories could be collapsed or areas redefined until that no longer occurred. If more than one population register existed then the resulting microdata could be subjected to additional cross tabulation. This solution should be recognized as being conservative since it is uniqueness in the population, rather than in the microdata file, which assures matchability. Thus, if possible, the multi-dimensional search for the unique case should be performed on the population register file rather than in the statistical microdata.

(4) Noise Introduction

The introduction of noise into microdata is a fourth alternative. In its simplest form it might involve adding or subtracting small amounts at random to values of continuous or interval variables. There are multiplicative as well as additive models, and a few ideas have also come out of the recent literature on randomized response. Clayton and Poole (1976) did some interesting research on the impact of a couple of error introduction techniques on certain univariate applications. But as yet there is little knowledge of the degree to which error introduction would degrade the more common multivariate analyses. If noise were introduced into data on age, for example, the user's concern is not just that age distributions can be faithfully reproduced, but that the noise does not distort sensitive relationships, such as between age and educational progress where one is attempting to

study the cohorts of students ahead of or behind "normal" progress defined by specific age-grade relationships.

(5) Removal of Well Known Individuals from the File

Finally, if disclosure potential lies primarily with a few people with unusual characteristics it is at least worth considering removing them from the file, rather than eliminating some of the information about all of the population. If more than a handful of such individuals is involved there must be concern about bias resulting from their removal. Of course, the originating agency could prepare summary statistics about the individuals removed. But such a procedure should not be relied on to the exclusion of other techniques since the existence of a large population register would make many people recognizable in a detailed file.

Disclosure Prevention Through Restrictions on Use

In the foregoing I have tried to identify ways in which a file may be made acceptable for unrestricted use. Invariable each bit of information removed from a file to make it disclosure-free reduces that file's usefulness for some research purpose. In fact, we at the Census Bureau are continually met with requests to relax our geographic restrictions on microdata to make this or that worthwhile research possible.

Life certainly would be simpler if we could just trust the data user not to misuse the file. Or, if not naive trust, surely strict contractual arrangements could bind the user of a restricted file to observe procedures which would maintain the confidentiality of the individual data.

Our subcommittee carefully considered what conditions could provide adequate protection, in terms of legal authority needed by the user, penalties for misuse, and a set of conditions agreed to by the receiving organization. The Social Security Administration is now releasing certain files on such a restricted basis--not files with individual identification, but files with too much disclosure potential for unrestricted dissemination.

Certain other agencies are not so ready to embrace the idea of restricted dissemination. The statutes of some agencies don't give them the flexibility SSA has. Furthermore, laws such as the Freedom of Information Act make it not altogether certain that regulations could be upheld if they allow one user access to a file but prohibit access to another.

In 1963 the Census Bureau placed certain restrictions on purchasers of its new 1-in-1000 sample. It wanted to keep careful records on the use of the file--for administrative rather than confidentiality reasons. Unfortunately, those signed agreements were soon forgotten by the purchasers, and the files in question passed freely from one to another. This experience certainly indicated to us that an agency could not successfully restrict use without specific attention to enforcement.

The most important reason, of course, for not relying on restricted-use agreements to enforce confidentiality is that there is a great deal to be gained, by the research community and by society at large, by broad and free access to microdata files such as we have discussed. Restricted release should be considered only where a file's disclosure potential cannot be reduced to an acceptable level while maintaining the usefulness of the file, and then, of course, only where the law allows and the restrictions can be successfully enforced.

Unfortunately, our subcommittee did not come up with a neat formula or simple package of rules to follow to produce microdata of optimum usefulness and confidentiality. Research--of both a theoretical and empirical nature--is needed. Our subcommittee report, then, is of greatest value when used as a study guide by responsible agency officials, simultaneously mindful of the importance of confidentiality and the societal benefits of broad access to public data.

References

Clayton, C. A. and Poole, W. K.

1976 Use of Randomized Response Techniques in Maintaining Confidentiality of Data. Draft Report. Research Triangle Institute.

Fellegi, I. P.

1972 "On the question of statistical confidentiality." Journal of the American Statistical Association, Vol. 67: 7-18.

FEDERAL AGENCY PRACTICES FOR AVOIDING STATISTICAL DISCLOSURE: FINDINGS AND RECOMMENDATIONS

by

Thomas B. Jabine, Social Security Administration
John A. Michael, National Center for Education Statistics
Robert H. Mugge, National Center for Health Statistics

Statisticians are becoming increasingly concerned over the need to avoid statistical disclosure, i.e., the revelation of confidential information about identifiable (but not identified) individual persons or organizations through published statistical tables and microdata tapes (computerized records pertaining to individual statistical units). For example: a published table might indicate that all male retirees in a given community receive the maximum social security benefit, thus disclosing the benefit amount for each retiree; or a published micro-data tape might give the details of health conditions of a female who according to the tape is over 100 years of age and there is only one such individual in the identified community.

This paper reports on an effort to examine statistical disclosure in the extensive and complex statistical programs of the Federal Government. People over the nation are constantly entrusting statistical agencies with various kinds of information about themselves, on the promise that the information will be used only in anonymous form, for purposes of statistical analysis. Federal agencies have a serious obligation to protect these data from statistical as well as any other kind of unauthorized disclosure.

What are Federal statistical agencies doing to prevent statistical disclosure, how well are they succeeding at it, and what more needs to be done on a government-wide basis to minimize the possibility of statistical disclosure? To answer these questions was the charge of the Subcommittee on Disclosure-Avoidance Techniques, established early in 1976 by the Federal Committee on Statistical Methodology, which is sponsored by the Statistical Policy Division of the Office of Management and Budget.1/

The Subcommittee began its work by studying the rules, regulations, and policy statements of Federal agencies relating to statistical-disclosure avoidance. The literature was then searched and relevant articles and reports were located and studied. The Subcommittee received reports on various relevant agency experiences and discussed and analyzed them. A number of actual examples of disclosure were found and considered. (To the best of our knowledge, none of these actually caused any harm, and none have ever been noted outside of the Subcommittee and the agencies which perpetrated them. However, some of them were considered by the Subcommittee to be unacceptable.) Finally, the chapters of the final report were drafted by Subcommittee members, these were intensively reviewed by the Subcommittee and revised, and the Subcommittee reached a reasonable degree of consensus on all points in the final report, which should be ready for the printer before the end of September 1977.

The Subcommittee's report is organized as follows:

The first chapter is an introduction, explaining the charge to the Subcommittee and its auspices and operating procedures.

Chapter II tackles the definition of statistical disclosure. Various previously used definitions are cited and evaluated. A definition proposed by Dalenius is found to be most useful: "If the releases of certain statistics makes it possible to determine a particular value relating to a known individual more accurately than is possible without access to those statistics, then a disclosure has taken place."

This definition is very broad and is not intended to be the basis for agency operating decisions. But neither do the definitions implied in the laws and regulations relating to confidentiality provide such a basis. In fact, absolutist definitions are useless in identifying disclosures which might be both necessary and acceptable for a given statistical program. It must be recognized that the release of some data in potentially identifiable form is justifiable under certain circumstances. Thus, the acceptability of disclosure risk in any given situation must be evaluated.

The Subcommittee found that published tabulations present quite a different set of conditions and problems concerning statistical disclosure as compared with public-use microdata tapes. Therefore, separate presentations are made. Chapter III deals with statistical disclosure in published tabulations. Different kinds of disclosure in statistical tabulations are defined and discussed.

Disclosures may be exact or approximate; they may be probability-based or certain; they may be direct or indirect; they may depend on external or internal data analysis; and they may relate to count data or magnitude data, each having a different set of implications. Depending upon the type of disclosure and its context, dh_ dJsk of actual revelation of confidential data may be great or small, so it is necessary to evaluate these risks before deciding what steps to take. Various disclosure-avoidance techniques which may be used in the case of tabulations are described and evaluated.

Chapter IV discusses potential disclosuri and their avoidance in connnection with the fast-burgeoning Federal agency programs involving the release of public-use microdata tapes. Several factors bear upon the likelihood of a disclosure taking place through a given microdata tape--the sampling fraction used in a survey, the detail of geographical descriptors, degree of detail given on the data subject's characteristics, existence of data for the same individuals in population registers, errors or noise in the data, and the age of the data. Two classes of risk are evaluated: first the risk of disclosure about a particular individual of interest; and second, the risk of disclosure of information on some identifiable individual through a "fishing expedition." Disclosure-avoidance techniques are described and evaluated --eliminiating small-group categories, allowing no unique cases, introducing noise into the data, removing known individuals from the file, and releasing files only for controlled, restricted usage.

For many statistical programs the only sure way to eliminate the risk of disclosure completely would be by refraining from any release of microdata tapes whatsoever, and by reducing published tables to a few broad and bland ones. Yet the release of public-use microdata tapes needed by the research community, together with far more detailed published tabulations, may entail a disclosure risk which, while not absolute zero, is extremely low. Decisions must be made on the proper balance between the community's needs for statistical information relevant to public policy issues and the individual's need for confidentiality protection.

Chapter V is devoted to this crucial question of balance. It reports on the Subcommittee's vain attempts to discover any cases in which an individual has been harmed through statistical disclosure, and it describes ongoing research into the public's attitudes on these questions.

The Subcommittee found that in actual practice, agencies are rarely confronted with problems arising from statistical disclosures, or even from public fears that such disclosures might take place. On the other hand, agencies receive many complaints from data users on the restrictions to data availability resulting from disclosure-avoidance practices.

The final chapter (VI) summarizes the Subcommittee's findings and lays out its recommendations to Federal agencies on the avoidance of statistical disclosure. The draft of Chapter VI is presented below in its entirety:

CHAPTER VI - Findings and Recommendations

A. The Concept of Statistical Disclosure

Findings: Several of the major Federal statistical agencies have developed and applied a variety of disclosure avoidance techniques in connection with the release of statistical tabulations and microdata files (files of individual records with identifiers removed). However, it appears that little attention has been given to defining exactly what constitutes disclosure and how to decide which disclosures are acceptable and which are not.

A few statisticians, notably Fellegi, Hansen and Dalenius have suggested formal definitions of statistical disclosure. This Subcommittee has adopted the definition proposed by Dalenius as a framework for its discussion and review of disclosure-avoidance techniques. The Dalenius definition is broad in scope. It was not the intention of Dalenius, nor is it ours, to recommend or imply that statistical disclosure so defined should never be permitted to occur. If that position were adopted, the present output of statistical information would be drastically reduced. We have adopted this broad definition because we believe it offers the best basis to

1. Identify all potential disclosures in connection with proposed releases.

2. Decide which of these potential disclosures are unacceptable.

3. Use appropriate techniques to prevent unacceptable disclosures.

Tne formal definition of disclosure adopted by the Subcommittee appears in Chapter II, pp.17-25. It can be summarized here by saying that disclosure takes place if the release of tabulation or microdata makes it possible to determine the value of some characteristic of an individual 2/ more accurately than would otherwise have been possible.

B. Deciding What to Release

Findings

1. Federal statutes and regulations governing the release of statistical information in the form of tabulations and microdata do not and cannot provide a clear basis for deciding in each case what must be done to avoid disclosure. Agencies that address this issue are obliged to strike a balance between the requirement to protect the confidentiality of information about individuals and the need for detailed statistical information and records for public policy purposes.

2. The use of microdata files by social scientists and others has developed rapidly since 1960. Microdata file users are becoming increasingly adept at handling these files and are applying sophisticated analytical techniques to exploit them fully. This development has significantly increased the utility of statistical data bases created by Federal agencies from censuses, surveys and administrative records and promises to do so even more.

3. The Privacy Act provision concerning the "disclosure" of certain microdata files (552 a(b)(5)) is ambiguous and has resulted in considerable uncertainty as to the circumstances under which microdata files can be released.

4. The Subcommittee has identified several examples of statistical disclosure which, in our opinion, were not acceptable. Some of those involved potential disclosures of salaries or benefit amounts of specific individuals. We also find, however, that most agencies that release statistical information are becoming increasingly sensitive to the disclosure issue, and that they have adopted or are in the process of adopting policies and procedures designed to avoid unacceptable disclosure (see agency statements in Appendix A).

Recommendations

B 1. All Federal agencies releasing statistical information, whether in tabular or microdata form, should formulate and apply policies and prodcedures designed to avoid unacceptable disclosures. Because there are wide variations in the content and format of information released, the Subcommittee does not feel that it is feasible to develop a uniform set of rules, applicable to all agencies, for distinguishing acceptable from unacceptable disclosures.

In formulating disclosure avoidance policies, agencies should give particular attention to the sensitivity of different data items. Financial data such as salaries and wages, benefits, and assets and data on illegal activities and on activities generally considered to be socially sensitive or undesirable require disclosure-avoidance policies that make the risk of statistical disclosure negligible.

Agencies should avoid framing regulations and policies which define unacceptable statistical disclosure in unnecessarily broad or absolute terms. Agencies should apply a test of reasonableness, i.e., releases should be made in such a way that it is reasonably certain that no information about a specific individual will be disclosed in a manner that can harm that individual.

B 2. Special care should be taken to protect individual data when releases are based on complete (as opposed to sample) files and when data are presented for small areas.

B 3. In formulating disclosure-avoidance policies and procedures, agencies should take into account the various kinds of disclosure discussed in Chapters III and IV of this report. Thus, these policies should deal with situations which can lead to unacceptable disclosures, such as:

a. In tabulations
   (1) Empty data cells.
   (2) Cells equal to marginal totals.
   (3) Cells representing a small number of cases.
   (4) Quantity data cells dominated by one or two units.
   (5) Sets of tables from which the above situations can be arrived at by algebraic manipulation.

b. In microdata files
   (1) Files containing data for all members of a defined population.
   (2) Files with detailed geographic information.
   (3) Files with very precise information, such as exact dates of events, or exact amounts of various kinds of income or assets.
   (4) Files containing substantial amounts of information which is likely to be duplicated in external sources containing identifiers.

B 4. With respect to the release of microdata files the Subcommittee believes that

a. There should be no restrictions or conditions attached to the release of microdata files when it is reasonably certain that no information for specific individuals will be disclosed as a result. The Subcommittee has referred to files released under these conditions as public-use files.

b. Where the test for a public-use microdata file is not met, but it appears that the public interest will be served by releasing microdata files for statistical and research purposes on a restricted basis to specific users, such releases should be permitted when all of the following conditions are met.3/

   (1) The receiving organization has authority and obligation to protect the file against mandatory disclosure equivalent to that of the releasing agency.
   (2) Responsible personnel of the receiving agency are subject to meaningful sanctions for violation of confidentiality provisions.
   (3) The receiving organization agrees to:
      (a) Use the file only for statistical and research purposes.
      (b) Not attempt to identify individual data subjects for any purpose.
      (c) Not release the file to anyone else without authorization from the releasing agency.
      (d) Maintain adequate security to protect the file from inadvertent or unauthorized disclosure.
      (e) Apply agreed-on disclosure-avoidance techniques before releasing tabulations based on the file.
      (f) Destroy or return the file within a specified period of time.

B 5. With respect to the release of tabulations, a distinction between unrestricted (public-use) and restricted releases, similar to that described for microdata files in recommendation B 4, would also be appropriate. Thus, for tabulations for which the risk of statistical disclosure is deemed too great to permit release to the general public, restricted releases might be made under conditions similar to those described in paragraph b of recommendation B 4, substituting "tabulations" for "file" wherever the latter word appears.

B 6. To insure compliance with its disclosure-avoidance policies and procedures, each agency that releases statistical information should establish appropriate internal clearance procedures. There should be a clear assignment of individual responsibilities for compliance. Staff members responsible for compliance should be encouraged to become familiar with the materials summarized in this report, and to take advantage of relevant training activities (see recommendation C 2).

B 7. In order to guide their disclosure-avoidance policies, agencies should systematically document the consequences of these policies. In particular they should investigate and record:

a. The details of any cases in which data subjects or others allege that statistical disclosure has occurred.
b. Requests for tabulations and microdata files without identifiers that have been denied or only partially met because of agency disclosure-avoidance policies.

B 8. The Statistical Policy Division, OMB, should encourage agencies that release tabulations and microdata to develop appropriate

policies and guidelines for avoiding disclosure, and to review these policies periodically. To the extent feasible, SPD should help agencies to obtain technical assistance in the development of disclosure-avoidance techniques. SPD should also be prepared to assist and advise agencies in cases where unacceptable disclosures are alleged to have occurred and in cases where potential users, including other Federal agencies, feel that agency disclosure-avoidance policies are unnecessarily restrictive.

C. Disclosure-Avoidance Techniques

Findings

1. In recent years, many different effective techniques for avoiding disclosure have been developed and used. No one technique is ideal for all types of releases.

2. While these techniques have been applied in several instances in the United States and other countries, they are not generally known or accessible to many agency personnel responsible for the release of statistical information. In this report, we have tried to provide a systematic summary description of useful disclosure-avoidance techniques and references to more detailed information.

Recommendations

C 1. This report should be given wide circulation to Federal agencies that release statistical information, whether based on surveys or on program records.

C 2. Based on the material covered in this report, the Statistical Policy Division, OMB, should conduct periodic training seminars for Federal agency personnel who are responsible for developing and applying statistical disclosure-avoidance procedures. These seminars could be organized in much the same way as OMB's recent seminar on presentation of errors in statistical data. Participants would be expected to train and provide technical assistance to appropriate persons in their agencies.

C 3. Disclosure-avoidance procedures should be described, in a general way, in connection with publications or other releases of data to which the procedures have been applied. However, such descriptions should not include details whose publication would tend to reduce the degree of protection provided by the particular procedures used.

C 4. To minimize disclosure risks, agencies that release data based on samples should, where feasible, refrain from publishing information that would make it easier for others to determine which individuals were included in the sample. For example, if a sample is based on ending digits of social security numbers, the particular pattern of ending digits used to select the sample should not be published.

D. Effects of Disclosure on Data Subjects and Users

Findings

1. While we have found some examples of what we consider to be unacceptable statistical disclosures, we have not been able, in spite of a fairly systematic effort, to locate a single instance in which an individual (natural person) alleged that he or she was harmed or might be harmed in any way by statistical disclosure resulting from data released by Federal agencies. The same statement cannot be made for legal persons (corporations, partnerships, etc.) as data subjects. Several companies included in the Federal Trade Commission's Line of Business Surveys have sought legal relief from mandatory response, asserting that publication of tabulations as planned by FTC would result in damaging disclosures of individual company data.

2. There have been a number of cases in which users of data for both natural and legal persons have been unable to obtain the amount of detail desired from tabulations or microdata files because of agency disclosure-avoidance policies. Many such restrictions occur because of limitations on the minimum size (population) of geographic area which may be separately identified. In the case of microdata files, these restrictions, in addition to limiting the availability of data as such, sometimes make it impossible for the user to calculate sampling errors for the statistics of interest when such information is not provided by the releasing agency.

Recommendations

D 1. With respect to agency policies for releases, in statistical form, of information about individuals (natural persons), consideration should be given to the present apparent imbalance where there have been no instances of harm to individuals but several cases where requests for data have been denied. It is recommended that agencies review their policies to determine whether there are ways to respond more fully to user needs without violating statutory requirements or risking harm to individual data subjects. Some agencies may wish to try new data release procedures, such as controlled remote access to restricted microdata files, on a trial or experimental basis, with careful monitoring.

D 2. With respect to data for legal persons (corporations, etc.), both data subjects and data users have expressed some dissatisfaction with current agency disclosure-avoidance policies. The Subcommittee believes that continuing review of these policies is warranted, but it does not have any specific recommendations for change at this time.

E. Needs for Research and Development

Findings

1. Insufficient theoretical or empirical research has been carried out to determine the vulnerability of different classes of data to disclosure or the effects of disclosure-avoidance techniques on the utility of statistical data.

2. The Privacy Protection Study Commission 4/ has recommended, "That the National Academy of Sciences, in conjuction with the relerelevant Federal agencies and scientific and professsional organizations, be asked to develop and promote the use of statistical and procedural techniques to protect the anonymity of an individual who is the subject of any information or record collected or maintained for a research or statistical purpose."

Recommendation

E 1. The Subcommittee would welcome a program of relevant research and development in the area of disclosure-avoidance techniques. Some particular areas that deserve attention are:

a. How disclosure risks in tabulations and microdata are related to varying sampling fractions.

b. How disclosure risks are related to the number of variables in the data base and to their individual and joint distributions.

c. Software systems for providing controlled online access to microdata files.

1/ Membership of the Subcommittee included the three authors together with Richard A. Bell of the Social Security Administration; Tore E. Dalenius, consultant to the Statistical Policy Division; William J. Smith, Jr., of the Internal Revenue Service; Mervyn R. Stuckey of the Statistical Reporting Service, USDA, and Paul T. Zeisset of the Bureau of the Census. Maria Elena Gonzalez of the Statistical Policy Division worked with the Committee in her capacity as chairperson of the Federal Committee on Statistical Methodology. Michael chaired the Subcommittee. Jabine gave oversight to the project on behalf of the Federal Committee on Statistical Methodology.

2/ Except where otherwise specified, the word "individual" as used in this chapter is meant to cover all types of reporting units--natural persons, corporations, partnerships, fiduciaries, etc.

3/ The Subcommittee recognizes that some agencies cannot make this kind of restricted release under existing law.

4/ Privcy Protection Study Commission, Personal Privacy in an Information Society, Washington D.C.: U.S. Government Printing Office, 1977, p. 587.

SUPPRESSION METHODOLOGY IN STATISTICAL DISCLOSURE ANALYSIS

Lawrence H. Cox, U.S. Bureau of the Census

INTRODUCTION

There are various classes of problems within the realm of statistical disclosure analysis and to each is associated a set of disclosure avoidance techniques. This paper is concerned with one specific disclosure avoidance technique, cell suppression, and the disclosure problems to which this technique applies. This limitation of scope does not, however, extend to the techniques we describe for analysis of the network defining the tabulation cells, as these techniques admit application in a variety of settings in and out of statistical disclosure analysis. In particular, they may be employed to define a bottom-to-top tabulation system for the network.

The suppression problem is discussed and solved here deterministically and completely within the context of the publication network, according to techniques and analyses developed by the author. This deterministic analysis is prerequisite to any associated stochastic or extra-network analysis, in particular because it provides the proper context for such analyses. The emphasis of this paper will be to highlight the relevant methodological problems posed in the application of suppression techniques in disclosure avoidance. Dut to limitations of space, it will not deal with the relevant issues and problems in the design and development of an automated system to effect these analyses and the practical experience gained from the development of a disclosure analysis system for the 1977 Economic Censuses currently underway at the U.S. Bureau of the Census.

The reader may refer to [2] for a discussion of other techniques of disclosure avoidance and to [1] for further explication of the terminology.

THE SUPPRESSION PROBLEM

To protect the confidentiality of the identity or response of each respondent to a set of statistical publications, a test of sensitivity must be applied to each tabulated cell for each statistic to be published. This is accomplished according to an operant definition of sensitive cell for this statistic. In general, a cell is sensitive for a particular statistic if the value of the cell for this statistic could be employed to yield an unacceptably close upper estimate of the contribution of any one respondent to the total cell value.. An unacceptable estimate of this response would by definition breach the confidentiality of the respondent by effectively publishing its response or providing information which could lead directly to a determination of the respondent's identity. For example, when the data are categorical (qualitative), so that each respondent contributes 1 to the cell value if the respondent is a member of the cell and 0 otherwise a threshold rule defines a cell to be sensitive if the cell contains n or fewer respondents, for n a fixed (small) positive integer. In applications involving aggregate (quantitive) data, such as the U.S. Economic Censuses, a dominance rule defines a cell to be sensitive if n or fewer respondents in the cell contribute greater than k% of the total cell value, for fixed parameters n and k; n is a small positive integer and $0 < k \leq 100$. If respondent data are assumed positive, then threshold rules for quantitative data are dominance rules with k = 100.

According to suppression methodology, the values of sensitive cells are not published (i.e., are "suppressed from publication") for the statistics for which they are sensitive. As linear relationships usually exist between tabulation cells in a publication network, upper and lower estimates of the values of the suppressed sensitive cells may be obtained by linear techniques and, in some instances, precise determination of the value of a sensitive cell may be made. As a result, a consistent definition of what constitutes an acceptable estimate of the value of a suppressed sensitive cell must be made in order that additional, appropriately chosen, linearly related non-sensitive cells, called complementary suppressions, may also be suppressed from publication. These complementary suppressions are made to insure that only acceptable estimates of the values of sensitive cells may be obtained from the network. Equally important, the complementary suppression process must be performed so as to minimize its adverse impact on the information content of the publications.

Each of the above concepts must be made precise to the extent that they may be measured in a predetermined and meaningful sense. These several issues will be dealt with in separate sections of this paper. Interrelationships between them will be discussed at appropriate points.

DEFINING ACCEPTABLE ESTIMATES OF SUPPRESSED SENSITIVE CELLS

Assuming the respondent date are non-negative, if the value of a cell or union of cells containing a particular individual respondent to a cell is known, then this value is an upper bound of the value of this respondent's datum. Similarily, zero is a lower bound on this value. In general, therefore, an interval estimate of the value of each individual response to each cell exists. Sensitivity rules are developed to identify those estimates of individual respondent data which are unacceptable according to established criteria. Acceptable estimates of sensitive cells therefore must be defined so that the estimates of the value of individual respondent data they provide conform to the corresponding estimates obtainable for respondent data from non-sensitive cells. Acceptable estimates must be determinable from the sensitivity rule and, ideally, one should be able to pass from formulae for acceptable upper and lower estimates of sensitive cells to a formula which describes the sensitivity rule.

If cell sensitivity for categorical data is defined by a threshold rule, then it follows that an unacceptable lower estimate of the value of a suppressed sensitive cell should be defined as zero, and an unacceptable upper estimate of its value should be defined to be greater than the parameter n. This results from the fact that a threshold rule is applied to categorical data to prevent any individual from being classified in a group of fewer than $n+1$ respondents.

To determine acceptable estimates of suppressed sensitive cells in a publication network of quantitative data, one must examine the available methods of estimation of cell values from above and below and the corresponding estimates of individual respondent data which can be made for respondents in non-sensitive cells. In general, dominance criteria are employed because, if there is dominance of a cell X by a small number n respondents, then it is possible for one of the dominating respondents to subtract its contribution from the total cell value $V(X)$, thereby obtaining an undesirably close upper estimate of the total value of the responses of the other dominating respondents, and thereby a refined upper estimate of the contribution of each of these other (n-1) dominating respondents. Indeed, it is the value of $D(X)$, the total contribution of the n largest respondents, which must in general be protected. If X is sensitive, $V(X)$ is suppressed only because it represents an unacceptably close upper estimate of $D(X)$. For cells X in which the total contribution $D(X)$ of the n largest contributing respondents lies below the dominance threshold (i.e., $D(X) < (k/100)V(X)$), $V(X)$ is by definition an acceptable upper estimate of the value of the response of any of the n dominating respondents. In particular, this is true when $D(X) = (k/100)V(X)$, in which case publishing $V(X)$ protects $D(X)$ by $((V(X)-D(X))/D(X))\%$ of its value, i.e., by $((100-k)/k)\%$ of the value of $D(X)$. For sensitive cells, therefore, it is reasonable to define an acceptable upper estimate of the value $V(X)$ of a sensitive cell X to be greater than or equal to $(100/k)D(X)$, so that the dominant portion $D(X)$ of the sensitive cell X will receive proportionately at least as much protection from above as does the corresponding $D(Y)$ for a cell Y on the dominance threshold.

Lower estimates of $D(Y)$ or $D(X)$ are obtained in a much more complex manner. As $D(X)$, considered as a cell (although in general it is not a tabulation cell), is the aggregate response of n or fewer respondents, then $D(X)$ and any subcell of $D(X)$ is sensitive and thus suppressed. Therefore, lower estimates of $D(X)$ are obtainable only through lower estimates of the corresponding $V(X)$, in the following manner. If a lower estimate n' of n and an upper estimate t' of the number of respondents t to a non-sensitive and published cell Y are known, then one may conclude $D(Y) \geq (n'/t')V(Y)$. In many publications, t is published or may be straightforwardly inferred from published data regardless of whether $V(Y)$ is published or not. As a result, analysis of the t's corresponding to published and suppressed cells would most certainly lead a serious data analyst to a precise determination of the value of the parameter n. Therefore, under the assumption that n and t are precisely known, the relative equivocation from below afforded $D(Y)$ by publishing $V(Y)$ for a non-sensitive cell Y equals $(D(Y) - (n/t)V(Y))/D(Y)$. For Y on the sensitivity threshold $D(Y)=(k/100)V(Y)$, this relative equivocation from below equals $1 - (n/t)(100/k)$.

Remark. For published cells Y, other lower estimates of $D(Y)$ may be obtained from $V(Y)$. However, under the mild restriction $(k/100) \geq n/t$ (recall that $t \geq n+1$ for non-sensitive Y), the lower estimate $D(Y) \geq (n/t)V(Y)$ was best possible among those considered.

If X is sensitive so that $V(X)$ is suppressed, then lower estimates $L(D)$ of $D(X)$ may be obtained from lower estimates $L(X)$ of $V(X)$ provided a lower estimate k' of k is known. As X is sensitive by assumption, then $D(X) \geq (k/100)V(X) \geq (k'/100)V(X)$. As $L(X)$ is a lower estimate of $V(X)$, then $V(X) \geq L(X)$ and hence $D(X) \geq (k'/100)L(X) = L(D)$.

To provide at least the same relative equivocation from below to $D(X)$ for sensitive X as to $D(Y)$ for Y on the sensitivity threshold, we define an acceptable lower estimate of $V(X)$ for a sensitive cell X to be any lower estimate which is less than or equal to

$$\left\{ \begin{array}{ll} 0 & t \leq n \\ (n/t)(100/k)^2 D(X), & t > n \end{array} \right\}$$

Remark. It would be useful to determine upper and lower sensitivity measures S^+ and S^- for which $S^-(X)$ and $S^+(X)$ measure the amount of additional suppression necessary to protect $D(X)$ from above and below, respectively. Theoretical and practical considerations indicate the desirability of requiring these measures to be subadditive and superadditive, respectively, as the following inequalities demonstrate. If X is sensitive and Y is a candidate cell for complementary suppression, then the union XUY will be non-sensitive if $S^-(XUY) \leq V(XUY)$; and a lower estimate $L(XUY)$ of the union XUY will be acceptable if $L(XUY) \leq S^-(X) + S^-(Y) \leq S^-(XUY)$. We may construct a subadditive function on the set of cells by assigning to each cell Y the minimum acceptable upper estimate of its corresponding $D(Y)$ i.e., by defining $S^+(Y) = (100/k)D(Y)$. However, the corresponding function which assigns to each cell Y the maximum acceptable lower estimate of its corresponding $D(Y)$ as determined above is not a subadditive or superadditive function. In terms of defining a sensitivity measure in the sense of [4], it would be desirable to determine a superadditive minorant $S^-(Y)$ of this function.

THE PUBLICATION NETWORK AND LOGICAL TABLES

By the term tabulation cell we shall mean any cell whose value for a particular statistic is either tabulated for publication or, although not explicitly tabulated, may be determined from the values of tabulated cells by linear techniques; and the term publication network shall

denote the set of all tabulation cells together with the collection of all linear relationships between them. A publication network is definable in terms of one or more independent parameters, such as membership in certain of several geographic sets, industry groups or industry types.

The publication network may be realized as a directed linear graph representing set-subset relationships between classes of tabulation cells. These set-subset relationships and the linear relationships between the tabulation cells mutually define each other. Each point on the directed graph corresponds to a class of tabulation cells and each directed line segment between graph points (nodes, vertices) corresponds to a set of linear equations between the members of the corresponding classes of tabulation cells. For example, the four geographic parameters United States, State, County and City-within-County are related hierarchically, so that the graphical representation of an associated publication network would consist of four points arranged vertically in the order above, with directed line segments from the points corresponding to United States to State, State to County and County to City-within-County. As each graph point has at most one superior on the graph, then this network is one-dimensional. A two-dimensional network would result if these geographically defined cells were further disaggregated by another strictly hierarchical set of parameters. For example, if, as in the U.S. Census of Manufactures, the responding universe comprises all manufacturing establishments, each classified according to geographic location of place of business and industry type (by 6-digit within 4-digit within 3-digit within 2-digit Standard Industry Code), then the publication network would be two-dimensional. The corresponding directed graph would consist of the four points of the strictly geographic graph previously mentioned, together with the sixteen possible combinations of each of four geographic types with the four industry types, with corresponding directed line segments between these 20 points.

As the maximum number of directed segments terminating at any graph point in the preceding example equals two, the publication network is two-dimensional. For example, the graph point corresponding to County by 3-digit industry type has precisely two directed segments terminating at it, one emanating from each of the graph points County by 2-digit industry type and State by 3-digit industry type. Each of these directed segments represents a class of linear equations, namely those equations between a specific county by a specific 2-digit industry type and this county by the 3-digit industry types which make up the given 2-digit industry type, and those equations between the state containing the county by one of those 3-digit industry types and all counties within this state by this particular 3-digit industry type. These two classes of linear equations may be brought together to form a class of two-dimensional statistical tables, each table of which displays the two-way disaggregation of a particular state by a specific 2-digit industry type for a given statistic by means of the counties within the state and the 3-digit industry types which make up the particular 2-digit industry type. This situation admits a straightforward generalization, subject to the following definition. A tabulation cell in a statistical table is an internal cell if it is not a marginal total or partial marginal total (i.e., cannot be disaggregated by subsets) in the table.

General Observation. Given a publication network and its associated directed graph, the tabulation cells and the linear relationships between these which define the publication network may be organized for each statistic into tables so that each tabulation cell appears as an internal cell in precisely one such table. Moreover, the dimension of this table is less than or equal to the number of directed segments terminating at the graph point corresponding to the tabulation cell.

One dimension of each of these tables represents the disaggregation of a tabulation cell corresponding to a superior graph point of the given interior graph point by the tabulation cells it comprises at the inferior graph point. For example, a state is broken down by its counties or a particular 2-digit industry group is broken down by its 3-digit industry groups as in the previously mentioned example. These tables may be constructed inductively from the "top" (the maximal points) of the graph downwards, and shall be referred to as the logical tables of the publication network. This definition is motivated in part to distinguish the logical tables from other tabular displays of the data. The importance of the logical tables become clear when the suppression problem is viewed globally, i.e., in the context of the entire publication network.

An ideal global solution to the suppression problem in a publication network may be described as follows. Associate a variable to each suppressed tabulation cell in the publication network and associate to each unsuppressed tabulation cell its value. These variables and constants are substituted into the linear equations defining the publication network. The publication network is thus realized as a system of linear equations. Through application of linear programming techniques, best-possible upper and lower estimates of the values of suppressed sensitive cells and sensitive unions of suppressed cells are obtained to yield best-possible interval estimates of the values of these cells. (Sensitive unions of suppressed cells are formed under dominance criteria within a linear relationship between sensitive and nonsensitive cells may be derived in which the largest n respondents dominate. Since the linear equation corresponding to this cell union is derivable, then the value of the cell union is effectively published). If the interval estimate thus obtained for any suppressed sensitive cell is unacceptable, then, according to an established suppression methodology, additional cells are suppressed (i.e., additional variables are introduced into the system) until no unacceptable interval estimates of suppressed sensitive cells may be obtained within the network. This suppression methodology must also be

sensitive to predetermined rankings of cells as candidates for complementary suppression, to historical precedent and to relevant policy to the extent that attention to these does not diminish the information content of the publications in disproportionate measure to their importance. Above all, this methodology should minimize over-suppression of cells so that as few cells of the smallest possible value be suppressed complementarily in the network.

Unfortunately, the computational enormity of the process just described renders this process virtually impossible to implement in all but the smallest and simplest (e.g. strictly hierarchical) poublication networks. To render the problem tractable in general (for example, in censuses or large surveys), the problem must be organized into a set of local problems for which valid local techniques can be developed, together with controls for maintaining consistency between these local analyses. The General Observation previously stated provides this organization.

As previously described, the network is organized into collections of logical tables for which each tabulation cell appears as an internal cell in precisely one logical table. Beginning with the logical tables formed at the maximal points on the directed graph and proceeding downwards through the graph (with respect to the partial ordering of the graph points imposed by the directed line segments), the logical tables are subjected to an intra-table disclosure analysis which performs complementary suppressions if necessary in each logical table until each incoming suppressed cell can only be acceptably estimated within the logical table. (The algorithmics of such intra-table techniques will be discussed in the next section). As each logical table completes disclosure processing, best-possible interval estimates of all suppressed cells are computed and acceptable interval estimates of the value of each complementary suppression created within this logical table are defined in terms of the relationship between such estimates and interval estimates of the values of the suppressed sensitive cells in the logical table. The acceptable interval estimates of the complementary suppressions thus defined are passed to any subsequently processed logical table in which the complementary suppression appears as a marginal total. This is done to insure that only acceptable estimates made be made of suppressed sensitive cells within the network. As each tabulation cell appears as an internal cell precisely one logical table, this processing sequence can be completed in one pass (i.e., without "backtracking" to reprocess a particular table) if the operant sensitivity criterion and the defined acceptable estimates resulting make it possible to adequately protect any sensitive cell in a logical table by suppressing only internal cells in the table. In general, to control the disclosure analysis and suppression process theoretically and operationally and to minimize over-suppression, it is advisable to adopt a suppression methodology which suppresses cells on the margins in logical tables only when no combination of suppressed internal cells within the table will suffice to protect the table's sensitive cells.

INTRA-TABLE COMPLEMENTARY SUPPRESSION METHODOLOGY

The problem of intra-table disclosure advoidance and complementary suppression methodology in a publication network is to adequately protect all cells and unions of cells which have been designated as suppressed in a logical table through the process of complementary suppression, while minimizing the adverse impact of this process on the quality of the published data. It is therefore necessary that adequate upper and lower levels of protection for these suppressed cells and that the adequacy of individual unsuppressed cells as complementary suppression candidates can be determined. Our major assumption is that the quality of the published data is adversely affected more by the suppression of a larger number of cells than by the suppression of fewer cells of perhaps larger aggregate value. This assumption is justified in a large publication network by the cascading effect of cell suppression, i.e., suppresisons at higher levels in the network force, in an unpredictable manner, more suppressions at lower levels in the network. Therefore, although in particular cases it may seem that the quality of the data is least affected by the suppression of many small cells in favor of suppressing a few large ones, the fact that each of these complementary suppressions must be protected at lower levels of the network and may force the suppression of large cells at lower levels indicates that suppressing fewer cells is the better strategy in general.

This strategy may be mitigated by preassigning a Prefer (for suppression) or Disallow (from suppression) status to individual suppression candidates prior to the intra-table analysis. These assignments should be respected unless they serve to render the intra-table problem intractable, in which case they must be selectively relaxed or ignored.

The objectives of study in intra-table complementary disclosure analysis are unions and differences of suppressed cells for which the value of the cell union or difference is effectively published (i.e., can be obtained from the values of published cells by linear techniques). As each complementary suppression is performed in the table in turn, the set of unions and differences of suppressed cells is changed. When this set is such that the value of none of its members may be derived as an unacceptable upper or lower estimate of the value of a sensitive or other suppressed cell, the intra-table analysis and complementary suppression process is complete for this logical table. A suppression methodology must be developed for which this sequence of complementary suppression terminates in a minimum or near-minimum number of complementary suppressions. This problem is significantly more difficult in three and higher dimensional logical tables than it is in one or two dimensions. Although operational programs based upon heuristic algorithms are being developed to

complementary suppression in three and higher dimensional tables, the subsequent discussion will be limited to the two dimensional case (of which the one dimensional case is a particular application). This limitation does not, however, apply to the techniques of linear estimation employed, which easily generalize to higher dimensional problems.

Although upper estimates of suppressed cells in a two-dimensional logical table can be obtained from the linear equations corresponding to the row and column containing the suppressed cell (i.e., the cell is estimated from above by the difference between the row or column marginal total, if it is published, and the sum of all published cells on the row or column), it is the set of all linear combinations of these line estimates which comprise all linear estimates of the value of the suppressed cells obtainable from the logical table. By means of these linear combinations, better upper estimates and nontrivial (i.e., positive) lower estimates of the values of suppressed internal cells in a logical table may be obtained. Techniques for obtaining such estimates are described in [1]. The problem of obtaining best-possible upper and lower estimates of cells in a logical table may be posed as a generalized transportation problem as studied in the field of operations research.

In the classical transportation problem, there are q supply points each with fixed supply and p demand points each with fixed demand. There is a transportation cost per unit delivered associated with each supply point-demand point association. Assuming total supply equals total demand, the transportation problem is to assign supply to demand so that the total transportation cost (the cost function) is minimized. The problem is represented by a $(p \times 1)^x$ $(q \times 1)$ array. The i-th row of this array corresponds to the i-th demand point, $1 \leq i \leq p$, the j-th column of the array corresponds to the j-th supply point, $1 \leq j \leq q$, the entry in position (i.j) is a variable x_{ij} representing the amount supplied by the j-th supply point to the i-th demand point, while the entries (i, q + 1) and (p + 1, j) are, respectively, the total demand at the i-th demand point and the total supply at the j-th supply point, $1 \leq i \leq p$ and $1 \leq j \leq q$. The entry (p + 1, q + 1) equals the common value of total supply and total demand. The reader is referred to [3] for a discussion of various classes of transportation problems and their solutions. In the disclosure application, each published cell in the logical table is replaced by its value. Unlike the classical transportation problem, some of the row and column marginal totals may be variables. The costs associated with each variable in the cost equation are taken from the discrete set {- 1, 0, 1}, so that, for example, if we seek to determine the minimum value (i.e., the best lower estimate) of the cell in the (1, 1) position, we minimize the cost function x_{11}. If we seek the maximum value of this cell (i.e., its best upper estimate), we find the minimum value of the cost function $-x_{11}$, and take its negative. Optimal estimates of cell unions and differences may be obtained by minimizing or maximizing the analagous linear relationships between their corresponding variables. Standard transportation problem techniques may be employed to determine these minima and maxima. The significant computational difference between this application and the classical transportation problem is that several iterations of the techniques may be necessary in the disclosure application before a feasible solution to the problem is reached (see [3]).

In [1], the author describes techniques for determining interval estimates of the values of suppressed cells in a logical table using an algorithm tailored to the disclosure problem. This algorithm begins with a line estimate of a particular cell or cell union and systematically generates cell unions and differences related to this cell, comparing the upper and lower estimates of the cell value thus obtained with previously obtained estimates. The algorithm operates quite efficiently and has never failed to obtain best-possible estimates. It remains to prove that this algorithm always generates best-possible interval estimates of the values of suppressed cells in a logical table (e.g., that this algorithm is equivalent to existing transportation algorithms). This is under investigation.

Although methods for determining best-possible interval estimates have been established, an area of research which remains open is that of determining a minimal set of complementary suppressions given a set of specified suppressions and their acceptable upper and lower estimates. An exhaustive examination of the alternative complementary suppression patterns is out of the question in all but the smallest logical tables; and no acceptable branch and bound procedure has yet been devised, although these remain under investigation. A geometric approach to the problem is indicated to provide guidance and control in the choice of complementary suppressions. Geometrically, we may view the disclosure problem as represented by a 0 - 1 matrix in which the position corresponding to a suppressed cell or a cell disallowed as a complementary suppression candidate contains a 0 and those corresponding to candidates for suppression contain a 1. For the moment ignoring the cell values and assuming that any one candidate for complementary suppression in a row or column will suffice to protect that row or column (i.e., the union of this cell with all suppressed cells on the row or column is non-sensitive), then a partial geometric solution of the suppression problem is provided by the following theorem.

Theorem. Let R and C denote the number of rows and columns, respectively, in a logical table which require additional suppression (the unprotected rows and columns). Assume that one additional suppression in an unprotected row or column will suffice to protect this row or column. If R=C=1, then at most three additional suppressions are necessary in the logical table to protect all rows and columns. Otherwise, Max (R,C) additional suppressions suffice. Assume for definiteness that R=Max (R,C). Then the

first C of these complementary suppressions must be chosen so that one is in each of the C unprotected columns and each is in a different row. The remaining R-C complementary suppressions are chosen with one in each of the remaining unprotected rows and each may be chosen in any column, provided that, if one is chosen in a column not containing any suppressions, then at least one other is chosen in the same column.

It results that the number of such solutions grows like the factorial of Max (R,C), so that many alternative suppression patterns exist. This theorem, when applied in conjunction with specified Prefer and Disallow suppression options and branch and bound techniques has proven effective in determining optimal or near-optimal suppression patterns which protect cells in their rows and columns in real disclosure settings (i.e., where one complementary suppression on a row or column may not suffice to protect the row or column, and where n respondent dominance in cell unions is a factor). If, after the Theorem has been applied, improved estimates of suppressed cells are obtained through linear combinations of row and column equations (i.e., from cell unions or differences which are formed through linear combinations of rows and columns), the suppression pattern generated by application of the Theorem is appropriately augmented. A generalization of the Theorem which identifies all single variable linear equations obtainable from a given suppression pattern and the corresponding set of covering suppressions is under investigation.

THE SYSTEM AS IMPLEMENTED

An automated system to perform disclosure analysis and complementary suppression for the 1977 Economic Censuses of Manufactures, Construction Industries and Wholesale and Retail Trade is currently completing development at the U.S. Bureau of the Census. This system is written in Fortran and its initial implementation will be on Univac 1100-series computers. The system applies the methodology described in this paper, with the following important exception.

There are four parameters employed to define the statistical cells in these publications, of which as many as three may be cross-tabulated to define a particular tabulation cell. These parameters are Geography, Standard Industry Code, Sales Type and Type of Establishment. The latter three of these are strictly hierarchical (i.e., one-dimensional), but the geographic parameter, owing to overlapping geographic regions, is two-dimensional. As almost all statistics are cross-tabulated by Geography, this implies that almost all logical tables will be at least three-dimensional. As no three or higher dimensional analog of the Theorem cited in the preceding section exists, it was decided to develop a methodologically sound two-dimensional complementary suppression computer program and to process only two-dimensional logical tables. This procedure is feasible in three-dimensional publication networks for which Geography is a defining parameter because data for overlapping portions of geographic regions are not published. Therefore, the corresponding cells may be employed as available suppressions, so that problems in the third dimension may be made to occur infrequently, and the constituent two-dimensional tables may be processed separately. When problems in the third dimension do occur, the processing order is backtracked in a well-defined manner.

The only four-dimensional tables constructed are those of Geography by SIC by Sales Type. Owing to the backtrack technique previously described, these four-dimensional tables can be regarded as sets of three-dimensional tables of one geographic dimension by SIC by Sales Type. To process these three-dimensional tables, each three-dimensional table will be partitioned into a collection of two-dimensional tables, one for each Sales Type. These will be processed separately by the two-dimensional suppression program. At various stages in this analysis, the collection of two-dimensional tables comprised by the original three-dimensional table will undergo a three-dimensional disclosure analysis reconciliation.

BIBLIOGRAPHY

[1] Cox, L., Statistical Disclosure in Publication Hierarchies, 1976 Proceedings of the Statistical Computing Section - American Statistical Association, pp. 130-136.

[2] Interim Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Subcommittee on Disclosure-Avoidance Techniques, Federal Committee on Statistical Methodology, Statistical Policy Division, Office of Management and Budget, 1977. (unpublished)

[3] Dantzig, G., Linear Programming and Extensions, Princeton University Press, Princeton, 1963.

[4] Sande, G., Towards Automated Disclosure Analysis for Enterprise Based Statistics, Statistics Canada, 1977. (unpublished)

ALTERNATIVE TYPES OF RECORD MATCHING: COSTS AND BENEFITS

Daniel B. Radner, Social Security Administration
Hans J. Muller, Bureau of the Census

I. Introduction

This paper reports on work being done by a Subcommittee on Matching Techniques associated with the Federal Committee on Statistical Methodology.1/ Because the topic of record matching2/ is so broad, we can only give an overview. At a later date the Subcommittee will issue a final report which will expand upon the discussion presented here.

The matching of data files is a very useful technique for many purposes. In this paper, we are interested only in matching for research and statistical purposes. Matching for other purposes, e.g., administrative, will not be considered. In the matching considered here, identification of individuals, if needed at all, is only necessary to make the match. After matching, that identification can be removed.

When we are considering only the accuracy of the matched data, the preferred method of matching is ordinarily what is commonly called "exact matching,"3/ i.e., combining data for the same individuals from different data sources, usually by means of personal identifiers (e.g., name, address, Social Security Number).4/ The use of the term "exact" match is not meant to suggest that such matches are made without error; problems associated with exact matching are mentioned later.

In many cases, for technical or legal reasons, or both, exact matches cannot be carried out. For example, both files might be samples which have few persons in common; or, the information might not be sufficient to identify the individuals in both files. Legal restrictions on exact matching, which have existed for some time, have been increasing in recent years (e.g., the Privacy Act of 1974 and the Tax Reform Act of 1976). These limitations on the use of exact matching have led to interest in alternative methods of matching.

This paper focuses on one such alternative approach, what is commonly called "statistical matching."5/ In a statistical match, the information brought together from the different files (ordinarily) is not for the same person, but is for similar persons. The match is made on the basis of similar characteristics,6/ rather than personal identifying information, as in the usual exact match.

The distinction between exact and statistical matches is not always clear-cut. In this paper, matches in which the aim is to link data for the same person from two files are defined to be exact matches. As defined here, exact matches can be carried out using similar characteristics, but ordinarily personal identifiers are used. Matches in which the aim (for all or most records) is to link data of similar persons, rather than the same person, are defined to be statistical matches. In general, statistical matches have been carried out in situations in which an exact match was not possible.

II. Overview of Matching Applications

The Subcommittee has collected many examples of matching of data files, most by government agencies and most from the U.S. This overview is based upon the examples we have collected, only a few of which can be mentioned in this paper. We will separate the applications of matching, somewhat arbitrarily, into two broad types: (1) adding more variables or additional reports on the same variables; and (2) comparing the presence of units in two files. Within type (1), several different kinds of applications can be identified. One application is the addition of more variables to make possible analyses which otherwise could not be done or to enrich analyses with more variables. Both exact and statistical matching have been used in this application. A cross-section example of one such exact match is the addition of Social Security Administration (SSA) age, race, and sex data to federal individual income tax returns in order to provide better income and tax data by those characteristics. In another cross-section example, a statistical match was carried out between observations from a household survey and a sample of federal individual income tax returns in order to add more detailed and more accurate income information to the household survey data [8]. A longitudinal example of exact matching is the linkage of hospital admission and separation records into cumulative health histories [27].

Another kind of application within type (1) is the evaluation of data, in which initial variables are compared with added variables, or with additional reports on the same variables--from other existing sources or from special evaluation surveys. Evaluation of the accuracy of data was carried out using the 1973 Current Population Survey--Internal Revenue Service--SSA Exact Match Study. In that work, the income data from the different data sources were compared and response and reporting errors were analyzed (e.g. [3]). Definitional differences were examined in Sweden using exact matching. Two different definitions of unemployment--from a household survey and from the labor market board--were compared by matching survey responses and labor market board records [10].

In type (2), two different kinds of applications can be identified: evaluation of coverage and construction of more comprehensive lists. The Bureau of the Census has conducted numerous coverage evaluation studies in connection with the Decennial Censuses. For example, in connection with the 1960 Population Census, samples from 1950 Census records, registered births, and other sources were matched with 1960 Census records, and coverage was assessed [19]. In such matches, the emphasis is upon the presence of units in the files, rather than upon the relationships between data in the two files. In an example of list construction, the Statistical Reporting Service of the U.S. Department of Agriculture used exact matching in the construction of a master list sampling frame of farms in each state. This master list was constructed from several different lists, and exact matching was used to detect duplication between

(and within) the different lists [9]. Statistical matching has not been used in type (2) applications, and is not appropriate for such applications.

In most of the applications mentioned above, one possible effect of matching was a reduction of response "burden"--i.e., to get the same information without matching, a considerable amount of direct data collection would have been necessary. Also, in some of those applications, cost reduction was a beneficial effect--i.e., matching was less expensive than direct collection of the same combination of data would have been. The Office of Management and Budget recently has suggested the use of statistical matching to reduce response burden and cost by means of what are called "nested surveys." In such surveys, different samples from the same population are asked different sets of questions, with a core of questions in common. The data from these different samples can then be matched statistically to obtain relationships between the items not in the common core of questions [17].

III. Exact Matching

For exact matching it is necessary that all or most of the individuals in one file ("base file") be included in the other file ("reference file"). However, rarely do both source files include enough identifiers to allow unique identification of all individuals; the identifiers that are used are usually missing from some records and reported inaccurately or with variations in some other records; each file may--correctly or incorrectly--include some persons absent from the other file. As a consequence, an apparently matched pair of records with the same or very similar identifiers usually links the records of the same person in both files ("true match"), but it may link the records of two different persons ("false match" or "mismatch"). On the other hand, if a record in one file appears to have no match in the other file, that may be because there really is no record for that unit in the second file ("true nonmatch"), or there really may be records for the same person in both files but one or both records may include errors or spelling variations that prevent them from being recognized as a match ("false nonmatch").

In many cases the true match status could only be ascertained at great expense or not at all; generally, a matched file must be assumed to contain some errors. The relative importance of false matches and false nonmatches varies in accordance with the purpose of each project. Techniques have been developed for designing the matching process for a particular study in such a way that the type of error most harmful in the context of that study can be minimized and the remaining error can be estimated.

An exact matching procedure generally includes the following steps (although they may not always be clearly distinguishable).7/8/

1. Data preparation: Transfer to machine-readable form, resequencing, reformatting, elimination of out-of-scope records, and other editing steps. If one or both of the files do not already exist, this step includes data collection.

2. Selection of matching variables and tolerances: Ideally, the most accurately reported and the most discriminating variables are preferred, but these are often conflicting requirements. Confidentiality restrictions may interfere by making identifiers such as names or Social Security Numbers unavailable. Because of the inaccuracies in the source files, strict agreement on such variables as age or on name spelling cannot always be required. More or less elaborate techniques have been used for selecting, for a particular matching project, the combination of matching variables and tolerances that will keep the probability of matching errors as low as feasible [14, 33].

3. File blocking: In order to avoid having to compare each base file record with all reference file records, relatively small portions of both files are selected for intensive searching, (e.g., all records with addresses in the same city block, or all records with a certain group of last names, including variant spellings of the same name). Ideally, these "comparison classes" or "blocks" should be formed on the basis of characteristics that will virtually never disagree in the case of true matches, and will almost always disagree in the case of true nonmatches [32, 33].

4. Weights and thresholds: Since a block ("comparison class") may include several possible reference file matches with the same base record ("comparison pairs"), some rules are needed for deciding which pair--if any--is accepted as a match. Each pair contains a particular configuration of agreements and disagreements on the matching variables; explicitly or implicitly, the decision is based on the probability of that configuration occurring if the pair were truly matched, or truly not matched (paired at random).

The rules for making that decision need to take into account the fact that different variables contribute different amounts of relevant information. This is done by assigning different weights to various degrees of agreement or disagreement on each variable, and deriving a total weight for each comparison pair. For carrying this out in practice, a great variety of procedures have been used, ranging from the intuitive judgment of a researcher to mathematical models of the matching process that require a computer for their application. The weights can be based on external evidence or derived from special pilot studies or from thorough investigation of samples, or their derivation can be incorporated in the computer program that uses them.

Finally, once it has been determined how likely or unlikely it is that a particular comparison pair constitutes a true match and which of several possible pairs is the most likely match, it must be decided whether it is likely enough to be accepted as a match, taking into consideration the purpose of the project.

This final decision, explicitly or implicitly, takes the form of setting a threshold that divides the range of total weight scores into "matched" and "not matched". This is not an isolated decision; it is affected by the previous decisions on matching variables, tolerances, and weights. All of these decisions must be coordinated with the aim of achieving the

results that are optimal in terms of the purpose of the particular matching project [9, 11, 19, 27, 29].

5. Except for very small studies, it is practically impossible to clear up all doubts and avoid all matching errors. In well planned matching studies, the probable impact of such errors may be estimated and, if necessary, appropriate adjustments may be made in the results [16, 23, 25].

IV. Statistical Matching

To the best of our knowledge, the vast majority of the statistical matches and of the developmental work carried out has been in the field of economics.9/ The most common application has been to combine data from a household survey with data from income tax returns where there was little overlap between the two files. Early statistical matches were performed at the Bureau of Economic Analysis of the U.S. Department of Commerce in connection with estimates of the size distribution of family personal income [5, 6, 7] and the Brookings Institution in connection with analysis of the tax system [18]. More recent matching work has been done at Statistics Canada [1], Yale University [22], the Office of Tax Analysis of the U.S. Treasury Department [30], Brookings [4], and the Office of Research and Statistics of the Social Security Administration [21]. 10/

Because statistical matching is not a well-known technique, the theoretical steps involved in the most common case will be summarized.11/ We begin with two microdata sets of observations on variables for units in a universe, U; these sets, A and B, are the sets we want to match statistically. A and B are assumed to be probability samples from U. It is also assumed that very few units from U are in both A and B. For example, A might be the persons interviewed in a household sample survey, and B might be an independent sample of income tax returns. Some variables from U may be contained in both A and B, while at least some are contained in only one set.

It is assumed that at least some of the variables in A and B contain errors, while in U they do not. Because of different error components, a variable from U which appears in both A and B can have different values in the two sets for the same unit in U. For example, even if wage income were defined identically in the household survey and the tax return, the survey response might differ from the amount shown on the tax return.

We now define C, a hypothetical data set which represents the results of an exact match (carried out without error) between A and B, if the units in A were also in B. The set C is hypothetical because that exact match cannot be carried out, since very few of the units in A are also in B. By assumption, C contains all variables from A and all variables from B, including their error terms. Because a statistical match is an approximation of an exact match, C is the data set which we try to approximate when we perform a statistical match. In our example, for each unit in A, C contains the survey response given by that A unit and the data from the tax return filed by that A unit. As noted above, that tax return probably does not appear in B.

When we actually want to make a match, we do not know C. Therefore we make an estimate of C, called L, using whatever information is available. Estimated values (for the B information) might be obtained by assumption. For example, for a given A unit, it might be assumed that the value for a given B variable should be equal to the value for a given A variable. We could say that wage income in B should be identical to wage income in A. This would be valid if wage income were defined identically and had identical error patterns (e.g., response and reporting error) in A and B. Ordinarily, this is not the case. Estimated values can also be obtained by other means, for example, by regression techniques or by using cross-tabulations from an exact match between sets similar to A and B. In our example, for each unit in A, L contains that unit's survey response data and estimates of (some or all of) the variables in the tax return filed by that A unit.12/

We now introduce M, the result of statistically matching A and B (in some unspecified way). Using our example, for each unit in A, M contains that unit's survey response data and the tax return data from the B unit assigned to that A unit in the statistical match. It is not necessary that every B unit be used in the match solution; some B units can be used more than once in the solution.13/ It follows from the definition of a statistical match that the variables assigned to a given A unit in the match are all from one B unit.

In making a statistical match we choose among alternative solutions; each alternative solution is characterized by the particular set of B units assigned and the particular A unit(s) to which each is assigned. We choose the solution in which M approximates L as closely as possible, in terms of the variables and relationships of greatest importance in the results of the match. This approximation can be viewed in terms of a distance function which measures the distance of M from L. The distance is defined in a subjective way according to the purpose of the match. The statistical match solution which minimizes this distance is the optimal match result.14/

In practice, many different statistical matching methods have been used. In most cases the variables in both files were separated into "matching variables" (which were similar in the two files and were used to carry out the match) and "nonmatching variables" (which were the "added" variables). In most matches, both files were separated into comparable subsets of units. Within each subset, rules were specified for the choice of a record from the second file to be assigned to each record from the first (or "base") file. The selection of the record within the subset usually was based upon a distance function by which a distance was computed between a given base set record and each potential match in the other set. The distance was based upon differences between matching variables in the two files.15/ In some cases, these differences were weighted according to the relative importance of the variables and the comparability of the pairs of variables for which values were

compared. The potential match with the smallest distance ordinarily was chosen as the match; a maximum distance has been used to define a subset of potential matches from which a random choice was made. In some cases, subsets were defined so narrowly that most subsets contained only one record. In other cases, the choice within subsets was random.

Very little work on the reliability of statistical matching has been done.16/ Given this lack, we will merely attempt to identify several types of errors which can arise in statistical matching, assuming that the matching is done in an optimal way. "Error" is defined as the difference between the "true" joint distribution of A variables and B variables that would be obtained from an exact match (carried out without error) between A and B, if such a match were possible, and the estimated joint distribution of those variables obtained from a statistical match. The following three sources of error can be identified. First, because of lack of comparability between matching variables in the two sets (i.e., the variables are not defined identically and/or have different error patterns), we cannot know with certainty the values of the matching variables that we are searching for in B. Second, even if we knew those values with certainty, often we could not find a B record with such values because B is a sample which ordinarily does not contain the true match. Third, even if we could find a B record with such values (assuming it is not the true match), the values for nonmatching variables in B probably would differ from the true values because those nonmatching variables are not "completely explained" by the matching variables.

V. Summary of Costs and Benefits

In this section the costs and benefits (or advantages and disadvantages) of exact and statistical matching are summarized. Three aspects of this topic will be touched upon: (1) the reliability of the data resulting from the match; (2) the confidentiality problems involved; and (3) the resource cost of the match. Of course, it is very difficult to generalize, since matches vary widely in these aspects. But we feel that some general statements contrasting exact and statistical matching can be made.

Reliability--Error at a single record level will be discussed first; then error on an aggregate level will be mentioned. Initially it will be assumed that the same persons are in the two sets to be matched; therefore an exact match of all units in the base set is possible. Under this assumption we can compare sources of error for an exact match using personal identifying information and an exact match using characteristics (which is a statistical matching type of technique). In this case, error in the data used to match is the main source of error in the match result. In most cases, the personal identifying information has been more reliable than characteristics in making the match; thus we could generalize and say that, in this case, exact matching is more accurate than a statistical matching type of technique. It should be noted that we are considering not only whether the match for any given record is correct, but how far the values are from the true match values if a mismatch is made.

We will now assume that the persons in the two files are all different, and examine additional sources of error in statistical matching. In this case, statistical matching faces what might be called the "proxy" problem; that is, how good a proxy for the true match can be found. Even if we assume that the characteristics used to match on are defined identically and have identical error patterns, the proxy found is not likely to have values which are identical to the true match values. Even if it did have such values for the matching variables, the values for nonmatching variables probably would not be identical to the true match values.

On an aggregate level, it is difficult to identify generally applicable measures of accuracy. Unless the statistical match is constrained to use all non-base file records, the means of variables in the non-base set can be biased (e.g., because amounts are matched too low more often than too high, even though the best match for any base file record is chosen). Or, the variance of the values in the records chosen from the non-base set can be too low (e.g., if records with extreme values are not chosen often enough in the match). In exact matching, biases can arise from false matches and from false nonmatches. In general, the reliability of the results can be estimated in exact matching more easily than in statistical matching. With both methods, it may be necessary to adjust the matched file to a set of independently established control totals.

Confidentiality--The confidentiality problems clearly are much greater for exact matches than for statistical matches for two reasons. First, if personal identifiers are used (as they usually are in exact matching), persons must be identified, at least at some stage of the matching. Second, in an exact match (assuming that the true match is found), the matched file contains more information regarding the person than either of the original files matched. Thus, there is a greater risk of a record in the matched file being identifiable even after the removal of the personal identifiers. Protective measures against disclosure can be taken in both cases, but they usually entail greater expense and/or some loss of information. These problems ordinarily do not exist in the case of statistical matching.

Resource Costs--It is very difficult to generalize regarding cost differences between exact and statistical matches. Costs can vary for many reasons, depending upon, for example, the amount of computer time used, the amount of clerical time used, the lengths of the files, the complexity of the statistical matching program, and the amount of preliminary data analysis and reformatting that need to be carried out. Because it is so difficult to make meaningful comparisons, no generally valid conclusions regarding cost comparisons can be made here; the costs of possible alternative procedures must be evaluated specifically for each project.

In discussing the comparisons in this section, we have assumed situations in which either exact or statistical matching might be

useful. However, there are many situations in which statistical matching would not be useful. In addition to the type (2) applications (comparison of presence of units in two files) mentioned earlier, statistical matching also can be inappropriate for many type (1) applications. For example, if we want to compare the earnings of persons who have had a given training program with those who have not, we can use an exact match between a list of trainees and earnings records. However, a statistical match between those two data sets would not be useful unless the earnings observations could be separated into those who had been trained and those who had not.

VI. Summary and Conclusions

Exact matching is extremely useful in a variety of research and statistical applications. In many of those applications, statistical matching is not an acceptable alternative because the resulting data would not be useful. However, statistical matching has been useful in a few limited applications. When statistical matching can be used, the data obtained from the statistical match in general contain far more error than those from an exact match. Statistical matches can be as expensive as, or more expensive than, exact matches, but statistical matches do have the important advantage that they are carried out without the use of personal identifying information and that they ordinarily do not bring together information for the same person. Thus, statistical matching does not pose the same confidentiality difficulties that exact matching does.

The data which result from matched files should be used with caution because matching, exact or statistical, is not error free. This is particularly true for statistical matching. A substantial literature on exact matching and its nature and reliability exists. However, little has been written regarding the nature and reliability of statistical matching. A great deal of research into the reliability of statistical matching is needed; of particular importance is an examination of the effects of lack of comparability between matching variables. One possible approach which has been suggested is to compare the results of exact and statistical matching of the same files.

FOOTNOTES

1/ The authors are greatly indebted to the members of the Subcommittee, particularly the ex officio members, Maria Gonzalez, Thomas Jabine, and Tore Dalenius, for their many helpful comments.

2/ Other terms have also been used, e.g., "record linkage."

3/ Other terms have also been used, e.g., "actual" and "object" matching.

4/ Although most of the discussion in this paper is in terms of matching information for persons, the discussion also applies to other units for which confidentiality can be an issue (e.g., business firms, hospitals).

5/ Other terms have also been used, e.g., "attribute," "data," "stochastic," and "synthetic" matching.

6/ It is possible to match on characteristics which are not similar; the requirement is that for one or more variables in one set, corresponding values of one or more variables in the other can be identified.

7/ Some of these steps can be executed efficiently by computer. For some applications, a prepared program is available that works with user-specified variables, weights, and tolerances [31].

8/ In addition to the references cited for various steps, [15] includes a very comprehensive treatment of all aspects of exact matching. Brief overviews of exact matching procedures and problems are given in [12, 28].

9/ Related work on matching (or "pairing") samples to reduce extraneous variation has been done outside economics (e.g., [2]). Also, the imputation of values to nonrespondents in household surveys is a closely related technique.

10/ For several comments and replies on statistical matching and an overview article on matching, see the July 1972 and April 1974 issues of the Annals of Economic and Social Measurement. [13] and [34] are somewhat more theoretical papers on statistical matching.

11/ This formulation was suggested in [20].

12/ L can also include constructed variables for both A and B.

13/ Some matching methods do require that every B unit must be used in the match solution, and used only once [20, 30]. In some matching methods, more than one B unit can be assigned to an A unit.

14/ This is not meant to suggest that any given match should be carried out using a distance function, or that a distance function method is the best method in theory.

15/ The matching variables ordinarily were chosen partly because they were (thought to be) significantly correlated with important variables which could not be used to make the match. In exact matches, such a correlation has not been an important factor in the choice of information used to make the match.

16/ See [26] and [34] for examples of work that has been done.

REFERENCES

[1] Alter, Horst E. (1974). "Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970." Annals of Economic and Social Measurement (April) 2: 373-394.

[2] Althauser, Robert P., and Rubin, Donald (1969). "The Computerized Construction of a Matched Sample." American Journal of Sociology (September) 76: 325-46.

[3] Alvey, Wendy and Cobleigh, Cynthia (1975). "Exploration of Differences Between Linked Current Population Survey and Social Security Earnings Data for 1972," 1975 Proceedings of the ASA, Social Statistics Section, 121-28.

[4] Armington, Catherine, and Odle, Marjorie (1975). "Creating the MERGE-70 File: Data Folding and Linking." Research on Microdata Files Based on Field Surveys and Tax Returns, Working Paper I, The Brookings Institution (June). Mimeographed.

[5] Budd, Edward C. (1971). "The Creation of a Microdata File for Estimating the Size Distribution of Income." Review of Income and Wealth (December) 17: 317-33.

[6] Budd, Edward C., and Radner, Daniel B. (1969). "The OBE Size Distribution Series: Methods and Tentative Results for 1964." American Economic Review (May) LIX: 435-49.

[7] Budd, Edward C., and Radner, Daniel B. (1975). "The Bureau of Economic Analysis and Current Population Survey Size Distributions: Some Comparisons for 1964," in James D. Smith, ed., The Personal Distribution of Income and Wealth, Studies in Income and Wealth, 39: 449-558.

[8] Budd, Edward C.; Radner, Daniel B.; and Hinrichs, John C. (1973). "Size Distribution of Family Personal Income: Methodology and Estimates for 1964." Bureau of Economic Analysis Staff Paper No. 21. U.S. Department of Commerce (June).

[9] Coulter, Richard W. (1977). "An Application of a Theory for Record Linkage." Paper presented at the April 6 meeting of the Washington Statistical Society, Washington, D.C.

[10] Dalenius, Tore (1974). "Tva matare av arbetslosheten. En studie 1 svensk arbetsmarknadsstatistik." Report No. 81 of the research project "Errors in Surveys," Department of Statistics, University of Stockholm.

[11] Fellegi, Ivan P., and Sunter, Alan B. (1969). "A Theory for Record Linkage." JASA 64: 1183-1210.

[12] Hansen, Morris H. (1971). "The Role and Feasibility of a National Data Bank, based on Matched Records and Interviews." Report of the President's Commission on Federal Statistics 2: 1-63. Washington.

[13] Kadane, Joseph B. (1975). "Statistical Problems of Merged Data Files," OTA Paper 6, Office of Tax Analysis, U.S. Treasury Department (December 12).

[14] Madigan, Francis C., and Wells, H.B. (1976). "Report on Matching Procedures of a Dual Record System in the Southern Philippines." Demography (August) 13: 381-95.

[15] Marks, Eli S.; Seltzer, William; and Krotki, Karol J. (1974). Population Growth Estimation - A Handbook of Vital Statistics Measurement. The Population Council, New York.

[16] Neter, John; Maynes, E.S.; and Ramanathan, R. (1965). "The Effect of Mismatching on the Measurement of Response Errors." JASA 60: 1005-1027.

[17] Office of Management and Budget (1977). "Standards for Statistical Methodology." Statistical Reporter, No. 77-9 (June), pp. 423-24.

[18] Okner, Benjamin A. (1972). "Constructing a New Data Base from Existing Microdata Sets: the 1966 Merge File." Annals of Economic and Social Measurement (July) 1: 325-42.

[19] Perkins, Walter M., and Jones, Charles D. (1965). "Matching for Census Coverage Checks." 1965 Proceedings of the ASA, Social Statistics Section, 122-41.

[20] Radner, Daniel B. (1974). "The Statistical Matching of Microdata Sets: The Bureau of Economic Analysis 1964 Current Population Survey--Tax Model Match." Ph.D. dissertation, Department of Economics, Yale University. Microfilm.

[21] Radner, Daniel B. (1977). "Federal Income Taxes, Social Security Taxes, and the U.S. Distribution of Income, 1972." Paper prepared for the 15th General Conference of the International Association for Research in Income and Wealth, University of York, England, August 19-25.

[22] Ruggles, Nancy, and Ruggles, Richard (1974). "A Strategy for Merging and Matching Microdata Sets." Annals of Economic and Social Measurement (April) 2: 353-72.

[23] Scheuren, Fritz and Oh, H. Lock (1975). "Fiddling Around with Nonmatches and Mismatches." 1975 Proceedings of the ASA, Social Statistics Section, 627-33.

[24] "Selected Bibliography on the Matching of Person Records from Different Sources." 1974 Proceedings of the ASA, Social Statistics Section, 151-54.

[25] Seltzer, William and Adlakha, Arjun (1969). "On the Effect of Errors in the Application of the Chandrasekar-Deming Technique." (Reprinted as Laboratories for Population Statistics Reprint Series No. 14. Chapel Hill, 1974.)

[26] Sims, Christopher A. (1972). "Comments." Annals of Economic and Social Measurement (July) 1: 343-46.

[27] Smith, Martha E., and Newcombe, H.B. (1975). "Methods for Computer Linkage of Hospital Admission-Separation Records into Cumulative Health Histories." Methods of Information in Medicine (July) 14: 118-25.

[28] Steinberg, Joseph, and Pritzker, Leon (1967). "Some Experiences with and Reflections on Data Linkage in the United States." Bulletin of the I.S.I. 42:786-805.

[29] Tepping, Benjamin J. (1968). "A Model for Optimum Linkage of Records." JASA 63: 1321-32.

[30] Turner, J. Scott, and Gilliam, Gary B. (1975). "Reducing and Merging Microdata Files," OTA Paper 7, Office of Tax Analysis, U.S. Treasury Department (October).

[31] "Unimatch 1 Users Manual--A Record Linkage System" (1974). Bureau of the Census, Census Use Study. Washington, March.

[32] U.S. Dept. of Agriculture, Statistical Reporting Service (1977). "Selection of a Surname Coding Procedure for the SRS Record Linkage System." (B.T. Lynch and W.L. Arends). Paper presented at the April 6 meeting of the Washington Statistical Society.

[33] U.S. Dept. of Commerce, National Bureau of Standards (1977). "Accessing Individual Records from Personal Data Files Using Non-Unique Identifiers." NBS Special Publication 500-2.

[34] Wolff, Edward N. (1974). "The Goodness of Match," National Bureau of Economic Research Working Paper No. 72 (December).

DISCUSSION

I.P. Fellegi, Statistics Canada

This discussion is based on an interim report of the Subcommittee, issued by OMB on May 27, 1977. Wherever possible, I will discuss the papers as a group, they being (I presume) largely the product of the Subcommittee. I want to state at the outset that in my opinion the authors have carried out an extremely thorough job, producing a most comprehensive review of the "state of the art". Notwithstanding some of the critical comments made below, I largely agree with their findings and certainly admire their thoroughness.

1. My first, and most important point relates to the implication of the definition of disclosure accepted by the authors. This definition, originally proposed by Dalenius, is extremely broad. Working with such a broad definition is useful at least from one point of view: it enables them to provide an excellent and comprehensive discussion of every conceivable disclosure--of great educational value! It is also very limiting: in fact, the definition is so broad that in the case of quantitative variables clearly every tabulation cell is a "disclosure"--as defined.

Perhaps the most extreme illustration of the implication of the breadth of definition of disclosure relates to what Bell calls "probability-based disclosure". The example quoted refers to a county in which a table shows that over 80% of the persons are earning income in the range of $2,000+ -- the conclusion being that "it is very likely that a given person in the county has a monthly income in excess of $2,000" and that consequently probability-based disclosure would occur. It seems to me that this somewhat stretches the issue: if the income class $2,000+ were broken out in more detail, no clear majority would fall into any given class. So by showing more detail, the apparent probability-based disclosure can always be remedied -- a result which is not intuitively too appealing.

Starting with their very broad definition of disclosure, clearly the authors found it very difficult to formulate guidelines with respect to disclosure-avoiding approaches. In fact, the operational aspects of the guidelines can be summarized, somewhat simplistically, as follows: there are no federal guidelines, each agency should formulate its own policies, and internal procedures to implement them. These should be reasonable and should always prevent exact disclosure of financial and related information; in formulating their own guidelines agencies should be aware of the educational material developed by OMB. Summarized in this somewhat crude way, it might sound to some like an admission of failure. I would certainly not agree with that assessment. Instead, it is an honest admission of the fact that the legal framework does not provide an operationally useful definition of disclosure, that the logical framework of Dalenius is too broad to be of operational (as opposed to educational!) value, and that, therefore, the operationalization of the concept of disclosure must be based on pragmatic considerations.

One can certainly protest that relying on individual judgements is intellectually not satisfying, but I tend to agree with the authors that it is the only realistic course, given our current state of knowledge of the issues. By analogy, our legal system survived well without the concept of "guilt" ever having been precisely defined. Convictions are based on being found guilty beyond reasonable doubt. The formal legal framework identifies various actions as being "crime" or "tort", but the definitions involved are usually somewhat abstract, and few except the most obvious cases coming to courts represent "perfect fits"--thus the need for the personal judgement of judges and juries. Pursuing the analogy, an important aspect of the legal system is that it rests on precedents--similarly, the guidelines encourage agencies to document, on the one hand, the details of any alleged disclosure and, on the other hand, requests for tabulations or microdata which were refused on grounds of potential disclosure. This is clearly a sound recommendation. In the absence of anything analogous to the Supreme Court, the guidelines propose that the Statistical Policy Division of OMB assist and advise agencies in cases of allegation of either disclosure or unnecessarily restrictive disclosure avoidance policies.

The analogy with the judicial/legal system breaks down in two fundamental ways: judgements can usually be appealed and reversed. However, disclosure, once it occurs, cannot be reversed--published data cannot effectively be withdrawn, nor the resulting damage to the statistical system easily repaired. For this reason, I tend to disagree with the implied criterion for balancing the "right of privacy versus the need to know". Indeed, the paper of Michael et al argues that there has been no documented case of a person having been harmed as a result of statistical disclosure and that, by contrast, this does not appear to be the case with respect to companies. Based on this observation, the paper states that, with respect to population data, there appears to be an "imbalance where there have been no instances of harm to data subjects but several cases where requests for data have been denied"; and that in the business sector, "there is a better balance between the interests of data subjects and users". Thus, it would appear that the state of equilibrium recommended by the paper would occur where the dissemination program, through gradual liberalization, begins to result in documented harm being caused to persons. Of course, it may well be true that some agencies are too conservative with respect to their dissemination program--I would simply argue (quite strenuously) against the implied criterion of equilibrium.

2. My next comment concerns the treatment in the Bell paper of the issue of sensitivity of data and the assurances given to respondents. I would be wary of classifying variables into "sensitive" and "non-sensitive" classes, presumably with the intention of being more liberal with

respect to the disclosure of non-sensitive variables. There are few variables, at least relating to people, which can safely be assumed to be non-sensitive. Even such basic demographic variables as age and relationship to head can be extremely sensitive: they can have a significant impact on, for example, social welfare eligibility. Moreover, when we promise confidentiality to respondents, we do not restrict our promise to some unspecified "sensitive" variables. We can hardly have a dissemination policy which is in conflict with our promise to the public at the time of collection.

3. My next point relates to the treatment in the papers of disclosure within the complex of federal government departments. One of the guidelines in the paper by Michael et al deals with the release of micro data files which do not meet the criteria of public-use microdata files. The same proposals surface also in the paper by Zeisset. The guideline, in effect, states that such files can be released if the receiving agency has the authority and obligation to protect the microdata files, with appropriate sanctions for violation of confidentiality provisions. Not being totally familiar with the legal framework under which U.S. federal statistical agencies work, I can only express a visitor's opinion that without an umbrella Statistics Act, which would establish "statistical enclaves" (to use Mr. Duncan's terminology) within the different departments, all subject to the same confidentiality protection statutes, this guideline might not be particularly useable. In the absence of such a Statistics Act, it is important to regard potential disclosure within the federal establishment as being just as serious as disclosure to non-governmental bodies or persons. At least at one place in the paper of Zeisset I could detect a distinction being made in favour of federal departments. The paper argues that in order to recognize unidentified persons on a microdata file, an extensive population register is required. It goes on to state that "in this country the best lists would be in the hands of the Internal Revenue Service and the Social Security Administration, but these are not available to the public". I find this argument quite unconvincing: the administrative (as opposed to the statistical research) arms of SSA and IRS might be precisely the agencies which the public might most strenuously wish to ensure do not get access to identifiable statistical records of other agencies.

4. One of the few areas where the educational material of the papers is, I believe, relatively incomplete relates to complementary disclosure. Very little is said about it in any of the papers except that by Dr. Cox. The proposed guidelines suggest only that agency policies should deal with situations where sets of tables can be algebraically manipulated in such a fashion that the result is an unacceptable disclosure. The truth of the matter is that, as demonstrated in my 1972 paper, the detection of such disclosure is mathematically equivalent to the comparison of the ranks of two typically hugh matrices--in other words not feasible in general. In spite of the very great difficulties involved, most statistical offices carry out a valiant effort to check their publication programmes for residual disclosure. This effort, although undoubtedly not complete, has nevertheless been largely successful so far--at least if the absence of complaints can be accepted as a yardstick. Thus, although agencies could not guarantee that all residual disclosure is detected, they managed to keep at least one step ahead of the risk. I think the educational value of the papers could be significantly enhanced by the inclusion of a substantive discussion of the problems related to residual disclosure, together with a documentation of the best agency practices in the field.

The paper by Dr. Cox deals with a particular procedure designed to prevent residual disclosure in business surveys. It is a description of a proposed algorithm--thus it is not, nor is it designed to be, a substitute for the educational type discussion mentioned above. In fact, the detection and avoidance of complementary disclosure can be considered as a process involving three steps. The first is the detection of complementary disclosure. The paper avoids this problem since it assumes that the classificatory variables which define statistical tables are sufficiently small in number so that all possible logical tables can explicitly be displayed and considered. For example, in business surveys if all tabulation cells are defined strictly in terms of, say, geography and SIC, then the maximum disaggregation of the data is defined by the finest level of geography cross-classified by the finest level of SIC. If there is no disclosure at this level of disaggregation, then of course there can be no disclosure at higher levels of aggregation. The next step involves checking the disclosure status of any proposed or derivable tabulation cell. This is a relatively easy step. The last is the remedial step. In other words, should a potential tabulation cell be a disclosure, it would have to be suppressed, together with enough other cells sufficient to prevent the calculation of the suppressed cell as a linear combination of the published ones. It is this last, and very difficult step, which Cox addresses explicitly. The author describes an algorithm designed to create a suppression pattern within a predetermined set of publications so as to protect against all would-be disclosures, while taking great pains to avoid over-protection (i.e. over-suppression). The great advantage of the algorithm is that it seems to work. However, its theoretical properties are as yet largely unexplored: is all residual disclosure indeed avoided, and is it avoided at minimal cost in terms of unnecessary suppressions? A more practical question relates to the dimensionality of tables involved in the publication program: the algorithm can deal with tables of relatively low dimensions, such as those defined by geography and SIC. What if other classificatory variables are involved in the definition of tables: such as employment size groups, assets in terms of ranges, use of different forms of energy, etc. Conceptually, every one of the questions on the Economic Census forms is a candidate for defining an additional dimension of the tables. At what point would the algorithm break down or become prohibitively expensive to apply? This question is of considerable interest: in the Population Census publications almost every question on the

questionnaire is actually used as a classificatory variable in at least some of the tables.

Raising these questions should not be conceived as a criticism of Dr. Cox's achievement: he has taken a giant step toward the absolutely necessary development of mass production residual disclosure analysis, corresponding to the mass production of statistical tables. I am looking forward with great anticipation to further contributions from him.

5. This brings me to my next point. With a few exceptions, the material of the papers, taken together, deals with two kinds of dissemination programs: the usual printed publications, and public use tapes. A third kind of dissemination will, I believe, enjoy increasing importance in the future: ad hoc, custom-made retrievals. As indicated elsewhere, I strongly believe that the nature of surveys and censuses will change in an important way: instead of being vehicles for the production of some predetermined tabulations, they will be viewed as sources of statistical tabulations to be used and reused. Thus the relative importance of user-requested ad hoc retrievals will increase. If I am correct in this assumption, then some important consequences follow. First of all, as the amount of information in the public domain increases, the problem of detecting residual disclosure will increase exponentially. Second, each released data point represents a potential restriction placed on future retrievals, therefore posing for statistical offices a whole new class of problems: how to balance the extent of planned publications in relation to future, and therefore unspecified, ad hoc retrieval requests.

At least in the case of our 1971 and 1976 Census dissemination program, we came to the conclusion that the only way we could deal with this problem is to literally eliminate it. In effect, by random rounding every data aggregate disseminated from the census, the problem of residual disclosure largely disappears--whether in the context of pre-planned publications or with respect to subsequent ad hoc retrievals. Of course, this introduces another trade-off over and above that of "the right to privacy vs. the need to know": namely that of the amount of data that can be disseminated before residual disclosure de facto chokes off the data supply, versus a marginal increase in the mean squared error for each disseminated data point. In light of the basic importance of this trade-off, I fully support the recommendation of Michael et al relating to a program of research and development on "the impact of deliberately introduced random noise on statistical analysis as well as on disclosure risk". I also welcome the proposed research on "software systems for providing controlled on-line access to microdata files". The provision of such on-line access would truly unlock federal statistical micro-data for extensive utilization going far beyond the pre-planned publication program, provided that software can be developed which would prevent the retrieval of data involving statistical disclosure. Having said this, I disagree with Bell with respect to the somewhat simplistic treatment of the impact of random noise on the reliability of the published data: it deals with this additional source of error in isolation rather than in the context of the overall MSE. It may well be that random rounding has a rather small effect on the MSE of reasonably large aggregates (because for large numbers the relative rounding error is small), and has a moderate effect on even relatively small aggregates because for these the sampling and non-sampling errors are generally large to begin with.

6. My last point relates to the issue of statistical matching, discussed by Radner and Muller. I largely agree with their discussion. I would want to be a little more cautious then they are with respect to this procedure. In a situation where social scientists are so hungrily looking for increasingly rich data bases, statistical matching is a dangerously attractive procedure for creating files containing the logical union of the variables involved in either of the component files. Of course, the issue is not the marginal distribution of any single variable: the two files separately can produce these. If statistical matching is carried out, it is to create a file from which the joint distribution of the variables in the component files can be studied. But it is precisely here where statistical matching, at the present time, is largely based on typically unsubstantiated assumptions. I would like to see a good deal of empirical evaluation of the validity of such joint distributions before I would suggest removing the label from this procedure: "DANGEROUS - USE WITH CAUTION".

In conclusion, I must emphasize once again my admiration of the authors and of the Statistical Policy Division of OMB for having undertaken this study. The subcommittee is dealing with one of the truly most difficult conceptual issues facing statistical offices. It is dealing with the problem with great insight and sensitivity and is clearly in the process of producing educational material of the highest quality.

MISCLASSIFICATION PROBLEM AND ITS RELATION TO THE CONTINGENCY TABLE WITH SUPPLEMENTAL MARGINS

T. Timothy Chen, The Upjohn Company

1. Introduction. In many studies, data may have errors. It could happen if we use fallible and inexpensive, rather than exact and expensive, devices to measure some variables. For example, in epidemiological studies, data are usually collected from an inexpensive interview instead of physicians' examination or laboratory chemical tests. If the data are categorical, this problem is called the misclassification problem.

Suppose we are interested in one variable only, which has r possible categories; due to using a fallible and inexpensive device, we observe a different variable with same r categories. Let us use a two-dimensional r x r contingency table to represent the situation, the first dimension is the fallible classification and the second dimension is the correct or true classification; let the probability of any observation having (i,j) as its fallible and correct classification be π_{ij}, and $\Sigma\Sigma\pi_{ij} = 1$. The elements $\{a_{i,j}\}$ of misclassification matrix A is defined as $a_{i,j} = \pi_{ij}/\pi_{+j}$, which is the conditional probability of any observation having i as the fallible classification given that it has j as the true classification. If $a_{i,j}$'s do not depend on j, then we have a random misclassification and $\pi_{ij} = \pi_{+j}\pi_{i+}$.

Now instead of just one variable, we are interested in the interrelationship between two variables, where the first variable X has r possible categories and is subjected to misclassification, and the second variable Y has t possible categories and can be easily determined without error. Let us use a three-dimensional r x r x t contingency table to describe the situation; the first and the second dimensions represent the fallible and the correct classifications of the variable X, and the third dimension represents the variable Y. The misclassification matrix A is a r by rt matrix with elements $\{a_{i,jk}\}$, where $a_{i,jk} = \pi_{ijk}/\pi_{+jk}$, which is the conditional probability of any observation having i as the fallible X value, given j,k are the true X and Y values. If $a_{i,jk}$'s do not depend on k, then the misclassification is the same for any Y value and we have

$$\pi_{ijk} = \pi_{ij+}\,\pi_{+jk}/\pi_{+j+}, \qquad (1.1)$$

which is the model of conditional independence of the first and the third dimensions in each layer of the second dimension. Let us denote this model by $H_{(12,23)}$, where the 12-marginal and the 23-marginal counts are the complete minimal sufficient statistics under the Poisson or multinomial sampling schemes (see Bishop, Fienberg, and Holland (1975) and Haberman (1974)). From equation (1.1), we can see that independence on the 23-margin implies independence on the 13-margin, but not vice versa unless the matrix A has r as its rank.

Diamond and Lilienfeld (1962), Newill (1962), and Rogot (1961) considered the above model in the case r = t = 2 and they showed that

$$\left|\frac{\pi_{+11}}{\pi_{++1}} - \frac{\pi_{+12}}{\pi_{++2}}\right| > \left|\frac{\pi_{1+1}}{\pi_{++1}} - \frac{\pi_{1+2}}{\pi_{++2}}\right|, \qquad (1.2)$$

and

$$\left|\frac{\pi_{+11}\pi_{+22}}{\pi_{+12}\pi_{+21}} - 1\right| > \left|\frac{\pi_{1+1}\pi_{2+2}}{\pi_{1+2}\pi_{2+1}} - 1\right|. \qquad (1.3)$$

In epidemiology, if Y represents two different populations, and X represents having disease or not, then the above two equations say that the true risk difference is greater than the fallible or stated risk difference, and the true approximate relative risk (true odds ratio or its inverse whichever is greater than 1) is greater than the fallible or stated approximate relative risk. But these will not be true with probability one when we substitute the population π_{ijk}'s by the observed proportions. Equations (1.2) and (1.3) can be explained intuitively by the log-linear representation of the model $H_{(12,23)}$,

$$\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{23(jk)}, \qquad (1.4)$$

where we see no u_{13}-terms; hence, the risk difference and the approximate relative risk on the 13-margin are smaller than those of the 23-margin.

Since it is very expensive to observe the true X values, we usually only collect the fallible X and the true Y values; i.e., we only observe the 13-margin of a three-dimensional contingency table. Bross (1954), Rubin, Rosenbaum, and Cobb (1956), and Mote and Anderson (1962) discussed about the inference of the relationship between true X and true Y (23-margin) in this situation. They concluded that the usual chi-square test of independence or homogeneity on the observed 13-margin is a correct α-level test with less power for the independence or homogeneity on the unobserved 23-margin, provided that the model $H_{(12,23)}$ is true and the misclassification matrix A has r as its rank.

Now let us discuss the situation where the variable Y is also subjected to misclassification.

Let the fallible and the true X be the first and the third dimensions, the fallible and the true Y be the second and the fourth dimensions of a four-dimensional contingency table. The misclassification matrix A is a rt by rt matrix with each element $a_{ij,kl} = \pi_{ijkl}/\pi_{++kl}$. If the element of the matrix A, $a_{ij,kl}$ is a product of two probabilities: one is the probability of any observation being fallibly classified into i on X variable given that its true X is k, and the other is the probability of any observation being fallibly classified into j on Y variable given that its true Y is l, then we have a model of independent misclassification and

$$\pi_{ijkl} = (\pi_{i+k+}\ \pi_{+j+l}\ \pi_{++kl})/(\pi_{++k+}\ \pi_{+++l}). \quad (1.5)$$

From the above equation, it's clear that independence on the 34-margin implies independence on the 12-margin, but not vice versa unless the matrix A is non-singular. Also, if we only look at the 134-marginal table, then the misclassfication matrix is independent of the variable Y. For the 234-marginal table, the misclassification matrix is independent of the variable X. We denote this model by $H_{(13,24,34)}$ and we have a log-linear reppresentation:

$$\log \pi_{ijkl} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(l)} + u_{13(ik)} + u_{24(jl)} + u_{34(kl)}, \quad (1.6)$$

where we don't have u_{12}-terms. Keys and Kihlberg (1963) and Gullen, Bearman, and Johnson (1968) discussed the above model in the case r=t=2 and they showed that

$$\left|\frac{\pi_{++11}}{\pi_{+++1}} - \frac{\pi_{++12}}{\pi_{+++2}}\right| > \left|\frac{\pi_{11++}}{\pi_{+1++}} - \frac{\pi_{12++}}{\pi_{+2++}}\right|. \quad (1.7)$$

It can also be shown that

$$\left|\frac{\pi_{++11}\pi_{++22}}{\pi_{++12}\pi_{++21}} - 1\right| > \left|\frac{\pi_{11++}\pi_{22++}}{\pi_{12++}\pi_{21++}} - 1\right|. \quad (1.8)$$

If Y represents two populations, and X represents having disease or not, above equations say that the risk difference and the approximate relative risk on 12-margin (fallible X and Y) are smaller than those of 34-margin (true X and Y), which can be explained intuitively by equation (1.6). When we observe only the fallible classifications for both variables, under assumptions of independent misclassification and non-singularity of the matrix A, Assakul and Proctor (1967) showed that the usual chi-square test of independence on the observed 12-margin would give us a correct α-level test of independence on the unobserved 34-margin, but misclassification reduced the power of this test comparing to the direct test on the 34-margin. In case of non-independent errors they showed that the test on the 12-margin, in general, would have a larger type I error for the independence hypothesis on the 34-margin.

Above discussion shows that log-linear models provide a class of models which give meaningful interpretation for the misclassification matrix, and under some models the test on the observed fallible data provide a correct test for the unobserved true data. But unless from past experience or from examination of some data which have both the fallible and the true values, we are not sure about the applicability of a particular model for the misclassification matrix. Therefore, besides observing the inexpensive fallible data, we should also collect both fallible and true data on some observations. This is the double sampling scheme proposed recently by Tenenbein (1970, 1971, 1972) and Chiaccheierini and Arnold (1977); the data collected can be presented as a full contingency table of both fallible and true data with a supplemental lower dimensional margin of fallible data.

2. Double Sampling Scheme. In this section we will discuss how to analyze categorical data with misclassification and double sampling. The detail of analysis will be shown for a three-dimensional contingency table with the first and the second dimensions representing the fallible and the true X, and the third dimensions representing the true Y variable. Suppose we observe n subjects with all three dimensions, and N-n subjects for the first and the third dimensions; the observed counts in the full table are denoted by x_{ijk} and the observed counts in the supplemental 13-margin are denoted by V_{ik} (where $\Sigma\Sigma\Sigma x_{ijk} = n$, and $\Sigma\Sigma V_{ik} = N-n$). We assume all x_{ijk} are greater than zero for simplicity. The main inference is about the independence of the true X and the true Y variables, but specifying a correct structure of misclassification may give us a better power for the test. The structures of misclassification we want to investigate are those log-linear models having u_{23}-terms like $H_{(123)}$, $H_{(12,13,23)}$, $H_{(12,23)}$, $H_{(13,23)}$, and $H_{(1,23)}$. The first model $H_{(123)}$ can be expressed as

$$\log \pi_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}, \quad (2.1)$$

with each set of subscripted u-terms adds to zero when summed over any subscript. This is the unrestricted (saturated) model where we have no restriction on π_{ijk}. The second model is a model of no second-order interaction with $u_{123(ijk)} = 0$ in (2.1). The third and the fourth models are models of conditional independence as explained in section 1. The fifth model is a model of independence between the first dimension and the other two dimensions, which is equivalent to (2.1) with $u_{12(ij)}$, $u_{13(ik)}$, $u_{123(ijk)}$ all set to zero.

Since we have double sampling data, the expected counts for x_{ijk} and V_{ik} are $n\pi_{ijk}$ and $(N-n)\pi_{i+k}$ respectively. Under the unrestricted model $H_{(123)}$, we have the following ML equations:

$$N\hat{\pi}_{ijk} = x_{ijk} + V_{ik}\ \hat{\pi}_{ijk}/\hat{\pi}_{i+k}, \quad \forall i,j,k, \quad (2.2)$$

where the right hand side is the observed count in the cell (i,j,k) plus a proportional allocation of supplemental marginal count to that cell based on the MLE's $\{\hat{\pi}_{ijk}\}$. For the no second-order interaction model, $H_{(12,13,23)}$, the ML equations are given by:

$$N\hat{\pi}_{ij+} = x_{ij+} + \sum_k V_{ik}\ \hat{\pi}_{ijk}/\hat{\pi}_{i+k}, \quad \forall i,j, \quad (2.3)$$

$$N\hat{\pi}_{i+k} = x_{i+k} + V_{ik}, \quad \forall i,k, \quad (2.4)$$

$$\text{and } N\hat{\pi}_{+jk} = x_{+jk} + \sum_i V_{ik}\hat{\pi}_{ijk}/\hat{\pi}_{i+k}, \quad \forall j.k. \quad (2.5)$$

Next for the model $H_{(12,23)}$, the ML equations are given by equations (2.3) and (2.5). For the model $H_{(1,23)}$, the ML equations are given by equation (2.5) and

$$N\hat{\pi}_{i++} = x_{i++} + V_{i+}, \quad \forall i. \quad (2.6)$$

In general the ML equations will correspond to the highest order subscripted u-terms in the model. We can use an iterative procedure such as the one described below to get a numerical solution to the ML equations. The iterative procedure we propose is an extension of the iterative proportional fitting used by Bishop et al (1975), Goodman (1970) and Haberman (1974). For all models, we take the same initial value: $\pi^{(0)}_{ijk} = 1/r^2t$ for all i,j,k. For a given log-linear model each cycle consists of a set of pairs of steps, each pair corresponding to one of the sets of ML equations for the model. For example, for the model $H_{(12,13,23)}$, each cycle of the iterative procedure consists of the following six steps:

$$\pi^{(v+1)}_{ij+} = \left(x_{ij+} + \Sigma_k V_{ik}\ \pi^{(v)}_{ijk}/\pi^{(v)}_{i+k}\right)/N, \quad \forall i,j, \quad (2.7)$$

$$\pi^{(v+1)}_{ijk} = \pi^{(v)}_{ijk}\ \pi^{(v+1)}_{ij+}/\pi^{(v)}_{ij+}, \quad \forall i,j,k, \quad (2.8)$$

$$\pi^{(v+2)}_{i+k} = \left(x_{i+k} + V_{ik}\right)/N, \quad \forall i,k, \quad (2.9)$$

$$\pi^{(v+2)}_{ijk} = \pi^{(v+1)}_{ijk}\pi^{(v+2)}_{i+k}/\pi^{(v+1)}_{i+k}, \quad \forall i,j,k \quad (2.10)$$

$$\pi^{(v+3)}_{+jk} = \left(x_{+jk} + \Sigma_i V_{ik}\ \pi^{(v+2)}_{ijk}/\pi^{(v+2)}_{i+k}\right)/N, \quad \forall j,k, \quad (2.11)$$

$$\pi^{(v+3)}_{ijk} = \pi^{(v+2)}_{ijk}\ \pi^{(v+3)}_{+jk}/\pi^{(v+2)}_{+jk}, \quad \forall i,j,k. \quad (2.12)$$

For the model $H_{(12,23)}$, each cycle of the iterative procedure consists of the four steps given by equations (2.7), (2.8), (2.11) and (2.12) with $\pi^{(v+2)}_{ijk} = \pi^{(v+1)}_{ijk}$.

Once we have the MLE's for cell probabilities, we can compute either the Pearson or likelihood ratio statistics to test the goodness-of-fit of the model:

$$X^2 = \Sigma\Sigma\Sigma\ \frac{(x_{ijk} - n\hat{\pi}_{ijk})^2}{n\hat{\pi}_{ijk}} + \Sigma\Sigma\ \frac{(V_{ik}-(N-n)\hat{\pi}_{i+k})^2}{(N-n)\hat{\pi}_{i+k}}, \quad (2.13)$$

and

$$G^2 = 2\Sigma\Sigma\Sigma\ x_{ijk}\ \log\frac{x_{ijk}}{n\hat{\pi}_{ijk}} + 2\Sigma\Sigma\ V_{ik}\ \log\frac{V_{ik}}{(N-n)\hat{\pi}_{i+k}}, \quad (2.14)$$

with appropriate degrees of freedom. For the model $H_{(12,13,23)}$ we have estimated u_1, u_2, u_3, u_{12}, u_{13}, u_{23} terms; hence, we have $(r^2t-1) + (rt-1)-2(r-1)-(t-1)-2(r-1)(t-1)$ d.f. for the tests.

We will first fit the model $H_{(123)}$ to the data $\{x_{ijk}\}$, $\{V_{ik}\}$, to find out whether they are consistent to each other, i.e., whether $\{x_{ijk}\}$ and $\{V_{ik}\}$ are both random samples from the same target population. After showing this model fits the data, we can fit the next simple model $H_{(12,13,23)}$. We can examine both unconditional test and conditional test (which is the difference between two unconditional tests) statistics

to decide whether to accept this model. We can proceed like this to choose a most appropriate and simplest model to describe the data. The general step-wise procedure of fitting models for a contingency table has been described in Goodman (1971).

After a final model for the full table which still has u_{23}-terms has been chosen, i.e., we have chosen a model for the misclassification matrix, we can now test the independence (or homogeneity) of true X and true Y in the 23-margin ($H^*_{(2,3)}$). We will again obtain the MLE's $\{\hat{\pi}_{ijk}\}$ under a particular model for the full table plus the model $H^*_{(2,3)}$.

Under the model $H_{(123)}$ and $H^*_{(2,3)}$, we have the following ML equations:

$$N\, E_{H^*}(\hat{\pi}_{ijk}) = x_{ijk} + V_{ik}\hat{\pi}_{ijk}/\hat{\pi}_{i+k}, \quad \forall i,j,k, \tag{2.15}$$

where

$$E_{H^*}(\hat{\pi}_{ijk}) = N\hat{\pi}_{ijk}[\Sigma_i(x_{ijk} + V_{ik}\hat{\pi}_{ijk}/\hat{\pi}_{i+k})]/\left\{[\Sigma_{i,j}(x_{ijk} + V_{ik}\hat{\pi}_{ijk}/\hat{\pi}_{i+k})][\Sigma_{i,k}(x_{ijk} + V_{ik}\hat{\pi}_{ijk}/\hat{\pi}_{i+k})]\right\} \tag{2.16}$$

which is the adjustment of $\hat{\pi}_{ijk}$ by the independence hypothesis on the 23-margin. Under the model $H_{(12,13,23)}$ and $H^*_{(2,3)}$, the ML equations are given by (2.3), (2.4), and (2.5) with the left hand sides substituted by $N \sum_k E_{H^*}(\hat{\pi}_{ijk})$, ...etc. These three ML equations can be solved by the following iterative procedure with

$$\pi^{(o)}_{ijk} = \pi^{*(o)}_{ijk} = 1/r^2t, \quad \forall i,j,k, \text{ then}$$

$$\pi^{*(v+1)}_{ij+} = (x_{ij+} + \Sigma_k V_{ik}\pi^{(v)}_{ijk}/\pi^{(v)}_{i+k}) / N, \quad \forall i,j, \tag{2.17}$$

$$\pi^{*(v+1)}_{ijk} = \pi^{*(v)}_{ijk}\pi^{*(v+1)}_{ij+}/\pi^{*(v)}_{ij+}, \quad \forall i,j,k \tag{2.18}$$

$$\pi^{(v+1)}_{ijk} = \pi^{*(v+1)}_{ijk}\, \pi^{*(v+1)}_{+j+}\, \pi^{*(v+1)}_{++k}/\pi^{*(v+1)}_{+jk} \quad \forall i,j,k, \tag{2.19}$$

and the other six steps are just similar modifications of (2.9), (2.10), (2.11), and (2.12) into procedures like (2.17), (2.18), and (2.19). The rationale behind the whole procedure is that we first obtain $\underset{\sim}{\pi}^{*(v)}$ in the parameter space specified by the model for the full table, then we adjust $\underset{\sim}{\pi}^{*(v)}$ to $\underset{\sim}{\pi}^{(v)}$, which is in the intersection of the above space and the space specified by $H^*_{(2,3)}$. The convergence can be achieved if there is no empty cell in the full table, since the likelihood function is concave and bounded above.

Once the MLE's $\{\hat{\pi}_{ijk}\}$ are obtained, we can test the goodness-of-fit of the model by computing either the Pearson or likelihood ratio statistics as (2.13) and (2.14). For the model $H_{(12,13,23)}$ $H^*_{(2,3)}$, since we have 23-marginal independence constraints on those u-terms, we reduce the number of free u-terms by $(r-1)(t-1)$, so we have $(r^2t-1)+(rt-1)-2(r-1)-(t-1)-(r-1)^2-(r-1)(t-1)$ d.f. for the tests. We will decide whether $H^*_{(2,3)}$ is true or not conditioning upon a particular model for the misclassfication matrix. The value of this conditional test statistic does depend on the model we've specified for the full table.

It should be noted here, the model $H_{(12,23)}$ $H^*_{(2,3)}$ is equivalent to $H_{(12,3)}$, similarily $H_{(13,23)}$ $H^*_{(2,3)}$ is $H_{(13,2)}$ and $H_{(1,23)}$ $H^*_{(2,3)}$ is $H_{(1,2,3)}$, which is mutual independence of three dimensions. The model $H_{(12,13)}$ does not have u_{23}-terms, hence the ML equations for $H_{(12,13)}$ $H^*_{(2,3)}$ are not the type specified in (2.15) and (2.16).

The method described above can be extended easily to higher dimensional table with many variables subjected to misclassification. We will first build log-linear models for the full table (including both fallible and true classifications), which have u-terms corresponding to the lower dimensional margin of true classifications. The method for this was explained in detail in Chen (1972). After a model is finally chosen for the full table, we can then build log-linear models for the lower dimensional margin of true classifications using similar procedure as explained in this paper. The iterative procedures proposed herein are examples of the generalized EM algorithm given in Dempster, Laird, and Rubin (1977). A computer program, which is an extension of Haberman (1972), has been written according to the method in this paper to give MLE's of cell

probabilities and counts under different models and produce both goodness-of-fit statistics with appropriate degrees of freedom. It is available to any interested person upon request.

Tenenbein (1970, 1971, 1972) first proposed using a double sampling scheme to make inference about categorical data with misclassification. He only discussed the estimation problem in one variable case without any assumption on the misclassification matrix. The estimates he obtained are similar to those obtained in Chen and Fienberg (1974). He derived formula to determine the optimum double sampling ratio (n/N) so that the variances of estimates are smallest; his formulas may be used in our model building problem. Chiacchierini and Arnold (1977) discussed a test of independence for the two variable case with r=t=2, which is our conditional test of $H^*_{(3,4)}$ given that $H_{(1234)}$ is true.

3. An Example. Cobb and Rosenbaum (1956) reported an arthritis study in the Arsenal Health District of Pittsburgh. A household morbidity survey was conducted in July, 1952, using a random sample of 3,000 households. All the persons over 14 years old in these households were classified into three strata, based on the information regarding rheumatism and arthritis obtained by non-medical interviewers: Stratum 1 consisted of individuals who were recorded as having arthritis or rheumatism; Stratum 2 consisted of individuals who were recorded free of arthritis or rheumatis, but were reported to have some rheumatic symptoms; Stratum 3 was made up of the remainder who were not recorded as suffering from rheumatis, arthritis, or related manifestations. A random sample of persons was selected for each sex separately and within each strata. The sampling rate was 60% for males and 30% for females in the Strata 1 and 2, 7% for both males and females in Stratum 3; this resulted in a total sample of 798 persons. Each person thus sampled was visited in his home by a non-medical interviewer equipped with the detailed arthritis questionnaire, and the individuals who were interviewed were urged to have an examiniation by physicians in the arthritis clinic. Some persons refused the interview, or were unavailable for interview, and some did not return to the clinic for examination. The final data included 478 people with both the interview and the examination. The data about whether the person had joint pain is given in Table 1. The two "unknown" rows were not reported in Cobb and Rosenbaum (1956); instead, they are generated artificially as supplemental data to demonstrate the methodology. Let the first dimension be the interview result, the second dimension be the physician's history, and the third and fourth dimensions be the strata and the sex.

1. Number of Persons Having Joint Pain by Sex and Stratum as Obtained by Physicians vs by Non-Medical Interviewers

	Interview Result					
Physician's	Yes			No		
Examination	Stratum 1	2	3	Stratum 1	2	3
	a.Males					
Yes	65	24	35	5	12	20
No	2	5	16	2	10	41
Unknown	69	25	45	10	23	70
	b. Females					
Yes	64	23	36	4	11	25
No	3	2	7	1	5	60
Unknown	69	27	37	8	20	75

We first fit the model $H_{(1234)}$ just to see whether the supplemental data are consistent with the data in the full 2 x 2 x 3 x 2 table. This model fits the data very well with $X^2 = 4.19$ and $G^2 = 4.20$, 11 d.f. We then try to fit the models which will give us nice interpretations for the misclassification matrix. Among the models $H_{(123,234)}$, $H_{(124,234)}$, $H_{(134,234)}$, the model $H_{(123,234)}$ fits the data well with $X^2 = 12.00$ and $G^2 = 11.87$, 17 d.f. When we try to fit simpler models which have the misclassification probabilities in explicit formula of the marginal probabilities, $H_{(12,234)}$ and $H_{(13,234)}$, both fail to fit the data. We then try to fit the model $H_{(12,13,234)}$, and this model fits well with $X^2 = 12.59$ and $G^2 = 12.59$, 19 d.f.; therefore, we will use it to interpret the misclassification matrix. Under this model we have

$$\pi_{ijkl} = \pi_{ijk+}\pi_{+jkl}/\pi_{+jk+}, \ \forall\, i,j,k,l, \qquad (3.1)$$

$$\text{or } \pi_{ijkl}/\pi_{+jkl} = \pi_{ijk+}/\pi_{+jk+}, \ \forall i,j,k,l \qquad (3.2)$$

Hence, the misclassification matrix are uniform over sex, and only dependent on strata.

Now we try to investigate relationship among the 234-margin of true joint pain, strata, and sex, given that the model $H_{(12,13,234)}$ is true; it turns out that the simplest model, which still has good fit, is $H_{(12,13,234)}$ $H^*_{(23,4)}$ with $X^2 = 16.10$, $G^2 = 16.14$, 24 d.f. But, since we have the fixed sex by strata margin (34-margin) originally, we have to settle on the model $H_{(12,13,234)}$ $H^*_{(23,34)}$ as the final model: the joint pain and the sex are conditionally independent given the

strata. The conclusion is that the prevalence rate of joint pain is not a function of sex, but only a function of strata. The estimates of proportions of classification errors, and the estimates of prevalence rates for joint pain under the final model $H_{(12,13,234)}$ $H^*_{(23,24)}$ are given in Table 2 by stratum.

2. The Estimates of Proportion of Classification Errors and the Estimates of Prevalence for Joint Pain by Stratum Under the Model $H_{(12,13,234)}$ $H^*_{(23,24)}$

Stratum	1	2	3
	a. Classification Errors		
False Negatives	.08	.33	.41
False Positives	.63	.24	.18
	b. Prevalence Estimates		
Physicians'	.94	.75	.48
Interviewers'	.90	.57	.38

REFERENCES

Assakul, K., and Proctor, C. H. (1967), "Testing Independence in Two-Way Contingency Tables with Data Subject to Misclassification," Psychometrika, 32, 67-76.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975), Discrete Multivariate Analysis: Theory and Practice, Cambridge, Massachusetts: MIT Press.

Bross, I. (1954), "Misclassification in 2 x 2 Tables," Biometrics, 10, 478-86.

Chen, T.T. (1972), "Mixed-up Frequencies and Missing Data in Contingency Tables," Unpublished Ph.D. Dissertation, Dept. of Statistics, Univ. of Chicago.

---, and Fienberg, S.E. (1974), "Two-Dimensional Contingency Tables with Both Completely and Partially Cross-Classified Data," Biometrics, 30, 629-42.

Chiacchierini, R.P., and Arnold, J.C. (1977), "A Two-Sample Test for Independence in 2 x 2 Contingency Tables with Both Margins Subject to Misclassification," J. Amer Statist. Assoc. 72, 170-4.

Cobb, S., and Rosenbaum, J. (1956), "A Comparison of Specific Symptom Data Obtained by Nonmedical Interviewers and by Physicians," J. Chronic Diseases, 4, 245-52.

Dempster, A.P., Laird, N.W., and Rubin, D.B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," J. R. Statist. Soc., Ser. B., 39, 1-38.

Diamond, E.L., and Lilienfeld, A.M. (1962), "Effects of Errors in Classification and Diagnosis in Various Types of Epidemiological Studies," Amer. J. Public Health, 52, 1134-44.

Goodman, L.A. (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," J. Amer. Statist. Assoc. 65, 226-56.

Goodman, L.A. (1971), "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classification," Technometrics 13, 33-61.

Haberman, S.J. (1972), "Log-Linear Fit for Contingency Tables," Applied Statis., 21, 218-25.

Haberman, S.J. (1974), The Analysis of Frequency Data, Univ. of Chicago Press, Chicago.

Keys, A., and Kihlberg, J.K. (1963), "Effects of Misclassification on Estimated Relative Prevalence of a Characteristic," Amer. J. Public Health, 53, 1656-65.

Mote, V.L., and Anderson, R.L. (1962), "An Investigation of the Effect of Misclassification on the Properties of χ^2-Test in the Analysis of Categorical Data," Biometrika, 52, 95-109.

Newell, D.J. (1962), "Errors in the Interpretation of Errors in Epidemiology," Amer. J. Public Health, 52, 1925-8.

Rogot, E. (1961), "A Note on Measurement Errors and Detecting Real Differences," J. Amer. Statist. Assoc., 56, 314-9.

Rubin, T., Rosenbaum, J., and Cobb, S. (1956), "The Use of Interview Data for the Detection of Associations in Field Studies," J. Chronic Diseases, 4, 253-66.

Tenenbein, A. (1970), "A Double Sampling Scheme for Estimating from Binomial Data with Misclassification," J. Amer. Statist. Assoc., 65, 1350-61.

---(1971), "A Double Sampling Scheme for Estimating From Binomial Data With Misclassification: Sample Size Determination," Biometrics, 27, 935-44.

---(1972), "A Double Sampling Scheme for Estimating From Misclassified Multinomial Data With Application to Sampling Inspection," Technometrics, 14, 187-202.

LINEAR FLOW GRAPHS FROM CONTINGENCY TABLES: A CONDITIONAL PROBABILITY APPROACH TO LAZARSFELDIAN CAUSAL ANALYSIS*

James R. Beniger
Department of Sociology
Princeton University

Because is in your mind.
--Screamin' Jay Hawkins
(1968), a painting by
Karl Wirsum, American

In contingency table analysis, the concept of " because " -- which may well be only a state of mind--must be introduced by means of an assymetric model. Unlike symmetric models based on odds ratios, which have already captured the attention of social scientists and applied statisticians (see Goodman 1970; 1972; an encyclopedic overview is provided by Bishop, Fienberg and Holland 1975), asymmetric models--based on proportions rather than odds ratios -- have been slower to gain acceptance (but see Goodman 1963; Coleman 1970; Davis 1975b). Nevertheless, asymmetric models seem to hold considerable promise for the type of causal and dynamic contingency table analysis developed by Paul Lazarsfeld and his Columbia colleagues in the 1950s (e.g., Kendall and Lazarsfeld 1950), and still predominant in much of survey, communications and market research.

A particularly promising approach to asymmetric analysis is that of linear flow graphing; this application was first suggested by Huggins and Entwisle (1968). Stinchcombe (1968) introduced the technique into systematic social theory construction, and Heise (1975) employed it to develop major principles of path analysis. Davis (1975a) is a comprehensive treatment of d-system flow graphing.

The purpose of this paper is to motivate the asymmetric analysis of contingency tables using proportions (differences in row and column percentages) in terms of conditional probability. This approach affords a natural causal interpretation, in the sense of changes in future probabilities, for linear flowgraph analysis of nominal or categorical variables. Extensions of the concept of causality to logical and set theoretic notions is also suggested.

Section 1 introduces the notion of partial or contributing cause, for which a measure (the coefficient d_{BA}) is proposed in Section 2 . This measure is extended to contingency tables in Section 3 , and to causal flow graphs in Section 4. Examples using the data of J.A. Davis (1975a) on region, education and racial tolerance from the NORC General Social Survey are given for the two-event case in Sections 5 and 6 , and the three-event case in Sections 7 and 8.

1. Introduction

Consider two events, A and B, such that A can be assumed, for extra-mathematical reasons, to be a contributing or partial cause (i.e., neither a necessary nor sufficient cause) of B (e.g., A is temporally prior to B) . Then the probability of B , given that A has occurred, is greater than it is when A has not occurred, i.e.,

$$P(B/A) > P(B/A^*) \quad . \qquad (1)$$

This is a necessary but not sufficient condition for A to be a cause of B , controlling for all confounding effects of other events. (The case in which A has a negative or dampening effect on B, i.e., where $P(B/A) < P(B/A^*)$, may be of equal substantive interest; this case can be treated as equivalent to event (1) by reversing the definitions of A and A*). Note two special cases of (1): when

$$P(B/A^*) = 0 \quad , \qquad (2)$$

A is said to be a necessary cause of B, and when

$$P(B^*/A) = 0 \quad , \qquad (3)$$

A is said to be a sufficient cause of B ; A is said to be a necessary and sufficient cause of B whenever the intersection of events (2) and (3) obtains.

2. Measuring Partial Causes

Given that A is a partial cause of B , it is often of substantive interest to measure the degree or strength of the relationship between A and B. This task might be seen as one of decomposing the probability that B will occur, P(B), into "explained" (by the occurrence of A) and "unexplained" probabilities. The unexplained probability, call it d_{BA} (as in regression and path notations, the first subscript (B) denotes the dependent event or effect , the second subscript (A) denotes the independent event or cause), will be a function of P(A). What is the expression for d_{BA} ?

From the definition of conditional probability,

$$P(B/A) = P(B \cap A)/P(A) \quad , \qquad (4)$$

and the fact that

$$P(B) = P(B \cap A^*) + P(B \cap A) \quad , \qquad (5)$$

it follows that

$$P(B) = P(B/A^*)P(A^*) + P(B/A)P(A) \quad . \qquad (6)$$

Substituting 1-P(A) for P(A*) , equation (6) becomes

$$P(B) = P(B/A^*) + P(A)[P(B/A) - P(B/A^*)] \quad . \qquad (7)$$

Equation (7) is in the desired form, namely,

*This paper was written while the author was a graduate student in the Departments of Sociology and Statistics, University of California, Berkeley. The research was funded, in part, by fellowship 1 F31 DA 05082-01, awarded by the National Institute on Drug Abuse, DHEW.

$$P(B) = \underset{\text{unexplained}}{P(B/A^*)} + \underset{\text{explained}}{d_{BA}P(A)} , \qquad (8)$$

where

$$d_{BA} = P(B/A) - P(B/A^*) . \qquad (9)$$

The coefficient d_{BA} has several desirable properties as a measure of the degree of P(B) "explained" by A. When A and B are independent, i.e., when $P(B) = P(B/A) = P(B/A^*)$, then $d_{BA} = 0$. When A is a necessary cause of B (i.e., when (2) holds), the "unexplained" term $P(B/A^*)$ equals 0 , and d_{BA} becomes P(B/A), i.e., the entire P(B) is explained by A.

3. Contingency Tables

Readers familiar with contingency table analysis will recognize d_{BA} from equation (9) as a difference in proportions in a two-by-two table. Consider the table

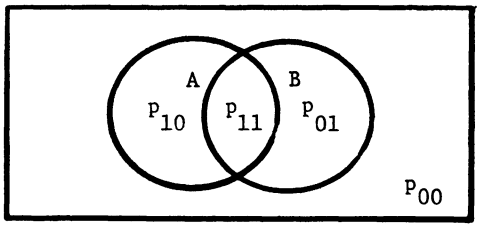
A \ B	0	1	
0	p_{00}	p_{01}	$p_{0\cdot}$
1	p_{10}	p_{11}	$p_{1\cdot}$
	$p_{\cdot 0}$	$p_{\cdot 1}$	

Here the marginal probabilities are the probabilities of individual events,

$$p_{0\cdot} = P(A^*) \qquad p_{\cdot 0} = P(B^*)$$
$$p_{1\cdot} = P(A) \qquad p_{\cdot 1} = P(B) .$$

The cell probabilities are the probabilities of the four possible intersections of A and B,

$$p_{00} = P(A^* \cap B^*)$$
$$p_{10} = P(A \cap B^*)$$
$$p_{01} = P(A^* \cap B)$$
$$p_{11} = P(A \cap B) .$$



The row and column probabilities are the eight possible conditional probabilities involving A and B , by the definition of conditional probability in (4):

$$n_{00}/n_{\cdot 0} = P(A^*/B^*) = P(A^* \cap B^*) \;/\; P(B^*)$$
$$n_{10}/n_{\cdot 0} = P(A/B^*) = P(A \cap B^*) \;/\; P(B^*)$$
$$n_{01}/n_{\cdot 1} = P(A^*/B) = P(A^* \cap B) \;/\; P(B)$$
$$n_{11}/n_{\cdot 1} = P(A/B) = P(A \cap B) \;/\; P(B)$$

*****

$$n_{00}/n_{0\cdot} = P(B^*/A^*) = P(B^* \cap A^*) \;/\; P(A^*)$$
$$n_{01}/n_{0\cdot} = P(B/A^*) = P(B \cap A^*) \;/\; P(A^*)$$
$$n_{10}/n_{1\cdot} = P(B^*/A) = P(B^* \cap A) \;/\; P(A)$$
$$n_{11}/n_{1\cdot} = P(B/A) = P(B \cap A) \;/\; P(A)$$

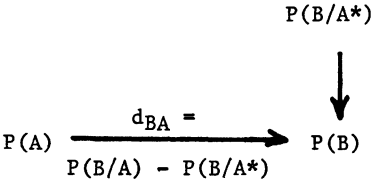
Now d_{BA} can be seen as one of the eight possible differences in row or column proportions, namely

$$d_{BA} = n_{11}/n_{1\cdot} - n_{01}/n_{0\cdot} . \qquad (10)$$

This concept has a venerable tradition in contingency table analysis , particularly in the social sciences ; hence the coefficient d_{BA} for the degree of P(B) "explained" by event A will have intuitive appeal for analysts working in this tradition.

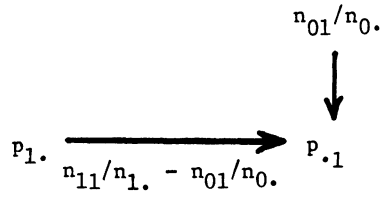
4. Causal Flow Graphs

Equations (8) and (9) can also be represented as a causal flow graph,



by applying four conventions: (1) probabilities of temporally prior or causative events (here P(A)) become values at source nodes, i.e., ones with outgoing arrows; (2) probabilities of dependent events or effects (here P(B)) become sink nodes, i.e., ones with incoming arrows; (3) the d_{ij} or "explained" probabilities (here d_{BA}) become coefficients of arrows running from source i to sink j ; and (4) "unexplained" probabilities (here P(B/A*)) become "dummy" source nodes, i.e. , ones with arrows running directly into a sink.

Causal flow graphs can be constructed directly from two-by-two tables using the following expressions:



Flow graphs have no unique mathematical properties; they merely translate equations like (8) into visual language. They do, however, facilitate substantive interpretations of data which might be less obvious in tabular or equation forms. The direct relationship between contingency tables, decomposition of conditional probabilities and causal flow diagrams has now been demonstrated.

5. Example with Two Events

The development of causal flow graphs owes much to the work of J.A. Davis (1975a). In order to facilitate comparisons between the conditional probability approach introduced here and the work of Davis , the data set used by him in illustrative examples will be adopted here. These data are pooled from the 1972 , 1973 and 1974 National Opinion Research Center (NORC) General Social Surveys (GSS) of Americans age 18 and older; the sample sizes are 1613, 1504 and 1484, respectively, for a total of 4601.

To illustrate the two-event example discussed thus far,

A is living outside of the American South (in U.S. Census regions East and West South Central and South Atlantic) at age 16;

B is stated opposition to laws against marriages between Blacks and Whites.

The hypothesis is that upbringing outside of the South (A) , an experience temporally prior to opinions expressed in 1972-4, constitutes a partial cause of tolerance on racial issues (B) . (Black respondents, and those raised in foreign countries , or failing to answer one or both questions, are excluded from this example, thus lowering the sample size from 4601 to 3786).

The cross-tabulation of A and B, from Davis' published data, is

A	B = 0	B = 1	
0	.550 597 (.158) .410	.450 489 (.129) .210	1086 (.287)
1	.318 858 (.227) .590	.682 1842 (.486) .790	2700 (.713)
	1455 (.384)	2331 (.616)	3786 (1.000)

The four values required for the causal flow graph can be computed directly from this contingency table:

$$P(A) = p_{1.} = 2700/3786 = .713 \quad (11)$$

$$P(B) = p_{.1} = 2331/3786 = .616 \quad (12)$$

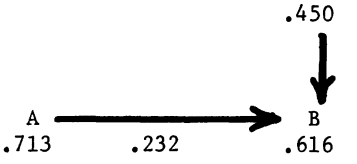
$$P(B/A^*) = n_{01}/n_{0.} = 489/1086 = .450 \quad (13)$$

$$d_{BA} = n_{11}/n_{1.} - n_{01}/n_{0.} = 1842/2700 - 489/1086 = .232 \quad (14)$$

Substituting in equation (8),

$$P(B) = .616 = \underset{\text{unexplained}}{.450} + \underset{\text{explained}}{(.232 * .713)} \quad (15)$$

The causal flow graph becomes:



It has been previously stated that such flow graphs facilitate substantive interpretations of data which might be less obvious in tabular or equation forms. The graph above, for example, might be given the following interpretation: Nonsouthern upbringing (A) is a partial cause of racial tolerance (B). A occurs with probability of .713 in the U.S.; when it does, the probability that B will also occur -- and would not have occurred otherwise -- is .232 . B occurs with probability .616 -- with probability .450 in the absence of A, and with an additional probability of .166 (.713 * .232) as a result of A.

Worth noting here is the interpretation of d_{BA} in terms of conditional probability -- as an additional probability, or the probability that an A will produce a B that would not have otherwise occurred. This interpretation can be given formal statement:

d_{BA} is the probability that a B will accompany A that would not have occurred in the absence of A.

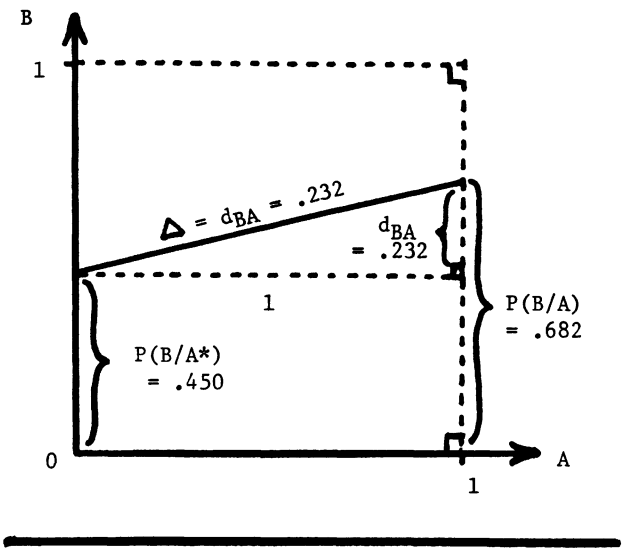
i.e., to repeat (9) , $d_{BA} = P(B/A) - P(B/A^*)$. Because of its interpretation , d_{ij} is often termed the graph transmittance value from j to i.

6. Coordinate Plots

When binary variables like A and B are assigned the values 0 and 1 , as in the tables here, at least three interpretations may be made in terms of coordinate plots: (1) the conditional probability of a 1-value on the dependent variable (B), given a 0-value on the independent variable (i.e., P(B/A*)), is the y-intercept of a coordinate plot of B against A , or the constant in a linear equation like (8) ; (2) the conditional probability of a 1-value on B, given a 1-value on A (i.e., P(B/A)), is the intercept of line A = 1; and (3) the difference in proportions d_{BA} (i.e., P(B/A) - P(B/A*)) is the slope of the linear relationship between B and A, or the coefficient in the linear equation . These graphic interpretations are illustrated in the coordinate plot at the head of the next page.

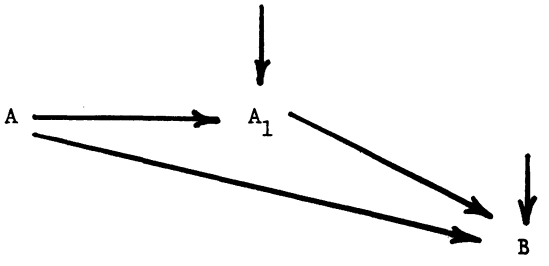
7. Three Events with No Interactions

The flow graph approach extends to systems with any number of events or variables . Consider the addition of a third event, A_1 , intervening in time between A and B. Again using the data set of Davis (1975a),



A_1 is educational attainment of at least a high school diploma.

The hypothesis is that, because upbringing outside the South (A) is more likely to result in high school education (A_1), this intervening event will at least partially "explain" the relationship between A and racial tolerance (B) in the Lazarsfeldian sense. The causal flow diagram of this hypothesis is



Because A is a direct partial cause of A_1, the relationship between A and A_1 is the same as that between A and B in equation (8), namely,

$$P(A_1) = \underset{\text{by A}}{P(A_1/A^*)} + \underset{\text{by A}}{d_{A_1A}P(A)} \quad , \qquad (16)$$

where

$$d_{A_1} = P(A_1/A) - P(A_1/A^*) \quad . \qquad (17)$$

Because B is partially caused by both A and A_1, it is helpful--to keep the flow graph analysis relatively unencumbered in this example--to make a simplifying assumption, namely, that there are no interactions between A and A_1 in determining B. Stated formally, the assumptions of no interactions between A and A_1 are:

$$P(B/A\cap A_1) - P(B/A^*\cap A_1) = P(B/A\cap A_1^*) - P(B/A^*\cap A_1^*) \quad , \qquad (18)$$

i.e., the direct effect of A on B is independent of A_1, and

$$P(B/A\cap A_1) - P(B/A\cap A_1^*) = P(B/A^*\cap A_1) - P(B/A^*\cap A_1^*) \quad , \qquad (19)$$

i.e., the direct effect of A_1 on B is independent of A. These assumptions are equivalent to the fact that d_{BA} is the same for both A_1 and A_1^*, and that d_{BA_1} is the same for both A and A^*.

Analogously to (9) and (17), because both A and A_1 are direct partial causes of B, and each has the same effect independent of the other factor (i.e., there are no interaction effects),

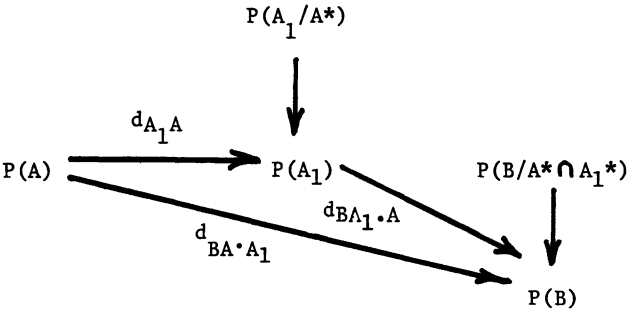
$$d_{BA\cdot A_1} = \begin{matrix} P(B/A\cap A_1) - P(B/A^*\cap A_1) \\ P(B/A\cap A_1^*) - P(B/A^*\cap A_1^*) \end{matrix} \quad , \qquad (20)$$

$$d_{BA_1\cdot A} = \begin{matrix} P(B/A_1\cap A) - P(B/A_1^*\cap A) \\ P(B/A_1\cap A^*) - P(B/A_1^*\cap A^*) \end{matrix} \quad . \qquad (21)$$

The complete equation for P(B), which combines (20) and (21), is

$$P(B) = \underset{\text{by A or } A_1}{P(B/A^*\cap A_1^*)} + \underset{\text{by A}}{d_{BA\cdot A_1}P(A)} + \underset{\text{by } A_1}{d_{BA_1\cdot A}P(A_1)} \quad . \qquad (22)$$

Equations (16) and (22) can be represented by the three-event causal flow graph



by applying the same conventions as above for source nodes and values, sink nodes, coefficients of arrows and "dummy" source nodes.

In terms of a three-way contingency table, marginal probabilities are the probabilities of individual events,

$$p_{0\cdot\cdot} = P(A^*) \quad p_{\cdot 0\cdot} = P(A_1^*) \quad p_{\cdot\cdot 0} = P(B^*)$$

$$p_{1\cdot\cdot} = P(A) \quad p_{\cdot 1\cdot} = P(A_1) \quad p_{\cdot\cdot 1} = P(B) \quad .$$

Cell probabilities are the probabilities of the eight possible intersections of A, A_1 and B,

$$p_{000} = P(A^*\cap A_1^*\cap B^*)$$

$$p_{100} = P(A\cap A_1^*\cap B^*)$$

$$p_{110} = P(A\cap A_1\cap B^*)$$

etc.

Row and column probabilities are the 24 possible probabilities of the 0- and 1-values of A, A_1 and B, each conditioned on the four possible combinations of the other two events:

$$p_{000}/p_{\cdot 00} = P(A^*/A_1^* \cap B^*)$$

$$p_{100}/p_{\cdot 00} = P(A\ /A_1^* \cap B^*)$$

$$p_{010}/p_{\cdot 10} = P(A^*/A_1\ \cap B^*)$$

$$\vdots$$

$$p_{011}/p_{01\cdot} = P(B\ /A^* \cap A_1)$$

$$p_{110}/p_{11\cdot} = P(B^*/A\ \cap A_1)$$

$$p_{111}/p_{11\cdot} = P(B\ /A\ \cap A_1)$$

8. Three-Event Example

The eight-fold cross-tabulation of A, A_1 and B, from Davis' published data, is

		B = 0	B = 1	
A = 0	A_1 = 0	341	122	463
	A_1 = 1	256	367	623
		597	489	1086
A = 1	A_1 = 0	454	383	837
	A_1 = 1	404	1459	1863
		858	1842	2700
		1455	2331	3786

The values need , in addition to (11) and (12), for the new three-variable flow graph can be computed directly from this table:

$$P(A_1) = p_{\cdot 1\cdot} = 2486/3786 = .657 \qquad (23)$$

$$P(A_1/A^*) = n_{01\cdot}/n_{0\cdot\cdot} = 623/1086 = .574 \qquad (24)$$

$$P(B/A^* \cap A_1^*) = n_{001}/n_{00\cdot} = 122/463 = .263 \qquad (25)$$

$$d_{A_1 A} = n_{11\cdot}/n_{1\cdot\cdot} - n_{01\cdot}/n_{0\cdot\cdot}$$

$$= 1863/2700 - 623/1086 = .116 \qquad (26)$$

$$d_{BA\cdot A_1} = n_{111}/n_{1\cdot\cdot} - n_{01\cdot}/n_{0\cdot\cdot}$$

$$= 1459/1863 - 367/623 = .194 \qquad (27)$$

$$= n_{101}/n_{10\cdot} - n_{001}/n_{00\cdot}$$

$$= 383/837 - 122/463 = .194 \qquad (28)$$

$$d_{BA_1\cdot A} = n_{111}/n_{11\cdot} - n_{101}/n_{10\cdot}$$

$$= 1459/1863 - 383/837 = .326 \qquad (29)$$

$$= n_{011}/n_{01\cdot} - n_{001}/n_{00\cdot}$$

$$= 367/623 - 122/463 = .326 \qquad (30)$$

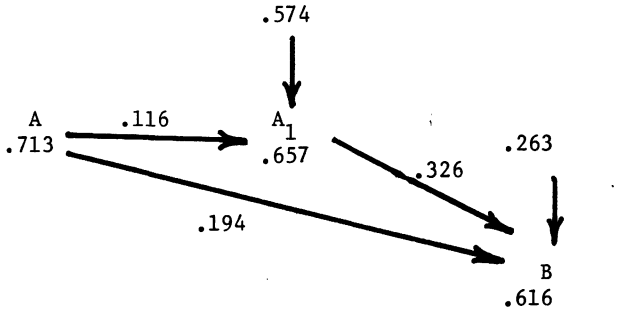
Substituting in equations (16) and (22),

$$P(A_1) = .657 = \underset{\text{unexplained}}{.574} + \underset{\text{explained by A}}{(.116 * .713)}, \qquad (31)$$

and

$$P(B) = .616 = \underset{\substack{\text{unexplained} \\ \text{by A or } A_1}}{.263} + \underset{\text{explained by A}}{(.194 * .713)} + \underset{\text{explained by } A_1}{(.326 * .657)}. \qquad (32)$$

The three-event causal flow graph becomes



This flow graph might be given the following substantive interpretation: Nonsouthern upbringing (A) is both a direct partial cause and indirect cause of racial tolerance (B) , the latter in that A is a partial cause of attaining an education of high school or above (A_1), which in turn is a partial cause of tolerance. A occurs with probability .713 ; when it does, the probabilities that A_1 and B will also occur -- and would not have occurred otherwise--are .116 and .194, respectively. A_1 occurs with probability .657, of which .083 (.713 * .116) is due to A and .574 is "unexplained" in the model. B occurs with probability .616 , which can be attributed to four factors: .138 to A acting directly (.713 * .194), .027 to A acting through A_1 (.713 * .116 * .326), .187 to A_1 acting directly (.574 * .326), and .263 to a component that remains "unexplained" in the model.

Analogously to the interpretation of d_{BA} in Section 5, $d_{BA\cdot A_1}$ may be interpreted in terms of conditional probability--as an additional probability, or the probability that an A will produce a B that would not otherwise have occurred. Similarly, $d_{BA_1\cdot A}$ can also be interpreted in terms of conditional probability--as the probability that an A_1 will produce a B that would not otherwise have occurred. The uniqueness of

$d_{BA.A_1}$ and $d_{BA_1.A}$ depends on the simplifying assumptions (18) and (19), respectively, namely, that there are no interactions between A and A_1 in determining B. In other words, d_{BA} is independent of A_1 (or the same within categories of A_1),

$$d_{BA/A_1} = d_{BA/A_1*} , \quad (33)$$

as shown in (27) and (28) , and d_{BA_1} is independent of A_1 ,

$$d_{BA_1/A} = d_{BA_1/A*} , \quad (34)$$

as shown in (29) and (30). When conditional ds are equal, as in (33) and (34), then d will have the same algebraic properties as coefficients in linear equations , or partial slopes in linear plots, as suggested by Section 6.

The interpretations of $d_{BA.A_1}$ and $d_{BA_1.A}$ can be given formal statement:

$d_{BA.A_1}$ is the probability that a B will accompany A that would not have occurred in the absence of A, independently of A_1 ;

$d_{BA_1.A}$ is the probability that a B will accompany A_1 that would not have occurred in the absence of A_1, independently of A,

which is to repeat (20) and (21) in words. Because of this interpretation , $d_{ij.k}$ is often termed the graph transmittance value from j to i, controlling for (or "within categories of") k.

In analyzing actual data, assumptions (18) and (19) would be subject to empirical verification. There is good reason for the perfect fit (i.e., total lack of interaction) in Davis' data: he began with raw figures from the NORC surveys, estimated parameters and constants, tested for interactions (finding none to be significant), and then adjusted the data to fit the resulting model (1975a, p. 129). As he reports, "With small samples , data with no significant interactions can be bouncy; with large samples, models can fit quite well despite interactions that are statistically significant" (p. 130).

9. Discussion and Summary

Conditional probability serves to motivate an asymetric interpretation of contingency tables based on differences in row and column proportions. This approach affords a natural causal interpretation, in the sense of changes in future probabilities, for linear flowgraph analysis of nominal or categorical variables like the "d system" of Davis (1975a). Motivation in terms of conditional probability will be particularly useful for survey and market researchers working with the elaboration model of the Lazarsfeldian school (involving "interpretation," reinforcers, suppressor variables, specification, spurious correlation, etc.; see Rosenberg 1968), and also as an introduction to path analysis for students familiar with statistical tables. Path diagrams involve variances and covariances , however, while flow graphs define absolute, partial and conditional probabilities. This difference makes the flow graph approach more nearly like regression, particularly in the importance of asymmetric assumptions, and indeed (as suggested by Section 6) the two procedures give identical results for data free of interaction effects. For the treatment of such effects using flow graphs, the reader is referred to Davis (1975a, pp. 125-38).

10. References

Bishop, Yvonne M.M., Stephen E. Fienberg and Paul W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

Coleman, James S. 1970. "Multivariate Analysis for Attribute Data." Pp. 217-45 in Sociological Methodology 1970, edited by E. Borgatta and G. Bohrnstedt. San Francisco: Jossey-Bass

Davis, James A. 1975a. "Analyzing Contingency Tables with Linear Flow Graphs: D Systems." Pp. 111-45 in Sociological Methodology 1976, edited by D. Heise. San Francisco: Jossey-Bass.

_____. 1975b. "Communism, Conformity, Cohorts, and Categories: American Tolerance in 1954 and 1972-73." American Journal of Sociology 81: 491-513.

Goodman, Leo A. 1963. "On Methods for Comparing Contingency Tables." The Journal of the Royal Statistical Society, Series A, 126: 94-108.

_____. 1970. "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications." Journal of the American Statistical Association 65: 225-56.

_____. 1972. "A General Model for the Analysis of Surveys." American Journal of Sociology 77: 1035-86.

Heise, David R. 1975. Causal Analysis. New York: Wiley-Interscience.

Huggins, William H., and Doris R. Entwisle. 1968. Introductory Systems and Design. Waltham, Mass.: Blaisdell.

Kendall, Patricia L., and Paul F. Lazarsfeld. 1950. Pp. 133-96 in Continuities in Social Research: Studies in the Scope and Method of the American Soldier, edited by R.K. Merton and P.F. Lazarsfeld. New York: Free Press.

Rosenberg, Morris. 1968. The Logic of Survey Analysis. New York: Basic Books.

Stinchcombe, Arthur L. 1968. Constructing Social Theories. New York: Harcourt Brace Jovanovich.

MEASURES OF ASSOCIATION FOR MULTIVARIATE CONTINGENCY TABLES

James M. Sakoda, Brown University

Need for a Measure of Association

In analyzing multidimensional contingency tables the goodness of fit of various models is generally tested via Pearson or likelihood ratio chi square. The acceptance or rejection of a model on the basis of a significance test alone runs the risk of allowing the number of cases to determine, at least in part, the number of parameters deemed to be significant. As in other test situations, judgment of the existence of a relationship should be dependent on the strength of the relationship as well as its statistical significance. If a sizable relationship is indicated the acceptable significance level might be raised to .10, say, to avoid rejecting a potentially meaningful source of variation. Conversely, effects which are extremely small, even though statistically significant, might be eliminated from a model. Measures of association are also useful in comparing tables with different numbers of cases.

Several measures of association for contingency tables have been developed for two-way tables, and one of the problems is to select an appropriate one for use with higher dimensions. We conclude that X^2 divided by the maximum X^2 for a table serves as a suitable basis for a measure of association. A second problem is the application of the chosen measure of association to higher dimensioned tables. In some situations multidimensional tables can be related to two-way tables. In the multiple correlation type of situation a dependent variable can be related to a combination of categories of independent variables by using a two-way table. In the partial correlation situation, two-way tables can be averaged over a set of control variables. For higher order effects, such as a three-way or four-way effect or for combinations of effects, reduction to two-way tables is not possible. The task is then to find the maximum X^2 for higher order effects so that they can be compared with the obtained X^2 for a given effect. Goodman (1971) suggested using a proportional reduction in X^2 as a method of calculating multiple or partial correlation coefficients. The approach suggested here differs from his in that higher order effects are analyzed in terms of the maximum X^2 rather than an arbitrarily-selected empirical X^2.

The Choice Among Measures of Association

For analysis of multidimensional tables the most convenient measures of association are those based on X^2, since data analysis is performed using X^2. It is possible to partition higher order X^2 into their component parts and to relate X^2 for two-way tables to higher order ones. Goodman (1971), for example, suggests using a proportional reduction in X^2 as a method of calculating multiple and partial correlation coefficients. X^2 is suitable with either ordered or unordered categories. Measures of association requiring ordered categories, such as Kendall's tau, Somer's D and Goodman and Kruskal's gamma are too specialized for routine contingency table analysis, since they apply only to certain tables. Moreover, gamma and its 2 x 2 table version, Yule's Q, have tendencies to be high in comparison to other coefficients when marginals are distributed unevenly.

Another advantage of X^2-based measures of association is their symmetric nature, requiring only a single measure regardless of the direction of relationship or prediction. There are a number of asymmetric measures of association which are developed on different bases and which are meaningful in different ways. These are Goodman and Kruskal's proportional reduction of errors of prediction measures lambda and tau, Margolin and Light's analysis of variance measure BSS/TSS of proportion of row variation attributable to column variation, and the proportional reduction of uncertainty measure based on information theory. Lambda cannot be recommended for tables with uneven marginals since a zero coefficient results when the largest frequencies in each column fall in one row and other measures of association show a relationship. This fault is not shared by tau, even though it is also a measure developed on the principle of proportional reduction in errors of prediction. On the other hand, tau is numerically identical to Margolin and Light's (1974) BSS/TSS measure, showing that proportional reduction in error can be quite similar to proportion of explained variation. According to Bishop, Fienberg and Holland (1975: 391), BSS/TSS and tau involve a Pearson X^2-like expression, and when the row sums are equal, they are equal to $\phi^2 / (I-1)$ and hence to $X^2 / N(I-1)$. $N(I-1)$ is maximum X^2 when $I \leq J$. The relative reduction in uncertainty measure utilizes likelihood statistics to express uncertainty, which is "variance-like" (Hays, 1973). These measures, except for their asymmetric nature (some have symmetric versions), have a great deal in common both meaningfully and numerically, with Cramér's V^2, which represents X^2 divided by maximum X^2. The interpretation of these measures is not any easier than the interpretation of X^2-based measures, as is sometimes claimed. In fact, these measures produce very small coefficients generally in comparison with measures such as V or the contingency coefficient which resemble the Pearson r rather than r^2 as these measures do.

One of the oldest X^2-based measures of association is Karl Pearson's mean square contingency or contingency coefficient:

$$C = \sqrt{\frac{\phi^2}{1 + \phi^2}} = \sqrt{\frac{X^2}{X^2 + N}},$$

where ϕ^2 is estimated by X^2 / N. "Karl Pearson showed that, if the items are capable of interpretation as a quantitatively ordered series, if the distributions are normal, and if the regression is rectilinear, C becomes identical with r as the number of categories is indefinitely in-

creased." (Peters and vanVoorhis, 1940: 392). In other words, C is an estimate of the Pearson r, but can be applied even when categories are unordered and the relationships are not linear. Its shortcoming is that its maximum value does not reach unity. But maximum ϕ^2 is the minimum of I-1 or J-1 and C_{max} can be calculated:

$$C_{max} = \sqrt{\frac{\min(I-1,J-1)}{\min(I,J)}} .$$

It is possible to correct C to achieve unity by calculating C/C_{max}, although it is not a standard practice. Tschuprow proposed the use of

$$T = \sqrt{\frac{\phi^2}{\sqrt{(I-1)(J-1)}}}$$

but achieved a maximum of 1 only for square tables. Its use is therefore not recommended.

Cramér (1946) suggested norming ϕ^2 by dividing it by its maximum value:

$$V = \sqrt{\frac{\phi^2}{\min(I-1,J-1)}} .$$

Estimating ϕ^2 by X^2/N,

$$V = \sqrt{\frac{X^2}{N \min(I-1,J-1)}} .$$

The denominator term is maximum X^2 for a two-way I x J table so that V^2 can be given a proportion of maximum variation due to interaction interpretation. V has an acceptable interpretation via V^2 and unlike C and T varies between 0 and 1. It is our choice as a suitable measure for application to higher order tables, with C/C_{max} a second possibility.

Maximum X^2

According to Cramér (1946: 443), the maximum V^2 of unity is obtained "when and only when each row (when $r \leq s$) or each column (when $r \geq s$) contains one single element different from zero." An example of an arrangement of cell frequencies for a maximum V^2 is shown in Fig. 1 for a 3 x 4 table with $I < J$. Cramér's condition can be expressed as

$$x_{ij} = x_{+j} .$$

30	-	-	-	30
-	15	5	-	20
-	-	-	10	10
30	15	5	10	60

Fig. 1. A 3 x 4 Table with Cramér's V = 1.0.

We start with the formula for X^2:

$$X^2 = N \left(\Sigma\Sigma \frac{x_{ij}^2}{x_{i+}\, x_{+j}} - 1\right) .$$

With cancellation of x_{ij} and x_{+j},

$$X^2_{max} = N \left(\Sigma\Sigma \frac{x_{ij}}{x_{i+}} - 1\right) .$$

By definition $\Sigma x_{ij} = x_{i+}$. Therefore,

$$X^2_{max} = N (\Sigma 1 - 1) = N (I-1) ,$$

or when $J \leq I$,

$$X^2_{max} = N (J-1) .$$

Hence,

$$X^2_{max} = N \min(I-1,J-1) .$$

For example,

$$X^2 = 60\left(\frac{30^2}{30\cdot 30} + \frac{15^2}{20\cdot 15} + \frac{5^2}{20\cdot 5} + \frac{10^2}{10\cdot 10} - 1\right)$$
$$= 60(1 + 1 + 1 - 1)$$
$$= 60(3-1) = 120 .$$

Analogue of Multiple Correlation

The analogue to the multiple correlation, in which Variables 2, 3 and 4, for example, are related to Variable 1, the dependent variable, can be set up by means of a two-way table. The row variable is Variable 1 with I categories and the column variable consists of all possible combinations of categories of Variables 2, 3 and 4 with J x K x L categories in all. In terms of the loglinear model this table represents an independence model (1 x 234), which tests all effects of 2, 3 and 4 on 1: (12), (13), (14), (123), (124), (134), (1234). The model can be run on a program like ECTA by fitting 1 and 234. With X^2 for the table available it is a simple matter to compute Cramér's V and it can be treated as an analogue to the multiple correlation coefficient:

$$V_{1\cdot 234} = \sqrt{\frac{X^2(1x234)}{N \min(I-1,JKL-1)}} .$$

In Fig. 2 is shown a 3 x 2 x 4 table arranged to give a maximum X^2 for Variable 1 against 2 and 3 combined. In filling the cells the distinctions between categories of Variable 2 and 3 are ignored. Maximum X^2 is equal to N x (3-1) or 240.

10	-	-	20	-	-	10	-	40
-	20	-	-	10	-	-	20	50
-	-	10	-	-	20	-	-	30
10	20	10	20	10	20	10	20	120

Fig. 2. A 3 x 2 x 4 Table Arranged for Maximum $X^2(1x23)$.

Partial Correlation

Partial correlation is generally defined as a measure of association between two variables holding constant the effects of a third variable. For continuous variables the effects of a third variable can be removed by taking the residualized score and correlating these. For discrete variables categories are generally unordered and this approach cannot be used. Instead, the alternative of setting up separate subtables for each level of the third variable is used (Agresti, 1977). For each two-way table a measure of association is calculated and these are weighted and averaged to obtain an overall measure of partial association. When using X^2 it is necessary either to assume that higher order interactions do not exist or to remove the effects.

Given a three-way table, we set up I x J tables for relationships for Variables 1 and 2 for each level of Variable 3. For each table

$$V^2_k = \frac{X^2_k}{N_k \min(I-1, J-1)} .$$

For an overall measure each V^2_k can be weighted by N_k/N, the proportion of the total number of cases in each subtable:

$$V^2 = \sum_k \frac{N_k}{N} \frac{X^2_k}{N_k \min(I-1, J-1)} .$$

The N_k's cancel out and

$$V^2 = \frac{\Sigma X^2_k}{N \min(I-1, J-1)} .$$

The numerator is the X^2 for the partial association model, (1x2|3) and tests the effects (12)(123), and can be obtained by fitting (12), (13). The denominator term is the maximum X^2 value for the partial correlation problem. From the numerator the higher order interaction must be removed, leaving only the (12) effect. $X^2(123)$ can be obtained by fitting (12), (13), (23). $X^2(1x2|3) - X^2(123)$ leaves $X^2(12)$. Hence,

$$V_{12\cdot 3} = \sqrt{\frac{X^2(12)}{N \min(I-1, J-1)}} .$$

The interpretation of $X^2(12)$ is the conditional test for (1x2|3), given no three-factor effect (Bishop, Fienberg and Holland, 1975: 171). When higher order interaction exists it is desirable to examine subtables individually.

In Fig. 3 is shown an analysis of the partial association of Preference x Use given Temperature and Softness in the Reis-Smith data. Higher order interactions are calculated by fitting (12)(134)(234). The interactions are significant at the .153 level and V of .089 and C/C_{max} of .126 indicate the existence of an appreciable amount of higher order interaction.

Model	Fitted Parameters	X^2	df	p	V	C/C_{max}
(1x2|34)	(134)(234)	27.81	6	.000	.166	.232
(123)(124)(1234)	(12)(134)(234)	8.05	5	.153	.089	.126
(12)		9.76	1	.000	.140	.196

Fig. 3. Analysis of Partial Association for the Reis-Smith Data

For the individual subtables the V_{k1}'s are

.122 .261 .225
.005 .108 .202

and they show how the partial $V_{12.34}$ is only an average and cannot reflect the full range of variations among subtables.

Higher Order Interactions

To apply Cramér's V to higher order interactions it is necessary to find the maximum X^2 corresponding to them. In Fig. 4 is shown a 3 x 2 x 2 table with frequencies arranged internally to obtain a maximum (123) interaction.

10	-	-	10	20
-	10	10	-	20
-	10	10	-	20
10	20	20	10	60

Fig. 4. A 3 x 2 x 2 Table with Maximum (123) Effect.

$X^2(123)$ obtained by fitting (12)(13)(23) is given by the ECTA computer program as 60 for the Pearson version and 76.38 for the likelihood ratio one. Evidently, the maximum X^2 for three-way interaction is given by N times the minimum of the three, I-1, J-1, K-1. Hence,

$$V_{123} = \sqrt{\frac{X^2(123)}{N \min(I-1, J-1, K-1)}} .$$

The formula for the maximum X^2 applies to the Pearson X^2 and only approximately to the likelihood ratio X^2. Hence, it is prudent to use Pearson X^2 for the numerator terms in calculating Cramér's V.

In Fig. 5 an example of maximum three-way effect for a 3 x 3 x 4 table is shown. $X^2(123)$ for this table is 240 and the likelihood ratio X^2 is 263.67. The 240 agrees with the formula:

$$X^2_{max} = 120\ (3-1) = 240 .$$

10	-	-	-	-	10	-	10	-	-	10	-
-	10	-	10	-	-	-	-	10	-	-	10
-	-	10	-	10	-	10	-	-	10	-	-

Fig. 5. A 3 x 3 x 4 Table with Maximum (123) Effect.

By analogy maximum X^2 for four-way interaction can be calculated as

$$X^2_{max} = N \min(I-1, J-1, K-1, L-1) .$$

In Fig. 6 is shown an example of maximum four-way effect for a 3 x 2 x 2 x 2 table.

10	-	-	10	-	10	10	-
-	10	10	-	10	-	-	10
-	10	10	-	10	-	-	10

Fig. 6. A 3 x 2 x 2 x 2 Table with Maximum (1234) Effect.

Pearson X^2 for this table is 120, which agrees with the formula; likelihood ratio X^2 is 152.76. An indication that this is indeed the maximum value is shown by the fact that other separate effects are zero.

To set up a table for maximum X^2 for a three-way effect one can use a latin square with the number of treatment equal to the smallest dimension. Each treatment appears only once in each row and column. Each treatment is then set up as a separate table as in Fig. 4 or 5. If there are additional rows, columns or blocks, one of the rows, columns or blocks is arbitrarily duplicated. In Fig. 4 the last row is a replicate and in Fig. 5 the last block is. For maximum four-way effects not only the rows and columns are arranged in latin-square form, but also the tables themselves. For example, in Fig. 6 Table 1 is followed by Table 2 and then Table 2 by Table 1. This forms a latin square of the form 1, 2, 2, 1.

There is no reason why Cramér's V cannot be applied to models representing a combination of effects. It would seem reasonable that the maximum X^2 would be determined by the component with the highest maximum. For example, given a model representing $X^2(123) + X^2(124) + X^2(1234)$ for a 3x3x3x4 table, the largest maximum would be for $X^2(124)$. This is N x min(I-1,J-1,L-1) or N x 2.

The calculation of C/C_{max} is possible if X^2 and X^2_{max} are available. If variables are basically continuous in nature, although tabled as discrete categories, and if an estimate of the Pearson r is desired, C/C_{max} can be calculated.

Summary

For application to multidimensional contingency tables X^2-based measures of association are the most convenient. Of the available measures based on X^2 for the two-way table Cramér's V is the most appropriate. It is applicable to both ordered and unordered categories, it is a symmetric measure, and V^2 can be interpreted as X^2 divided by the maximum possible X^2. Maximum X^2, which is given as N min(I-1,J-1) is easy to calculate. Cramér's V can be applied to the multiple correlation situation and the analogue of the partial correlation coefficient. It can also be applied to the three-way, four-way and other higher order interactions, as well as to X^2 based on a combination of effects. A second candidate is C/C_{max}.

References

Agresti, Alan (1977). Considerations in Measuring Partial Association for Ordinal Categorical Data. Journal of the American Statistical Association, 72, 37-45.

Bishop, Yvonne M. M., Fienberg, Stephen E. and Holland, Paul W. (1975). Discrete Multivariate Analysis, Theory and Practice. Cambridge, Mass., The MIT Press.

Costner, Herbert (1965). Criteria for Measures of Association. American Sociological Review, 30, 341-353.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton, N.J., Princeton University Press.

Goodman, Leo A. (1971). The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimate Methods of Building Models for Multiple Classifications. Technometrics, 13, 33-61.

Hays, William L. (1973). Statistics for the Social Sciences. New York, N.Y., Holt, Rinehart and Winston.

Margolin, Barry H. and Light, Richard J. (1974). Analysis of Variance for Categorical Data II: Small Sample Comparisons with Chi Square and Other Competitors. Journal of the American Statistical Association, 69, 755-764.

Peters, Charles C. and van Voorhies, Walter R. (1940). Statistical Procedures and Their Mathematical Bases. New York, N.Y., McGraw-Hill.

An Alternative Strategy for Estimating the Parameters of Local Areas

Steven B. Cohen, University of North Carolina at Chapel Hill
William D. Kalsbeek, Research Triangle Institute, North Carolina

1. Introduction

The ever-growing need for good estimates of the social, political, economic, and health parameters has been rapidly gaining recognition. The allocation of federal aid to both states and municipalities is often dependent upon information pertaining to population, unemployment, and housing. Candidates vying for political office are particularly concerned with obtaining reliable estimates of voter preference and participation at the sub-national level. Similarly, rather precise small area estimates of retail trade are essential indicators for the commercial sector.

Some useful information has been obtained from sources which include the decennial census and vital registration systems. Generally, federal agencies have relied upon sample surveys to provide estimates of the data they require, though such estimates pertain to the entire United States or each of its four broad geographical regions. Estimates of data for small areas are unavailable primarily due to sample size requirements which are prohibitive with respect to cost and strata designs which often cross state and county limits. Consequently, several procedures have been developed which utilize available data from large areas, local data on population and accessible local data on ancillary (symptomatic) variables, in order to produce synthetically the desired estimates. Synthetic estimation is perhaps the most well known, defined by the United States Bureau of the Census as "the method of reference to a standard national distribution." Gonzalez (1974) has offered a more comprehensive explanation - "An unbiased estimate is obtained from a sample survey for a large area; when this estimate is used to derive estimates for subareas on the assumption that the small areas have the same characteristics as the larger area, we identify these estimates as synthetic estimates." Developed at the National Center for Health Statistics, the method was initially used to provide synthetic state estimates of disability from the results of the National Health Interview Survey (H.I.S.).

Procedurally, a number of demographic variables are selected (i.e., race, income, sex, age), and when possible, national sample surveys are used to determine estimates of a characteristic (criterion variable) of interest for each of the G mutually exclusive and exhaustive domains defined by the respective demographic cross-classifications. To produce the synthetic estimate of a criterion variable (Y) for local area ℓ, the NCHS model takes the form of a weighted average

$$Y_\ell^* = \sum_{j=1}^{G} P_{\ell j}\, Y_{\cdot j} \qquad (1.1)$$

where $P_{\ell j}$ is the proportion of local area ℓ's population represented by domain j so that $\sum_j^G P_{\ell j} = 1$, and $Y_{\cdot j}$ is the probability estimate of the criterion variable for domain j obtained from a national sample. The more detailed estimating equation includes a regional adjustment.

Considering the underlying model's structure, the synthetic estimates are biased. A popular measure used to assess their reliability is the average mean squared error (M.S.E.)

$$E[1/N \sum_{i=1}^{N} (Y_\ell^* - Y_\ell)^2]$$

calculated over all N local areas defined by the survey population. Gonzalez and Waksburg (1973) have derived an approximation for this expression, assuming that

i) the $P_{\ell j}$'s are fixed and measured without error, and
ii) the $\mathrm{Cov}(Y_{\cdot j}, Y_{\cdot k}) = 0$ for $j \neq k$.

Due to the nature of their derivation, the synthetic estimates will generally cluster near the mean for a specific geographic region. Consequently, the method is not particularly sensitive to many of the internal forces operating at the local level. By assuming the small areas share the same characteristics as a standard national distribution, they can only be distinguished by their respective demographic configurations. Recognizing this inherent limitation, Levy (1971) proposed a method which utilized available information at the local level on predictor (symptomatic) variables in conjunction with the NCHS estimator. The following model was considered:

$$Y_\ell^{**} = \alpha + \beta X_\ell + \varepsilon_\ell \qquad (1.2)$$

where X_ℓ is the value of the symptomatic variable for the ℓ^{th} subarea,

$$Y_\ell^{**} = (Y_\ell - Y_\ell^*)/Y_\ell^* \times 100$$

where ε_ℓ is a term representing random error, and α and β, regression coefficients to be estimated. Here, the percentage difference between the synthetic estimate and the true value is treated as a linear function of some related predictor variable X_ℓ. Were the estimates $\hat{\alpha}$ and $\hat{\beta}$ available and ε_ℓ omitted, an estimator $\hat{Y}_\ell$ of Y_ℓ could be derived from (1.2), taking the form:

$$\hat{Y}_\ell = Y_\ell^*[(\hat{\alpha} + \hat{\beta} X_\ell)/100 + 1] \qquad (1.3)$$

It is assumed that X_ℓ is available for every local area, but since Y_ℓ^{**} is a function of the true value Y_ℓ (which is unknown), a different strategy is used to estimate the linear coefficients. Briefly, α and β are estimated by least squares after combining local areas to form strata. The method can be extended to consider X_ℓ as a vector of symptomatic data, whereby $\hat{Y}_\ell$ is treated as a multiple regression estimator.

Ericksen (1974) developed another technique

for computing local area estimates which, unlike the NCHS estimator, solely combines symptomatic information and sample data into a multiple regression format (assuming an underlying linear model). Referred to as the regression-sample data of local area estimation, the procedure can be outlined as follows:

1. Initially, a sample of n local areas, referred to as primary sampling units (PSU's), is selected from the N local areas in the population. Estimates of the criterion variable are then computed for the respective PSU's in the sample.
2. Collect symptomatic information for both sample and non-sample PSU's. Typical predictor variables are the number of births, deaths, and school enrollment.
3. Compute the linear least squares regression estimate using data for the sample PSU's only. Estimates for all subareas are then determined by substituting values of the symptomatic indicators, whether included in the respective sample or not.

The model assumes the availability of criterion variable estimates for each of n sample PSU's and the values of p symptomatic indicators for the universe of N local areas. It takes the matrix representation:

$$Y = XB + u \qquad (1.4)$$

where Y, an $n\times1$ vector, is the criterion variable consisting of a set of actual unobserved values; X, an $n\times(p+1)$ matrix denoting the set of predictor variables;

B, the $(p+1)\times1$ vector of regression coefficients; and

u, an $n\times1$ vector, a stochastic error term.

Under the assumption of linearity, B could be estimated by ordinary least squares regression were the Y values observed. Because the individual observations of Y are affected by sampling variability, the model may be revised to explain the within-PSU sampling error in the following manner:

$$Y_0 = XB + u + v \qquad (1.5)$$

where v is an $n\times1$ vector of sampling error deviations and Y_0 the observed values.

The regression equation is then computed, substituing the observed values of Y_0 for Y. Hence, the regression coefficients are unbiased in the absence of correlations between v and Y. The mean square error of the regression estimates is expressed as:

$$E(Y-\hat{Y})'(Y-\hat{Y})/n = [(n-p-1)\sigma_u^2/n] + [(p+1)\sigma_v^2/n] \qquad (1.6)$$

where σ_u^2 is the between-PSU variance unexplained by the predictor variables, and σ_v^2 is the within PSU variance.

This method was tested for counties and states using 1970 census data on population growth. The resulting estimates were found to be more accurate than estimates computed by standard demographic procedures for the same period.

2. An Alternative Strategy

2.1 Methodology

The method advanced by Ericksen is most feasible when the linearity assumption is satisfied and the observed multiple correlation is high. But what decision is reached when the multiple correlation level is moderate (.5-.8) and a non-linear model is more suitable? The inclusion of all possible symptomatic variables into the regression would increase the R^2 but most probably at the expense of an "over-fit" model which increases the mean square error of the final estimate. More generally, in those situations where assumptions are too strict or unrealistic, the need for a more flexible approach is most obvious. Kalsbeek (1973) has developed one such procedure in which the most limiting assumption is the availability of good symptomatic information.

It has usually been common practice to treat the local area units as the smallest level for which the estimates are made. Contrarily, Kalsbeek suggests breaking up the local unit into constituent geographical sectors called "base units," such as townships, enumeration districts, or other geographical subunits of a county. The local area for which a variable of interest is to be estimated is referred to as the "target area" and further subdivided into "target area base units." Unlike other methods which use symptomatic information directly for the purposes of estimation, this procedure uses the information to group base units (sample base units) from the total population. The symptomatic information is also used to classify "target area base units" into the appropriate group.

Initially, a random sample of n base units is selected from the total population of N base units. The sample base units (possibly including some "target area base units") are required to possess both symptomatic and criterion information. These units are divided into K groups (strata) using either or both types of the information available. The object is to form groups which are most homogeneous within while dissimilar between themselves. Grouping can be handled be any one of several iterative procedures in cluster analysis (i.e., Automatic Interaction Detection (A.I.D.), Multivariate Iterative K-Means Cluster Analysis (MIKCA)). It is noteworthy that the respective groups may be defined by either rectilinear or non-rectilinear boundaries.

All "target area base units" belonging to the local area in question are then assigned (classified) to one of the K groups with respect to symptomatic information. Consequently, each "target area base unit" is associated with a group of base units both similar to itself and internally homogeneous. An estimate for each of the "target area base units" with respect to the criterion variable is obtained from the sample base units in the group to which it has been assigned. These estimates are then pooled to arrive at a final estimate for the respective target area.

2.2 Notation

Consider a population consisting of L local areas, indexed by ℓ=1, 2, ..., L, which have further been subdivided into constituent geographical sectors called "base units." There are $_{\ell}N_{.}$ base units in the ℓ^{th} local area, and

$$\sum_{\ell=1}^{L} {}_{\ell}N_{.} = {}_{..}N_{.}$$

in the population, individually indexed by i=1, 2, ..., $_{\ell}N_{.}$, to denote the i^{th} base unit from the ℓ^{th} local area. When the local area reference is dropped, each base unit is indexed by i=1, 2, ..., $_{..}N_{.}$. Furthermore, each base unit i consists of a cluster of M_i smaller units referred to as elements. Hence, there are

$$_{\ell}M_{.} = \sum_{i=1}^{{}_{\ell}N_{.}} M_i$$

elements in the ℓ^{th} local area and

$$_{..}M_{.} = \sum_{\ell=1}^{L} {}_{\ell}M_{.} = \sum_{i=1}^{{}_{..}N_{.}} M_i$$

elements in the population. Let y_{ij} represent the observed value of the criterion variable for the j^{th} element within the i^{th} base unit, where

$$Y_i = \sum_{j=1}^{M_i} y_{ij}$$

is the i^{th} unit total.

In practice, a multi-stage sampling design is most appropriate. To facilitate the presentation, we assume a two-stage sampling design whereby a simple random sample of n base units (first stage units) is initially drawn from the $_{..}N_{.}$ base units in the population. A subsample of m_i out of the M_i elements is then selected with equal probabilities of selection from each of the chosen sample base units. Here, the subunits are chosen independently in different units. The units are then divided into K groups (strata), indexed by g=1, 2, ..., K, by one of the aforementioned procedures (Section 2.1). Consequently, estimates of the group means are obtained by a method which most closely resembles post-stratification. To determine the criterion variable estimator for the ℓ^{th} local area, each "target base unit" is assigned to the group most similar with respect to symptomatic information. Thus, we have a two-way classification of all base units in the population by respective strata and local areas, where $_{\ell}N_g$ is the total number of base units in the g^{th} strata from the ℓ^{th} local area.

2.3 Representation of the Model

The local area estimator of the criterion variable may be expressed in terms of an average, a proportion, or a total. Initially, we direct attention to the mean per element representation.

Assuming a two-stage sampling design with sub-units of unequal sizes, we define

$$\bar{y}_i = \sum_{j=1}^{m_i} \frac{y_{ij}}{m_i}$$

as the sample mean per element in the i^{th} base unit and

$$\bar{Y}_i = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i}$$

as the overall mean per element in the i^{th} base unit. To obtain an estimate of the g^{th} stratum mean per element, we also define the indicator variables I_{gi} (once more dropping the local area reference), such that

I_{gi} = 1 if the (first stage) base unit falls in the g^{th} stratum;

= 0, otherwise

for g=1, 2, ..., K and i=1, 2, ..., $_{..}N_{.}$. Here, $\sum_{i=1}^{n} I_{gi} = n_g$, the number of sample base units belonging to the g^{th} stratum, and

$$\sum_{i=1}^{{}_{..}N_{.}} I_{gi} = {}_{.}N_g .$$

Consequently, let

$$\hat{\bar{y}}_g = \frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} = \frac{\sum^{n_g} M_i \bar{y}_i}{\sum^{n_g} M_i}$$

(summed only over the n_g sample base units from the g^{th} stratum) be our (post-stratified) estimator of the g^{th} stratum mean per element. To facilitate the presentation, we assume the values of M_i in the sample are known. Since $\hat{\bar{y}}_g$ is a ratio estimator of

$$\bar{\bar{Y}}_g = \frac{\sum_{i=1}^{{}_{..}N_{.}} I_{gi} M_i \bar{Y}_i}{\sum_{i=1}^{{}_{..}N_{.}} I_{gi} M_i} = \frac{\sum^{{}_{.}N_g} M_i \bar{Y}_i}{{}_{.}M_g}$$

(where the sum is over the $_{.}N_g$ base units assigned to the g^{th} stratum), it is biased to the order of 1/n. Yet, when n is large (i.e., n $\geq$ 100), the bias is negligible and the expectation of $\hat{\bar{y}}_g$ is approximately equivalent to $\bar{\bar{Y}}_g$,

$$E(\hat{\bar{y}}_g) \doteq \bar{\bar{Y}}_g \quad , \quad g = 1, 2, \ldots, K \quad .$$

Returning to the ℓ^{th} local area, we focus attention on the "target base unit" alignment in order to weight appropriately the stratum estimators $(\hat{\bar{y}}_g)$ by the proportion of base units so classified. Therefore, the estimator of the criterion variable for the ℓ^{th} local area takes the following form:

$$_{\ell}\hat{\bar{y}}^{*} = \sum_{g=1}^{K} \frac{{}_{\ell}M_g}{{}_{\ell}M_{.}} \hat{\bar{y}}_g \qquad (2.3.1)$$

such that

$$E(\hat{\bar{y}}^{*}_{\ell}) = \sum_{g=1}^{K} \frac{\ell M_g}{\ell M_.} E(\hat{\bar{y}}_g) \doteq \sum_{g=1}^{K} \frac{\ell M_g}{\ell M_.} \bar{\bar{Y}}_g$$

when n is large. Often the sizes of ℓM_g and $\ell M_.$ are only known approximately. When this occurs, the respective estimators of the strata means are weighted by the ratio of available estimates $\ell \hat{M}_g$ and $\ell \hat{M}_.$, or by the cruder ratio $\ell N_g / \ell N_.$.

Due to the nature of its derivation, the local area estimator $\hat{\bar{y}}^{*}_{\ell}$ of $\bar{\bar{Y}}_{\ell}$ is biased. The observed value of the criterion variable mean per element is

$$\bar{\bar{Y}}_{\ell} = \frac{\sum_{\ell N_.} M_i \bar{Y}_i}{\sum_{\ell N_.} M_i} = \frac{\sum_{\ell N_.} M_i \bar{Y}_i}{\ell M_.}$$

summed across only those base units in the ℓ^{th} local area. The bias,

$$B = [E(\hat{\bar{Y}}^{*}_{\ell}) - \bar{\bar{Y}}_{\ell}]$$

can be approximated by

$$B' = [\sum_{g=1}^{K} \frac{\ell M_g}{\ell M_.} \bar{\bar{Y}}_g - \frac{\sum_{\ell N_.} M_i \bar{Y}_i}{\ell M_.}]$$

Similarly, to express the local area estimator in terms of a proportion, y_{ij} is redefined, so that

y_{ij} = 1 when the j^{th} element in the i^{th} base unit has the characteristic of interest;

= 0 otherwise,

so that

$$\sum_{j=1}^{M_i} y_{ij} = Y_i$$

is the total number of elements in the i^{th} base unit with the characteristic of interest. Model (2.3.1) can then be used.

2.4 An Expression for the Mean Squared Error of the Local Area Estimator

It has already been observed that the local area estimator $\hat{\bar{y}}^{*}_{\ell}$ is biased. Consequently, the mean squared error term takes the form:

$$E[(\hat{\bar{y}}^{*}_{\ell} - \bar{\bar{Y}}_{\ell})^2] = E(\hat{\bar{y}}^{*}_{\ell} - E(\hat{\bar{y}}^{*}_{\ell}))^2 + (E(\hat{\bar{y}}^{*}_{\ell}) - \bar{\bar{Y}}_{\ell})^2$$

$$= \mathrm{Var}(\hat{\bar{y}}^{*}_{\ell}) + \mathrm{Bias}^2 \quad . \qquad (2.4.1)$$

By assuming

$$E(\hat{\bar{y}}^{*}_{\ell}) \doteq \sum_{g=1}^{K} \frac{\ell M_g}{\ell M_.} \bar{\bar{Y}}_g \quad ,$$

where $\hat{\bar{y}}^{*}_{\ell}$ is a linear combination of the ratio estimators $\hat{\bar{y}}_g$, g=1, 2, ..., K with neglible bias, the variance of $\hat{\bar{y}}^{*}_{\ell}$ can be approximated by

$$\mathrm{Var}(\hat{\bar{y}}^{*}_{\ell}) =$$

$$\sum_{g=1}^{K} (\frac{\ell M_g}{\ell M_.})^2 \mathrm{Var}(\hat{\bar{y}}_g) + \sum_{g \neq g'} \sum (\frac{\ell M_g}{\ell M_.})(\frac{\ell M_{g'}}{\ell M_.}) \mathrm{Cov}(\hat{\bar{y}}_g, \hat{\bar{y}}_{g'})$$

If we also assume

$$\frac{\sum_{i=1}^{n} I_{gi} M_i \bar{y}_i}{\sum_{i=1}^{n} I_{gi} M_i} - \bar{\bar{Y}}_g \doteq \frac{\sum_{i=1}^{n} I_{gi} M_i (\bar{y}_i - \bar{\bar{Y}}_g)}{n(\frac{M_g}{N_.})} \quad ,$$

then

$$\mathrm{Var}(\hat{\bar{y}}_g) \doteq \frac{(N_. - n)}{n N_.} (\frac{N_.^2}{M_g^2}) \frac{\sum_{i=1}^{N_.} I_{gi}^2 M_i^2 (\bar{Y}_i - \bar{\bar{Y}}_g)^2}{(N_. - 1)}$$

$$+ \frac{N_.}{n M_g^2} \frac{\sum_{i=1}^{N_.} I_{gi}^2 M_i^2}{m_i} (1 - \frac{m_i}{M_i}) \frac{\sum_{j=1}^{M_i} (y_{ij} - \bar{Y}_i)^2}{(M_i - 1)}$$

This is the standard form of the approximate variance of ratio estimator for a two-stage sampling design where the base units have equal probabilities of selection. Here, the first term represents the between base unit component of the variance, whereas the second denotes the within-base unit contribution. A nearly unbiased sample estimate of $\mathrm{Var}(\hat{\bar{y}}_g)$ takes the form:

$$\mathrm{var}(\hat{\bar{y}}_g) = (\frac{N_. - n}{N_. n})(\frac{N^2}{M_g^2}) \frac{\sum_{i=1}^{n} I_{gi}^2 M_i^2 (y_i - \hat{\bar{y}}_g)^2}{(n-1)}$$

$$+ \frac{N_.}{n M_g^2} \frac{\sum_{i=1}^{n} I_{gi}^2 M_i^2}{m_i} (1 - \frac{m_i}{M_i}) \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{(m_i - 1)}$$

Since our sampling design requires the independent selection of subsamples from different sample base units, and the respective strata estimators are defined in terms of the indicator variable I_{gi}, it can also be shown that $\mathrm{Cov}(\hat{\bar{y}}_g, \hat{\bar{y}}_{g'}) \doteq 0$. Hence, the mean squared error of our small area estimator can be expressed as:

$$\mathrm{MSE}(\hat{\bar{y}}^{*}_{\ell}) \doteq \sum_{g=1}^{k} (\frac{\ell M_g}{M_g})^2 \mathrm{Var}(\hat{\bar{y}}_g) + (\mathrm{Bias})^2$$

3. An Illustrative Example

The availability of Census data on population and per capita income for 1970 allowed for an examination of the method's accuracy. State estimates of population growth (from 1960-1970) and per capita income were generated by Kalsbeek, using the Current Population Survey (a national multi-stage probability sample of the U.S. con-

ducted monthly) as the source of sample information. Here, the sample base units correspond to the first stage primary sampling units (PSU's) in the C.P.S., which are counties or groups of contiguous counties. The symptomatic variables considered when estimating population growth include total school enrollment, live births, and deaths, all expressed in ratio form (1970 total/1960 total). Those considered in the per capita income example include the percent natural increase in population between 1960 and 1965, the 1960 per capita aggregate income, and the 1964 percent of the population on public assistance (all obtained from the 1967 County-City Data Book).

The grouping of the sample base units (PSU's) was done using the Automatic Interaction Detector, version II (AID II), which is essentially a clustering algorithm that uses both symptomatic and criterion variable information. Since the respective groups (strata) formed have rectilinear boundaries, the "target base units" (here counties) are assigned to the group whose boundaries include the observation's symptomatic values. Hence, each target base unit takes on the group estimate of the criterion variable to which it is assigned. For the ℓ^{th} state, one considers the respective target base unit alignment, and weights the group (strata) estimators $(\hat{\bar{y}}_g)$ by the proportion of the state's population in the target base units so classified. Here, the 1960 county populations were used.

The method was compared with Ericksen's procedure since both are applicable under essentially the same circumstances. The criterion for measuring the accuracy of the estimates was the relative absolute deviation from the true value: $\left|\frac{\text{Estimated-True}}{\text{True}}\right|$. Ericksen's procedure would be expected to give better results for the population growth example due to the inclusion of three symptomatic variables with a high level of multiple correlation and an underlying linear relationship. Still, the proposed method yielded more accurate estimates for more than 25 percent of the states considered (11 out of 42). It was observed that the results tended to improve with increases in population size for both methods. The proposed method did much better in generating state estimates of per capita income, yielding more accurate results in 29 of the 47 states considered. In general, the proposed method produced better results with moderate per capita income states, while Ericksen's approach was more successful at the extremes.

4. Summary

Reliable estimates at the local level are generally difficult, if not impossible, to obtain from sample surveys, primarily due to the constraints of sample size and design. Yet, the very nature of the problem has served as the motivating force in the development of several alternative procedures. The strategy suggested here offers a quick though not consistently clean method of generating the desired estimates. Here, a trade-off exists between the considerations of cost and accuracy. Generally, one is willing to sacrifice a degree of exactness when confronted with the harsh realities of limited resources. This strategy is particularly attractive in that no particular functional model between the criterion and symptomatic variables must be specified. Estimates for the base units of the "target areas" are available as a by-product of the technique. Finally, the method performs reasonably well even for a linear setting, though here it would be better to choose Ericksen's approach.

ACKNOWLEDGEMENTS

The research was partially supported by the National Institute of Child Health and Human Development (Grant HD-00371) and the U.S. Bureau of the Census (JSA-76-93). The authors wish to thank Gary G. Koch for his discussions of the paper and Jean McKinney for her conscientious typing of this manuscript.

REFERENCES

Cochran, W. G. Sampling Techniques. New York: John Wiley and Sons, 1963.

Ericksen, E. P. "A Regression Method for Estimating Population Changes of Local Areas." Journal of the American Statistical Association, 69 (1974), 867-875.

Gonzalez, M. E. and Waksberg, J. "Estimation of the Error of Synthetic Estimates." Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria, 1973.

Gonzalez, M. E. "Use and Evaluation of Synthetic Estimates." American Statistical Association, Proceedings of the Social Statistics Section (1975).

Gonzalez, M. E. and Hoza, C. "Small Area Estimation of Unemployment." American Statistical Association, Proceedings of the Social Statistics Section (1975).

Kalsbeek, W. D. "A Method for Obtaining Local Postcensal Estimates for Several Types of Variables." Unpublished doctoral dissertation, University of Michigan (1973).

Koch, G. G. "An Alternative Approach to Multivariate Response Error Models for Sample Survey Data with Applications to Estimators Involving Subclass Means." Journal of the American Statistical Association, 68 (1973), 328-331.

Levy, P. S. "The Use of Mortality Data in Evaluating Synthetic Estimates." American Statistical Association, Proceedings of the Social Statistics Section (1971), 328-331.

U. S. Bureau of the Census. "Federal State Cooperative Program for Local Population Estimates: Test Results - April 1, 1970." Current Population Reports, P-26, No. 21, 1973.

U. S. Bureau of the Census. "Coverage of the

Population in the 1970 Census and Some Implications for Public Programs." Current Population Reports, P-23, No. 56, 1975.

U. S. National Center for Health Statistics. Synthetic Estimates of Disability, PHS publication, No. 1759, 1968.

ALTERNATIVE ESTIMATORS OF VARIANCE FOR SYSTEMATIC SAMPLING

Kirk M. Wolter and Shana McCann
U.S. Bureau of the Census

1. Introduction

This paper is concerned with the problem of estimating the variance of the sample mean, say $\bar{y}_{sy}$, when the sample is drawn systematically from a finite population of size N. We shall only consider equal probability systematic sampling with a single, random start. Unequal probability systematic sampling or sampling with two or more random starts will not be treated here.

In the 1940's several authors addressed the issue of variance estimation for systematic samples, including Osborne (1942), Cochran (1946), Matérn (1947), and Yates (1949). One of the most comprehensive discussions is given by Cochran (1963). A more recent reference is Koop (1971). Little in the way of empirical comparisons of alternative estimators is available in this literature. In recent years, the topic appears to have received little attention, no doubt because systematic sampling is often used at the last stage of sampling, a case where rigorous estimates of the variance can be given. However, there remain many surveys where an estimate of $\mathrm{Var}\{\bar{y}_{sy}\}$ is required. In such cases we have noticed a tendency on the part of many researchers to regard the sample as random, and, in the absence of knowing what else to do, to estimate the variance using random sample formulae. This practice often leads to badly biased estimates of variance, and to incorrect inferences concerning the population mean.

In the remainder of this paper we shall empirically investigate eight estimators of the variance of $\bar{y}_{sy}$. Our goal is to provide some guidance about when a given estimator may be more appropriate than other estimators. The estimators are defined in Section 2. In Section 3, the various populations used in our study are described. The results of the comparison are then summarized in Sections 4 and 5.

2. Description of the Estimators

Throughout our investigation we assume N=nk where n is the sample size and k is an integer. We let Y_{ij} denote the value of the y-variable for the j-th unit in the i-th systematic sample, where i=1, ...,k and j=1,...,n. Then, the eight estimators of variance for the i-th selected sample are defined as follows:

1. $v_{sy1}(i) = \frac{N-n}{Nn} \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{sy})^2$.

2. $v_{sy2}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-1} (y_{ij} - y_{i,j+1})^2/2(n-1)$.

3. $v_{sy3}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n/2} (y_{i,2j-1} - y_{i,2j})^2/n$.

4. $v_{sy4}(i) = \frac{N-n}{N} \frac{n'}{n^2} \sum_{j=1}^{n-2} (y_{ij} - 2y_{i,j+1} + y_{i,j+2})^2/6(n-2)$,

where $\frac{n'}{n^2} = \left(\frac{1}{n} + \frac{2i-k-1}{2(n-1)k}\right)^2 + \frac{n-2}{n^2} + \left(\frac{1}{n} - \frac{2i-k-1}{2(n-1)k}\right)^2$.

5. $v_{sy5}(i) = \frac{1}{4}(\bar{y}_A - \bar{y}_B)^2$, where $\bar{y}_A$ is the is the mean of the even numbered members of the sample and $\bar{y}_B$ is the mean of the odd numbered members.

6. $v_{sy6}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-4} c_{ij}^2/3.5(n-4)$, where

$$c_{ij} = \frac{1}{2}y_{ij} - y_{i,j+1} + y_{i,j+2} - y_{i,j+3} + \frac{1}{2}y_{i,j+4} .$$

7. $v_{sy7}(i) = \frac{N-n}{Nn} \sum_{j=1}^{n-8} d_{ij}^2/7.5(n-8)$, where

$$d_{ij} = \frac{1}{2}y_{ij} - y_{i,j+1} + y_{i,j+2} - y_{i,j+3} + y_{i,j+4} - y_{i,j+5} + y_{i,j+6} - y_{i,j+7} + \frac{1}{2}y_{i,j+8} .$$

8. $v_{sy8}(i) = \frac{N-n}{Nn} s^2 \left\{1 + \frac{2}{\ln\hat{\rho}_k} + \frac{2}{\left(\hat{\rho}_k^{-1}-1\right)}\right\}$,

$= \frac{N-n}{Nn} s^2$ if $\hat{\rho}_k > 0$
if $\hat{\rho}_k \leq 0$,

where $\hat{\rho}_k = \sum_{j=1}^{n-1} (y_{ij} - \bar{y}_{sy})(y_{i,j+1} - \bar{y}_{sy}) / (n-1)s^2$.

$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{sy})^2$.

v_{sy1} is the estimate of variance for simple random sampling. v_{sy2} and v_{sy3} are based on overlapping and nonoverlapping differences, respectively. v_{sy4}, v_{sy6}, and v_{sy7} are based on higher order contrasts. Koop's (1971) estimator, v_{sy5}, is obtained by splitting the systematic sample into equal halves. v_{sy8} was devised from an assumption about the correlogram (cf. Cochran (1946)).

3. Description of the Populations

3.1 The Artificial Populations

Sixteen artificial populations, each of size N=1000, were generated according to the simple model

$$Y_{ij} = \mu_{ij} + u_{ij}, \qquad (2.1)$$

where the μ_{ij} denote fixed constants and the errors u_{ij} are drawn from some infinite superpopulation. The reader will recognize (2.1) as the model employed by Cochran (1963). The eight estimators were evaluated using the sixteen populations and four sampling fractions: $f=k^{-1}$=.01,.02,.1 and .25. Due to limited space, only seven populations and two sampling fractions, f=.01 and .02, will be discussed here.

The seven populations for which results will be presented and the specific assumptions about the μ_{ij} and u_{ij} are described in Table 1. For notational convenience, we shall employ the

Table 1. Description of the Artificial Populations

Code	Description	μ_{ij}	u_{ij}
A1	Random	0	u_{ij} iid $\Gamma(2,11.32)$
A2	Linear Trend	$i+(j-1)k$	u_{ij} iid N(0,2.25)
A3	Stratification Effects	$\mu_{.j}$	u_{ij} iid N(0,9)
A4	Stratification Effects	$\mu_{.j}$	u_{ij} iid N(0,9)
A5	First Order Autocorrelated	0	$u_{ij}=\rho u_{i-1,j}+e_{ij}$; $u_{11}\sim N\left(0,\frac{55.43}{(1-\rho^2)}\right)$; e_{ij} iid N(0,55.43); ρ = .9
A6	First Order Autocorrelated	0	$u_{ij}=\rho u_{i-1,j}+e_{ij}$; $u_{11}\sim N\left(0,\frac{190.84}{(1-\rho^2)}\right)$; e_{ij} iid N(0,190.84); ρ = .5
A7	Periodic	$2\sin\frac{\pi i}{2}$	u_{ij} iid N(0,.07)

population codes in future references to these populations. Population A3 was only used with the f=.01 sampling fraction and the $\mu_{.j}$'s took the values 8, 42, 70, 90, 99, 96, 81, 57, 24, and 8. Similarly, population A4 was only used with the f=.02 sampling fraction and the $\mu_{.j}$'s took the values 0, 17, 34, 50, 64, 76, 86, 94, 98, 100, 98, 94, 86, 76, 64, 50, 34, 17, 0, and 17. Each of the remaining populations was studied for both f=.01 and .02.

Of the 9 populations for which results are not being presented, three were random, three had a linear trend, two had stratification effects, and one was autocorrelated.

3.2 The Real Populations

The estimators of variance were also compared on the basis of six real populations obtained from Census Bureau files. The first two populations, R1 and R2, were comprised of 6900 fuel oil dealers from the 1972 Economic Census. The y-variable was annual sales in both cases. R1 was sorted by multi- versus single-unit firms, by State, and by ID number. The nature of the ID number was such that within a given class of firms within a given State, the sort was essentially random. R2 was sorted by annual payroll.

The remaining four populations were from the Income Supplement to the March, 1975 Current Population Survey (CPS). A one-in-five sample of persons in the civilian labor force and living in SMSA's of 250,000 population or more was the basis for these populations. For R3 and R4 the y-variable was the unemployment indicator

$$y = \begin{cases} 1, & \text{if unemployed} \\ 0, & \text{if employed} \end{cases}$$

while in R5 and R6 the y-variable was total income. R3 and R5 were in sort by two census tract characteristics: "non-whites as a percent of the total population" and "persons with four or more years of high school as a percent of all persons 25 years old or older." R4 and R6 were in sort by the census tract characteristic "median family income." Populations R3, R4, R5, and R6, were each of size N=11300.

4. Empirical Results

Some of the results of our investigation are presented in Tables 2, 3, and 4. Tables 2 and 3 give the relative biases and relative mean square errors (MSE) of the eight estimators, respectively. Table 4 presents the actual proportion of confidence intervals which contained the true population mean, where the confidence interval for the α-th estimator is of the form

$$\left(\bar{y}_{sy} - t_{n-1,.025}v_{sy\alpha}(i)\ ,\ \bar{y}_{sy} + t_{n-1,.025}v_{sy\alpha}(i)\right)$$

and $t_{n-1,.025}$ denotes the .025 percentage point of Student's t distribution with n-1 degrees of freedom. As noted in Section 3, the populations and sampling fractions reported in the tables comprise less than half of those actually studied. When describing the results, however, our remarks shall apply to all of the study populations, not merely the illustrative ones.

An important observation regarding these results is that our sample of populations is far too small to conclusively demonstrate estimator behavior. As a result, we shall not try to claim too much from our results. Our remarks will be limited to instances where, in our view, a reasonably consistent pattern of behavior was established.

Many additional commentaries could be given beyond those presented in this section. For example, one may wish to observe certain patterns of bias depending on the value of the intraclass correlation coefficient. However, such analyses will be left to the reader as our space is limited.

Table 2. Relative Bias of Eight Estimators of $V\{\bar{y}_{sy}\}$

Population Code	Sampling Fraction f	Estimator of Variance								Intraclass Correlation
		v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}	
R1	0.01	0.239	-0.416	-0.991	-0.587	-0.719	-0.582	-0.726	-0.705	-0.00292
R1	0.02	0.713	-0.321	-0.984	-0.513	0.792	-0.564	-0.552	-0.696	-0.00311
R2	0.01	0.505	0.218	-0.257	0.146	0.042	-0.034	-0.182	-0.253	-0.00429
R2	0.02	0.313	0.155	-0.328	0.142	-0.105	0.096	-0.065	-0.403	-0.00184
R3	0.01	-0.094	-0.121	0.122	-0.123	-0.218	-0.129	-0.138	-0.376	0.00082
R3	0.02	-0.146	-0.152	0.100	-0.144	-0.096	-0.134	-0.122	-0.388	0.00065
R4	0.01	0.106	0.111	0.109	0.114	0.269	0.105	0.086	-0.167	-0.00093
R4	0.02	0.114	0.101	0.140	0.096	0.035	0.084	0.073	-0.196	-0.00053
R5	0.01	0.381	0.065	-0.461	-0.007	0.245	-0.057	-0.113	-0.404	-0.00251
R5	0.02	0.349	0.073	-0.453	0.016	-0.023	-0.064	-0.123	-0.540	-0.00121
R6	0.01	-0.068	-0.069	-0.109	-0.063	-0.032	-0.072	-0.076	-0.307	0.00055
R6	0.02	-0.041	-0.048	-0.076	-0.050	0.078	-0.052	-0.040	-0.329	0.00010
A1	0.01	0.022	0.056	0.012	0.099	0.101	0.169	0.230	-0.204	-0.00317
A1	0.02	-0.068	-0.036	-0.008	-0.017	-0.147	-0.034	-0.090	-0.190	0.00255
A2	0.01	9.901	-0.405	-0.404	-1.000	2.008	-1.000	-1.000	-0.353	-0.10001
A2	0.02	19.652	-0.705	-0.705	-1.000	2.010	-1.000	-1.000	-0.441	-0.05001
A3	0.01	133.928	27.361	26.435	2.310	1.613	2.804	2.869	11.446	-0.11021
A4	0.02	146.639	9.616	9.824	1.401	4.627	0.787	0.712	3.875	-0.05226
A5	0.01	0.866	0.762	0.896	0.816	0.982	0.865	0.408	0.216	-0.04926
A5	0.02	1.011	0.532	0.318	0.360	0.226	0.126	-0.022	-0.235	-0.02361
A6	0.01	0.118	0.137	0.107	0.195	0.184	0.181	0.011	-0.144	-0.01161
A6	0.02	0.222	0.166	0.138	0.140	0.211	0.088	0.020	-0.229	-0.01000
A7	0.01	-0.996	-0.996	-0.996	-0.996	-0.996	-0.996	-0.997	-0.997	0.96502
A7	0.02	32.668	61.868	61.891	83.143	631.178	140.949	262.807	32.444	-0.05102

Table 3. Relative Mean Square Error (MSE) of Eight Estimators of $V\{\bar{y}_{sy}\}$

Population	Sampling Fraction f	v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}
R1	0.01	4.349	2.002	0.982	1.482	1.235	1.638	0.988	1.109
R1	0.02	4.123	1.435	0.969	1.094	12.722	1.080	1.160	0.809
R2	0.01	3.598	2.988	3.004	2.764	5.162	2.086	1.865	1.960
R2	0.02	1.391	1.298	1.389	1.364	1.245	1.431	1.131	0.823
R3	0.01	0.078	0.082	0.185	0.088	1.528	0.105	0.144	0.242
R3	0.02	0.053	0.054	0.078	0.055	1.596	0.061	0.076	0.215
R4	0.01	0.092	0.102	0.228	0.112	2.246	0.132	0.195	0.169
R4	0.02	0.055	0.070	0.161	0.079	2.842	0.099	0.132	0.201
R5	0.01	0.557	0.206	0.496	0.160	3.235	0.184	0.265	0.494
R5	0.02	0.275	0.137	0.343	0.115	1.963	0.103	0.119	0.398
R6	0.01	0.212	0.241	0.395	0.274	1.677	0.343	0.376	0.358
R6	0.02	0.138	0.143	0.166	0.142	2.146	0.149	0.213	0.235
A1	0.01	0.561	0.736	0.830	1.046	3.037	2.603	4.133	0.601
A1	0.02	0.238	0.339	0.394	0.449	1.624	0.684	1.183	0.343
A2	0.01	98.034	0.164	0.164	1.000	4.033	1.000	1.000	0.125
A2	0.02	386.213	0.497	0.497	1.000	4.043	1.000	1.000	0.194
A3	0.01	17990.670	751.789	724.542	6.066	9.285	8.959	11.640	131.725
A4	0.02	21532.965	93.282	99.893	2.476	45.284	1.266	1.942	15.102
A5	0.01	1.391	1.365	2.127	1.892	4.266	3.000	2.118	0.721
A5	0.02	1.359	0.593	0.398	0.501	1.418	0.454	0.582	0.393
A6	0.01	0.303	0.460	0.578	0.711	2.344	1.275	2.036	0.397
A6	0.02	0.209	0.299	0.250	0.380	3.596	0.563	1.190	0.413
A7	0.01	0.993	0.993	0.993	0.992	0.992	0.993	0.993	0.994
A7	0.02	2132.668	7678.381	7693.850	13882.034	799450.766	39904.320	138758.197	2132.783

Table 4. Proportion of Times that the True Population Mean Fell within the Confidence Interval formed Using One of Eight Estimators of Variance

Population Code	Sampling Fraction f	Estimator of Variance								$V\{\bar{y}_{sy}\}$
		v_{sy1}	v_{sy2}	v_{sy3}	v_{sy4}	v_{sy5}	v_{sy6}	v_{sy7}	v_{sy8}	
R1	0.01	0.99	0.75	0.19	0.71	0.47	0.67	0.56	0.59	$6.135 \cdot 10^{3}$
R1	0.02	1.00	0.84	0.18	0.84	0.82	0.80	0.78	0.64	$2.197 \cdot 10^{3}$
R2	0.01	0.91	0.88	0.79	0.87	0.65	0.86	0.84	0.74	$5.423 \cdot 10^{3}$
R2	0.02	0.90	0.86	0.80	0.86	0.68	0.86	0.84	0.76	$2.862 \cdot 10^{3}$
R3	0.01	0.93	0.93	0.92	0.93	0.58	0.93	0.93	0.83	$7.8 \cdot 10^{-4}$
R3	0.02	0.90	0.90	0.98	0.90	0.64	0.92	0.90	0.84	$4.1 \cdot 10^{-4}$
R4	0.01	0.96	0.96	0.89	0.96	0.69	0.95	0.94	0.92	$6.4 \cdot 10^{-4}$
R4	0.02	0.92	0.92	0.92	0.92	0.74	0.94	0.94	0.82	$3.1 \cdot 10^{-4}$
R5	0.01	0.97	0.97	0.83	0.96	0.73	0.96	0.93	0.83	$4.247 \cdot 10^{5}$
R5	0.02	0.88	0.88	0.82	0.86	0.74	0.86	0.84	0.76	$2.148 \cdot 10^{5}$
R6	0.01	0.92	0.92	0.89	0.92	0.74	0.91	0.90	0.85	$6.275 \cdot 10^{5}$
R6	0.02	0.92	0.90	0.90	0.90	0.70	0.90	0.92	0.84	$3.020 \cdot 10^{5}$
A1	0.01	0.97	0.94	0.91	0.94	0.74	0.89	0.76	0.87	$2.435 \cdot 10^{1}$
A1	0.02	0.92	0.90	0.88	0.88	0.62	0.82	0.74	0.82	$1.314 \cdot 10^{1}$
A2	0.01	1.00	1.00	1.00	0.02	1.00	0.01	0.00	1.00	$8.323 \cdot 10^{2}$
A2	0.02	1.00	0.66	0.66	0.02	1.00	0.02	0.02	0.90	$2.075 \cdot 10^{2}$
A3	0.01	1.00	1.00	1.00	1.00	0.89	1.00	1.00	1.00	$9.328 \cdot 10^{-1}$
A4	0.02	1.00	1.00	1.00	1.00	0.94	0.98	0.98	1.00	$4.071 \cdot 10^{-1}$
A5	0.01	0.99	0.97	0.97	0.98	0.90	0.94	0.80	0.93	$1.393 \cdot 10^{1}$
A5	0.02	1.00	0.98	1.00	0.96	0.84	0.88	0.86	0.84	$6.259 \cdot 10^{0}$
A6	0.01	0.95	0.93	0.90	0.93	0.71	0.89	0.73	0.90	$2.251 \cdot 10^{1}$
A6	0.02	0.96	0.92	0.94	0.94	0.70	0.92	0.90	0.86	$1.018 \cdot 10^{1}$
A7	0.01	0.49	0.48	0.45	0.48	0.36	0.43	0.38	0.45	$2.005 \cdot 10^{0}$
A7	0.02	1.00	1.00	0.96	0.96	0.92	0.94	0.94	0.96	$3.17 \cdot 10^{-3}$

4.1 Random Populations (A1)

Estimators v_{sy1}, v_{sy2}, v_{sy3}, and v_{sy4} were comparable and each displayed acceptable properties. v_{sy5} had a larger bias than $v_{sy1},\ldots,v_{sy4}$; its mean square error was extremely large; and it led to unacceptable confidence intervals. v_{sy8} tended to have the largest bias, but one of the smaller MSE's. v_{sy8} also produced slightly low confidence levels. v_{sy6} and v_{sy7} behaved similarly, with larger mean square error than $v_{sy1},\ldots,v_{sy4}$ and similar confidence levels to v_{sy8}.

4.2 Populations with Linear Trend (A2)

Remarkably, estimator v_{sy8} always produced the smallest bias, the smallest MSE, and the best confidence intervals (in the sense that confidence levels were nearest to 95 percent). v_{sy2} and v_{sy3} were comparable, producing lower confidence levels and larger bias and MSE than v_{sy8}. Estimators v_{sy1} and v_{sy5} had particularly bad properties.

4.3 Populations with Stratification Effects (A3,A4)

Estimators v_{sy2}, v_{sy3}, and particularly v_{sy1} were consistently bad both in terms of bias and MSE. This was undoubtedly due to our construction of the populations, with very large differences between successive values of $\mu_{\cdot j}$. The behavior of v_{sy5} and v_{sy8} was not firmly established with both estimators displaying relatively large and small biases for various populations. Estimators v_{sy4}, v_{sy6}, and v_{sy7} were comparable, usually having smaller bias and MSE than the other estimators. Confidence levels tended to be too high for all estimators except v_{sy5}, and all estimators tended to overestimate the true variance.

4.4 Autocorrelated Populations (A5, A6)

The results for autocorrelated populations depended largely on the value of ρ, i.e. the first order autocorrelation coefficient, and on the sampling fraction. For small to moderate values of ρ, the estimators behaved as they did for the random populations. For large ρ, v_{sy8} tended to have the smallest absolute bias and by far the smallest MSE. There was very little to choose between v_{sy1}, v_{sy2}, v_{sy3}, and v_{sy4} for large ρ: their MSE's were about twice that of v_{sy8} and likewise their biases. Confidence levels for intervals formed from v_{sy1}, v_{sy2}, v_{sy3}, v_{sy4} were very near to the nominal level of 95 percent, however. The differences between the estimators seemed to decrease as the sampling fraction increased.

4.5 Periodic Populations (A7)

As one would expect of periodic populations, the behavior of the estimators depended exclusively on the sampling fraction. However, for all sampling fractions studied, the eight estimators possessed nearly identical bias and MSE. In one case v_{sy1} and v_{sy8} had smaller MSE than the other estimators, but the MSE's were so large that it would make little practical difference which estimator was used.

4.6 Fuel Oil Dealers Sales (R1, R2)

Sort by Multi- versus Single-Unit by State and by ID Number: Estimator v_{sy1} overestimated the true variance, while the remaining estimators possessed a negative bias. v_{sy1} or v_{sy2} had the

smallest absolute bias, and v_{sy3} and v_{sy5} the largest. v_{sy3}, v_{sy7}, and v_{sy8} tended to have the smallest MSE. v_{sy4} and v_{sy6} had MSE's which were comparable to those of v_{sy3}, v_{sy7}, and v_{sy8} for large sampling fractions. The MSE's of the remaining estimators were larger. In spite of its small MSE, v_{sy3} led to extremely poor confidence intervals owing to its large bias. Except for v_{sy1}, whose MSE was too large, each of the estimators led to lower confidence levels than the anticipated 95 percent. v_{sy2}, v_{sy4}, and v_{sy6} seemed to give the best confidence intervals.

Sort by Annual Payroll: Estimator v_{sy6} tended to have the smallest absolute bias; v_{sy4}, v_{sy5}, and v_{sy7} also had relatively small absolute bias, but larger than v_{sy6}. The MSE's of v_{sy6}, v_{sy7} and v_{sy8} tended to be smaller than those of the other estimators. In particular, v_{sy5} had a very large MSE when f=.01. Among those estimators with relatively small bias, v_{sy4}, v_{sy6}, and v_{sy7} produced good confidence intervals, though the coverage rate was lower than expected. The population in this sort seemed to follow the linear model

$$Y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{ij} \text{ , where } E\{u_{ij}|x_{ij}\} = 0,$$

$E\{u_{ij}^2|x_{ij}\} = \sigma^2 x_{ij}^g$, $g \in [\frac{1}{2}, 2]$, and x_{ij} denotes the annual payroll of the (i,j)-th unit.

4.7 CPS Unemployment (R3, R4)

Sort by % Nonwhite Etc. of Census Tract: The absolute biases of v_{sy1}, v_{sy2}, v_{sy3}, v_{sy4}, v_{sy6}, and v_{sy7} were comparable and relatively small, usually less than around 15%. The bias of v_{sy5} was also small when f=.02, but for f=.01 it exceeded 20%. The absolute bias of v_{sy8} was larger. Most estimators tended to underestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. v_{sy1}, v_{sy2}, and v_{sy4} had the smallest MSE's, closely followed by v_{sy3}, v_{sy6}, and v_{sy7}. The MSE's of v_{sy8} and particularly v_{sy5} were larger. Most estimators led to acceptable confidence intervals except v_{sy5} and v_{sy8}, where the confidence levels were very low and slightly low, respectively.

Sort by Median Family Income of Census Tract: Estimator v_{sy8} tended to have larger bias than the other estimators. Also the bias of v_{sy8} was negative, while all other estimators tended to overestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. v_{sy1}, v_{sy2}, v_{sy6} and v_{sy4} tended to have the smallest MSE, followed by v_{sy8}, v_{sy7}, and v_{sy3}. The MSE of v_{sy5} was much larger. All estimators produced acceptable confidence intervals, with the exception of v_{sy5} whose confidence level was too low.

4.8 CPS Income (R5, R6)

Sort by % Nonwhite Etc. of Census Tract: v_{sy2}, v_{sy4}, v_{sy6}, and v_{sy7} tended to have the smallest bias, though v_{sy5} also had small bias when f=.02. v_{sy1}, v_{sy3}, and v_{sy8} had larger biases. The MSE's of v_{sy2}, v_{sy4}, v_{sy6}, and v_{sy7} were comparable and relatively small. v_{sy1}, v_{sy3}, v_{sy8}, and particularly v_{sy5} had larger MSE's. Confidence intervals formed from v_{sy3}, v_{sy5}, and v_{sy8} had low coverage rates.

Sort by Median Family Income of Census Tract: Most of the estimators tended to underestimate $\mathrm{Var}\{\bar{y}_{sy}\}$. The biases of all estimators were of the same order of magnitude except v_{sy8}, which was larger. v_{sy1}, v_{sy2}, and v_{sy4} tended to have the smallest MSE. v_{sy3}, v_{sy6}, v_{sy7}, and v_{sy8} also displayed consistently small MSE, while the MSE of v_{sy5} was much larger. Each of the estimators except v_{sy5} gave acceptable confidence intervals, though the confidence levels were lower than 95 percent.

5. Detailed Analysis of Populations With Linear Trend

One of the interesting aspects of the results in Section 4 was the performance of the estimator v_{sy8}. In a variety of circumstances this estimator had relatively small bias and MSE and gave useable confidence intervals. This was particularly true of the populations with linear trend, even though v_{sy8} was constructed for another purpose (i.e. autocorrelated populations). This led us to question whether the behavior observed was a unique attribute of the particular populations studied, or was a more general result characteristic of all populations with linear trend. A partial answer to this question can be provided by obtaining the expected bias of each estimator of variance.

Towards this end, we assume the finite population is generated according to (2.1), with

$$\mu_{ij} = \beta_0 + \beta_1(i + (j-1)k)$$

and

$$u_{ij} \text{ iid } (0, \sigma^2).$$

If we let E denote the expectation with respect to the superpopulation, then the expected bias and expected relative bias of the α-th estimator are defined by

$$B\{v_{sy\alpha}\} = E\,E\{v_{sy\alpha}\} - EV\{\bar{y}_{sy}\}$$

and

$$R\{v_{sy\alpha}\} = B\{v_{sy\alpha}\} \,/\, EV\{\bar{y}_{sy}\} \ ,$$

respectively. It can then be shown that

$$R\{v_{sy1}\} = \frac{\beta_1^2(N-1)}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.1)$$

$$R\{v_{sy2}\} = \frac{\beta_1^2(6k-N-n)/n}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.2)$$

$$R\{v_{sy3}\} = R\{v_{sy2}\} \ , \qquad (5.3)$$

$$R\{v_{sy4}\} = \frac{-\beta_1^2(k+1)}{\beta_1^2(k+1) + 12\sigma^2/N} \ , \qquad (5.4)$$

$$R\{v_{sy5}\} = \frac{\beta_1^2(2k^2+1)}{\beta_1^2(k^2-1) + 12(k-1)\sigma^2/N} \ , \qquad (5.5)$$

$$R\{v_{sy6}\} = R\{v_{sy4}\} \ , \qquad (5.6)$$

$$R\{v_{sy7}\} = R\{v_{sy4}\} \ , \qquad (5.7)$$

and

$$R\{v_{sy8}\} \doteq \left\{\beta_1^2 k(n+1)\left[1 + \frac{2}{\ln\gamma(1)/\gamma(0)} + \frac{2}{\gamma(0)/\gamma(1)-1}\right] - \beta_1^{2(k+1)}\right\} \Big/ \left\{\beta_1^2(k+1) + 12\sigma^2/N\right\} \quad , \quad (5.8)$$

where

$$\gamma(1) = \beta_1^2 k^2 (n-3)(n+1)/12 - \sigma^2/n$$

$$\gamma(0) = \beta_1^2 k^2 n(n+1)/12 + \sigma^2 \quad .$$

The expression for $R\{v_{sy8}\}$ was derived by approximating the expectation, E, of the function v_{sy8} $(s^2, \hat{\rho}_k s^2)$ by the same function of the expectations $E\{s^2\}$ and $E\{\hat{\rho}_k s^2\}$, where we have used an expanded notation for v_{sy8}. In deriving this result it was also assumed that $\hat{\rho}_k > 0$ with probability one. This assumption is quite modest and guarantees that terms involving the operator ln (•) are well defined.

From (5.1),...,(5.8), it can be seen that the value of the intercept, β_0, has no effect on the relative biases, while the error variance has only slight impact since terms in σ^2 are of lower order than the remaining terms. Similarly, the value of β_1 has little effect on the relative bias, unless β_1 is extraordinarily small. Note that $R\{v_{sy1}\},\ldots,R\{v_{sy7}\}$ converge to zero as $\beta_1 \to 0$. Thus, v_{sy1} through v_{sy7} are unbiased when the population is random. As $\beta_1 \to 0$, the assumption that $\Pr\{\hat{\rho}_k > 0\} = 1$ will not hold, and the expression for $R\{v_{sy8}\}$ in (5.8) will not be valid. For large populations where β_1 is not extremely close to 0, (5.1), (5.4), (5.5), (5.6), and (5.7) suggest the following useful approximations:

$$R\{v_{sy1}\} \doteq n$$
$$R\{v_{sy4}\} \doteq -1$$
$$R\{v_{sy5}\} \doteq 2$$
$$R\{v_{sy6}\} \doteq -1$$
$$R\{v_{sy7}\} \doteq -1 \quad .$$

We have also derived expressions for the relative biases under the more general assumption that the u_{ij} are mutually independent with zero mean and heterogeneous variance σ_{ij}^2. The observations made in the previous paragraph also apply to this model.

The results for population A2 in Table 2 agree well with the expressions for the expected relative biases. For example, letting N=1000, n=10, k=100, σ=1.5, β_1=1, and β_0=0, we find that equations (5.1),...,(5.8) take the values 9.888, -0.406, -0.406, -1.000, 2.000, -1.000, -1.000, and -0.355, respectively.

As further confirmation of the expressions for the relative biases, 100 populations of size N=1000 were generated according to the superpopulation model for each of the following values of $(\beta_0, \beta_1, \sigma)$: (0,.5,1.5), (0,1,1.5), (0,2,1.5), and (0,1,5). The bias, MSE, and significance level (associated with confidence intervals which used the multiplier $t_{n-1,.025}$) of each estimator of variance was then found for both f=.01 and .02 for each population. To illustrate, the results for the case $(\beta_0, \beta_1, \sigma)$=(0,1,5) with f=.01 are summarized in Table 5. The second and third columns give the quantities

$$\left(\Sigma(E\{v_{sy\alpha}\} - V\{\bar{y}_{sy}\})/100\right) \Big/ \left(\Sigma\, V\{\bar{y}_{sy}\}/100\right)$$

and

$$\left(\Sigma\{E(v_{sy\alpha} - V\{\bar{y}_{sy}\})^2\}/100\right) \Big/ \left(\Sigma(V\{\bar{y}_{sy}\})^2/100\right)$$

respectively, where α=1,...,8 and the summations are taken over the 100 populations. The fourth column gives the average significance level for each estimator. Note that the bias results agree well with the expressions in (5.1),...,(5.8). Clearly, the only estimators with acceptable properties are v_{sy2}, v_{sy3}, and v_{sy8}: the remaining estimators have either large MSE or lead to unacceptably low confidence levels. And among these estimators, v_{sy8} has the smallest bias and MSE. The results for the other values of $(\beta_0, \beta_1, \sigma)$ are similar.

Table 5. Monte Carlo Estimates of Expected Bias, Expected MSE, and Expected Confidence Levels

Estimator	Expected Relative Bias	Expected Relative MSE	Expected Confidence Level
v_{sy1}	9.856	97.168	100.00
v_{sy2}	-0.405	0.164	99.11
v_{sy3}	-0.405	0.164	99.06
v_{sy4}	-0.997	0.994	6.62
v_{sy5}	1.993	4.007	100.00
v_{sy6}	-0.997	0.994	6.14
v_{sy7}	-0.997	0.994	5.54
v_{sy8}	-0.355	0.126	99.94

It would be hazardous at this point for the reader to draw very general conclusions about the eight estimators, since the investigation in this section assumed a very specific model which may not be obtained in practice. In the future, we will be investigating models with a higher order polynomial trend and other alternative specifications. Our continuing goal in this work will be to establish conditions under which the various estimators have acceptable properties.

References

[1] Cochran, W.G. "Relative Accuracy of Systematic and Stratified Random Samples for a Certain Class of Populations" Ann. Math. Statist. 17 (1946): 164-177.

[2] Cochran, W.G. Sampling Techniques. New York: John Wiley and Sons, 1963.

[3] Koop, J.C. "On Splitting a Systematic Sample for Variance Estimation." Ann. Math. Statist. 42 (1971): 1084-1087.

[4] Matérn, B. "Methods of Estimating the Accuracy of Line and Sample Plot Surveys." Medd. fr. Statens. Skogsforskningsinstitut 36 (1947); 1-138

[5] Osborne, J.G. "Sampling Errors of Systematic and Random Surveys of Cover-Type Areas." J. Amer. Statist. Assoc. 37 (June 1942): 256-264.

[6] Yates, F. Sampling Methods for Censuses and Surveys . London: Griffin, 1949.

RELATIVE EFFICIENCY OF SOME TWO-FRAME ESTIMATORS

H. Huang, Minnesota State Department of Education

1. Introduction

In sample surveys, a complete frame is often unavailable or too expensive to construct. When these situations arise, a survey practitioner may use multiple frames. One of the first applications of the multiple frame procedure appeared in the "Sample Survey of Retail Stores" conducted by the United States Bureau of the Census in 1949, reported by Bershad [1]. Hartley [5]gave a complete description of multiple frame concepts. Cochran [2,3], Lund [7], and others have also considered the problem.

Fuller and Burmeister [4] proposed some alternative estimators. In this study, agricultural data is used to illustrate their multiple regression estimators for population totals. The relative efficiencies of these estimators to Hartley's estimator are presented.

2. Notation and Estimators for Population Totals

We assume that two frames, A and B, containing N_A and N_B elements respectively, are available. We denote by N_{ab} the number of elements included in both frame A and frame B, by N_a the number of elements occurring only in frame A, and by N_b the number of elements occurring only in frame B. Thus

$$N_A = N_a + N_{ab} ,$$

$$N_B = N_b + N_{ab}$$

and the total number of elements in the population is given by

$$N = N_a + N_b + N_{ab} = N_a + N_B = N_b + N_A .$$

We refer to the elements contained only in Frame A as domain a, the elements only in frame B as domain b and those elements in both frames A and B as domain ab. Domain ab is sometimes called the overlap domain.

Given that simple random samples of size n_A and n_B are selected from frame A and frame B, respectively, Hartley [5] proposed the following estimator of the population total for the characteristic, Y:

$$\hat{Y}_H = \hat{Y}_a + \hat{Y}_B + P(\hat{Y}'_{ab} - \hat{Y}''_{ab}) \qquad (2.1)$$

where

$$\hat{Y}_B = \hat{Y}_b + \hat{Y}''_{ab}$$

$\hat{Y}_a$ is the estimator of the total of Y for domain a obtained from the sample from frame A,

$\hat{Y}'_{ab}$ is the estimator of the total of Y for domain ab obtained from the sample from frame A,

$\hat{Y}_b$ is the estimator of the total of Y for domain b obtained from the sample from frame B,

$\hat{Y}''_{ab}$ is the estimator of the total of Y for domain ab obtained from the sample from frame B, and

P is the number chosen to minimize the variance of the estimator.

Fuller and Burmeister [4] suggested the estimator:

$$\tilde{Y}_r = \hat{Y}_a + \hat{Y}_B + b_1 (\hat{N}'_{ab} - \hat{N}''_{ab}) + b_2 (\hat{Y}'_{ab} - \hat{Y}''_{ab}) , \qquad (2.2)$$

where

$\hat{N}'_{ab}$ is an estimator of the number of elements in domain ab estimated from the sample from frame A,

$\hat{N}''_{ab}$ is an estimator of the number of elements in domain ab estimated from the sample from frame B,

b_1 and b_2 are numbers chosen to minimize the variance of the estimator.

The estimators $\hat{N}'_{ab} - \hat{N}''_{ab}$ and $\hat{Y}'_{ab} - \hat{Y}''_{ab}$ are unbiased estimators of zero. Both $\tilde{Y}_r$ and $\hat{Y}_H$ are recognizable as multiple regression estimators. Therefore, Hartley's estimator, $\hat{Y}_H$, is inefficient relative to the Fuller-Burmeister estimator $\tilde{Y}_r$, if the partial correlation between $\hat{Y}_a + \hat{Y}_B$ and $\hat{N}'_{ab} - \hat{N}''_{ab}$, after adjusting for $\hat{Y}'_{ab} - \hat{Y}''_{ab}$, is not zero.

In our application of the theory frame A is a stratified list frame and frame B is a complete area frame. The sample elements selected from the area frame can be identified as belonging or not belonging to the list frame A. The Hartley estimator remains the same for a stratified list, but the Fuller-Burmeister estimators can be extended to include additional unbiased estimators of zero. We define

$$\tilde{Y}_{mR} = \hat{Y}_B + \sum_{i=1}^{L} b_{1i} (\hat{Y}'_{iab} - \hat{Y}''_{iab}) + \sum_{j=1}^{m} b_{2j} (\hat{N}'_{jab} - \hat{N}''_{jab}) , \qquad (2.3)$$

where

$\hat{N}'_{jab}$ is an estimator of the number of elements in domain ab of the jth subgroup

obtained from the sample of frame A,

$\hat{N}''_{jab}$ is an estimator of the number of elements in domain ab of the j^{th} subgroup obtained from the sample of frame B,

$\hat{Y}'_{iab}$ is an estimator of the total of Y for domain ab of the i^{th} stratum obtained from the sample of frame A,

$\hat{Y}''_{iab}$ is an estimator of the total of Y for domain ab of the i^{th} stratum obtained from the sample of frame B,

L is the total number of strata,

and

m is the number of subgroups on which the estimator of the number of elements in domain ab are obtained and included in the estimator.

We note that $\hat{N}'_{jab} - \hat{N}''_{jab}$ may be an estimator of zero obtained from a particular stratum or from a combination of several strata. We also define n_{Ai}, i = 1, L, as the size of sample selected from the i^{th} stratum of frame A.

When frame B is a complete area frame, the variance of $\hat{Y}_H$ and $\tilde{Y}_r$ are given as follows:

$$V(\hat{Y}_H) = V(\hat{Y}_B) - \frac{[\mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab})]^2}{V(\hat{Y}'_{ab}) + V(\hat{Y}''_{ab})} , \qquad (2.4)$$

$$V(\tilde{Y}_r) = V(\hat{Y}_B) - b_1 \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{ab}) - b_2 \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab}) \qquad (2.5)$$

where

$$\begin{pmatrix} b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} V(\hat{N}''_{ab}) & \mathrm{Cov}(\hat{N}''_{ab}, \hat{Y}''_{ab}) \\ \mathrm{Cov}(\hat{N}''_{ab}, \hat{Y}''_{ab}) & V(\hat{Y}'_{ab}) + V(\hat{Y}''_{ab}) \end{bmatrix}^{-1} \begin{bmatrix} \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{ab}) \\ \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{ab}) \end{bmatrix} .$$

To obtain the variance of $\tilde{Y}_{mR}$, we write (2.3) as

$$\tilde{Y}_{mR} = \hat{Y}_B + \hat{X}\hat{\beta} \qquad (2.6)$$

where

$$\hat{\beta}' = (b_{11}, b_{12}, \ldots, b_{1L}, b_{21}, b_{22}, \ldots, b_{2m})$$

$$\hat{X} = \hat{X}_1 - \hat{X}_2 = (\hat{Y}'_{1ab} - \hat{Y}''_{1ab}, \hat{Y}'_{2ab} - \hat{Y}''_{2ab}, \ldots, \hat{Y}'_{Lab} - \hat{Y}''_{Lab}, \hat{N}'_{1ab} - \hat{N}''_{1ab}, \ldots, \hat{N}'_{mab} - \hat{N}''_{mab}) .$$

Then

$$V(\tilde{Y}_{mR}) = V(\hat{Y}_B) - \hat{\beta}' \mathrm{Cov}(\hat{Y}_B, \hat{X}_2) \qquad (2.7)$$

where

$$\hat{\beta} = V^{-1} \mathrm{Cov}(\hat{Y}_B, \hat{X}_2),$$

$$\mathrm{Cov}(\hat{Y}_B, \hat{X}_2) = (\mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{1ab}), \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{2ab}), \ldots, \mathrm{Cov}(\hat{Y}_B, \hat{Y}''_{Lab}), \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{1ab}), \ldots, \mathrm{Cov}(\hat{Y}_B, \hat{N}''_{mab}))' ,$$

and V is the covariance matrix of $\hat{X}$.

3. Application of Two-Frame Estimators to California Fruit Data

3.1. Description of the frames

Some data on fruit collected by USDA in California in 1972 are used to illustrate the relative efficiency of the Fuller-Burmeister estimator to Hartley's estimator. These data represent a complete listing of acreages of certain fruits organized on an area basis. The basic unit is an area segment. The area segments are grouped into clusters to form an area frame of 187 area clusters. Some of the clusters contain no acreage in fruit.

A "list frame" of area segments was constructed using the list of segments. This list was constructed to simulate the type of list that might be constructed using producer lists. Such lists traditionally contain a larger fraction of the large operators. Therefore the list frame contained 95% of the segments with area over 500 acres devoted to fruits, 60% of the segments having fruit acreage greater than or equal to 100 acres but less than 500 acres, and 28% of the segments having some fruit acreage but less than 100 acres. The list frame created in this manner contained a total of 310 segments, representing 50% of the non-zero area segments.

Two characteristics, the number of acres under fruit and the number of fruit trees (in hundreds), are studied.

3.2. Simple Random Sampling From List Frame

In the first study, we assume selection of simple random samples of segments from the list frame (frame A) and of clusters from the area frame (frame B). Variances of the estimated totals of the two characteristics for various sample sizes were computed both with and without the finite population correction (fpc) for both frames. The variances were computed using the optimal values of p for Hartley's estimator and

optimal values of b_1 and b_2 for the Fuller-Burmeister estimator.

The percentage gain in efficiency of the Fuller-Burmeister estimator, $\tilde{Y}_r$, relative to the Hartley estimator, $\hat{Y}_H$, is defined by $100[V(\hat{Y}_H) - V(\tilde{Y}_r)]/V(\tilde{Y}_r)$. The results for selected sample sizes with fpc, are given in Table 1. Substantial gains are evident for most sample combinations. The gain increases as the fraction of the sample selected from the area frame increases.

The procedure used in the 1949 'Sample Survey of Retail Stores' consisted of observing only that portion of the area frame that fell in the non-overlap domain. If a screening process is applied and the data on that portion of the area frame sample elements belonging to the overlap domain not collected, then the Hartley estimator reduces to

$$\hat{Y}_c = \hat{Y}'_{ab} + \hat{Y}_b \quad . \tag{3.1}$$

The Fuller-Burmeister estimator for this particular situation is

$$\hat{Y}_{cr} = \hat{Y}'_{ab} + \hat{Y}_b + \hat{\beta}_c(N_{ab} - \hat{N}''_{ab}) \quad . \tag{3.2}$$

The gains in efficiency from using $\hat{Y}_{cr}$, rather than $\hat{Y}_c$, for the set of sample sizes given in Table 1 were computed. The largest gain was 26% associated with a list sample size of 60 and area sample size of 10. For a fixed sample size selected from the list frame, the gain decreases as the size of the sample selected from the area frame increases. This is also apparent from the efficiency gain formula,

$$\frac{V(\hat{Y}_c) - V(\hat{Y}_{cr})}{V(\hat{Y}_{cr})} = \frac{[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2}{V(\hat{N}''_{ab})} \cdot \left[V(\hat{Y}'_{ab}) + V(\hat{Y}_b) - \frac{[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2}{V(\hat{N}''_{ab})} \right]^{-1} \tag{3.3}$$

Since $[\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2 [V(\hat{N}''_{ab})]^{-1}$ and $V(\hat{Y}_b) - [\mathrm{Cov}(\hat{Y}_b, \hat{N}''_{ab})]^2 [V(\hat{N}''_{ab})]^{-1}$ are multiples of n_B^{-1}, the ratio must decrease as n_B increases.

3.3. Stratified Sampling From the List Frame

To investigate efficiencies for stratified sampling of the list frame, we divided the list frame into three strata on the basis of our original construction of the frame. The three strata were sampled in the ratio 4:2:1.

Three forms of Fuller-Burmeister estimators, $\tilde{Y}_{mR}$, were considered. They are

$$\tilde{Y}_{1R} = \hat{Y}_B + b_{11}(\hat{N}'_{ab} - \hat{N}''_{ab}) + b_{12}(\hat{Y}'_{ab} - \hat{Y}''_{ab}) \tag{3.4}$$

$$\tilde{Y}_{2R} = \hat{Y}_B + b_{21}(\hat{N}'_{ab} - \hat{N}''_{ab}) + b_{22}(\hat{Y}'_{1ab} - \hat{Y}''_{1ab}) + b_{23}(\hat{Y}'_{2ab} - \hat{Y}''_{2ab}) + b_{24}(\hat{Y}'_{3ab} - \hat{Y}''_{3ab}) \tag{3.5}$$

$$\tilde{Y}_{3R} = \hat{Y}_B + b_{31}(\hat{N}'_{1ab} - \hat{N}''_{1ab}) + b_{32}(\hat{N}'_{2ab} - \hat{N}''_{2ab}) + b_{33}(\hat{N}'_{3ab} - \hat{N}''_{3ab}) + b_{34}(\hat{Y}'_{1ab} - \hat{Y}''_{1ab}) + b_{35}(\hat{Y}'_{2ab} - \hat{Y}''_{2ab}) + b_{36}(\hat{Y}'_{3ab} - \hat{Y}''_{3ab}) \tag{3.6}$$

where $\hat{Y}_B$, $\hat{Y}'_{iab}$, $\hat{Y}''_{iab}$, $\hat{N}'_{ab}$, and $\hat{N}''_{ab}$ are previously defined, while $\hat{N}'_{iab}$ and $\hat{N}''_{iab}$ are the estimators of the number of elements in domain ab of the i^{th} stratum obtained from the sample of frame A and frame B respectively.

The optimal p's of the Hartley estimator and the optimal b's of Fuller-Burmeister estimators for various sample sizes and the associated variances of the estimators, $V(\hat{Y}_H)$, $V(\hat{Y}_{1R})$, $V(\hat{Y}_{2R})$, and $V(\hat{Y}_{3R})$ were computed retaining the finite population correction. The gains in efficiency from using $\tilde{Y}_{1R}$, $\tilde{Y}_{2R}$, and $\tilde{Y}_{3R}$ relative to Hartley's estimator, $\hat{Y}_H$, are shown in Tables 2-4.

The gains from including additional estimators of zero in the estimator for the total are substantial. As before the gain increases as the area sample size increases.

A summary of the efficiency of $\tilde{Y}_r$ in simple random sampling, and $\tilde{Y}_{3R}$ in stratified sampling, relative to the Hartley estimator is presented in Table 5.

3.4. Optimum Allocation

For any given cost structure, we can obtain the gain in efficiency under optimum allocation among the two frames for each estimator. We now assume the cost for each unit in the area sample is six times as great as that for a unit in the list sample. We study optimal allocation only for the data of acreage in fruit. In simple random sampling, ignoring the finite population correction terms, the optimum allocation for the Hartley estimator is specified by the ratio $n_A/n_B = 4.34$. For the Fuller-Burmeister estimator the optimal ratio is $n_A/n_B = 3.12$. The gain in

efficiency of the Fuller-Burmeister procedure relative to the Hartley procedure given optimum allocation for each procedure is 13.64%.

We now investigate the behavior of these estimators under the optimum allocation among the strata. We assume the cost of a unit in one stratum is the same as that of a unit in other strata. Using the iteration procedure, we found that, for $\hat{Y}_H$, the optimum stratum allocation is 49:45:6 and the optimum frame sample ratio is $n_A/n_B = 2.18$, while, for $\tilde{Y}_{3R}$, the optimum stratum allocation is 62:37:1 and the optimum frame sample ratio is $n_A/n_B = 0.79$. Under these best conditions for each estimator, the gain in efficiency from using $\tilde{Y}_{3R}$ relative to $\hat{Y}_H$ is 19.26%.

By comparing the gains in efficiency under the best conditions for each estimator with the data in Table 4, we can see that the relative efficiency of the Hartley estimator is slightly better under optimum sample allocation than under nonoptimum allocation. That is, as we improve the efficiency with which we select the sample, the potential for reduction in variance associated with the inclusion of estimators of zero is reduced.

4. Summary

The variances of alternative multiple-frame estimators are compared using data collected in a census of fruit trees in California in 1972.

In one comparison, we assumed the selection of a simple random sample of individual segments from the list frame and of clusters of segments from the area frame. The gain in efficiency of the Fuller-Burmeister estimator relative to the Hartley estimator was a function of the relative rates at which the two frames were sampled. The gain in efficiency increases as the sampling rate in the area frame increases. In a second comparison the optimum sampling procedure for a fixed budget was used for each estimator under reasonable cost assumptions, the gain of the Fuller-Burmeister estimator relative to the Hartley estimator is about fourteen percent.

The efficiency of the Fuller-Burmeister estimators were also investigated for stratified sampling. When stratified sampling is used, there are a number of estimators of zero that can be used in the regression estimator. The regression estimators displayed considerable gains in efficiency when several estimators of zero were used. As in simple random sampling, the gain in efficiency from using the Fuller-Burmeister estimators is largest for samples where the ratio of the size of the list sample to the size of the area sample size is small. When the optimum sample allocation is used for each estimator, the gain is about nineteen percent.

REFERENCES

[1] Bershad, M. A., "The Sample of Retail Stores," in Hansen, Hurwitz, and Madow, Sample Survey Methods and Theory, Vol. I. Wiley (1953), 516-558.

[2] Cochran, R. S., "Multiple Frame Sample Surveys." Proceedings of the Social Statistics Section of the American Statistical Association (1964), 16-19.

[3] _____, "The Estimation of Domain Sizes When Sampling Frames are Interlocking." Proceedings of the Social Statistics Section of the American Statistical Association (1967), 332-335.

[4] Fuller, W. A. and Burmeister, L. F., "Estimators for Samples Selected from Two Overlapping Frames." Research Report for the Bureau of the Census, Iowa State University, Ames, Iowa (1973).

[5] Hartley, H. O., "Multiple Frame Surveys." Proceedings of the Social Statistics Section of the American Statistical Association (1962), 203-206.

[6] Huang, H. T., "The Relative Efficiency of Some Two-Frame Estimators." A report for the Statistical Reporting Service, USDA, Iowa State University, Ames, Iowa (1974).

[7] Lund, R. E., "Estimators in Multiple Frame Surveys." Proceedings of the Social Statistics Section of the American Statistical Association (1968), 282-288.

Table 1. Percentage Gain in Efficiency of the Fuller-Burmeister Estimator ($\tilde{Y}_r$) relative to the Hartley Estimator ($\hat{Y}_H$) for Various Sample Sizes for California Fruit Data.

List frame sample size (n_A)	Area frame sample size (n_B)				
	10	15	20	25	30
Acres in Fruit					
20	25.30	38.67	51.87	64.77	77.33
30	16.18	25.14	34.36	43.71	53.12
40	11.63	18.11	24.96	32.08	39.40
50	8.95	13.87	19.18	24.78	30.63
60	7.21	11.07	15.29	19.81	24.59
No. of trees					
20	7.35	12.69	17.90	22.90	27.69
30	3.75	7.29	10.97	14.69	18.39
40	2.05	4.50	7.22	10.06	12.98
50	1.12	2.87	4.92	7.15	9.48
60	0.60	1.84	3.41	5.17	7.07

Table 2. Percentage Gain in Efficiency of $\hat{Y}_{1R}$ Relative to the $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	6.07	9.41	13.10	17.08	21.32
16	8	4	4.51	6.82	9.41	12.25	15.33
20	10	5	3.62	5.33	7.27	9.41	11.76
32	16	8	2.39	3.25	4.24	5.35	6.58
40	20	10	2.01	2.62	3.32	4.10	4.97
No. of trees							
12	6	3	2.86	5.86	9.06	12.35	15.67
16	8	4	1.50	3.55	5.88	8.38	10.96
20	10	5	0.79	2.22	3.98	5.91	7.98
32	16	8	0.07	0.54	1.31	2.29	3.41
40	20	10	0.00	0.17	0.60	1.22	1.98

Table 3. Percentage Gain in Efficiency of $\tilde{Y}_{2R}$ Relative to $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	12.07	20.19	28.80	37.65	46.60
16	8	4	8.30	14.05	20.42	27.19	34.23
20	10	5	6.18	10.45	15.34	20.68	26.35
32	16	8	3.41	5.45	7.98	10.90	14.16
40	20	10	2.66	4.01	5.74	7.81	10.17
No. of trees							
12	6	3	24.87	35.16	43.84	51.20	57.50
16	8	4	19.51	28.48	36.52	43.66	50.02
20	10	5	16.00	23.85	31.21	38.01	44.24
32	16	8	10.45	15.93	21.58	27.23	32.78
40	20	10	8.61	13.06	17.87	22.86	27.93

Table 4. Percentage Gain in Efficiency of $\hat{Y}_{3R}$ Relative to the Hartley Estimator $(\hat{Y}_H)$ for Stratified List Sampling.

List frame stratum sample size			Area frame sample size (n_B)				
n_{A1}	n_{A2}	n_{A3}	10	15	20	25	30
Acres in Fruit							
12	6	3	15.07	26.50	39.08	52.48	66.57
16	8	4	9.89	17.74	26.68	36.43	46.86
20	10	5	7.07	12.75	19.44	26.88	34.96
32	16	8	3.56	6.11	9.38	13.23	17.58
40	20	10	2.70	4.30	6.48	9.14	12.22
No. of trees							
12	6	3	32.33	41.26	48.81	55.46	61.50
16	8	4	28.04	36.14	43.11	49.27	54.87
20	10	5	25.19	32.68	39.27	45.15	50.51
32	16	8	20.49	26.78	32.71	38.25	43.44
40	20	10	18.85	24.60	30.28	35.76	41.02

Table 5. Efficiency of Fuller-Burmeister Estimator Relative to the Hartley Estimator.

n_A/n_B	Acres in fruit		No. of trees	
	Simple random	Stratified	Simple random	Stratified
6.0	107	103	101	120
5.0	109	104	101	121
4.0	111	106	102	124
3.0	116	109	104	128
2.0	125	115	107	132
1.3	139	129	113	141
1.0	152	139	118	149
0.8	165	152	123	155
0.7	177	167	128	162

EXACT CONFIDENCE INTERVALS FOR FINITE POPULATION PARAMETERS IN TWO-STAGE SAMPLING

Richard K. Burdick, Arizona State University
Robert L. Sielken, Jr., Texas A&M University[1]

INTRODUCTION

The linear least-squares prediction approach has recently been applied by Royall [1976] in two-stage sampling from finite populations. Royall develops alternative estimators and their variances for the finite population total and compares them under various situations. This paper considers a special case of the super-population model assumed by Royall and discusses a technique for the unbiased estimation of variance and construction of an exact confidence interval on the finite population total.

THE MODEL FOR TWO-STAGE SAMPLING

A finite population of K elements is separated into N clusters of size M_i where $\sum_{i=1}^{N} M_i = K$. Letting y_{ij} denote the value associated with the j^{th} element in cluster i, the model describing the super-population from which the K elements are assumed to have been selected is

$$y_{ij} = \mu + \eta_{ij} \tag{1}$$

where the η_{ij}'s are normal random variables with mean zero and

$$\begin{aligned} E(\eta_{ij}\,\eta_{k\ell}) &= \tau^2 + \sigma^2, && i = k,\ j = \ell, \\ &= \tau^2, && i = k,\ j \neq \ell, \\ &= 0, && i \neq k. \end{aligned} \tag{2}$$

This two-stage model has previously been used by Fuller [1973] to estimate parameters of the super-population. It is also a special case of the model used by Royall [1976] and Scott and Smith [1969] in which the variance of y_{ij} is constant for all i. Royall also uses this simplified model when comparing alternative estimators for the population total and when considering efficient sample designs.

The methodology used by Royall [1976] in estimating the finite population total involves selecting a random sample s of n clusters, and from the M_i elements in each of the sampled clusters selecting a random sample s_i of size m_i. The finite population total is then partitioned into the sum of sampled elements, the sum of non-sampled elements from sampled clusters, and the sum of non-sampled elements from non-sampled clusters. The sum of the non-sampled elements is then estimated using the combined knowledge of s and the assumed super-population model.

One of the estimators for the population total, $T = \sum_{i}^{N} \sum_{j}^{M_i} y_{ij}$, suggested by Royall is

$$\hat{T}_H = \sum_{i\varepsilon s} \sum_{j\varepsilon s_i} y_{ij} + \sum_{i\varepsilon s} (M_i - m_i)\bar{y}_i + \sum_{i\varepsilon s} (m_i\bar{y}_i/k) \sum_{i\notin s} M_i \tag{3}$$

where $\bar{y}_i = \sum_{j\varepsilon s_i} y_{ij}/m_i$, $i\varepsilon s$, (4)

and $k = \sum_{i\varepsilon s} m_i$.

This estimator will be used to illustrate a new technique which constructs an exact confidence interval on T. Under the super-population model assumed in (1), the variance of $(\hat{T}_H - T)$ is

$$V(\hat{T}_H - T) = \tau^2 \left\{ \sum_{i\notin s} M_i^2 + \frac{\left(\sum_{i\notin s} M_i\right)^2 \sum_{i\varepsilon s} m_i^2}{k^2} \right\} + \sigma^2 \left\{ \sum_{i\varepsilon s} \frac{M_i^2}{m_i} - \frac{\left(\sum_{i\varepsilon s} M_i\right)^2}{k} + K\left(\frac{K}{k} - 1\right) \right\}. \tag{5}$$

A comment should be made about the preceding results and those to follow in this section. When clusters are of unequal size, even though $M_1,\ldots,M_N$ are fixed and assumed known, in a strict probabilistic sense the $M_1,\ldots,M_n$ corresponding to the sampled clusters are really random variables whose realization depends upon which clusters are sampled. Hence, the arguments used above and those to follow are really conditional arguments for given values of $M_1,\ldots,M_n$. However, since the unbiasedness of $\hat{T}_H$ and the confidence level of the corresponding confidence interval will not depend upon the values of $M_1,\ldots,M_n$, these properties will also apply in an unconditional sense.

The problems of estimating a linear combination of variance components such as $V(\hat{T}_H - T)$ for the unbalanced case are well known and the interested reader is referred to Searle [1971] for a complete discussion. However, new results due to Burdick and Sielken [1977] can be used to construct an exact confidence interval on T. The method considers the random variable $U_i = c_1 y_i + c_{2i} d_i$ for $i\varepsilon s$, where $d_i = \sum_{j\varepsilon s_i} \ell_{ij} y_{ij}$, $\sum_{j\varepsilon s_i} \ell_{ij} = 0$,

$c_{3i} = \sum_{j\epsilon s_i} \ell_{ij}^2$, and the ℓ_{ij}'s, c_1, c_{2i}'s, and c_{3i}'s are constants. Under model (1), $V(U_i) = c_1^2\tau^2 + (c_{2i}^2\, c_{3i} + c_1^2/m_i)\,\sigma^2$. Thus, with

$$c_1^2 = \sum_{i\notin s} M_i^2 + \left(\sum_{i\notin s} M_i\right)^2 \sum_{i\epsilon s} m_i^2/k^2 \quad (6)$$

and

$$c_{2i}^2\, c_{3i} = \sum_{i\epsilon s} \frac{M_i^2}{m_i} - \frac{\left(\sum_{i\epsilon s} M_i\right)^2}{k} + K\left(\frac{K}{k} - 1\right) - c_1^2/m_i \quad (7)$$

the U_i's are independent identically distributed $N(c_1\mu,\ V(\hat{T}_H - T))$. Letting $a = \{i \mid c_{2i}^2\, c_{3i} \geq 0\}$ and b denote the number of elements in set a, then $\sum_{i\epsilon a} (U_i - \bar{U})^2/V(\hat{T}_H - T) \sim \chi^2_{(b-1)}$ where $\bar{U} = (1/b) \sum_{i\epsilon a} U_i$. An unbiased estimator for $V(\hat{T}_H - T)$ is therefore

$$v_H = \frac{1}{b-1} \sum_{i\epsilon a} (U_i - \bar{U})^2. \quad (8)$$

For the special case where all $m_i = m$, $b = n$.

Burdick [1976] has shown that when $M_i = M$ and $m_i = m$ for all i, $(\hat{T}_H - T)$ is independent of $(b-1)\, v_H/V(\hat{T}_H - T)$. Thus, since $(\hat{T}_H - T) \sim N(0, V(\hat{T}_H - T))$ and $(b-1)\, v_H/V(\hat{T}_H - T) \sim \chi^2_{(b-1)}$, it follows that in this case $(\hat{T}_H - T)/\sqrt{v_H}$ will have an exact t-distribution with (b-1) degrees of freedom and that an exact $100(1 - \delta)\%$ confidence interval on T is

$$[\hat{T}_H \pm t_{\delta/2;\ b-1} \sqrt{v_H}]. \quad (9)$$

It should be noted that (9) is an exact confidence interval for any choice of ℓ_{ij} as long as $\sum_{j\epsilon s_i} \ell_{ij} = 0$ and equations (6) and (7) are satisfied for all i. Since the length of the confidence interval is determined by the value of $\sqrt{v_H}$, it would seem to be important to minimize this quantity when selecting the ℓ_{ij}. However, since the distribution of v_H does not depend on ℓ_{ij}, any convenient set of ℓ_{ij} may be used.

For example, if m is even, let

$$\ell_{ij} = -1,\ j = 1,\ldots,\frac{m}{2},$$
$$= +1,\ j = \frac{m}{2} + 1,\ldots,\ m, \quad (10)$$

and, if m is odd, let

$$\ell_{ij} = -1,\ j = 1,\ldots,\frac{m-1}{2},$$
$$= 0,\ j = \frac{m+1}{2},$$
$$= +1,\ j = \frac{m+3}{2},\ldots,m, \quad (11)$$

for all i. As discussed by Burdick and Sielken [1977], these values represent a good choice with respect to the robustness of the confidence interval to model breakdown.

In the more general case where either all of the M_i's or all of the m_i's are not equal $(\hat{T}_H - T)$ is not necessarily independent of v_H and the confidence interval given by (9) is only approximate. Furthermore, when all of the m_i's are not equal it is possible that $b < n$. This implies that some of the observations used in calculating $\hat{T}_H$ would be ignored in the calculation of v_H. This weakness can sometimes be avoided by using a "pooling" procedure to estimate v_H as suggested by Burdick and Sielken [1977].

FOOTNOTES

[1]This work was done at Texas A&M and supported by a grant from the Army Research Office, contract number DAHC04-74-C0018.

REFERENCES

Burdick, R. K. [1976]. "A Super-Population Approach to Multi-Stage Sampling" Ph.D. dissertation, Institute of Statistics, Texas A&M University.

Burdick, R. K., and Sielken, R. L. [1977]. Exact Confidence Intervals for Linear Combinations of Variance Components in Nested Classifications. Article presently under review by J. Amer. Statist. Assn.

Fuller, W. A. [1973]. Regression Analysis for Sample Surveys, paper presented at the Vienna meeting of the International Institute of Survey Statisticians.

Royall, R. M. [1976]. The Linear Least-Squares Prediction Approach to Two-Stage Sampling. J. Amer. Statist. Assn. 71, 657-664.

Scott, Alastair, and Smith, T. M. F. [1969]. Estimation in Multi-Stage Surveys. J. Amer. Statist. Assn. 64, 830-840.

Searle, S. R. [1971]. Topics in Variance Components Estimation. Biometrics 27, 1-76.

PROPERTIES OF ORDINARY AND WEIGHTED LEAST SQUARE ESTIMATORS OF REGRESSION COEFFICIENTS FOR TWO-STAGE SAMPLES

Cathy Campbell, University of Minnesota

1. Introduction

In this paper we present a simple regression model for two-stage cluster samples. It is hypothesized that the model may be appropriate for many situations in which both the clusters and elements within the clusters are assumed to be sampled from infinite populations. As such, it is not a model for sampling from finite populations, but may also be considered as a super-population model for two-stage samples taken from finite populations.

The model of interest can be given as

$y_{ij} = \mu + \underline{x}_{ij}'\beta + u_{ij}$ $(i=1,\ldots,b;\ j=1,\ldots,n_i)$,

or in matrix notation as

$$\underline{y} = \underline{1}\mu + X\underline{\beta} + \underline{u} \qquad (1.1)$$

where $\underline{1}$ is an nx1 vector of 1's,

$\underline{y} = (y_{11},\ldots,y_{bn_b})'$ is an nx1 vector of observed dependent variables,

X is an nxp matrix of observed predictor variables,

μ and $\underline{\beta}$ (px1) are unknown parameters

$\underline{u}$ is an nx1 vector of unobserved random variables with

$E(\underline{u}|X) = 0$,

$\mathrm{Var}(\underline{u}|X) = \sigma^2 V$,

$$V = (1-\rho_y)I + \rho_y \begin{bmatrix} J_{n_1} & & \emptyset \\ & \ddots & \\ \emptyset & & J_{n_b} \end{bmatrix} \qquad (1.2)$$

$$= (1-\rho_y)I + \rho_y X_b X_b', \qquad (1.3)$$

J_{n_i} is an $n_i x n_i$ matrix of 1's,

$X_b = \begin{bmatrix} \underline{1}_{n_1} & & \emptyset \\ & \ddots & \\ \emptyset & & \underline{1}_{n_b} \end{bmatrix}$ is the matrix of indicator variables identifying the cluster from which each element was sampled,

ρ_y is the intraclass correlation of the residuals around the regression line,

$$\frac{-1}{n_b-1} < \rho_y \leq 1.$$

From (1.2) and (1.3), it is clear that

$$\mathrm{Var}(y_{ij}) = \sigma^2$$
$$\mathrm{Cov}(y_{ij},y_{ik}) = \rho_y\sigma^2 \quad (j\neq k)$$
$$\mathrm{Cov}(y_{ij},y_{k\ell}) = 0 \quad (i\neq k).$$

Since the constant σ^2 appears as a constant multiplier on all variance expressions, and will cancel from all ratios, for convenience we assume $\sigma^2=1$.

Our interest in this model lies in studying the estimation of $\underline{\beta}$; μ is considered a nuisance parameter. Conditional on X, the weighted least squares (WLS) estimator of $\underline{\beta}$ is BLU, but is rarely used because it depends on unknown parameters and is difficult to compute. More often, because of availability of computer programs and the familiarity of the technique, ordinary least squares (OLS) is used to estimate $\underline{\beta}$. In this paper we wish to consider two aspects of the estimation of $\underline{\beta}$.

(1) Sometimes cluster samples are taken for convenience or economy, sometimes from necessity. What would be the effect on the variance of the parameter estimates if a simple random sampling procedure were used instead? In sampling terminology we wish to study the design effect for the OLS estimator of $\underline{\beta}$.

(2) Is OLS an efficient estimation procedure when model (1.1) holds? If OLS is extremely inefficient, then perhaps some form of approximate WLS, using an estimate of ρ_y, should be considered as an alternative.

For convenience, we restrict our results here to models with one or two predictor variables and consider the issue of design effects first.

2. Design Effect for Simple Linear Regression

When p=1, model (1.1) becomes

$$\underline{y} = \underline{1}\mu + \underline{x}\beta + \underline{u}. \qquad (2.1)$$

We assume that $\underline{x}$ has been transformed so that $\underline{x}'\underline{1} = 0$. Then the OLS estimator of β is given by

$$\hat{\beta}_o = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{y}, \qquad (2.2)$$

and

$$\mathrm{Var}(\hat{\beta}_o|\underline{x}) = (\underline{x}'\underline{x})^{-2}\underline{x}'V\underline{x}. \qquad (2.3)$$

Following Frankel (1971), we define the design effect of $\hat{\beta}_o$, $\mathrm{Deff}(\hat{\beta}_o)$, as the ratio of the variance of $\hat{\beta}_o$ under model (1.1) to the variance of $\hat{\beta}_o$ under the assumption of a simple random selection of elements of the same overall sample size. As $\mathrm{Var}(y_{ij})=\sigma^2=1$, $\mathrm{Var}(\hat{\beta}_o|\underline{x})$ with simple random sampling is $(\underline{x}'\underline{x})^{-1}$. Therefore

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = \frac{(\underline{x}'\underline{x})^{-2}\underline{x}'V\underline{x}}{(\underline{x}'\underline{x})^{-1}} = (\underline{x}'\underline{x})^{-1}\underline{x}'V\underline{x}. \qquad (2.4)$$

More correctly, the expression (2.4) should be called a conditional design effect since the same $\underline{x}$ is used in both numerator and denominator.

Without loss of generality, we may assume $\underline{x}'\underline{x} = 1$ in (2.4) and obtain

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = \underline{x}'V\underline{x} \quad (\underline{x}'\underline{x} = 1,\ \underline{x}'\underline{1} = 0). \qquad (2.5)$$

Substituting (1.3) for V in (2.5) yields

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\underline{x}'X_bX_b'\underline{x}-1)\rho_y. \qquad (2.6)$$

To make (2.6) more easily comparable with the usual expressions for design effects, it is convenient to express $\underline{x}'X_bX_b'\underline{x}$ in terms of the intra-

class correlation, ρ_x, of the observed $\underline{x}$. We first note that $\underline{x}'X_b X_b'\underline{x}$ is the sum of squares of the b cluster totals of $\underline{x}$, and can be expressed as

$$\underline{x}'X_b X_b'\underline{x} = \sum_{i=1}^{b} n_i^2 \bar{x}_i^2 . \qquad (2.7)$$

Following Murthy (1967), we use the following definition of ρ_x that is applicable for unequal cluster sizes:

$$\rho_x = \frac{\sum_{i=1}^{b}\sum_{k\neq j}^{n_i}\sum^{n_i}(x_{ij}-\bar{x})(x_{ik}-\bar{x})}{\sum_{i=1}^{b} n_i(n_i-1)\sigma_x^2} . \qquad (2.8)$$

Using the relationships $\sigma_x^2 = \frac{1}{n}$ and $\bar{x} = 0$, (2.8) reduces to

$$\rho_x = \frac{\Sigma n_i^2 \bar{x}_i^2 - 1}{\frac{\Sigma n_i^2}{\Sigma n_i} - 1} ,$$

which gives

$$\Sigma\, n_i^2 \bar{x}_i^2 = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_x . \qquad (2.9)$$

Substituting (2.9) in (2.6) gives

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_x\rho_y \qquad (2.10)$$

$$= 1 + \left(\frac{\mathrm{Var}(n_i)}{\bar{n}} + \bar{n} - 1\right)\rho_x\rho_y \qquad (2.11)$$

where $\bar{n} = \sum_{i=1}^{b} n_i/b$ is the average sample size.

Noting that (2.9) is the design effect (see (2.13)) for estimating the mean of $\underline{x}$, we can also obtain

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\mathrm{Deff}(\bar{x}) - 1)\rho_y . \qquad (2.12)$$

We now wish to make the following points about $\mathrm{Deff}(\hat{\beta}_o|\underline{x})$:

(i) When the sample sizes are all equal,

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) = 1 + (\bar{n}-1)\rho_x\rho_y . \qquad (2.13)$$

(ii) To obtain the design effect for estimating μ under model (1.1), we let $\underline{x}$ in (2.10) be $\frac{1}{\sqrt{n}}\underline{1}_n$ and define the intraclass correlation for a column of 1's as 1. Then (2.13) reduces to

$$\mathrm{Deff}(\hat{\mu}_o) = 1 + \left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right)\rho_y , \qquad (2.14)$$

which also shows why (2.9) is $\mathrm{Deff}(\bar{x})$.

(iii) With equal sample sizes, (2.14) becomes

$$\mathrm{Deff}(\hat{\mu}_o) = 1 + (\bar{n}-1)\rho_y , \qquad (2.15)$$

the well-known design effect for cluster samples.

(iv) If ρ_x and ρ_y have the same sign, then $\mathrm{Deff}(\hat{\beta}_o|\underline{x}) > 1$, while the converse holds if ρ_x and ρ_y have opposite signs.

(v) If either ρ_x or ρ_y is 0, then $\mathrm{Deff}(\hat{\beta}_o) = 1$.

(vi) If $\rho_x > 0$ and $\rho_y > 0$, then

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) \leq \mathrm{Deff}(\hat{\mu}_o)$$

and

$$\mathrm{Deff}(\hat{\beta}_o|\underline{x}) \leq \mathrm{Deff}(\bar{x}) .$$

The last point is an important piece of theoretical evidence in support of Kish and Frankel's (1974) observation that design effects for complex statistics (including regression coefficients) tend to be less than design effects for means.

The fact that the design effects for means obtained from this model reduce to those used in practice for balanced samples is encouraging as is the fact that the empirical observation of Kish and Frankel is supported by the use of model (1.1).

Unfortunately, at the time of this writing, we do not have empirical values of the design effects for single variable regressions with which to compare (2.13) or (2.11). Therefore, it is not yet possible to verify the applicability of these results to sample survey situations.

3. Design Effects in a Two-Variable Regression

The model we use here is

$$\underline{y} = \underline{1}\mu + \underline{x}_1\beta_1 + \underline{x}_2\beta_2 + \underline{u} \qquad (3.1)$$

with $\underline{x}_1'\underline{1} = \underline{x}_2'\underline{1} = 0$, and $\underline{x}_1'\underline{x}_1 = \underline{x}_2'\underline{x}_2 = 1$. The variance of the OLS estimator of $(\beta_1,\beta_2)'$ is

$$\mathrm{Var}\begin{pmatrix}\hat{\beta}_{1o}\\ \hat{\beta}_{2o}\end{pmatrix} X = (X'X)^{-1}X'VX(X'X)^{-1} , \qquad (3.2)$$

where $X = [\underline{x}_1,\underline{x}_2]$.

With the restriction $\underline{x}_1'\underline{x}_1 = \underline{x}_2'\underline{x}_2 = 1$,

$$X'X = \begin{bmatrix}1 & r\\ r & 1\end{bmatrix} \qquad (3.3)$$

and

$$(X'X)^{-1} = \frac{1}{1-r^2}\begin{bmatrix}1 & -r\\ -r & 1\end{bmatrix} . \qquad (3.4)$$

The center matrix, X'VX, is

$$X'VX = (1-\rho_y)X'X + \rho_y\begin{bmatrix}\Sigma\, n_i^2\bar{x}_{1i}^2 & \Sigma\, n_i^2\bar{x}_{1i}\bar{x}_{2i}\\ \Sigma\, n_i^2\bar{x}_{1i}\bar{x}_{2i} & \Sigma\, n_i^2\bar{x}_{2i}^2\end{bmatrix} . \qquad (3.5)$$

Using (2.9), the diagonal elements of (3.5) can be easily represented in terms of the intraclass correlations ρ_{x_1} and ρ_{x_2}. By generalizing (2.8), we define the intraclass "co-correlation" of $\underline{x}_1$ and $\underline{x}_2$ as

$$\rho_{X_1X_2} = \frac{\sum_{i=1}^{b}\sum_{k\neq j}^{n_i}\sum^{n_i}(x_{1ij}-\bar{x}_1)(x_{2ik}-\bar{x}_2)}{\sum_{i=1}^{b} n_i(n_i-1)\sqrt{\sigma^2_{X_1}\sigma^2_{X_2}}}, \qquad (3.6)$$

which reduces to

$$\rho_{X_1X_2} = \frac{\sum n_i^2 \bar{x}_{1i}\bar{x}_{2i} - r}{\frac{\sum n_i^2}{\sum n_i} - 1}. \qquad (3.7)$$

We note that the sign of $\rho_{X_1X_2}$ does not depend on the sign of the covariance between the cluster totals, but on whether this covariance is larger or smaller than the overall correlation between $\underline{x}_1$ and $\underline{x}_2$. If the cluster totals are uncorrelated, then

$$\rho_{X_1X_2} = \frac{-r}{\frac{\sum n_i^2}{\sum n_i} - 1}.$$

To form the design effect for $\hat{\beta}_{1o}$, we perform the necessary matrix multiplication in (3.2), using (2.9) and (3.7) in (3.5), to find

$$\text{Deff}(\hat{\beta}_o|X) = 1 + \left[\frac{\sum n_i^{2-2}x_{1i} - 2r\sum n_i^{2-}\bar{x}_{1i}\bar{x}_{2i} + r^2\sum n_i^{2-2}x_{2i}}{1-r^2} - 1\right]\rho_Y \qquad (3.8)$$

$$= 1 + \left(\frac{\sum n_i^2}{\sum n_i} - 1\right)\left(\frac{\rho_{X_1} - 2r\rho_{X_1X_2} + r^2\rho_{X_2}}{1-r^2}\right)\rho_Y. \qquad (3.9)$$

Due to the number of parameters involved, it is difficult to make general statements about the value of $\text{Deff}(\hat{\beta}_{1o}|X)$. However, we can notice that:

(i) if $r = 0$, then (3.9) reduces to the single variable design effect of (2.10);

(ii) if $\rho_Y > 0$, then $\text{Deff}(\beta_{1o}|X)$ increases with ρ_{X_1} and ρ_{X_2};

(iii) $\text{Deff}(\hat{\beta}_{1o}|X)$ is larger if r and $\rho_{X_1X_2}$ have opposite signs than if they have the same sign.

(iv) $\text{Deff}(\hat{\beta}_{1o}|X)$ becomes very large if r approaches 1 or -1.

Perhaps a more intuitive parametrization of $\text{Deff}(\hat{\beta}_{1o}|X)$ occurs when it is expressed in terms of the design effects of x_1 and x_2. By letting ρ_{12} be the correlation coefficient between the block totals, we obtain

$$\text{Deff}(\hat{\beta}_{1o}|X) = \qquad (3.10)$$

$$1 + \left[\frac{\text{Deff}(\bar{x}_1) - 2r\rho_{12}\sqrt{\text{Deff}(\bar{x}_1)\text{Deff}(\bar{x}_2)} + r^2\text{Deff}(\bar{x}_2)}{1-r^2} - 1\right]\rho_Y.$$

To evaluate (3.10), we used data from Frankel's (1971) three variable regressions, considering the variables pairwise. Values of $\sqrt{\text{Deff}(\bar{x}_i)}$ were included in his appendix E. Values of r were available from Table 5.1. Values for ρ_Y were obtained by assuming $\text{Deff}(\bar{y}) = 1 + (\bar{n}-1)\rho_Y$ and solving for ρ_Y. As sufficient data were not available for evaluating ρ_{12}, we assumed it was equal to r. Only data from the six strata designs are used.

Frankel considered 2 different three variable regressions. Since we used the variables in pairs in 2 variable equations, (3.10) was evaluated twice for each regression coefficient. The results of our calculations and Frankel's empirically obtained values are given below.

Comparison of Theoretical and Empirical Design Effects

Variable	$\text{Deff}(\hat{\beta}_{io})$ from (3.10)		$\text{Deff}(\hat{\beta}_{io})$ from Frankel
6	1.067,	1.063	1.089
7	1.092,	1.088	1.134
12	1.089,	1.089	.984
8	1.058,	1.057	1.093
11	1.126,	1.128	1.080
17	1.251,	1.252	1.432

Variables 6, 7, and 12 were predictor variables in one equation while 8, 11, and 17 were included in the other equation. Except for variables 12 and 11, the results from (3.10) are somewhat smaller but ordered approximately the same as Frankel's results. Variable 12 is clearly an anomaly for which we have no explanation at this time. From Frankel's data we found variables 6 and 7 were highly correlated with each other and correlated only slightly with variable 12. Until results are obtained for three variable regressions, we do not know whether this explains the small design effect for variable 12.

In this section and the preceding one, we have presented expressions for conditional design effects for regression coefficients. These were obtained by assuming the data follow a simple linear model appropriate for two-stage sampling from infinite populations. It is hoped that these results may also shed some light on the properties of regression coefficients obtained from finite populations.

The comparisons in the above table are not totally discouraging. Further investigation is needed to determine whether the discrepancies are due to differences between two-variable and three-variable regressions or from some oversimplification in the assumed model.

4. Relative Efficiency of OLS for Cluster Samples

In this section we study the efficiency of OLS with respect to WLS when model (1.1) holds. We consider only single variable regressions and define the relative efficiency as

$$E^* = \frac{\text{Var}(\hat{\beta}_w|\underline{x})}{\text{Var}(\hat{\beta}_o|\underline{x})}, \qquad (4.1)$$

where $\hat{\beta}_w$ is the second element of

$$\begin{pmatrix} \hat{\mu}_w \\ \hat{\beta}_w \end{pmatrix} = (Z'V^{-1}Z)^{-1}Z'V^{-1}\underline{y}$$

with $Z = [\underline{1},\underline{x}]$. $\mathrm{Var}(\hat{\beta}_w)$ is the (2,2) element of $(Z'V^{-1}Z)^{-1}$. As before we assume $\underline{x}'\underline{1} = 0$ and $\underline{x}'\underline{x} = 1$. The efficiencies given here are a pessimistic reflection of the efficiency of OLS since $\mathrm{Var}(\hat{\beta}_w)$ can never be achieved.

It can be shown that

$$E^* = \frac{(1-\rho_Y)}{[1-\underline{a}_2'D\underline{a}_2][1+(\underline{a}_2'N\underline{a}_2-1)\rho_Y]}, \qquad (4.2)$$

where $\underline{a}_2' = [\sqrt{n_1}\bar{x}_1,\ldots,\sqrt{n_b}\bar{x}_b]$

$$N = \mathrm{diag}\,(n_1,\ldots,n_b)$$

$$D = \mathrm{diag}\left(\frac{n_1\rho_Y}{1+(n_1-1)\rho_Y},\ldots,\frac{n_b\rho_Y}{1+(n_b-1)\rho_Y}\right).$$

We note that

$$0 \le \underline{a}_2'\underline{a}_2 = \Sigma\, n_i\bar{x}_i^2 \le 1$$

is the between-cluster sum of squares of $\underline{x}$, since $\bar{x} = 0$. As such it is the length of the projection of $\underline{x}$ into the subspace spanned by X_b. The vector $\underline{a}_2$ contains the between-cluster information for regressing $\underline{y}$ on $\underline{x}$. If $\underline{a}_2 = \underline{0}$, all cluster means are 0 and $\underline{x}$ varies only within the clusters.

Rather than discussing the properties of E^*, we give some graphs of it in simple situations. We choose to represent $\underline{a}_2$ as

$$\underline{a}_2 = \sqrt{k}\,\underline{u}$$

where $\underline{u}'\underline{u} = 1$ and $k = \underline{a}_2'\underline{a}_2$. Using this representation, $\underline{a}_2'N\underline{a}_2 = k\underline{u}'N\underline{u}$, where the quantity $\underline{u}'N\underline{u}$ is a weighted average of the n_i and must satisfy

$$n_1 \le \underline{u}'N\underline{u} \le n_b . \qquad (4.3)$$

The restriction $\underline{x}'\underline{1} = 0$ translates to $\underline{u}'\underline{\ell} = 0$

where $\ell' = \left(\sqrt{\frac{n_1}{n}},\ldots,\sqrt{\frac{n_b}{n}}\right)$.

With this restriction on $\underline{u}$, equality on the left in (4.3) can be attained only if $n_1 = n_2$ and on the right only if $n_{b-1} = n_b$.

We also point out that k is a linear function of the intraclass correlation of $\underline{x}$ via

$$k = \frac{\rho_X\left(\frac{\Sigma n_i^2}{\Sigma n_i} - 1\right) + 1}{\underline{u}'N\underline{u}} . \qquad (4.4)$$

For the illustrations, we consider only balanced samples and use

$$E^* = \frac{(1-\rho_Y)\,(1-\rho_Y+\bar{n}\rho_Y)}{(1-\rho_Y+\bar{n}\rho_Y(1-k))(1-\rho_Y+\bar{n}\rho_Y k)} \qquad (4.5)$$

which is obtained from (4.2) with $n_1 = \ldots = n_b = \bar{n}$. The relationship between k and ρ_X simplifies to

$$k = \frac{1 + (\bar{n}-1)\rho_X}{\bar{n}} \qquad (4.6)$$

for balanced samples.

In figures 1 and 2 we present graphs of E^* versus k (or ρ_X) for different values of ρ_Y. Figure 1 contains results for $\bar{n} = 100$ and Figure 2 for $\bar{n} = 50$. Small values of ρ_Y and k such as are commonly found in sample survey data were used in the calculations.

If we define "reasonable efficiency" as $E^* \ge 0.75$, then with $\bar{n} = 100$ the efficiency of OLS could be unreasonably low if $\rho_Y > 0.05$ unless ρ_X is very small. With $\bar{n} = 50$, the values of E^* remain high until $\rho_Y > 0.10$. Given the small values of ρ_Y and ρ_X commonly present in social science data, OLS should be reasonably efficient for most ρ_X and ρ_Y when $\bar{n} \le 50$. With large values of $\bar{n}$, some inefficient estimates may result if OLS is used consistently.

We also point out that with large total sample sizes, $\bar{n}$ and/or b both large, then $\mathrm{Var}(\hat{\beta}_o)$ and $\mathrm{Var}(\hat{\beta}_w)$ may both be acceptably small even though the efficiency of OLS is low - making it not worthwhile to attempt a WLS analysis.

In conclusion, with the presence of clustering as modelled in (1.1), it appears that OLS is a reasonably efficient estimator of a single regression coefficient for many parameter values commonly obtained in social science data. Therefore standard computer programs can usually be used for calculating point estimates of regression coefficients. The properties of the estimated standard error of $\hat{\beta}_o$ provided by an OLS routine when model (1.1) holds have not been investigated at this time.

REFERENCES

Frankel, Martin R. (1971) Inference from Sample Surveys: An Empirical Investigation. Ann Arbor: Institute for Social Research, The University of Michigan.

Kish, Leslie and Frankel, Martin R. (1974). "Inference from complex samples." The Journal of the Royal Statistical Society, Series B, 36. No. 1, pp. 1-37.

Murthy, M.N. (1967). Sampling: Theory and Methods. Calcutta: Statistical Publishing Society.

ACKNOWLEDGMENT

This research is based, in part, on the author's Ph.D. dissertation which was completed at Southern Methodist University.

Figure 1

E* vs. $K(\rho_x)$ for Different Values of ρ_y: $\bar{n} = 100$

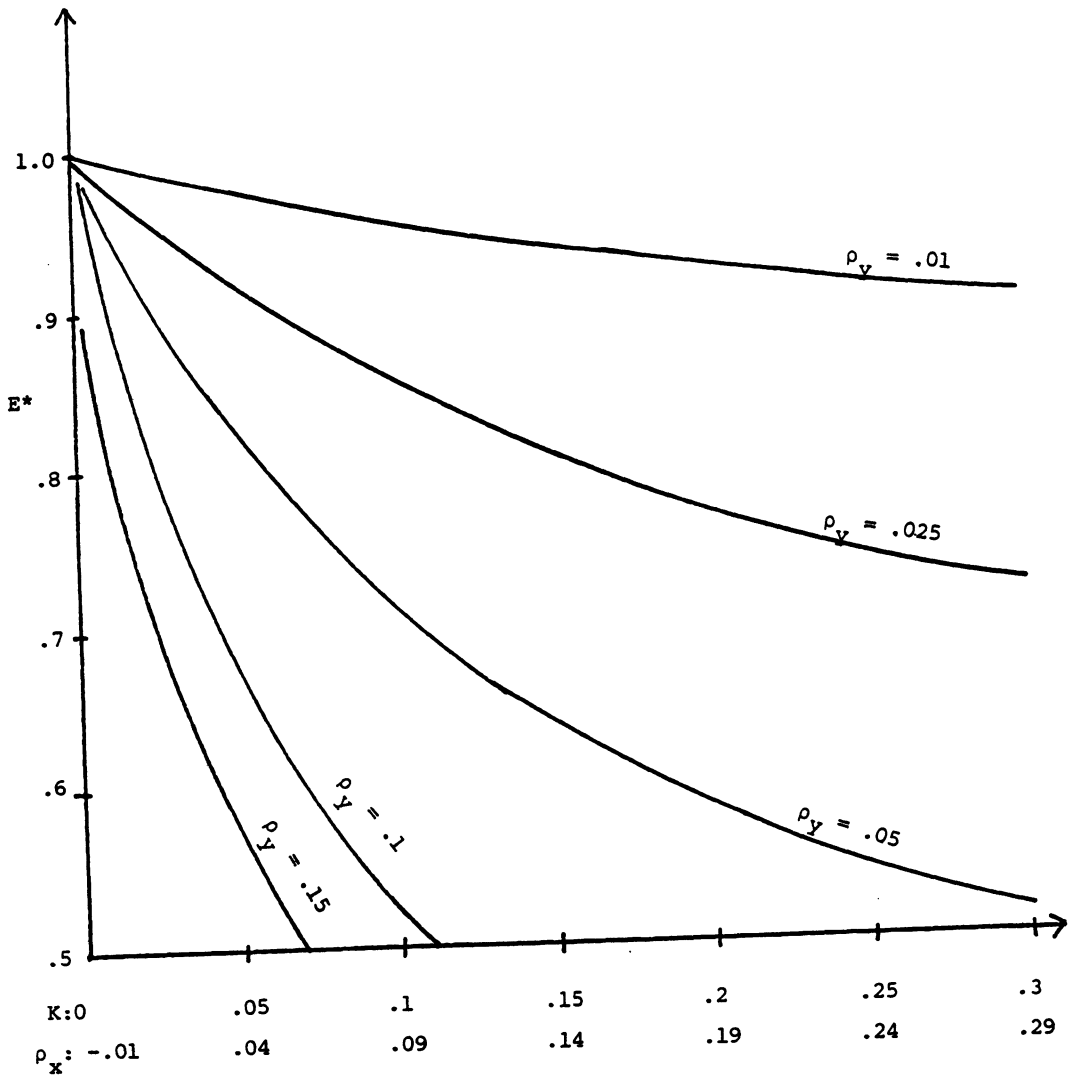
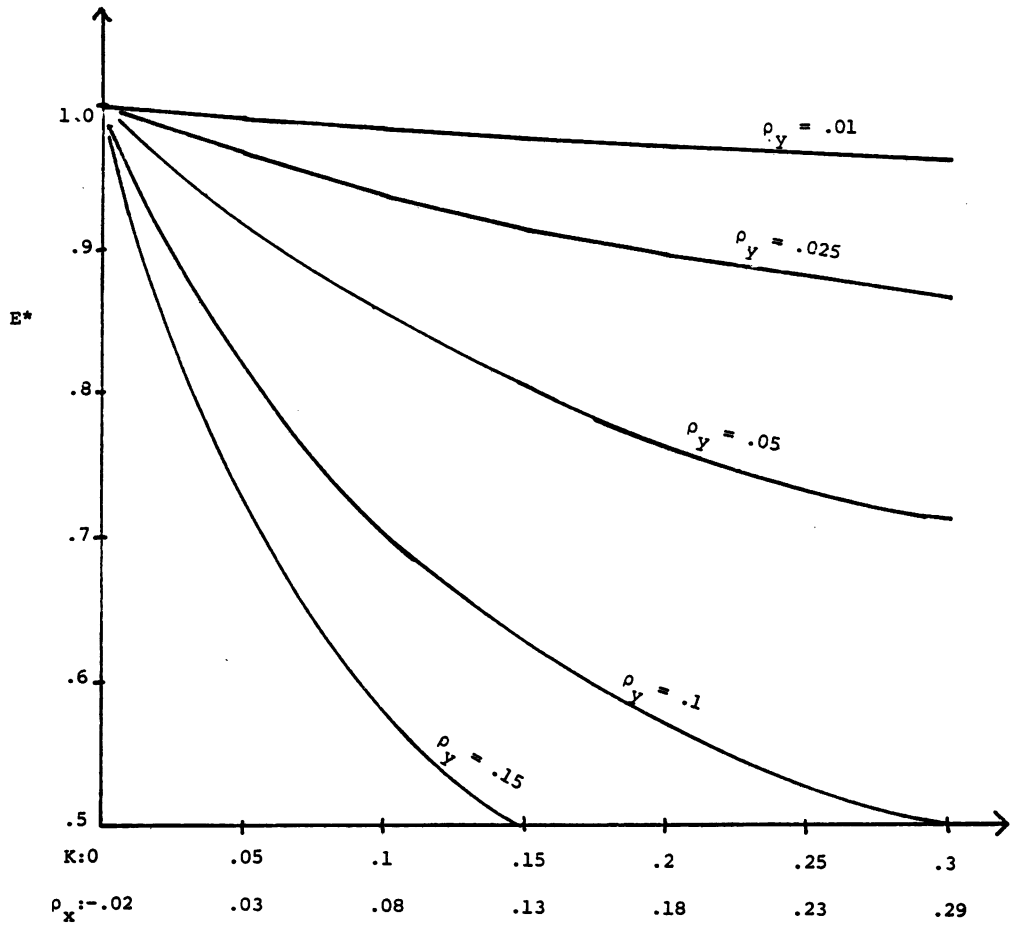


Figure 2

E* vs. ρ_x for Different Values of ρ_y: $\bar{n}$ = 50



JOINT ESTIMATES OF CURRENT LEVEL AND CHANGE IN THREE-STAGE SUCCESSIVE SAMPLING

Dharam S. Rana, Old Dominion University

1. Introduction - In many repeat surveys, it is required to report estimates of population mean on the current occasion and the immediately preceding occasion. These simultaneous estimates of mean and change are called joint or combined estimates. It has been seen that sometimes the required timings of the survey estimates is such that the estimate of the mean for the first occasion can wait until data for the second occasion is available. In such cases the estimate of mean on the first occasion can also be improved by using data from the second sample and the difference between these two estimated means gives the best linear estimate (i.e., unbiased estimate with minimum variance) of change that can be obtained from the data from the two samples. However, in many situations, estimates for the first occasion must be made before sample results from the second occasion are made available. In such cases, the population mean on the first occasion has to be estimated from the first sample only, and it may not be feasible to revise this initial estimate later on. It is the latter case that will be considered here to develop joint estimates of mean and change.

In three-stage successive sampling, there are many ways to alter the composition of the first sample on the second occasion. In the present paper only four important alternatives (sampling procedures) have been selected to obtain the joint estimates. It is assumed that the units in the population of interest are fixed and the sample size remains same on each occasion. The study is confined to two occasions only, but the results obtained can be extended to more than two occasions. On the first occasion, a simple random sample of n primary stage units (PSU's) is selected from the population of interest. Within each of n PSU's, a random selection of m second-stage units (SSU's) is made and in each of these nm SSU's a random sample of k third-stage units (TSU's) is taken. Selection at each stage is carried out by simple random sampling without replacement, and it is the same in case of all the four procedures.

2. Notations

Let N = Number of PSU,s in the population, M = Number of SSU's in each PSU, K = Number of TSU's in each SSU within PSU's, and $y_{hij\ell}$ = value of the ℓ-th tertiary unit in the j-th second-stage unit located in i-th first-stage unit.

$$S_{bh}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_{i..} - \bar{Y}...)^2 \ , \ h = 1, 2$$

= true variance among PSU means on the h-th occasion.

$$S_{wh}^2 = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_{ij.} - \bar{Y}_{i..})^2 \ , \ h = 1, 2$$

= true variance among SSU means on the h-th occasion.

$$S_{th} = \frac{1}{NM(K-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\ell=1}^{K}(y_{ij\ell} - \bar{Y}_{ij.})^2 \ , \ h = 1,2$$

= true variance among TSU's on the h-th occasion.

$$\rho_b S_{b1} S_{b2} = \frac{1}{N-1}\sum_{i=1}^{N}(\bar{Y}_{1i..} - \bar{Y}_{1...})(\bar{Y}_{2i..} - \bar{Y}_{2...})$$

$$\rho_w S_{w1} S_{w2} = \frac{1}{N(M-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{Y}_{1ij.} - \bar{Y}_{1i..})(\bar{Y}_{2ij.} - \bar{Y}_{2i..})$$

and

$$\rho_t S_{t1} S_{t2} = \frac{1}{NM(K-1)}\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{\ell=1}^{K}(y_{1ij\ell} - \bar{Y}_{1ij.})(y_{2ij\ell} - \bar{Y}_{2ij.})$$

represent true covariances among PSU mean, SSU means and TSU's respectively between first and second occasion. In the above relations, ρ_b, ρ_w and ρ_t denote true correlations among PSU means, SSU means and TSU's respectively $\bar{Y}_{h...}$, $\bar{Y}_{hi..}$ and $\bar{Y}_{hij}$ stand for population means of TSU's, i-th PSU and j-th SSU within i-th PSU on the h-th occasion (h = 1,2) respectively. For convenience sake, dots to denote $\bar{Y}_{h...}$, $\bar{Y}_{hi..}$, etc. will be dropped.

2.1 Joint Estimates of Mean and Change by Procedure (1)- All the PSU's of the first sample are retained on the second occasion but only a fraction r of SSU's with their sample of TSU's in each of these PSU's is retained. The remaining fraction s of SSU's is selected afresh in a random manner so that r + s = 1. Under this sampling plan, an initial estimate of $\bar{Y}_1$ the population on the first occasion can be written as

$$\bar{y}_{1(1)} = r\,\bar{y}'_{1(1)} + s\,\bar{y}^*_{1(1)}$$

The joint linear estimates of the population mean on the second occasion and the change that occured in the characteristic of interest between first and second occasion may be expressed as

$$\bar{y}_{2(1)} = a\,y_{1(1)} + b\,\bar{y}^*_{1(1)} + c\,\bar{y}'_{2(1)} + d\,\bar{y}^*_{2(1)}$$

and

$$\Delta_{(1)} = e\,\bar{y}'_{1(1)} + f\,\bar{y}^*_{1(1)} + g\,\bar{y}'_{2(1)} + h\,\bar{y}^*_{2(1)} \qquad (2.1.1)$$

where

$$\bar{y}'_{h(1)} = \frac{1}{nrmk}\sum_{i=1}^{n}\sum_{j=1}^{rm}\sum_{\ell=1}^{k} y_{hij\ell}, \ h = 1, 2$$

= mean per TSU based on nrmk matched TSU's. $\Delta_{(i)}$ = estimate of change between first and second occasion by i-th procedure (i = 1,2,3, 4) and

$$\bar{y}^*_{h(1)} = \frac{1}{nsmk}\sum_{i=1}^{n}\sum_{j=1}^{sm}\sum_{\ell=1}^{k} y_{hij\ell}, \ h = 1, 2$$

= mean per TSU based on nsmk TSU's. It is to be noted here that the number within parenthesis indicates the sampling procedure.

Suppose it is desired that the estimate of change between first and second occasion equals the difference between estimated means on these occasions, that is,

$$\Delta_{(1)} = \bar{y}_{2(1)} - \bar{y}_{1(1)}$$

Using the above relation the condition of unbiasedness, the joint estimates in (2.1.1) may be rewritten as

$$\bar{y}_{2(1)} = (e+r)\bar{y}'_{1(1)} - (e+r)\bar{y}^*_{1(1)} + c\bar{y}'_{2(1)} + (1-c)\bar{y}^*_{2(1)}$$

$$\Delta_{(1)} = e\bar{y}'_{1(1)} - (1+e)\bar{y}^*_{1(1)} + c\bar{y}'_{2(1)} + (1-c)\bar{y}^*_{2(1)} \qquad (2.1.2)$$

It may be useful to determine the weights c and e so as to minimize a linear function of variances of $\bar{y}_{2(1)}$ and $\Delta_{(1)}$. Let $\mathrm{Var}[\Delta_{(1)}] + \lambda\, \mathrm{Var}[\bar{y}_{2(1)}]$ represent one such linear function, where λ is a specified positive number. It will be assumed throughout this study that N, M and K are large and the true variances on the two occasions are equal, that is, $S^2_{b1} = S^2_{b2} = S^2_b$, $S^2_{w1} = S^2_{w2} = S^2_w$ and $S^2_{t1} = S^2_{t2} = S^2_t$. Yates [5] argues that wherever the successive sampling is likely to be used the assumption of equality of variances on consecutive occasions holds. Under the simplifying assumptions of equal variances and that the finite population correction factors e.g., $\frac{n}{N}$, $\frac{m}{M}$ and $\frac{k}{K}$ are negligible, it can be shown that

$$\begin{aligned}\mathrm{Var}\,\Delta_{(1)} &= e^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nrm} + \frac{S^2_t}{nrmk}\right) + (1+e)^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nsm} + \frac{S^2_t}{nsmk}\right) + c^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nrm} + \frac{S^2_t}{nrmk}\right) + (1-c)^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nsm} + \frac{S^2_t}{nsmk}\right) - 2e(1+e)\frac{S^2_b}{n} + 2ec\left(\rho_b\frac{S^2_b}{n} + \rho_w\frac{S^2_w}{nrm} + \rho_t\frac{S^2_t}{nrmk}\right) + 2e(1-c)\rho_b\frac{S^2_b}{n} - 2(1+e)c\,\rho_b\frac{S^2_b}{n} - 2(1+e)(1-c)\rho_b\frac{S^2_b}{n} + 2c(1-c)\frac{S^2_b}{n} \text{ and}\\ \mathrm{Var}\,y_{2(1)} &= (e+r)^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nrm} + \frac{S^2_t}{nrmk}\right) + \left(\frac{S^2_b}{n} + \frac{S^2_w}{nsm} + \frac{S^2_t}{nsmk}\right) + c^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nrm} + \frac{S^2_t}{nrmk}\right) + (1-c)^2\left(\frac{S^2_b}{n} + \frac{S^2_w}{nsm} + \frac{S^2_t}{nsmk}\right) - 2(e+r)^2\frac{S^2_b}{n} + 2(e+r)c\,\rho_b\frac{S^2_b}{n} + \rho_w\frac{S^2_w}{nrm} + \rho_t\frac{S^2_t}{nrmk} - 2(e+r)c\,\rho_b\frac{S^2_b}{n} + 2c(1-c)\frac{S^2_b}{n}\end{aligned}$$

The optimum weights c_o and e_o that will minimize the linear function $\mathrm{Var}[\Delta_{(1)}] + \lambda\, \mathrm{Var}[\bar{y}_{2(1)}]$ are obtained by solving the following equations for c_o and e_o:

$$\frac{\partial}{\partial c}\,\mathrm{Var}(\Delta_{(1)}) + \lambda\,\mathrm{Var}(\bar{y}_{2(1)}) = 0$$

$$\frac{\partial}{\partial e}\,\mathrm{Var}(\Delta_{(1)}) + \lambda\,\mathrm{Var}(y_{2(1)}) = 0$$

The optimum values of weights are

$$c_o = \frac{r\left(S^2_w + \frac{S^2_t}{k}\right)}{\left(S^2_w + \frac{S^2_t}{k}\right) - s\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)} - \frac{\lambda}{(1+\lambda)}\,\frac{rs\left(S^2_w + \frac{S^2_t}{k}\right)\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)}{\left[\left(S^2_w + \frac{S^2_t}{k}\right)^2 - s^2\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)^2\right]}$$

and

$$e_o = \frac{-r\left(S^2_w + \frac{S^2_t}{k}\right)}{\left(S^2_w + \frac{S^2_t}{k}\right) - s\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)} + \frac{\lambda}{(1+\lambda)}\,\frac{rs^2\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)^2}{\left[\left(S^2_w + \frac{S^2_t}{k}\right)^2 - s^2\left(\rho_w S^2_w + \rho_t\frac{S^2_t}{k}\right)^2\right]}$$

The joint estimators of $\bar{Y}_2$ and $\bar{Y}_2 - \bar{Y}_1$ with optimum weights, are given by

$$\begin{aligned}\bar{y}_{2(1)} &= \frac{-rs\beta_o}{(\alpha_o^2 - s^2\beta_o^2)}\left(\alpha_o + \frac{s\beta_o}{1+\lambda}\right)\left[\bar{y}'_{1(1)} - \bar{y}^*_{1(1)}\right] + \frac{r\alpha_o}{(\alpha_o^2 - s^2\beta_o^2)}\left(\alpha_o + \frac{s\beta_o}{1+\lambda}\right)\left[\bar{y}'_{2(1)} - \bar{y}^*_{2(1)}\right] + \bar{y}^*_{2(1)}\\ \Delta_{(1)} &= \frac{-r}{(\alpha_o^2 - s\beta_o^2)}\left[\alpha_o(\alpha_o + s\beta_o) - \frac{\lambda\, s\beta_o^2}{(1+\lambda)}\right]\left[\bar{y}'_{1(1)} - \bar{y}^*_{1(5)}\right] + \frac{r\,\alpha_o}{(\alpha_o^2 - s\beta_o^2)}\left(\alpha_o + \frac{s\beta_o}{(1+\lambda)}\right)\left[\bar{y}'_{2(1)} - \bar{y}^*_{2(1)}\right] + \bar{y}^*_{2(1)} - \bar{y}^*_{1(1)}\end{aligned}$$

where

$$\alpha_o = S^2_w + \frac{S^2_t}{k} \text{ and } \beta_o = \rho_w S^2_w + \rho_t\frac{S^2_t}{k}$$

For the special case of $\lambda = 1$, the variances of the joint estimators with optimum weights are

reduced to

$$\mathrm{Var}[\bar{y}_{2(1)}] = \frac{S_b^2}{n} + \frac{1}{nsm}\left(S_w^2 + \frac{S_t^2}{k}\right) - \frac{r(S_w^2 + \frac{S_t^2}{k})[(S_w^2 + \frac{S_t^2}{k})^2 - \frac{s^2}{4}(\rho_w S_w^2 + \rho_t \frac{S_t^2}{k})^2]}{nsm[(S_w^2 + \frac{S_t^2}{k})^2 - s^2(\rho_w S_w^2 + \rho_t \frac{S_t^2}{k})^2]}$$

and

$$\mathrm{Var}[\Delta_{(1)}] = \frac{2}{n}(1-\rho_b)S_b^2 + \frac{2}{nsm}(S_w^2 + \frac{S_t^2}{k}) + r(S_w^2 + \frac{S_t^2}{k})[s^2(\rho_w S_w^2 + \rho_t \frac{S_t^2}{k}) - 8(S_w^2 + \frac{S_t^2}{k})\{(1+s\rho_w)S_w^2 + (1+s\rho_t)\frac{S_t^2}{k}\}][4nsm\{(S_w^2 + \frac{S_t^2}{k})^2 - s^2(\rho_w S_w^2 + \rho_t \frac{S_t^2}{k})^2\}]^{-1}$$

2.2 Joint Estimates of Mean and Change by Procedure (2) - Under this sampling plan, a partial replacement of units is carried out at second and third stage. Retain all the first-stage units from the first sample but retain only a fraction r of the PSU's retained and make a fresh random selection of the remaining fraction s of SSU's (such that r + s = 1) on the second occasion. Within each of the nrm SSU's retained, further retain only a fraction t of TSU's, and supplement the remaining fraction u of TSU's selected at random so that t + u = 1.

An initial estimate of $\bar{Y}_1$, based on first sample only, is given by

$$\bar{y}_{1(2)} = rt\,\bar{y}'_{1(2)} + ru\,\bar{y}^{**}_{1(2)} + s\,\bar{y}^{*}_{1(2)}$$

The joint linear and unbiased estimators of $\bar{Y}_2$ and $\bar{Y}_2 - \bar{Y}_1$ may be written as

$$\bar{y}_{2(2)} = a\,\bar{y}'_{1(2)} + b\,\bar{y}^{**}_{1(2)} - (a+b)\,\bar{y}^{*}_{1(2)}$$

and

$$\Delta_{(2)} = f[\bar{y}'_{2(2)} - \bar{y}_{1(2)}] + g[\bar{y}^{**}_{2(2)} - \bar{y}^{**}_{1(2)}] + (1-f-g)[\bar{y}^{*}_{2(7)} - \bar{y}^{*}_{1(2)}]$$

where

$$\bar{y}'_{h(2)} = \frac{1}{nrmtk}\sum_{i=1}^{n}\sum_{j=1}^{rm}\sum_{\ell=1}^{tk} y_{ij\ell}\,,\ h = 1, 2$$

$$\bar{y}^{**}_{h(2)} = \frac{1}{nrmuk}\sum_{i=1}^{n}\sum_{j=1}^{rm}\sum_{\ell=1}^{k} y_{ij\ell},\ h = 1, 2$$

$$\bar{y}^{*}_{h(2)} = \frac{1}{nsmk}\sum_{i=1}^{n}\sum_{j=1}^{sm}\sum_{\ell=1}^{k} y_{ij\ell},\ h = 1, 2$$

If we impose the condition that $\Delta_{(2)} = \bar{y}_{2(2)} - \bar{y}_{1(2)}$ then by comparing coefficients on both sides it follows that a = rt - d, b = ru - e and f = d, g = e. So the joint linear unbiased estimators can be rewritten as

$$\bar{y}_{2(2)} = (rt - d)\bar{y}'_{1(2)} + (ru - e)\bar{y}^{**}_{1(2)} - (r - d - e)\bar{y}^{*}_{1(2)} + d\bar{y}'_{2(2)} + e\bar{y}^{**}_{2(2)} + (1 - d - e)\bar{y}^{*}_{2(2)}$$

$$\Delta_{(2)} = d[\bar{y}'_{2(2)} - \bar{y}'_{1(2)}] + e[\bar{y}^{**}_{2(2)} - \bar{y}^{**}_{1(2)}] + (1 - d - e)[\bar{y}^{*}_{2(2)} - \bar{y}^{*}_{1(2)}] \qquad (2.2.1)$$

and their variances are

$$\mathrm{Var}\ y_{2(2)} = [(rt - d)^2 + d^2]\alpha' + [(ru - e)^2 + e^2]\beta' + [(r - d - e)^2 + (1 - d - e)^2]\gamma' + 2[(rt - d)(ru - e) + de]\alpha^* + 2(rt - d)d\,\delta' + 2[(r - d - e)e + (ru - e)d]\delta^* + 2(1 - d - e)(d + e) - (r - d - e)^2\frac{S_b^2}{n} - 2(d + e)(r - d - e)\rho_b\frac{S_b^2}{n}$$

and

$$\mathrm{Var}\ \Delta_{(2)} = 2d^2(\alpha' - \delta') + 2e^2(\beta' - \delta^*) + 2(1 - d - e)^2(\gamma' - \rho_b\frac{S_b^2}{n}) + 4de[(1 - \rho_b)\frac{S_b^2}{n} + (1 - \rho_w)\frac{S_w^2}{nrm}] + 4d(1 - d - e)(1 - \rho_b)\frac{S_b^2}{n} + 4e(1 - d - e)(1 - \rho_b)\frac{S_b^2}{n} \qquad (2.2.2)$$

where

$$\alpha' = \frac{S_b^2}{n} + \frac{S_w^2}{nrm} + \frac{S_t^2}{nrmtk},\ \beta' = \frac{S_b^2}{n} + \frac{S_w^2}{nrm} + \frac{S_t^2}{nrmuk}$$

$$\delta' = \rho_b\frac{S_b^2}{n} + \rho_w\frac{S_w^2}{nrm} + \rho_t\frac{S_t^2}{nrmtk},\ \alpha^* = \frac{S_b^2}{n} + \frac{S_w^2}{nrm}$$

and $\delta^* = \rho_b\frac{S_b^2}{n} + \rho_w\frac{S_w^2}{nrm}$. The optimum weights that will minimize the linear function $\mathrm{Var}[y_{2(2)}] + \mathrm{Var}[\Delta_{(2)}]$ are

$$d_o = \frac{rt[(3+(1-s\rho_w)(1-u\rho_t))S_w^2 + (4-(s+ru)\rho_t)\frac{S_t^2}{k}]}{[(1-u\rho_t)(1-s\rho_w)S_w^2 + (1-(s+ru)\rho_t)\frac{S_t^2}{k}]}$$

and $e_o =$

$$\frac{ru[(3(1-\rho_t)+(1-s\rho_w)(1-u\rho_t))S_w^2 + (4-3\rho_t-(s+ru)\rho_t)\frac{S_t^2}{k}]}{[(1-u\rho_t)(1-s\rho_w)S_w^2 + (1-(s+ru)\rho_t)\frac{S_t^2}{k}]}$$

By substituting the optimum values of the weights in equations (2.2.1) and (2.2.2) the joint linear unbiased estimators mean and change and their variances with optimum weights can be obtained.

2.3 Joint Estimates of Mean and Change by Procedure (3) - In this plan, only a fraction p of the PSU's along with their samples of SSU's and TSU's from the first sample is retained and a fresh random selection of the remaining fraction q of PSU's is made on the second occasion. Note that p + q = 1 so that the sample size remains same on the two occasions.

Using data from first sample only, a linear unbiased estimate of $\bar{Y}_1$ is

$$\bar{y}_{1(3)} = p\,\bar{y}'_{1(3)} + q\,\bar{y}''_{1(3)}$$

The joint linear unbiased estimators of $\bar{Y}_2$ and $\bar{Y}_2 - \bar{Y}_1$ subject to the constraint $\Delta_{(3)} = \bar{y}_{2(3)} - \bar{y}_{1(3)}$ may be expressed as

$$\bar{y}_{2(3)} = (e+p)\bar{y}'_{1(3)} - (e+p)\bar{y}''_{1(3)} + c\bar{y}'_{2(3)} + (1-c)\bar{y}''_{2(3)}$$

and

$$\Delta_{(3)} = e\bar{y}'_{1(3)} - (1+e)\bar{y}''_{1(3)} + c\bar{y}'_{2(3)} + (1-c)\bar{y}''_{2(3)} \qquad (2.3.1)$$

where

$$\bar{y}'_{h(3)} = \frac{1}{npmk}\sum_{i=1}^{np}\sum_{j=1}^{m}\sum_{\ell=1}^{k} y_{ij\ell}\,, \quad h = 1, 2$$

$$\bar{y}''_{h(3)} = \frac{1}{nqmk}\sum_{i=1}^{nq}\sum_{j=1}^{m}\sum_{\ell=1}^{k} y_{ij\ell}\,, \quad h = 1, 2$$

The variances of joint estimators are

$$\mathrm{Var}[\bar{y}_{2(3)}] = \frac{(e+p)^2}{np}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right] + \frac{(e+p)^2}{nq}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right]$$
$$+ \frac{c^2}{np}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right] + \frac{(1-c)^2}{nq}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right]$$
$$+ \frac{2(e+p)c}{np}\left[\rho_b S_b^2 + \rho_w \frac{S_w^2}{m} + \rho_t \frac{S_t^2}{mk}\right]$$

and

$$\mathrm{Var}[\Delta_{(3)}] = \frac{e^2}{np}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right] + \frac{(1+e)^2}{nq}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right] + \frac{c^2}{np}\left[S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk}\right] + \frac{(1-c)^2}{nq}\left[S_b^2 + \frac{S_t^2}{mk}\right] + \frac{2ec}{np}\left[\rho_b S_b^2 + \rho_w \frac{S_w^2}{m} + \rho_t \frac{S_t^2}{mk}\right] \qquad (2.3.2)$$

The optimum weights that will minimize a linear function $\mathrm{Var}[\Delta_{(3)}] + \mathrm{Var}[y_{2(3)}]$ are obtained in the usual manner.

$$c_o = \frac{p\,\alpha}{\alpha - q\delta} - \frac{pq\alpha\delta}{2(\alpha^2 - q^2\delta^2)}$$

and

$$e_o = \frac{-p\alpha}{\alpha - q\delta} + \frac{pq^2\delta^2}{2(\alpha^2 - q^2\delta^2)} \qquad (2.3.3)$$

where

$$\alpha = S_b^2 + \frac{S_w^2}{m} + \frac{S_t^2}{mk} \text{ and } \delta = \rho_b S_b^2 + \rho_w \frac{S_w^2}{m} + \rho_t \frac{S_t^2}{mk}$$

From equations (2.3.2) and (2.3.3) the optimum variances of the joint estimators are given by

$$\mathrm{Var}[y_{2(3)}] = \frac{\alpha}{4n}\,\frac{(4\alpha^2 - 3q\delta^2 - q^2\delta^2)}{(\alpha^2 - q^2\delta^2)}$$

and

$$\mathrm{Var}[\Delta_{(3)}] = \frac{\alpha}{4n}\,\frac{(8\alpha^2 - 8q\delta^2 + pq\delta^2 - 8p\alpha\delta)}{(\alpha^2 - q^2\delta^2)}$$

From equations (2.3.1) and (2.3.3) the joint estimators of mean and change with optimum weights can be obtained.

2.4 Joint Estimates of Mean and Change by Procedure (4) - Partial replacement of units at primary as well as secondary stage is considered in this procedure. Only a fraction p of the PSU's from the first sample is retained on the second occasion and the rest of the fraction q of PSU's is selected anew. Within each of the np PSU's retained, only a fraction r of SSU's with their samples of TSU's is retained and the remaining fraction s of SSU's is selected afresh in a random manner. Again, p + q = 1 and r + s = 1 so as to keep the same sample size on the two occasions.

An initial estimate of $\bar{Y}_2$ based on first sample only is given by

$$\bar{y}_{1(4)} = pr\,\bar{y}'_{1(4)} + ps\,\bar{y}^*_{1(4)} + q\,\bar{y}''_{1(4)}$$

Subject to the constraint $\Delta_{(4)} = \bar{y}_{2(4)} - \bar{y}_{1(4)}$ the joint linear unbiased estimators of $\bar{Y}_2$ and $\bar{Y}_2 - \bar{Y}_1$ may be written as

$$\bar{y}_{2(4)} = (pr-d)\bar{y}'_{1(4)} + (ps-e)\bar{y}^*_{1(4)} - (p-d-e)\bar{y}''_{1(4)} + d\,\bar{y}'_{2(4)} + e\,\bar{y}^*_{2(4)} + (1-d-e)\,\bar{y}''_{2(4)}$$

and

$$\Delta_{(4)} = d[\bar{y}'_{2(4)} - \bar{y}'_{1(4)}] + e[\bar{y}^*_{2(4)} - \bar{y}^*_{1(4)}] + (1 - d - e)[\bar{y}''_{2(4)} - \bar{y}''_{1(4)}] \qquad (2.4.1)$$

The variances of the joint estimators in (2.4.1) are given by

$$\mathrm{Var}[y_{2(4)}] = [(pr-d)^2 + d^2]\alpha'' + [(ps-e)^2 + e^2]\beta'' + [(p-d-e)^2 + (1-d-e)^2]\alpha'' + 2(pr-d)d\,\delta'' + 2\frac{S_b^2}{np}[(pr-d)(ps-e) + (pr-d)e\,\rho_b + (ps-e)d\,\rho_b + \rho_b + de]$$

and

$$\mathrm{Var}[\Delta_{(4)}] = 2d^2(\alpha'' - \delta'') + 2e^2\left(\beta'' - \rho_b \frac{S_b}{np}\right)$$

$$+ 2(1 - d - e)^2\gamma'' + 4\,de(1 - \rho_b)\frac{S_b^2}{np} \qquad (2.4.2)$$

where

$$\alpha'' = \frac{S_b^2}{np} + \frac{S_w^2}{nprm} + \frac{S_t^2}{nprmk}, \quad \beta'' = \frac{S_b^2}{np} + \frac{S_w^2}{npsm} + \frac{S_t^2}{npsmk}$$

$$\gamma'' = \frac{S_b^2}{nq} + \frac{S_w^2}{nqm} + \frac{S_t^2}{nqmk} \quad \text{and} \quad \delta'' = \frac{\rho_b S_b^2}{np} + \frac{\rho_w S_w^2}{nprm} + \frac{\rho_t S_t^2}{nprmk}$$

To determine the suitable values of weights d and e that will minimize the linear function $\mathrm{Var}[y_{2(4)}] + \mathrm{Var}[\Delta_{(4)}]$, it can be shown that the solution of the simultaneous equations

$$\frac{\partial}{\partial d}\{\mathrm{Var}\ y_{2(4)}] + \mathrm{Var}[\Delta_{(4)}]\} = 0$$

and

$$\frac{\partial}{\partial e}\{\mathrm{Var}\ y_{2(4)}] + \mathrm{Var}[\Delta_{(4)}]\} = 0$$

provides the following optimum values: $e_o =$

$$\frac{ps[S_b^2\{4(\alpha_o-\beta_o)-q\rho_b(\alpha_o-s\beta_o)+r\beta_o\}+\frac{\alpha_o}{m}\{4(\alpha_o-\beta_o)+pr\beta_o\}]}{4[S_b^2(1-q\rho_b)(\alpha_o-s\beta_o)+\frac{\alpha_o}{m}\{\alpha_o-\beta_o(q+ps)\}]}$$

and

$$d_o = \frac{pr[S_b^2\{4\alpha_o-q\rho_b(\alpha_o-s\beta_o)-s\beta_o\}+\frac{\alpha_o}{m}\{4\alpha_o-(q+ps)\beta_o\}]}{4[S_b^2(1-q\rho_b)(\alpha_o-s\beta_o)+\frac{\alpha_o}{m}\{\alpha_o-\beta_o(q+ps)\}]} \qquad (2.4.3)$$

where α_o and β_o are defined in section 2.1. The joint estimators of mean and change with optimum weights can be obtained from equations (2.4.1) and (2.4.3) and their optimum variance can be obtained from equations (2.4.2) and (2.4.3).

2.5 Comparison - Combined estimates of mean and change have been obtained by four different sampling plans. It is important to find out which of the plans is more efficient. Relative performance of these plans is studied here for the same overall replacement fraction, say q^*. By equating total number of units replaced, it follows that $q^* = s + u - su$ for procedure (2) and $q^* = q + s - qs$ for procedure (4). From these preceding relations, it is seen that $q = u$. In case of procedure (1) and (3), it is obvious that $q^* = s$ and $q^* = q$ respectively. For convenience, the ratios S_w^2/S_b^2 and S_t^2/S_b^2 are denoted by ϕ and ψ respectively. It can be shown that for a three-stage sampling design to be useful, ϕ and ψ must satisfy the following conditions:

$$0 < \phi < M \quad \text{and} \quad 0 < \psi < K\phi$$

Let RJM34 = $\mathrm{Var}[\bar{y}_{2(4)}]/\mathrm{Var}[\bar{y}_{2(3)}]$ represent the relative efficiency of the jointly estimated mean by procedure (3) with respect to the joint estimate of mean by procedure (4). The symbols RJM31 and RJM32 have similar meanings. Similarly the symbols RJC14, RJC13 and RJC14 represent the relative efficiency of sampling procedure (1) with respect to procedures (4), (3) and (2) respectively in the combined estimation of change. The relative efficiencies of the four sampling plans are studied numerically for an arbitrarily selected range of values of the parameters and design quantities. Some of the results are arranged in Tables 1 through 4. Some important observations made from these tables are as follows:

(i) As ρ_b increases from 0.5 to 0.9, RJM34 increases moderately and gains in RJM32 and RJM31 are relatively more significant.

(ii) As ρ_w and ρ_t increase from 0.5 to 0.9, the changes produced in the values of RJM34, RJM31 and RJM32 are negligibly small.

(iii) When ϕ changes from 0.5 to 10, RJM34 remains practically unaltered, but RJM31 decreases moderately and RJM32 shows a slight decline with ϕ.

(iv) All three relative efficiencies register a slight decline as overall replacement fraction q^* changes from 0.65 to 0.85.

Some of the results from the numerical investigation of the joint estimates of change are presented in Tables 3 and 4. The following observations are made from these tables:

(i) RJC14 and RJC13 increase as ρ_b increases from 0.5 to 0.9, and RJC12 decreases slightly with ρ_b in most cases.

(ii) As ρ_w and ρ_t increase from 0.5 to 0.9, RJC13 increases slowly, however RJC12 and RJC14 remain almost unaltered.

(iii) All the three relative efficiencies decrease as ϕ increases from 0.5 to 10.

(iv) When q^* changes from 0.65 to 0.85, RJC13 shows significant gains but RJC12 and RJC14 remain practically unchanged.

2.6 Sample Allocation - In a design problem, it is important to study optimum allocation of sample. In the present section, optimum distribution of sample will be considered for a special case that is two-stage successive sampling. Assuming that travel cost among units is unimportant, one possible cost function for procedure (1) may be of the form

$$c = c_1 n + (c_2 + c_2' r + c_2'' s)nm$$

where c is total cost for two occasions, c_1 is the cost of preparing frame and c_2, c_2' and c_2'' are enumeration cost. The optimum values of m and n that will minimize the linear function $\mathrm{Var}[\bar{y}_{2(1)}] + \mathrm{Var}[\Delta_{(1)}]$ subject to the above cost function are

$$m_o = \frac{[12c_1S_w^2 - rc_1(12+8s\rho_w-2s^2\rho_w^2)S_w^2/(1-s^2\rho_w^2)]^{\frac{1}{2}}}{[4(c_2+c_2'r+c_2''s)sS_b^2(3-2\rho_b)]^{\frac{1}{2}}}$$

and

$$n_o = c[c_1 + m_o(c_2 + c_2'r + c_2''s)]^{-1}$$

A possible cost function for procedure (3) may be written as

$$c = (c_1 + c_2'q)n + (c_2 + c_2'p + c_2''q)nm$$

The optimum values of n and m that will minimize $Var[y_{2(3)}] + Var[\Delta_{(3)}]$ are obtained by method of successive approximation. The relative performance of sampling plans (1) and (3) are studied numerically and it is noted that procedure (1) is more efficient than procedure (3).

Table 1. $q = .5$, $s = .3$, $q^* = .65$

ρ_b	ρ_w	ρ_t	$\phi = .5$ x	y	z	$\phi = 10$ x	y	z
	.5	.5	102	105	105	102	103	104
.5	.7	.7	102	105	105	102	103	105
	.9	.9	102	105	105	102	101	106
	.5	.5	104	111	111	103	107	108
.7	.7	.7	104	111	112	104	107	110
	.9	.9	104	111	112	104	107	112
	.5	.5	109	125	125	105	112	113
.9	.7	.7	109	125	126	107	114	117
	.9	.9	109	126	126	108	116	122

Table 2. $q = .5$, $s = .7$, $q^* = 85$

ρ_b	ρ_w	ρ_t	$\phi = .5$ x	y	z	$\phi = 10$ x	y	z
	.5	.5	100	103	103	101	102	102
.5	.7	.7	100	103	103	100	101	101
	.9	.9	100	103	103	98	99	99
	.5	.5	100	107	107	101	104	104
.7	.7	.7	100	108	108	101	105	105
	.9	.9	100	107	107	100	103	104
	.5	.5	105	121	121	103	109	108
.9	.7	.7	106	121	122	104	111	111
	.9	.9	106	122	122	105	114	115

In Tables 1 and 2, x = RJM34, y = RJM31, z = RJM32, m = 16, k = 8, $\psi = 0.5$.

Table 3. $q = .5$, $s = .3$, $q^* = .65$

ρ_b	ρ_w	ρ_t	$\phi = .5$ x	y	z	$\phi = 10$ x	y	z
	.5	.5	133	147	99	118	126	90
.5	.7	.7	133	148	99	120	131	87
	.9	.9	134	150	94	131	149	88
	.5	.5	154	183	99	125	137	87
.7	.7	.7	154	184	98	126	141	83
	.9	.9	156	187	99	138	163	84
	.5	.5	189	247	96	131	147	83
.9	.7	.7	190	249	95	129	145	75
	.9	.9	195	260	96	137	164	72

Table 4. $q = .5$, $s = .7$, $q^* = .85$

ρ_b	ρ_w	ρ_t	$\phi = .5$ x	y	z	$\phi = 10$ x	y	z
	.5	.5	133	171	99	117	136	94
.5	.7	.7	133	171	99	118	140	90
	.9	.9	134	174	99	123	163	87
	.5	.5	154	237	99	126	155	93
.7	.7	.7	154	239	99	125	159	88
	.9	.9	155	245	98	129	187	83
	.5	.5	190	388	98	135	178	90
.9	.7	.7	189	388	96	131	173	83
	.9	.9	191	407	96	130	192	75

In Tables 3 and 4, x = RJC14, y = RJC13, z = RJC12, m = 16, k = 8, $\psi = 0.5$.

References

[1] Chakrabarty, R. P. and D. S. Rana (1974). Multi-stage Sampling with Partial Replacement of the Sample on Successive Occasions. Proceedings of the Amer. Stat. Assoc. Social Statistics Section, 1974, pp. 262-268.

[2] Rana, Dharam S. and R. P. Chakrabarty (1977). Three-stage Sampling on Successive Occasion. Submitted to Communications in Statistics.

[3] Hansen, M. H., Hurwitz, W. N. and Madow, W. Y. (1953). Sample Survey Methods and Theory. Vol. II. John Wiley and Sons, Inc., New York.

[4] Singh, D. (1968). Estimates in Successive Sampling Using a Multi-stage Design. Jour. Amer. Stat. Assoc. 63, pp. 99-112.

[5] Yates, F. (1960). Sampling Methods for Censuses and Surveys. Third edition, Charles Griffin and Co., Ltd., London.

DISCUSSION OF PAPERS ON SAMPLING PROBLEMS I

R. P. Chakrabarty, Jackson State University

Professor Monroe Lerner is to be congratulated for arranging the contributed papers sessions of the Social Statistics Section very neatly subjectwise. This made the deliberations of the contributed papers sessions very interesting and meaningful. The papers in this session deals with research problems in Survey Sampling.

Systematic sampling is often used for convenience sometimes from necessity in sample surveys. The estimation of the variance is, however, a long standing problem. Mr. Wolter and Ms. McCann are to be commended for taking up this important question. They have made an emperical study of the performance of the eight variance estimators available in literature using several artifical and real populations. Koop's variance estimator is comparable to the jackknife variance estimator. Its relatively poor performance is not surprising. Similar results for jackknife variance estimator in estimation of ratios were obtained by Chakrabarty and Rao (1967) and Rao and Rao (1971).

It is rather surprising to see that Cocharan's variance estimator derived for auto-correlated populations is superior in populations with linear trend and even in populations in random order. It gives smallest mean square error in almost all cases they have studied. As the authors mentioned there is thus need for further study along this line using different population models and live data to evaluate the performances of the variance estimators now available in literature and to provide a guideline about the choice of an estimator in a given situation.

In survey-sampling a complete 'frame' (list of sampling units) is sometimes either unavailable or too expensive to construct. In such situations the sample from an incomplete list may be supplemented by another sample from a complete areal frame to gain increased accuracy and to reduce costs. Since Hartley's (1962) paper outlining the theory of multiple frame surveys several researchers have proposed some alternative estimators in two frame surveys.

The paper by H. Huang compares the efficiency of the Fuller-Burmeister estimator relative to that of Hartley's using real data. His emperical study shows that the estimator given by Fuller-Burmeister is more efficient. This result is to be expected since Fuller-Burmeister estimate uses better estimates of post-strata sizes than given by Hartley. This is relatively a new area in survey sampling and further research in this area is needed. I would also like to mention that recently, Hartley (1974) gave a more general theory of multiple frame surveys.

The paper by Richard K. Burdick and Robert L. Sielken is an useful contribution to the new estimation techniques in finite population sampling developed by Royall. Professor Royall looks at the estimation problem in sampling as a problem of prediction for un-sampled units and uses linear least squares prediction method. One would like to see how the exact confidence intervals obtained by Burdick and Sielken compare with the exact confidence intervals that may be obtained using classical method of estimation under the same super population model.

Survey statisticians design complex sampling plans appropriate for estimation of parameters like population mean, total or ratios. Social scientists use data collected from surveys for research problems dealing with inter-relations of different variables. They often use statistical packages for analysis of survey data assuming such data as a random sample from an infinite population. This raises the question of design effect. Kish and Frankel have made extensive emperical studies of design effects. Campbell's paper is perhaps the first paper that deals with the theoretical study of design effect. Campbell provides the theoretical evidence to support Kish and Frankel's emperical results that the design effect for higher order statistics like regression estimates is generally less than the design effects for first order statistics like means.

The critical analysis of survey data is often done using methods appropriate for random samples from normal population because computer programs for data analysis geared to complex survey designs are generally not available. We hope that the organizations like the International Association of Survey Statisticians, Bureau of Census, Statistics-Canada and Survey Research Centers will develop statistical packages for critical analysis of survey data.

REFERENCES

Chakrabarty, R. P. and J. N. K. Rao (1967). "The Bias and Stability of Jackknife Variance Estimator in Ratio Estimation." Proceedings of American Statistical Association (Social Statistics Section). 326-331.

Hartley, H. O. (1974). "Multiple Frame Methodology and Selected Applications." SANKHYA, C36, 99-118.

Rao, P. S. R. S. and J. N. K. Rao (1971). "Small Sample Results for Ratio Estimators." Biometrika 58, 625-630.

INTERSTATE DIFFERENCES IN EDUCATION AND INCOME DISTRIBUTION

Joe Won Lee, Howard University and Institute for Policy Analyses

Introduction

During the last decade and a half the national economy has achieved a continuous growth despite its frequent ups and downs. However, the growth and improvement have not been even among different parts of the nation.

This paper attempts to analyze the redistributional patterns of the growth and improvement in income and education between the different parts (states) in the Union. The analyses are based on the 1960 and 1970 censuses of population and the 1976 Survey of Income and Education (SIE). Although I am well aware that SIE data are not strictly comparable with the two decennial census data, I have tried to incorporate the data in the analysis in order to put the findings in a current perspective.

The second objective of this paper is to analyse the relationship between parents' educational attainments, income levels, and pupils' (school-age children, 5-17 years) educational achievements. This analysis is based on the cross-sectional regression analysis taking the state data as observation units.

Educational Attainment Trend Among The States

A frequent measure of educational attainments of the population in an area has been represented by the median number of "school years completed." During the 1960's the general level of education for the U.S. adult population rose by 1.2 years in terms of their school years completed from 10.6 years in 1960 to 11.8 years in 1970.

It is interesting to note that the initially lower areas, especially the Southeast and the Southwest, made relatively faster increases in their educational attainments. The states in the Mountain and Far West regions enjoyed their highest (i.e., 11.7 years of school completed in 1960), but their rate of improvement was not as fast as the low attainment states. However, these Western states in general still enjoy their highest position.

Texas and North Dakota made an extremely rapid improvement; i.e., 1.9 and 1.8 years improvement, respectively. However, it must be noted that these two states started at the lowest level in their respective regions. In other words, these two states made the fastest improvement in the nation during the period, but they were still at the relativelylow side within their respective regions.

The forces which are responsible for interstate differences in educational attainments seem to be both internal and external. On one hand, there is a political force in each state to push its educational performance to the regional average (this force may be termed interstate competition.) The second important force is operating at the federal level, which attempts to equalize education attainments across the country.

The change between 1970 and 1976 showed a somewhat different trend in the interstate differences in the adults' education attainment levels. That is to say, the indication is that during the more recent period, (i.e., 1970-76 period) the interstate differences widened slightly. This assertion can be demonstrated in terms of coefficient of variation, as follows:

Coefficient of Variation
(Standard deviation/mean)

1960	1970	1976
8.3%	4.8%	11.8%

This recent trend seems attributable to the shift in interstate migration patterns, i.e., significant out-migration away from the nation's large metropolitan areas, especially toward good climate states.

Relative strength in terms of the trend in adults' educational attainments is shown for many states in the Plains, Mountain and Far West regions; especially, Minnesota, Iowa, Nebraska, Montana, Wyoming, Utah, Colorado, Washington, Oregon, California, Nevada and Alaska. The Southeast region generally has not changed its relative position yet. A noticeable improvement in the 1970-76 period has been witnessed by Arizona.

The observed relative strength among states seems due to industrial structure, high rate of economic growth and demographic composition of respective states.

Income or Earnings Trend Among the States

An average person earned $2,668 and $3,436 in 1959 and 1969, respectively. However, interstate differences were enormous. In 1959, an average person in Mississippi earned only $1,204, while an average person in New Jersey earned $3,641, which amounts to three times the average of the Mississippians.

By 1969 the interstate differences in earnings have diminished in relative sense. The average Mississippian worker earned $2,614 in 1969, while the counterpart in Alaska earned $5,351 in the same year. In the relative sense the Alaskan's earning was approximately twice that of the Mississippian. However, in absolute dollar terms the differences between the lowest and the highest states widened. That is, the differences were $2,437 and $2,737, in 1959 and 1969, respectively.

At this moment we have not obtained comparable data (median earnings by persons) from the 1976 SIE tabulation. However, I have chosen a proxy from the SIE tabulation in order to examine the general trend in the interstate differences in income and/or earnings. The SIE tabulation provides for each state median family income for persons 25 years or older. In 1975 an average family in Arkansas had income of $9,649, while the counterpart in Alaska had $23,206 in the same year. This abnormally high median family income in Alaska is due to the boom attributable to the pipe line construction. Let us take the next highest state, which is Hawaii whose median family income was $18,614 in 1975. The Hawaiian median family income is twice that of Arkansas.

As can be seen in Table I, the interstate differences in income and/or earnings levels are being diminished over time. However, there is some persistent force at work, which keeps the income levels high in Middle Atlantic, Great Lakes and Far West regions. These three regions are highly industrialized regions. Connecticut and Massachusets in New England also belong to these highly industrialized areas. The relatively cheap labor and weak labor unions have helped

the southern states to expand their productive capacity more rapidly than the rest of the country. However, the industrial mix in these less industrialized states are not favorable in the sense that their industrial structure is heavily concentrated in those industries which are not expanding rapidly at the national level (such as textile industries.) On the other hand, the aforementioned highly industrialized states have to pay relatively high wages. Therefore, these states are at a disadvantage in competition with the southern, less industrialized states. However, the northern industrialized states have favorable industrial structure in the sense that these states contain those industries whose capacity is expanding more rapidly at the national level (such as service-oriented industries.)

As mentioned above, despite the persistent forces, the general trend which narrows the income gaps between states has been reinforced by deliberate public policies as well as the more or less natural economic forces stemming from the expanding markets in the presently less industrialized states. The narrowing trend of the interstate income gaps (in relative sense) can be demonstrated in terms of the coefficients of variation (standard deviation divided by the national mean income level for each of the observed years), as follows:

1960	1970	1976
22.0%	17.2%	16.3%

Educational Attainments and Income Levels

It may be assumed that the higher the educational attainments in an area the higher the earnings level would be. This hypothesis has been tested utilizing the 1960, 1970, and 1976 data. The following three equations show the relationship:

$$Y_{60} = -2381.1 + 474.9E_{60} \quad . \; . \; . \; (1)$$
$$Y_{70} = -3744.8 + 651.8E_{70} \quad . \; . \; . \; (2)$$
$$Y_{76} = 1730.8 + 187.3E_{76} \quad . \; . \; . \; (3)$$

where Y_{60} and Y_{70} represent median earnings in 1960 and 1970 respectively. Y_{76} stands for median family income, as reported in SIE. E_{60} and E_{70} stand for median number of school years completed; E_{76} stands for percentage of population who have completed high school education as of 1976. R squares were .50, .31 and .40 for 1960, 1970 and 1976, respectively. Although they are not very high, the relationships are significant at 95 percent confidence level. The results seem to indicate that income level (or earnings) of individuals are only partially determined by their educational attainments (in terms of number of years spent for formal education.) Beside the formal education, there seem to be a host of factors influencing the earnings level of workers. These might include the individual's ability to succeed, his training on the job, amount of wealth accumulated or inherited or both, industrial characteristics, and quality of the education in different states. It must be noted that this regression model is based on state observation not individual person's observations. If we take a sample of individuals' educational attainments and their earnings as observation units, the correlation between the two indicators may be much higher than the aforementioned results.

Pupils' Educational Achievements

There are many ways of measuring pupils' educational achievements. In this analysis, however, I have taken two measures of pupils' achievement levels by state. One is their enrollment rates. In other words, the enrollment rate for each state has been derived by those children, aged 5-17 inclusive, who are enrolled divided by total number of the school-age children in each state. The second measure may be termed "deficient rate." This rate has been computed by identifying the modal grade for each age of children. For example, a 7 year old child is normally supposed to be in the second grade. If he or she is enrolled in that grade, he or she is given zero percent credit. If enrolled in the 3rd grade, the child is given one point (100 percent) positive credit. If the child is enrolled in the first grade, he is given negative one point (minus 100 percent) credit. In this way, I computed a "weighted" average "deficient rate" by age and by sex. And, finally, I derived an overall "weighted" average deficient rate (weighted by number of children in each age cohort.) [1]

At the national level, the enrollment rate rose from 92.0 percent in 1960 to 93.3 percent in 1970. And it rose to 95.4 percent by 1976, when the SIE survey was taken. In the following the mean (unweighted average) enrollment rate, standard deviation, coefficient of variation (standard deviation/mean), minimum rate and maximum rate are presented.

Pupils' Enrollment Rates
(all in percent)

	1960	1970	1976
Mean	92.0	93.3	95.4
Standard Deviation	1.6	1.8	1.2
Coefficient of Variation	17.7	19.6	12.7
Minimum	87.5	87.5	92.2
Maximum	95.1	96.2	97.4

A noticeable improvement has been made during the 1970-76 period. And the interstate differences have been narrowed significantly during the recent period, while the interstate differences widened generally in the 1960's.

Another noticeable observation is that not only the interstate (or interregional) gaps have been narrowed in terms of enrollment rate of school age children, the improvement of the southern states (both in Southeast and Southwest) has been especially pronounced. In the south, especially the states of Virginia and South Carolina have achieved the most pronounced improvement in their enrollment rates. South Carolina was 5 percent below the national average in 1960, but by 1976 South Carolina reached the national average. Virginia's enrollment rate in 1960 was 3.3 percent below the national norm, but she exceeded the national average by one percentage point in 1976. Thus, although all states competed for excellence in their education attainments, their relative successes varied, depending upon a host of factors, such as respective states' priority ordering, relative economic performance, which in turn has been affected not only by states' own efforts but also federal policies.

A similar observation can be made in terms of

interstate differences in grade "deficiency rates". The following table presents an overall picture relative to the grade "deficiency rates". As the table indicates, the interstate differences are gradually narrowing over time.

	Deficiency Rates (%)		
	1960	1970	1976
Mean	-49.7	-51.5	-65.6
Standard Deviation	13.5	8.8	8.4
Coefficient of Variation	27.2	19.6	12.8
Minimum	-91.0	-73.0	-80.0
Maximum	-28.0	-32.0	-47.0

The table indicates that children's grades in which they are actually enrolled are approximately one half year below the grades in which they are supposed to be enrolled. It must be pointed out that the slightly higher "deficiency rate" for 1976 is somewhat exaggerated, because the available SIE tabulation did not include those children who are enrolled above their modal grade level. Moreover, the SIE was conducted during the months of April, May, and June 1976, while the decennial census data are recorded as of April. If this factor is taken into account, -65.6 percent for 1976 will be reduced to 57.4 percent. Thus, if the aforementioned two factors are combined, it is probable that the true mean deficiency rate for 1976 would be about the same level as those for 1960 and 1970.[2]

Parents' Education and Childrens' Education

I have attempted to quantify the effect of parents' educational attainments on childrens' education, by means of cross-sectional regression, utilizing state data for 1960, 1970 and 1976. Here the dependent variable represents childrens' enrollment rates and the explanatory variable is the median number of school years completed by the population 14 years and over in each state. The results of the regression analysis using children's grade "deficiency rate" as the dependent variable, have been quite parallel to the results shown here.

$$ENR60 = \underset{(37.0)}{79.027} + \underset{(6.1)}{1.216}ED60 \quad . \; . \; . \; (4)$$
$$R^2 = 0.42$$

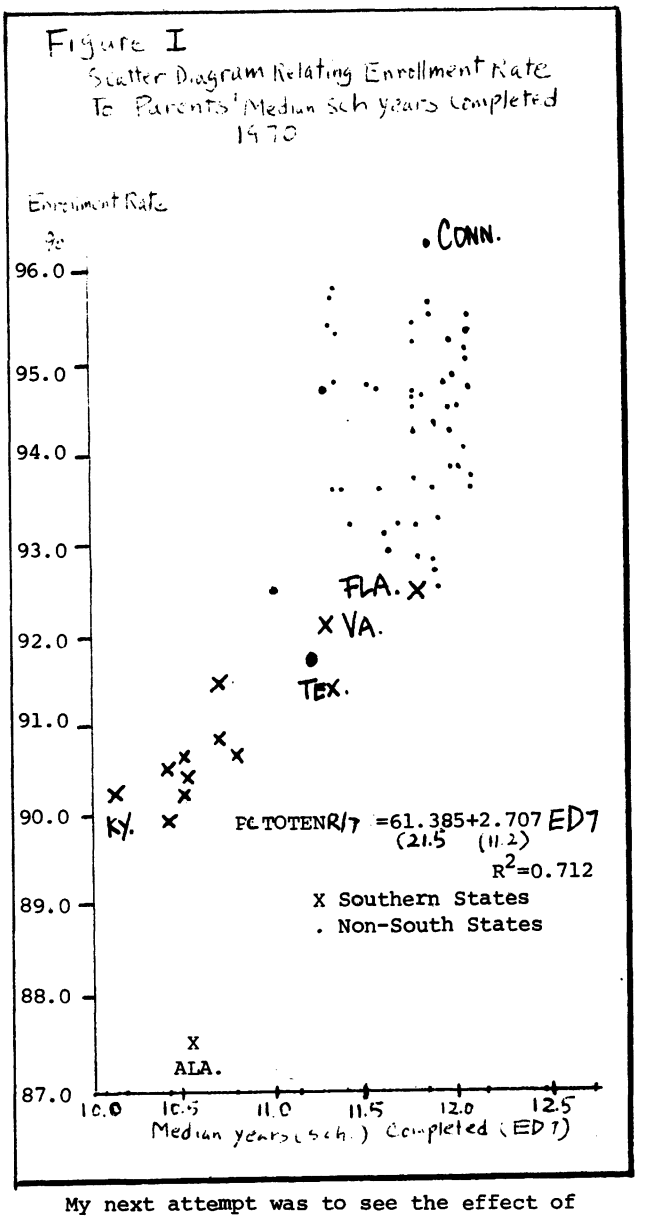
$$ENR70 = \underset{(21.5)}{61.385} + \underset{(11.2)}{2.707}ED70 \quad . \; . \; . \; (5)$$
$$R^2 = 0.71$$

$$ENR76 = \underset{(18.7)}{91.445} + \underset{(3.0)}{0.062}ED76 \quad . \; . \; . \; (6)$$
$$R^2 = 0.15$$

The independent variable for 1976 (ED76) is percentage of population (25 years and over) who completed high school education. Therefore, the 1976 result cannot be compared with the two previous years.

An interesting result is that the relation between childrens' enrollment rates and adults' educational attainment levels has become stronger in the 1960's (1970 census) than in the 1950's (1960 census). This may be attributable to the relatively prosperous economic conditions during 1960's. That decade witnessed a rapid growth of college enrollment, which endorsed indirectly the utility of education.

Figure I shows a scatter diagram of the 1970 data. As can be seen in the diagram, if the abnormal value of Alabama had been eliminated, a semi-logarithmic specification should have improved the relationship substantially.



My next attempt was to see the effect of parents' educational attainments on enrollment rates for children of different ages. In Figure II the horizontal axis represents children's age from 5 to 17. The vertical axis represents the R square value resulting from the regression relating the enrollment rate of children of a particular age (e.g., 5 years old) with adults' educational attainments (the median number of school years completed).

The figure suggests that parents with high educational attainment levels seem to be more successful in keeping more children in school. The lines (2) and (4) suggest this assertion.

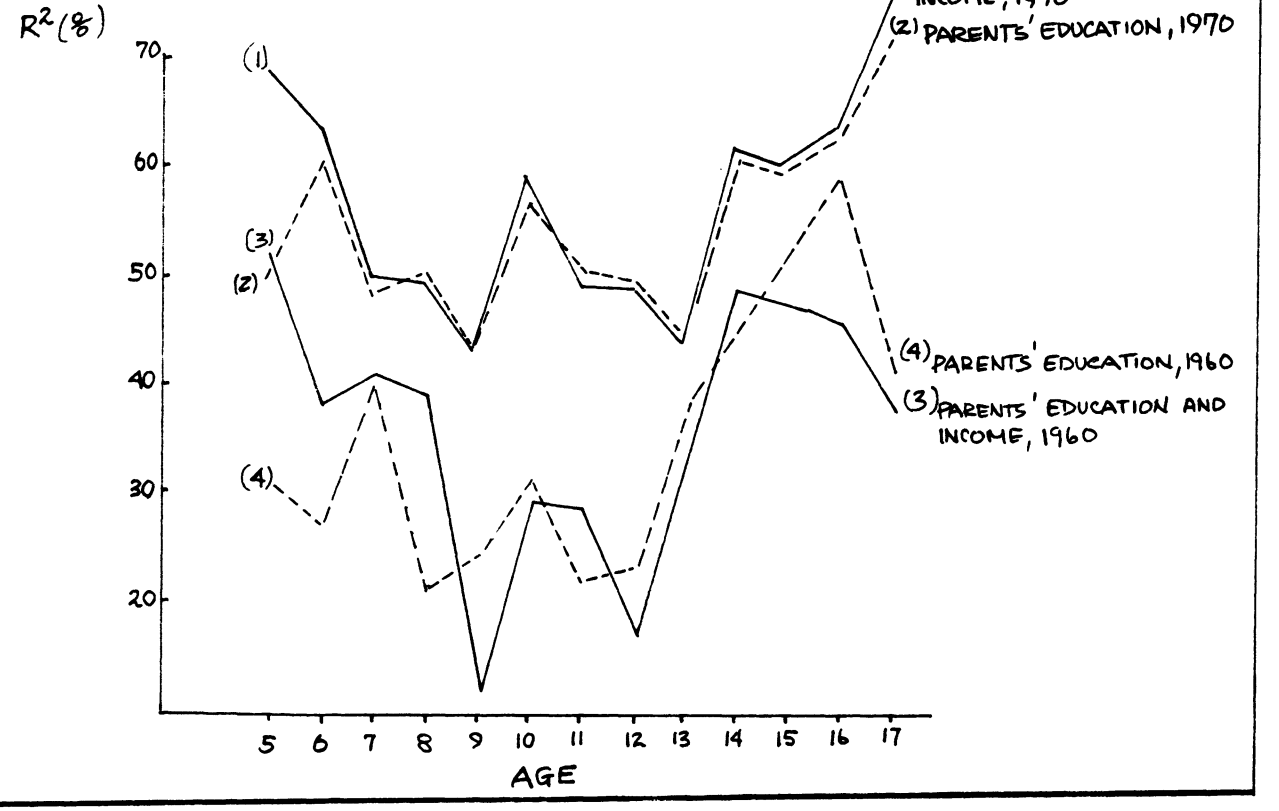
One more interesting observation can be made. When parents' income is included as an explanatory variable, in addition to their educational attainment levels, the income effect is extremely significant for 5 and 6 year old children. But the income effect on the older children is negligible as lines (1) and (3) suggest in Figure II.

Conclusions

Interstate differences in all aspects of income and education as examined above, have been

FIGURE II

CORRELATION BETWEEN CHILDREN ENROLLMENT RATES (BY AGE) AND PARENTS' EDUCATION AND INCOME LEVEL



narrowing substantially during the past 15 years; especially pronounced improvement has been achieved during the 1970's.

Interstate differences in income level have been explained at least partially, by the interstate differences in educational attainment levels. As noted earlier, interstate differences in income and/or earnings are attributable to many factors, in addition to the differences in educational attainments of working people. However, the most important point is that the income elasticity with respect to the educational attainments is much greater than unity -- 1.89 and 1.95 for 1960 and 1970, respectively. This indicates that one percent improvement in education (in terms of educational period) will bring about approximately two percent increase in income and/or earnings.

As Figure II indicates, influences of parents' educational level has been significant on childrens' enrollment rates during the 1960-70 period. The influences are stronger in both ends -- youngest children, i.e., 5 and 6 years old, and the children who are completing their high school education. The "U" shape, which is applicable both to the 1960 and 1970 data, seems to reflect the normal human nature. Parents try to put and keep their children in schools when their children reach the school age. And when their children approach the end of high school education, parents with higher income and educational attainment also try harder to make their children complete the high school education, possibly looking forward to their entrance into college.

In addition to parents' education, their income level also exerts a significant influence on both ends of school-age childrens' enrollment rates, since the parents have the means to do so.

The higher correlations for 1970 than 1960 seem to be attributable to two factors. First, the 1960's "new frontier" and "great society" concepts encouraged people to recognize the high return to investment in human capital. Secondly, the prosperity prevailing in the 1960's seems to have reinforced the momentum.

The more rapidly narrowing gap (observed during the 1970-76 period) in the interstate differences in income and education signifies the time lag involved in the long process of policy pronouncement, legislative enactment and administrative implementation in our political process.

FOOTNOTES

[1] The term "deficient rate" has a somewhat misleading connotation. However, as expected, the percentage credit for each state, by age and sex has turned out to be negative percentage. This is mainly due to the registration rule in each

TABLE I School Enrollment Rate, Deficiency Rate, Median School Years Completed and Median Earnings, by State, 1960, 1970 and 1976

	1960				1970				1976			
	1	2	3	4	1	2	3	4	I	II	III	IV
Maine	92.1	.60	10.7	2.3	93.1	.61	12.0	3.5	95.8	64.9	.79	11.6
New Hampshire	91.5	.44	10.7	2.7	92.4	.46	12.1	4.1	94.6	68.7	.66	14.0
Vermont	92.3	.46	10.7	2.3	92.7	.44	12.1	3.7	95.6	67.2	.57	11.9
Massachusetts	92.1	.32	11.3	3.0	95.1	.41	12.2	4.4	97.4	69.8	.58	15.4
Rhode Island	91.0	.42	10.1	2.7	94.6	.49	11.5	3.9	95.7	58.5	.54	14.3
Connecticut	93.7	.36	10.8	3.6	96.2	.42	12.1	5.1	97.4	67.7	.60	16.3
New York	93.3	.28	10.8	3.4	94.5	.38	12.0	4.9	96.8	63.4	.54	14.8
New Jersey	93.3	.37	10.6	3.6	95.3	.44	12.0	5.0	96.5	63.4	.62	16.6
Pennsylvania	92.0	.30	10.4	3.0	94.6	.36	11.8	4.3	96.1	60.5	.53	13.9
Delaware	91.1	.41	10.9	3.2	94.2	.43	12.1	4.5	92.2	66.9	.58	16.0
Maryland	91.3	.39	10.5	3.3	94.4	.44	12.1	5.0	96.7	67.4	.54	17.5
District of Columbia	91.0	.54	11.6	3.3	93.5	.60	12.1	5.0	96.1	63.7	.65	13.4
Ohio	92.5	.48	10.8	3.3	94.5	.56	12.0	4.5	95.7	64.9	.71	14.8
Indiana	91.7	.47	10.7	3.1	93.6	.60	12.0	4.4	94.6	63.8	.80	14.3
Illinois	92.4	.35	10.6	3.5	94.6	.41	12.0	4.9	96.2	63.0	.56	15.7
Michigan	94.4	.40	10.8	3.4	95.1	.45	12.0	4.9	96.9	65.5	.61	15.4
Wisconsin	93.7	.42	10.6	2.9	95.5	.48	12.1	4.0	97.4	67.4	.66	14.6
Virginia	88.9	.70	10.1	2.4	92.1	.68	11.5	4.0	96.5	61.8	.77	14.5
West Virginia	90.0	.52	9.1	2.3	87.5	.58	10.7	3.3	93.4	49.5	.72	11.6
Kentucky	88.6	.60	8.9	2.0	90.2	.49	10.3	3.3	92.8	49.5	.63	11.0
Tennessee	89.7	.58	9.2	1.9	90.4	.55	10.7	3.4	94.8	51.2	.65	11.1
North Carolina	90.0	.66	9.5	1.9	90.8	.60	10.9	3.4	93.0	51.9	.71	11.8
South Carolina	87.5	.74	9.2	1.8	90.2	.66	10.7	3.4	95.6	53.3	.72	12.2
Georgia	90.8	.61	9.5	1.9	90.6	.51	11.0	3.6	94.4	55.4	.62	12.2
Florida	91.5	.50	10.7	2.4	92.8	.45	12.0	3.6	97.0	62.6	.52	11.7
Alabama	90.3	.70	9.5	1.8	90.6	.64	10.7	3.2	93.1	53.2	.73	11.5
Mississippi	91.0	.91	9.1	1.2	89.9	.63	10.6	2.6	94.9	49.1	.58	9.8
Louisiana	91.3	.60	9.3	2.0	91.4	.55	10.9	3.3	95.5	53.9	.58	12.5
Arkansas	90.7	.62	9.2	1.5	90.5	.63	10.6	2.8	93.8	53.2	.74	9.6
Minnesota	93.8	.33	10.9	2.6	95.4	.49	12.1	3.8	96.8	69.3	.74	14.1
Iowa	94.4	.47	11.2	2.5	94.1	.58	12.1	3.6	96.0	69.4	.77	14.0
Missouri	91.3	.42	10.1	2.5	93.5	.57	11.6	3.7	95.2	61.2	.66	12.6
North Dakota	92.8	.36	10.0	2.2	93.5	.47	11.8	3.1	94.0	62.0	.68	13.5
South Dakota	93.0	.46	10.5	2.0	94.1	.51	12.0	2.9	95.8	64.1	.69	11.7
Nebraska	94.4	.45	11.5	2.5	94.7	.53	12.2	3.5	95.2	70.9	.68	13.8
Kansas	92.8	.35	11.5	2.6	94.4	.44	12.2	3.6	95.5	69.7	.72	13.4
Oklahoma	91.9	.50	10.6	2.2	93.1	.51	11.9	3.3	95.7	61.7	.67	12.0
Texas	89.8	.76	9.5	2.4	91.7	.73	11.4	3.7	95.2	61.6	.81	12.6
Arizona	90.8	.56	10.9	2.8	92.6	.47	12.1	4.0	96.1	70.7	.56	13.6
New Mexico	91.0	.61	10.9	2.8	92.7	.50	12.0	3.6	95.4	63.3	.62	12.0
Montana	92.5	.52	11.3	2.7	93.7	.57	12.2	3.5	95.3	69.7	.72	13.3
Idaho	93.1	.54	11.3	2.5	92.7	.56	12.1	3.3	94.6	68.9	.76	13.0
Wyoming	92.5	.58	11.8	3.0	93.7	.61	12.2	3.7	94.7	72.6	.80	14.6
Utah	92.6	.47	12.1	3.0	95.4	.49	12.3	3.6	95.9	78.7	.67	14.8
Colorado	92.8	.44	11.9	2.9	94.9	.52	12.3	3.9	95.7	76.7	.75	14.8
Washington	93.5	.52	11.8	3.1	95.0	.56	12.3	4.3	95.1	74.7	.72	14.5
Oregon	93.8	.44	11.5	2.9	94.4	.49	12.2	3.8	95.3	73.5	.63	13.5
California	94.2	.31	11.9	3.5	95.2	.40	12.3	4.7	96.7	72.4	.57	14.7
Nevada	94.0	.44	12.0	3.6	93.6	.40	12.3	5.2	95.5	74.3	.61	14.7
Alaska	90.3	.73	12.0	3.3	93.5	.63	12.3	5.4	94.4	78.7	.71	23.2
Hawaii	95.1	.31	11.2	3.1	94.6	.32	12.3	4.8	97.0	68.6	.47	18.6
Mean	91.9	.50	10.6	2.7	93.2	.52	11.7	3.9	95.4	64.4	.66	13.8
S.D.	1.6	.14	0.9	0.6	1.8	.09	0.6	0.7	1.2	7.6	.08	2.3
C.V.	1.8	.27%	8.3	22%	2.0	20%	4.8	17%	1.3	11.8	13%	16.3
Min.	87.5	.28	8.9	1.2	87.5	.32	10.3	2.6	92.2	49.1	.47	9.6
Max.	95.1	.91	12.1	3.6	96.2	.73	12.3	5.4	97.4	78.7	.80	23.2

(1) Enrollment Rate (%)
(2) Grade "Deficiency Rate" (year)
(3) Median No. of School Years Completed
(4) Median Earnings (thousand $)

(I) Enrollment Rate (%)
(II) Percent of Population (25+) comp.High Sch
(III) Grade "Deficiency Rate" (year)
(IV) Median Family Income

state, which provides that unless a child has reached a certain biological age, he can not be enrolled.

[2]The author is grateful to Mr. Gerald Kahn at NCES/DHEW for his comment and suggestion concerning the deficiency rates.

REFERENCES

(1) Al-Samarrie, A., and H.P. Miller, "State Differentials in Income Concentration," American Economic Review, March 1967,pp.59-72.

(2) Conlisk, J. "Determinants of School Enroll-Ment and School Performance,"Journal of Human Resources, Spring, 1969.

(3) _____: "A Bit of Evidence on Income-Education-Ability Interrelation," Journal of Human Resources, Summer, 1971, pp.358-62.

(4) Hines, F.; Tweeten, L. and Redfern, M. "Social and Private Rates of Return to Investment in Schooling, by Race-Sex Groups and Regions," Journal of Human Resources, Summer 1970, pp.318-40.

(5) Johnson, G.E. and Stafford, F.P. "Social Returns to Quantity and Quality of Schooling,"Journal of Human Resources, Spring, 1973, pp.139-55.

(6) Liebenberg, M. "Nomographic Interpolation of Income Size Distributions,"The Review of Economics and Statistics,August 1956, pp. 258-72.

(7) Masters, S.H. "The Effect of Family Income on Children's Education: Some Findings on Inequality of Opportunity," Journal of Human Resources, Spring, 1969.

(8) Pryor, Frederic "Simulation of the Impact of Social and Economic Institutions on the Size Distribution of Income and Wealth," The American Economic Review, March 1973, pp. 50-72.

(9) Tolley, G.S. and E. Olson. "The Interdependence between Income and Education," Journal of Political Economy, May/June 1971, pp. 460-80.

(10) U.S. Department of Commerce, Bureau of the Census: United States Census of Population (Detailed Characteristics), 1960.

(11) _____: 1970 Census of Population (Detailed Characteristics)

(12) _____: Survey of Income and Education, 1976 (Unpublished tabulation).

TRENDS IN COLLEGE ENROLLMENT BY FAMILY INCOME FOR REGIONS OF THE UNITED STATES

Larry E. Suter, Bureau of the Census

Many studies of the determinants of college participation have established that families of higher educational and income levels have been more likely than those of families with lower socioeconomic backgrounds to send their offspring to college, regardless of their ability levels (Condition of Education, 1977 edition). This relationship persists despite increasing amounts of federal and state funding available to aspiring college students in the 1960's and 70's. Higher income of families is related to several factors which affect the rate of college enrollment of college age members. For example, high income families are probably more likely to provide a home environment which encourages reading, the development of intellectual skills, and the selection of occupations which require college education. In general, high income families are probably more likely to develop the attitudes and values in their offspring that make a college education seem necessary, as well as supply the money to support a student through the expensive college years (see Jencks, Inequality, p. 138). Attempts to separate the determinants of college attendance have not successfully divided the direct effect of family income from personal motivation and other factors that determine college entrance. However, all studies agree that level of family income in the family of origin influences both the amount and the quality of higher education received by family members. To the extent that money is a direct factor in the decision to attend college, the opportunity to attend college would vary within the United States as the cost of living and the costs of education varies between regions. The purpose of this paper is to assess the level of differences in college enrollment rates between regions by income level of the family.

Variation in college level participation by region and changes in participation rates by income level during the 6 years since 1970 may illustrate the influence (or lack of it) of governmental programs on equalizing college participation by income groupings. Federal programs provide assistance to college students unadjusted for variations in regional cost of living. The regional patterns may also reflect, of course, cultural and opportunity differences between region, as well as the cost of attending college. This paper will not be able to supply all the necessary evidence to separate each of the causative effects (money, motivation, values, or home encouragement) known to be determinants of college participation in each region, but these data will provide a solid basis upon which to discuss the actual meaning of differences between regions.

This paper will present some new statistical data from the October Current Population Surveys for 1970 and 1976 on college participation rates of persons 18 to 24 years old by family income level in four U.S. regions. The family income measure from the CPS has been adjusted to 1967 dollars to preserve the relationship of income categories to each other (the BLS cost of living index for the U.S. was used as the deflator).

Before examining the measures of college participation, two characteristics of the population represented by the survey design which may affect generalization of the results should be outlined. The October CPS is representative of only the civilian population excluding inmates of institutions. Although the universe represents nearly all women of college age, there have been major changes in the CPS coverage for men following the decline in the Armed Forces population since 1970. The proportion of all men 18 to 24 years old not represented by this universe because they were in the Armed Forces was 15 percent in 1970 and decreased to 8 percent in 1976 (for men 20 and 21 years old, the decline was a dramatic 23 percent to 9 percent). The effect of the decrease in proportion of men in the Armed Forces has been to reduce the proportion of civilian men enrolled in college. Thus any analysis of changes in enrollment rates for men during the period which encompasses the Vietnam War Era must consider the possible confounding effects of the changing population base. The changes in the proportion of men who were inmates of institutions has not been sufficient to affect any of the enrollment rates and can be safely ignored.

Another issue in the use of CPS data is the correct specification of the universe reporting family income, and thus, a proper accounting for the amount of income available to the college age person. College students living away from home who are considered by their family as household members (i.e. that the sample address is the usual place of residence for that person) are reported as members of their parent's family and thus it is their parent's income which is reported as available to them. However, the married 18- to 24-year-old or the student who lives away from home permanently (even though both of these may receive financial assistance from their parents) report themselves in their own household and thus report only that income received by members of this household. A "dependent family member" is defined for purposes of the analysis of the October CPS as an 18- to 24-year-old relative of the household head (except the wife). In effect, of course, most dependent family members are the sons or daughters of a family head, although some may be other relatives living temporarily in a household. Dependent family members accounted for 52 percent of all 18- to 24-year-olds in 1975, which is 59 percent of men and 47 percent of women (see Current Population Reports, P-20, No. 303). Married persons account for 27 percent of men and 12 percent of women; and primary individuals account for 14 percent of men and 12 percent of women in this age. Thus, the family income of the primary family of a large portion of college students is not reported in the October CPS. To avoid confusing reported household income with the true level of income available to potential college students, the analysis will be restricted to dependent family members; for that group there is a stronger reason to believe that the family income

of parents reported in the survey is the source of financial support for their college attendance.

Data Analysis

Graduation from high school is a necessary step to college attendance; therefore the regional and family income differences in high school graduation will be examined before turning to college enrollment. High school graduation rates for dependent family members 18 to 24 years old in the United States had reached 82 percent by 1976, only a 2 percentage point increase since 1970 according to the October Current Population Survey (table 1). The lowest level of high school graduation was in the Southeastern States (78 percent) and the highest was in the Northeastern States (86 percent).[1]

The income level of the family strongly affects the probability of high school graduation of dependent family members. The rates varied from around 57 percent at the lowest level to around 95 percent at the highest income level. However, the greater change occurs in the income levels below $10,000 (1967 dollars) than above that range. In all regions, high school graduation rates were higher for persons in families with highest income levels; in fact, the rates are very similar throughout all regions at every income level (see figure 1).

Low high school graduation rates in 1976 for the lowest income groupings are probably not directly due to the lack of money in the family since public high schools require few funds for attendance. More likely, they result from cultural factors in the family that determine the attitudes and values of its members toward education. The lack of large differences in high school graduation rates among the regions, within similar income categories, suggests that the abilities or attitudes toward completing high school are similar in all regions, but that only the lowest income families do not, or cannot, encourage their dependent members to always complete high school.

Attendance in college, on the other hand, is contingent on high school graduation and is much more likely to be dependent upon the ability of the family to support a student in college. This dependence on income is apparent in the relationship of college enrollment rates for each family income level shown in table 1 for the four regions. Unlike the slope of the lines for high school graduation in figure 1, which rise sharply in the lowest portion of the income distribution, college enrollment increases with income in almost a perfect linear fashion in each region. Regional differences in the effect of family income on college attendance is best measured by the combination of those enrolled and those who have already completed some college. Many persons, especially at the upper age ranges, have completed some college although they have remained dependent family members. Figure 2 indicates the enrollment rates by income for those enrolled in 1976 and table 1 shows the combined enrollment rates for those enrolled or who had already completed some college.

College participation rates by family income do not indicate any strong evidence of differential opportunity to attend college between the various regions of the country. In fact, the differences in participation rates, once income level of family is considered, are not as great as might be expected. There is clear evidence that dependent persons living in the West are more likely to attend college. However, even this statement is not true for every income level (the highest income level is somewhat ambiguous, see figure 2). It is likely that the large junior college system of California is responsible for the higher participation rates in that State.[2]

To some extent the cost of attending college might explain why the Western States have slightly higher participation levels than other regions. The average cost for attending college in the Western States[3] is lower than in other regions. However, the cost of living while in college is higher in that region, making the total cost of living for in-State students (about 80 percent of all students) slightly higher than in other regions. Average reported student tuition and living expenses as found in the October 1973 Current Population Survey are shown in the table below.

Table A.

Mean Student Expenses in College for Full-Time, In-State College Students by Region: October 1973

Region	Total	Living costs	Tuition	Books, transp., etc.
Northeast	$5,700	$4,200	$1,100	$400
North Central	$5,300	$4,100	$ 800	$400
South	$5,300	$4,300	$ 600	$400
West	$5,600	$4,900	$ 400	$300

Differences in cost of living between regions might also be responsible for some of the differences in college attendance rates, as suggested earlier in this paper. However, the evidence now available does not support that contention. An index of differences in cost of living in the areas of the United States is provided by the BLS Urban Family Budget (BLS release 77-369) which is an estimate of hypothetical annual family budgets for selected metropolitan areas. Assuming that the intermediate budget for a 4-person family reflects the constraints faced by families with college age members, an average of the cities was computed, assuming equal weights for cities. This somewhat crude measure shows that for the U.S. as a whole, the West is about the same as the U.S., the North Central is about 1 percent lower, and the Southern cities are about 8 percent lower than for the U.S.

Thus, the cost of tuition and expenses of college and not the cost of living has a greater apparent association with lower participation rates of middle income college age persons. Conclusions regarding the possible impact of college costs should not be too hastily accepted because not all of the factors that affect the costs of education or of living in each of the regions have been fully considered here. For example, differences in the number of public and private schools in each State and the number of students who attend college in another State, may affect the overall participation rates. Since college students are counted at the address of their

parents for the CPS, even though they may be attending college in another State, the actual costs of their schooling cannot be derived without detailed surveys of student costs.

The possibility that the college enrollment rates for persons at the lower income level in Southern States were low because of differences in enrollment rates of Whites and Blacks was examined. The overall enrollment rates for Black dependent family members is much lower than for Whites (see table 2). The income level of Blacks is also much lower than for Whites (see table 3). In fact, more than one-half of the 18- to 24-year-olds in families with incomes less than $5,000 living in the Southeastern region were Black (about 55 percent in 1976). The results show that the college enrollment of Blacks is about the same as for Whites in the same income grouping in the Southeastern region, and in the Northeast may even be higher. Thus, the income level and not the race of the family appears to be the major determinant in the participation levels of college age persons who are members of families in recent years. The fact that the participation rates for lower income persons in Southern States is especially low, while rates at higher income levels are very high, suggests that decisions other than availability of money, perhaps cultural values, are influencing the level of college participation in that region. Whether income level is so strongly related to college participation because of the implications of costs of college attendance or because of the values toward higher education in each income level cannot be adequately determined with these data, but the strength of the statistics suggest the strong importance of family regardless of region of residence or race.

This exercise in examining the relationship of family income to the college participation of its college age members has shown that only small differences exist between major regions of the country. Future studies of the reasons for variations in college attendance might be more fruitfully applied to specific State systems. Anderson and Bowman pointed out in their study of college attendance in 4 States that the variation in history of educational institutions and application of financial aid is so great between the States of the Union that a stronger understanding of variations in college attendance requires a comprehensive analysis of each State system. Information soon to be received from the Survey of Income and Education will establish, at least, whether the participation rates in college by dependent family members vary between States once level of income has been controlled by more than was found for the major regions. The conclusion of this paper must be that in all areas of the country two processes govern the level of college attendance of college age family members: the ability to complete high school and the attributes in the family that are associated with the level of income in that family. Neither the variations in costs of living nor the racial composition of region are strong determinants of participation rates within equivalent income categories. This study can provide only weak evidence that those areas with lowest participation rates are those in which the costs of public college may be higher.

FOOTNOTES

1/ See table 1 for a definition of "Southeastern" and "Northeastern" regions. In this table the Census Bureau definition of regions were not used. Instead, the combination of States used by the Bureau of Economic Analysis and the National Assessment of Educational Progress was employed in this table. Oklahoma and Texas are added to the West; Delaware, D.C., and Maryland are in the Northeast.

2/ Although, as Anderson, Bowman, and Tinto point out, the direction of the cause cannot be easily inferred. Possibly California's junior college system is not the cause of large college enrollments, but is a result of demand for higher education by the State population. C. Arnold Anderson, Mary Jean Bowman, and Vincent Tinto, Where Colleges Are and Who Attends: Effects of Accessibility on College Attendance. McGraw-Hill, New York: 1972. Either way, the level of higher education received by residents of California is possibly increased by the general availability of a low cost educational system.

3/ The regions in this table are comparable to the 1970 census definitions.

REFERENCES

C. Arnold Anderson, Mary Jean Bowman, and Vincent Where Colleges Are and Who Attends: Effects of Accessibility on College Attendance. McGraw-Hill, New York: 1972.

Christopher Jencks, et al., Inequality: A Reassessment of the Effect of Family and Schooling in America, Basic Books, Inc., New York: 1972.

U.S. Bureau of the Census, Current Population Reports, Series P-20, No. 303, "School Enrollment--Social and Economic Characteristics of Students: October 1975," Washington, D.C.

Table 1

College Enrollment of Primary Family Members 18 to 24 Years Old by Family Income: October 1976

(Numbers in thousands. Family income in 1967 constant dollars)

Year, region, and college participation	Total 18 to 24 years old	Under $3,000	$3,000 to $4,999	$5,000 to $7,499	$7,500 to $9,999	$10,000 to $14,999	$15,000 or more
1976							
U.S. Total	14,222	1,425	1,629	2,325	2,312	2,872	2,176
Percent:							
High school graduate	82.0	56.8	69.1	81.1	86.4	88.9	95.5
Enrolled in college	38.8	20.4	24.2	32.2	40.4	47.5	58.2
Not enrolled	43.2	36.4	44.9	49.0	46.1	41.4	37.3
Attended college 1+ yrs.	13.0	6.7	8.1	12.8	13.9	14.3	18.0
Northeast	4,160	293	433	720	708	859	669
Percent:							
High school graduate	84.8	56.0	75.5	84.9	87.6	89.5	96.6
Enrolled in college	39.0	20.1	28.9	31.3	40.1	46.9	54.4
Not enrolled	45.8	35.8	46.7	53.6	47.5	42.6	42.2
Attended college 1+ yrs.	14.8	9.9	9.0	12.9	14.5	15.7	21.2
Southeast	3,100	550	440	542	430	480	404
Percent:							
High school graduate	76.4	55.8	64.3	77.7	85.1	87.5	93.1
Enrolled in college	34.9	18.9	19.8	30.8	41.9	49.0	59.7
Not enrolled	41.4	36.9	44.5	46.9	43.3	38.5	33.4
Attended college 1+ yrs.	10.9	4.0	4.8	10.5	12.3	12.7	18.8
Central	3,763	239	385	590	660	828	650
Percent:							
High school graduate	84.4	64.4	68.1	80.8	86.4	89.0	96.5
Enrolled in college	39.0	24.7	21.0	32.9	37.6	43.8	58.3
Not enrolled	45.4	39.7	47.0	48.0	48.8	45.2	38.2
Attended college 1+ yrs.	12.4	7.1	8.6	13.4	12.6	12.4	16.5
West	3,199	344	372	474	512	655	504
Percent:							
High school graduate	80.8	53.5	67.5	80.8	86.1	88.7	94.6
Enrolled in college	41.9	19.8	26.9	34.6	43.2	50.8	61.9
Not enrolled	38.9	33.7	40.6	46.2	43.0	37.9	32.7
Attended college 1+ yrs.	13.5	7.0	10.5	14.8	16.8	15.6	15.3

Note: The total includes persons not reporting on family income, which is not shown separately. States are combined into the set of regions defined for use by the National Assessment of Educational Progress Program. The following States make up the National Assessment regions used in Table 1:

NORTHEAST: Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont.

SOUTHEAST: Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia, West Virginia.

CENTRAL: Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin.

WEST: Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oklahoma, Texas, Utah, Washington, Wyoming.

Table 2

Percent of 18- to 24-Year-Old Dependent Family Members Enrolled in College by Income, Region, and Race: 1970 and 1976

(Income in 1967 dollars)

Region and race	Persons (thousands)	Total percent enrolled	Under $3,000		$3,000 to $4,999		$5,000 to $7,499		$7,500 to $9,999		$10,000 or more	
			1970	1976	1970	1976	1970	1976	1970	1976	1970	1976
United States	(1976)	(1976)										
White	11,834	40.4	19.9	20.0	26.1	24.9	33.0	31.4	40.6	39.9	54.0	50.4
Black	1,990	28.1	15.2	20.2	19.4	22.1	25.5	33.9	31.4	40.3	35.9	53.2
Northeast												
White	3,611	39.4	19.7	17.0	27.2	28.1	32.7	30.7	37.1	39.1	49.5	49.8
Black	452	34.8	11.5	26.3	15.0	27.1	20.0	35.9	24.5	46.2	29.1	57.7
Southeast												
White	2,163	39.6	16.1	19.4	23.1	19.5	27.9	30.7	39.7	42.0	61.8	54.5
Black	885	22.7	15.1	17.4	17.8	20.3	26.7	30.2	42.1	38.6	21.7	46.5
Central												
White	3,344	40.3	25.4	24.5	27.0	23.9	34.4	32.2	39.9	37.7	51.0	50.2
Black	377	29.6	17.5	19.7	23.4	16.0	26.7	33.3	25.6	33.7	32.5	53.2
West												
White	2,716	42.6	18.8	19.7	27.1	26.8	36.2	32.2	47.1	42.4	61.0	56.4
Black	277	32.9	19.5	23.1	26.7	28.8	40.5	42.2	48.4	43.9	58.5	56.0

Note: Figures for Blacks are an average of survey data for 1975 and 1976.

Table 3

Family Income Distribution of Dependent Family Members 18 to 24 Years Old by Region and Race: 1976

(Income in 1967 dollars)

Region and race	Persons (thousands)	Total percent enrolled	Under $3,000	$3,000 to $4,999	$5,000 to $7,499	$7,500 to $9,999	$10,000 or more
United States							
White	11,834	100.0	7.0	11.0	17.9	19.3	14.5
Black	1,990	100.0	33.9	24.4	19.9	11.2	10.6
Northeast							
White	3,611	100.0	5.2	10.1	18.9	20.1	45.7
Black	452	100.0	25.8	24.5	21.9	12.6	14.9
Southeast							
White	2,163	100.0	9.1	12.4	19.6	17.6	41.2
Black	885	100.0	42.8	25.1	17.3	8.8	6.1
Central							
White	3,344	100.0	5.5	9.5	17.3	20.4	47.3
Black	377	100.0	21.6	24.9	23.0	14.0	16.3
West							
White	2,716	100.0	7.8	10.5	13.1	14.7	35.4
Black	277	100.0	34.5	20.8	21.8	13.2	10.0

Note: Figures for Blacks are an average of survey data for 1975 and 1976.

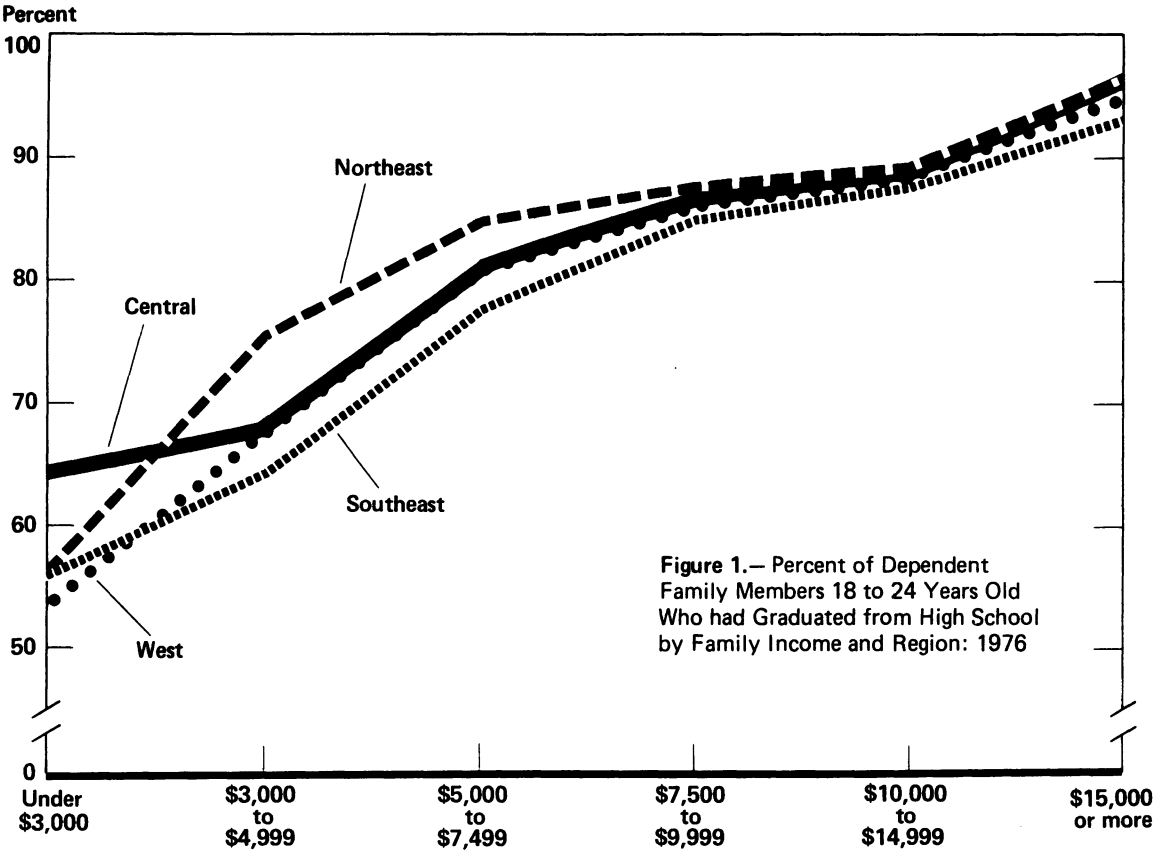


Figure 1.— Percent of Dependent Family Members 18 to 24 Years Old Who had Graduated from High School by Family Income and Region: 1976

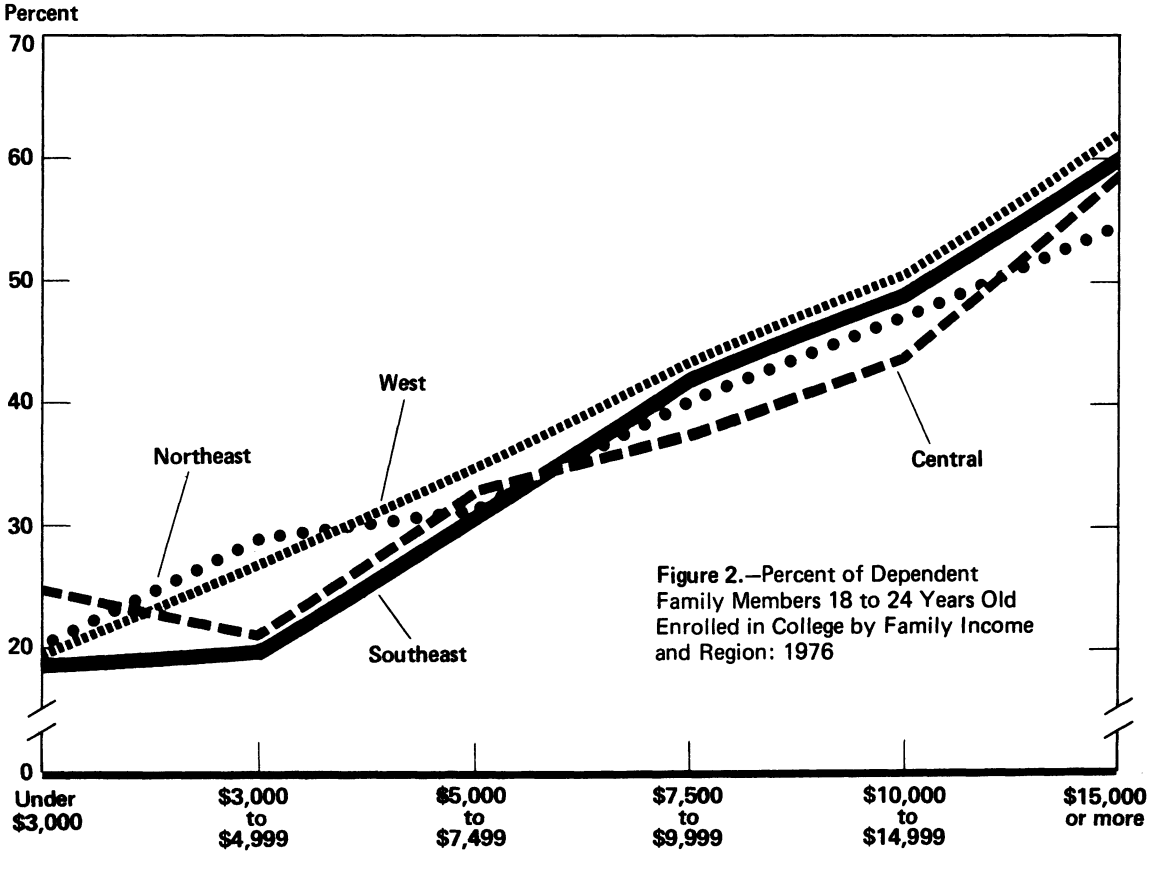


Figure 2.—Percent of Dependent Family Members 18 to 24 Years Old Enrolled in College by Family Income and Region: 1976

ESTIMATION OF CHILDREN IN POVERTY IN NEW JERSEY

Thomas Vietorisz, New School for Social Research
and
Robert Mier, University of Illinois at Chicago Circle

I. INTRODUCTION

Public Law 93-389, The Elementary and Secondary Education Act, mandates the annual distribution of close to $2 billion of federal funds to local school districts.[1] The primary basis for the distribution of these funds is the number of children in poverty (CIP), aged 5-17 inclusive, in each school district at the time of the last decennial census (1970). This basis is slightly modified by the inclusion of children in families receiving Aid to Families with Dependent Children (AFDC) if the family income exceeds the Government's most recent official poverty line after receipt of such welfare funds. The principal objective of this study, one of several commissioned by the Department of Health, Education and Welfare (HEW) to test the feasibility of updating CIP counts, has been the testing and determination of an improved method for estimating the number of children in poverty at the sub-state level based on relationships between labor market attachment, poverty, and specific quality of life indicators uncovered by earlier work of the Research Center for Economic Planning. This earlier study established strong statistical associations between labor market attachment, poverty, and a series of indicators measuring deterioration of the quality of life and the incidence of such social pathologies as family break-up, ill physical and mental health, crime, and poor housing.[2] The final report inferred a causal link between family incomes earned in the labor market and the prevalence of poverty, even though such causality, for well-known fundamental reasons, can never be conclusively established on the basis of statistical observations alone.[3]

II. METHODS OF PROCEDURE

The specific objectives of the study reported here included:

* Establishing regression specifications that had a high degree of explanatory power in describing the county-by-county distribution of children in poverty, as counted by the 1970 census.
* Estimating the regression parameters for the State of New Jersey.
* Applying the estimated parameters for prediction of the county-by-county distribution of children in poverty in New Jersey in 1974 and 1976.
* Comparing the predictions with the Khan-Miller-BEA estimates for New Jersey counties for 1974 and 1976.[4]

The original intent in undertaking this study was to utilize, among the findings of the earlier work, the relationship between labor market attachment and poverty, while using information on quality of life and social pathologies for background purposes only. Constraints on data availability, coupled with the tight time schedule of the project, forced a modification of this strategy, shifting the main emphasis to quality of life and social pathology indicators as predictors of poverty.

The critical data set in regard to this decision was the Continuous Work History Sample of the United States Social Security Administration. The methodology proposed for quantification of the labor market attachment of principal wage earners in households turned largely on the use of the 1-percent sample annual Employee-Employer ("Ee-Er") file. Immediately following initiation of the project, it was learned that the Social Security Administration had temporarily embargoed the release of all information in this file to new users. This embargo applied both directly, in the form of suspension of tape file sales, and indirectly, by denying permission for existing users to share their information with new users. The embargo had its origin in administrative relationships between the Internal Revenue Service and the Social Security Administration. In mid-March, at the time of initiation of the study, this embargo was expected to remain in force at least until May, with no assurance that far longer delays were excluded.

Since the Social Security tapes could not become available in time for this project, labor market attachment could be quantified on a current basis only by means of data from the Unemployment Insurance and Employment Service systems of the state. As it turned out, these data were also severely limited in availability. Even in states where such data are more readily available, such as New York, unemployment and related data are poor predictors of long-term poverty. In the earlier study of the Research Center for Economic Planning, referred to above, the fraction of poverty directly explained by unemployment variation turned out to be of the order of one percent, and was totally without statistical significance.[5]

An indicator related to unemployment that does, in fact, predict poverty far more effectively is "subemployment," a concept that broadly reflects both a lack of work opportunities and a prevalence of jobs that offer less than family support level wages. The subemployment concept combines officially defined unemployment, the long-term discouraged jobless who have stopped looking for work, involuntary part-time workers who are unable to find full-time jobs, and workers earning substandard wages (by one of several possible definitions).[6] If subemployment data could be obtained on a current basis, they would be considerably more useful than unemployment for predicting poverty. Unfortunately, the necessary components are available only once a year, and then only on a nationwide basis, from the Current Population Survey.

This is why the present study intended to rely on the Social Security file tapes for defining labor market attachment. The subemployment concept embodies the effects of "secondary" labor markets in generating poverty, but is not avail-

able on a current basis by small areas. The Social Security sample tapes are available on a reasonably current basis and include information --industry codes, occupation codes, and wage levels--that allow construction of a proxy for the relative prevalence of primary and secondary jobs in an area. Unemployment data, even in combination with some related wage or benefit exhaustion information, cannot yield such a proxy.

Given this severe data problem, it has become necessary to shift emphasis to another aspect of the nexus between labor market attachment, poverty, and the quality of life uncovered by the earlier study of the Research Center for Economic Planning. Fortunately, this earlier study disclosed that just as good a profile of poverty can be assembled from quality of life indicators as from labor market attachment indicators, and--as it has turned out--the components of quality of life indicators are available on a far more timely basis for sub-state areas than are components of labor market attachment indicators. Guided by the findings compiled and extensive experience with poverty data gained in this earlier study, a major effort was thus made to collect the best possible indicators for the conditions that are known to be fundamentally involved in the existence and reproduction of poverty.

Among the many conditions known to be involved in both being caused by and causing poverty, the principal ones are physical and mental health; educational levels; drug and crime problems; and housing conditions. The data collection effort was therefore centered on obtaining a good representation of such variables in the data set on which the regressions and projections were based, and are presented in Table 1.

TABLE 1

Explanatory Variables: Data Availability at the County Level in New Jersey

Series No.	Item	Years
1	Population (census and estimates)	1970-75
2	Deaths under 1 year	1965-75
3	Deaths under 28 days	1965-75
4	Maternal deaths	1965-75
5	Stillbirths	1965-75
6	Illegitimate births	1965-75
7	Homicides	1970-75
8	Suicides	1964-75
9	Syphilis cases, all stages	1965-75
10	Syphilis cases, Pri-sec	1965-75
11	Syphilis, early latent	1965-75
12	Gonorrhea cases	1965-75
13	No. of AFDC children	1965-76
14	AFDC Assistance payments	1965-76
15	Pub. Schl enrollment, K-12	1965-76
16	P.S. enrollment K-12, black	1962-76
17	P.S. enrollments K-12, hispanic	1962-76
18	Paroch. Schl. enrollment 1-8	1972-75
19	Unemployment insurance, all weeks claimed	1968-76
20	Unemployment rate, seas. adjusted	1970-76

III. MODEL SPECIFICATION AND TESTING

The objective in the statistical correlation and regression specifications was to identify variables that met the following requirements:

* They were highly correlated with the 1970 CIP count.
* They also showed a high correlation even when lagged by several years. This is a very desirable characteristic, since current data often become available only with a long time lag; therefore, if an explanatory variable is to serve for project purposes, it should predict well for some years ahead.
* It is plausible to assume that they were causally interrelated with poverty. To the extent that a variable satisfies this requirement, the statistical relationship can be expected to hold broadly across the United States as a whole, rather than being tied to the characteristics of the state of New Jersey, for which this case study was being undertaken.
* They were available within the data set for a sufficient number of years prior to the 1970 Census to allow the confirmation of lagged correlations referred to above.

Of the variables included in the set, the following correlate approximately at the .95 level, with time lags of at least two years, with the 1970 CIP count:

Variable Number	Description
2	Deaths under 1 year
3	Deaths under 28 days
4	Maternal deaths
5	Stillbirths
6	Illegitimate births
9	Syphilis, all stages
10	Syphilis, primary and secondary
11	Syphilis, early latent
12	Gonorrhea
13	Number of AFDC children
14	AFDC assistance payments
16	Public school enrollment, black

All of these correlations, including the last one, were statistically significant at the .05 level.

By contrast, the unemployment rate correlated with the CIP count only with a coefficient of .5 and was statistically not significant.

Using the methods of Daniel and Wood, the following three linear regression estimates were found appropriate in terms of functional form and model efficiency.[7] Standard deviations are shown beneath the respective coefficient in parentheses; single, double, and triple stars denote coefficients that are statistically significant at the .05, .01, and .001 levels respectively.

$$CIP = -1165.5 + \underset{(5.0)***}{70.2\ DEATH < 28}; \quad R^2 = .91;$$

$$F = 195.43.$$

Explanatory variable lagged three years; coefficient Statistically significant at the .001 level.[8]

$$CIP = 584.2 + \underset{(10.0)**}{36.1\ STILLB} + \underset{(3.24)***}{19.9\ SYPH};$$

$$R^2 = .96; \text{ overall } F = 205.95$$

All explanatory variables lagged three years; coefficients significant at the .01 level.

$$CIP = -3842.1 - 222.7\ DEATH{<}28 - 16.3\ STILLB$$
$$(66.1)^{**} \qquad (20.4)$$
$$+ 212.6\ DEATH{<}1 + .9\ UNEMP - 12.5\ ILLEG$$
$$(60.3)^{**} \qquad (.52) \qquad (5.75)^{*}$$
$$+ 1581.1\ MATDEATH + 39.0\ SYPH$$
$$(403.0)^{***} \qquad (12.9)^{**}$$
$$R^2 = .98;\ \text{overall } F = 115.0$$

All variables except unemployment lagged three years; unemployment lagged two years. Not all coefficients statistically significant.

As the first regression result shows, a single explanatory variable, that of deaths under 28 days, explains 91 percent pf the variation in CIP counts; two variables, stillbirths and total syphilis cases, esplain 96 percent; and seven variables explain 98 percent. In the regression with seven variables, the coefficients of some of the variables are seen to be negative. This is at first sight counter-intuitive, since deaths under 28 days, stillbirths, and illegitimate births are individually each positively correlated with CIP. The phenomenon becomes readily understandable, however, when it is noted that all the variables that individually are highly correlated with CIP are in turn highly intercorrelated among themselves. This is illustrated in Appendix A showing the results of principle components analysis of the data set.

The principle components analysis yielded two primary factors. Factor 1 can be interpreted as a quality-of-life factor. It has very high loadings of the eleven health and welfare variables: four load above .98 and all but one above .90, with the remaining one--maternal deaths--still showing a high loading at .83. The two unemployment variables, conversely, show low loadings on this factor. The second factor is identified as an unemployment factor. Both unemployment variables load at .73 on the second factor, while the health and welfare variables show very low loadings, rarely reaching .20 at the most.

The interpretation of these results is entirely in line with the anticipated causal relationship between quality of life and social pathology indicators and poverty. It is particularly noteworthy that data taken from birth and death certificates offer such a powerful explanatory device. This leads directly to an important consideration flowing from the current study. Since data from birth and death certificates are available on a current basis, and since they are such excellent explanatory variables for CIP, it might be useful to undertake an effort to obtain uniform country-wide tabulations of birth and death certificates by area of residence on a timely basis. With a moderate effort, data from such certificates could be tabulated directly on a county-by-county or even a school district-by-school district basis, and could be used as CIP correlates at a local area level, hopefully down to the local school district.

IV. ESTIMATES OF CHILDREN IN POVERTY

Estimates of children in poverty (CIP) for the year 1974 and 1976 have been prepared by projecting the cross-sectional regression results obtained for the year 1970. In each case, the 1970 CIP count by county in New Jersey was used as the base, and positive or negative changes were added as calculated from the regression equation, substituting into this equation the corresponding changes in the explanatory variables.

Two different regression specifications were used for projection purposes. Both used the same three explanatory variables--deaths under 28 days, weeks of unemployment insurance claimed, and the count of AFDC children. In specification A, the variables were averaged over several years, while in specification B, they referred to single years. Details will be found in Appendix

Appendix B shows the count-by-count projection results. The most notable feature of the projections is the contrast between the very high R^2's of the cross-sectional regression equations and the changeability of the coefficients over time. For reference purposes Appendix B includes the 1974 and 1976 Khan-Miller-BEA estimates.

A) State Total Counts

As can be seen in the row of state totals, these totals are highly sensitive to the time period to which the explanatory variables refer. Specification A, whose explanatory variables are averaged over a period lagged from 1 to 5 years (in the case unemployment, 1 to 2 years), leads to a projection of a state total less than half of that projected by specification B. The latter has explanatory variables that are lagged only 2 or 3 years.

This sensitivity has its origin in the time trands of some of the explanatory relationships. The number of deaths under 28 days has, for example, decreased substantially in recent years, and these decreases do not necessarily follow a linear pattern. Changes in the AFDC caseload, likewise, cannot be taken to change in a smooth, regular manner over time.

It is, therefore, concluded that the projections must be normalized in order to be useful. The very strong cross-sectional relationship between CIP by county and the explanatory variables can be used to predict each county's share in the state-wide CIP count, but the rapid changes of coefficients over time preclude the direct longitudinal use of cross-sectional coefficients. For the present projections, the state SIE count was used as a normalizer; in future work, normalization must be tied to the total CIP count for the U.S. as a whole, since only this total is available as a yearly basis from the CPS.

B) County-by-county Proportions

Comparison of the Khan-Miller-BEA estimates with the projections obtained shows that the projections may overstate the county-by-county changes in the CIP count since 1970. These changes have their root in shifting spatial patterns of poverty. The Khan-Miller-BEA estimates do not allow for any autonomous county-by-county

poverty shifts; they simply infer the consequences of an overall population and income increase on a spatially fixed 1970 CIP pattern.[9] Therefore, it is not possible to judge whether the differences between the Khan-Miller-BEA county-by-county proportions of CIP and the projected proportions obtained by regression methods have their origin in:

- understatement of spatial shifts by the Khan-Miller-BEA method;
- overstatement of spatial shifts by the regression method; or
- a combination of both

It is, therefore, concluded that the quality-of-life variables should be used as a group in future work, in the form of factor scores calculated for individual counties. This will merge the information contained in these explanatory variables and will greatly reduce the effect of random variations on county-by-county projections.

V. CONCLUSIONS

The conclusions flowing from the work under this contract can be summarized as follows:

* A broad data set of quality-of-life and unemployment variables has been identified that correlates strongly with CIP and is available for projections by individual counties on a reasonably current basis.
* Regressions specified on this data set in a number of different ways, using lagged or time-averaged explanatory variables, give excellent cross-sectioned explanations of the 1970 county-by-county CIP count.
* These regressions give statistically significant coefficients, at the .95 level or better, for many of the quality-of-life variables; the coefficients of the unemployment variables are, however, generally less significant.
* A factor analysis of the explanatory variables yields a quality-of-life factor with very high loadings of eleven health and welfare variables, and an unemployment factor with high loadings of the two unemployment variables.
* County-by-county projections based on the cross-sectional regression results indicate that the regression coefficients change over time. Therefore, the cross-sectional regressions should be normalized to a state or national total.
* County-by-county projections may also show instabilities owing to random variations in individual explanatory variables, especially for the smaller counties. Merging sets of variables by the device of using factor scores should improve stability in this regard.

It is, therefore, recommended that:

* The high explanatory power of the cross-sectional regressions should be confirmed, using the same data set as applied to a different state.
* The regressions should be run with variables expressed as proportions of state-wide totals, and merged into quality-of-life and unemployment factors.
* The explanatory power of the unemployment factor should be improved by including further variables, using the Social Security sample tapes if and when available.
* The regressions should be sharpened by distinguishing at least urban and rural counties in a large data set.
* If the results of future tests in one or more other states are promising, a national data set should be tested, consisting of all 3,000 counties. Within such a data set, different types of urban areas and different regions of the country should be distinguished, and the absolute CIP counts should be normalized to the U.S. total, as well as the total of all metropolitan areas.

FOOTNOTES

1. This research was sponsored under contract from the Office of the Assistant Secretary for Education (Policy Development), Department of Health, Education, and Welfare to the Research Center for Economic Planning, New York, N.Y. Assistance in statistical computation and analysis has been provided by David Less.

2. Thomas Vietorisz, "Earned Family Incomes and the Urban Crisis", a research report submitted to the Center for the Study of Metropolitan Problems, National Institute for Mental Health, January, 1976; see also Robert Mier, Thomas Vietorisz, and Jean-Ellen Giblin, "Indicators of Labor Market Functioning and Urban Social Distress", Social Economy of Cities, ed. Gary Gappert and Harold M. Rose (Beverly Hills, Sage Publications, 1975), pp. 361-394.

3. For example, David Layzer, "Is There Any Real Evidence that I.Q. Tests are Heritable?", Scientific American, July, 1975, pp. 126-128. In any regards, a "culture-of-poverty" thesis such as Banfield's would suggest a similar model for estimating children in poverty. Edward Banfield, Unheavenly City Revisited, (Boston, Little, Brown, 1974), pp. 52-76.

4. Herman P. Miller and Abdul Khan, "Methodology for Estimating the Number of Children in Poverty for States and Counties", Business Uses of Small-Area Statistics and Education's Needs and Methods for Estimating Low-Income Population (U.S. Department of Commerce, Bureau of the Census, June 1976), pp. 40-46

5. Mier, et. al.

6. Thomas Vietorisz, Robert Mier, and Jean-Ellen Giblin, "Subemployment; Exclusion and Inadequacy Indexes", Monthly Labor Review, May 1975 pp. 3-12.

7. Cuthbert Daniel and Fred S. Wood, Fitting Equations to Data (New York, Wiley-Interscience, 1971).

(Footnotes Continued)

8. The regressions equations were also specified with the CIP counts normalized as proportions to total school enrollments, and the explanatory variables normalized as proportions to population. These regressions equations yielded no improvement over those discussed in the text, either in terms of R^2s or in terms of standard deviations of the coefficients.

9. Miller and Khan

APPENDIX A

PRINCIPLE COMPONENTS RESULTS

Vbl. No.	Vbl. Name	Rotated Factor Pattern Factor 1	Factor 2
13	AFDC Children	0.98132	0.09468
14	AFDC Payments	0.98249	0.06832
20	Unemployment rate	-0.13816	0.73272
19	Unemployment insurance	0.35064	0.73508
2	Death < 1	0.93673	0.20010
3	Death < 28	0.94196	0.15960
4	Maternal death	0.83416	-0.08902
9	Syphillis All	0.98331	0.09303
10	Syphillis 1,2	0.07183	0.00095
11	Syphillis early	0.95971	0.12614
12	Gonorrhea	0.95456	-0.07979
5	Stillborn	0.90628	0.20548
6	Illegal births	0.98163	0.12196

Orthogonal Transformation Matrix

	1	2
1	0.99349	0.11392
2	-0.11392	0.99349

APPENDIX B

ESTIMATES OF CHILDREN IN POVERTY

County	Estimate for 1975 (SIE total)	1970 CIP (census)	1974 BEA	1976 BEA	1974 RCEP(A)	1976 RCEP(A)	1974 RCEP(B)	1976 RCEP(B)
		#	#	#	#	#	#	#
Atlantic	9409	5703	6268	6956	5583	5612	11512	11659
Bagen	3880	7382	7231	8461	2975	1695	10587	9391
Burlington	6111	7176	6545	7911	3567	3733	10013	10289
Camden	26869	12445	13293	15381	14105	17912	27787	32001
Conmay	2425	1477	1558	1689	1357	1530	2640	2898
Cumberland	16402	4051	4597	5249	3632	4039	7650	8157
Essex	37151	36793	39685	40691	23057	21240	61966	60427
Gloucester	4074	3796	4264	4662	2819	2124	6699	6104
Hudson	20176	19723	21208	24595	11162	12896	31750	33836
Houstendon	1649	909	1250	1266	888	1043	1543	1709
Mason	9700	7088	7241	8168	5959	5672	13279	13349
Middlesex	9700	7353	7597	7984	7633	6293	19558	14416
Monmouth	9797	9228	10551	11343	7671	6324	17377	16102
Morris	1067	3224	3536	3524	751	526	3978	3968
Ocean	8051	4153	6280	5763	4807	4800	9045	9263
Passaic	18818	11287	11653	12261	11910	10530	24279	22686
Salem	2231	2227	2343	2843	1338	1363	3273	3318
Somerset	1940	3027	2274	2276	1325	986	3163	2799
Sussex	2231	1294	1697	2001	1372	1297	2426	2393
Union	6984	7850	8299	9030	3941	4412	12004	12851
Weber	1455	1232	1300	1505	838	930	1809	1878
TOTAL	194,000	155,690	169,172	183,505	116,688	114,957	278,382	279,496

NOTES: (A) $CIP_{74,76} = CIP_{70} + 33.4 \Delta DEATH < 28_{Ave} + 0.341 \Delta UNEMP_{Ave} + 0.59 \Delta AFDC_{Ave}$

(B) $CIP_{74,76} = CIP_{70} + 32.8 \Delta DEATH < 28 + 0.541 \Delta UNEMP + 0.63 \Delta AFDC$

(C) Average of 1974 and 1976 A-type estimated percentages applied to 1975 SIE state total.

APPENDIX C

ESTIMATING METHODS FOR PROJECTIONS

Estimating Equation A

$CIP_{1970} = 435.43 + 33.42\ DEATH < 28^{*}_{Ave\ 65-69} + 0.34\ UNEMP_{Ave\ 68-69} + 0.60\ AFDC^{**}_{Ave\ 65-69}$ $R^2 = .973$

T - Statistic	(3.51)	(0.96)	(6.93)
Standard error	(9.52)	(0.35)	(0.09)

*Significant at .05 level
**Significant at .01 level

F - statistic for the entire equation is 202.49. It is significant at better than the .01 level.

$DEATH < 28_{Ave\ 65-69}$ = Average Annual Number of Infant Deaths occuring between birth and 28 days in the period 1965-1969.

$UNEMP_{Ave\ 68-69}$ = Average Monthly Insured Unemployed in the period 1968-1969.

$AFDC_{1967}$ = Average Monthly Number of AFDC Children Assisted in the period 1965-1969.

In difference, or estimatory, form this equation is:

$$CIP_{1974} = CIP_{1970} + 33.42\ (DEATH < 28_{Ave\ 69-73} - DEATH < 28_{Ave\ 65-69}) + 0.34\ (UNEMP_{Ave\ 72-73} - UNEMP_{Ave\ 68-69}) + 0.60\ (AFDC_{Ave\ 69-73} - AFDC_{Ave\ 65-69})$$

$$CIP_{1976} = CIP_{1970} + 33.42\ (DEATH < 28_{Ave\ 71-75} - DEATH < 28_{Ave\ 65-69}) + 0.34 - (UNEMP_{Ave\ 74-75} - UNEMP_{Ave\ 68-69}) + 0.60\ (AFDC_{Ave\ 71-75} - AFDC_{Ave\ 65-69})$$

Estimating Equation B

$CIP_{1970} = 423.6 + 32.85\ DEATH < 28^{*}_{1967} + 0.55\ UNEMP_{1968} + 0.63\ AFDC^{**}_{1967}$ $R^2 = .971$

T Statistic	(3.37)	(1.43)	(6.28)
Standard Error	(9.76)	(0.38)	(0.10)

*Significant at .05 level
** Significant at .01 level

F Statistic for the entire equation is 188.08. It is significant at better than the .01 level.

$DEATH < 28_{1967}$ = Infant deaths occuring between birth and 28 days in 1967.

$UNEMP_{1968}$ = Average Monthly Insured Unemployed in 1968.

$AFDC_{1967}$ = Average Monthly Number of AFDC Children Assisted in 1967.

In difference of estimating form, this equation is:

$$CIP_{1974} = CIP_{1970} + 32.85\ (DEATH < 28_{1971} - DEATH < 28_{1967}) + 0.55\ (UNEMP_{1972} - UNEMP_{1968}) + 0.63\ (AFDC_{1971} - AFDC_{1967})$$

$$CIP_{1967} = CIP_{1970} + 32.85\ DEATH < 28_{1973} - DEATH < 28_{1967}) + 0.55\ (UNEMP_{1974} - UNEMP_{1968}) + 0.63\ (AFDC_{1973} - AFDC_{1967})$$

ESTIMATING THE VARIANCE OF THE SLOPE OF A LINEAR REGRESSION IN A STRATIFIED RANDOM SAMPLE WITH THE BALANCED HALF-SAMPLE TECHNIQUE

Stanley Lemeshow, University of Massachusetts/Amherst

Abstract

Estimation of the variance of the slope of the linear regression under a variety of computer generated situations with the Balanced Half-Sample procedure is considered. Three estimates for the population slope β, each of which is optimal for different situations, are presented. The method of applying the Balanced Half-Sample technique with each of these estimates is investigated and then evaluated with a Monte Carlo experiment.

The results of the investigation show that variance estimates of the slope are highly biased and very unstable unless sizeable numbers of observations are selected from each stratum. The choice of the best estimator of β from the three presented depends on the particular situation under consideration.

1. Introduction

The balanced half-sample (BHS) technique has been used for some time to estimate the variance of the combined ratio estimate in such large-scale sample surveys as the Health Examination Survey (HES) and the Health Interview Survey (HIS) of the National Center for Health Statistics (NCHS). Other large-scale surveys have used variance estimation techniques such as a Taylor Series expansion or the linearization method for the same purpose. Proponents of the BHS technique have claimed that an estimate of the variance of any non-linear estimate of interest could be obtained without having to derive new expressions for the approximations to the variances as would be the case with the linearization method. The properties of the BHS technique have been documented by McCarthy (1966, 1969) and Lemeshow and Epp (1977) and its properties for the ratio estimate have been presented by Lemeshow and Levy (1977).

This paper considers the slope of the linear regression as a particular non-linear estimate. The BHS technique is used to estimate its variance in a variety of computer generated situations. The ability of this method to effectively estimate the variance of the slope is carefully considered and evaluated through the use of Monte-Carlo experiments. This is done in the context of a stratified random sample.

Specifically, consider a population subdivided into L strata of equal weight. A random sample of size n is drawn from each stratum and observations denoted (x_{ij},y_{ij}), $i=1,\ldots,L$, $j=1,\ldots,n$ are made. The pertinent population parameters are denoted by μ_x, μ_y, σ_x^2, σ_y^2, σ_{xy}, $\rho=\sigma_{xy}/\sigma_x\sigma_y$ and $\beta=\sigma_{xy}/\sigma_x^2$. The strata parameters and observations are illustrated as below:

Strata	Pop. Size	Strata Parameters	Sample Observations
1	N_1	$\mu_x^{(1)},\mu_y^{(1)},\sigma_{xx}^{(1)},\sigma_{yy}^{(1)},\sigma_{xy}^{(1)},\beta^{(1)},\rho^{(1)}$	$(x_{11},y_{11}),\ldots,(x_{1n},y_{1n})$
2	N_2	$\mu_x^{(2)},\mu_y^{(2)},\sigma_{xx}^{(2)},\sigma_{yy}^{(2)},\sigma_{xy}^{(2)},\beta^{(2)}\ \rho^{(2)}$	$(x_{21},y_{21}),\ldots,(x_{2n},y_{2n})$
.	.	.	.
.	.	.	.
.	.	.	.
L	$\frac{N_L}{N}$	$\mu_x^{(L)},\mu_y^{(L)},\sigma_{xx}^{(L)},\sigma_{yy}^{(L)},\sigma_{xy}^{(L)},\beta^{(L)},\rho^{(L)}$	$(x_{L1},y_{L1}),\ldots,(x_{Ln},y_{Ln})$

2. Estimating β

Let the population slope be defined as

$$\beta = \sigma_{xy}/\sigma_x^2 . \tag{2.1}$$

The following three estimates are considered for this parameter:

(i)
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{L}\sum_{j=1}^{n}(x_{ij}-\bar{x}..)(y_{ij}-\bar{y}..)}{\sum_{i=1}^{L}\sum_{j=1}^{n}(x_{ij}-\bar{x}..)^2} \tag{2.2}$$

where the deviations are taken about the overall means.

(ii)
$$\hat{\beta}_2 = \frac{\sum_{i=1}^{L}\left[\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i}.)(y_{ij}-\bar{y}_{i}.)\right]}{\sum_{i=1}^{L}\left[\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i}.)^2\right]} \tag{2.3}$$

where the deviations are taken about the within-stratum means.

(iii)
$$\hat{\beta}_3 = \frac{\frac{1}{L}\sum_{i=1}^{L}\{\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i}.)(y_{ij}-\bar{y}_{i}.)}{\sum_{j=1}^{n}(x_{ij}-\bar{x}_{i}.)^2\}} \tag{2.4}$$

which is the average of the strata slopes.

The following theorem is presented without proof:

Theorem 1: If n observations are randomly selected from each of L strata and β is defined as in (2.1), then as $n\to\infty$,

(i) $\hat{\beta}_1 \xrightarrow{p} \beta$ irrespective of the distribution of x or y

(ii) $\hat{\beta}_2 \xrightarrow{p} \beta$ if $\mu_x^{(i)}=\mu_x^*$ for all i, where μ_x^* is any arbitrary constant

(iii) $\hat{\beta}_3 \xrightarrow{p} \beta$ if $\mu_x^{(i)}=\mu_x^*$ and $\sigma_{xx}^{(i)}=\sigma_{xx}^*$ for all i, where μ_x^* and σ_{xx}^* are any abribrary constants.

A proof of the theorem is given by Lemeshow (1976).

The choice of the appropriate method of estimating β is not always clear because the parameters of the independent variable in each stratum are often unknown. In certain cases the choice is clear. For instance, if x and y are bivariate normal, and if the distribution of x is the same in each stratum, then $\hat{\beta}_3$ is the maximum likelihood estimate of β and as such is known to be the minimum variance unbiased estimate. If we only have $\mu_x^{(i)}=\mu_x^*$ in each stratum, then both $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent. Consistency is always assured using $\hat{\beta}_1$, but clearly, use of this estimate may provide an unnecessary loss of precision.

3. Estimating the Variance of $\hat{\beta}$ with the BHS Technique

In the half-sample method, assume the n observations from each stratum are divided into two groups of r=n/2 observations each. Let x_{ijw}=w^{th} observation in the j^{th} group of stratum i, i=1,...,L, j=1,2,w=1,...,r. The balanced half-sample method can be used to estimate the variance of $\hat{\beta}$ when $\hat{\beta}$ is computed using any of the three estimates (2.2), (2.3), or (2.4).

3.1 Method 1 (deviations computed about overall means):

Let $\hat{\beta}_{(p)}$ be the p^{th} half-sample estimate of β corresponding to the estimate defined in (2.2). That is,

$$\hat{\beta}_{(p)} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}(x_{ijw}-\bar{x}_{...}^{(p)})(y_{ijw}-\bar{y}_{...}^{(p)})}{\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}(x_{ijw}-\bar{x}_{...}^{(p)})^2}$$

where

$$\bar{x}_{...}^{(p)} = \frac{1}{Lr}\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}x_{ijw} \quad ,$$

$$\bar{y}_{...}^{(p)} = \frac{1}{Lr}\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}y_{ijw} \quad ,$$

and

$$\delta_{ij}^{(p)}\begin{cases}1 & \text{if the } j^{th} \text{ group of } i^{th} \text{ stratum is in the } p^{th} \text{ half-sample}\\ 0 & \text{if the } j^{th} \text{ group of } i^{th} \text{ stratum is not in the } p^{th} \text{ half-sample}\end{cases} \qquad (3.1)$$

Then, letting M = total number of half-samples computed,

$$\hat{V}_{B1}(\hat{\beta}_1) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\bar{\beta})^2 \quad , \quad \bar{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{(i)}$$

and

$$\hat{V}_{B2}(\hat{\beta}_1) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\hat{\beta}_1)^2 \quad , \quad \hat{\beta}_1 \text{ defined in (2.2).}$$

3.2 Method 2 (deviations computed about within-stratum means):

Let $\hat{\beta}_{(p)}$ be the p^{th} half-sample estimate of β corresponding to the estimate defined in (2.3).

$$\hat{\beta}_{(p)} = \frac{\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}(x_{ijw}-\bar{x}_{ij.})(y_{ijw}-\bar{y}_{ij.})}{\sum_{i=1}^{L}\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{w=1}^{r}(x_{ijw}-\bar{x}_{ij.})^2}$$

where

$$\bar{x}_{ij.} = \frac{1}{r}\sum_{w=1}^{r}x_{ijw}$$

$$\bar{y}_{ij.} = \frac{1}{r}\sum_{w=1}^{r}y_{ijw}$$

and $\delta_{ij}^{(p)}$ is defined as in (3.1).

Then,

$$\hat{V}_{B1}(\hat{\beta}_2) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\bar{\beta})^2 \quad , \quad \bar{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{(i)}$$

and

$$\hat{V}_{B2}(\hat{\beta}_2) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\hat{\beta}_2)^2 \quad , \quad \hat{\beta}_2 \text{ defined in (2.3).}$$

3.3 Method 3 (average of the strata slopes):

Let $\hat{\beta}_{(p)}$ be the p^{th} half-sample estimate of β corresponding to the estimate defined in (2.4).

$$\hat{\beta}_{(p)} = \frac{1}{L}\sum_{i=1}^{L}\left\{\frac{\sum_{j=1}^{2}\delta_{ij}^{(p)}\sum_{k=1}^{r}(x_{ijw}-\bar{x}_{ij.})(y_{ijw}-\bar{y}_{ij.})}{\sum_{i=1}^{2}\delta_{ij}^{(p)}\sum_{k=1}^{r}(x_{ijw}-\bar{x}_{ij.})^2}\right\},$$

where

$\bar{x}_{ij.}$, $\bar{y}_{ij.}$, $\delta_{ij}^{(p)}$ are defined as before.

Then,

$$\hat{V}_{B1}(\hat{\beta}_3) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\bar{\beta})^2 \quad , \quad \bar{\beta} = \frac{1}{M}\sum_{i=1}^{M}\hat{\beta}_{(i)}$$

and

$\hat{V}_{B2}(\hat{\beta}_3) = \frac{1}{M}\sum_{i=1}^{M}(\hat{\beta}_{(i)}-\hat{\beta}_3)^2$, $\hat{\beta}_3$ defined in (2.4).

4. The Sampling Experiment

The sampling experiment consists of randomly selecting n observations for each of L strata of infinite size whose parameters are precisely specified. On the j-th draw from the i-th stratum the random pair (x_{ij},y_{ij}) is observed where

$$\begin{pmatrix} x_{ij} \\ y_{ij} \end{pmatrix} \sim N\left[\begin{pmatrix} \mu_x^{(i)} \\ \mu_y^{(i)} \end{pmatrix}, \begin{pmatrix} \sigma_{xx}^{(i)} & \sigma_{xy}^{(i)} \\ \sigma_{xy}^{(i)} & \sigma_{yy}^{(i)} \end{pmatrix}\right] .$$

Although in this experiment, the parameters are known, estimates of them are obtained. The process was then repeated K times and the distribution of the estimates of β were studied.

The strata correlations, $p^{(i)}$ are set equal to .9 for all strata and values are specified for $\mu_x^{(i)}$, $\sigma_{xx}^{(i)}$, $\beta^{(i)}=\sigma_{xy}^{(i)}/\sigma_{xx}^{(i)}$ and $\alpha^{(i)}=\mu_y^{(i)}-\beta(i)\mu_x^{(i)}$. By fixing these parameters, the values of $\mu_y^{(i)}$, $\sigma_{yy}^{(i)}$ and $\sigma_{xy}^{(i)}$ are determined.

A variety of "situations," covering a range of parameters, were considered. These can be summarized as follows:

Situation (i) L=3, n=20, K=1200:

	$(\mu_x^{(1)},\mu_x^{(2)},\mu_x^{(3)})$	$(\sigma_{xx}^{(1)},\sigma_{xx}^{(2)},\sigma_{xx}^{(3)})$	$(\beta^{(1)},\beta^{(2)},\beta^{(3)})$	$(\alpha^{(1)},\alpha^{(2)},\alpha^{(3)})$
Situation (i-1)	(5,5,5)	(1,1,1)	(1,1,1)	(0,0,0)
Situation (i-2)	(5,5,5)	(1,1,1)	(1,2,3)	(0,0,0)
Situation (i-3)	(5,5,5)	(1,1,1)	(1,1,1)	(0,1,2)
Situation (i-4)	(5,5,5)	(1,1,1)	(1,2,3)	(0,1,2)
Situation (i-5)	(5,5,5)	(1,2,3)	(1,1,1)	(0,0,0)
Situation (i-6)	(5,5,5)	(1,2,3)	(1,2,3)	(0,0,0)
Situation (i-7)	(5,5,5)	(1,2,3)	(1,1,1)	(0,1,2)
Situation (i-8)	(5,5,5)	(1,2,3)	(1,2,3)	(0,1,2)
Situation (i-9)	(5,10,15)	(3,6,9)	(1,1,1)	(0,0,0)
Situation (i-10)	(5,10,15)	(3,6,9)	(1,2,3)	(0,0,0)
Situation (i-11)	(5,10,15)	(3,6,9)	(1,1,1)	(0,1,2)
Situation (i-12)	(5,10,15)	(3,6,9)	(1,2,3)	(0,1,2)

Situation (ii) L=3, n=2, repeated for all 12 sets of parameters as in Situation (i)

Situation (iii) L=3, n=4, " " " " " " " " "

Situation (iv) L=3, n=8, " " " " " " " " "

Situation (v) L=3, n=12, " " " " " " " " "

Situation (vi) L=3, n=16, " " " " " " " " "

Situation (vii) L=3, n=100, " " " " " " " " "

Situations (viii-1)-(viii-12) correspond, for L=4 strata, to the situations described in Situations (i-1)-(i-12).

Situations (ix-1)-(ix-12) correspond, for L=15 strata, to the situations described in Situations (i-1)-(i-12).

The method used for generating the random pair (x_{ij},y_{ij}) is not described in detail here. All normal deviates were independently generated by the method of Marsaglia (1973).

The validity of the sampling experiments were checked in a variety of ways. These are described by Lemeshow (1976). There was close agreement between theoretical and simulated results providing reassuring evidence that the simulations reported here operated correctly.

5. Results

To assess the relative advantages of the different estimates of β defined in (2.2), (2.3) and (2.4), it is necessary to compute the population value of β in each of the sampling situations under consideration. This value, in terms of the parameters fixed in each stratum, is as follows:

$$\beta = \frac{\sigma_{xy}}{\sigma_x^2} = \frac{\sum_{i=1}^{L}\sigma_{xy}^{(i)} + \sum_{i=1}^{L}(\mu_x^{(i)}-\mu_x)(\mu_y^{(i)}-\mu_y)}{\sum_{i=1}^{L}\sigma_{xx}^{(i)} + \sum_{i=1}^{L}(\mu_x^{(i)}-\mu_x)^2} . \quad (5.1)$$

The discussion here is restricted to the situations with L=3. Situations (i-1)-(i-4) correspond to having $\mu_x^{(i)}=\mu_x^*$ and $\sigma_{xx}^{(i)}=\sigma_{xx}^*$, i=1,...,L. From Theorem 1 we expect all three estimates of β to be consistent and, since sampling is from bivariate normal populations, $\hat{\beta}_3$ should be the minimum variance unbiased estimate. Situations (i-5)-(i-8) have $\mu_x^{(i)}=\mu_x^*$, i=1,...,L. From Theorem 1, only $\hat{\beta}_1$ and $\hat{\beta}_2$ will be consistent estimates of the population β. Situations (i-9)-(i-12) correspond to all strata having different means and variances. From Theorem 1, only $\hat{\beta}_1$ should be consistent for β.

Values of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ were calculated for each of the twelve situations (i-1)-(i-12). The means, $\hat{E}(\beta_i)$, and variances, $\hat{V}(\beta_i)$, i=1,2,3, from the 1200 repetitions were computed. Table 1 presents these results along with the value of β computed using (5.1).

It is clear that these results agree with the conclusions of Theorem 1. In Situations (i-1)-(i-4), $\hat{E}(\beta_i)\approx\beta$, i=1,2,3. That is, each of the β_i appears to be an unbiased estimate of β. In addition, except in the rather uninteresting Situation (i-1) in which all strata are identical, $\hat{\beta}_1$ has larger variances than $\hat{\beta}_2$ or β_3. In Situations (i-2) and (i-4) where each stratum has a different slope, $\hat{\beta}_3$ has the minimum variance of the unbiased estimates. In Situations (i-5)-(i-8), β_1 and $\hat{\beta}_2$ are consistent by Theorem 1 and, in the sampling experiment, $\hat{E}(\hat{\beta}_i)\approx\beta$, i=1,2. However, in Situations (i-6) and (i-8) in which the strata slopes are not all equal, $\hat{E}(\hat{\beta}_3)\neq\beta$. Note that $\hat{V}(\hat{\beta}_2)$ never exceeds $\hat{V}(\beta_1)$. In Situations (i-9)-(i-12), $\hat{\beta}_1$ appears unbiased and has smaller variance than the others. $\hat{\beta}_2$ and $\hat{\beta}_3$ are biased whenever the strata have different linear regressions.

The conslusions of Theorem 1 and this sampling experiment is that if an estimate of the population β is desired, $\hat{\beta}_1$ is consistent and asymptotically unbiased. The variance of $\hat{\beta}_1$ is generally larger than the variances of the alternative estimates. If it can be assumed that the strata have the same mean for the independent variable, then, in the sampling experiment, $\hat{\beta}_2$

is a better estimate of β than $\hat{\beta}_1$ since it is also unbiased but much less variable. Use of $\hat{\beta}_3$ is not recommended since the necessary assumptions may be too restrictive.

Note that β is estimated by computing deviations about some mean. When this mean is a within-group mean as in $\hat{\beta}_2$ or $\hat{\beta}_3$, at least two observations are needed in each of the 2 groups which were established for use with the balanced half-sample method. Moreover, it is of interest to determine the minimum number of observations per stratum necessary to introduce some degree of stability into the variance estimation calculations.

As described earlier, the sampling experiment was repeated for the L=3 strata situations, with n=4, 8, 12, 16, 20 or 100 observations per stratum. By taking at least n=4 observations per stratum, we are assured of having at least 2 observations in each of the established groups.

Table 2 presented the absolute relative bias and variance of the two estimates of $V(\hat{\beta}_i)$, i=1,2,3, in Situation (i-2), (iii-2), (iv-2), (v-2), (vi-2) and (vii-2). That is, we present the results for those situations in which the means and variances of the independent variable are the same for each of the L=3 strata but, while the intercepts are the same, the slopes differ for the linear regressions in each stratum.

In Table 2 we see that selecting small samples from each stratum can result in estimates of extremely low precision and high variability. For instance, when $\hat{\beta}_1$ was used to estimate β with n=4 observations per stratum, the balanced half-sample estimate of $\hat{V}(\hat{\beta}_1)$ missed this target variance by 93%. The variance of these estimates were quite high. Using an absolute relative bias of .05 as an acceptable level of precision, we observe that, when using $\hat{\beta}_1$ to estimate β, more than 100 observations per stratum were needed. When using $\hat{\beta}_2$ as an estimate of β, at least 20 observations should be used. About 100 observations should be used for the variance estimates when $\hat{\beta}_3$ is used to estimate β. As a general rule, at least n=20 observations from each stratum are needed in order to introduce stability into the variance estimates. When using $\hat{\beta}_3$, variance estimates show greater sensitivity to small n, and a larger n should be selected if possible. All twelve parameterizations were studied in the same way with similar results.

A study of the other situations indicates that contrary to the results of the linear case (Lemeshow and Epp (1977)) and the combined ratio estimates (Lemeshow and Levy (1977)), the two half-sample estimates of the variance of $\hat{\beta}_i$ differed. In the "balanced" situation (i.e., L=3), neither one of variance estimates was consistantly better or worse than the other over all situations considered.

In the "full-matrix" situation (i.e., L=4) it was noted that $\hat{E}[\hat{V}_{B1}(\hat{\beta}_i)] < \hat{E}[\hat{V}_{B2}(\hat{\beta}_i)]$, i=1,2,3. In fact, $\hat{V}_{B1}(\hat{\beta}_i)$ always had a negative bias. However, the variance and mean square error using this balanced half-sample estimate were never greater that the corresponding measurement for $\hat{V}_{B2}(\hat{\beta}_i)$. This corresponds to results for linear and combined ratio estimates presented in the references cited above. When L=15 strata were used, the results described above for L=3 appear to apply.

6. Conclusions

The sampling experiments have demonstrated that each of the variance estimation techniques appear to have the potential of providing usable estimates of the target variance for the slope provided a large enough sample is selected from each stratum. This is very different than previous results for using the balanced half-sample method in the linear case or with the combined ratio estimate. There, even as few as two observations per stratum would result in minimal bias. Here, however, at least 20 observations per stratum are necessary in order to introduce some stability into the variance estimation process for the slope of the linear regression.

Acknowledgements

This research is based, in part, on the author's Ph.D dissertation, which was undertaken at UCLA. Computing assistance was obtained from the Health Sciences Computing Facility (UCLA) under NIH Special Resources Grant RR-3.

References

Lemeshow, S., An Evaluation of the Performance of the Balanced Half-Sample and Jackknife Variance Estimation Techniques. Unpublished doctoral dissertation, University of California, Los Angeles (1976).

Lemeshow, S. and Epp, R., "Properties of the Balanced Half-Sample and Jackknife Variance Estimation Techniques in the Linear Case." To be published in Communications in Statistics A, Vol. 6, Issue 13 (1977).

Lemeshow, S. and Levy, P. "Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Sampling Units Per Stratum - A Comparison of Balanced Replication and Jackknife Techniques." Submitted for publication.

Marsaglia, G., Ananthanarayanan, K. and Paul, N., How to Use the McGill Random Number Package 'SUPER-DUPER,' four pages typed description, School of Computer Sciences, McGill University, Montreal, Quebec (1973).

McCarthy, P.J., Replication. An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. NCHS, Series 2, No. 14 (1966).

McCarthy, P.J., Pseudoreplication: Half Samples. Review of the International Statistical Institute, Vol. 37, No. 3: 239-264 (1969).

_____, Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique. Vital and Health Statistics. Series 2, No. 31 (1969).

Table 1: Population β for each of the twelve situations with L=3 strata described in means, $\tilde{E}(\beta_i)$, and variances, $\tilde{V}(\hat{\beta}_i)$, of $\hat{\beta}_i$, i=1,2,3, as estimated from the sampling experiment are presented.

		Means			Variances		
Situation	Pop β	$\tilde{E}(\hat{\beta}_1)$	$\tilde{E}(\hat{\beta}_2)$	$\tilde{E}(\hat{\beta}_3)$	$\tilde{V}(\hat{\beta}_1)$	$\tilde{V}(\hat{\beta}_2)$	$\tilde{V}(\hat{\beta}_3)$
(i- 1)	1.00000	1.00078	1.00034	1.00027	.00410	.00425	.00459
(i- 2)	2.00000	2.00553	2.00030	2.00095	.33297	.04640	.02094
(i- 3)	1.00000	1.00092	.99994	1.00073	.01523	.00417	.00453
(i- 4)	2.00000	2.01199	1.99562	2.00025	.50021	.04376	.02234
(i- 5)	1.00000	1.00016	.99926	1.00062	.00457	.00476	.00424
(i- 6)	2.33333	2.34308	2.32874	1.99972	.18958	.05090	.02015
(i- 7)	1.00000	.99703	.99811	.99680	.01089	.00503	.00477
(i- 8)	2.33333	2.30799	2.31779	1.99410	.27262	.05287	.02220
(i- 9)	1.00000	.99936	1.00159	1.00132	.00110	.00479	.00430
(i-10)	3.55882	3.56746	2.32645	2.00300	.01732	.05320	.02076
(i-11)	1.14706	1.14880	1.00009	1.00123	.00122	.00490	.00447
(i-12)	3.70588	3.72134	2.32342	2.00092	.01911	.04999	.02025

Table 2: Results of sampling experiment in which n=4, 8, 12, 16, 20 or 100 observations were selected from each of L=3 strata. Estimated values based on the sampling experiment are presented for absolute relative bias, and variance using the three methods of estimation. In all cases, $\mu_x^{(i)}=5$, $\sigma_{xx}^{(i)}=1$, $\alpha^{(i)}=0$, $\rho^{(i)}=.9$, i=1,2,3. $\beta^{(1)}=1$, $\beta^{(2)}=2$, $\beta^{(3)}=3$.

Method 1

	Absolute Relative Bias (I)		Variance (I)	
n	B1	B2	B1	B2
4	.93	1.11	35.04468	51.53947
8	.34	.38	1.85307	1.99540
12	.24	.27	.54089	.58215
16	.13	.15	.26443	.27707
20	.10	.12	.14486	.14973
100	.07	.08	.00460	.00462

Method 2

	Absolute Relative Bias (I)		Variance (I)	
n	B1	B2	B1	B2
4	1.99	3.15	3.09580	5.48564
8	.20	.42	.02398	.02919
12	.16	.31	.00537	.00626
16	.01	.11	.00308	.00349
20	.01	.06	.00185	.00197
100	.02	.04	.00005	.00005

Method 3

	Absolute Relative Bias (I)		Variance (I)	
n	B1	B2	B1	B2
4	*	*	**	**
8	1.53	2.87	.28248	.84464
12	.49	.92	.00540	.00979
16	.23	.51	.00201	.00323
20	.25	.49	.00083	.00123
100	.03	.06	.00001	.00001

*Absolute relative bias >100

**Variance >100

THE BEHAVIOR OF BALANCED HALF-SAMPLE VARIANCE ESTIMATES FOR LINEAR AND COMBINED RATIO ESTIMATES WHEN STRATA ARE PAIRED TO FORM PSEUDO-STRATA

Edward J. Stanek III and Stanley Lemeshow, University of Massachusetts, Amherst

Summary

Expressions are developed illustrating the effect that pairing of strata into pseudo-strata has on balanced half-sample estimates of the variance for estimates of the stratified mean. The development is extended to variance estimates of the combined ratio estimate by using a Monte Carlo sampling experiment. An evaluation is then made of the effect that pairing strata into pseudo-strata has on variance estimates for heights and weights from Cycle 2 of the Health Examination Survey.

The results of this investigation demonstrate that in certain situations, pairing of strata into pseudo-strata can result in highly variable and biased variance estimates of linear and combined ratio estimates. Variance estimates of heights and weights in Cycle 2 of the Health Examination Surveys were found to be insensitive to different pairings of strata, regardless of whether or not the pairings were done in a homogeneous fashion.

1. Introduction

Variance estimation has been a problem in the past when dealing with large, complex surveys. Exact expressions for the variance of parameter estimates in such surveys are often unknown and intractable. The balanced half-sample method has grown to be a popular method of estimating variances and is currently used by the National Center for Health Statistics (NCHS) in its Health Examination Survey (HES) and Health Interview Survey (HIS).

As originally developed by McCarthy (1966), the balanced half-sample method requires what may be thought of as the simplest of all designs: a simple stratified sample with two independent observations per stratum. Often, large scale sample surveys are designed so that one cluster of observations (PSU) is selected from each stratum. In these surveys, in order to conform to the balanced half-sample method, strata were paired forming the "pseudo-strata" that were used in the subsequent variance estimation. There were two PSU's per pseudo-stratum. These two PSU's were treated as the two independent observations required by the balanced half-sample variance estimation method.

Properties of the balanced half-sample method have been investigated by numerous authors for a variety of estimates (McCarthy (1966, 1969), Kish and Frankel (1968, 1970), Frankel (1971), Bean (1975), Lemeshow and Epp (1977), Lemeshow and Levy (1977), Lemeshow and Stanek (1977)). All of these studies have started with the assumption upon which the balanced half-sample technique is based: there are two independent observations per stratum. Kish and Frankel (1968, p. 21) suggest that "a model of two independent primary selections per stratum is probably the most basic design that conforms adequately." Nevertheless, all of the investigators are cognizant of the fact that in situations where the balanced half-sample technique is being applied, the assumption of two independent observations per stratum is seldom met. There is no documented evidence that the techniques used in forming pseudo-strata in the HES produce observations which conform adequately to the assumption of independent observations.

This paper investigates the effect pairing of strata into pseudo-strata has on estimates of the variance for a linear stratified mean, and a combined ratio estimate. The results of this investigation are then placed in the context of height and weight measurements made on children in Cycle 2 of the HES.

2. The Linear Case

Consider a situation with $2L$ strata having means μ_j, $j=1,\ldots,2L$ and common variance σ^2. As shown by Stanek (1977), for a particular arrangement of the $2L$ strata into L pseudo-strata, the expected value of the variance estimate is given by

$$E\left[\hat{var}(\bar{x}_{st})\right] = \frac{\sigma^2}{2L} + \frac{1}{4L^2}\sum_{i=1}^{L}(\mu_{ij}-\mu_{ij'})^2 \qquad (2.1)$$

and the variance of this estimate is given by

$$var\left[\hat{var}(\bar{x}_{st})\right] = \frac{\sigma^4}{2L^3} + \frac{\sigma^2}{2L^4}\sum_{i=1}^{L}(\mu_{ij}-\mu_{ij'})^2 \qquad (2.2)$$

where j and j' represent the two original strata which were combined to form the i^{th} pseudo-stratum. In these expressions, we appeal to the fact that for linear estimates, the balanced half-sample method produces variance estimates identical to usual stratified sampling formulae. Both Kish (1965, p. 283) and Cochran (1963, p. 141) have noted similar effects in a slightly different context.

Clearly, this process of forming pseudo-strata has certain inherent dangers. Normally, if many of the pseudo-strata in a particular arrangement were comprised of heterogeneous strata, the resulting estimates of variance could be highly biased and variable.

3. The Combined Ratio Estimate

Through the use of a sampling experiment, Stanek (1977) investigated the effect that pairing strata to form pseudo-strata has on estimates of variance for combined ratio estimates. The balanced half-sample estimate considered was defined to be

$$\hat{V}(r) = \frac{1}{M}\sum_{i=1}^{M}(\hat{r}^{(m)}-\hat{r})^2 \qquad (3.1)$$

where $\hat{r}^{(m)}$ is the estimate of the combined ratio estimate for the m^{th} half-sample and $\hat{r}$ is the estimate obtained using all the sample observations.

Specifically, let us now assume we have 6 strata with two pairs of observations per stratum.

	Strata				
1	2	3	...	6
(x_{11}, y_{11})	(x_{21}, y_{21})	(x_{31}, y_{31})	...	(x_{61}, y_{61})
(x_{12}, y_{12})	(x_{22}, y_{22})	(x_{32}, y_{32})	...	(x_{62}, y_{62})

The combined ratio estimate of $R = Y/X$ is

$$\hat{R} = \frac{\sum_{i=1}^{6} w_i \bar{y}_i}{\sum_{i=1}^{6} w_i \bar{x}_i}$$

where $\bar{y}_i = \sum_{j=1}^{2} \frac{y_{ij}}{2}$ and $\bar{x}_i = \sum_{j=1}^{2} \frac{x_{ij}}{2}$.

Throughout the sampling experiment, we assume that we have normally distributed strata of equal weights. We restric ourselves to a situation where for each stratum, $Y = R^{(i)}X$, i.e., the regression line of Y on X for each stratum passes through the origin. With this restriction $\hat{R}$ becomes an unbiased estimate of the ratio R. We will assume the correlation coefficient across all strata is constant and equal to $\rho=0.9$. Without loss of generality, the mean of X is held constant across the strata. We will also assume that the variance of X is held constant across the strata. The four values of the variance of X as considered in the sampling experiment were $\sigma_x^2=0.1, 0.001, 0.001, 0.0001$.

The strata were generated so that the first and second strata had the same set of population parameters with ratio $R^{(1)}$, the third and fourth strata had the same set of population parameters different from the first with ratio $R^{(2)}=dR^{(1)}$, the fifth and sixth strata has the same set of population parameters different from the previous four with ratio $R^{(3)}=dR^{(2)}$. A balanced half-sample estimate of the variance of the combined ratio estimate was made on the basis of all 6 strata. The 6 strata were then paired in all possible ways (i.e., $\frac{6!}{2^3 3!} = 15$) to form 3 pseudo-strata. Since pairs of strata had the same parameters, these 15 different pairings represented only 5 distinctly different situations. If we assume the strata were paired at random to form pseudo-strata, each of the 15 arrangements would be equally likely. The 5 distinct arrangements of the pseudo-strata along with their probability of occurrence are given below.

k^{th} Arrangement	Prob (k)	Pseudo-strata 1		Pseudo-strata 2		Pseudo-strata 3	
1	1/15	A	A	B	B	C	C
2	2/15	A	B	A	B	C	C
3	2/15	A	C	B	B	A	C
4	2/15	A	A	B	C	B	C
5	8/15	A	B	A	C	B	C

To obtain a variance estimate using one arrangement of the pseudo-strata, each pair of observations generated per stratum was averaged and then considered as an observation for the pseudo-strata. In such a manner, two pairs of observations were created for each pseudo-stratum from the original data. A balanced half-sample estimate of the variance of the combined ratio estimate was then made on the basis of the 3 pseudo-strata. The constant "d" was chosen such that $\sum_{i=1}^{3} R^{(i)}=5$. This restriction kept the true ratio R constant throughout the sampling experiment. The values of d considered along with the corresponding values of $R^{(1)}$ are given below.

d	$R^{(1)}$
1	5
1.001	4.9950
1.050	4.7581
1.100	4.5317
1.200	4.1209

Balanced half-sample variance estimates produced from the 5 distinct arrangements of pseudo-strata were compared to estimates obtained based on all six strata.

The results of the sampling experiment closely followed results that were derived in the linear case. As the variance of X became smaller, the relative-bias of variance estimates became larger for a given pairing of the strata. As strata were paired more heterogeneously, the rel-bias of balanced half-sample variance estimates increased. As the difference in strata ratios became larger, balanced half-sample variance estimates became more biased. These results are presented in Table 1.

Similar results held for the change in the variance of the variance estimates (Stanek, 1977).

4. Application to the HES

The increased bias and variability of balanced half-sample variance estimates based on pseudo-strata is only of interest to the extent that surveys actually use this variance estimation method in practice. As has been mentioned earlier, the Health Examination Survey has paired strata into pseudo-strata and used the balanced half-sample method of variance estimation in the past. The Health Interview Survey is presently pairing strata into pseudo-strata and using the balanced half-sample method to estimate the variance. This section will consider

the effect that pairing of strata has on variance estimates for data collected in Cycle 2 of the HES. Specifically, we will investigate the effect that pairing of strata has on variance estimates of height and weight for white children from the ages of 6 to 11. We will use as a reference, variance estimates published on height and weight from the HES. (NCHS, 1972, p. 42.)

The HES was designed with 40 strata formed "in a manner which maximized the degree of homogeneity within superstrata with respect to population size, geographic proximity, degree of industrialization, and degree of urbanization." (NCHS, 1973, p. 6.) One ultimate cluster of observations was chosen from each stratum, and approximately 180 subsequent observations were taken within the ultimate cluster. An estimate of the effect that pairing of strata has on variance estimates cannot be made through a comparison of balanced half-sample variance estimates based on 40 strata with balanced half-sample variance estimates based on 20 pseudo-strata. Differences in these two estimates may stem from the effects of pairing or from the effects of the covariance of observations within the ultimate cluster. Estimates of the effect of pairing strata into pseudo-strata can be made, however, through a comparison of variance estimates under a number of plausible rearrangements of the strata. It is in this manner that we will assess the effect that the formation of pseudo-strata had on variance estimates for heights and weights of children.

The investigators in the HES were cognizant of the potential dangers in forming pseudo-strata to estimate the variance. An effort was made to pair strata as homogeneously as possible. They were paired "on the basis of (1) some subjective determination of the homogeneity of the population in which the primary considerations were population density, region, rate of growth, and industry and (2) concern that strata of approximately equal size would be paired." (NCHS, 1973, p. 27.)

Population density along with rate of growth was defined on a sliding scale for each of 4 geographic regions. The resulting pairings of strata into pseudo-strata for Cycle 2 of the HES are given in Table 2. Clearly, the specific pairing of strata into pseudo-strata as in Cycle 2 of the HES was not the only possible way of forming homogeneous pseudo-strata. A comparison of variance estimates resulting from the HES's pairing with estimates based on other possible homogeneous arrangements (arrangements A-C) and an extremely heterogeneous arrangement (arrangement D) will give a measure of the effect pairing strata has on variance estimates. Details of the criteria for alternative pairing of strata are given by Stanek (1977). It should be noted that the first pairing given in Table 2 is the same as was used by the HES except that when forming their estimates, the HES divided the self-representing strata (pairs 17 through 20) by segments to form new strata. These new strata were used by the HES as pairs 1 through 4 for variance estimation.

Table 3 presents estimates of the standard error of heights and weights of white boys and girls (6-11 years of age) based on the different rearrangements of strata into pseudo-strata. It is important to note that in this regard, there is no asymptotic variance estimate or target value with which to compare variance estimates based on various arrangements of pseudo-strata. Differences in variance estimates for different arrangements may be due to the heterogeneity of strata composing the pseudo-strata, or due to the random variability of samples selected. If consistently large differences were to occur in variance estimates for different arrangements of pseudo-strata, we would suspect that variance estimates were sensitive to pseudo-strata formation. A comparison of estimates based on these alternative homogeneous pairings of strata into pseudo-strata indicates whether variance estimates are highly sensitive to strata pairing. A comparison of variance estimates made when strata are paired heterogeneously with variance estimates made with the HES's pairing should detect gross effects due to the formation of pseudo-strata.

Table 3, which presents a comparison of the standard error estimates for 12 age-sex categories, demonstrates the insensitivity of estimates to alternative pairings of the strata. In most cases, estimates of the standard error based on different arrangements of pseudo-strata differed from published estimates by less than 25%. Due to the small value of the coefficient of variation, this difference would seldom be of practical significance. The largest differences from published estimates of the standard error occurred for arrangements C and D for height of 10 year old white girls. In those cases, estimates of the standard error differed from published estimates by 112%.

Estimates of the standard error based on arrangement D of the pseudo-strata were anticipated to be larger than estimates based on other arrangements. The hypothesis that standard error estimates based on arrangement D of the pseudo-strata were equal to standard error estimates based on another arrangement of pseudo-strata was tested against the one sided alternative that standard error estimates based on arrangement D were greater than standard error estimates based on the other arrangement. The tests were based on Freedman rank sums. (Hollander and Wolfe, 1973, p. 155.) The tests were made on standard error estimates for height and for weight. In each test, the published variance estimates were included as a comparison group. In none of the tests was the hypothesis of equality of standard error estimates rejected in favor of the one sided alternative at $\alpha=.05$.

5. Conclusions

In summary, balanced half-sample estimates of the variance of mean heights and weights of children were not found to be highly dependent on the arrangement of pseudo-strata for the specific age-color-sex classes considered from Cycle 2 of the HES. Differences did occur due to the arrangement of strata into pseudo-strata but

these differences were no greater than were found by using the complements of the appropriate Plackett-Burman matrices. (See Stanek, (1977).) Rarely did an alternative estimate of the standard error exceed twice the published estimate. Since the coefficient of variation for heights and weights was extremely small for these measurements, differences in estimates of the standard error may not be of practical significance. Only a limited number of situations were considered using HES data. Sampling experiment results demonstrated that in certain situations, large biases could be introduced through the formation of pseudo-strata. Variables whose sample measurements differ widely from stratum to stratum will be more susceptible to these biases. Caution should be exercised in using such variance estimates. Care should be taken to avoid the design of surveys which face this problem in the future.

Acknowledgements

This research is based, in part, on the thesis of the first author at the University of Massachusetts, Amherst. Computing assistance was obtained from the University Computing Center at the University of Massachusetts.

Bibliography

Bean, Judy A. (1975). "Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution." Data Evaluation and Methods Research. NCHS, Series 2, #65. DHEW Publication No. (HRA) 75-1339.

Cochran, W.G. (1962). Sampling Techniques. New York: John Wiley and Sons.

Hollander, M. and Wolfe, D.A. (1973). Non Parametric Statistical Methods. New York: John Wiley and Sons.

Kish, Leslie (1965). Survey Sampling. New York: John Wiley and Sons.

Kish, Leslie and Frankel, Martin R. (1968). "Balanced Repeated Replication for Analytical Statistics." Proceedings of the Social Statistics Section of American Statistical Association: 2-10.

Kish, Leslie and Frankel, Martin R. (1970). "Balanced Repeated Replication for Standard Errors." Journal of the American Statistical Association, Vol. 65, No. 331, pp. 1071-1094.

Lemeshow, S. and Epp, R. (1977). "Properties of the Balanced Half-Sample and Jackknife Variance Estimation Techniques in the Linear Case." Communications in Statistics A, Vol. 6, Issue 13.

Lemeshow, S. and Levy, P. (1977). "Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Sampling Units Per Stratum - A Comparison of Balanced Replication and Jackknife Techniques." Submitted for publication.

Lemeshow, Stanley and Stanek, Edward J. (1977). "Estimating the Variance of the Slope of a Linear Regression in a Stratified Random Sample with the Balanced Half-Sample Technique." Submitted for publication.

McCarthy, Philip J. (1966). "Replication. An Approach to the Analysis of Data from Complex Surveys." Vital and Health Statistics. NCHS, Series 2, #14.

McCarthy, Philip J. (1969). "Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique." Vital and Health Statistics. NCHS, Series 2, #31.

National Center for Health Statistics (1972). "Height and Weight of Children: Socio-economics Status, United States." Vital and Health Statistics. NCHS, Series 11, #119.

National Center for Health Statistics (1973). "Sample Design and Estimation Procedures for a National Health Examination Survey of Children." Vital and Health-Statistics. NCHS, Series 2, #43.

Plackett, R.L. and Burman, J.P. (1946). "The Design of Optimum Multifactorial Experiments." Biometrika, Vol. 33 (pt. IV): pp. 305-325.

Stanek, E.J. (1977). "The Properties of Balanced Half-Sample Variance Estimates in Complex Surveys when Strata are Paired to Form Pseudo-Strata." Biostatistics Technical Reports, No. 77-8, Division of Public Health, School of Health Sciences, University of Massachusetts/Amherst.

TABLE 1

Results of a sampling experiment for $\hat{V}_2(\hat{R})$ in which N=2 observations were selected from each of L=6 strata. The strata were paired in k=1,...,5 arrangements of pseudo-strata. Six multiplicative factors, d, (1, 1.001, 1.010, 1.050, 1.100, and 1.200) and four rel-variances of X, (0.1, 0.01, 0.001, 0.0001) were used. In each case, ρ=0.9.

		Rel-Bias					
Rel-Var (X)	k	d					
		1	1.001	1.010	1.050	1.100	1.200
.1000	1	-.926E-01	-.924E-01	-.909E-01	-.842E-01	-.760E-01	-.609E-01
.1000	2	-.646E-01	-.642E-01	-.583E-01	.110E-01	.173E+00	.618E+00
.1000	3	-.537E-01	-.528E-01	-.361E-01	.225E+00	.918E+00	.310E+01
.1000	4	-.361E-01	-.361E-01	-.346E-01	.217E-01	.199E+00	.843E+00
.1000	5	-.513E-01	-.510E-01	-.416E-01	.139E+00	.639E+00	.224E+01
.0100	1	-.586E-01	-.585E-01	-.568E-01	-.501E-01	-.427E-01	-.318E-01
.0100	2	-.541E-01	-.525E-01	-.143E-01	.618E+00	.222E+01	.675E+01
.0100	3	-.446E-01	-.419E-01	.791E-01	.261E+01	.969E+01	.324E+02
.0100	4	-.405E-02	-.440E-02	.171E-01	.642E+00	.257E+01	.947E+01
.0100	5	-.257E-01	-.241E-01	.635E-01	.194E+01	.723E+01	.243E+02
.0010	1	-.502E-01	-.501E-01	-.484E-01	-.415E-01	-.342E-01	-.239E-01
.0010	2	-.508E-01	-.440E-01	.259E+00	.631E+01	.222E+02	.674E+02
.0010	3	-.424E-01	-.273E-01	.109E+01	.261E+02	.969E+02	.325E+03
.0010	4	.730E-02	.793E-02	.261E+00	.672E+01	.263E+02	.961E+02
.0010	5	-.188E-01	-.787E-02	.825E+00	.196E+02	.727E+02	.244E+03
.0001	1	-.478E-01	-.476E-01	-.459E-01	-.389E-01	-.316E-01	-.216E-01
.0001	2	-.499E-01	-.963E-02	.278E+01	.624E+02	.220E+03	.672E+03
.0001	3	-.418E-01	.802E-01	.110E+02	.260E+03	.968E+03	.325E+04
.0001	4	.110E-01	.317E-01	.270E+01	.679E+02	.265E+03	.965E+03
.0001	5	-.168E-01	.744E-01	.825E+01	.195E+03	.726E+03	.244E+04

TABLE 2

Arrangements of strata into pseudo-strata

	HES	A	B	C	D
Boston, Mass.	1	1	3	3	1
Newark, N.J.	1	4	10	10	2
Jersey City, N.J.	2	4	10	10	3
Allentown, Pa.	2	3	9	6	4
Poughkeepsie, N.Y.	3	5	14	11	5
Hartford, Conn.	3	3	8	6	6
Columbia, S.C.	4	7	11	12	7
Charleston, S.C.	4	8	11	12	8
Marked Tree, Ark.	5	10	17	15	4
Georgetown, Del.	5	10	17	16	9
Barbourville, Ky.	6	9	15	15	10
West Liberty, Ky.	6	9	15	16	3
Cleveland, Ohio	7	12	7	5	11
Minneapolis, Minn.	7	12	8	7	2
Lapeer, Mich.	8	15	18	17	11
Ashtabula, Ohio	8	14	18	18	12
San Francisco, Calif.	9	17	6	5	12
Denver, Colo.	9	18	7	8	13
Prowers, Colo.	10	20	19	20	1
Mariposa, Calif.	10	19	19	19	14
Atlanta, Ga.	11	6	4	9	15
Houston, Tex.	11	17	6	8	10
Des Moines, Iowa	12	13	12	14	13
Wichita, Kans.	12	18	13	13	16
Birmingham, Ala.	13	7	9	9	17
Grand Rapids, Mich.	13	13	12	14	18
Clark, Wis.	14	15	20	17	6
Grant, Wash.	14	20	16	19	15
Portland, Maine	15	5	14	11	19
Ottumwa, Iowa	15	14	20	18	17
Sarasota, Fla.	16	8	13	13	20
Brownsville, Tex.	16	19	16	20	20
Philadelphia, Pa.	17	1	3	3	16
Baltimore, Md.	17	6	4	7	18
Chicago, Ill.	18	11	5	4	5
Detroit, Mich.	18	11	5	4	19
Los Angeles, Calif.	19	16	2	2	7
Los Angeles, Calif.	19	16	2	2	14
New York, N.Y.	20	2	1	1	8
New York, N.Y.	20	2	1	1	9

Table 3

Standard error estimates for heights in CM and weights in KG for white boys and girls in 6 age categories from Cycle 2 of the health examination survey. Standard error estimates are presented for each of 5 alternative arrangements of strata into psuedo-strata, (HES, A, B, C, and D). The published st. error-estimates and mean heights are given for comparison. (NCHS, 1972)

	HEIGHT											
	Boys Age						Girls Age					
	6	7	8	9	10	11	6	7	8	9	10	11
Published Mean Height	118	124	130	135	140	146	118	123	129	135	141	147
Published St. Error	.30	.38	.29	.50	.37	.30	.32	.17	.39	.36	.34	.37
HES	.37	.35	.26	.46	.37	.38	.28	.21	.51	.44	.47	.37
A	.37	.35	.26	.44	.23	.35	.31	.26	33	.45	.67	.45
B	.41	.32	.25	.46	.40	.39	.41	.21	.25	.45	.39	.24
C	.28	.37	.25	.39	.36	.31	.22	.25	.30	.42	.72	.36
D	.29	.36	.31	.46	.36	.33	.31	.22	.32	.37	.72	.29
	WEIGHT											
	Boys Age						Girls Age					
	6	7	8	9	10	11	6	7	8	9	10	11
Published Mean Height	22	25	28	31	34	39	22	24	28	31	35	40
Published St. Error	.17	.21	.25	.47	.30	.40	.25	.20	.26	.43	.44	.36
HES	.19	.22	.28	.46	.31	.46	.25	.18	.28	.48	.47	.42
A	.22	.20	.24	.38	.22	.54	.21	.22	.27	.40	.52	.41
B	.20	.22	.23	.31	.28	.43	.27	.24	.21	.51	.40	.53
C	.15	.19	.28	.37	.29	.36	.25	.26	.24	.45	.64	.44
D	.16	.23	.29	.35	.36	.46	.24	.20	.27	.46	.47	.31

EVALUATION OF BALANCED HALF-SAMPLE AND JACKKNIFE ESTIMATES OF COMBINED RATIO ESTIMATES FOR NON-NORMALLY DISTRIBUTED POPULATIONS

David Hislop, U.S. Bureau of the Census
Stanley Lemeshow, University of Massachusetts/Amherst

Abstract

The performance of the Balanced Half-Sample and Jackknife methods for estimating the variance of the combined ratio estimate is evaluated using artificially generated non-normally distributed populations. In a Monte-Carlo design two variations of the balanced half-sample technique and three variations of the jackknife are examined within a framework which permits the manipulation of the underlying distributions of the random variables. The variance estimates are empirically evaluated using one symmetric and two skewed non-normal distributions which are related to the well documented results based on the normal distribution.

The results of this investigation demonstrate that the variance estimates of the combined ratio estimate are highly biased and quite unstable when the underlying distribution is non-normal and the balanced half-sample method is used. The jackknife estimates are shown to be considerably better, particularly when estimates are desired for domains of interest containing few observations.

This paper examines the performance of the performance of the Balanced Half-Sample and Jackknife techniques for estimating the variance of the combined ratio estimate when the underlying distributions of the random variables under consideration are non-normal. Previous work by McCarthy (1966), Frankel (1971), Bean (1975), Lemeshow and Epp (1977), and Lemeshow and Levy (1977) concerned the ability of these techniques to accurately estimate the variance of both linear and non-linear estimates from complex multi-stage survey designs such as the Health Examination Survey (HES) and Health Interview Survey (HIS) of the National Center for Health Statistics (NCHS). To date, all research in this area has dealt with populations whose distributional characteristics were either unknown or specifically normal. This study will evaluate the two variance estimation techniques by means of Monte-Carlo methods in which samples are selected from populations whose parameters are precisely specified.

1. Background

The Balanced Half-Sample technique is currently used by the NCHS for variance estimation of population estimates from the HES and HIS. The Jackknife, originally due to Quenouille (1956), has been gaining popularity in recent years. Both methods have been thoroughly examined by Lemeshow and Epp (1977) and Lemeshow and Levy (1977) in Monte-Carlo sampling experiments under the assumption of normality.

The nature of much of the data collected by sample survey methods, especially in the health sciences, is well documented. Data gathered in the HES, for example, are in many instances found to be non-normally distributed. Clearly, evaluation of techniques designed specifically for data from such complex sample surveys as HES should include examination of specifically non-normal populations.

Research into the Balanced Half-Sample and Jackknife variance estimation methods has been in response to the fact that precise formulae for the variance of non-linear parameters in highly complex surveys do not exist. The effect of non-normality on the ability of the techniques under consideration to provide precise variance estimates is, to date, unexplored.

2. The Sampling Experiment

To obtain the sample, observations are drawn at random from L strata of infinite size. The distribution of these observations is known and specified. This sample is used to estimate the population ratio. Subsequent samples are drawn and from them estimates are made of this population parameter. This process is repeated $M=1000$ times and the distribution of the sample estimates is studied. This Monte-Carlo computer simulation is patterned after the work of Lemeshow and Levy (1977).

For the two variations of the Balanced Half-Sample technique considered, half-sample estimates are constructed such that,

$$\hat{R}_p = \frac{\sum_{h=1}^{L} (\delta_{ph} X_{h1} + (1-\delta_{ph}) X_{h2})}{\sum_{h=1}^{L} (\delta_{ph} Y_{h1} + (1-\delta_{ph}) Y_{h2})} ,$$

is the p^{th} half-sample estimate of R, the population ratio, where δ_{ph} is an element from the p^{th} row and h^{th} column of the appropriate matrix given by Plackett and Burman (1946), and (X_{hi}, Y_{hi}) is the i^{th} sample observation from the h^{th} stratum.

The two variations of the Balanced Half-Sample variance estimate considered here are,

$$(1) \qquad \hat{V}_{B1}(\hat{R}) = \frac{1}{\ell} \sum_{p=1}^{\ell} (\hat{R}_p - \bar{R})^2$$

where $\bar{R} = \frac{1}{\ell} \sum_{p=1}^{\ell} \hat{R}_p$ and ℓ is the number of half-samples formed, and $\hat{R}$ is the sample estimate of R, and

$$(2) \qquad \hat{V}_{B2}(\hat{R}) = \frac{1}{\ell} \sum_{p=1}^{\ell} (\hat{R}_p - \hat{R})^2$$

where $\hat{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} X_{hj} / \sum_{h=1}^{L}\sum_{j=1}^{2} Y_{hj}$.

In the sampling experiment the observations in each stratum are grouped into two primary units of equal size.

For the combined ratio estimate the jackknifed estimate of R are,

$$\hat{R}_{hj} = \frac{\sum_{u=1}^{L}\sum_{v=1}^{2} X_{uv} - (X_{hj} - X'_{hj})}{\sum_{u=1}^{L}\sum_{v=1}^{2} Y_{uv} - (Y_{hj} - Y'_{hj})} ,$$

where (X'_{hj}, Y'_{hj}) is the observation left in the h^{th} stratum following the delection of (X_{hj}, Y_{hj}).

The three variations of the jackknife variance estimate considered are,

(1) $$\hat{V}_{J1}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \bar{R})^2$$

where $\bar{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} \hat{R}_{hj}/2L$,

(2) $$\hat{V}_{J2}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \hat{R})^2$$

where $\hat{R} = \sum_{h=1}^{L}\sum_{j=1}^{2} X_{hj} / \sum_{h=1}^{L}\sum_{j=1}^{2} Y_{hj}$,

and

(3) $$\hat{V}_{J3}(\hat{R}) = \frac{1}{2}\sum_{h=1}^{L}\sum_{j=1}^{2} (\hat{R}_{hj} - \hat{R}_{i\cdot})^2$$

where $\hat{R}_{i\cdot} = \frac{1}{2}\sum_{j=1}^{2} \hat{R}_{hj}$.

Two situations are considered:

I. L=3 strata with n=2 observations per strata.

II. L=3 strata with n=10 observations per strata.

In naturally occuring health related data sets one may find cases in which the ratio of the variables under consideration differs greatly in each stratum. Conversely, it is possible to find data in which virtually no spread across strata ratios occurs. Into this experiment are designed two cases: "No Spread" and "High Spread." No Spread is the case where the probability distribution in each stratum is precisely the same yielding equal location parameters. High Spread is characterized by large differences between strata with respect to the stratum ratios.

Four families of distributions are considered in this experiment: the Uniform, the Chi-Square, the F and the Normal Distributions. Note that two are skewed and two are symmetric. Figure 1 presents the parameters chosen for each distribution by spread.

FIGURE 1

PARAMETERS OF THE DISTRIBUTIONS TO BE CONSIDERED

		SPREAD		
		NO	LOW	HIGH
DISTRIBUTION	(Parameters)	(a,b)	(a,b)	(a,b)
U(a,b)	Stratum 1	(100,150)	(90,140)	(60,110)
	Stratum 2	(100,150)	(100,150)	(100,150)
	Stratum 3	(100,150)	(110,160)	(140,190)
	(Parameter)	(n)	(n)	(n)
χ^2 n=df	Stratum 1	10	9	2
	Stratum 2	10	10	10
	Stratum 3	10	11	18
	(Parameters)	(ν_1,ν_2)	(ν_1,ν_2)	(ν_1,ν_2)
$F_{(\nu_1,\nu_2)}$	Stratum 1	(6,14)	(6,12)	(6,10)
	Stratum 2	(6,14)	(6,14)	(6,14)
	Stratum 3	(6,14)	(6,16)	(6,18)
	(Parameters)	(μ,σ^2)	(μ, σ^2)	(μ, σ^2)
$N(\mu,\sigma^2)$	Stratum 1	(50,5)	(45,5)	(30,5)
	Stratum 2	(50,5)	(50,5)	(50,5)
	Stratum 3	(50,5)	(55,5)	(70,5)

3. Evaluation of the Variance Estimators

To evaluate the performance of the Balanced Half-Sample and Jackknife estimators of $V(\hat{R})$ one would like a precise value for $V(\hat{R})$. For the purpose a "target value," $\hat{V}(\hat{R})$, is used. This value is the variance of the M=1000 values of $\hat{R}$ as computed in the sampling experiment. Also estimated from the sampling experiment are the expected values, variances, and absolute relative biases of the variance estimation techniques under consideration.

4. Results

Since the populations used in the experiment were artifically generated several checks were implemented to verify the performance of the computer processes. First a goodness of fit test provided information confirming that the basic U(0,1) numbers generated were random. Subsequent goodness of fit tests supported the claim that the transformation utilized provided populations having the specified F, Chi-Square and Normal Distributions.

As a check on the validity of the experiment the final results are presented only after several independent trails, each using a different set of random numbers were done. On each occasion the results were comparable.

As one check on the operation of the sampling experiment the expected value of the combined ratio estimate using all 2L observations over the M=1000 trails, $\hat{E}(\hat{R})$, was compared to the theoretical value. The two were in close agreement confirming the reliability of the simulation. First the case where n=2 will be examined.

In this research a criterion is established for considering the magnitude of the estimated absolute relative bias to be "acceptable" at 10%. Table 1 shows that for n=2 when Y, the variable in the denominator of the combined

ratio estimate, has the uniform distribution, both the Jackknife and the Balanced Half-Sample method yield estimates which have low bias. The absolute relative bias, ARB, was less than or equal to 9% regardless of the distribution of X, the numerator variable. However, in virtually every other instance a pattern was found to develop. The jackknife estimates were consistently less biased than the Balanced Half-Sample estimates and yielded values which were acceptable with ARB$\leq$9%. In each of the four situations with skewed, non-normally distributed variables in the denominator, the Balanced Half-Sample estimates were shown to be highly biased. For example, when the denominator distribution was $F_{(\nu_1,\nu_2)}$ the Jackknife produced estimates generally within acceptable bounds while the balanced half-sample proved to be highly biased with ARB ranging from 37% to 69% regardless of the distribution of the variable in the numerator.

Table 2 presents $\hat{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3, for selected representative distributional combinations for n=2. Clearly, the three jackknife estimates of the variance of the combined ratio estimate are less variable than either of the balanced half-sample estimates.

When there are n=2 observations per stratum and the distribution is non-normal the three jackknife estimates are shown to provide better estimates of $V(\hat{R})$, with respect to amount of bias and variability, than the two balanced half-sample estimates. This is not surprising since the jackknife techniques use more of the available information from a stratified sample in constructing estimates of the variance of the combined ratio estimate than does the balanced half-sample method. Each of the 2L jackknife estimates of the population ratio omits only one observation from a specified stratum adding twice the value of the observation left in that stratum to all the information contained in the remaining strata. This should be compared to a balanced half-sample estimate which uses one of the two observations in each stratum to estimate the combined ratio estimate. Also note that for the L=3 strata case considered here only ℓ=4 half-sample estimes of the combined ratio estimate are used for $\hat{V}_I(R)$, I=B1, B2, as opposed to the 2L=6 jackknife estimates that are used for $\hat{V}_I(\hat{R})$, I=J1, J2, J3. The next result is that, for n=2, the Jackknife technique produced a more stable estimate of the variance of the combined ratio estimate than is possible using the Balanced Half-Sample method, particularly when the data are from dispersed or highly skewed distributions.

Hislop (1977) demonstrated that the Spread factor has little effect upon the results of this investigation and, therefore, for brevity, only one case was selected for presentation.

When n=10 observations per stratum are used with two primary units in each stratum the results appear similar to those obtained in the linear case insofaras the ARB falls within acceptable bounds, (ARB$\leq$10%). This is seen in Table 3. A possible explanation for this may be attributed to the central limit theorem, since summary measures are calculated in each stratum yielding two primary units per stratum when the number of observations exceeds two. Each primary unit is the mean of half the observations in the stratum. Thus, regardless of the distribution of the original observations, as n increases, results much like those obtained when the underlying distribution is normal are expected. Table 3 shows, for the normal case, ARB$\approx$.8%.

Table 4 presents the target value as well as the expected value of the estimates over the M=1000 trials, $\hat{E}[\hat{V}_I(\hat{R})]$, I=B1, J3, for n=10 observations per stratum. Upon visual inspection it is clear that in many cases the balanced half-sample and jackknife methods are producing estimates of the target value which are strikingly similar to the findings for the normal case regardless of the distribution of the variables comprising the random pair. For several particular cases, notably when the numerator distribution is U(100,150) and the denominator distribution is $F_{(6,14)}$, the variability of the estimates was found to be high. The most variable families of distributions considered in this work are the uniform, U(a,b), and the $F_{(\nu_1,\nu_2)}$. This is shown in Table 5.

5. Conclusions

It is proposed that, as n increases, no matter what the distribution of the original observations, one may appeal to the central limit theorem and the estimates under consideration will yield values similar to those found when the distribution is normal.

For most situations considered, however, with n=10 the two techniques under consideration are shown to yield estimates whose variability is of the magnitude found in previous research for the case where the underlying distribution is normal. This is a key point for it is supportive of the use of the balanced half-sample techniques for estimating the variance of the combined ratio estimate regardless of the underlying distribution when the number of observations per stratum is equal to 10. This implies that surveys such as the HES are correct in using balanced replication since in most cases, the sample size is much larger than n=10. Notably, the Jackknife once again out performs the Balanced Half-Sample but the difference is not as pronounced.

In the complex multi-stage surveys presently in use, comparisons within domains of interest many times effectively reduce the sample size under consideration. In these cases, when the distribution of the variables of interest are non-normal or unknown, with n<10 observations per stratum, the jackknife estimate of the variance of the combined ratio estimate is to be preferred. As brought out in this research, the effect of small stratum sample size and non-normally distributed populations on the Balanced Half-Sample technique is quite serious producing estimates which are highly biased and

unstable. When n is large, however, both techniques considered here are shown to perform well regardless of the distribution of the variables under consideration.

Acknowledgements

The authors would like to acknowledge a grant from the University of Massachusetts/ Amherst Computer Center which made the programming and computations of this research possible.

References

Bean, Judy A. (1975). Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples: An Empirical Distribution. Data Evaluation and Methods Research. NCHS. Series 2, Number 65. DHEW Publication No. (HRA) 75-1339.

Frankel, Martin R. (1971). Inference from Survey Samples. Ann Arbor: Institute for Social Research, The University of Michigan.

Hislop, David A. (1977). Evaluation of the Balanced Half-Sample and Jackknife Variance Estimates for Linear and Combined Ratio Estimates for Non-Normal Populations. Biostatistics-Epidemiology Program Series No. 77-6. University of Massachusetts, Amherst.

Lemeshow, Stanley and Epp, Robert (1977). Properties of the Balanced Half-Sample and Jackknife Variance Estimation Techniques in the Linear Case. Communications in Statistics A, Vol. 6, Issue 13.

Lemeshow, Stanley A. and Levy, Paul S. (1977). Estimating the Variance of Ratio Estimates in Complex Sample Surveys with Two Primary Sampling Units per Stratum - A Comparison of Balanced Replication and Jackknife Techniques. Submitted for publication.

Marsaglia, G., Ananthanarayanan, K. and Paul, N. (1973). How to Use the McGill Random Number Package 'SUPER-DUPER,' four page typed description, School of Computer Sciences, McGill University, Montreal, Quebec.

McCarthy, Philip J. (1966). Replication. An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics. NCHS. Series 2, No. 14.

_____ (1969). Pseudoreplication: Half-Samples. Review of the International Statistical Institute, Vol. 37, No. 3: 239-264.

_____ (1969). Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique. Vital and Health Statistics. Series 2, No. 31.

Plackett, R.L. and Burman, J.P. (1946). The Design of Optimum Multifactoral Experiments. Biometrika, Vol. 33 (Pt. IV): 305.325.

Quenouille, M.H. (1956). Notes on Bias in Estimation. Biometrika, Vol. 43: 353-360.

TABLE 1. Absolute Relative Bias, ARB, as estimated from a sampling experiment with m=1000 trials. Values are given by distribution and spread for $\hat{V}_I(\hat{R})$, I=B1, B2, J1, J2, J3, for n=2.

Absolute Relative Bias

Numerator* Distribution and Spread		B1	B2	J1	J2	J3
		Denominator Distribution χ^2_n No spread				
χ^2_n	High	.175247E 00	.217068E 00	.355795E-01	.454862E-01	.217410E-01
N	No	.248767E 00	.319661E 00	.611118E-01	.803036E-01	.409610E-01
U	No	.399185E-01	.928543E-01	.101506E-01	.871287E-01	.115853E-01
F	No	.197346E 00	.226171E 00	.904778E-01	.981040E-01	.805292E-01
		Denominator Distribution $F_{(\nu_1,\nu_2)}$ No spread				
χ^2_n	High	.372152E 00	.525033E 00	.518855E-01	.875958E-01	.203471E-01
N	No	.494171E 00	.695482E 00	.496440E-01	.937074E-01	.708737E-02
U	No	.466603E 00	.658083E 00	.442817E-01	.873108E-01	.114631E-01
F	High	.464392E 00	.588241E 00	.118826E 00	.145909E 00	.805166E-01
		Denominator Distribution U(a,b) No spread				
χ^2_n	High	.670399E-02	.762607E-02	.129730E-02	.994899E-03	.180999E-02
N	No	.731554E-01	.765164E-01	.490608E-01	.501436E-01	.483511E-01
U	No	.903523E-01	.925498E-01	.801824E-01	.809013E-01	.795514E-01
F	No	.423911E-03	.106765E-02	.223983E-02	.202627E-02	.263540E-02

* χ^2_n: Chi square U: uniform(a,b)
N: normal(μ,σ) F: $F_{(\nu_1,\nu_2)}$

TABLE 2. Variance of the variance estimates of the combined ratio estimate from a sampling experiment, $\tilde{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3. Values are given by distribution of the random variables and spread for n=2.

$\tilde{V}[\hat{V}_I(\hat{R})]$

Numerator* Distribution and Spread		B1	B2	J1	J2	J3
		Denominator Distribution χ^2_n No spread				
N	No	.226662E 01	.274572E 01	.120096E 01	.128264E 01	.110110E 01
U	No	.864569E 02	.103471E 03	.403767E 02	.426828E 02	.380012E 02
F	No	.242117E-04	.255654E-04	.197749E-04	.200323E-04	.194153E-04
χ^2	High	.126978E-01	.142676E-01	.750100E-02	.775114E-02	.722465E-02
		Denominator Distribution $F_{(\nu_1,\nu_2)}$ No spread				
N	No	.290601E 06	.416092E 06	.731203E 05	.838236E 05	.580908E 05
U	No	.968188E 07	.138746E 08	.291988E 07	.354533E 07	.231239E 07
F	High	.148071E 01	.173652E 01	.768299E 00	.706854E 00	.726055E 00
χ^2	No	.115364E 04	.157357E 04	.297568E 03	.331110E 03	.250521E 03
		Denominator Distribution U(a,b) No spread				
N	No	.118490E-06	.119759E-06	.103981E-06	.104347E-06	.103793E-06
U	No	.132358E-04	.133280E-04	.126253E-04	.126540E-04	.126062E-04
F	No	.265109E-08	.265154E-08	.264356E-08	.264371E-08	.264328E-08
χ^2	High	.580932E-07	.582082E-07	.565221E-07	.565594E-07	.565493E-07

* N: Normal(μ,σ), U: Uniform(a,b), F: $F_{(\nu_1,\nu_2)}$, χ^2: χ^2_n

TABLE 3. Absolute Relative Bias, ARB, as estimated from a sampling experiment with m=1000 trials. Values are given by distribution for $\hat{V}_I(\hat{R})$, I=B1, B2, J1, J2, J3, n=10 and No spread.

Absolute Relative Bias

Distribution*	B1	B2	J1	J2	J3
U/U	.702727E-01	.707533E-01	.686316E-01	.687913E-01	.684745E-01
χ^2/U	.507675E-01	.509781E-01	.482317E-01	.483012E-01	.481224E-01
F/F	.254325E-01	.431291E-01	.274835E-01	.220894E-01	.334131E-01
χ^2/F	.481465E-01	.229696E-01	.116021E 00	.108459E 00	.124244E 00
U/F	.593404E-01	.929300E-01	.203678E-01	.101377E-01	.304936E-01
χ^2/χ^2	.725569E-02	.115231E-02	.196865E-01	.176952E-01	.217388E-01
U/χ^2	.573058E-01	.683600E-01	.300167E-01	.335846E-01	.269165E-01
F/χ^2	.546991E-02	.506174E-01	.761548E-01	.748622E-01	.777854E-01
N/N	.796754E-02	.790266E-02	.841999E-02	.839838E-02	.844212E-02

* U/U: Uniformly distributed variable in both numerator and denominator, U(100,150).

χ^2: Chi Square n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).

TABLE 4. Expected value of the estimate of the variance of the combined ratio, $\tilde{E}[\hat{V}_I(\hat{R})]$, I=B1, J3 and the Target value for those estimates, $\tilde{V}(\hat{R})$. Results of a sampling experiment are given by distribution for No spread and n=10.

Distributions*	$\tilde{V}(\hat{R})$	$\tilde{E}[\hat{V}_I(\hat{R})]$ B1	J3
U/U	.887925E-03	.950520E-03	.948723E-03
χ^2/U	.464299E-04	.487870E-04	.486642E-04
F/F	.427233E-01	.438099E-01	.412958E-01
χ^2/F	.213145E 01	.202883E 01	.186663E 01
U/F	.258163E 03	.243701E 03	.236270E 03
χ^2/χ^2	.140054E-01	.139037E-01	.137009E-01
U/χ^2	.118276E 01	.125054E 01	.121460E 01
F/χ^2	.389343E-03	.368046E-03	.354058E-03
N/N	.133671E-03	.132606E-03	.132543E-03

* U/U: Uniformly distributed variable in numerator and denominator, U(100,150).

χ^2: Chi Square n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).

TABLE 5 Variance of the variance estimates of the combined ratio estimates from a sampling experiment, $\tilde{V}[\hat{V}_I(\hat{R})]$, I=B1, B2, J1, J2, J3. Values are given by distribution for No spread and for n=10.

$\tilde{V}[\hat{V}_I(\hat{R})]$

Distributions*	B1	B2	J1	J2	J3
U/U	.613459E-06	.614459E-06	.614383E-06	.614729E-06	.614092E-06
χ^2/U	.147141E-08	.147224E-08	.145501E-08	.145528E-08	.145268E-08
F/F	.195193E-02	.205731E-02	.161293E-02	.163940E-02	.158894E-02
χ^2/F	.401505E 01	.441704E 01	.303707E 01	.313801E 01	.291710E 01
U/F	.667917E 05	.747136E 05	.478116E 05	.496628E 05	.457670E 05
χ^2/χ^2	.149780E-03	.153090E-03	.139113E-03	.140117E-03	.138225E-03
U/χ^2	.106813E 01	.110353E 01	.923549E 00	.940041E 00	.918151E 00
F/χ^2	.171114E-06	.175258E-06	.153689E-06	.154274E-06	.153101E-06
N/N	.123622E-07	.123651E-07	.123188E-07	.123197E-07	.123179E-07

* U/U: Uniformly distributed variable in numerator and denominator, U(100,150).

χ^2: Chi Square, n=10 df, F: $F_{(6,14)}$, N: Normal(50,5).

ESTIMATION WHEN THE SAMPLING RATIO IS A LINEAR FUNCTION OF THE DEPENDENT VARIABLE

John Bishop, Institute for Research on Poverty

It is not uncommon for economists and sociologists to use data bases where the probability that a random individual will be in the sample depends upon his income, occupation, or education. Often these data bases are used to estimate models predicting these same success indicators. The application of ordinary least squares (OLS) to such data, however, yields inconsistent estimators of models predicting income, occupation, or education.

The biased nature of OLS estimators when the sample selection is based on the dependent variable, often called truncation bias in the literature, has been pointed out frequently (Bishop 1974; Cain 1975; Crawford 1975; Hausman and Wise 1977; Manski and Lerman 1976; Taubman and Wales 1974, ch. 4, app. F, L). Sometimes the sampling process results in an absolute truncation (i.e., absolutely no one with initial year incomes above 1.5 times the poverty line, as in the Rural Income Maintenance Experiment). Estimation techniques for this situation have been developed by Crawford (1975) and Hausman and Wise (1977).

This paper tackles the situation where all observations in the population have some probability of being in the sample and the probability is a linear function of the dependent variable. I calculate and apply formulae that relate the bias to the strength of success selectivity and the R^2 of the true relationship.

Data bases where sampling ratio depends upon income are of two types: follow-up surveys with substantial nonresponse rates, and interview surveys that oversample people in low or high income neighborhoods. Follow-up surveys may fail to obtain information from many of the people in its defined sample for a variety of reasons: death, inability to find a current address, or refusal by the respondent to fill out the questionnaire. Refusals are the primary cause of success bias.

One heavily used data set with substantial refusal problem is Project Talent. The combined 1 and 5 year follow-ups of the male 11th graders had a response rate of 52% to the series of mail questionnaires. A special intensive follow-up of a 5% sample of mail questionnaire nonrespondents which obtained a 90% response rate allows us to establish the extent to which success affects the probability of responding to a mail questionnaire. Stratifying by the social status of each student's parents, college attenders were 1.5 to 1.6 times as likely to respond to at least one of the two follow-ups (Bishop 1974). Given college attendance status, the student's family background had no systematic impact on his response rate.

Another very important data set that potentially has a success bias is the National Bureau of Economic Research (NBER)-Thorndike sample. Thorndike took a random sample of 17,000 from a population of Army Air Corps volunteers for pilot, navigator, and bombardier training programs who passed a preliminary screening test. By 1955, 1500 had died and of the living, 2000 military and 9700 civilians responded to a mail questionnaire. The response rate was therefore about 75%. This is a high response rate and is attributed by Taubman and Wales (1974) to the accurate current addresses generally available from the Veterans Administration and the use of Retail Credit Bureau to find some of the nonrespondents.

The 1969 data is a survey of the 1955 respondents. Of those for whom current addresses were obtained and who had not died, 70% responded. Taubman and Wales found that while the 1955 income of 1969 nonrespondents was lower than for the respondents, it was not lower when ability and schooling were controlled. From this they argued that any selection process that existed was based on the independent and not the dependent variables. It has been shown that when the true model has homogeneous coefficients, differential sampling ratios that depend on included right hand side variables do not bias the estimates of structural parameters (Porter 1973; Taubman and Wales 1974).

However, their test applies only to the response rate conditional upon having responded in 1955. There may still be success bias in the 1955 response rate. Their test also depends upon the assumption that success persists over time and that income is as good a measure of success at age 29 as at age 44. If the conditional probability of responding in 1969, given that one responded in 1955, is a function of the change in one's relative income over the period, the test used by Taubman and Wales will miss the success bias. An alternative way to test for success bias in the 1969 data would be to compare those who responded as soon as they received a questionnaire to those who required reminders. But even this requires some strong assumptions. Because of the lack of an intensive follow-up by retail credit or phone, we can never be sure there is no success bias in the NBER-Thorndike data. However, it may be possible to put limits on the effects a success bias could have.

Another type of data set in which this problem arises is when black neighborhoods have been oversampled, as in the 1966-67 Survey of Economic Opportunity (SEO); when low income neighborhoods have been oversampled, as in the Census Employment Surveys; or when low family incomes relative to the poverty line are oversampled, as in the Michigan Panel Study of Income Dynamics. These data sets have been used to estimate models predicting success variables like hours worked, weeks worked, and earnings. A widely publicized finding using these surveys has been that rates of return to schooling are lower in low income neighborhoods than for samples of people drawn from the metropolitan area as a whole or the nation (Harrison 1972). Since living in a poverty neighborhood is a consequence of earnings, restricting one's sample to these neighborhoods or oversampling in them results in a simultaneous equations bias when estimating the structural parameters of models that predict earnings and other success variables.

In the next section of this paper, I calculate the bias to be expected in OLS estimates of structural models of earnings, work effort, or

status attainment when the probability of being in the sample is a linear function of the dependent variable. If we adopt the conventional assumption that the true relationship has a homoskedastic error structure, we find that the ratio of the true to the estimated coefficient is a simple positive function of the R^2 of the true relationship and a negative function of the absolute size of the proportionate change in sampling probability for a standard deviation change in the dependent variable. When the right hand side variables are symmetric (the third moment = 0), the bias is independent of whether the sampling proportion is a positive or negative function of the dependent variable. To demonstrate the importance and relevance of these findings, the final section of this paper compares the schooling coefficients estimated in different subsamples of the SEO in models predicting yearly earnings.

1. Statistical Model

Porter (1973) and others have shown that if sampling ratios are independent of the disturbances of the model to be estimated and the coefficients of that model are homogeneous over the population, OLS estimators of structural parameters are unbiased. In other words, sampling ratios that are functions of included independent variables (correlated with y only because of the joint dependence of x and y) do not produce a selection bias in OLS estimators. The problem dealt with in this paper is sampling ratios that are linear functions of the dependent variable. Sampling proportions correlate with independent variables solely as a result of their joint association with y.

Analytical solutions are not difficult to obtain for models with only one independent variable. Let the true model be

$$1)\quad y_i = \beta x_i + u_i$$

$$2)\quad p_i = (1 + \gamma y_i + v_i)\, n_s/n.$$

Then

$$3)\quad E_s(y) = \frac{\Sigma_1^n p_i y_i}{\Sigma_1^n p_i} = \frac{\Sigma_1^n (1 + \gamma y_i + v_i) y_i}{\Sigma_1^n (1 + \gamma y_i + v_i)}$$

$$= \gamma V(y)$$

where i indexes each observation in the population (i = 1 . . . n)

- y_i and x_i are defined as deviations from their population mean
- u_i is homoskedastic and independent of x_i and v_i
- p_i = probability the "i"th observation will be selected
- n_s/n = the average sampling ratio = the number of observations selected for the the sample (n_s) divided by the total number in the population (n)
- v_i is independent of x_i and consequently independent of y_i
- γ = the increased probability of being sampled per unit of y divided by the average sampling proportion
- E is the expectation operator
- s subscript indicates the mean, variance, or covariance indicated is for the nonrandom sample.

We note that all summations are over the entire population, i = 1 . . . n, and drop the limits from our notation. The sample mean of x is

$$4)\quad E_s(x) = \frac{\Sigma\, p_i x_i}{\Sigma\, p_i} = \frac{\Sigma(1 + \gamma y_i + v_i)x_i}{\Sigma(1 + \gamma y_i + v_i)}$$

$$= \gamma \mathrm{Cov}(xy) = \gamma\beta V(x).$$

Noting that $\Sigma x = \Sigma y = 0$, the sample variances and covariances have the following expectations:[1]

$$E(V_s(x)) = \frac{\Sigma[1 + \gamma y_i + v_i][x_i - \gamma\beta V(x)]^2}{\Sigma(1 + \gamma y_i + v_i)}$$

$$= \frac{\Sigma[x_i - \gamma\beta V(x)]^2}{n} + \frac{\gamma}{n}\Sigma y[x - \gamma\beta V(x)]^2$$

$$= V(x) + \gamma^2\beta^2 V(x)^2 + \frac{\gamma\Sigma y_i x_i^2}{n} - 2\gamma^2\beta V(x)\mathrm{Cov}(xy)$$

$$5)\quad E(V_s(x)) = V(x)\left[1 + \frac{\gamma\beta\Sigma x^3}{nV(x)} - \gamma^2\beta^2 V(x)\right]$$

$$E(\mathrm{Cov}_s(xy)) = \frac{1}{n}\Sigma[x_i - \gamma\beta V(x)][y_i - \gamma V(y)] + \gamma\frac{1}{n}\Sigma y[x_i - \gamma\beta V(x)][y_i - \gamma V(y)]$$

$$= \mathrm{Cov}(xy) + \gamma^2\beta V(x)V(y) + \frac{\gamma}{n}\Sigma x_i y_i^2 - \gamma^2 V(y)\mathrm{Cov}(xy) - \gamma^2\beta V(y)V(x)$$

$$6)\quad E(\mathrm{Cov}_s(xy)) = \mathrm{Cov}(xy)\left[1 + \frac{\gamma\beta^2\Sigma x^3}{n\ \mathrm{Cov}(xy)} - \gamma^2 V_{(y)}\right].$$

The probability limit of the sample estimate of β is

$$7)\quad b_s = \frac{\mathrm{Cov}(xy)[1 - \gamma^2 V(y) + \gamma\beta^2\Sigma x^3/n\mathrm{Cov}(xy)]}{V(x)[1 - \gamma^2\beta V(x) + \gamma\beta\Sigma x^3/nV(x)]}$$

$$8)\quad \frac{b_s}{\beta} = \frac{1 + D - \gamma^2 V(y)}{1 + D - \gamma^2 V(y)R^2},$$

where $D = \gamma\beta\Sigma x^3/nV(x) = \gamma\beta$ times the ratio of the third and second moments of x

R^2 = the propostion of the variance explained by the true relationship.

Since $R^2 \leq 1$, b_s/β is necessarily less than or equal to 1. Selection on the dependent variable attenuates the parameter estimates. The amount of attenuation depends upon three factors: the direction and degree of skewness of x(D), the strength of the relationship between y and the probability of selection (γ), and the R^2 of the underlying relationship.

The D term in (8) depends upon the interaction of the sample selection process with the skewness of x. Since skewness is defined as $a_3 = \Sigma x^3/n\sigma_x^3$ = the third moment of a variable over the cube of its standard deviation, we may rewrite

$D = a_3 \cdot \gamma\beta\sigma_x = a_3 \cdot \gamma\sigma_y r_{xy}$. The expression, $\gamma\sigma_y r_{xy}$ times 100, can be interpreted as the percentage change in the probability of an observation's selection into the sample that is associated with a standard deviation change in x. It is positive when γ and r_{xy} have the same sign, as in earnings functions estimated on Project Talent or NBER-Thorndike data sets. Thus, if the distribution of x in the population has positive skew, D is positive, which reduces bias. In SEO and Census Employment Survey data sets where families in black or low income neighborhoods are oversampled, $\gamma\sigma_y r_{xy}$ is negative because here γ and r_{xy} have opposite signs. In these surveys a positive skew to x causes D to be negative, thus increasing the bias.

The distribution of years of schooling--the x variable upon which we are focussing in this paper--can be skewed in either direction, depending on the year and population studied. People educated in the early twentieth century have positively skewed educational attainment distributions. The most recent cohorts have negatively skewed distributions. Men between the ages of 30 and 35 in 1974 have an $a_3 = -.53$. Distributions for adults of all ages are very close to being symmetric. When compared to the skewness of a zero-one variable with a mean of .1, whose $a_3 = 2.67$, skewness for all adults is quite small: .04 for white males and -.14 for black males in the 1967 CPS. Since the term measuring the impact of a standard deviation change in x on the probability of selection, $\gamma\sigma_y r_{xy}$, must have an absolute value of substantially less than one, schoolings skewness does not have an important effect upon the magnitude of the selection bias in first order statistics of relationships between schooling and income. From this point on we will, therefore, neglect the impact of skewness and assume that all independent variables are symmetric ($a_3 = 0$). When all variables are assumed symmetric, it is possible to derive a simple formula for the selection bias in the coefficients of regressions with two independent variables. (The mathematical derivation is carried out in the Appendix.) The formula that results is the same as the formula for order regression coefficients when x is symmetric:

$$9) \quad \frac{b_s}{\beta} = \frac{1 - \gamma^2 V(y)}{1 - \gamma^2 V(y) R^2},$$

where R^2 is the coefficient of determination in the multi or bivariate regression in the full population.[2] The sign of γ indicates whether the sampling ratio is positively or negatively associated with the dependent variable. It is squared in the final terms of both the numerator and denominator. Consequently, the size of the bias is not affected by whether more income raises or lowers the probability of selection. The probability limit of the ratio of estimated to true parameters when the independent variables are symmetric is presented in Table 1 for alternative γ's and R^2's.

If the $R^2 = 1.0$, there is no bias, for selecting the sample on the dependent variable is equivalent to selecting on the independent variables. As the R^2 declines, the bias increases in size for a $|\gamma\sigma_y|$ of .4, an R^2 of .6 implies a bias ratio of .929. An R^2 of .3 implies a bias ratio of .882 or a 12% attenuation of regression coefficients. An R^2 of .1 implies a bias ratio of .853 or a 15% attenuation of the coefficients. In the limit as R^2 approaches zero, the bias ratio approaches its maximum of $b_s/\beta = 1 - \gamma^2 V(y)$. Thus, when the bias in first order coefficients is compared across alternative right hand side variables, the proportionate attenuation is larger in variables that have a weak relationship with y. Since in a trivariate relationship bias depends upon the multiple correlation coefficient, the coefficients of both independent variables attenuate by an identical proportionate amount.

The expression, $\gamma\sigma_y$, is the change in the probability of inclusion in the sample associated with a standard deviation change in y divided by the average probability of inclusion. The smaller $|\gamma\sigma_y|$ the smaller the bias. Since γ must approach zero as the proportion of a population that is sampled approaches one, selection bias must decline as a survey's response rate approaches 100%. For a given R^2, the attenuation of regression coefficients rises roughly in proportion to the square of $\gamma\sigma_y$. At an R^2 of .30, a $\gamma\sigma_y$ of .2 causes a 3% attenuation, a $\gamma\sigma_y$ of .4 causes an attenuation of 12%, an a $\gamma\sigma_y$ of .707 yields an attenuation of 41%.

Biases of even larger magnitudes are possible if selection probabilities have a nonlinear relation ($\ln \frac{P}{1-P} = \gamma y$, for instance) with the dependent variable. As long as the sampling ratio is defined as a linear function of y, it is not possible for our model to handle truly powerful selection biases. The derivations would be internally inconsistent if predicted sampling ratios fell outside the zero-one interval. They will not fall outside this interval if γ is sufficiently small and the y distribution sufficiently compact. A rectangular distribution for y would require a $|\gamma\sigma_y| < .81$, if $n_s/n < .5$, and a $|\gamma\sigma_y| < 2(.81)(1-n_s)/n$ for $n_s/n > .5$. All other single modal distributions of y will require that γ be smaller than these limits.

2. Application to Earnings Functions in the Survey of Economic Opportunity

Our statistical model predicts that when the sampling ratio is dependent of income, the schooling coefficients in an earnings function will be lower than the true population coefficient. Table 2 tabulates estimated relative sampling ratios by earnings for alternative subsamples of the SEO. Not surprisingly, the probability of living in a low income neighborhood is negatively associated with the level of one's earnings. For whites, the probability of living in a predominantly black area is also negatively associated with earnings. For blacks, however, there was no visible relationship. Therefore, we do not expect blacks in the special sample of predominantly black neighborhoods to have lower schooling coefficients than a national sample of blacks. We do expect, however, that whites living in these neighborhoods will have a smaller schooling coefficient than a national sample of whites. Also, rates of return to schooling estimated for both blacks and whites

living in low income neighborhoods in urban areas are expected to be smaller than the rates of of return for all urban residents. An examination of Table 3 indicates, as expected, that schooling coefficients of whites in predominantly black and low income areas are substantially smaller than those in the national sample. For whites the unbiased coefficient of .0879 falls to .0701 in black areas and to .0643 when the sample is limited to low income neighborhoods. The schooling coefficients for blacks are smaller only for the low income areas. Furthermore, the drop in the schooling coefficients is larger for models with low R^2 (those without measures of work effort on the right hand side).

For blacks in low income areas the linear specification of the sampling mechanism predicts the coefficient changes well. For whites, the impact of income on the sampling ratio is so powerful that the estimates of γ produced are too high. Some high earnings individuals will have negative predicted sampling ratios, in which case the analysis becomes internally inconsistent. If predicted coefficients are calculated, nevertheless, we overpredict the reduction in the schooling coefficients.

The problem is that for whites the sampling ratio-earnings relationship for predominantly black or low income neighborhoods is nonlinear. It looks like a logistic specification would serve better than a linear specification. Simple analytic results are not obtainable, however, when the sampling ratio is a nonlinear function of the dependent variable.

Where does that leave the researcher? If data availability forces one to use a data set in which sampling ratios are nonlinear functions of the dependent variable, how can consistent estimators be obtained? The solution that suggests itself is a two stage process. First, estimate a model of the sampling process. If sampling ratios depend directly on some of the independent variables as well as the dependent variable, these variables should be included in the model along with y. The main requirement of this model is that the error in predicting the sampling ratio be independent of the disturbances of the structural model. In Census Employment Surveys this could be done by comparing the low income area's population to that of the SMSA as a whole. In follow-up surveys a data set with an intensive follow-up of a sample of nonrespondents is required.

The second step is to estimate the structural model, using the inverse of these predicted sampling ratios as weights. Manski and Lerman (1976) have shown that when probabilities of inclusion in the sample are a function of a categorical dependent variable, weighting each observation by the inverse of its sampling ratio yields unbiased and efficient estimators of the coefficients of a logistic model. Where sampling ratios are known (as for follow-up surveys with intensive follow-ups of a small sample) weighted least squares using these ratios from the sampling frame is another alternative. It is safe from misspecification of the sampling model but it becomes highly sensitive to the observations in the nonrespondent sample, since just a few observations carry a major share of the variance to be explained. Both approaches reduce bias only at the cost of increasing heteroskedasticity. The advantage of using predicted sampling ratios rather than sampling frame ratios is that the heteroskedasticity created by weighting will be less serious. Heteroskedasticity, however, does not bias coefficients, it only lowers the precision with which they are estimated.

This paper presents a suggested route for exploration. I leave the rigorous development of the properties of such estimators to a later time, and to others.

NOTES

[1]Homoskedasticity and the independence of x and u makes it possible to simplify $\Sigma y^2 x$ and $\Sigma y x^2$:

$$\Sigma y^2 x = \Sigma x(\beta x + u)^2 = \Sigma\, \beta^2 x^3 + 2\beta x^2 u + \Sigma x u^2$$

$$= \beta^2 \Sigma x^3$$

$$\Sigma y x^2 = \Sigma x^2(\beta x + u) = \beta \Sigma x^3 + \Sigma x^2 u = \beta \Sigma x^3.$$

[2]In recent, as yet unpublished work, Arthur Goldberger (1975) has proved a result that is in many ways more general. When the right hand side variables are multi-normally distributed, truncation or selection bias results in a proportionate shrinkage of all regression slopes by $\theta^2/1 - (1 - \theta^2)R^2$, where θ^2 is the ratio of the restricted sample variance of y to the population variance of y. Note that $(1 - \theta^2)$ corresponds to $\gamma^2 V(y)$ in our notation. Thus, for the special case of bivariate and trivariate regressions when there is a linear relation between y and the probability of selection, this paper generalizes Goldberger's result to symmetric right hand side variables.

REFERENCES

Bishop, John. 1974. The private demand for places in higher education. Ph.D. dissertation, University of Michigan. Available from University Microfilms.

Cain, Glen. 1975. The challenge of dual and radical theories of labor market to orthodox theory. Discussion Paper 255-75. Institute for Research on Poverty, University of Wisconsin-Madison.

Crawford, David. 1975. Estimating earnings functions from truncated samples. Discussion Paper 287-75. Institute for Research on Poverty, University of Wisconsin-Madison.

Goldberger, Arthur. 1975. Linear regression in truncated samples. Unpublished manuscript.

Harrison, Bennett. 1972. Education and underemployment in the urban ghetto. American Economic Review 62:796-812.

Hausman, Jerry, and Wise, David. 1977. Social experimentation truncated distributions, and efficient estimation. Econometrica 45:919-938.

Johnston, J. 1963. Econometric methods. New York: McGraw-Hill.

Kmenta, Jon. 1971. Elements of econometrics. New York: MacMillan.

Manski, Charles, and Lerman, Steven. 1976. The estimation of choice probabilities from choice based samples. Unpublished paper. School of Urban and Public Affairs, Carnegia-Mellon University.

Masters, Stanley, and Ribich, Thomas. 1972. Schooling and earnings of low achievers: comment. American Economic Review 82:755.

Porter, Richard D. 1973. On the use of survey sample weights in the linear model. Annals of Economic and Social Measurement 2:141-158.

Taubman, Paul and Wales, Terence. 1974. Higher education and earnings. New York: McGraw-Hill.

The Appendix is available by writing the Institute for Research on Poverty, University of Wisconsin, Madison, Wisconsin 53706, and requesting Discussion Paper 430-77.

Table 1

Values of b_s/β as a Function of R^2 of the True Relationship and the Strength of Selection on y

	R^2 = 1.0	.8	.6	.5	.4	.3	.2	.10	0
$\lvert\gamma\sigma_y\rvert$ = .707	1	.833	.714	.667	.625	.588	.555	.526	.5
$\lvert\gamma\sigma_y\rvert$ = .5	1	.938	.882	.853	.833	.811	.789	.769	.75
$\lvert\gamma\sigma_y\rvert$ = .4	1	.965	.929	.913	.897	.882	.868	.853	.84
$\lvert\gamma\sigma_y\rvert$ = .2	1	.992	.984	.979	.975	.971	.967	.964	.96

Note: All independent variables are symmetric.

$\gamma\sigma_y$ is the proportionate increase in the sampling probability per standard deviation of the dependent variable.

Table 2

Estimated Sampling Ratio Conditional Upon Income Relative to the Average Sampling Ratio

Earnings	0-2	2-3	3-4	4-5	5-6	6-7	7-8	8-10	10-14	14-20
Whites in predominantly black areas	2.24	2.24	2.47	1.53	1.41	.83	.90	.64	.48	*
Blacks in predominantly black areas	1.00	.91	1.02	.97	.83	1.09	.91	.88	*	*
Whites in low income areas	2.31	2.45	3.00	2.17	1.37	.87	.70	.56	.21	*
Blacks in low income areas[1]	1.07	1.08	1.27	1.12	1.12	.88	.63	.52	*	*

*means n of the Current Population Survey base is below 10.

[1]Since low income areas were defined only for Standard Metropolitan Statistical Areas the comparison base is all blacks living in SMSA's.

Table 3

Schooling Coefficients in Different Samples

	Low Income Area Coef.	Predom. Black Area Coef.	CPS Coef.	CPS R^2	CPS $\gamma\sigma$* Predom. Black	CPS $\gamma\sigma$* Low Income
Yearly Earnings						
Whites						
0-20 yrs schooling	.0643	.0701	.0879	.23	-.95	-1.06
0-15 yrs schooling	.0500	.0656	.0889	.17	-1.00	-1.11
Blacks						
0-20 yrs schooling	.0525	.0628	.0610	.08	.45	-.20
0-15 yrs schooling	.0447	.0530	.0621	.07	.45	-.20
Hourly Earnings, 0-20 yrs Schooling						
Whites	.0588	.0556	.0743	.40	-.95	-1.06
Blacks	.0410	.0504	.0462	.54	.45	-.20

Note: The dependent variable is the log of yearly earnings. Samples were limited to nonfarm males not in school with at least six years of experience. The schooling coefficients are from regressions with experience, experience squared, SMSA residence, and SMSA size as controls. The hourly earnings coefficients have additional controls: log of weeks worked last year, part time last year, and last week. The Black CPS sample was limited to SMSA residents.

*Estimates of γ were obtained from unweighted regressions of the ratio of the observed conditional sampling ratio to the average sampling ratio on the log of yearly earnings. The CPS provides the estimate of the population distribution of earnings. Weighted regressions yield more negative estimates of γ.

DOUBLE SAMPLING IN MULTI-AUXILIARY REGRESSION ESTIMATION BASED ON CONDITIONAL SPECIFICATION

Grace O. Esimai and Chien-Pai Han
Iowa State University

1. Introduction

Consider a $(p + 1)$ random vector $\begin{pmatrix} Y \\ \underline{X} \end{pmatrix}$ which follows a multivariate normal distribution where Y is a scalar and $\underline{X}$ is a $p \times 1$ vector $(p \geq 1)$. In estimating the population mean μ_y of Y, it is well known that the precision of the estimator can be increased if auxiliary information is available. In this paper, we shall consider the linear regression estimator of μ_y with $\underline{X}$ as the auxiliary variable. To use the regression estimator we need to know the population mean $\underline{\mu}_X$ of $\underline{X}$. When $\underline{\mu}_X$ is unknown, we may take a preliminary sample to estimate it. This sampling procedure is the double sampling technique. In certain situations, an investigator may have partial information about $\underline{\mu}_X$ and suspects that $\underline{\mu}_X = \underline{\mu}_0$. In order to utilize this partial information, the investigator can perform a preliminary test about the hypothesis H_0: $\underline{\mu}_X = \underline{\mu}_0$ versus H_1: $\underline{\mu}_X \neq \underline{\mu}_0$. As an example, consider estimating the average yield per acre of a certain crop. It is known that the yield is highly correlated with the moisture and nitrogen content of the soil. Hence, the moisture and nitrogen content can be used as the auxiliary variable, $\underline{X}$. The experimenter usually does not know $\underline{\mu}_X$; but from the amount of rainfall reported by the weather bureau or other sources and from analysis by the soil science department, he believes that $\underline{\mu}_X$ should be $\underline{\mu}_0$. Once a preliminary sample is available, the investigator may test H_0. He then will use $\underline{\mu}_0$ in the regression estimator if H_0 is accepted; otherwise he uses the sample mean based on the preliminary sample. This estimator is usually known as the preliminary test estimator. If the investigator's prior information or experience is reliable, then the true mean $\underline{\mu}_X$ of $\underline{X}$ will be expected to be very close to $\underline{\mu}_0$. In this situation, the efficiency of the preliminary test estimator is high. Thus in practice, it is desirable to use the preliminary test estimator when partial information is available to the investigator.

Preliminary test estimator was first studied by Bancroft (1944). It belongs to the area of inference based on conditional specification. A bibliography on inference based on conditional specification was recently compiled by Bancroft and Han (1977).

Let $\begin{pmatrix} Y \\ \underline{X} \end{pmatrix} \sim N(\underline{\mu}, \underline{\Sigma})$; $\underline{\mu} = \begin{pmatrix} \mu \\ \underline{\mu}_X \end{pmatrix}$ and $\underline{\Sigma} = \begin{pmatrix} \sigma^2 & \underline{\Sigma}_{12} \\ \underline{\Sigma}_{21} & \underline{\Sigma}_{22} \end{pmatrix}$. We assume $\underline{X}$ is cheaply observed while Y is more expensive to observe. Let $(Y_i, X_{1i}, X_{2i}, \ldots, X_{pi})'$ $i = 1, \ldots, n_2$ be a random sample from $N(\underline{\mu}, \underline{\Sigma})$. This is supplemented by $n_1 - n_2$ $(n_1 > n_2)$ more independent observations on $\underline{X} = (X_1, \ldots, X_p)'$. In practice, the sample of n_2 observations is usually a subsample from the sample of n_1 observations. From all the observations, we define

$\bar{\underline{X}}_1 = (1/n_1)\left(\sum_{i=1}^{n_1} X_{1i}, \ldots, \sum_{i=1}^{n_1} X_{pi}\right)'$, and from the subsample, we define $\bar{y} = 1/n_2 \sum_{i=1}^{n_2} y$,

and $\bar{\underline{X}}_2 = (1/n_2)\left(\sum_{i=1}^{n_2} X_{1i}, \ldots, \sum_{i=1}^{n_2} X_{pi}\right)'$. If the vector $\underline{\mu}_X$ and $\underline{\Sigma}$ are known, then given $\underline{X}$ an unbiased estimator of μ_y is $\hat{\mu}_{y|\underline{X}}$

$= \bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,(\underline{\mu}_X - \bar{\underline{X}}_2)$ with variance

$(1/n_2)\{\sigma^2 - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}\}$. If $(1/n_2)\,\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}$ is considerably large, we have an appreciable gain in precision.

If $\underline{\mu}_X$ is unknown and partial information about $\underline{\mu}_X$ is available, without loss of generality we let $\underline{\mu}_0 = \underline{0}$, the linear regression preliminary test estimator is defined as

$$\hat{\mu}_{lr} = \begin{cases} \bar{y} - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_2 \\ \quad \text{if } n_1\,(\bar{\underline{X}}_1'\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_1) \leq \chi^2_{p,\alpha} \\ \bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) \\ \quad \text{if } n_1(\bar{\underline{X}}'_1\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_1) > \chi^2_{p,\alpha} \end{cases} \qquad (1.1)$$

where $\chi^2_{p,\alpha}$ is the $100(1-\alpha)$ percent point of the Chi-squared distribution with p degrees of freedom and α is the level of significance of the preliminary test. Han (1973) studied the estimator $\hat{\mu}_{lr}$ when $p = 1$. This paper will consider the general case when $p \geq 1$. The bias,

mean squared error (MSE) and relative efficiency of $\hat{\mu}_{\ell r}$ are derived in Esimai (1977) and are given in Section 2. The optimal sample design is discussed in Section 3.

When Σ is unknown, the linear regression preliminary test estimator is

$$\tilde{\mu}_{\ell r} = \begin{cases} \bar{y} - \underline{S}_{12}\,\underline{S}_{22}^{-1}\,\bar{\underline{X}}_2 \\ \qquad \text{if } m_1 n_1 (\bar{\underline{X}}_1' \underline{S}_{22}^{-1} \bar{\underline{X}}_1) \le T_0^2 \\ \bar{y} + \underline{S}_{12}\,\underline{S}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2) \\ \qquad \text{if } m_1 n_1 (\bar{\underline{X}}_1' \underline{S}_{22}^{-1} \bar{\underline{X}}_1) > T_0^2 \end{cases} \tag{1.2}$$

where $m_1 = n_1 - 1$, T_0^2 is the $100(1-\alpha)$th percentile of the Hotelling's T^2 distribution with m_1 degrees of freedom. We define

$$\underline{S} = \begin{pmatrix} S_{11} & \underline{S}_{12} \\ \underline{S}_{21} & \underline{S}_{22} \end{pmatrix} \text{ where } S_{11} = \sum_{i=1}^{n_2} (y_i - \bar{y}),$$

$$\underline{S}_{12} = \sum_{i=1}^{n_2} (y_i - \bar{y})(\underline{X}_i - \bar{\underline{X}}_2)',$$

$$\underline{S}_{22} = \sum_{i=1}^{n_1} (\underline{X}_i - \bar{\underline{X}}_1)(\underline{X}_i - \bar{\underline{X}}_1)' \text{ and } \bar{y},\ \bar{\underline{X}}_1 \text{ and } \bar{\underline{X}}_2 \text{ are as defined above.}$$

2. Bias, MSE and Relative Efficiency of $\hat{\mu}_{\ell r}$

The joint distribution of $(\bar{\underline{X}}_1', \bar{\underline{X}}_2', \bar{y})$ is normal. Denote the acceptance region for the preliminary test by A and its complement by $\bar{A}$ and let $\chi^2_{p,\alpha} = b$.

$$\begin{aligned} E(\hat{\mu}_{\ell r}) &= E\{(\bar{y} - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\bar{\underline{X}}_2) | A\}\, P(A) \\ &\quad + E\{[\bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)] | \bar{A}\}\, P(\bar{A}) \\ &= \mu_y - \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\, E\{\bar{\underline{X}}_1 | \bar{A}\}\, P(\bar{A}) \\ &= \mu_y + B_1 \end{aligned} \tag{2.1}$$

B_1 is evaluated in Esimai (1977) and found to be

$$B_1 = -\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\, H_{p+2}(b;\ \delta) \tag{2.2}$$

where $H_{p+2}(b;\ \delta)$ is the cumulative distribution function of the noncentral chi-squared distribution with $p+2$ degrees of freedom and noncentrality parameter $\delta = n\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X$.

Without loss of generality, we let $\underline{\Sigma}_{22} = I$ and $\sigma^2 = 1$. Since B_1 changes sign with $\underline{\Sigma}_{12}$ and $\underline{\mu}_X$, we need only study the bias for $\mu_X > 0$ and $\rho > 0$ for $p = 1$. The bias was also studied by Han (1973) where the bias was expressed in terms of the cumulative distribution function of the standard normal distribution. The two expressions are equivalent as they should be. The general behavior of $-B_1$ is as follows. The bias is zero when $\mu_X = 0$ which is when the null hypothesis is true. It is an increasing function of ρ, but a decreasing function of α. For fixed n_1, α and ρ, the bias increases from zero and then decreases to zero as μ_X increases from zero to one. The values of $-B_1$ for $n_1 = 30$, $p = 2$ and certain values of $\underline{\Sigma}_{12}$, $\underline{\mu}_X$ and α are given in Table 1. The properties of the bias are found to be identical with those recorded for $p = 1$.

Table 1. Values of $-B_1$ for $p = 2$, $n_1 = 30$

$\underline{\mu}_X$ \ $\underline{\Sigma}_{12}'$	$\begin{pmatrix}.7\\0\end{pmatrix}$	$\begin{pmatrix}.7\\.7\end{pmatrix}$	$\begin{pmatrix}-.5\\.7\end{pmatrix}$
	$\alpha = .05$		
(0, 0)	0	0	0
(.5, 0)	.07	.07	-.05
(.5, .5)	.06	.12	.02
(1, 0)	0	0	0
	$\alpha = .10$		
(0, 0)	0	0	0
(.5, 0)	.04	.04	-.03
(.5, .5)	.04	.07	.01
(1, 0)	0	0	0

The MSE of $\hat{\mu}_{\ell r}$ was found to be $M_1 = MSE(\hat{\mu}_{\ell r}) = g_1 + h_1$ where

$$\begin{aligned} g_1 &= (1/n_2)\sigma^2 + (1/n_1)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21} \\ &\quad -(1/n_2)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}, \end{aligned} \tag{2.3}$$

$$\begin{aligned} h_1 &= \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}\,\underline{\Sigma}_{21} \\ &\quad -(1/n_1)\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}\, H_{p+2}(b;\ \delta) \\ &\quad -2\,\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}[1 - H_{p+2}(b;\ \delta)] \\ &\quad +\underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}\,\underline{\mu}_X\,\underline{\mu}_X'\,\underline{\Sigma}_{22}^{-1}\,\underline{\Sigma}_{21}[1 - H_{p+4}(b;\ \delta)]. \end{aligned}$$

Now we compare the performance of the preliminary test estimator, $\hat{\mu}_{\ell r}$ with the usual linear regression estimator, $\bar{y} + \underline{\Sigma}_{12}\,\underline{\Sigma}_{22}^{-1}(\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$,

when the information of $\underline{\mu}_X$ is ignored. The relative efficiency of $\hat{\mu}_{\ell r}$ to the linear regression estimator is defined as

$$e_1 = \frac{MSE\ (\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)}{MSE\ (\hat{\mu}_{\ell r})} = \frac{g_1}{g_1 + h_1} \quad (2.4)$$

Table 2. Values of e_1 for $p = 2$, $n_1 = 30$, $n_2 = 10$

$\underline{\Sigma}_{12}$ \ $\underline{\mu}_X$	$\begin{pmatrix}0\\0\end{pmatrix}$	$\begin{pmatrix}.5\\0\end{pmatrix}$	$\begin{pmatrix}.5\\.5\end{pmatrix}$	$\begin{pmatrix}1.0\\0\end{pmatrix}$
		$\alpha = .05$		
(.7, 0)	1.24	.69	.70	1.0
(.5, .5)	1.25	.83	.52	1.0
(.7, .7)	4.07	.56	.22	1.0
(-.5, .7)	1.64	.80	1.04	1.0
		$\alpha = .10$		
(.7, 0)	1.19	.78	.79	1.0
(.5, .5)	1.20	.88	.64	1.0
(.7, .7)	2.71	.67	.32	1.0
(-.5, .7)	1.48	.87	1.02	1.0
		$\alpha = .25$		
(.7, 0)	1.11	.91	.92	1.0
(.5, .5)	1.11	.95	.83	1.0
(.7, .7)	1.61	.85	.57	1.0
(-.5, .7)	1.24	.95	1.01	1.0

Without loss of generality we let $\underline{\Sigma}_{22} = I$ and $\sigma^2 = 1$. The values of e_1 for $p = 1$ are given in Han (1973) and will not be given here. The values of e_1 for $p = 2$, $n_1 = 30$, $n_2 = 10$ and certain values of $\underline{\Sigma}_{12}$, α and $\underline{\mu}_X$ are given in Table 2. It is seen that e_1 assumes maximum value at $\underline{\mu}_X = \underline{0}$. The maximum value of e_1 is an increasing function of ρ for fixed α, n_1 and n_2. The value of e_1 decreases to a minimum and then increases to unity as $\underline{\mu}_X'$ increases from $(0, 0)$.

The estimator $\tilde{\mu}_{\ell r}$ in (1.2) is given when $\underline{\Sigma}$ is unknown. The bias, B_2, and the mean square error, M_2, are derived in Esimai (1977) and are omitted here. The behavior of B_2 is the same as that of B_1 and the behavior of M_2 is similar to that of M_1.

3. The Optimal Sample Design

We shall now consider the problem of finding the optimum allocation of the sample sizes n_1 and n_2 for the estimator $\hat{\mu}_{\ell r}$ the cost function is

$$C = n_1 c_1 + n_2 c_2 \quad (3.1)$$

where c_1 is the cost of observing the vector $\underline{X}$ and c_2 is the cost of observing Y. The optimum values of n_1 and n_2 are obtained by minimizing M_1 subject to the constraint (3.1). We recall that in practice, under the supposition of a conditional specification, the experimenter has only partial information based on which he believes that $\underline{\mu}_X$ is close to $\underline{0}$. The relative efficiency of $\hat{\mu}_{\ell r}$ is the largest at $\underline{\mu}_X = \underline{0}$ and so it would be reasonable to consider the problem of optimum allocation under the optimum situation by letting $\underline{\mu}_X = \underline{0}$ in M_1. When $\underline{\mu}_X = \underline{0}$, M_1 becomes

$$M_1 = k_1/n_1 + k_2/n_2 \quad (3.2)$$

where

$$k_1 = \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}\ [1 - H_{p+2}\ (b;\ 0)]$$

$$k_2 = \sigma^2 - \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$$

Minimizing (3.2) subject to (3.1) we find

$$n_1 = \frac{C\sqrt{k_1}}{\sqrt{k_2 c_1 c_2} + c_1\sqrt{k_1}},$$

$$n_2 = \frac{C\sqrt{k_2}}{\sqrt{k_1 c_1 c_2} + c_2\sqrt{k_2}} \quad (3.3)$$

and the optimum value of M_1 is

$$M_{1,\ opt} = \frac{(\sqrt{k_1 c_1} + \sqrt{k_2 c_2})^2}{C} \quad (3.4)$$

We now compare $M_{1,\ opt}$ with the optimum value of the MSE of $\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$, the regression estimator under double sampling without using the preliminary test. If we denote the MSE of $\bar{y} + \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ (\bar{\underline{X}}_1 - \bar{\underline{X}}_2)$ by M

$$M = k_1'/n_1 + k_2'/n_2 \quad (3.5)$$

where $k_1' = \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$, $k_2' = \sigma^2 - \underline{\Sigma}_{12}\ \underline{\Sigma}_{22}^{-1}\ \underline{\Sigma}_{21}$ and the optimum value of M is

$$M_{opt} = \frac{(\sqrt{k_1' c_1} + \sqrt{k_2' c_2})^2}{C} \quad (3.6)$$

To compare (3.4) and (3.6) we note from (3.2) that $(1 - H_{p+2}\ (b;0))$ is a decreasing function of b with a maximum equal to unity at $b = 0$. Hence the numerator of $M_{1,\ opt}$ at most as large as that of M_{opt} and $M_{1,\ opt} \leq M_{opt}$ with equality holding for $b = 0$, i.e. when the two estimators coincide.

References

Bancroft, T. A. (1944). On biases in estimation due to the use of preliminary tests of significance. Ann. Math. Stat. 15, 190-204.

Bancroft, T. A. and Han, C. P. (1977). Inference based on conditional specification: A note and a bibliography. Accepted for publication in International Statistical Review.

Esimai, Grace O. (1977). Regression estimation for multivariate normal distributions. Unpublished Ph.D. thesis, Iowa State University, Ames, Iowa.

Han, C. P. (1973). Double sampling with partial information on auxiliary variables. J. Amer. Statist. Assoc. 68, 914-918.

DISCUSSION

Judy A. Bean, University of Iowa

First, I wish to thank the authors for presenting a set of interesting papers dealing with a range of problems that survey statisticians encounter. In order to allow time for discussion from the audience and on account of my own research interests, I intend to restrict myself to making specific comments on the three papers concerning the use of the balanced half-sample technique.

I am sure everyone here is aware of the increasing use of surveys to collect data. As more research is accomplished with surveys, fundamental and philosophical issues are being raised regarding the validity of inference. As the first step in the process of making statements about population parameters, variances of the estimates of the parameters are necessary. One such technique that has been employed to estimate variances is the balanced half-sample method. Today, we have heard the results of three investigations concerning the properties of the estimates of the procedure. For brevity I will refer to the balanced half-sample technique as BHS.

I. Estimating the Variance of the Slope of a Linear Regression in a Stratified Random Sample with the Balanced Half-Sample Technique

This paper provides us with more evidence on the behavior of the BHS technique for estimating the variance of a slope along with exposing us to a new form of the BHS method and to three ways of estimating the slope itself. Because of its flexibility the BHS method has been employed by survey statisticians, for example, Leslie Kish of the Survey Research Center at the University of Michigan, to estimate the variance of the slope but just what sample size is needed to yield an adequate stable estimate is not known. These results for the particular sample design used are the beginnings of guidelines needed by practicing statisticians. Naturally, more work in the area is needed.

As far as the comparison between the two forms of the BHS estimator, it would have been interesting to have included the estimator which is an average of the estimate obtained from the half-samples and the estimate from the complement half-samples since other investigations have shown this average to be better than the estimate obtained from the half-samples only. However, I suspect that in this situation the complement estimate would almost be identical to the usual half-sample estimate. Also, I think it may have been helpful to the reader if the expected value of this different form of the BHS method for the linear case would have been given. Unfortunately, as often happens, the results of the sampling experiments do not give a definite answer to the question of which form should be used. I wonder, for the "full-matrix" case, if the conclusion can be made that this different form will yield the same or smaller variances and mean square errors than the usual BHS estimator.

II. The Behavior of Balanced Half-Sample Variance Estimates for Linear and Combined Ratio Estimates When Strata Are Paired to Form Pseudo-Strata

When designing sample surveys, practicing statisticians often wish to select one primary unit per stratum in order to take full advantage of possible stratification gains. Thus, one does not have a satisfactory method for estimating variance from the sample itself; on the other hand, selecting two or more units from a stratum may obliterate potential gains in stratification. I am therefore delighted to see a study in which the problem is approached both mathematically and empirically.

As one would expect, the simulation results indicate that, when the method of collapsed strata is used, the BHS estimator of the variance of a ratio estimator is biased with the magnitude of the bias depending upon the formation of pairs of strata. The results are useful in establishing a direction for more research effort. The next step would be to examine the magnitude of the bias as the number of strata increases.

The other aspect of the problem discussed is whether or not for a large scale survey actually using both the method of collapsed strata and the balanced half-sample technique, different schemes of pairing strata has a practical effect. For this limited case, the answer was no. However, further study of the problem needs to be done. George Schnack of the National Center for Health Statistics and I have recently finished a feasibility study of the application of the BHS method for estimating variance components [Bean and Schnack (1977)]. As a substudy of that investigation, we noted that the formulation of the pairs of strata does affect the estimates of variance components for the Health Interview Survey. Thus, collapsing strata may seriously influence the estimates of components of variance necessary for designing purposes.

III. Evaluation of the Balanced Half-Sample Estimates for Linear and Combined Ratio Estimates for Non-Normally Distributed Populations

This paper correctly points out that the BHS technique is used to estimate variances from sample variables that are known to have non-normal distributions. Prior to this work, no study of the BHS estimator using Monte Carlo sampling from specified non-normal distributions has been performed. Thus, Dr. Hislop's work will add to the growing body of knowledge about the behavior of BHS estimators.

The results for the linear case together with the findings of other investigations have shown conclusively, I think, that when the estimator of the population parameter is linear, regardless of the underlying distribution of the variable, BHS estimates are satisfactory.

I was glad to learn that when the sample size is at least ten per stratum (so that the total sample size is at least 30) the results for the combined ratio estimate support the use of the balanced half-sample technique. The fact the estimates are extremely biased when the sample size is only two per stratum for three strata

does not alarm me since in practice for this situation one would not use the method. However, because of the magnitude of bias, a survey statistician should be cautious in using the technique when studying subdomains having small sizes as the paper indicates.

It would have been helpful if specific examples of variables having distributions studied here were given. Also, it is important to know if the numerator and denominator of the ratio for the various combinations of distributions are independent.

References

Bean, Judy A. And Schnack, George A., "A Feasibility Study of Estimating Variance Components Using the Balanced Repeated Replication Method." University of Iowa (Internal Memorandum).

A SURVEY OF MATHEMATICAL MODELS FOR GRAPHS

Stanley S. Wasserman, University of Minnesota

Introduction

The basic mathematical concept in this paper is the directed graph, or digraph, which is defined as a set V of nodes or "points" and a set L of directed arcs or "lines," connecting pairs of nodes. The set V contains g distinct elements, $v_1, v_2, \ldots, v_g$, and the set L contains C arcs, $\ell_1, \ell_2, \ldots, \ell_C$. We further require that no two distinct lines be in parallel; i.e., there exists at most one line ℓ_i connecting node v_j to node v_k. For convenience and to adhere to the established convention, a loop, a line connecting v_j to v_j, is not allowed in the digraph.

Digraphs differ from the more common undirected graphs because they have the additional characteristic that every line has an orientation or direction. Digraphs in which an arc from v_j to v_k implies the existence of an arc from v_k to v_j are symmetric. Symmetric digraphs are, of course, undirected graphs. When we desire to denote a directed line in terms of its two points, we write $\ell_i = v_j v_k$ for the directed line running from v_j to v_k. We let D_g be a specific digraph on g nodes. Note that D_g is a zero-one, or binary directed graph. The strengths or intensities attached to each arc are irrelevant, since our definition does not allow for the existence of valued lines. In what follows, we discuss mathematical representations for both directed graphs and undirected graphs, although we concentrate on the more general directed graph.

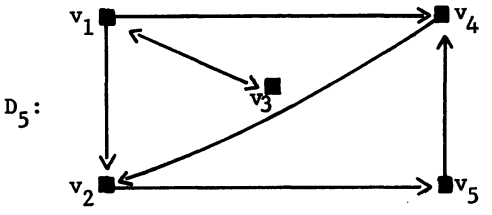
A digraph D_g is easily represented by a (gxg) matrix. We define a matrix $\underset{\sim}{X}$, with elements

$$X_{ij} = \begin{cases} 1, & \text{if } v_i v_j \in L \\ 0, & \text{otherwise.} \end{cases}$$

The matrix $\underset{\sim}{X}$ is called the adjacency matrix of D_g and has one row and one column for every node in V. An adjacency matrix for an undirected graph if, of course, symmetric. A different ordering of the elements in V produces an adjacency matrix that differs from $\underset{\sim}{X}$ by a simultaneous row-column permutation. Two digraphs with g nodes whose adjacency matrices differ by such a row-column rearrangement are called isomorphic. Note that since loops are not allowed, the diagonal elements of $\underset{\sim}{X}$, X_{ii}, i=1, 2,...,g, are set to zero.

Two sets of quantities are particularly interesting. The outdegree of node v_i, written r_i, is the number of arcs originating at node v_i. The indegree of node v_j, written c_j, is the number of arcs terminating at node v_j. Every element in $\underset{\sim}{r}$ and $\underset{\sim}{c}$ takes on a value between 0 and (g-1). Figure 1 shows an example of a digraph and associated adjacency matrix, including indegrees and outdegrees. The standard reference for these concepts is Harary, Norman, and Cartwright [1965].

This discussion of mathematical models for graphs is both a literature review and a collection of future suggestions for graph modelling. We present several models originally developed for processes other than graphs, giving model assumptions and a few derived results. We also comment on the applicability of these models to directed graphs, and in particular, social networks. Quite a few ideas for future research are given. We feel that a thorough understanding of existing models applicable to directed graphs is an essential prerequisite for the development of relevant and encompassing stochastic models for social networks.



$\underset{\sim}{X}$:

	v_1	v_2	v_3	v_4	v_5	r_i
v_1	0	1	1	1	0	3
v_2	0	0	0	0	1	1
v_3	1	0	0	0	0	1
v_4	0	1	0	0	0	1
v_5	0	0	0	1	0	0
c_j	1	2	1	2	1	

Figure 1: Digraph and Associated Adjacency Matrix.

Categorizing Mathematical Models

In addition to reviewing mathematical models specifically developed for graphs, we examine, in some detail, mathematical models originally constructed for entities other than directed graphs. The structure considered are models for processes from such natural science fields as statistical physics, organic chemistry and biology, and biophysics, easily modified to become models for graphs. The models from the literature for graphs themselves, are, by and large, from the social sciences, and postulate mathematical representations for social networks, specific sets of social relations linking members of well-defined groups. Mathematically, a social network may be defined as a binary directed graph with nodes for individual group members and arcs for the relational links.

Before we present the various models, we note that all mathematical models (for many kinds of processes) can be dichotomized twice: one by a deterministic vs. stochastic division and once by a static vs. time-dependent split. The first

dichotomy is a function of whether or not the model under consideration incorporates probabilistic assumptions. Deterministic models allow no opportunity for the graph to deviate from a prescribed pattern, usually given by a system of differential equations, or substantive theory. Stochastic models by definition do not allow the current or future structure of a graph to be predicted with certainty. The second dichotomy is solely a function of what the model postulates about time. Does the model in question assume that a graph evolves over time (time-dependence) or not (stasis)?

Deterministic, Static Models

Until very recently, the analysis of directed graphs, particularly in the social science context, was static and deterministic. Networks were not explicitly assumed to evolve over time, and conclusions drawn from a single network were deterministic, or precisely defined. The sociological version of Heider's balance theory (see Heider [1958]) was the prevalent paradigm. Leinhardt [1977] discusses the beginnings of network analysis, focusing on Heider's contribution. Heider's research was generalized by Cartwright and Harary [1956] in a paper where formal graph theory was introduced to social network research.

The predictions of Cartwright and Harary's structural balance theory did not accord with reality. Davis [1967], referring to the lack of empirical support for the dichotomous cliquing of groups predicted by Cartwright and Harary's theorem, further elaborated on the balance paradigm, extending it to multiple clusters of individual. However, the deterministic nature of the theory was retained, and consequently, the model's fit to empirical data remained poor. What was needed was a model incorporating probabilistic assumptions on the relations among group members. In a series of papers, Davis, Leinhardt, and Holland built a stochastic component into the paradigm.

Stochastic, Static Models

The Davis-Holland-Leinhardt methodology involves computing conditional uniform distributions on the space of all directed graphs. The most highly conditioned distribution controls for the dyad census, or the number of mutual, asymmetric, and null arcs in a digraph (see Holland and Leinhardt [1975]). Essentially, one computes the first two moments of the 16 component triad census, a count of the isomorphism classes of the $\binom{g}{3}$ triads in a digraph, and compares the empirically determined triad census with its expectation. Davis [1977] reviews this line of research, and Wasserman [1977a] discusses other random directed graph distributions.

This approach is static in time, since it concentrates on only one adjacency matrix. It is, however, stochastic. The analysis can even be compared to current methodology on stochastic processes. Holland and Leinhardt essentially compute equilibrium distributions for digraphs, and assume that data on the digraph process follow these distributions. One outstanding question is whether any of these "equilibrium" distributions are true equilibrium distributions obtained from some stochastic process. Further research may clarify this issue.

Deterministic, Time-Dependent Model

Several deterministic models for directed graphs have been proposed. Differential equations are the driving forces of such models, in which the effect of any change in the system can be predicted with certainty. However, in the social sciences, and to a lesser extent in the natural sciences, changes in a system cannot be predicted with certainty, usually because of the unpredictable nature of the objects being modelled. This uncertainty is best modelled through the use of probability distributions on random variables instead of the "controlling" mathematical variables of a system of differential equations. (A blend of the two approaches would be promising, but no such model has been developed.) We prefer to concentrate on the more realistic set of stochastic models, and we merely refer the reader to Bernard and Killworth [1977] for a recent review of deterministic models.

Stochastic, Time-Dependent Models

For the remainder of this paper, we discuss stochastic, time-dependent models, first from the social sciences, and then from the natural sciences.

The first model is the "Dynamic Model" of Holland and Heinhardt [1977a]. The Holland-Leinhardt stochastic model is actually an encompassing framework for the modelling of graphs, more general than an explicit statement on the evolution of digraphs through time. The framework operates on the individual arcs in L, the most elementary and basic level of a digraph. In Wasserman [1977b], we develop this modelling system theoretically, and discuss several simple parameterizations and estimation of structural parameters.

We next present three other models for social networks. These are a model in discrete time by Katz and Proctor [1959], a model based on learning theory of Rainio [1966], and a more recent model of Sørenson and Hallinan [1976].

Following the social science models, we discuss three models from the natural sciences. The first model that we shall discuss is for percolation processes of the flow of fluid through a medium. Broadbent and Hammersley [1957] give a mathematical formulation of percolation theory as it applies to crystals and mazes. Frisch and Hammersley [1963] present a thorough review of the theory, giving definitions and listing some of the results available at the time and unsolved problems.

Secondly, we shall describe a stochastic model for polymerization, or the evolution of polymers in organic chemistry. Polymers are "units" (or atoms) which associate into clusters and are also capable of disassociation. The model is Whittle's [1965a, 1965b], and is based on both the Gibbs equilibrium distribution for an ensemble of particles, and the deterministic kinetic equations of thermodynamics. The blend of these two approaches produces a unique set of stochastic kinetic equations as a model for polymerization.

Next, we discuss a model for neural networks of biophysics proposed by Rapoport. Rapoport's models of random and biased nets are not stochastic in nature; however, we include them here because the various types of biased nets are parallel to the simple stochastic models discussed by Wasserman [1977b]. Rapoport's notion of

"biases" may even be considered as the theoretical forerunner of the structural parameters of the Holland-Leinhardt framework. These models are presented in a group of papers written in the 1950's by Rapoport, in the Bulletin of Mathematical Biophysics. Rapoport [1957] reviews the contributions to the theory of random and biased nets, and Rapoport [1963] discusses the importance of nets to the theory of social interaction.

Throughout this section, we let $\underset{\sim}{X}(t)$ be the adjacency matrix representing the state of the digraph at time t. The binary-valued matrix $\underset{\sim}{X}(t)$ has elements $(X_{ij}(t))$ where

$$X_{ij}(t) = \begin{cases} 1, \text{ if } v_i v_j \in L \text{ at time } t \\ 0, \text{ otherwise.} \end{cases} \quad (1)$$

The time parameter, t, is assumed continuous, $t \geq 0$. Throughout, we set the g diagonal terms, $(X_{ii}(t))$, to 0 for all i and t.

We let $\underset{\sim}{x}$ be a single state of the continuous time stochastic process $\underset{\sim}{X}(t)$. The process has a finite state space $\underset{\sim}{S}$ of all possible $2^{g(g-1)}$ binary-valued (gxg) matrices with zero diagonal. We shall let $\underset{\sim}{w}$, $\underset{\sim}{x}$, $\underset{\sim}{y}$, $\underset{\sim}{z}$, ... denote elements of the state space.

Social Science Models

1. Holland-Leinhardt Framework

The Holland-Leinhardt framework is merely two simple assumptions regarding the stochastic nature of the arcs $X_{ij}(t)$. The first assumption is that $\underset{\sim}{X}(t)$ is Markov chain. Thus by Assumption 1, the current state of the process is all that we need to predict future behavior of the process.

We make an additional assumption regarding conditional independence of the elements of $\underset{\sim}{X}(t+h)$ given $\underset{\sim}{X}(t)$ for small h, conditional choice independence.

This is a critical assumption and in unique to this framework. It states that for very small intervals of time, the changes in a digraph are statistically independent. Consequently, the probability that any two arcs change simultaneously is essentially zero. This assumption is crucial for theoretical results, since it greatly simplifies the mathematics.

The infinitesimal transition rates depend on the entire adjacency matrix at time t, and may imply complex interrelations among the elements of $\underset{\sim}{x}$. Holland and Leinhardt [1977a, 1977b] and Wasserman [1977b] give examples of various functions specifically postulating that the infinitesimal transition rates of the digraph process are linear functions of various graph-theoretic quantities.

This line of inquiry into the nature of social structure and evolution of social networks is unique because of the proposed framework for parameterization. By assuming that, for small intervals of time, the arcs of a digraph operate in a statistically independent manner, we are able to assume various functional forms for the infinitesimal transition rates of the process. Thus, a researcher may define "social structure" by a set of graph-theoretic quantities, and combine these in a linear fashion to form the change rates of the process. This aspect of the framework is an important contribution to mathematical sociology, being an explicit statement on the evolution of a digraph as a continuous time Markov chain and providing a "wide open" framework for quantifying social structure.

2. Other Models from Social Science

There have been other attempts at modelling social networks as stochastic processes. In an early paper, Katz and Proctor [1959] analyze a sociomatrix at the level of dyads, or pairs of nodes in the digraph. The authors assume that the $\binom{g}{2}$ dyads are independent observations on a time discrete Markov chain, and therefore test whether a specific data set is compatible with the assumptions of a Markov chain. Unfortunately, no explicit structural model is developed for the evolution of a network over time.

Rainio [1966] develops a stochastic theory of social interaction. He posits a vector of probabilities, summing to unity, that regulates the frequency of interaction between individual i and the remaining (g-1) individuals in the group. These g vectors, one for each individual, evolve over time. This model is applied to a group of twelve girls, and the individual learning parameters (α,β) varied to provide the best fit of the model to the data. The model is very similar to the learning theory models developed by Bush and Mosteller [1955], and although its discrete time nature is a great simplification, it is an important contribution.

More recently, Sørenson and Hallinan [1976] hypothesize that each triad, or triple of nodes, in a network is a continuous time Markov chain. However, unlike the $\binom{g}{2}$ dyads in a network, the $\binom{g}{3}$ triads are not independent, and the assumption that the set of triads are independent observations on a basic Markov chain is incorrect. Unlike the model of Sørenson and Hallinan, the Holland-Leinhardt framework operates at the level of individual choices, the most elementary and basic level of a network. Placing a stochastic mechanism on the dyads or triads and ignoring subgraphs of lesser order is indeed less accurate in describing the operational behavior of a group.

Natural Science Models

1. Percolation Processes

Percolation theory seeks to describe the spread of a fluid throughout a medium. The random mechanism can either be attributed to the fluid or the medium: the former alternative is easily recognized as a diffusion process, while the latter is a percolation process. By its nature, percolation theory is more deterministic than diffusion theory, being subject to more restrictive assumptions, and certainly less widely known. The examples of percolation processes are many, ranging from fractures of crystals, and water absorption in a porous solid, to spread of blight in an orchard. Percolation theory stands apart from general epidemics (see Bartholomew [1973], Chapter 9 and 10) in that the medium under consideration is constrained by a particular geographic structure.

Percolation theory considers the following problem:

Let C be a connected graph with a countable set of nodes $\{X_i\}_{i=0}^{\infty}$ and arcs $\{L_{ij}\}$ joining X_i to X_j. Each arc L_{ij} is blocked, so that no fluid may traverse it, with probability $1-p_{ij}$, and unblocked, with probability p_{ij}. We then supply fluid to a random set of arcs, and study the flow of fluid through the system. This is the simplest case in percolation theory. More complicated situations are given in Frisch and Hammersley [1963].

As one can see, percolation theory might be quite important to the study of diffusion of innovations through a social network. Unfortunately, because of the complicated mathematics, it is virtually inaccessible to social scientists, and very rarely referenced.

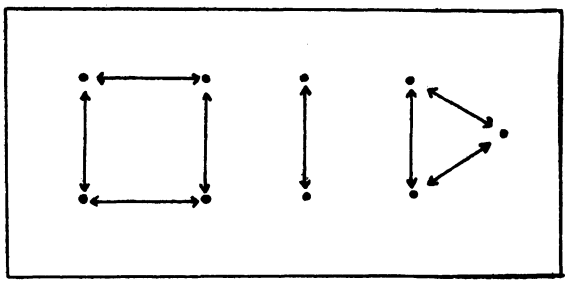
2. Polymerization Processes

We shall now consider Whittle's model for polymerization. The polymerization process has a state space of all symmetric graphs whose adjacency matrices can be permuted to block diagonal form, a space much smaller than S. An example of polymerization will be illustrated.

Suppose at time t, we have a group of size g that is composed of $k \leq g$ distinct clusters of nodes or cliques or polymers, such that no arcs exist between cliques, and within each clique, all arcs are present. Thus, each clique is strongly connected, and the adjacency matrix for the digraph can be permuted to a matrix with blocks of ones along the diagonal, one block per clique, and zeros elsewhere. Figure 2 depicts a situation with 9 nodes and 3 clusters.

When two-polymers come together, all arcs between the two come into existence, so that the new clique is also strongly connected. When a polymer disassociates into two new polymers, all arcs between the two smaller polymers disappear. Thus we always have a symmetric digraph with block-diagonal-permutable adjacency matrix. Note that the second assumption of the Holland-Leinhardt framework does not apply to Whittle's polymerization process, since, in general, a large number of arcs change simultaneously whenever polymers associate or disassociate.

Recently, in sociology, there has been renewed interest in clique formation. One of the proposed models for cliques in social networks (Breiger, Boorman, and Arabie [1975] and White, Boorman, and Breiger [1976]), the blockmodel, is not stochastic. Merging Whittle's stochastic model of group structure with the blockmodel proposed by White, et al, would be a substantial contribution to the analysis of social networks. There are other, simpler stochastic models for group changes. Morgan [1976] gives a review of these models, in addition to extending Whittle's results by proving the polymerization model to be reversible.



0	1	1	1	0	0	0	0	0
1	0	1	1	0	0	0	0	0
1	1	0	1	0	0	0	0	0
1	1	1	0	0	0	0	0	0
0	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	0	0	0	0	0	0	1	1
0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	1	1	0

Figure 2: Polymerization Digraph and Permuted Adjacency Matrix.

3. Random and Biased Nets

Next, we discuss Rapoport's models for random and biased nets. Rapoport defines a random net as a binary directed graph with each node assigned a fixed outdegree a. This "fixed choice" adjacency matrix is only random in the conditional sense that every node in the group is equally likely to receive one of the a arcs of node i. The adjacency matrix is conditioned to have a fixed vector of outdegrees $(a,a,\ldots,a)'$.

Rapoport assumes a single adjacency matrix, with some fixed outdegree, and examines a tracing of the network. A tracing is merely a path through the net, beginning at an arbitrary number of nodes. We then record these nodes as the initial set and all new nodes that are chosen by the initial set are termed first remove,. This tracing is continued, all the while recording the fraction of the population present in the initial set, first remove, second remove, etc. By examining these fractions, Rapoport estimates a by α, the apparent choice or axone density. This choice density was found to deviate from a empirically.

In an attempt to explain this deviation, Rapoport defines certain biases, operating in nets, that could cause the discrepancy. These biases include distance bias, symmetry bias, and transitivity biases. Distance bias decreases the chance that two individuals far apart from one another in the constructed "social space" will interact. Symmetry bias increases the chance of a choice $j \to i$, if the choice $i \to j$ is present, and as defined, is identical to the μ_0 and μ_1 terms in the mutuality

model of Wasserman [1977b]. Transitivity biases have a similar interpretation as τ_0, τ_1, and τ_2 in the model presented in Holland and Leinhardt [1977a].

Unfortunately, Rapoport is able to do little with these biases mathematically, except to estimate gross statistical features of the graph. There is a strong relation between Rapoport's work and the models of mutuality, popularity, and expansiveness discussed in Wasserman [1977b] utilizing the modelling framework. A further study of his relation would be useful.

Concluding Remarks

We have discussed several models for graphs and compared the stochastic, time-dependent models to the new modelling framework proposed by Holland and Leinhardt [1977a]. Directions for future research are indicated throughout this paper, specific ideas concerning how existing models could be accurately represented by this new framework. This research should prove both promising and exciting, with additional insight into the evolution of social networks over time as an added benefit.

Acknowledgements

Support was provided by NSF Grant SOC73-05489 to Carnegie-Mellon University. I thank Stephen Fienberg, Maureen Hallinan, Kenneth Land, Samuel Leinhardt, and Anatol Rapoport for helpful suggestions.

References

Bartholomew, D.G. [1973] Stochastic Models for Social Processes, London: John Wiley & Sons.

Bernard, H.R. and P.D. Killworth [1977] "Deterministic models of social networks" Social Networks: Surveys, Advances, and Commentaries, edited by P.W. Holland and S. Leinhardt, New York: Academic Press, to appear.

Breiger, R.L., S.A. Boorman, and P. Arabie [1975] "An algorithm for clustering relational data with application to social network analysis and comparison with multidimensional scaling," J. of Math. Psych. 12, pp. 328-383.

Broadbent, S.R. and J.M. Hammersley [1957] "Percolation processes. Crystals and mazes," Proc. Cambridge Phil. Soc. 53, pp. 629-641.

Bush, R.R. and F. Mosteller [1955] Stochastic Models for Learning, New York: John Wiley & Sons.

Cartwright, D. and F. Harary [1956] "Structural balance: A generalization of Heider's theory," Psychological Review 63, pp. 277-293. Also in Social Networks: A Developing Paradigm, edited by S. Leinhardt. New York: Academic Press, 1977.

Chung, K.L. [1967] Markov Chains with Stationary Transition Probabilities, 2nd Edition, New York: Springer-Verlag.

Davis, J.A. [1967] "Clustering and structural balance in graphs," Human Relations 20, pp. 181-187. Also in Social Networks: A Developing Paradigm, edited by S. Leinhardt, New York: Academic Press, 1977.

Frisch, H.L. and J.M. Hammersley [1963] "Percolation processes and related topics," J. of Society for Indust. and Applied Math. 11, pp. 894-918.

Harary, F., R.Z. Norman, and D. Cartwright [1965] Structural Models: An Introduction to the Theory of Directed Graphs, New York: John Wiley & Sons.

Heider, F. [1958] The Psychology of Interpersonal Relations, New York: John Wiley & Sons.

Holland, P.W. and S. Leinhardt [1975] "Local structure in social networks," Sociological Methodology, 1976, edited by D.R. Heise, San Francisco: Jossey-Bass.

Holland, P.W. and S. Leinhardt [1977a] "A dynamic model for social networks," J. of Math. Soc. 5, pp. 5-20.

Holland, P.W. and S. Leinhardt [1977b] "Social structure as a network process," paper presented at the International Conference on "Mathematical Approaches in Social Network Analysis," Bad Homburg, West Germany.

Katz, L. and C.H. Proctor [1959] "The concept of configuration of interpersonal relations in a group as a time-dependent stochastic process," Psychometrika 24, pp. 317-327.

Leinhardt, S. [1977] "Social networks: A developing paradigm," in S. Leinhardt, editor, Social Networks: A Developing Paradigm, New York: Academic Press.

Morgan, B.J.T. [1976] "Stochastic models of grouping changes," Adv. in Applied Prob. 8, pp. 30-57.

Rainio, K. [1966] "A study on sociometric group structure: An application of a stochastic theory of social interaction," Sociological Theories in Progress, Vol. 1, edited by J. Berger, M. Zelditch, and B. Anderson, Boston: Houghton-Mifflin.

Rapoport, A. [1957] "Contribution to the theory of random and biased nets," Bull. of Math. Biophy. 19, pp. 257-277. Also in Social Networks: A Developing Paradigm, edited by S. Leinhardt, New York: Academic Press, 1977.

Rapoport, A. [1963] "Mathematical models of social interaction," Handbook of Math. Psych. Vol. II, edited by R.D. Luce, R.R. Bush, and E. Galanter, New York: John Wiley & Sons.

Sørenson, A.B. and M. Hallinan [1976] "A stochastic model for change in group structure," Social Science Research 5, pp. 43-61.

Wasserman, S.S. [1977a] "Random directed graph distributions and the triad census in social networks," J. of Math. Soc. 5, pp. 61-86.

Wasserman, S.S. [1977b] "Stochastic models for directed graphs," Ph.D. Dissertation, Harvard University.

White, H.C., S.A. Boorman, and R.L. Breiger [1976] "Social structure from multiple networks. I. Block models of roles and positions," Amer. J. of Soc. 81, pp. 730-780.

Whittle, P. [1965a] "Statistical processes of aggregation of a clustering process in the uncondensed phase," Proc. of R. Soc., Series A 285, pp. 501-519.

Whittle, P. [1965b] "The equilibrium statistics of a clustering process in the uncondensed phase," Proc. of R. Soc., Series A 285, pp. 501-519.

CLASSIFYING AN INCOMPLETE VECTOR

Tom Bohannon, Appalachian State University
W. B. Smith, Texas A.& M. University

Introduction.

Discriminant analysis is concerned with the problem of assigning an observation vector, Z, of unknown origin to one of several distinct populations on the basis of some classification rule. Hodges [1950], Cacoullos [1973], and Lachenbruch [1975] give comprehensive lists of some case studies of various applications of discriminant analysis.

In our study, we shall only consider the situation where there are two p-variate normal populations Π_1 and Π_2 with distributions denoted by $N_p(\mu,\Sigma)$ and $N_p(\omega,\Sigma)$, respectively. We shall also assume that the probabilities of misclassification and the costs of misclassification are equal for the populations, thus the optimum classification rule is given by

$$D(X) = [X - 1/2(\mu + \omega)]'\Sigma^{-1}(\mu - \omega) .$$

In practice these parameters are usually not known and are estimated. These estimates are then substituted into the discriminant function, $D(X)$, to yield what is often referred to as Anderson's discriminant function, $W(X)$.

$$W(X) = [X - \frac{(\hat{\mu} + \hat{\omega})}{2}]\Sigma^{X-1}(\hat{\mu} - \hat{\omega}) ,$$

where

$$\hat{\Sigma} = \frac{(n_1-1)\hat{\Sigma}_1 + (n - 1)\hat{\Sigma}_2}{n_1+ n_2- 1}$$

$\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ being the unbiased estimates of Σ_1 and Σ_2 based on the data collected from their respective populations. An additional unclassified observation Z will be classified into Π_1 if $W(Z)$ is non-negative, otherwise into Π_2.

We consider the problem of classifying an observation vector,

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}$$

where Z_1 is a vector of observations and Z_2 is a vector with the components missing. The following notation will be used to denote the partitioned mean vectors and variance-covariance matrix:

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \omega = \begin{pmatrix} \omega_1 \\ \omega_2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Similar notation will be used for the estimates of these vectors and matrices.

Review of Discriminant Analysis With Missing Data

The use of discriminant analysis techniques on incomplete data sets is an area where very little research has been done. Jackson [1968] investigated a problem which had missing values in a discriminant problem where both the number of variables and the number of observations were large. Estimation of missing values and the number of observations were large. Estimation of missing values using mean and regression techniques were tried for the problem under study. The estimation procedure utilizing missing data gave more realistic results than the often used procedure of ignoring observation vectors with missing values.

Chan and Dunn [1972] investigated the problem of constructing a discriminant function based on samples which contain incomplete observation vectors. Several methods of estimating the components of the incomplete vectors were used to contruct the discriminant function. The effect on the performance of the discriminant function for each method was studied and compared. They concluded that no method is best for every situation, and gave guidelines to use in choosing the best method. Chan and Dunn [1974] studied the asymptotic behavior of these methods when the variables are equally correlated. They found that the differences of the asymptotic probability of correct classification from maximum were found to be small for all methods. Chan, Gilman and Dunn [1976] studied two additional methods and recommended their modified regression method.

Srivastava and Zaatar [1972] derived the maximum likelihood rule for incomplete data when a common covariance matrix is assumed known. Smith and Zeis [1973], using a generalization to the maximum likelihood estimation technique of Hocking and Smith [1968] and an application of the likelihood ratio criterion generalized the results of Srivastava and Zaatar to unknown and unequal covariance matrices. Bohannon [1976] compared this procedure to the standard procedure of ignoring the incomplete observations in the construction of the classification rule. The comparisons were made on the basis of the probability of misclassification and the proposed method performed best in the simulation study.

Classification Rules

Marginal Rule

If one ignores the variables in the vector Z_2 then the optimum rule is

$$V_1 = Z_1 - \frac{\mu_1 + \omega_1}{2}' \Sigma_{11}^{-1}(\mu_1 - \omega_1) .$$

This function shall be referred to as the marginal discriminant function. This function also results if one uses the regression approach and estimates Z_2 by the regression equation

$$Z_2 = \frac{\mu_2+ \omega_2}{2} + \Sigma_{21}\Sigma_{11}^{-1}\left(Z_1 - \frac{\mu_1 + \omega_1}{2}\right)$$

and substitutes this value into the discriminant function of Z. That is, using $\frac{\mu_1 + \omega_1}{2}$ as the mean for Z_1 and $\frac{\mu_2+ \omega_2}{2}$ as the mean for Z_2 yields V_1 in the regression approach.

Two-Stage Rule

For the first stage of the classification rule use the marginal discriminant rule to classify Z_1 into population Π_1 or Π_2. Now we have the vector Z in either Π_1 or Π_2 and we shall utilize the Smith-Hocking estimation procedure to estimate the mean vector and covariance matrices. These new estimates shall be denoted

$\hat{\mu}_2$, $\hat{\Sigma}_2$ and $\hat{\omega}_2$. If Z was classified into Π_1 then Z_2 is estimated by $\hat{\mu}_2$ where

$$\hat{\mu}_2 = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix}$$

and in a similar manner we estimate Z_2 by $\hat{\omega}_2$ if Z was classified into Π_2 by V_1. The preliminary simulation studies indicate this procedure performs better than the rule utilizing only V_1.

Application

A student at Tarleton State University whose curriculum requires college algebra may take the College Algebra course for credit if the student is prepared or may take the Fundamentals of College Algebra for credit and then follow with the College Algebra for credit. It is most beneficial for the student who is prepared for the College Algebra course to be counselled into that course, rather than to lose interest in both mathematics and a semester in college by taking the remedial Fundamentals of College Algebra course. Even more serious, is the incorrect placement of a student in College Algebra who is deficient in his mathematical training and should be placed in the remedial course in algebra.

In an attempt to aid in the proper placement of the students, Bohannon [1976] utilized discriminant analysis to place the students. The variables used in the analysis are defined in Table 1.

Table I

Variables Utilized for Classification

Variable	Description of Variables
X_1	Student's High School Algebra I Score
X_2	Student's SAT Math Score
X_3	Student's SAT Verbal Score
X_4	Student's High School Geometry Score
X_5	Student's High School Algebra II Score

The samples utilized to construct a discriminant function were the students who enrolled at Tarleton State University during the Fall Semester of 1973 without having taken the Fundamentals of College Algebra course. Population one is defined to be the set of students who receive or will receive a grade of C or better in the College Algebra course, and the students who receive below a C, or drop the course, constitutes population two.

The study found estimates for the error of misclassification and for the purpose of cross-validation of the discriminant function, similar samples were drawn from the students who enrolled in the College Algebra in the Fall of 1974. The samples contained partial data records for some of the students and hence the two-stage discriminant rule was applied to those vectors and the results are shown in Table 2.

Table 2

Prediction Results

Marginal Rule

Group	Predicted Group 1	2
1	12	9
2	7	19

Two-Stage Rule

Group	Predicted Group 1	2
1	13	8
2	5	21

Thus we observe that the two-stage rule yields slightly better results than the marginal rule.

REFERENCES

1. Bohannon, Tom [1976]. "Discriminant Analysis with Missing Data." Ph.D. dissertation, Texas A.& M. University, College Station, Texas.

2. Cacoullos, T. [1972]. Discriminant Analysis and Applications. Academic Press, New York.

3. Chan, L. S. and Dunn, L. J. [1972]. "The treatment of missing values in discriminant analysis-I. The sampling experiment." J.A.S.A. 67, 473-477.

4. Chan, L. S. and Dunn, L. J. [1974]. "A note on the asymptotic aspect of the treatment of missing values in discriminant analysis." J.A.S.A. 69, 672-673.

5. Chan, Gilman and Dunn [1976]. "Alternative approaches to missing values in discriminant analysis." J.A.S.A. 71, 842-844.

6. Hocking, R. R. and Smith, W. B. [1968]. "Estimation of parameters in the multivariate normal distribution with missing observations." J.A.S.A. 63, 159-173.

7. Jackson, E. L. [1968]. "Missing values in linear multiple discriminant analysis." Biometrics 24. 835-844.

8. Lachenbruch, P. A. [1975]. Discriminant Analysis. Hafner Press, New York.

9. Smith, W. B. and Zeis, L. D. [1973]. "On classification for incomplete multinormal data." Communications in Statistics 2, 85-92.

10. Srvastava, J. N. and Zaatar, M. K. [1972]. "On the maximum likelihood classification rule for incomplete multivariate samples and its admissibility." Journal of Multivariate Analysis 22, 115-126.

OPTIMAL DESIGN FOR PROGRAM EVALUATIONS[1]

Stuart H. Kerachsky, Mathematica Policy Research
Charles D. Mallar, Mathematica Policy Research and John Hopkins University

A. INTRODUCTION

This paper presents a methodology for selecting a control sample that can be used in program evaluations when the program treatment group is both geographically clustered and preselected.[2] The criteria used to choose a program evaluation design are the classical ones of minimizing bias and maximizing efficiency in the estimation of treatment effects. These criteria are satisfied if (1) there are no systematic differences between the treatment and control groups, and (2) the other variables that explain the behavioral outcomes are observed and included in the statistical analysis, in order to obtain precise estimates of standard errors. Generally, random assignment of the target population to treatment and control groups at the point of entry into the program will minimize bias and maximize efficiency, ceteris paribus. If the assignment deviates from a random one according to known and measured characteristics, unbiased but less efficient estimates can be obtained through multivariate techniques, as long as the treatment and control groups are sufficiently similar so that their behavioral relationships are structurally identical. (See Goldberger, 1972a and b; Cain, 1975; Pitcher, forthcoming; and Conlisk, forthcoming.) The quasi-experimental design developed below for treatment samples that are geographically clustered and preselected approximates random assignments at the point of entry into the program and will generally have the same properties if successfully applied.

B. THE GENERAL PROBLEM AND RESOLUTION

Geographically clustered programs include both those that are size-specific and those that draw heavily from only certain areas of the country. It is assumed that the selection of program sites can be either arbitrary or controlled for in the evaluation design. Evaluations of geographically clustered programs with treatment samples that have been preselected are common in the social sciences. Examples can be found in the evaluations of:

(1) Employment and training programs

(2) Education programs and projects

(3) Variants of state unemployment insurance programs

(4) Different public assistance programs

(5) Local transportation programs.

Many of these evaluations have been of ongoing programs for which random assignments of participants to treatment and control groups at the time of enrollment are not feasible. The potential for political, ethical, budgetary, and operational problems when intervening in the selection process for an ongoing program often precludes random assignment as a viable approach.[3] Consequently, the program treatment group is often preselected.

Previous evaluations have often relied on comparisons between the behavior of the program treatment group and another sample composed of some combination of the following:

(1) People on waiting lists for an oversubscribed program

(2) Early dropouts from the program

(3) Friends or relatives of those in the program

(4) People who have opted not to enroll (including "no shows"), or who have been screened out of the program

(5) Preprogram observations of the treatment group

(6) General population samples, including (at least some) program participants.[4]

The findings from such comparison-group evaluation studies have, in turn, been disputed because the assumptions needed to show unbiasedness and efficiency are not plausible.[5] Two likely sources of bias are unobserved differences in the sample (e.g., in terms of motivation) and overlap between the treatment and comparison groups. Even if there are no unobserved differences and no overlap between the groups so that unbiased measures of the treatment effect can be obtained, observed differences have often reduced the efficiency of estimates of the treatment effects. Furthermore, very disparate samples also strain the credibility of the underlying assumption that the treatment and comparison samples have the same behavioral structures (i.e., that the same equation is applicable to both groups).

Because of the geographic clustering, however, another approach can be developed. A random sample of program participants, combined with a sequentially matched sample from nonprogram sites, can approximate a random assignment strategy and thereby avoid bias and maximize efficiency. This sequential matching involves two distinct steps. First, a random sample of sites similar to those of the program are chosen for the control sites. Second, within these control sites, an appropriate sampling frame is set up, and individuals are randomly selected from the sampling frame for the comparison group. Throughout the remaining discussion it is assumed that the treatment sample for this quasi-experimental design is a random sample of people in the program.[6]

For the control group, a sample of sites must be selected that are similar to, but outside, the areas in which the program is clustered.

Program sites are excluded to minimize biases that result both from self-selection into the program (e.g., unobserved differences in motivation) and from treatments affecting the behavior of persons not in the program (especially for saturation programs). Selection probabilities are then assigned to the remaining sites (the nonprogram sites) in proportion to their similarity to the program site.

Once the control sites are chosen, a selection process similar to the de facto program selection process is then set up within the control sites, to yield a sampling frame of persons with observed and unobserved characteristics similar to program participants. The comparison group is then randomly chosen from the sampling frame with selection probabilities for individuals that are proportional to their similarity to program participants.

The sequential process of obtaining an appropriate comparison sample can be summarized as follows:

(1) Eliminate program sites from which participants are principally recruited.

(2) Assign probabilities of selection to nonprogram sites in proportion to their similarity to program sites.

(3) Randomly select the control sites based on the probabilities as assigned in step (2).

(4) Within control sites, eliminate any program participants.

(5) Assign probability of selection to other persons in proportion to their similarity to program participants.

(6) Randomly select individuals for the comparison group based on the probabilities as assigned in step (5).

This quasi-experimental design will yield treatment and control groups for which the assumptions needed to obtain unbiased and efficient estimates of treatment effects are usually plausible. The two groups are unlikely to differ systematically in either observed or unobserved characteristics, and there is no overlap in the samples. Finally, any observed differences that remain between the treatment and comparison groups can be controlled for in a multivariate estimation framework.

In some instances, the quasi-experimental design developed here will be preferable to random assignments at the time of enrollment. For example, randomization across sites is desirable when the fraction of the population being served is so large that the behavior of a within-site control group could be affected. This would of course be true for saturation programs in which a large portion of the eligible population is enrolled in the program.

C. AN APPLICATION TO AN EVALUATION OF THE JOB CORPS

The methodology developed above had recently been applied in a design for an evaluation of the economic impact of the Job Corps program on its participants (see Kerachsky and Mallar, 1977, for more details). The Job Corps program provides education, training, and support services in a residential setting to youths who come from severely disadvantaged families (youths age 16 to 23). Random assignments of potential enrollees to a control group were not feasible because of operational and other considerations. Therefore, the sequential matching process outlined above was instituted to obtain an appropriate comparison group.

First, program sites--both zip-code regions saturated by Job Corps participation (i.e., high proportions of eligible youths in the program) and zip-code regions proximate to Job Corps centers--were eliminated. Then the remaining regions were assigned selection probabilities in proportion to their similarities to the home regions of Job Corps members, based primarily on the poverty and racial compositions of the regions. Once the control sites were chosen, youths living in the relevant areas were assigned selection probabilities in proportion to their similarity to Job Corps participants, based primarily on their poverty, age, race, and educational status.[7] A sample of youths was then chosen for interviewing. Finally, the baseline questionnaire was designed to measure any observed differences that remained and which are now important for explaining the economic outcomes that are being studied.

This quasi-experimental design seems appropriate for the Job Corps evaluation and should lead to precise estimates of the economic impacts of the program. The assumptions needed for unbiased and efficient estimates of the program treatment effects seem plausible. There is no overlap, and with a large number of observations, the program treatment group should differ from a comparison sample only in terms of access both to information about Job Corps and to Job Corps centers. Therefore, a feasible program evaluation has been designed even within the constraints of an ongoing program.

D. CONCLUSIONS

A widely applicable technique for evaluating ongoing programs has been developed. The strategy for obtaining the comparison group sample is feasible and should lead to unbiased and efficient estimates of program treatment effects. The assumptions needed for minimizing bias and maximizing efficiency are plausible. There should be no overlap between the treatment and comparison samples, unobserved differences between the samples should be minimized, and observed differences should be small enough to be controlled for with a multivariate estimation technique, with only a small loss in efficiency.

FOOTNOTES

1. This paper summarizes a quasi-experimental design that was first developed for and applied to an evaluation of the economic impact of the Job Corps program on its participants (see Kerachsky and Mallar, 1977).

2. See the next section for precise definitions of "geographically clustered" and "preselected."

3. These problems are, of course, less important for experimental and demonstration programs.

4. The closer the match between these general population samples and the program sample, the greater the overlap between the samples--hence, the greater the biases.

5. See Goldstein (1973) for summaries and criticisms of several of the studies of employment and training programs.

6. Random selection as discussed here can be with or without stratifications.

7. Females were oversampled in the comparison group relative to Job Corps participants to increase the efficiency of separate estimates for females.

REFERENCES

Cain, Glen G. "Regression and Selection Models to Improve Nonexperimental Comparisons." In Evaluation and Experiment, edited by Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press, 1975.

Conlisk, John. "Choice of Sample Size in Evaluating Manpower Programs." In Research in Labor Economics, Supplement I: Evaluating Manpower Training Programs, edited by Farrell Bloch. Greenwich, Connecticut: JAI Press, forthcoming.

Goldberger, Arthur S. "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations." Institute for Research on Poverty Discussion Paper No. 123-72, Madison: University of Wisconsin, 1972a.

Goldberger, Arthur S. "Selection Bias in Evaluating Treatment Effects: The Case of Interactions." Institute for Research on Poverty Discussion Paper No. 129-72, Madison: University of Wisconsin, 1972b.

Goldstein, Jon H. "The Effectiveness of Manpower Training Programs: A Review of Research on the Impact of the Poor." Paper No. 3 of Studies in Public Welfare, Subcommittee on Fiscal Policy, Joint Economic Committee of the Congress of the United States. Washington, D.C.: Government Printing Office, 1973.

Kerachsky, Stuart H. and Charles D. Mallar. "Design of an Evaluation of the Job Corps." MPR Working Paper No. C-12, Princeton, New Jersey: Mathematica Policy Research, Inc., 1977.

Pitcher, Hugh. "A Sensitivity Analysis to Determine Sample Size for Performing Impact Evaluations." In Research in Labor Economics, Supplement I: Evaluating Manpower Training Programs, edited by Farrell Bloch. Greenwich, Connecticut: JAI Press, forthcoming.

COMPARISON OF ALTERNATIVE REGRESSION MODELS FOR PREDICTING CHANGE

Robert B. Bendel, Washington State University

ABSTRACT:

The consulting statistician frequently encounters problems in which an initial score (pre-test) and a final score (post-test) are observed. This paper contrasts three regression models which use the final score and change score (final score minus initial score) as dependent variables. It has been noted that for most problems the initial score should be used as an independent variable or covariate. When the two regression models which have the same independent variables but different dependent variables are contrasted, the models differ only in their multiple correlation coefficients but the standard error of estimate and other important statistics are the same. Tests of hypotheses conditional on the initial score are also the same for both models. An example is given and related topics encountered in the behavioral and animal sciences are discussed.

INTRODUCTION:

The problem is to see the mathematical relationships between three regression models with a view toward choosing the most appropriate model. Measurements on some variable are taken before (initial score) and after (final score) treatment. Treatment can be either quantitative (a regression problem) or qualitative (an ANOVA/ANCOVA problem). The mathematical aspects will be displayed in the more general context of a regression model but the results also apply to the more popular (special case) ANCOVA model. In the context of psychometrics, say, the problem can be viewed as the regression analogue of the pre-intervention-post design:

Pre-intervention-Post Design

	Pre	Post
Control		
Treatment		

Question researcher asks is: Is there more change in the treatment group than in the control group?

Statistical hypothesis: Test for equality of gains for the two or more groups.

Assumptions: At this point, we assume that all classical assumptions are satisfied, including random assignment to treatment groups.

DEFINITIONS:

IS = initial score
FS = final score
G = FS-IS = gain score

R^2 = coefficient of determination
= square of the multiple correlation coefficient
$s_{y.x}$ = standard error of estimate
= square root of the estimated variance about the regression line
RSS = residual sum of squares
edf = degrees of freedom for RSS

PROBLEM:

Which of the three regression models do we choose and what differences are there in the regression statistics, R^2 and $s_{y.x}$?

Model	Dependent Variable	Independent Variables
FS	FS	IS + $\underset{\sim}{x}_2$
G	G = FS-IS	IS + $\underset{\sim}{x}_2$
$\overline{IS}$	G = FS-IS	$\underset{\sim}{x}_2$ only

MODEL EXAMPLE:

For the G model, let $\underset{\sim}{x}_2$ represent the independent variables excluding IS. Then, for example, $\underset{\sim}{x}_2$ could represent age and a treatment variable, trt. The G model could then be represented by the equation

$$G = \beta_0 + \beta_1 \times IS + \beta_2 \times Age + \beta_3 \times trt.$$

We may be interested in predicting gain or we may be interested in seeing how treatment affects gain after adjustment for IS and age.

MODEL PREFERENCE:

Choose either the FS or G model since $s_{y.x}$ is the same for both models. R^2 is generally different. The $\overline{IS}$ model is usually inadequate since IS is frequently related to G or FS.

THEOREM:

The residuals, RSS and $s_{y.x}$ are the same for the FS and G models.

IDEA OF PROOF:

$$Var\ (FS-IS|IS) = Var\ (FS|IS)$$

That is, the variance about the population regression line is the same for the FS and G models since both have the same independent variables and both vary the same at each value of IS. To illustrate, consider the (IS, FS) data (1, 2), (1, 3), (1, 4), (2, 3), (2, 4), (2, 5) consisting

of three FS values at each of the two values of IS. A plot of IS versus FS and a separate plot of IS versus G indicate that $s_{y.x}$ is equal to one for both plots whereas $R^2 = .273$ for the (IS, FS) data and $R^2 = 0$ for the (IS, G) data.

PROOF:

Basic idea is to show that the regression coefficients for $\underset{\sim}{x}_2$ are the same for both models and that the regression coefficient for IS satisfies the condition b = 1 + g (This has been noted by Werts and Linn, 1970). Without loss of generality, let $\underset{\sim}{x}_2$ consist of one independent variable x_2. Then

$$RSS = \min \Sigma(FS - b_0 - b_1 \times IS - b_2 x_2)^2$$

for FS model and

$$RSS = \min \Sigma(FS - IS - g_0 - g_1 \times IS - g_2 x_2)^2$$

$$= \min \Sigma(FS - g_0 - (1 + g_1) \times IS - g_2 x_2)^2$$

for G model. Assuming there are no singularity problems, the Gauss-Markov theorem says that the least squares estimates of the regression coefficients are unique so,

$$b_0 = g_0$$

$$b_1 = 1 + g_1$$

$$b_2 = g_2$$

from which it follows that the residuals, $FS - b_0 - b_1 \times IS - b_2 x_2$, are the same for the FS and G models.

Hence, RSS and

$$s_{y.x} = (RSS/edf)^{1/2}$$

are also the same since edf = n - #parameters estimated is the same for both models.

EXAMPLE FROM PSYCHOLOGY (Mental Retardation):

Score = Adaptive Behavior (AB)
IS = initial score, AB at time 1
FS = final score, AB at time 2
$\underset{\sim}{x}_2$ = vector of independent variables, e.g., IQ, age, treatment = environmental factor score.

See Table 1. Note the following:

a) .88 = -.12 + 1.0 since $b_1 = g_1 + 1$

b) the standard errors are the same for the FS and G models.

c) the t values are the same (except for IS); for x_{23}, t = 5.83; so a test for significance of the partial regression coefficient of trt after adjustment for IS, age, etc. is highly significant.

d) $s_{y.x}$ is the same for the FS and G models.

e) R^2 is different; R^2 is usually higher for the FS model since the variance of FS is higher than the variance of G whenever IS and G are positively correlated - as is usually the case.

Note: In general, it can be shown that tests of hypotheses conditional on IS are the same for the FS and G models. To see this, let H_0: $\underset{\sim}{\beta}_q = \underset{\sim}{0}$ where $\underset{\sim}{\beta}_q$ is a vector of regression coefficients which does not include IS. Then the F test can be written as

$$F = (RSS' - RSS)\, edf/RSS\, (edf' - edf)$$

where RSS' and edf' denote the RSS and edf under the null hypothesis. Since IS is "included" in RSS and RSS' and edf and edf' are the same for both the FS and G models, it follows that tests of hypothesis, H_0: $\underset{\sim}{\beta}_q = \underset{\sim}{0}$, are identical for both the FS and G models.

EXAMPLE FROM ANIMAL SCIENCE:

IW = initial weight of steer
FW = final weight of steer
G = FW-IW = gain in weight
x_2 = treatment coded as 1, 2, 3, 4, which is simply a ranking of the amount of concentrate in the diet. For ANCOVA and ANOVA of gain scores, x_2 is used as a qualitative grouping variable.

Table 2 shows the results for comparing ANOVA of Gain, repeated measures (RM) ANOVA, ANCOVA with FW and G as the dependent variables and multiple regression using the treatment variable as a quantitative (1, 2, 3, 4) variable. Note that the ANOVA of Gain and the time by treatment interaction in the RM ANOVA test the same hypothesis (that the gain is the same for each treatment) so that the F value is the same as that in the ANOVA of Gain. In comparing the ANCOVA models, the same results are true for the ANCOVA models as are true for the regression models, i.e., that $s_{y.x}$ is the same, R^2 is generally different, and that the regression coefficient for IW is one more for the FW model than for the G model.

When all the assumptions are met including random assignment to treatment groups and the covariate and independent variables are measured without error, it has been established (Bock, 1975) that ANCOVA is more powerful than ANOVA of Gain scores (and repeated measures since the same F value is obtained when testing the treatment by time interaction in a repeated measures ANOVA).

DIFFICULTIES INVOLVED WHEN THE ASSUMPTIONS ARE VIOLATED:

The educational, psychological, and sociological literature contain many papers discussing the use of gain scores and ANCOVA when the assumptions are not met. The paper by Cronbach and Furby entitled, "How should we measure "change" or should we?" is a classic and Lord's Paradox (Lord and Novick, 1968; Bock, 1975) is also a controversial paper. In short, it is felt that there is some agreement that ANCOVA can be used with caution when there are intact groups (no random assignment to treatment). See Elashoff (1969), Kenny (1975), and Alwin and Sullivan (1975). Also, when the covariate is measured with error, there is a general agreement that some form of adjustment should be made, but as Cochran (1968) discusses, this could depend on the assumed model (Is there a linear regression of FS on "true" IS or on IS measured with error?). Werts and Linn (1970) and Bergman (1971) discuss alternative models to use when dealing with change. The problems with some of these models is that they require an estimate of the reliability of the covariate and/or independent variable. The reliability, $\tilde{R}$, is defined as the ratio of the variance of the true value to the variance of the observed value. To be specific, let $X = x + e$, where X is the observed score, x is the true score, and e is the error of measurement. Assuming that x and e are independent, it follows that

$$\tilde{R} = \sigma_x^2/\sigma_X^2 = \sigma_x^2/(\sigma_x^2 + \sigma_e^2).$$

CONCLUSION:

Use either the FS or G models, but there is some controversy and some unanswered problems when the assumptions are violated.

REFERENCES:

Alwin, D. F. and Sullivan, M. J., 1975. Issues of design and analysis in evaluation research. Sociological Methods and Research, 4(August): 77-100.

Bergman, L. R., 1971. Some univariate models in studying change. Reports from the Psychological Laboratories. The University of Stockholm Supplement Series 10.

Bock, R. D., 1975. Multivariate Statistical Methods in Behavioral Research. McGraw-Hill, New York.

Cochran, W. G., 1968. Errors of Measurement in Statistics. Technometrics, 13 (No. 4):637-666.

Cronbach, L. J. and Furby, L., 1979. How should we measure "change" - or should we? Psychological Bulletin, 74 (No. 1):68-80.

Elashoff, J. D., 1969. Analysis of covariance: a delicate instrument. American Educational Research Journal, 6 (No. 3):383-401.

Kenny, D. A., 1975. A quasi-experimental approach to assessing treatment effects in the nonequivalent control group design. Psychological Bulletin, 82 (No. 3): 345-362.

Lord, F. M. and Novick, M. R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, Mass.

Porter, A. C., 1973. Analysis strategies for some common evaluation paradigms. Occasional Paper No. 21, College of Education, Michigan State University.

Werts, C. E. and Linn, R. L., 1970. A general linear model for studying growth. Psychological Bulletin, 73 (No. 1):17-22.

Table 1.
An Example From MR Comparing the Three Regression Models

	Coefficient			Standard Error			t-value		
Model:	FS	G	$\overline{IS}$	FS	G	$\overline{IS}$	FS	G	$\overline{IS}$
IS	.88	-.12	-	.062	.062	-	14.20	-2.02	-
x_{21}	-.10	-.10	-.14	.057	.057	.053	-1.74	-1.74	-2.64
x_{22}	-1.05	-1.05	-.56	.640	.640	.602	-1.64	-1.64	-.93
x_{23}	1.57	1.57	1.23	.269	.269	.211	5.83	5.83	5.82
x_{24}	1.22	1.22	1.07	.460	.460	.457	2.65	2.65	2.34
x_{25}	-2.33	-2.33	-2.328	1.189	1.189	1.200	-1.96	-1.96	-1.94

Model:	R	$s_{y \cdot x}$	edf	SS(TOTAL)	RSS	F for testing R
FS	.932	11.16	205	193078	25523	225.9
G	.528	11.16	205	35374	25523	13.2
$\overline{IS}$	.514	11.24	206	35374	26032	14.8

Table 2.
Comparison of Alternative Models for
Assessing Differences in Treatment Gains for Angus Steers

Model	edf	R^2	$s_{y \cdot x}$	b_{IW}, s_b, t^b	b_t, s_b, t^c	F^d
ANOVA of Gain	28	.71	55.39	-	-	22.4
RM ANOVA	28	-	-	-	-	22.4
ANCOVA [G]	27	.72	55.40	.18, .18, 1.0	-	21.1
ANCOVA [FW]	27	.82	55.40	1.18, .18, 6.4	-	21.1
REGR [G]	29	.52	69.28	.28, .22, 1.3	-59, 11, -5.4	28.7
REGR [FW]	29	.70	69.28	1.28, .22, 5.8	-59, 11, -5.4	28.7

[b]The three numbers, b_{IW}, s_b, t, represent the regression coefficient for IW, the standard error and the t value.

[c]The three numbers, b_t, s_b, t, represent the regression coefficient for the treatment variable, the standard error, and the t value.

[d]The value of F given in the table represents the F value associated with the main (treatment) hypothesis of interest. For the ANOVA of Gain, the hypothesis is the equality of mean treatment gains; for the RM ANOVA, it is the time by treatment interaction; for the ANCOVA models, it is the equality of the adjusted treatment means; for the regression models, the hypothesis is testing for significance of the treatment partial regression coefficient.

INTERPOLATING, EXTRAPOLATING, AND FORECASTING QUALITATIVE ATTRIBUTES BY LOG-LINEAR MODELS

Clifford C. Clogg, The Pennsylvania State University

ABSTRACT

Log-linear models provide methods for the time-trending of qualitative attributes on the basis of information contained in crosstabulations obtained at successive points in time. Simple time-trend models are presented whereby the logits ϕ_{it} of some polytomous variable with $i = 1,\ldots,I$ classes are linked to time scores $t = 1,\ldots,T$ by means of a polymonial equation $\phi_{it} = A_i + B_{i1}t + \ldots + B_{i(T-1)}\, t^{T-1}$ of order T-1 or less. Once a suitable model is found, the predicted logit for any time t* is obtained by substituting t* for t in the final model. Formulae for the standard error of interpolated or extrapolated logits (or proportions) are developed such that the variance of a prediction depends upon the distance of that prediction from the mean of the original time scores. These models, except in some special cases, require use of a Newton-Raphson type algorithm for maximum likelihood estimation. Examples of varying complexity show the utility of these methods.

Keywords: Log-linear models; Newton-Raphson algorithm.

1. Introduction

In this paper log-linear models are used for the time-series analysis of qualitative attributes. Attribute data observed over time are common in social research, perhaps the most common example being the repeated cross-sectional survey. Time-series of qualitative attributes form much of the empirical base for the study of "social indicators" [13, 15], and so we can expect this type of data to become even more common in the future. The reasons for analyzing such data are evidently two-fold. The first objective is usually to parameterize the time-trend actually exhibited over the interval spanned by observed data. Statistical models are necessary for this task, since our understanding of the past is usually conditional on sample (rather than population) characteristics. Hopefully, an economical interpretation of the past will emerge from the analysis of only a few parameters of well chosen models.

A second objective of time series analysis, closely related to the first, is the forecasting or "prediction" of the future. Our expectations for the future are often conditioned upon information about the past. Social and economic policy, by its very nature oriented to the future, is most wisely formulated when explicit forecasts (based on past experience) are readily at hand. Time series methods which satisfy these needs are not to our knowledge now available. For attribute or frequency data the more usual time-series methods [e.g., 4, 14] are not strictly appropriate. These other methods can, however, motivate the corresponding methods suited for attribute data. The methods which we propose are applied to labor force data from the Current Population Survey of the United States [5].

2. Time-Trend Models Using Polynomial Equations for the Logits

Let us begin by describing the data which will be analyzed here. Table 1 classifies the civilian population of the United States aged 14 and over into four mutually exclusive and exhaustive categories based upon sample data from the March Current Population Survey for years 1969-1973. These data are easily seen to comprise a 4 x 5 crosstable. The comumn variable will be denoted by a "T," referring explicitly to the Time variable, and has categories $t = 1,\ldots,5$. The scores attaching to the categories of the T variable could just as well be rearranged to -2, -1, 0, +1, +2, but in either alternative the ordered and interval nature of the T variable is to be taken into account.

For simplicity we shall regard each of the five time-period observations as simple random samples (i.e., multinomial samples), but actually these data derive from a very complicated sampling scheme. The methods presented here can be extended to deal with sampling arrangements different from simple random samples. The reader is referred to Haberman [10, 11] for the proper extensions to some situations of possible interest. Note should be taken in Table 1 of the marginals of the T variable, $f_{.t}$, since these are fixed by sampling design. Acceptable models for these data (i.e., models which will generate the frequencies in Table 1) will need to fit this marginal in order not to violate the sampling design.

The attribute under investigation here pertains to the labor force status of persons. These statuses (i.e., categories) will be unfamiliar to some, so we make brief comment about them here. We let U refer to the labor force status variable in the row of Table 1, and we denote its classes by $i = 1,\ldots,4$. Category 1 refers to "adequate employment," and was actually defined as a residual category left over after the measurement of categories 2, 3, and 4. Category 4 refers to "economic inactivity," and comprises all persons who are not seeking out work at the time of the survey. In most respects, this status is similar to the "not-in-labor-force" category widely used in federal statistics. Category 3 denotes a status which we shall refer to as "economic underemployment," and comprises all persons who are unemployed, part-time unemployed, or working full time but receiving sub-standard wages. Category 2 refers to persons whose work wages are satisfactory, but whose skill level (measured by years completed education) is considerably greater than the mean skill level (educational level) of other workers in similar occupations. The labor force statuses contained in Table 1 are not now part of the federal government's system of labor force statistics; even the nomenclature chosen to

describe the categories is different from that of customary labor force reports. A detailed justification for this scheme of measurement is presented in [5]. It will suffice here to note that Table 1 is an example of time series attribute data, and data of this kind appear often in social research. Even though the column variable T (perhaps specified as an "independent" variable) is quantitative, the row variable U (a "dependent" variable) is qualitative.

To begin a time series analysis of Table 1 we first consider a simplified table derived by combining the 1st and 4th categories of U and the 2nd and 3rd categories of U. The result is Table 2 where now the dichotomous row variable U has category 1 denoting "not underemployed" and category 2 denoting "underemployed."

The usual time series models begin with a quantitative variable y, scores for which are observed at t = 1,...,T points in time. The across time variation in y is then "explained" by certain kinds of linear models [e.g., 14]. If $\underline{y}_T$ represents the vector of observations, the standard approach is to consider a model

$$y_t = f(\underline{y}_T) + e_t, \quad t = 1,\ldots,T,$$

where the e_t are assumed to be normally distributed error terms with constant variance and zero autocorrelation. The functional form f in those applications with which we are familiar is a linear function chosen in such a way to ensure that the e_t have regular properties. E.g., autoregressive-moving average models (ARMA models) reduce to certain variations on the linear model shown above. When a suitable function f can be found to purge the error term of undesirable properties, the forecast of y into the future for any t' > T is given as the projection along the trend curve fit to the original observations. Of course, we could also use the estimated function f to provide interpolated values of y for points t' < T, if there were sufficient reason to believe that f could be used to predict the trend in y for all points interior to the T points actually observed in the data. The forecasted score (or the interpolated score) $y_{t'}$ will represent an "optimal" prediction to the extent to which the chosen function f has ensured regular properties to the disturbances, and to the extent to which time-trend observed in the past can serve as a prediction of scores which are not yet known.

One kind of time series model appropriate for the attribute data in Table 2 is based upon a trending of logits. Let the observed frequencies in Table 2 be denoted as f_{it} and the expected frequencies given some model as F_{it}, i = 1, 2; t = 1,...,T. First consider a model for the U x T cross-classification whereby the expected logits $\phi_t = \log (F_{1t}/F_{2t})$ are related to the time scores t = 1,...,T by the following polynomial equation:

$$\phi_t = A + B_1 t + \ldots + B_{T-1} t^{T-1} = \underline{t}^{(T-1)'} \underline{B}, \qquad (2.1)$$

where $\underline{B}' = (A, B_1,\ldots,B_{T-1})$ and $\underline{t}^{(T-1)'} = (1, t,\ldots,t^{T-1})$. Equation (2.1) is a polynomial of degree T-1 linking the expected logits of U to the time scores, and it will be desirable to find models which fit the data and in which several of the B_i are zero. Models of this kind are considered by Bock [2, Ch. 8], Goodman [9], and Haberman [10, 11], but by considering the time series nature of (2.1) we shall obtain some new results. In this approach to the analysis of time series we have made weaker assumptions about the distribution of the dependent variable (i.e., it is binomial) and can appeal to maximum likelihood methods generally associated with log-linear models.

Given (2.1) above, a forecast for the logits of U for time points t' > T is straightforward. First we find a suitable representation of the time trend in our observed table. Suppose this model is

$$\phi_t = \underline{t}^{(p)'} \underline{B}^{(p)}, \qquad (2.2)$$

where $\underline{t}^{(p)}$, $\underline{B}^{(p)}$ are subsets of $\underline{t}^{(T-1)}$, $\underline{B}^{(T-1)}$, respectively. The predicted logit is then merely

$$\phi_{t'} = \underline{t}^{(p)'} \underline{b}^{(p)}, \quad t' = T + 1, T + 2,\ldots, \qquad (2.3)$$

where $\underline{t}^{(p)}$ is the same subset of $\underline{t}^{(T-1)}$ that appeared in (2.2) with the modification that t' replaces t. The vector $\underline{b}^{(p)}$ is the sample estimate of $\underline{B}^{(p)}$. The predicted proportions in the i-th category of U at t' are given by

$$P_{1t'} = \exp(\phi_{t'})/(1 + \exp(\phi_{t'}))$$
$$P_{2t'} = 1 - P_{it'}. \qquad (2.4)$$

If it were of interest to interpolate values of $\phi_{t'}$ for t' interior to the T sample points, then (2.3) and (2.4) are modified accordingly.

To estimate the model for the expected frequencies implied by (2.1), a model for the logits, several different strategies present themselves. For the column variable T a set of T-1 orthogonal polynomials are required to define the vector basis of variable T. Direct products of these with the simple deviation contrast vector (1/2, -1/2) for U define the appropriate interaction terms. For the case where the categories of T are equally spaced, standard computer programs such as the ECTA program of Goodman and Fay and the MULTIQUAL program of Bock and Yates [3] provide the necessary orthogonal polynomials. For cases where the number of time scores T is of moderate size, these may also be found in common statistical tables (e.g., [8]). For cases where the time scores are not equally spaced, the orthogonal polynomials can be obtained from formulae

reported by Bliss [1, pp. 2-27]. For the saturated model with zero degrees of freedom (where all of the B_i in (2.1) may be nonzero), the parameters can be calculated directly from formulae to be presented later. For the unsaturated model obtained by setting all of the B_i at zero, a model equivalent to the usual independence hypothesis for the two-way table, the constant A can also be estimated by elementary means. For various other models obtained from (2.1) by setting some (but not all) of the B_i at zero, the implied models for the frequencies are not equivalent to models based upon the fitting of marginals, and so computational methods for determining the F_{it} and the B_i different from the iterative proportional scaling algorithm have to be employed.

A Newton-Raphson algorithm can be used to find the maximum likelihood estimate of the F_{it} and B_i of (2.1). The approach suggests itself by considering the log-linear model for the frequencies implied by the linear model for the logits reported in (2.1). Letting $u = (\log F_{11}, \log F_{21}, \ldots, \log F_{2T})'$ we find that this model is

$$\underline{u} = X\underline{\beta} \qquad (2.5)$$

where $\underline{u}$ is 2T X 1, X is 2T X 2T (in the saturated model), and $\underline{\beta}$ is the 2T X 1 vector of coefficients. For various unsaturated models corresponding to (2.2), (2.5) will be modified by replacing the X matrix of contrasts by a corresponding 2T X (T + P) matrix of contrasts. The vector of logits $(\phi_1, \phi_2, \ldots, \phi_T)$ is obtained by premultiplying $\underline{u}$ in (2.5) by a matrix C with elements $C_{i,2i-1} = 1$, $C_{i,2i} = -1$, and all other $C_{ij} = 0$. That is,

$$\underline{\Phi} = C\underline{u}. \qquad (2.6)$$

From (2.1) we find that

$$\begin{aligned}\underline{\Phi} = {} & A + B_1 t_1 + \ldots + B_{T-1} t_1^{T-1} \\ & A + B_1 t_2 + \ldots + B_{T-1} t_2^{T-1} \\ & \vdots \\ & A + B_1 t_T + \ldots + B_{T-1} t_T^{T-1} \\ = {} & Z\underline{B}, \end{aligned} \qquad (2.7)$$

implying that $\underline{B}$ in (2.1) is given simply by

$$\begin{aligned}\underline{B} &= Z^{-1} C \underline{u} \\ &= Z^{-1} C X \underline{\beta}. \end{aligned} \qquad (2.8)$$

For unsaturated models corresponding to (2.2) Z will be of order T X p, but (2.8) will nonetheless provide the maximum likelihood estimate of $\underline{B}^{(p)}$ if $\underline{u}$ is a vector of maximum likehihood estimates. Equation (2.8) makes explicit some of the formulae which appear in Haberman [11], and shows how the coefficients in (2.1) can be estimated from computer output (e.g., MULTIQUAL output) providing $X\,\hat{\underline{\beta}}$.

The variance of a predicted logit $\phi_{t'}$ is easily seen to be

$$\text{Var}\,(\phi_{t'}) = \underline{t}^{(p)\prime}\ \text{Var}\,(\underline{b}^{(p)})\ \underline{t}^{(p)}, \qquad (2.9)$$

a formula familiar from regression analysis. Note that in (2.9) the $\underline{t}$ vector is composed of powers of powers of t', regardless of the value of t' (i.e., regardless of whether t' is an observed time score or an unobserved time score). Furthermore, the formula in (2.9) allows the variance of the predicted logit to depend on the distance of the prediction from the mean of the observed time scores, unlike some other asymptomatic variance formulae which might be used here. From (2.8) we have $\underline{b}^{(p)} = Z^{-1} C X \hat{\underline{\beta}}$ where the Z and X matrices are defined appropriately, and so

$$\begin{aligned}\text{Var}\,(\underline{b}^{(p)}) = {} & Z^{-1} C X\,(\text{Var}\,(\hat{\underline{\beta}})) \\ & X' C' Z'^{-1}. \end{aligned} \qquad (2.10)$$

As shown in [10, 11], $\text{Var}\,(\hat{\beta}) = (X' D(F) X)^{-1}$ where D(F) is the diagonal matrix with expected frequencies on the diagonal. Finally, the variance of predicted proportions in (2.4) can be approximated by application of the delta method. This shows how the polynomial time-trend model may be estimated and how the precision of a forecast can be obtained from it.

In sum, the approach to the time series analysis of qualitative attributes suggested here seems well suited to the interpolation of logits (or proportions) between time points actually sampled, and to the extrapolation or forecasting of logits (or proportions) into the future. While this approach has not to our knowledge been previously applied, there is little that is new in the log-linear time trend models suggested here.

As a first example consider the data in Table 2 where the 1973 sample is ignored. We consider the problem of forecasting the distribution in 1973 from the time trend 1969-1972. For simplicity, we assign scores -1.5, -.5, +.5, +1.5 to the four time-periods included. This choice of time scores only affects the value of the constant term A in (2.1). In Table 3 the degrees of freedom and the fit of various models are presented. The model H_0 where $\hat{\phi}_t = a$, equivalent to an hypothesis of independence between U and T, produces a likelihood-ratio Chi-square of 491.79 on 3 df, contradicting this simplest time-trend hypothesis. Introducing a linear term produces model H_1 where $\hat{\phi}_t = a + b_1 t$. With $L^2(H_1)$ of 13.69 on 2 df we have achieved a remarkable improvement in fit with addition of

only a single parameter. On such a large sample size as this (total n over 400,000), such a fit is certainly acceptable, even though the descriptive level of significance is approximately .001. We find by application of formulae presented earlier that a = 1.7344 and b_1 = -.0844, the latter term reflecting the decrease in economic opportunity 1969-1972.

We find when using H_1 and substituting the value t' = 2.5 (corresponding to 1973) in the equation ϕ_t = 1.7344 -.0844t that $\hat{\phi}_{1973}$ = 1.5264, implying a predicted proportion underemployed in 1973 of .1785. The observed logit and the observed proportion underemployed in 1973 were 1.5270 and .1642, respectively. (See Table 4.) We see that by virtue of the upturn in the economy during 1973 we have overestimated the number of underemployed persons by 1.42%.

Model H_2 in Table 3 corresponds to ϕ_t = $a + b_1t + b_2t^2$, and we see that this model does not significantly reduce Chi-square. Model H_3 corresponds to a linear and a cubic (but not a quadratic) term in the model. With $L^2(H_3)$ = .84 on 1 df we see that this model fits the data very well indeed. For H_3 we find a 1.7377, b_1 = -.1325, and b_3 = .0235. The predicted logit for 1973 is 1.5974, considerably worse than our first prediction. For these data the standard error of the forecasted proportion underemployed in 1973 would be virtually nil. Given the time-trend 1969-1972, the upturn in the economy during 1973 was totally unexpected.

Table 5 presents log-linear time trend models for the full 2 x 5 crosstable in Table 2. The Chi-square of 69.90 for the model with a linear time-trend parameter would be acceptable for most purposes. We see that addition of a quadratic term adds substantially, however, to the goodness-of-fit.

We now consider models for the 4 x 5 crosstable presented earlier in Table 1. Models for this table are generalizations of the one considered in (2.1), taking account of the polytomous U (dependent) variable. Models of the form

$$\phi_{it} = \log (F_{it}/F_{4t}) = A_i + B_{i1}t + \ldots + B_{i4}t^4, \; i = 1, 2, 3, \qquad (2.11)$$

are appropriate when U is unordered. To estimate models of the kind in (2.11) we generate the matrix X of contrasts in (2.5) by again using orthogonal polynomials for the T variable and using deviation contrasts implied by (2.11) for the U variable (see [3]). By following a hierarchy principle we might focus upon a subset of the wide range of models open to our choice where if B_{ik} = 0, then $B_{ik'}$ = 0 for k' > k, i = 1, 2, 3. The fit of some of these models is presented in Table 6. By restricting our attention to models of the kind in (2.11), interpolation or extrapolation is also straightforward, and can be carried out with the aid of the formulae presented earlier. We see from Table 6 that the model with only the linear terms B_{11}, B_{21}, B_{31}, is adequate (accounting for 79% of the variation in the data), but also that the inclusion of quadratic terms contributes in a substantial way to explaining time trend.

3. A Model Allowing Autocorrelation of the Logits

The models considered in the previous section linked the observed logits (or predicted logits) to scores reflecting the spacing of the time variable. For purposes of interpolation those models appear satisfactory. However, for purposes of forecasting (or extrapolation) beyond time points actually observed, the previous models can lead to unacceptable results. For example, in the analysis of the 2 x 4 crosstable presented in Tables 2 and 3, we found that $\hat{\phi}_t$ = 1.7374 -.0844t provided an acceptable summary of observed time trend. If we were to entertain this model seriously for purposes of forecasting, then the predicted logit for t' = 20.6 (=1.7374/.0844) would be zero, and the predicted distribution of the attribute would be a degenerate one. In this section we briefly consider a model which does not suffer this difficulty. This model is motivated by the simple autocorrelation model associated with the analysis of time series of quantitative variables [4], and suggests an alternative way of viewing time series attribute data.

A model where the expected logit at time t $\hat{\phi}_t$ depends only on the observed logit at time t-1, ϕ_{t-1} is the "first order autocorrelation of logits" model, viz.,

$$\hat{\phi}_t = \rho\phi_{t-1}, \qquad (3.1)$$

where ρ is the "autocorrelation" parameter. As in the usual time series approach to (3.1) (where the corresponding quantitative scores are substituted for the logits), the initial observation at t = 1 is considered as a given, and so we find that $\hat{\phi}_1 = \phi_1$, implying further that $f_{11} = F_{11}$, $f_{21} = F_{21}$. The model in (3.1) thus has some characteristics of a "quasi-independence" model, since the relation in (3.1) only pertains to a subset of the cells in the complete table. A least squares procedure, which in this case provides estimates almost equivalent to maximum likelihood, produced the results presented in Table 7. The model in (3.1) has an L^2 of 43.14 on two degrees of freedom, and provides an estimate of ρ of .9549. The predicted logit for 1973 is closer to the observed logit than was the case for the models considered in Section 2. (Cf. Table 4.) The advantage of model (3.1) is that forecasts of $\phi_{t'}$ for finite t' will result in nondegenerate predicted distri-

butions of the attribute. Because of this property these models deserve further consideration. Models of the kind in (3.1) can be modified to deal with certain other kinds of time series models (e.g., moving average models). We do not go into those details here. [Tables 5, 6 and 7 are available upon request from the writer]

REFERENCES

1. Bliss, C. I., Statistics in Biology, Vol. 2, New York: McGraw-Hill, 1970.

2. Bock, R. Darrell, Multivariate Statistical Methods in Behavioral Research, New York: McGraw-Hill, 1975.

3. Bock, R. Darrell, and Yates, George, MULTIQUAL: Log-Linear Analysis of Nominal or Ordinal Qualitative Data by the Method of Maximum Likelihood, Chicago: National Educational Resources, 1973.

4. Box, G. E. P. and Jenkins, G. M., Time Series Analysis, Forecasting and Control, San Francisco: Holden-Day, Inc., 1970.

5. Clogg, Clifford C., "Measuring Underemployment: Demographic Indicators for the U.S. Labor Force, 1979-1973," Ph.D. Dissertation, University of Chicago, 1977. Forthcoming by Academic Press.

6. Cochran, W. G. and Cox, G. M., Experimental Design, 2nd ed., New York: Wiley, 1957.

7. Cox, D. R., The Analysis of Binary Data, London: Methuen, 1970.

8. Fisher, R. A. and Yates F., Statistical Tables for Biological, Agricultural, and Medical Research, 6th ed., New York: Hafner, 1963.

9. Goodman, Leo A., "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," Technometrics 13, (1971):33-61.

10. Haberman, Shelby J., The Analysis of Frequency Data, Chicago: University of Chicago Press, 1974.

11. ------, "Log-Linear Models for Frequency Tables with Ordered Classifications," Biometrics 30 (1973):589-600.

12. Kendall, M. G. and Stuart A., The Advanced Theory of Statistics, 3rd ed., New York: Hafner, 1973.

13. Land, Kenneth and Spilerman, Seymour, eds., Social Indicator Models, New York: Russel Sage Foundation, 1975.

14. Nelson, Charles R., Applied Time Series Analysis for Managerial Forecasting, San Francisco: Holden-Day, 1973.

15. Stone, Richard, Demographic Accounting and Model Building, Paris: Organization for Economic Co-Operation and Development, 1971.

Table 1. Labor Force Status Over Time, Civilian Population Aged 14 and Over, 1969-1973.

Source: March Current Population Survey

	YEAR				
Labor Force Status	1969	1970	1971	1972	1973
1. Adequate Employment	48017 (44.2%)	45299 (43.6%)	44373 (41.8%)	42811 (41.7%)	42350 (42.1%)
2. Mismatch	5640 (5.2%)	5560 (5.4%)	6219 (5.9%)	6363 (6.2%)	6766 (6.7%)
3. Economic Underemployment	8971 (8.3%)	9184 (8.9%)	10571 (10.0%)	10592 (10.3%)	9748 (9.7%)
4. Not-in-Labor-Force	45887 (42.3%)	43705 (42.1%)	44956 (42.3%)	42939 (41.8%)	41685 (41.5%)
Total	108,515	103,748	106,119	102,705	100,549

Table 2. 2 X 5 Cross-Classification of Labor Force Status Over Time

Source: Table 1

	YEAR				
	1969	1970	1971	1972	(1973)
Not Underemployed [a/]	93904	89004	89329	85750	(84035)
Underemployed [b/]	14611	14744	16790	16955	(16514)
Total	108,515	103,748	106,119	102,705	(100,549)

[a/] Not Underemployed = Adequately Employed or Not-in-Labor-Force.

[b/] Underemployed = Mismatched or Economic Underemployed

Table 3. Log-Linear Time-Trend Models for the 2 X 4 Table (Ignoring 1973)

Model	Likelihood-Ratio Chi-Square	Goodness-of-Fit Chi-Square	Degrees of Freedom
H_o: $B_1 = B_2 = B_3 = 0$	491.79	491.12	3
H_1: $B_2 = B_3 = 0$	13.69	13.71	2
H_2: $B_3 = 0$	12.69	12.95	1
H_3: $B_2 = 0$	.48	.48	1

Table 4. Observed Logits Log (p_{1t}/p_{2t}) and Expected Logits Log $(\hat{p}_{1t}/\hat{p}_{2t})$ [a/] From Model H_1.

	1969	1970	1971	1972	(1973)
Observed	1.8605	1.7978	1.6715	1.6209	(1.6270) [b/]
Expected	1.8640	1.7796	1.6952	1.6108	(1.5264) [c/]

[a/] Expected logits obtained from $\Phi_t = 1.7374 - .0844t$.

[b/] Proportion underemployed in 1973 = .1642.

[c/] Predicted proportion underemployed in 1973 = .1785.

SOME ROBUSTNESS AND CONVERGENCE PROPERTIES OF THE KAPPA STATISTIC

Claude O. Archer, Brentwood Veterans Administration Hospital
and University of California, Los Angeles

Norman W. Reccius, Brentwood Veterans Administration Hospital

Introduction

During the last several years the statistic that has emerged as the dominant measure of agreement (as a form of reliability) for categorical data is the kappa statistic introduced by Cohen (1960). This special case of association uses the simple or observed proportion of agreement adjusted for occurrence by chance. Later, Cohen (1968) expanded the concept to include a weighted kappa. Others — Fleiss, Cohen, and Everitt (1969), Fleiss (1971), Fleiss and Cohen (1973), and Fleiss (1975) — have described some of the statistical properties of the kappa statistic, including exact and large sample standard errors, and equivalence to the intraclass correlation coefficient. More recently, Landis and Koch (1977a, 1977b) have expanded the concept of kappa-type statistics to a heirarchical variety to deal with the problem of agreement among multiple observers.

As a means of expanding our practical understanding of the kappa statistic beyond the indices of spread, and relation to correlation as mentioned above, we examine the variation of the kappa statistic as a function of the number of categories or scale steps that may be used in a study. This investigation covers four simple discrete distributions, and is carried out using proportion of agreement as the reference point. Knowledge is also developed to increase insight into the number of categories or scale steps that are mathematically optimal, while retaining consistency with earlier studies on reliability. For example, Nunnally (1967) stated that in terms of psychometric theory, the advantage is always with using more, rather than fewer, scale steps. The reliability of rating scales as a monotonically increasing function of the number of steps was further noted by Guilford (1954). Also, Garner (1960) reiterated essentially the same thing in relating the number of scale steps to the information or the amount of discrimination that was inherent in the scale. The comments of these authors were made with no mathematical justification. More recently, Green and Roe (1970) have taken a multidimensional-scaling approach to the problem, and Ramsey (1973) has investigated the precision of the estimation of scale values by using a maximum-likelihood approach while varying the number of categories and the amount of discrimination. Our study adds some mathematical justification to the literature for the agreement problem. It is limited to the case of unweighted kappa as first defined by Cohen (1960).

Discrete Distributions

When investigating the properties of a descriptive statistic, it is necessary to examine various distributions so that one sees the behavior of the statistic under a variety of conditions. This enables us to realize the scope of any inferences that we may make. The kappa statistic is a measure of agreement for categorical or nominal data first defined by Cohen (1960) as

$$\kappa = \frac{p_o - p_c}{1 - p_c} = 1 - \frac{\delta}{1 - p_c}$$

where

p_o = observed proportion of agreement
p_c = expected proportion of agreement

and

$\delta = 1 - p_o$.

Agreement is defined as identical categorization or rating by two individuals, which we visualize as the diagonal elements of a Person 1 by Person 2 categorization matrix. For our study, it is assumed that the observed row and column marginals of this matrix have independent identical distributions and hence, determine p_c. Under these assumptions the value of kappa is computed as a function of the number of categories for the four particular discrete distributions described below. The four distributions are described in terms of k successive proportions for the marginals.

1. Uniform
   1 : 1 : ··· : 1 (k times)
2. Triangular
   1 : 2 : ··· : k
3. Symmetric, Center Peak
   1 : 2 : ··· : (k+1)/2 : ··· : 2 : 1 ; k odd
   1 : 2 : ··· : (k/2) : (k/2) : ··· : 2 : 1 ; k even
4. Symmetric, Center Dip
   (k+1)/2 : ··· : 2 : 1 : 2 : ··· : (k+1)/2 ; k odd
   (k/2) : ··· : 2 : 1 : 1 : 2 : ··· : (k/2) ; k even

The coefficient kappa as a function of δ and k can now be computed for these four distributions. Since k = 1 yields the trivial case of complete agreement, we consider $k \geq 2$.

For the uniform distribution $p_c = 1/k$, hence kappa is

$$\kappa = 1 - \left[\frac{k}{k-1}\right]\delta, \quad k \geq 2$$

Moreover, note that if k is fixed, κ is a simple linear function of δ; also $\kappa \to 1 - \delta = p_o$ as $k \to \infty$.

For the triangular distribution, we sum from one to k as follows:

$$p_c = \frac{\sum j^2}{\left[\sum j\right]^2} = \frac{2}{3}\,\frac{(2k+1)}{k(k+1)}$$

Therefore

$$\kappa = 1 - \left[\frac{3k\,(k+1)}{3k^2 - k - 2}\right]\delta \; ; \; k \geq 2$$

As before, $\kappa \to 1 - \delta = p_o$ as $k \to \infty$.

The symmetric distribution with a central peak is considered next. In this case, we can take advantage of symmetry and sum from one to k/2, hence when k is even

$$p_c = \frac{2\sum j^2}{\left[2\sum j\right]^2} = \frac{4}{3}\,\frac{k+1}{k\,(k+2)}$$

and after some algebra,

$$\kappa = 1 - \left[\frac{3k^2 + 6k}{3k^2 + 2k - 4}\right]\delta.$$

When k is odd, symmetry can again be used; each summation occurs from one to (k-1)/2, and

$$p_c = \frac{2\sum j^2 + \left[\frac{k}{2}\right]^2}{\left[2\sum j + \frac{k}{2}\right]^2} = \frac{4}{3}\,\frac{(k^3 + 3k^2 - k)}{(k^2 + 2k - 1)^2}$$

hence,

$$\kappa = 1 - \left[\frac{3\,(k^2 + 2k - 1)^2}{3k^4 + 8k^3 - 6k^2 - 8k + 3}\right]\delta.$$

Continuing with the same methodology for the symmetric distribution with a central dip, the same result as for the symmetric distribution with a central peaking point is obtained when k is even. On the other hand, when k is odd, the summations are from one to (k-1)/2, so

$$p_c = \frac{2\sum j^2 - 1}{\left[2\sum j - 1\right]^2} = \frac{4}{3}\,\frac{(k^3 + 6k^2 + 11k - 6)}{(k^4 + 8k^3 + 14k^2 - 8k + 1)}$$

and

$$\kappa = 1 - \left[\frac{3\,(k^4 + 8k^3 + 14k^2 - 8k + 1)}{3k^4 + 20k^3 + 18k^2 - 68k + 27}\right]\delta.$$

Practical Implications

The formulas that were derived above examine the variability of the coefficient kappa as a function of the number of categories, k, for four discrete distributions. The practical implications of these results for psychosocial studies using a categorical data collection are related below.

The concept of reliability, and subsequently the more narrow concept of agreement, evolved out of a practical need for demonstrating how consistent a particular instrument was under varying conditions. The need had arisen out of the recognition that sources of error, such as the instrument being used, the variability of the persons doing the ratings, and the variability of the patients or things being rated were important considerations. The practical implication of this recognition has been to insist upon "high" reliability.

The literature has uniformly dealt with this problem in very loose terms. For example, it is generally felt that a reliability of .9 is great, .8 is good, and .5 is poor, but the means for more understanding is lacking. In this paper we hope to conceive a more solid, meaningful interpretation for the concept of agreement when the kappa statistic is used.

This is done by relating the coefficient kappa to the simpler concept of proportion of simple or observed agreement, that is, the number of times that two people agree out of the total number of possibilities of agreement and nonagreement. The comparisons are done for the aforementioned discrete distributions and values of k = 2, 3, 4, 5, 6, 8, 10, and 20.

Figure 1 shows these comparisons for the case

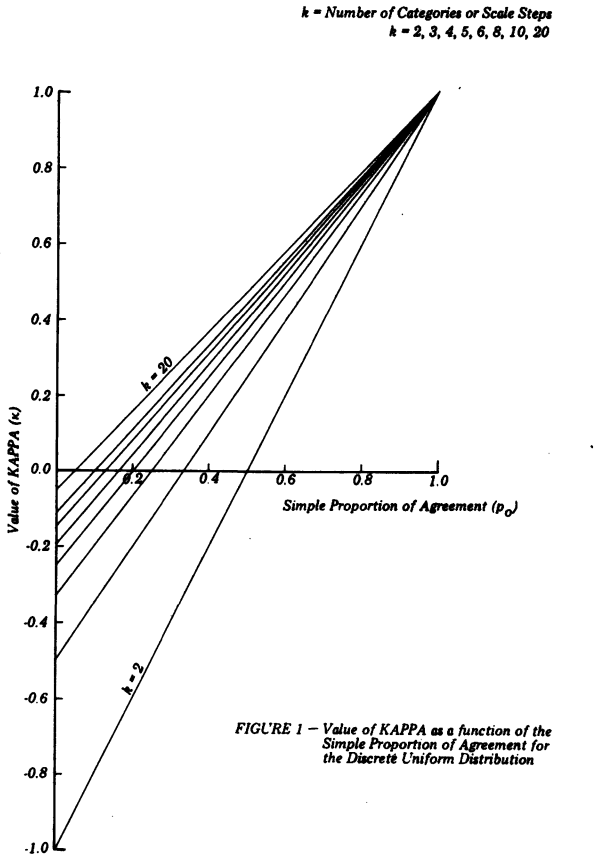


FIGURE 1 – Value of KAPPA as a function of the Simple Proportion of Agreement for the Discrete Uniform Distribution

of a uniform distribution. For example, when k = 2 we have a dichotomous distribution with a 50% chance of falling into each of the categories. Similarly, if k = 10 there is a 10% chance of falling into each of the categories. Studying Figure 1 more closely, assume an observed proportion of agreement of 50%. In other words, half of the time the two raters agree as to what they are rating or categorizing. Given a two-point dichotomous scale, kappa is zero, telling us that the agreement is exactly what is expected from pure chance. For a 10-point scale a kappa of approximately 0.45 is obtained; for a 20-point scale under the same situation, we get a kappa of approximately 0.48. Considering Figure 2, which is similar to Figure 1 except that the distributions assumed for the marginals are triangular,

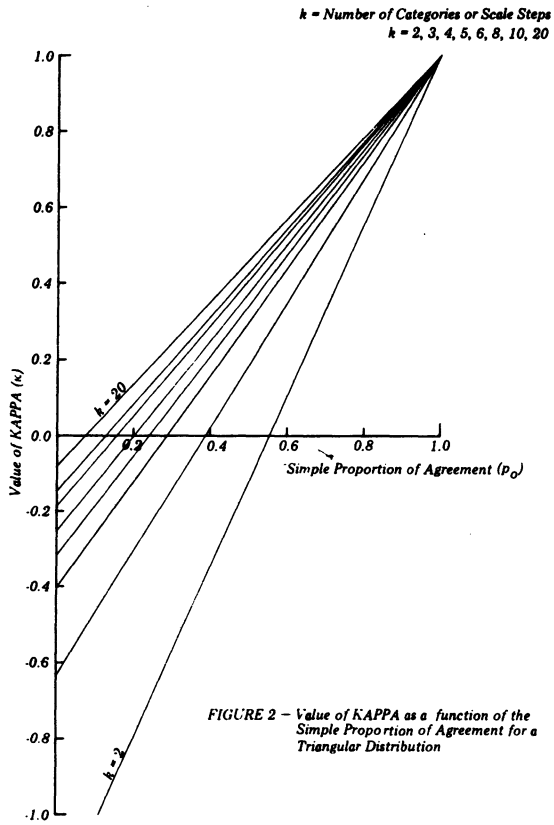


FIGURE 2 – Value of KAPPA as a function of the Simple Proportion of Agreement for a Triangular Distribution

TABLE I

A Partial Tabular Comparison of Kappa and the Simple Proportion of Agreement

Number of Categories k	Simple Agreement p_o	Kappa (κ) Uniform	Kappa (κ) Triangular
3	.5	.250	.182
5	.5	.375	.338
9	.5	.438	.418
20	.5	.474	.465
3	.7	.550	.510
5	.7	.625	.603
9	.7	.663	.651
20	.7	.684	.679
3	.9	.850	.836
5	.9	.875	.868
9	.9	.888	.884
20	.9	.895	.893

we note that when k = 2 and the proportion of simple agreement, p_o, is 0.5, kappa is equal to approximately -0.12. If k = 10 and p_o = 0.5, kappa = .43; if k = 20 we get an approximate value of 0.47 for kappa. For a simple agreement of about 0.9, and more than four categories kappa is between .86 and .90. These results from Figures 1 and 2 imply that the chance of getting a higher coefficient of agreement are better the more points or categories we have, even though the observed proportion of agreement is the same. (Note that we are not taking into account the ability of each person to place things equally well into 2, 6, 10, or 20 categories.) In addition, indications are that the more categories used, the closer the coefficient kappa is to the observed proportion of agreement, p_o.

Further illumination about what kappa means can be obtained by looking at some tabulations of k, p_o, and κ based upon our formulas (or Figures 1 and 2). For p_o = .5, .7, and .9, Table 1 illustrates that for a very good, highly reliable categorization scheme, the number of points does not matter nearly as much. Also, the magnitude of the difference between p_o and kappa is irrelevant for all practical purposes.

These two distributions, the uniform and triangular, have the widest disparity of the four discrete distributions considered, and since this disparity is not very broad the other two examples are not included in the illustrations.

Before we turn our attention to Figure 3, note that

$$\kappa = 1 - C_k \delta$$

where C_k is the quotient of two different polynomials in k for each of the discrete distributions introduced. Moreover, when $C_k = 1$, then $\kappa = 1 - \delta = p_o$, the simple proportion of agreement. Therefore, graphs of C_k as a function of the number of categories k and the discrete distribution considered are of interest. Figure 3 illustrates how rapidly C_k converges to one, and therefore how rapidly the kappa coefficient converges to the simple proportion of agreement. Only the two most dissimilar of the four discrete distributions are plotted here, since the other two distributions fell between these. The small differences between the curves give a strong indication of the robustness of kappa under the conditions considered. On Figure 3 note the very rapid change for small values of k up to 8 or 9, then a more gradual change to the end of the graph. Past k = 20, values of C_k for both distributions are very slowly asymptotic to one. Beyond k = 12, the practical difference of C_k and 1 is nil for all distributions considered. For example, for k = 12, simple agreement (p_o) on the order of .9 yields δ = .1 and kappa is about .89 for both the uniform and the triangular distributions; the only differences occurring in the third decimal place. The differences beyond k = 12 are even smaller. A further inference drawn from these results is that an optimal number of scale steps appears to be about eight or nine.

Conclusions

The agreement statistic kappa as a function of number of categories and the observed or simple proportion of agreement for the discrete

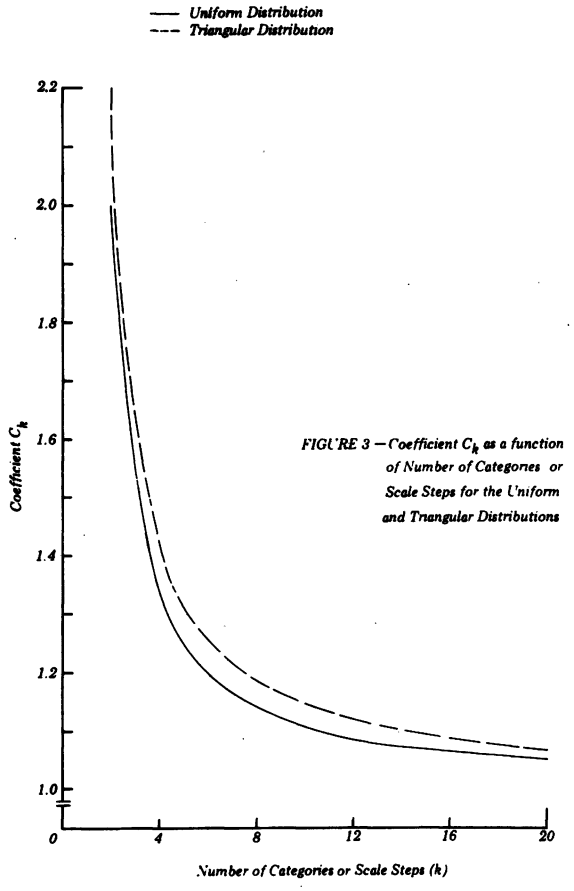


FIGURE 3 — Coefficient C_k as a function of Number of Categories or Scale Steps for the Uniform and Triangular Distributions

uniform, triangular, and symmetric with either center peak or center dip distributions has been studied. Findings indicate that for k moderately large (say $k \geq 8$), there is no practical difference between kappa and the simple or observed proportion of agreement. Also, for practical purposes, the differences between C_k for the distributions considered is negligible, indicating that kappa is a fairly robust indicator of agreement. We have also demonstrated empirically that $\kappa \to p_o$ monotonically as $k \to \infty$, hence a higher value of kappa is obtained with larger values of k, for a fixed amount of simple agreement.

References

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46, 1960.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70,* 213-220, 1968.

Fleiss, J.L. Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382, 1971.

Fleiss, J.L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659, 1975.

Fleiss, J.L. and Cohen, J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement, 33,* 613-619, 1973.

Fleiss, J.L., Cohen, J., and Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323-327, 1969.

Garner, W.R. Rating scales, descriminability and information transmission. *Psychological Review, 67,* 343-352, 1960.

Green, P.E. and Rao, V.R. Rating scales and information recovery — how many scales and response categories to use? *Journal of Marketing, 34,* 33-39, 1970.

Guilford, J.P. *Psychometric Models.* (New York: McGraw-Hill) 1954.

Landis, J.R. and Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174, 1977a.

Landis, J.R. and Koch, G.G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33,* 363-374, 1977b.

Nunnally, J.C. *Psychometric Theory.* (New York: McGraw-Hill) 1967.

Ramsay, J.O. The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrica, 38,* 513-532, 1973.

SOME COMPARISONS BETWEEN LEAST-SQUARES PREDICTION AND UNRESTRICTED RANDOM SAMPLING WHEN THERE ARE TWO CHARACTERISTICS TO BE ESTIMATED

G. W. Lynch, University of Ottawa

ABSTRACT

For sample sizes 4, 6, 8 and 12, Monte Carlo techniques are used to generate 2,000 random samples (without replacement) from a "real" finite population which has two auxiliary variables, x_1 and x_2, and two characteristics, $N\bar{Y}_1$ and $N\bar{Y}_2$, to be estimated. The mean square errors (mse) of the population total obtained by these methods are compared to those of the predictive sampling approach. The results indicate that the ratio estimator, under conventional unrestricted random sampling, yield mean square errors which are of the same order of magnitude (for each sample size) when x_1 and x_2 are used as auxiliary variables; and are decreasing with increasing sample size. Similar results are not obtained under least-square prediction. Additionally, regardless of sample size, unrestricted random sampling is more efficient than the corresponding extreme sample except when information from x_2 is used in the estimation of $N\bar{Y}_2$.

PURPOSIVE SAMPLING

Recently, Royall [1] has presented a methodology, based on least-squares prediction, of sampling from finite populations. The precise sampling scheme is to choose those n units whose x-values are largest (hence an "extreme" or "purposive" sample) and, for this sample, estimate the population total, $Y = N\bar{Y}$, by

$$N\hat{\bar{Y}} = [\sum_s y_j + \hat{\beta} \sum_{\bar{s}} x_j]$$

where the first sum is over the sample units, the second sum is over the units not in the sample,

$$\beta = [\sum_s (x_j y_j / v(x_j))] / [\sum_s (x_j^2 / v(x_j))] ,$$

and $v(x_j)$ is the variance of x_j.

When $v(x_j) \alpha x_j$, $\hat{\beta}$ is given by

$$\hat{\beta} = \bar{y}/\bar{x}$$

and $N\hat{\bar{Y}} = N \frac{\bar{y}}{\bar{x}} \bar{X}$.

Thus, in the precise situations for which the ratio estimator is optimal (see Cochran [2]), the classical ratio estimator and the estimator obtained from the predictive sampling approach are identical.

Note also that the "extreme" or "purposive" sample is one of the possible samples under unrestricted random sampling--but it is purposely, not randomly, chosen.

How does purposive sampling compare with unrestricted random sampling? On each of 16 natural populations where there was one characteristic to be estimated, Royall [1] compared the mean square errors obtained under each of the sampling procedures. His results suggest that the predictive sampling scheme generally produced smaller mse's.

It is our contention, however, that multipurpose surveys (rather than unipurpose surveys) are the usual practice. Thus, the natural question to ask is: How will the predictive sampling approach compare with the classical unrestrictive random sampling procedures when there is more than one characteristic to be estimated? To answer this question, we utilized an existing natural population for which there were two quantities to be estimated and computed the mse's under each of the sampling plans for each characteristic.

THE POPULATION

In a survey conducted in late 1973 (Lynch [3]), we had gathered information from the residents of King and Pierce Counties in the State of Washington. The information on all 350 sample units (N = 350) included: (a) the number of persons in each sample unit (x_1), (b) the number of households in each sample unit (x_2), (c) the number of females (18 years and older) who had ever had a pap smear (y_1), and (d) the number of females, 18 years and older, who had had a recent, 1972 or 1973, pap smear (y_2).

CHARACTERISTICS OF THE POPULATION

TABLE 1: *Population Means, Ranges, Standard Deviations and Coefficients of Variation*

	Variables			
	x_1	x_2	y_1	y_2
Mean	9.9	3.6	3.0	2.4
Range	25	8	8	7
St. Dev.	5.2	1.1	1.4	1.3
Coeff. of Var.	0.53	0.31	0.46	0.55

$C_{x_1 x_2} = 0.09$, $C_{y_1 y_2} = 0.20$

The population means, standard deviations, ranges and coefficients of variation are presented in Table 1. Here, it is evident that the coefficients of variation range from about 0.1 to approximately 0.55 and that the coefficients of variation of x_1, y_1 and y_2 are approximately equal, while that of x_2 is less.

TABLE 2: *Correlation Coefficients*

Variables	x_1	x_2	y_1	y_2
x_1	1	0.58	0.65	0.54
x_2	0.58	1	0.75	0.61
y_1	0.65	0.75	1	0.78
y_2	0.54	0.61	0.78	1

From Table 2 which presents the correlation coefficients, it is evident that all correlations are greater than 0.5. Also, scattergram plots of the data (not shown, in the interest of brevity) revealed that the intercepts were small. Thus, we have the conditions under which the ratio estimator is useful.

METHODS

Because of issues of bias and variability in small samples, it was decided to cover a range of small sample sizes--that is, n = 4, 6, 8 and 12. Due to limitations on available computer time and financial resources, it was immediately apparent that not all ${}_N C_n$ samples could be generated. Since the computer program, written by Dr. Kronmal [4] and later modified by the author, generated the ${}_N C_n$ samples in a random order, it was decided that a selection of 2,000 random samples for each sample size would provide the desired precision.

For the purposive sampling scheme, we use the n largest units of x_1 (Extreme-x_1) to compute the mse's for $N\bar{Y}_1$ and $N\bar{Y}_2$. This procedure was repeated for the n largest units of x_2.

RESULTS

The results shown in Table 3 indicate that, when information from either x_1 or x_2 is used in the estimation of $N\bar{Y}_1$ (see first four rows of Table 1), the purposive sampling plan yields larger mse's at all sample sizes. When x_1 was used in the estimation of $N\bar{Y}_2$, the univariate ratio estimator yielded the smallest mse's at all sample sizes; the reverse was true when x_2 was employed.

TABLE 3: *Comparison of Mean Square Error Results for the Extreme and the Univariate Ratio Estimators in a Multipurpose Survey*

Estimator	Mean Square Error, $N\bar{Y}_1$			
	n = 4	n = 6	n = 8	n = 12
Ratio -x_1	60874	37811	27560	16314
Extreme-x_1	104431	106697	94080	122769
Ratio -x_2	35664	17285	12249	8096
Extreme-x_2	31813	27360	19268	10266
	Mean Square Error, $N\bar{Y}_2$			
Ratio -x_1	57087	36039	27435	15681
Extreme-x_1	134260	121250	94638	101025
Ratio -x_2	35410	23068	16706	10635
Extreme-x_2	11519	7302	2288	3382

The classical univariate ratio estimator yields mse's which are of the same order of magnitude in the estimation of $N\bar{Y}_1$, and then $N\bar{Y}_2$. Similar results were not always obtained under the purposive sampling scheme (see Extreme-x_2).

Thirdly, under unrestricted random sampling, the ratio estimator yields mse's which are decreasing with increasing sample size. This does not appear to be evident under the purposive sampling plan.

CONCLUSIONS

Of course, one should be cautious about drawing general conclusions from the results of a single population. However, on the basis of estimating two characteristics from this population, the results would seem to suggest that the unrestricted random sampling plan has some desirable properties which are not evident under the purposive sampling scheme.

ACKNOWLEDGMENTS

I would like to express sincere thanks to Professor D. J. Thompson for his comments and suggestions. Much of this work was done while I was a Graduate Student at the University of Washington and supported by a National (Canada) Student Health Fellowship. Currently, the author is supported by the grant, RD10, of the Government of Ontario.

REFERENCES

1. Royall, R.M. (1970): "On finite population sampling under certain linear regression models." *Biometrics*, 57: 377-387.

2. Cochran, W.G. (1963): *Sampling Techniques. 2nd Edition.* John Wiley and Sons, Inc., New York.

3. Lynch, G.W. (1974): "An evaluation of the accuracy and efficiency of an area sample of King and Pierce Counties." Unpublished M.Sc. Thesis, University of Washington.

4. Kronmal, R. (1975): Personal Communication.

A MONTE CARLO ASSESSMENT OF THE STABILITY OF LOG-LINEAR ESTIMATES IN SMALL SAMPLES

Mark Evers, Duke University
N. Krishnan Namboodiri, University of North Carolina at Chapel Hill

Any reasonably complex contingency table will frequently contain empty or zero cells, merely due to sampling fluctuations. Of course, the number of zero cells is negatively related to sample size, and positively related to the number of cells in the contingency table. Thus, theoretically, zero cells can be "removed" either by obtaining a larger sample or by collapsing categories of some of the variables. However, the typical situation is one in which the investigator has only one sample of a given size, and in which collapsing the table is an unattractive alternative. Thus, we have the need for techniques to handle contingency tables with empty cells.

In a situation where there are only a few zero cells, Grizzle, Starmer, and Koch (1969) recommend inserting in each empty cell the value 1/r, where r is the number of response categories. For the iterative maximum likelihood procedure developed by Goodman and others, the following procedure has been suggested in the literature. For each model of interest, examine the marginals to be fitted, and discard all models that require fitting one or more zero marginals. This obviously is not a satisfactory strategy since investigators may wish to estimate parameters for models chosen on a priori grounds. Several options are open if the chosen model requires fitting empty marginals. (1) Use the technique advocated by Grizzle, Starmer and Koch, namely add the quantity 1/r to zero cells and analyze the data with their method. (2) Replace zero cells with small numbers, such as 1/r, and analyze the data using the iterative maximum likelihood technique. (3) Follow the strategy suggested in Bishop et al. (1975), chapter 12, which requires the assumption of a priori cell probabilities.

In this paper we examine strategy (2) in an effort to shed light on the resulting biases in parameter estimates. We refer to the small values added to observed zero cells as correction factors. We address the following questions, using the iterative maximum likelihood procedures as programmed in ECTA (Fay and Goodman 1973). First, does the size of the correction factor systematically affect the parameter estimates one obtains? Second, does the number of zero cells in the contingency table, which is closely related to sample size, influence the behavior of these estimates?

STUDY DESIGN

From the 1-in-100 Public Use Sample (PUS) of the 1970 U.S. Census, we first obtained data for about 219,000 women aged 14 to 44 years. From this data set, we created a four-way contingency table of children ever born by education by race by age. In this table, children ever born had four categories (0, 1, 2-4, 5+), education had three categories (less than 12 years, 12 years, more than 12 years), age had three categories (14-24, 25-34, 35-44), and race had two categories (white, nonwhite), thus giving a table with 72 cells. We specified a hierarchical model, which can be described in terms of the following three-way marginals to be fitted: children ever born by education by race, children ever born by age by race, and education by age by race. This model has 24 degrees of freedom and has a total of 48 independent parameters. In this paper, we examine only the 14 parameters which had the largest estimated values. Table 1 shows these parameter estimates, which we term the full sample estimates, since they are based on the full sample of women from the PUS.

From this full sample of women, we drew several sets of independent random samples: 100 samples of size 250, 100 samples of size 500, and 100 samples of size 1000. For each of these 300 subsamples, we constructed a contingency table with dimensions and categories identical to the table for the full sample of women described above. Every one of these contingency tables contained a number of empty cells, ranging from a minimum of 6 to a maximum of 37. For each of the subsample contingency tables, we used three different correction factors to replace the zero cells--.02, .2, and .5--and we used ECTA to obtain parameter estimates for the model that was fitted to the full sample of data. Thus, this design systematically varies sample size and correction factors, although the three different correction factors were applied to the same set of data. Because there are 100 samples in each set, we also have a reasonable amount of variation in the number of zero cells in each set.

The particular model we chose to fit to these sets of data did not fit well for the full sample of women. The chi-square value was 3304, which with 24 degrees of freedom has a proability of less than .001. It is therefore likely that whatever variation in the parameter estimates that we observe for the different subsamples may in fact be partly attributable to some unknown quantity of specification error. In order to deal with this problem, we took the expected cell counts based on the model fitted to the full sample, and simulated data which paralleled the same study design that was used for the data from the PUS. That is, we simulated 100 samples of each of the three sample sizes, and for each sample, applied each of the three correction factors, and used ECTA to obtain parameter estimates for our model.

RESULTS

Table 2 shows how the estimates of R, the main effect due to race, vary across sample size, correction factor, and number of zero cells in

the contingency table. Results are shown separately for the simulated data, and for the data drawn from the PUS. The bias of the estimates is calculated as the mean value of subsample estimates minus the full sample value. Thus, the value of .508 in the table (top of column 5) refers to the bias for the 61 samples of size 250 in the simulated set, which have between 19 and 28 zero cells, and which have the correction factor of .02 added to the zero cells. For this set of data, the bias is .508, indicating that the mean of the small sample parameter estimates was .508 higher than 1.108, the full sample value for R. The standard deviation for this group of 61 estimates is .273.

There are several patterns for both the bias and the standard deviation which deserve to be noted, since these are similar to the patterns for the other estimates we examined.

First, the correction factor is related to the bias in the following way: overall, the .02 correction tends to produce a positive bias, the .5 correction tends to produce a negative bias, and the .2 correction tends to produce the smallest bias, which hovers close to zero.

This finding makes sense, since, other things being equal, a large increment to zero cells would reduce the heterogeneity of a table and attenuate the value of an effect or a relationship. Hence, a large increment such as .5 would underestimate a positive effect and give a negative bias. On the other hand, a very small correction factor such as .02 clearly overestimates the effect, giving a positive bias.

The second observation about the pattern of bias in Table 2 is that for the .02 and .5 corrections, the amount of bias becomes smaller with increasing sample size. This apparent effect of sample size is most likely due to the number of zero cells in the table, which is strongly and negatively related to sample size. Other things being equal, an increase in the number of zero cells must be offset by larger entries in the remaining cells, giving a larger value for an effect or relationship. Since larger samples have fewer zero cells, we would expect the estimates to tend toward the full sample estimates. This effect is clear for the .02 correction factor, since the bias, or difference between the subsample estimates and the full sample estimate, becomes smaller with increasing sample size.

However, for the .5 correction factor, where the mean of the subsample estimates is consistently less than the full sample value, the bias becomes less negative with increasing sample size. Thus, the subsample estimates are getting larger with sample size, rather than smaller as we would predict by knowing the number of zero cells alone. We argue that the observed trend is due to the attenuating effect of the .5 correction factor. For larger samples, where there are fewer zero cells, there is less chance for this increment to attenuate the size of the estimates.

The third observation about the pattern of bias in Table 2 concerns the effect of zero cells, which we can detect by looking at the trend of bias within sample size. The bias generally becomes more positive as the number of zero cells increases, which gives support to the earlier argument that an increase in the number of zero cells will tend to increase the size of the estimate. Moreover, this effect of the number of zero cells is a good deal stronger for the .02 increment than for the .2 increment, and weakest of all for the .5 increment. It seems that the effect of the number of zero cells on the estimates simply cannot operate as strongly when the increment to these zero cells is larger, but the effect is very clear when the increment is close to zero.

The fourth observation about the pattern in Table 2 concerns the standard deviations of the estimates. We find a strong and negative effect of the size of the correction factor on the magnitude of the standard deviation. This finding is expected, since we know that larger corrections give the estimates more stability.

The last observation about the table is that we can find no major or systematic differences between the simulated data and the data from the Public Use Sample--specification error has no apparent effect on the patterns we observe.

The relation of the size of the estimates, sample size, and the correction factor, is shown in more detail in Table 3, for estimates of R based on the Public Use Sample. For this effect, the correction factor of .2 is likely to give the least bias for the two smaller sample sizes. Indeed, for sample size 250, the .02 and .5 increments do not approximate the full sample estimate of 1.108 for any of the 100 samples. Moreover, the .02 increment yields what most investigators would consider an unacceptably large amount of variation in the sampling distribution of the estimate. If one is willing to tolerate the slightly large standard deviation for the .2 correction factor, this correction yields estimates with relatively small bias, regardless of sample size. That the estimates based on a sample size of 250 can be this good is quite surprising, since the number of zero cells is so large, ranging from 19 to 37 out of a total of 72 cells in the table, and since the average frequency per cell is only 3.5.

Thus far we have considered only one parameter estimate out of the 14 we are examining here. One would naturally ask whether the findings just described can be generalized to the other effects, particularly where they are smaller in value than the estimate of R we have just discussed. The answer is that, generally, we find the same pattern for other effects. Evidence in support of this answer is found in Table 4, which shows the pattern of bias for four other parameters estimates, which differ markedly in size from one another. Close inspection of the table will show that the relationship of bias to sample

size follows the same pattern as we described earlier for the estimate of R. Regarding the correction factor, the value of .2 generally gives the least bias. In contrast to the finding for the estimate of R, this pattern holds even at the largest sample size.

In order to more systematically assess the apparent amount of bias that is linked with the three correction factors, we examined the relative amount of bias for each of the 14 parameter estimates we are considering. For each sample size and for both simulated and real data, we compared the amount of bias that resulted from using each of the three correction factors, and ranked the three factors as yielding high, medium, or low bias, for each estimate. The results of this tally are shown in Table 5. Across all sample sizes, the .2 correction factor consistently is the least likely of the three correction factors to give the highest amount of bias, and in all except the simulated data of sample size 250, the .2 correction factor is most likely to give the least bias. The .02 and .5 correction factors are both very likely to give estimates with a high degree of bias.

The results reported here, of course, concern only one method of dealing with zero cells in contingency tables. We are currently undertaking a Monte Carlo investigation to compare the bias of the estimates and validity of goodness-of-fit tests associated with the correction procedure described in this paper with those associated with the "pseudo-Bayes" procedures described in chapter 12 of Bishop et al. (1975).

REFERENCES

[1] Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. 1975. Discrete Multivariate Analysis: Theory and Practice. Cambridge, Mass.: MIT Press.

[2] Fay, R. and Leo A. Goodman. 1973. "Everyman's Contingency Table Analyzer (ECTA)." Department of Sociology, University of Chicago.

[3] Grizzle, James E., C. Frank Starmer, and Gary G. Koch. 1969. "Analysis of Categorical Data by Linear Models." Biometrics 25 (September): 489-504.

ACKNOWLEDGMENTS

We wish to thank Sharon Poss for extensive programming support, Karen Bothel, Richard Caston, Miltiades Damanakis, Dennis Gilligan, Eugenia Hatley, and Cathie Mayes Hudson for research assistance, and Valerie Hawkins for typing the manuscript. The research was supported by grant SOC76-02100 from the National Science Foundation.

TABLE 1

SELECTED PARAMETER ESTIMATES FOR MODEL FITTED TO CONTINGENCY TABLE BASED ON 1-in-100 PUBLIC USE SAMPLE

Description of Parameter		Full Sample Value (λ)
C_3:	C, third component	1.0214
A_1:	A, first component	-.5808
A_2:	A, second component	.3472
R	R	1.1083
E_2:	E, second component	.3613
E_1R:	E x R, first component	-.3136
C_1A_1:	C x A, first component	1.0385
C_2A_1:	C x A, second component	.9047
C_1A_2:	C x A, fourth component	-.3787
C_2A_2:	C x A, fifth component	-.2622
C_1E_1:	C x E, first component	-.4991
C_2E_1:	C x E, second component	-.2654
C_1A_1R:	C x A x R, first component	.3232
C_3A_1R:	C x A x R, third component	-.2105

Note: C = children ever born, A = age, E = education, R = race.

TABLE 2

BIAS AND STANDARD DEVIATION OF ESTIMATES OF R[a], BY CORRECTION FACTOR, SAMPLE SIZE, TYPE OF DATA, AND NUMBER OF ZERO CELLS

Sample Size	Type of Data	No. of 0 Cells	Correction Factor .02 Bias[b]	.02 Std. Dev.	.20 Bias	.20 Std. Dev.	.50 Bias	.50 Std. Dev.	(N)
250	Simulated	19-28	.508	.273	-.117	.086	-.325	.056	(61)
		29-31	.876	.179	.005	.060	-.261	.043	(29)
		32-37	.985	.167	.018	.080	-.260	.044	(10)
250	PUS	19-28	.596	.290	-.108	.056	-.316	.043	(14)
		29-31	.733	.192	-.051	.058	-.293	.040	(36)
		32-37	.980	.197	.021	.067	-.260	.041	(50)
500	Simulated	13-19	.248	.207	-.019	.081	-.145	.055	(50)
		20-21	.448	.199	.056	.070	-.108	.050	(24)
		22-27	.546	.185	.067	.058	-.117	.037	(26)
500	PUS	13-19	.268	.173	-.017	.065	-.150	.046	(26)
		20-21	.460	.231	.059	.088	-.103	.063	(30)
		22-27	.672	.236	.119	.060	-.079	.040	(44)
1000	Simulated	6-11	.055	.131	-.014	.071	-.061	.056	(47)
		12-13	.178	.114	.050	.056	-.025	.047	(29)
		14-19	.335	.172	.114	.081	.006	.056	(24)
1000	PUS	6-11	.074	.145	.002	.068	.050	.051	(29)
		12-13	.185	.129	.049	.068	-.022	.054	(34)
		14-19	.251	.145	.088	.060	-.003	.046	(37)

[a]Full sample value for R: $\lambda = 1.108$
[b]Bias = Mean value of subsample estimates minus full sample value.

TABLE 3

FREQUENCY DISTRIBUTION OF SUBSAMPLE ESTIMATES OF R[a], BY CORRECTION FACTOR AND SAMPLE SIZE, PUS DATA

Size of Estimate	Sample Size 250 .02	250 .20	250 .50	500 .02	500 .20	500 .50	1000 .02	1000 .20	1000 .50
0.60-0.79	0	0	29	0	0	0	0	0	0
0.80-0.99	0	19	71	0	2	45	5	3	8
1.00-1.19	0	76	0	5	59	55	30	65	91
1.20-1.39	2	5	0	19	39	0	39	32	1
1.40-1.59	8	0	0	31	0	0	25	0	0
1.60-1.79	17	0	0	20	0	0	1	0	0
1.80-1.99	24	0	0	15	0	0	0	0	0
2.00 +	49	0	0	10	0	0	0	0	0
Total	100	100	100	100	100	100	100	100	100
Mean	1.945	1.085	0.828	1.611	1.173	1.003	1.278	1.158	1.085
Bias[b]	.837	-.023	-.280	.503	.065	-.105	-.170	.050	-.023
Std. Dev.	.252	.078	.046	.273	.089	.057	.154	.073	.053

[a]Full sample value for R: $\lambda = 1.108$
[b]Bias = Mean value of subsample estimates minus full sample value.

TABLE 4

BIAS[a] FOR FOUR PARAMETER ESTIMATES OF DIFFERENT SIZE, BY SAMPLE SIZE, TYPE OF DATA, AND CORRECTION FACTOR

Sample Size	Type of Data	Correction Factor	Parameter C_3 (λ=1.021)	C_2A_1 (λ=.905)	C_1E_1 (λ=-.499)	E_1R (λ=-.314)
250	Simulated	.02	.244	.100	-.178	-.136
		.20	-.207	-.319	.044	.085
		.50	-.365	-.473	.138	.152
250	PUS	.02	.115	.155	.113	-.122
		.20	-.234	-.303	.097	.121
		.50	-.354	-.422	.119	.189
500	Simulated	.02	.254	.170	-.107	-.121
		.20	-.096	-.178	.010	.044
		.50	-.241	-.328	.073	.107
500	PUS	.02	.227	.323	.035	-.096
		.20	-.120	-.155	.046	.066
		.50	.253	-.340	.057	.129
1000	Simulated	.02	.161	.156	-.065	-.057
		.20	-.031	-.075	-.011	.017
		.50	-.129	-.193	.027	.057
1000	PUS	.02	.197	.164	-.011	.122
		.20	-.013	-.085	.028	.085
		.50	-.116	-.214	.053	.072

[a]Bias = Mean value of subsample estimates minus full sample value.

TABLE 5

FREQUENCY WITH WHICH DIFFERENT CORRECTION FACTORS RESULT IN HIGH, MEDIUM, OR LOW BIAS ACROSS 14 PARAMETER ESTIMATES

Type of Data	Correction Factor	Sample Size 250 Amount of Bias High	Medium	Low	500 Amount of Bias High	Medium	Low	1000 Amount of Bias High	Medium	Low
Simulated	.02	3	2	9	7	1	5	5	6	3
	.20	0	9	5	0	5	9	1	2	11
	.50	11	3	0	7	7	0	9[a]	4	1[b]
PUS	.02	4	4	6	5	5	4	8	3	3
	.20	0	6	8	0	6	8	0	6	8
	.50	10	4	0	9	4	1	6	5	3

[a]For one parameter, correction factors of .02 and .50 tied for "high" bias, .20 was assigned "low" bias.

[b]For one parameter, correction factors of .20 and .50 tied for "low" bias, .02 was assigned "high" bias.

REPLICATION OF SAMPLES FROM FINITE POPULATIONS

Barry L. Ford, USDA

1. Introduction

In the context of a simple random sample replication is randomly dividing a sample into groups so that each group is capable of estimating a population parameter. Replication has become an important strategy in sampling theory. Not only does replication simplify the calculations involved in a complex sampling scheme, but it also yields unbiased estimates of the variance of complex, nonlinear estimators.

When one has an infinite population or is sampling with replacement, a rationale for the number of replicates is given by Des Raj in his sampling text (2). The purpose of this paper is to extend the formulas to finite populations. Furthermore, it is demonstrated that one only needs a moderate population size in order to ignore the process of sampling without replacement and use the simpler formulas of sampling with replacement as an approximation.

When the sample design is a simple random sample of size n, the population total:

$$\tau = \sum_{i=1}^{N} x_i$$

is usually estimated by the sample statistic:

$$T_{simple} = \frac{N}{n} \sum_{i=1}^{n} x_i = N\bar{x}.$$

Another estimator of τ results from the technique of replication. If r replicates of size m are selected, replication yields the estimator;

$$T_{replicate} = \frac{\sum_{i=1}^{r} t_i}{r}$$

where

$$t_i = \frac{N}{m} \sum_{j=1}^{m} x_{ij}$$

Thus, the subscript of x refers to the j^{th} element in replicate i.

Obviously, the expected value of $T_{replicate}$ is:

$$E(T_{replicate}) = E(t)$$

$$= \frac{N}{m} \sum_{i=1}^{m} E(x)$$

$$= N \cdot E(x) \qquad (1.1)$$

$$= E(T_{simple}). \qquad (1.2)$$

When one selects the sample units *with replacement*, the replicates are independent and it is obvious that the variance of $T_{replicate}$ is:

$$Var(T_{replicate}) = \frac{M_2(t_i)}{r}$$

where $M_i(\cdot)$ represents the i^{th} central moment. Thus, when n = mr:

$$Var(T_{replicate}) = \frac{1}{r} M_2 \left(\frac{N}{m} \sum_{j=1}^{m} x_{ij}\right)$$

$$= \frac{N^2}{mr} M_2(x)$$

$$= \frac{N^2}{n} M_2(x) \qquad (1.3)$$

$$= Var(T_{simple}). \qquad (1.4)$$

If $u = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1}$, then one can estimate Var (T_{simple}) unbiasedly by:

$$Var(T_{simple}) = \frac{N^2}{n} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n - 1} \qquad (1.5)$$

$$= \frac{N^2}{n} u \qquad (1.6)$$

where $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$. Also, by allowing $u_t = \frac{\sum_{i=1}^{r} (t_i - \bar{t})^2}{r - 1}$, one can estimate Var $(T_{replicate})$ unbiasedly with:

$$var(T_{replicate}) = \frac{\sum_{i=1}^{r} (t_i - \bar{t})^2}{r(r - 1)} \qquad (1.7)$$

$$= \frac{u_t}{r} \qquad (1.8)$$

where $\bar{t} = \frac{\sum_{i=1}^{r} t_i}{r}$.

Up to this point there is no loss in efficiency by adopting a replicated design. However, *there is a loss of efficiency in replicated designs caused by a decrease in the precision of the variance estimate* (Raj, pg. 194). Remembering (1.6) one sees that the squared coefficient of variation of $var(T_{simple})$ is:

$$CV^2 \left[var(T_{simple})\right] = CV^2 (u) \qquad (1.9)$$

and

$$CV^2 \left[var(T_{replicate})\right] = CV^2 (u_t). \qquad (1.10)$$

Because the right side of expressions (1.9) and (1.10) are more easily written and comprehended, they are used in the following comparisons. It is well known (Raj, pg. 190) that:

$$CV^2 (u) = \frac{1}{n} \left\{\beta_x - \frac{n - 3}{n - 1}\right\} \qquad (1.11)$$

where $\beta = \frac{M_4(\cdot)}{M_2^2(\cdot)}$, the kurtosis of a distribution.

Thus,

$$CV^2(u_t) = \frac{1}{r}\left\{\beta_t - \frac{r-3}{r-1}\right\} \quad (1.12)$$

An easy calculation (Raj, 1964) yields:

$$\beta_t = \frac{1}{m}\left\{\beta_x + 3(m-1)\right\} \quad (1.13)$$

and therefore:

$$CV^2(u_t) = \frac{1}{n}\left\{\beta_x - 3 + \frac{2}{r-1}\right\}. \quad (1.14)$$

The result (Raj, pg. 195) is that:

$$CV^2(u_t) - CV^2(u) = \frac{2(n-r)}{(n-1)(r-1)} > 0. \quad (1.15)$$

For example, in a simple random sample of total size n = 100 with r = 10 the variance estimate using the replicates has a squared coefficient of variation which is approximately 0.20 greater than the squared coefficient of variation of the variance estimate, u. One can observe from (1.15) that r should be as large as possible.

2. The Stability of Variance Estimates When Sampling a Finite Population

Suppose a sample of size n is drawn *without replacement* from a population of size N. Then the most common estimator of τ:

$$T = N\bar{x}$$

remains of the same form as when sampling with replacement but the variance of T becomes:

$$Var(T) = \left(1 - \frac{n}{N}\right)\left(\frac{N}{N-1}\right)\frac{M_2(x)}{n}. \quad (2.1)$$

An unbiased estimator of $\frac{N}{N-1}M_2(x)$ (usually referred to by sampling theory texts as S^2) is:

$$u = \sum_{i=1}^{n}\frac{(x_i - \bar{x})^2}{n-1}. \quad (2.2)$$

One must derive the variance of u under the condition of sampling without replacement. The details are not given in this paper, but if they are requested will be furnished by the author. One can derive:

$$E(u^2) = \frac{1}{n}M_4(x) + \frac{1}{N-1}\left(\frac{n-1}{n}\right)\left(\frac{2}{(n-1)^2} + 1\right)\left\{NM_2^2(x) - M_4(x)\right\} + \frac{1}{N-1}\left(\frac{4}{n}\right)\left\{M_4(x)\right\} + \left(\frac{2(n-2)(n-3)}{(N-1)(N-2)(n)(n-1)}\right)\left\{NM_2^2(x) - 2M_4(x)\right\} + \left(\frac{3(n-2)(n-3)}{(N-1)(N-2)(N-3)(n)(n-1)}\right)\left\{NM_2^2(x) - 2M_4(x)\right\}. \quad (2.3)$$

The result, (2.3) can be found in a different form in a sampling text (Hansen, Hurwitz, Madow; page 101; Volume II).

By substracting the term $E(u)^2$ from both sides of (2.13) and remembering that:

$$E(u) = \frac{N}{N-1}M_2(x) \quad (2.4)$$

One finds:

$$Var(u) = M_4(x)\left\{\frac{N-n+2}{n(N-1)} \cdot \frac{2(n-2)(n-3)(2N-3)}{n(n-1)(N-1)(N-2)(N-3)} - \frac{2}{n(n-1)(N-1)}\right\} + NM_2^2(x)\left\{\frac{2}{n(n-1)(N-1)} + \frac{n-1}{n(N-1)} + \frac{(n-2)(n-3)(2N-3)}{n(n-1)(N-1)(N-2)(N-3)} - \frac{N}{(N-1)^2}\right\} \quad (2.5)$$

After a great deal of algebra (2.5) can be simplified into the form:

$$Var(u) = D_1 M_4(x) + D_2 M_2^2(x) \quad (2.6)$$

where:

$$D_1 = \frac{N(N-n)\left\{(N-1)(n-1) - 2\right\}}{n(n-1)(N-1)(N-2)(N-3)} \quad (2.7)$$

$$D_2 = \frac{N(N-n)(3N^2 - nN^2 - 6N + 3n + 3)}{n(n-1)(N-1)^2(N-2)(N-3)} \quad (2.8)$$

There are a few properties of D_1 and D_2 that should be pointed out.

Theorem 2.1: If n = N, then $D_1 = D_2 = 0$.

The proof is obvious.

Theorem 2.2: With simple random sampling from a a finite population:

$$\lim Var(u) = \frac{1}{n}M_4(x) - \frac{(n-3)}{(n-1)}M_2^2(x). \quad (2.9)$$

Again the proof is obvious. As expected, (2.9) is the variance of u when sampling with replacement (Raj, pg. 190) and will be used as a large size approximation to Var(u).

Theorem 2.3: If n strictly increases, the variance of u strictly decreases.

The proof is accomplished by showing that both D_1 and D_2 decrease as n increases. By rewriting (2.8):

$$D_2 = \frac{N(N-n)\left\{3N^2 - 6N - n(N^2-3) + 3\right\}}{n(n-1)(N-1)^2(N-2)(N-3)}$$

it is evident that as n increases the denominator increases and the numerator decreases if N > 3. (The restriction on N is inconsequestial because N must be greater than 3 to prevent division by zero.)

It is also true that as n increases, D_i decreases, but the proof required is more tedious. Suppose n increases by one then from (2.7)

$$(D_1|n) = \frac{N(N-n)\left\{(N-1)(n-1) - 2\right\}}{n(n-1)(N-1)(N-2)(N-3)} \quad (2.10)$$

$$(D_1|n+1) = \frac{N(N-n-1)\left\{(N-1)(n) - 2\right\}}{n(n+1)(N-1)(N-2)(N-3)} \quad (2.11)$$

Ignoring common factors, to prove $(D_1|n) > (D_1|n+1)$ one needs to show that:

$$\frac{(N-n)(nN-N-n-1)}{n-1} > \frac{(N-n-1)(nN-n-2)}{n+1} \quad (2.12)$$

Algebraically, (2.12) is equivalent to:

$$\frac{N^2n^2 - Nn - Nn^3 + n^3 + 2n^2 - N^2 - Nn^2 + n}{(n-1)(n+1)}$$

$$> \frac{N^2n^2 - Nn^2}{(n-1)(n+1)}$$

$$- \frac{Nn^3 + n^3 - 2n^2 - nN^2 + 2N - n - 2}{(n-1)(n+1)} \qquad (2.13)$$

After subtracting all terms on the left side of (2.12), one need only show for $n > 1$:

$$Q = N^2(n-1) - N(n+2) + 4n^2 + 2n + 2 > 0. \qquad (2.14)$$

When $n = 2$, (2.14) becomes $Q = N^2 - 4N + 22 > 0$ which is true for all N.

If n increases by one, then the change in Q, ΔQ, is:

$$\Delta Q = N^2 - N + 8n + 8 > 0$$

for $N > 0$, $n > 1$. Thus, one proves (2.14) which proves (2.12) which in turn proves that C_1 strictly decreases as n strictly increases given $N > 3$, $n > 1$. Therefore, one has the property that as n strictly increases, the variance of u strictly decreases.

Knowing the variance of u, formula (2.6) and remembering that $E(u) = \frac{N}{N-1} M_x(x)$; one finds:

$$CV^2(u) = C_1 \beta_x + C_2$$

where:

$$C_1 = \frac{(N-1)(N-n)\{(N-1)(n-1)-2\}}{n(n-1)N(N-2)(N-3)}$$

$$C_2 = \frac{(N-n)\{3N^2 - nN^2 - 6N + 3n + 3\}}{n(n-1)N(N-2)(N-3)}$$

It is easy to see that the limit as $N \to \infty$ of formula (2.15) is formula (1.11), the formula for with replacement sampling. Table 1 displays the values of N where the difference in these two formulas in less than 0.01. Thus, for population sizes larger than those in the table one can forget the condition of with replacement sampling and use the simpler formulas of without replacement sampling.

3. Determining the Number of Replicates

Now one should consider two situations that often arise in relicated sampling.

Case 1: *R relicates of size m are constructed (perhaps in a nonrandom manner) from a population. Assuming a without replacement structure within each replicate, how many replicates are needed to achieve a desired level of the coefficient of variation?*

A good example of this situation is where the population is ordered according to some arbitrary criteria and replicates are formed systematically. When replicates are not formed randomly, the obvious method of estimating any coefficient of variations is to consider each replicate as a sampling unit and to estimate the distribution of the replicates. Thus, to estimate the coefficient of variation of u_t (1.7 and 1.8), one uses (2.26) and substitutes the corresponding parameters from the population of replicates.

Case 2: *Suppose one must randomly select r replicates. Within each replicate units are chosen without replacement. However, each replicate may contain any unit in the population.*

One still uses the with replacement formula of Raj:

$$CV^2(u_t) = \frac{1}{r}\left\{\beta_t - \frac{r-3}{r-1}\right\} \qquad (3.1)$$

Now β_t is also subject to the laws of without replacement sampling. It is possible to derive β_t in terms of β_x. The derivation is again quite tedious, but details will be furnished by the author upon request.

One can derive:

$$M_4(t) = \frac{N^4}{m^4}\Bigg[m\,M_4(x) + \frac{3m(m-1)}{N-1}\left\{N M_2^2(x) - M_4(x)\right\} - 4\frac{m(m-1)}{(N-1)} M_4(x) + \frac{6m(m-1)(m-2)}{(N-1)(N-2)}\left\{2M_4(x) - NM_2^2(x)\right\} + \frac{3m(m-1)(m-2)(m-3)}{(N-1)(N-2)(N-3)}\left\{NM_2^2(x) - 2M_4(x)\right\}$$

Algebra yields the result:

$$M_4(t) = \frac{N^4(N-m)}{m^3(N-1)(N-2)(N-3)}\Big[3N(m-1)(N-m-1)M_2^2(x) + (N^2 - 6mN + 6m^2 + N)M_4(x)\Big] \qquad (3.2)$$

Thus, by dividing expression (3.2) by the square of $M_2(t)$ one finds that:

$$\beta_t = \frac{(N-1)}{m(N-m)(N-2)(N-3)}\Big[(N^2 - 6mN + 6m^2 + N)\beta_x + 3N(m-1)(N-m-1)\Big] \qquad (3.3)$$

One should note that when $m = 1$, $\beta_t - \beta_x$ and as N approaches ∞, expression (3.3) becomes (1.13). Table 2 shows those values of N such that the without replacement formula for β_t, (3.3), can be approximated by the with replacement formula, (1.13). These values of N are extremely low.

One should also note that in both Tables 1 and 2 the formulas are not monotonic functions of N but curves. When computer programs were written to compute the tables, this fact showed up as irregularities in the tables, However, corrections were made, and calculations were performed to insure that N was large enough to compensate for curves in the functions.

Table 1: Population sizes for which the coefficient of variation of the estimate of the population variance $CV(u_t)$, can be approximated by the formula $CV(u) = [\frac{1}{n}(\beta_x - \frac{n-3}{n-1})]^{\frac{1}{2}}$

β_x = Kurtosis

n = Sample Size	1.0	2.0	3.0	4.0	5.0	10.0	20.0	50.0	100.0
2	100	102	104	111	116	139	178	263	363
3	116	102	100	112	117	140	179	264	364
4	123	110	113	119	125	154	203	306	425
5	127	120	129	139	150	195	264	405	567
6	130	129	145	160	174	233	321	499	701
10	135	164	199	230	257	364	514	812	1146
15	138	200	255	300	339	490	699	1110	1570
25	140	258	342	408	465	681	978	1559	2210
50	145	368	500	603	691	1020	1470	2347	3328
100	147	528	725	877	1006	1488	2146	3426	4706
500	502	1260	1714	2065	2097	2657	3601	6237	9170
1000	1002	1468	1715	2066	2543	3543	5433	9398	13850

Table 2: Values of N, population size, for which the calculation of β_t (Kurtosis of replicates) differs by less than 0.1 between the with replacement and the without replacement formulas.

β_x = Kurtosis

m = Replicate Size	1.0	2.0	3.0	4.0	5.0	10.0	20.0	50.0	100.0
2	22	12	34	60	84	210	460	1210	2460
3	30	15	45	78	111	279	612	1611	3276
4	32	16	52	88	124	312	688	1812	3688
5	35	20	55	95	135	335	735	1935	3935
6	36	18	54	96	138	348	762	2016	4098
10	30	30	60	110	150	380	830	2180	4430
15	30	30	60	105	165	390	855	2265	4590
25	50	50	75	125	175	400	875	2325	4725
50	100	100	100	150	200	400	900	2400	4850
100	200	200	200	200	200	400	900	2400	4900
500	1000	1000	1000	1000	1000	1000	1000	2500	5000
1000	2000	2000	2000	2000	2000	2000	2000	2000	5000

Case 3: First one selects n units without replacement from the population. These n units are then randomly selected without replacement to form r replicates of size m.

Now one must use formula (2.16) to form:

$$CV^2(u_t) = \beta_t C_1 + C_2 \qquad (3.4)$$

where:

$$\beta_t = \frac{(N-1)}{m(N - m(N-2)(N-3)} \left[(N^2 - 6mN + 6m^2 + N)\beta_x + 3N(m-1)(N-m-1)\right] \qquad (3.5)$$

$$C_1 = \frac{(R-1)(R-r)(rR-R-r-1)}{r(r-1)R(R-2)(R-3)} \qquad (3.6)$$

$$C_2 = \frac{(R-r)(3R^2 - rR - 6R + 3r + 3)}{r(r-1)R(R-2)(R-3)} \qquad (3.7)$$

Question 1: If m is fixed, what should r be to insure a specific level, a, of $CV^2(u)$?

Because of theorem 2.3 it is possible to find the lowest value of r which satisfies the CV requirement by using a simple computer program. The computer program would use the method of bisection. It would:

1: solve equation (3.12) for r* = 2 (if $CV^2(u_t) \leq \alpha$ at r* = 2, then r* is the solution and the problem is solved)

2: solve for r** = A, where A is a large even number

3: solve (3.12) for $r' = \frac{r^{**} + r^*}{2}$ if r' is even and for $r' = \frac{r^{**} + r^*}{2} - 1$ otherwise

4: if $CV^2(u_t) \leq \alpha$ at r', then r** is set equal r' and return to step 3

5: if $CV^2(u_t) \geq \alpha$ at r, then r* is set equal to r' and return to step 3

6: continue until r** - r* = 2 and then set r' - r* + 1

7: if $CV^2(u_t) > \alpha$ at r', then r** is the solution; if $CV^2(u_t) \leq \alpha$ at r', then r' is the solution.

If N is large enough (see Table 1 and Table 2), one may use the simpler, with replacement formulas for β_t, C_1, and C_2.

If one can make the large population assumption, one has:

$$CV^2(u_t) = C_1 \beta_t + C_2 \qquad (3.8)$$

where:

$$C_1 = \frac{1}{r} \qquad (3.9)$$

$$C_2 = \frac{1}{r}\left(\frac{3 - r}{r - 1}\right) \qquad (3.10)$$

in place of equations (3.6) and (3.7) and:

$$\beta_t = \lambda_1 \beta_x + \lambda_2$$

where:

$$\lambda_1 = \frac{1}{m} \qquad (1.11)$$

$$\lambda_2 = \frac{3(m - 1)}{m} \qquad (1.12)$$

in place of equation (3.3). Thus, one can solve:

$$CV^2(u_t) = \frac{1}{r}\{(\lambda_1 \beta_x + \lambda_2) + \frac{3 - r}{r - 1}\}$$

for r by use of the quadratic formula:

$$r = \frac{1}{2a}[\alpha + \lambda_1 \beta_x + \lambda_2 - 1 \pm \{(\alpha + \lambda_1 \beta_x + \lambda_2 - 1)^2 - 4\alpha^2 (\lambda_1 \beta_x + \lambda_2 - 3)\}^{\frac{1}{2}}]$$

Example: Suppose $\beta_x = 17$ and one desires an α of 0.30 (i.e. $CV(u_t) \approx 0.55$).

Then:

$$\beta_t = \frac{17}{2} + \frac{3}{2} = 10$$

and

$$r = \frac{1}{0.60}[0.30 + 10 - 1 \pm \{(0.30 + 10 - 1)^2 - 4(0.30)^2 (7)\}^{\frac{1}{2}}]$$

$$r = \frac{1}{0.60}[9.3 \pm (86.49 - 2.52)^{\frac{1}{2}}]$$

$$r = 30.8 \quad \text{or} \quad r = 0.20.$$

Thus, one would select 31 replicates.

Question II: Suppose n is fixed, but m is not fixed. What combination of r and m is best?

When R is small, one can find a minimum r (thus, maximum m) by using the computer program outlined above and substituting n/r for m in equation (3.5). To get a maximum r (thus, minimum m) one should substitute n/m = r and proceed iteratively (beginning with m = 2) through the calculations.

Suppose from Tables 1 and 2 that large population approximations are appropriate. $CV^2(u_t)$ is maximized for fixed n when m = 2. One can see that:

$$CV^2(u_t) = \frac{1}{r}\left[\frac{\beta_x}{m} + \frac{3(m - 1)}{m}\right] + \frac{1}{r}\left(\frac{3 - r}{r - 1}\right)$$

$$= \frac{\beta_x}{n} + \frac{3(m - 1)}{n} + \frac{m}{n}\left(\frac{3 - \frac{n}{m}}{\frac{n}{m} - 1}\right)$$

$$= \frac{\beta_x}{n} + \frac{3(m - 1)}{n} + \frac{m(3m - n)}{n(n - m)} \quad .$$

It is obvious that $CV^2(u_t)$ will increase with an increase in m. Therefore, if n is fixed, m = 2 will yield the lowest $CV(u_t)$. If one has other restrictions on the size of m, one can proceed inductively with larger values of m until these restrictions are met or until $CV(u_t)$ exceed an acceptable level.

Question III: If m, r, and n are unkown what values should they have to attain a specific level, a, of $CV^2(u_t)$?

Certainly a minimum n is determined by a desired accuracy on the mean or total estimate. From this minimum n one can compute the calculations of $CV^2(u_t)$ for m = 2 (when using the with replacement formula). If $CV^2(u_t)$ is greater than the desired α, one can continue to n + 1 and so forth because m = 2 yields the minimum $CV^2(u_t)$ for a specific n. When a certain n satisfies the requirements, then one can proceed inductively on m.

When using the with replacement formulas such principles can not be applied because it can not be shown that m = 2 yields a minimum $CV^2(u_t)$ for a fixed m.

5. Conclusions

From Table 1 one recognized the fact that most large sample surveys which sample *without replacement* may use the *with replacement* formulas of Raj as a good approximation. When the population sizes are small enough to require the exact formulas presented here, one can estimate the size and number of replicates needed to stabilize the variance estimator. These two factors--size and number--are determined by a specific precision requirement on the estimated variance of a total. This paper only presents work on simple random samples.

6. Bibliography

1. Hansen, Morris H.; Hurwitz, William N.; and Madow, William G. Sample Survey Methods and Theory, New York, Wiley and Sons. 1953.

2. Raj, Des. Sampling Theory. New York, McGraw-Hill Book Company. 1968

PREDICTING THE POPULATION VALUE OF THE CANONICAL REDUNDANCY STATISTIC

Beth K. Dawson, Southern Illinois University School of Medicine

The basis for canonical correlation analysis was developed by Hotelling (1935, 1936). He defined "the most predictable criterion" as the linear combination of criterion variables that is predicted by a linear combination of predictor variables so that the two linear combinations have the highest possible correlation. When the influence of the first two linear combinations is partialled out, the process is repeated on the residuals, thus obtaining a sequence of pairs of variates with maximum correlations between them. These were denoted by Hotelling as canonical variates and canonical correlations, respectively.

Many authors subsequently noted that determination of the "most predictable criterion" is not always the appropriate goal of educational or psychological research. However, it was immediately appreciated that the technique developed by Hotelling provides a mechanism to study the number and nature of mutually independent relations between two sets of variables.

Darlington, Weinberg and Walberg (1973) described the manner in which canonical correlation analysis assists in researching such relationships. First, it determines the minimum number of traits needed to account for the important linear relationships between two batteries. For example, a researcher might hypothesize that there are r traits that describe the important relationships between a set of attitude variables and a set of performance variables. After performing a canonical correlation analysis, the number of significant relationships may be determined by testing whether the canonical correlation coefficient is greater than zero. Second, the standardized weights and factor structures assist in describing the nature of these traits.

An outline of the eigenanalysis procedure used in canonical correlation is presented to facilitate the following discussions.

Let $\underline{R}_{xx}$ be the matrix of correlations between the p predictor variables;
$\underline{R}_{yy}$ be the matrix of correlations between the q criterion variables;
$\underline{R}_{xy}$ be the matrix of correlations between the predictor and criterion variables.

The canonical correlation solution is obtained from eigenanalysis of the nonsymmetric matrix formed by the product $\underline{R}_{xx}^{-1}\underline{R}_{xy}\underline{R}_{yy}^{-1}\underline{R}_{yx}$.

Let $\underline{\Lambda}$ be the resulting diagonal matrix of eigenvalues with the ith diagonal element denoted λ_i^2;

$\underline{A}$ be the resulting matrix of eigenvectors for the predictor variables with the ith column denoted $\underline{a}_i$;

$\underline{B}$ be the resulting matrix of eigenvectors for the criterion variables resulting from eigenanalysis of $\underline{R}_{yy}^{-1}\underline{R}_{yx}\underline{R}_{xx}^{-1}\underline{R}_{xy}$ with the ith column denoted b_i.

Then $v_{xi} = \underline{a}_i'\underline{x} = a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{ip}x_p$ is the ith canonical variate for the predictor variables, i=1, ..., s, where 2=min(p,q);

$v_{yj} = \underline{b}_j'\underline{y} = b_{j1}y_1 + b_{j2}y_2 + \cdots + b_{jq}y_q$ is the jth canonical variate for the criterion variables, j=1, ..., 2.

The eigenvalues are the squared canonical correlation coefficients between each successive pair of canonical variates. The eigenvector matrices, $\underline{A}$ and $\underline{B}$, contain the standardized weights for the predictor and criterion variables, respectively, and are used to form the canonical variates. If eigenanalysis is performed using variance-covariance matrices in place of correlation matrices the eigenvalue matrix must be postmultiplied by the diagonal matrix of standard deviations of the appropriate set of variables, predictor or criterion, to obtain the standardized weights.

As with any statistical technique, researchers must have a mechanism to judge the statistical and practical significance of the results of canonical correlation analysis. The standardized weights, $\underline{A}$ and $\underline{B}$, and the factor structure correlations between each canonical variate and the original variables, $\underline{R}_{xx}\underline{A}$ and $\underline{R}_{yy}\underline{B}$, may be examined to interpret the relationship between canonical variates and the original measures. The squared correlation, λ^2, between each pair of canonical variates may be interpreted as the amount of variance shared by the two linear combinations of the predictor and criterion variables. However, these statistics fail to provide information regarding the amount of shared variation between the variables in the two batteries. Stewart and Love (1968) and Miller (1969), with the consultation of Paul Lohnes, developed a measure to permit this type of interpretation. Their statistic, denoted the bimultivariate redundancy index, or the canonical redundancy index in the special case of canonical correlation, has intuitive appeal. The canonical redundancy statistic is defined as the sum of successive products between the proportion of variance that the canonical variates of either battery explain in the canonical variates of the other (accounted for by the squared canonical correlations), and the proportion of variance absorbed by the canonical variates from their respective batteries (accounted for by the squared correlations between the variates and original variables). Using the above notation, the total canonical redundancy statistic for the predictor variables, given the criterion variables, is

$$Rd_x = \frac{1}{p}\Sigma_j^s (\underline{R}_{xx}\underline{a}_j)' (\underline{R}_{xx}\underline{a}_j) \lambda_j^2;$$

the total canonical redundancy statistic for the criterion variables, given the predictor variables, is similarly

$$Rd_y = \frac{1}{q} \Sigma_j^s (\underline{R}_{yy}\underline{b}_j)' (\underline{R}_{yy}\underline{b}_j) \lambda_j^2;$$

where $s = \min(p,q)$.

This sum determines the amount of redundancy in one battery of variables given the other. As such, it is directional and nonsymmetric and has a desirable range of zero to one.

Miller (1969) and Miller and Farr (1971) demonstrated the equivalence of the total redundancy measures based upon multiple regression of independently orthogonalized batteries, such as in principle component analysis, and the total redundancy measures based upon canonical correlation analysis, a simultaneous orthogonalization procedure. The two solutions differ, however, in the structural components of the batteries and, therefore, the redundancy measures for individual components or variates are not identical.

Gleason (1976) established the mathematical basis for a generalized version of the canonical redundancy statistic. One way of interpreting redundancy of one set of variables, given another set, is to reconstruct one set using only the information in the second set that is relevant to that in the first set. Gleason demonstrated that this approach leads to a mathematical expression that is equivalent to the canonical redundancy index as defined by Stewart and Love (1968), and thus provides the mathematical rigor for the definition of this measure.

The recent development of an index that describes the overlap or amount of redundancy between two batteries in canonical correlation analysis is extremely welcome. The need for such a measure is apparent, and researchers in education and psychology are generally eager to utilize measures that assist them in their studies. In the short period of time since the papers by Stewart and Love (1968) and Miller (1969) were published, the canonical redundancy statistic has been described and recommended in texts and articles by some of social sciences' leading authors. For examples, see Cooley and Lohnes (1971, p. 170-172; Tatsuoka (1973, p. 280-282); Cohen and Cohen (1975, p. 429-432); Timm (1975, p. 355-358) and Cooley and Lohnes (1976, p. 211-212). In addition, an entire session at the 1976 Annual Meeting of the American Education Research Association was devoted to applied research on the redundancy statistic.

It is almost a certainty that the redundancy statistic is positively biased. First, consider the dependence of this statistic on the squared canonical correlation coefficient. Second, the results by Miller (reported in Cooley and Lohnes, 1976) of a Monte Carlo analysis investigating the sampling distribution in the null case indicate bias of the median total redundancy statistic ranging from .06 to .09 for various combinations of numbers of predictor and criterion variables and sample sizes. No other data are available on the bias of this statistic. Thus the purpose of the present study was to investigate the empirical sampling distribution of the first squared canonical correlation coefficient and of the redundancy statistic using Monte Carlo methods; subsequently, an attempt was made to derive a formula to correct for bias in the total redundancy statistic.

The investigation entailed the systematic variation of the number of predictor and criterion variables, the sample size, the size of the intrabattery correlations and the size of the interbattery correlations since these are the six parameters that affect the magnitude of the canonical redundancy statistic.

Two levels of the number of variables were designated for each of the left and right sets: cases with five and nine variables for each set. Due to the symmetry of canonical correlation analysis, it was necessary to consider only the following three combinations:

$$\{(p,q)\} = \{(5,5), (9,5), (9,9)\}$$

Two sample sizes were selected for study, the case of a small sample and the case of a large sample. For small n, the value used was 5(p+q), a sample size frequently encountered in applied research. For the large sample, n equal to 20(p+q) was used to examine effects of a sample size frequently recommended. Thus the sample sizes were as follow:

$(p,q) = (5,5)$; $n_s = 50$, $n_\ell = 200$
$(p,q) = (9,5)$; $n_s = 70$, $n_\ell = 280$
$(p,q) = (9,9)$; $n_s = 90$, $n_\ell = 360$

Two conditions were chosen for the average intercorrelations of each of the matrices $\underline{P}_{xx}$ and $\underline{P}_{yy}$, and three conditions were selected for $\underline{P}_{xy}$. In applied research, variables in the predictor set often have medium to high correlations. Similarly for the intercorrelations between criterion variables. However, the correlations between predictor and criterion variables are quite often lower. In an attempt to reflect conditions often encountered in actual research, the off-diagonal elements of $\underline{P}_{xx}$ and $\underline{P}_{yy}$ were set to .30 and .60 to reflect medium and high correlations respectively. All of the elements of $\underline{P}_{xy}$ were set to .00, .20 and .40 to reflect the null case, low correlations and medium correlations, respectively. The inclusion of the null case for no relationship between the two sets of variables was important in this study to provide baseline information against which to compare bias in the non-null cases. The fact that $\underline{P}_{xy} = \underline{P}_{yx}'$ was utilized to form the $((p+q)x(p+q))$ supermatrix

$$\underline{P} = \begin{pmatrix} \underline{P}_{xx} & \underline{P}_{xy} \\ \underline{P}_{yx} & \underline{P}_{yy} \end{pmatrix}.$$

The above conditions lead to the definition of 36 population matrices. (Three combinations of numbers of predictor and criterion variables, 2 levels of $\underline{P}_{xx}$, 2 levels of $\underline{P}_{yy}$, and 3 levels of $\underline{P}_{xy}$.) Calculation of parameters and statistics based upon 2 sample sizes increases the number of specific situations under investigation to 72. The process used to define population conditions and generate sample matrices for the Monte Carlo

analysis is presented in schematic form in Figure 1.

The results of the Monte Carlo study show that considerable positive bias is obtained when a sample redundancy statistic is used to estimate the population value. In general, the amount of bias for the redundancy statistic defined on one battery tends to decrease as the number of variables increase in the second battery. Bias appears to be unaffected by the number of variables when this number is equal for both batteries. Bias of both the redundancy statistic and the largest squared canonical correlation coefficient is consistent for all levels of intrabattery correlation but decreases as interbattery correlations increase, indicating less bias in the non-null cases. The most dramatic parameter affective bias is, as might be expected, sample size. Bias increases approximately fourfold as the sample size increases by the same amount. It is not known whether this relationship is linear since a sufficient number of sample sizes were not considered in the present study.

It is useful in the case of a biased estimate to employ a formula that "corrects" the estimate and provides a value that is closer to the population parameter. The present study utilized two approaches to attempt to estimate the population value of the total redundancy statistic, given information about the sample. One approach applied two standard shrinkage formulae to the sample value; the other regressed the population value on sample information. The results of the regression analysis are presented first.

The intrabattery and interbattery correlations were recoded for purpose of the regression analysis. Two regression equations were calculated, the first using the following variables and values:

p: 5 or 9
q: 5 or 9
R_{xx}: 1 if R_{xx} = .30, 2 if R_{xx} = .60
R_{yy}: 1 if R_{yy} = .30, 2 if R_{yy} = .60
R_{xy}: 0 if R_{xy} = .00, 1 if R_{xy} = .20, 2 if R_{xy} = .40
n: 5(p+q) or 20(p+q).

The values for intrabattery and interbattery correlations were recoded as above to attempt development of a regression equation that would be more generalizable. Indeed, the matrix randomly generated from the population conditions did not have intra- and interbattery correlations precisely equal to the population values. The computer algorithm routine used (Montanelli, 1971) generates sample correlation matrices that would result from sampling random normal variables having the required population correlation structure. Thus the actual matrices used as the population had correlations that, upon repeated sampling, have expected values equal to the population values of .30, .60 for intrabattery correlations and .00, .20 and .40 for interbattery correlations, respectively.

The second regression analysis included the mean sample value of the redundancy statistic in addition to the predictor variables listed above. The 72 values of Rd_x and Rd_y were combined to give a sample of N=144 for the analysis. The regression analysis resulted in multiple correlation coefficients of R=.9925 and R=.7904 for the regression equations computed with and without the mean sample value, respectively. Beta weights were tested for significant contribution to the regression equation, and only those variables with corresponding beta weights significant at $\alpha \leq .01$ were retained. The significant beta weights were then rounded to three decimal places for all but one variable and a predicted value for each of the 144 sets of observations was computed. This value was then correlated with the population value to obtain an adjusted Multiple R.

The results of the regression analysis are contained in Table 1. The use of the reduced set of weights does not affect the Multiple R significantly. It is perhaps unfortunate that the population value is not better predicted without the sample mean. This may be an artifact of the restricted upper range of these statistics, although population values were not that large in the present study, the largest being 0.502. Another reason for the poor prediction without the sample mean is that the population value does not vary with the sample size while it is apparent from detailed Monte Carlo results that the degree of bias is greatly affected by this variable. Consequently, it is not surprising that sample size was not significant in the regression analysis omitting the sample mean. However, with the sample mean included in the set of predictors, the sample size is a significant predictor, as might be expected.

The results of the Monte Carlo Analyses indicate that the behavior of the bias of total redundancy statistic is quite similar to that of the squared canonical correlation coefficient. In addition, Miller (1975b, 1976) demonstrated that the redundancy statistics are approximated by an F distribution with modified degrees of freedom in the null case. It was therefore decided to apply the Wherry and the Olkin-Pratt (Kendall and Stuart, 1967) shrinkage formulae for the squared multiple correlation coefficient in the hopes that the population values of the redundancy statistics may be similarly estimated*. The significant increase in the Multiple R achieved when the sample mean value of the canonical redundancy statistic was included in the regression analysis was a further indication that investigation of these formulae, which utilize the sample value, might be worthwhile.

The formulae used were the following:

Let N be the sample size,
q be the number of variables in the criterion set,
Rd be the total redundancy statistic for the predictor set given the criterion set,

* The author would like to express her appreciation to John Pohlmann, Southern Illinois University, for his suggestion to examine the efficacy of these formulae.

Then

$$\text{Wherry correction} = 1 - \frac{N-1}{N-q-1}(1-Rd);$$

$$\text{Olkin-Pratt correction} = 1 - \frac{N-3}{N-q-1}(1-Rd) - \left(\frac{N-3}{N-q-1}\right)\left(\frac{2}{N-q+1}\right)(1-Rd)^2.$$

In the case of the total redundancy for the criterion set given the predictor set, the roles of p and q were exchanged to be consistent with the above.

Each formula was applied to the mean sample values of the total redundancy statistic and the difference between the population value and the shrunken estimate was calculated. In addition the mean absolute value of the difference was computed over all 144 values of Rd. Table 2 contains the results of this analysis.

Both formulae provide excellent approximations to the population value. The average absolute values of the difference between the population and the corrected value were 0.003347 for the Wherry formula and 0.001955 for the Olkin-Pratt formula. Although the Olkin-Pratt formula is better when considering only the residuals, Table 2 illustrates that this formula results in more overestimates of the population values. This is evidenced by the larger number (82) of negative differences as compared to only 12 overestimates using the Wherry formula. With the Wherry formula, only 3 of the 144 estimates vary from the population parameter by a value greater than 0.01, and these are all underestimating the parameter. Thus, while both formulae provide excellent corrections for the total redundancy statistic, the Wherry is recommended due to its tendency to provide a conservative estimate.

It would seem that the results of the Monte Carlo analysis do not justify the recommendation of the canonical redundancy statistic as an alternative to the squared canonical correlations on the basis of less bias. The redundancy statistic, in general, appears to exhibit a degree of bias quite similar to that of the squared canonical correlation. However, the bias is easily corrected by the Wherry or Olkin-Pratt formulae to estimate the true population value. What is perhaps more important are the interpretive characteristics of the canonical redundancy statistic as compared to those of the canonical correlation coefficient. If interest truly lies in the relationship between groups of variables as opposed to the relationship between linear combinations of variables, then the redundancy statistic provides a more realistic and meaningful measure for the conscientious education researcher.

References

Cohen, J. and Cohen, P. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: Wiley, 1975.

Cooley, W.W. and Lohnes, P.R. Multivariate Data Analysis. New York: Wiley, 1971.

Cooley, W.W. and Lohnes, P.R. Evaluation Research in Education. New York: Wiley, 1976.

Gleason, T.C. On redundancy in canonical analysis. Psychological Bulletin, 1976, 83: 1004-1006.

Hotelling, H. The most predictable criterion. Journal of Educational Psychology, 1935, 26: 139-142.

Hotelling, H. Relations between two sets of variates. Biometrika, 1936, 28: 321-377.

Kendall, M.G. and Stuart, A. The Advanced Theory of Statistics (2nd Ed.), Volume 2. New York: Harper, 1967.

Lohnes, P.R. and Cooley, W.W. Partitioning redundancy among predictor domains in reduced-rank models for school effects. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, 1976.

Miller, J.K. The development and application of bi-multivariate correlation. Unpublished doctoral dissertation, State University of New York at Buffalo, 1969.

Miller, J.K. The sampling distribution and a test for the significance of the bi-multivariate redundancy statistic: A Monte Carlo study. Multivariate Behavioral Research, 1975, 10: 233-244.

Miller, J.K. Testing hypotheses about the bi-multivariate redundancy statistic. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco, 1976.

Miller, J.K. and Farr, S.D. Bimultivariate redundancy: A comprehensive measure of inter-battery relationship. Multivariate Behavioral Research, 1971, 6: 313-324.

Montanelli, R.G., Jr. An investigation of the goodness of fit of the maximum likelihood estimation procedure in factor analysis. Unpublished doctoral dissertation, University of Illinois, Urbana, 1971.

Stewart, D. and Love, W. A general canonical correlation index. Psychological Bulletin, 1968, 70: 160-163.

Tatsuoka, M.M. Multivariate analysis in educational research. In F.N. Kerlinger (Ed.), Review of Research in Education. Itaska, IL: Peacock, 1973.

Timm, N.H. Multivariate Analysis with Applications in Education and Psychology. Monterey, CA: Brooks/Cole, 1975.

FIGURE 1: SCHEMATIC REPRESENTATION OF DESIGN OF MONTE CARLO ANALYSIS

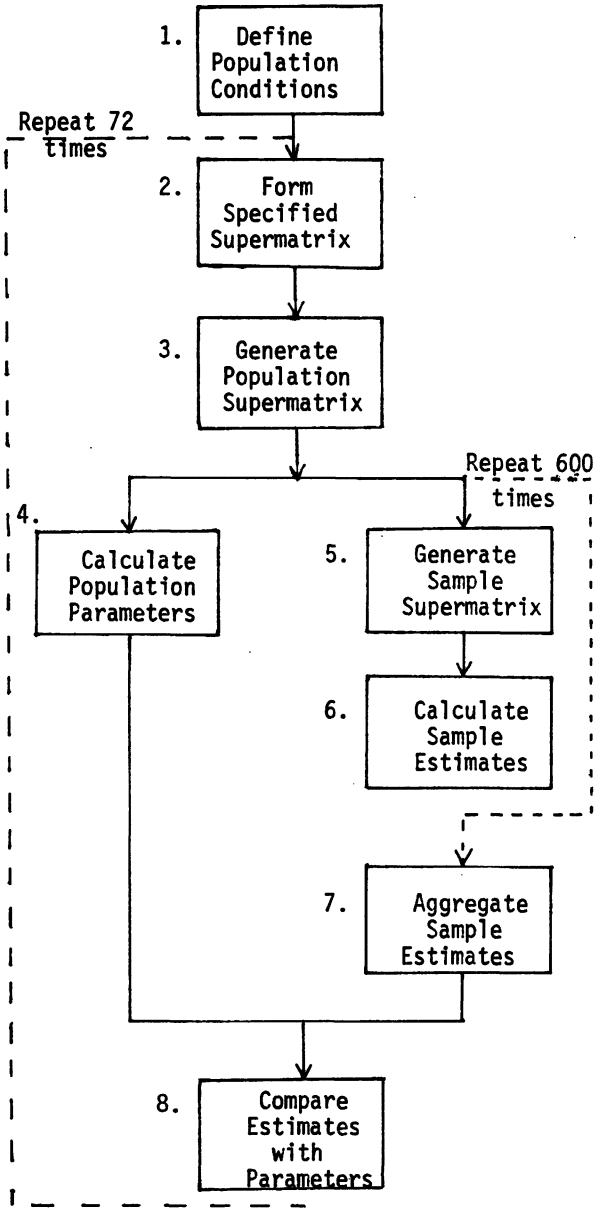


TABLE 1: REGRESSION OF TOTAL REDUNDANCY POPULATION VALUE ON SAMPLE INFORMATION

NOT INCLUDING THE SAMPLE MEAN AS A PREDICTOR:

VARIABLE	ORIGINAL β WEIGHT	F
p	0.0102	9.15
q	-0.0007	0.04
R_{xx}	-0.0496	21.01
R_{yy}	-0.0372	11.80
R_{xy}	0.0897	183.36
n	-0.0000	0.00
Constant	0.1775	

Multiple R = .7904

Adjusted Multiple R = .7903

INCLUDING THE MEAN SAMPLE VALUE AS A PREDICTOR:

VARIABLE	ORIGINAL β WEIGHT	F
p	-0.0025	12.18
q	-0.0033	23.86
R_{xx}	-0.0032	1.95
R_{yy}	-0.0024	1.11
R_{xy}	0.0056	7.84
n	0.0003	541.45
$\bar{X}$	0.9984	3261.739
Constant	-0.0484	

Multiple R = .9925

Adjusted Multiple R = .9911

TABLE 2

WHERRY AND OLKIN/PRATT CORRECTION FORMULAE APPLIED TO THE TOTAL REDUNDANCY STATISTIC

p	q	n	POPULATION VALUE	POPULATION-WHERRY	POPULATION-OLKIN/PRATT	p	q	n	POPULATION VALUE	POPULATION-WHERRY	POPULATION-OLKIN/PRATT
5	5	50	0.143963	0.003304	-0.004118	5	5	50	0.158364	0.003330	-0.004422
5	5	50	0.140473	-0.000261	-0.007684	5	5	50	0.159670	0.000585	-0.007257
5	5	50	0.409582	0.008356	-0.002657	5	5	50	0.368618	0.008368	-0.002417
5	5	50	0.143963	0.007461	0.000138	5	5	50	0.169623	0.005078	-0.002884
5	5	50	0.119790	0.004827	-0.001965	5	5	50	0.153155	0.010905	0.003447
5	5	50	0.251462	0.006330	-0.003142	5	5	50	0.339137	0.006596	-0.003964
5	5	50	0.148249	0.005880	-0.001581	5	5	50	0.158365	0.005209	-0.002500
5	5	50	0.097611	0.007781	0.001653	5	5	50	0.131802	0.007002	-0.000037
5	5	50	0.257855	0.004125	-0.005480	5	5	50	0.217230	0.004450	-0.004474
5	5	50	0.148248	0.000872	-0.006706	5	5	50	0.169623	0.005690	-0.002259
5	5	50	0.088609	0.003768	-0.002222	5	5	50	0.125883	0.006079	-0.000835
5	5	50	0.140648	0.007815	0.000580	5	5	50	0.188901	0.006999	-0.001329

TABLE 2 (continued)

WHERRY AND OLKIN/PRATT CORRECTION FORMULAE APPLIED TO THE TOTAL REDUNDANCY STATISTIC

p	q	n	POPULATION VALUE	POPULATION-WHERRY	POPULATION-OLKIN/PRATT	p	q	n	POPULATION VALUE	POPULATION-WHERRY	POPULATION-OLKIN/PRATT
5	5	200	0.143963	0.001337	-0.000041	9	5	280	0.185906	0.001482	0.000335
5	5	200	0.140473	0.001132	-0.000223	9	5	280	0.235254	0.001258	-0.000087
5	5	200	0.409582	0.002325	-0.000172	9	5	280	0.502448	0.001689	-0.000129
5	5	200	0.143963	0.002531	0.001161	9	5	280	0.178609	0.000362	-0.000758
5	5	200	0.119790	0.001502	0.000297	9	5	280	0.195848	0.002733	0.001549
5	5	200	0.251462	0.001453	-0.000545	9	5	280	0.468800	0.000971	-0.000843
5	5	200	0.148249	0.002995	0.001599	9	5	280	0.185910	0.000338	-0.000814
5	5	200	0.097611	0.001712	0.000675	9	5	280	0.185180	0.002177	0.001037
5	5	200	0.257855	0.001999	-0.000027	9	5	280	0.311414	0.000169	-0.001417
5	5	200	0.148248	0.001202	-0.000206	9	5	280	0.178601	0.002573	0.001463
5	5	200	0.088609	0.002333	0.001371	9	5	280	0.155714	0.000544	-0.000469
5	5	200	0.140648	0.001889	0.000538	9	5	280	0.255294	0.000875	-0.000543
5	5	200	0.158364	0.000991	-0.000485	9	9	90	0.146454	0.009138	0.005702
5	5	200	0.159670	0.001988	0.000509	9	9	90	0.211141	0.004578	0.000240
5	5	200	0.368618	0.003466	0.001054	9	9	90	0.464035	0.007861	0.001979
5	5	200	0.169623	0.001568	0.000023	9	9	90	0.146455	0.007755	0.004299
5	5	200	0.153155	0.001228	-0.000212	9	9	90	0.160521	0.002294	-0.001437
5	5	200	0.339137	0.001201	-0.001137	9	9	90	0.337630	0.003364	-0.002098
5	5	200	0.158365	0.000399	-0.001082	9	9	90	0.157883	0.013028	0.009483
5	5	200	0.131802	0.001810	0.000521	9	9	90	0.181501	0.000256	-0.003776
5	5	200	0.217230	0.003660	0.001847	9	9	90	0.452014	0.004943	-0.000930
5	5	200	0.169623	0.000722	-0.000829	9	9	90	0.157884	0.009940	0.006352
5	5	200	0.125883	0.000900	-0.000354	9	9	90	0.142738	0.002298	-0.001183
5	5	200	0.188901	0.001900	0.000238	9	9	90	0.262638	0.000827	-0.004083
9	5	70	0.126378	0.005372	0.000981	9	9	90	0.125934	0.007711	0.004559
9	5	70	0.177453	0.006378	0.001104	9	9	90	0.196522	-0.000012	-0.004232
9	5	70	0.404501	0.004114	-0.003453	9	9	90	0.379337	0.006317	0.000653
9	5	70	0.126366	0.006023	0.001645	9	9	90	0.118614	0.004445	0.001356
9	5	70	0.140718	-0.000907	-0.005678	9	9	90	0.153542	0.002280	-0.001355
9	5	70	0.291738	0.005802	-0.000980	9	9	90	0.453077	0.001585	-0.004293
9	5	70	0.135215	0.003134	-0.001464	9	9	90	0.130175	0.008485	0.005281
9	5	70	0.140720	-0.000963	-0.005735	9	9	90	0.150499	-0.000666	-0.004299
9	5	70	0.358482	0.011438	0.004148	9	9	90	0.325202	0.000843	-0.004557
9	5	70	0.135185	0.006471	0.001935	9	9	90	0.122898	0.006177	0.003048
9	5	70	0.117356	0.004498	0.000263	9	9	90	0.119296	0.001248	-0.001901
9	5	70	0.216770	0.007672	0.001819	9	9	90	0.248147	0.002326	-0.002432
9	5	70	0.185906	0.005070	-0.000397	9	9	360	0.146454	0.003555	0.002827
9	5	70	0.235254	0.006339	0.000183	9	9	360	0.211141	0.002434	0.001475
9	5	70	0.502448	0.004181	-0.003505	9	9	360	0.464035	-0.000382	-0.001786
9	5	70	0.178609	0.002624	-0.002767	9	9	360	0.146455	0.003094	0.002364
9	5	70	0.195848	0.001877	-0.003791	9	9	360	0.160521	0.000214	-0.000581
9	5	70	0.468800	0.007413	-0.000298	9	9	360	0.337630	0.001363	0.000092
9	5	70	0.185910	0.006682	0.001240	9	9	360	0.157883	0.005156	0.004390
9	5	70	0.185180	0.002131	-0.003371	9	9	360	0.181501	-0.000246	-0.001116
9	5	70	0.311414	0.009326	0.002363	9	9	360	0.452014	0.000755	-0.000644
9	5	70	0.178601	0.006940	0.001619	9	9	360	0.157884	0.005443	0.004678
9	5	70	0.155714	0.001594	-0.003436	9	9	360	0.142738	0.002636	0.001919
9	5	70	0.255294	0.007934	0.001547	9	9	360	0.262638	-0.000884	-0.001999
9	5	280	0.126378	0.000311	-0.000557	9	9	360	0.125934	0.002309	0.001657
9	5	280	0.177453	0.001906	0.000798	9	9	360	0.196522	0.000495	-0.000423
9	5	280	0.404501	0.001089	-0.000673	9	9	360	0.379337	0.001064	-0.000270
9	5	280	0.126366	0.001165	0.000301	9	9	360	0.118614	0.001321	0.000695
9	5	280	0.140718	0.001171	0.000234	9	9	360	0.153542	0.001024	0.000259
9	5	280	0.291738	0.001049	-0.000481	9	9	360	0.453077	0.001321	-0.000078
9	5	280	0.135215	0.000048	-0.000867	9	9	360	0.130175	0.003793	0.003130
9	5	280	0.140720	0.001432	0.000497	9	9	360	0.150499	-0.000119	-0.000876
9	5	280	0.358482	-0.000552	-0.002245	9	9	360	0.325202	-0.000193	-0.001445
9	5	280	0.135185	0.002048	0.001143	9	9	360	0.122898	0.004237	0.003605
9	5	280	0.117356	-0.000014	-0.000837	9	9	360	0.119296	0.005171	0.004559
9	5	280	0.216770	0.001037	-0.000240	9	9	360	0.248147	0.001237	0.000166

SOME ESTIMATORS OF POPULATION TOTALS FROM SIMPLE RANDOM SAMPLES CONTAINING LARGE UNITS

M.A. Hidiroglou and K.P. Srinath, Statistics Canada

1. INTRODUCTION

The problem considered in this paper is the estimation of the population total of some characteristic from a simple random sample containing a few large or extreme observations. These observations are true observations belonging to the population that is being sampled. The presence of these observations in the sample will tend to make the usual estimate of the population total $\hat{Y}_0 = N\bar{y}$ (where $\bar{y}$ is the sample mean and N the population size) exceed the population total Y by a considerable amount though the estimation procedure itself is unbiased. It is therefore important to deflate the weights for such units at the estimation stage once they have been sampled and identified.

Several techniques have been proposed to handle unusually large values. Tukey and McLaughlin (1963) considered trimmed and Winsorized sample means from symmetric distributions. Crow (1964) has studied weighting procedures for observations. Fuller (1960) studied one-sided Winsorized means, Winsorization being applied to the largest observations only, assuming that the right tail of the distribution is well approximated by the tail of a Weibull distribution. Censored sample procedures have been considered by numerous authors (see for example Dixon (1960)). Searls (1966) proposed an estimator that used information external to the sample to predetermine a point, T, which separates "large" sample observations from the rest.

Recently, in studying estimators for skewed populations, Jenkins, Ringer and Hartley (1973) have adopted biased estimators which were preferable to $N\bar{y}$. Their quadratic loss function incorporated both the squared bias and the variance of the estimators, i.e., the mean square error (MSE).

We confine our attention to estimators which involve only a change of the usual weights as this seems a realistic and practical approach in sample surveys. No knowledge of the number of large units (outliners) in the population is assumed. We propose three estimators which are designed to reduce the effect of these large observations. The efficiencies of these estimators are empirically investigated along with the efficiency of the post-stratified estimator which involves a knowledge of the number of outliers in the population. The criterion for comparison of the proposed estimators with the usual estimator $N\bar{y}$ is the ratio of the variance of the unbiased estimator to the mean square error of these estimators. It is shown that, in certain situations, these estimators will have a smaller mean square error than the usual estimator $N\bar{y}$.

2. THE ESTIMATORS

We assume that a population $\{Y_1, Y_2, \ldots, Y_N\}$ of size N contains T large units. It is assumed that T is unknown. These outliers are are elements of the population whose Y-value exceeds a prespecified value T. A simple random sample of size n is drawn without replacement from the population and t outliers are identified. The estimators which we consider are:

$$\hat{Y}_1 = \sum_{i=1}^{t} y_i + \frac{N-t}{n-t} \sum_{i=t+1}^{n} y_i, \qquad (2.1)$$

$$\hat{Y}_2 = \frac{N}{n} \sum_{i=1}^{n} y_i - \frac{Nt}{2n} \left(\frac{n-t}{n}\right) \left(\sum_{i=1}^{t} \frac{y_i}{t} - \sum_{i=t+1}^{n} \frac{y_i}{n-t} \right) \qquad (2.2)$$

and

$$\hat{Y}_3 = r \sum_{i=1}^{t} y_i + \frac{N-rt}{n-t} \sum_{i=t+1}^{n} y_i . \qquad (2.3)$$

Estimator (2.1) assigns weight one to the outlier units and adjusts the weights of the non-outliers so that the sum of the sample weights adds up to N. Estimator (2.2) assigns a weight to the outlier units which is dependent upon the number of outliers in the sample. Finally, estimator (2.3) generalizes estimator (2.1) in that it assigns an optimal weight r to the outlier and non-outlier units.

If T is known a priori, the post-stratified estimator is:

$$\hat{Y}_4 = \frac{T}{t} \sum_{i=1}^{t} y_i + \frac{N-T}{n-t} \sum_{i=t+1}^{n} y_i . \qquad (2.4)$$

The bias and the mean square error (MSE) of these estimators are given in the following section.

3. THE MSE OF THE ESTIMATORS

We shall first consider the usual estimator of the population total $\hat{Y}_0$. $\hat{Y}_0$ may be expressed as the sum of outlier units and non-outlier units as:

$$\hat{Y}_0 = \frac{N}{n} \left\{ \sum_{i=1}^{t} y_i + \sum_{i=t+1}^{n} y_i \right\} . \qquad (3.1)$$

The variance of Y_o in the form given in (3.1) is

$$V(\hat{Y}_0) = \{ f^{-1} T\left(\frac{N-T}{N-1}\right) (1-\delta)^2 + N(f^{-1} - 1) \frac{T-1}{N-1} C_2^2 \delta^2 + N(f^{-1} - 1) \frac{N-T-1}{N-1} C_1^2 \delta^2 \} \bar{Y}_\nu^2 \qquad (3.2)$$

where f is the sampling fraction, δ is the ratio of the mean of the outlier units $\bar{Y}_\mu$ in the population to the mean of the non-outlier $\bar{Y}_\nu$ units in the population, C_1 and C_2 are the coefficients

of variation for the non-oulier and outlier units in the population respectively.

It can easily be shown that the biases of $\hat{Y}_1$, $\hat{Y}_2$ and $\hat{Y}_3$, for $T \geq 1$ are

$$B(\hat{Y}_1) = -T(1-f)(\delta-1)\,\bar{Y}_\nu \,, \qquad (3.3)$$

$$B(\hat{Y}_2) = \frac{-T(\delta-1)(N-T)\,\bar{Y}_\nu}{2N} \,, \qquad (3.4)$$

$$B(\hat{Y}_3) = -T(1-rf)(\delta-1)\,\bar{Y}_\nu \,. \qquad (3.5)$$

Note that estimators (2.1) and (2.3) are consistent whereas estimator (2.2) is not. The mean square error (MSE) of these estimators can be presented in two ways, depending on T. For T=1, the mean square error can be derived exactly. For $T > 1$, the approximate MSE for $\hat{Y}_1$ and $\hat{Y}_3$ is obtained using $E(t) \doteq 1/E(t)$. For $T > 1$, the exact MSE for $\hat{Y}_2$ has been derived.

We first present the exact mean square errors associated with T = 1. Details of the derivations are not given here.

$$MSE(\hat{Y}_1) = \{(1-f)(1-\delta)^2 + [\frac{f(N-1)}{n-1}(N-n) + N(1-f)(f^{-1} - \frac{N}{N-1})]\, C_1^2\, \bar{Y}_\nu^2\} \qquad (3.6)$$

$$MSE(\hat{Y}_2) = \{\frac{N^2 n(1-f)}{f(N-1)(n-1)}[(1-f)(1 - \frac{n+1}{2n})^2 + 1]\, C_1^2 + (1-\delta)^2\,[(1-f)[1 - \frac{f^{-1}}{2}(1 + \frac{1}{n})]^2 + f]\}\, \bar{Y}_\nu^2 \qquad (3.7)$$

and

$$MSE(\hat{Y}_3) = \{[(1-f) + f(1-r)^2](1-\delta)^2 + [\frac{f(N-r)^2}{n-1}\,\frac{N-n}{N-1} + N(1-f)(f^{-1} - \frac{N}{N-1})]\, C_1^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.8)$$

The optimal value of r for (3.8) is given as

$$r_o = \frac{(1-f)\,C_1^2 + f(1-\delta)^2}{\frac{(1-f)}{N}\,C_1^2 + f(1-\delta)^2} \,.$$

Next, we provide expressions for MSE for $T > 1$.

$$MSE(\hat{Y}_1) \doteq (1-\delta)^2 f(1-f)\,T\,(1 - \frac{T}{N}) + (T-1)\, f(1-f)\, C_2^2\, \delta^2 + \frac{(1-f)}{f(N-T)}[(N-fT)^2 - f^2T]\, C_1^2 + T^2(1-f)^2(1-\delta)^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.9)$$

$$MSE(\hat{Y}_2) = [\frac{T(1-\delta)(N-T)(n-1)}{2n(N-1)}]^2 + \left(\frac{C_2\delta}{2f}\right)^2 [Et + \frac{2Et^2}{n} + \frac{Et^3}{n^2} - \frac{1}{T}(Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})] + \left(\frac{C_1}{2f}\right)^2 [(4n - \frac{3\,Et^2}{n} - \frac{Et^3}{n^2} - \frac{1}{N-T}(4n^2 - 4nEt - 3Et^2 + \frac{2Et^3}{n} + \frac{Et^4}{n^2})] + (\frac{1}{2nf})^2 (1-\delta)^2\, V(nt + t^2)\}\, \bar{Y}_\nu^2, \qquad (3.10)$$

where $V(nt + t^2) = n^2\, V(t) + 2n\, Cov\,(t,t^2) + V(t^2)$ and Et^k, k=1, 2,3,4, are moments obtained from the hypergeometric distribution given by

$$H(t|N,n,T) = \frac{\binom{N-T}{n-t}\binom{T}{t}}{\binom{N}{n}} \,, \quad 0 \leq t \leq T,\ N-T > n.$$

The mean square error of $\hat{Y}_3$ for $T > 1$ is

$$MSE(\hat{Y}_3) \doteq \{r^2(1-\delta)^2\, f(1-f)\, T(1 - \frac{T}{N}) + r^2(T-1)\, f(1-f)\, C_2^2\, \delta^2 + \frac{(1-f)}{f(N-T)}[(N-rfT)^2 - r^2f^2T]\, C_1^2 + T^2(1-rf)^2\,(1-\delta)^2\}\, \bar{Y}_\nu^2 \,. \qquad (3.11)$$

The optimal value of r for $T > 1$ is obtained by minimizing (3.11). Differentiating (3.11) with respect to r and solving for r, we obtain

$$r_o = \frac{g_1(N,f,T,\delta,C_1)}{g_2(N,f,T,\delta,C_1,C_2)} \qquad (3.12)$$

where

$$g_1(N,f,T,\delta,C_1) = (1-\delta)^2\, fT^2 + \frac{(1-f)TN}{N-T}\, C_1^2$$

and

$$g_2(N,f,T,\delta,C_1,C_2) = (1-\delta)^2\, fT\,[(1-f)(1 - \frac{T}{N}) + fT]$$

$$+ f(1-f)(T-1)[C_2^2 \delta^2 + \frac{T}{N-T} C_1^2].$$

The variance of the post-stratified estimator $\hat{Y}_4$ for T > 1 is given by

$$V(\hat{Y}_4) \doteq \{C_1^2 [f^{-1}(1-f)(N-T) + \frac{T}{nf}] + C_2^2 \delta^2 [f^{-1}(1-f)T + \frac{N-T}{nf}]\} \bar{Y}_\nu^2 . \quad (3.13)$$

4. AN EMPRICAL INVESTIGATION OF THE ESTIMATORS

To investigate the efficiency and utility of the proposed estimators, we have used a variety of artificial populations. We have studied the relative efficiency of these estimators for various values of C_1, C_2, δ, f, N and T. The relative efficiency is defined as the ratio of the variance of the usual estimator of the total $\hat{Y}_0$ to the mean square error of $\hat{Y}_i$, i=1, 2,3,4.

The empirical investigation has been extensive and in view of the difficulty of presenting a great number of tables, only six tables are presented. Tables 1 through 5 are constructed to reveal a difference in the behaviour of the estimators $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ and $\hat{Y}_4$ for various values of C_1, C_2, δ, f and T for a given value of N. Within each of these tables C_2 and T vary while, δ, f and C_1 are fixed. The tables differ from each other by having one of the variables δ, f or C_1 vary while the other two variables are fixed. Table 6 differs from the others in that a large value of N and a small sampling fraction f have been used. The conclusions drawn from these tables, in general, should apply to other populations.

Tables of Relative Efficiencies

Estimators	$\hat{Y}_1$		$\hat{Y}_2$		$\hat{Y}_3$		$\hat{Y}_4$	
1.			δ=5	f=0.3		C_1= 0.5	N=500	
T \ C_2	1.0	2.0	1.0	2.0	1.0	2.0	1.0	2.0
2	1.26	-	1.16	-	1.26 (1.10)	-	0.62	-
4	1.37	2.32	1.41	2.00	1.41 (1.49)	2.32 (0.78)	0.77	0.49
10	1.02	2.13	1.30	2.08	1.37 (2.15)	2.17 (1.32)	1.03	0.74
15	0.75	1.69	1.10	1.89	1.30 (2.43)	1.92 (1.62)	1.15	0.84
25	0.48	1.18	0.81	1.56	1.20 (2.72)	1.61 (2.02)	1.28	0.94
80	0.14	0.40	0.35	0.83	1.06 (3.12)	1.20 (2.77)	1.43	1.07
2.	δ=5	f=0.1		C_1=0.5			N=500	
2	1.37	-	1.28	-	1.37 (1.19)	-	0.40	-
4	1.75	3.22	1.56	2.13	1.75 (1.76)	3.22 (0.75)	0.53	0.30
10	1.85	4.17	1.85	2.63	2.04 (3.22)	4.17 (1.46)	0.78	0.53
15	1.53	3.70	1.85	2.63	1.96 (4.11)	3.84 (2.00)	0.92	0.65
25	1.06	2.78	1.67	2.00	1.69 (5.34)	3.12 (2.87)	1.09	0.79
3.	δ=10	f=0.3		C_1=0.5			N=500	
2	1.78	-	1.72	-	1.78 (1.16)	-	0.48	-
4	1.64	3.12	1.78	2.56	1.78 (1.58)	3.12 (0.89)	0.75	0.46
10	0.92	2.04	1.30	2.13	1.45 (2.24)	2.17 (1.47)	1.15	0.76
15	0.64	1.51	1.02	1.82	1.31 (2.51)	1.85 (1.79)	1.30	0.87
25	0.40	0.99	0.71	1.41	1.19 (2.78)	1.51 (2.18)	1.45	0.98
80	0.12	0.33	0.31	0.71	1.06 (3.14)	1.16 (2.86)	1.59	1.10

4.	δ=10		f=0.1		C_1=0.5			N=500	
2	2.43	-	1.92	-	2.43 (1.23)	-		0.27	-
4	3.03	6.25	2.32	2.94	3.12 (1.89)	6.25	(0.87)	0.46	0.27
10	2.08	5.00	2.27	2.94	2.56 (3.48)	5.26	(1.70)	0.82	0.53
15	1.51	3.70	2.08	2.78	2.13 (4.41)	4.17	(2.30)	1.00	0.66
25	0.94	2.44	1.69	2.50	1.72 (5.66)	3.03	(3.28)	1.20	0.81
5.	δ=5		f=0.3		C_1=1.0			N=500	
2	1.06	-	1.06	-	1.06 (1.16)	-		0.83	-
4	1.12	1.47	1.12	1.39	1.14 (1.53)	1.47	(0.80)	0.88	0.63
10	1.00	1.67	1.14	1.64	1.18 (2.17)	1.69	(1.33)	1.02	0.79
15	0.81	1.52	1.05	1.64	1.16 (2.44)	1.64	(1.64)	1.10	0.87
25	0.54	1.14	0.84	1.45	1.14 (2.73)	1.49	(2.03)	1.20	0.95
80	0.16	0.41	0.37	0.83	1.06 (3.12)	1.19	(2.77)	1.37	1.06
6.	δ=5		f=.01		C_1=0.5			N=10,000	
5	1.06	1.19	1.05	1.14	1.07 (2.88)	1.19	(1.09)	0.52	0.23
15	1.22	1.64	1.16	1.41	1.22 (6.41)	1.64	(2.39)	0.57	0.29
25	1.33	2.04	1.25	1.64	1.35 (9.75)	2.04	(3.70)	0.62	0.35
25	1.51	2.70	1.41	1.96	1.54(15.79)	2.70	(6.22)	0.70	0.45
65	1.61	3.12	1.54	2.17	1.67(21.08)	3.12	(8.61)	0.77	0.52
85	1.61	3.33	1.61	2.32	1.75(25.76)	3.45	(10.89)	0.83	0.58

Note: Dashes indicate that C_2 is non-existent for these cases. The numbers in brackets are the optimal r_o values given by (3.12).

It is seen from the above tables that, for fixed δ, f, C_1, C_2, and N, the efficiencies of of the estimators decrease after an initial improvement as T increases. The efficiency gain in using these estimators increases as the coefficient of variation C_2 of the outlier units increases. Comparing the values in Table 1 with those in Table 5, we see that as C_1 increases, the efficiencies of the estimators decrease for small values of T and increase after a certain number of outliers has been reached. Comparing values in Tables 1 and 3, we see that as δ increases from 5 to 10, gains in efficiency are not uniform. In fact, for large T, there is a greater loss in efficiency. This is due to the fact that the bias term of the estimators dominates the mean square error as δ increases. Referring to Tables 1 and 2, 3 and 4, it is seen that as f decreases, gains in efficiencies of the estimators increase.

To stress the effectiveness of these estimators, a fairly large population of N=10,000 and a small sampling fraction of f=0.01 have been used. The results are given in Table 6. Note that for a few number of outliers in the population, the gain in using these estimators is not very considerable. However, as the number of outliers in the population increases, the effectiveness of these estimators improves quite significantly.

It is possible to make the following general observations. The best estimator to use with respect to efficiency is $\hat{Y}_3$. $\hat{Y}_2$ has lower efficiency than $\hat{Y}_1$ for a small number of outliers, however, after a certain number of outliers has been reached, $\hat{Y}_2$ is superior to $\hat{Y}_1$. Hence, $\hat{Y}_2$ is to be preferred to $\hat{Y}_1$ in the presence of a moderate number of outliers. For a small number of outliers, the post-stratified estimator $\hat{Y}_4$ is not as good as the other estimators because the allocation between the post-strata is likely to be poor, being very different from the optimum allocation in such cases. But, as expected, once a certain number of outliers is reached, it is superior to all estimators including $\hat{Y}_0$.

$\hat{Y}_3$, the optimal estimator, requires a knowledge of T, C_1, C_2 and δ from the sample. We use these in the expression (3.12). Estimating r_o using sample values could imply a departure from optimal efficiency of $\hat{Y}_4$. To study this possible departure, the efficiency of $\hat{Y}_3$ has been investigated for different values of $r_o(1+\Delta)$, where $0.0 \leq \Delta < 1.0$. Two situations have been invest-

TABLES OF RELATIVE EFFICIENCIES OF $\hat{Y}_3$ FOR $r_o(1+\Delta)$

7.	δ=5	f=0.3	C_1=0.5		C_2=1.0	N=500
T	2	4	10	15	25	80
$\lvert\Delta\rvert$ \ r_o	1.10	1.49	2.5	2.43	2.72	3.12
0.0	1.26	1.41	1.37	1.30	1.20	1.06
0.1	1.26	1.41	1.35	1.26	1.15	0.94
0.2	1.26	1.40	1.31	1.19	1.04	0.69
0.3	1.26	1.38	1.23	1.09	0.89	0.47
0.4	1.25	1.35	1.15	0.97	0.74	0.33
0.5	1.25	1.32	1.05	0.85	0.61	0.24
0.6	1.23	1.29	0.95	0.74	0.51	0.18
0.7	1.23	1.25	0.86	0.64	0.42	0.14
0.8	1.21	1.21	0.77	0.56	0.35	0.11
0.9	1.20	1.16	0.69	0.48	0.29	0.09

8.	δ=5	f=0.01	C_1=0.5		C_2=1.0	N=10,000
T	5	15	25	45	65	85
$\lvert\Delta\rvert$ \ r_o	2.88	6.41	9.75	15.79	21.08	25.76
0.0	1.069	1.216	1.346	1.546	1.678	1.755
0.2	1.069	1.216	1.345	1.545	1.673	1.746
0.4	1.069	1.216	1.344	1.541	1.541	1.664
0.6	1.069	1.216	1.343	1.535	1.648	1.697
0.8	1.069	1.215	1.341	1.527	1.626	1.656

igated. The first one being a large population of size 10,000 with an associated small sampling fraction of 0.01 and the second being a small population size of 500 with a fairly high sampling fraction of 0.3. The results are given in Tables 7 and 8. From the preceding tables, it is seen that when there is a low number of outliers, the efficiency of $\hat{Y}_3$ is not significantly affected by departures from optimal r_o. As the number of outliers increases in the first (N=500) population, even small departures from optimal r_o result in low efficiency. Note that in the case of the second population (N=10,000), departures from optimal r_o are not significant even for large number of outliers in the population.

5. CONCLUSIONS

When the sampling fraction f and the number of outliers T are small, use of the estimator $\hat{Y}_1$ would result in substantial gains in efficiency. If f and T are moderately large, use of $\hat{Y}_2$ is recommended. $\hat{Y}_3$ can be used to advantage if values of C_1, C_2, δ and T are approximately known from previous surveys. Deviations from the optimal r_o associated with $\hat{Y}_3$ will not affect the efficiency if T is small. If T is large and known, it is obvious that the post-stratified estimator $\hat{Y}_4$ should be used.

REFERENCES

Bershad, M., "Some Observation On Outliers", Unpublished dittoed memorandum, 1960, Statistical Research Division, U.S. Bureau of Census.

Chinnappa, N., "A Preliminary Note On Methods of Dealing With Unusually Large Units In Sampling from Skew Populations", Unpublished, IASMD technical memorandum, February 1976.

Crow, Edwin L., "The Statistical Construction of a Single Standard from Several Available Standards", IEEEE Transactions On Instrumentation and Measurement, 13 (1964), 180-5.

Dixon, W.J., "Simplified Estimation From Censored Normal Samples", Annals of Mathematical

Statistics, 31 (1960), 385-91.

Fuller, W.A., "Simple Estimation for the Mean of Skewed Populations" 1960, U.S. Bureau of Census.

Hartley, Herman O., and Rao, J.N.K., "A New Estimation Theory for Sample Surveys", Biometrika, 55 (March 1968), 547-57.

Jenkins, O.C., Ringer, L.G., and Hartley, H.O., "Root Estimators", Journal of the American Statistical Association, 68 (1973) 414-19.

Rao, C.R., "Some Aspects of Statistical Inference in Problems of Sampling from Finite Populations", in Foundation of Statistical Inference. Holt, Rinehart and Winston of Canada Ltd., (1971) 177-202.

Searls, D.T., "An Estimator which Reduces Large True Observations", Journal of the American Statistical Association, (1966), 1200-4.

Tukey, J.W., and McLaughlin, D.H., "Less Vulnerable Confidence and Significance Procedures for Location Based On A Single Sample: trimming/ Winsorization 1", Sankhya, Series A, 25 (1963), 331-52.

A COMPARISON OF APPROXIMATE STRATIFICATION TECHNIQUES UNDER AN AREA SAMPLING FRAME, TWO EMPIRICAL STUDIES

Robert D. Tortora, USDA

Nicholas J. Ciancio, USDA

Introduction

Recently work has been done comparing various approximate optimum stratification techniques. Hess, et.al., [6] and Cochran, [2] compared various techniques for actual populations. In [2], Cochran was concerned with eight different populations ranging from income per tax return, population of US cities, resources of commercial banks, number of farms per area sampling unit, and proportion of gross bank loans. In [6], the stratification and primary estimation variable was size of hospital. A brief discussion examines other estimation variables with high correlations (> .9) with the stratification variable, Kish, et.al., [1] compared various stratification techniques for a specified bivariate population where the stratification is carried out on an auxiliary variable X and estimation is made for a variable Y. Kpedeko, [7] in a review of the literature on stratification techniques calls for further empirical studies to evaluate some of these methods for different types of data.

This paper compares five approximate optimum stratification techniques when an auxiliary variable is used for stratification and when one is interested in estimating crop acreage and livestock totals. The stratification techniques are 1) cum$\sqrt{f}$, 2) Durbin, 3) Ekman, 4) Sethi, and 5) Equal Aggregate Output (EAO). The Statistical Reporting Services' (SRS) area frame is used in two States to make the comparisons. In the area frame the land area is classified (stratified) according to land use in order to achieve homogeneity within strata. The sampling unit is a segment, which is a piece of land with boundaries delineated on a map. Every parcel of land within a segment is accounted for during a survey.

The stratification variable for the area frame is the percent of land under cultivation. For each segment this is defined as the total cropland in acres in the segment divided by the total acres in the segment times 100. The crops acreage variables of interest are the three most important income-producing crops for US farmers, vis., corn, wheat, and soybeans. Similarly, the important livestock variables are cattle and hogs and these variables will also be studied.

The data used in this study are from the 1975 June Enumerative Survey for Ohio and Kansas. The segments are from the agricultural strata and the population sizes are N=252 and N=435 segments in Ohio and Kansas, respectively. Even though the above five commodities are the most important within Ohio and Kansas, these States differ demographically and geographically. Kansas is a more homogeneous farm state with more area under irrigation. The average size of farm in Kansas is larger (616 acres vs. 150 acres). Ohio has more farms (117,000 vs. 81,000) and less land in farms (17.5×10^6 acres vs 50×10^6 acres). The segment size is Kansas ranges from 1 to 4 square miles while in Ohio the segment size is ½ to 1 square miles.

Optimum allocation for fixed sample size is used to determine sample sizes in the strata. This technique is the one used by SRS.

The comparisons are made for 2, 3, 4 or 5 strata. Currently SRS uses four strata with stratum boundary values 15%, 50%, and 75%. The variances for the approximate techniques will be compared to the variance under the current SRS technique. The total number of strata is held to 5 for two reasons. First, to strain the stratification techniques, which depend more or less on the assumption that the number of strata L is reasonably large, so that within a stratum the frequency function can be assumed to be rectangular. Secondly, for practical purposes of frame construction, it is very difficult to efficiently divide the area frame into a large number of strata.

The Approximate Methods

Let $X_0, X_1, \ldots, X_L$ be the stratum boundaries, the strata being numbered 1, 2, ..., L. let S_h be the standard deviation is stratum h and $W_h = N_h/N$ be the ratio of the number of sampling units in stratum h to the total number in the population. The usual estimate of the population total is

$$y_{st} = N\Sigma W_h \bar{y}_h$$

where $\bar{y}_h$ is the sample mean in statum h. Its variance is

$$V(y_{st}) = N^2 \Sigma W_h^2 S_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h}\right).$$

For a fixed total sample size of n, $V(y_{st})$ is minimized by taking $n_n = nN_h S_h / \Sigma N_h S_h$. The minimum variance is

$$(1)\quad V_{min}(y_{st}) = \frac{1}{n}(\Sigma N_h S_h)^2,$$

ignoring the fpc. Equation (1) becomes the basis for further caculations.

A discussion of the actual implementation of the approximate stratification techniques can be found in [2], [6] or [7].

The Study Variables

Table 1 gives the shapes of the stratification variables in the two States. It is simply the percentage of the total number of segments lying within each tenth of the range of the percent of land under cultivation.

Table 1: Percentage of segments falling into successive tenths of the stratification variable.

	Ohio	Kansas
0-10	8.73	10.11
10-20	7.14	3.91
20-30	5.56	5.06
30-40	5.16	4.60
40-50	9.52	7.82
50-60	7.52	8.74
60-70	9.52	10.57
70-80	14.68	12.64
80-90	15.87	15.40
90-100	16.28	21.15

The distribution in Ohio is closest to a rectangular distribution, while in Kansas the distribution is close to being two-tailed. Cum$\sqrt{f}$, Durbin and Ekman compute approximately the same stratum boundary values, as do EAO and Sethi. Also all techniques seem to be relatively insensitive to the distribution of the stratification variable although in Kansas the values are higher than in Ohio.

Table 2 and 3 give the shape of the frequency distribution of the variables to be estimated.

Table 2: Percentage of segments having crop acreage falling into given classes by State, (Ohio/Kansas).

acres	corn	wheat	soybeans
0-10	17.46/76.78	34.92/13.79	31.75/80.46
10-50	26.59/ 8.04	38.10/ 9.66	19.05/ 8.50
50-100	36.11/ 5/98	22.62/12.18	24.60/ 6.21
100-250	19.44/ 4.83	4.36/30/11	23.81/ 4.60
> 250	0.00/ 4.37	0.00/34.26	0.79/ 0.03

Table 3: Percentage of segments having livestock numbers falling into given classes by States, (Ohio/Kansas).

Number of livestock	cattle	hogs
0	69.65/85.29	28.97/31.49
1 - 50	21.43/ 8.28	40.48/25.98
51 - 100	3.97/ 2.53	18.65/22.30
101 - 500	4.36/ 3.45	11.90/19.54
> 500	1.59/ 0.45	0.00/ 0.69

The tables show that in Ohio corn and wheat appear unimodal while soybeans appear bimodal. In Kansas corn and soybeans are skewed to the left while wheat is skewed to the right. The general shape of the distribution for hogs are the same. For cattle, the distributions are skewed to the left with the distribution in Ohio being fatter.

Finally Table 4 lists the estimated correlation coefficients between the stratification variable and the variables of interest. Note that none of the correlations are near the correlations in [6] and thus should strain the techniques.

Table 4: Estimated correlation coefficients between the stratification variable and the variables of interest, (Ohio/Kansas).

	corn	wheat	soybeans
Percent of land under cultivation	.554 / .240	.614 / .609	.652 / .028
	cattle	hogs	
percent of land under cultivation	.118 / .077	-.216 / -.356	

The correlation coefficients between the crops and the auxiliary variable are consistent in Ohio while in Kansas wheat has the highest correlation. In both states the magnitude of the correlations between livestock and percent cultivated are small.

Comparison of the Rules

Equation (1) is used as the basis for the comparisons for each technique. The variance under the approximate stratification technique is compared to the variance obtained under the present stratification used by SRS. The results are presented in Table 5.

The separation of boundary values (cum$\sqrt{f}$, Durbin, and Ekman vs. EAO and Sethi) is reflected in Table 5. As the number of strata increases the differences between the techniques does not change as much for the crop variables, i.e., for those with higher correlation coefficient with the stratification variable and where the stratification variable is more nearly rectangular (Ohio). Whenever EAO or Sethi have a smaller ratio for 2 strata than either cum$\sqrt{f}$, Durbin, or Ekman they retain their smaller ratio as the number of strata increases. For the negatively correlated variables in both states cum$\sqrt{f}$, Durbin and Ekman perform much better than the other techniques. For the highest positively correlated variables (soybeans in Ohio and wheat in Kansas) cum$\sqrt{f}$, Durbin, and Ekman perform well. To get a feel for the performance of the techniques across all variables of interest a plot of the technique(s) (with smallest ratios from Table 5 vs. the correlation coefficients (r) is given in Figure 1. For the range of correlation coefficients we see that cum$\sqrt{f}$, Durbin and Ekman perform well for negatively correlated variables as well as for moderately correlated values (r > .5). For small positive values of r all techniques seem to perform on a par.

Table 5: Variance under the approximate stratification technique divided by the variance under the current SRS stratification (Ohio/Kansas).

technique	strata	corn	wheat	soybeans	cattle	hogs
cum$\sqrt{f}$	2	3.434 / 2.499	3.229 / 2.963	3.165 / 3.012	3.224 / 2.341	3.528 / 3.374
	3	1.334 / 1.133	1.221 / 1.316	1.261 / 1.282	1.090 / 1.149	1.546 / 1.470
	4	0.781 / 0.643	0.676 / 0.719	0.699 / 0.703	0.707 / 0.603	0.882 / 0.658
	5	0.471 / 0.394	0.434 / 0.460	0.421 / 0.442	0.437 / 0.342	0.773 / 0.461

Table 5: (Con't)

technique	strata	corn	wheat	soybeans	cattle	hogs
Durbin	2	3.595 2.499	3.959 2.963	3.334 3.012	3.239 2.341	3.514 3.374
	3	1.440 1.046	1.357 1.236	1.418 1.266	1.260 0.989	1.622 1.398
	4	0.806 0.643	0.724 0.719	0.721 0.703	0.718 0.603	0.861 0.658
	5	0.483 0.377	0.452 0.439	0.433 0.420	0.446 0.285	0.538 0.444
Ekman	2	3.909 2.499	3.229 2.963	3.932 3.012	3.024 2.341	3.527 3.374
	3	1.369 1.133	1.254 1.316	1.297 1.282	1.210 1.149	1.566 1.470
	4	0.753 0.714	0.656 0.798	0.638 0.750	0.650 0.720	0.865 0.889
	5	0.498 0.440	0.464 0.481	0.425 0.428	0.467 0.424	0.554 0.429
EAO	2	4.281 2.316	4.055 3.776	3.981 3.253	3.139 1.601	5.187 4.783
	3	1.955 0.818	1.639 1.604	1.642 1.451	1.541 0.803	2.494 2.404
	4	1.045 0.450	0.942 0.912	0.934 0.836	0.844 0.343	1.650 1.433
	5	0.723 0.250	0.548 0.632	0.497 0.522	0.560 0.200	1.099 1.034
Sethi	2	4.281 1.963	4.055 3.320	3.981 3.480	3.139 1.635	5.013 4.396
	3	1.915 0.896	1.610 1.556	1.439 1.698	1.511 0.783	2.425 2.165
	4	1.048 0.497	0.925 0.847	0.945 0.937	0.842 0.339	1.696 1.284
	5	0.758 0.260	0.590 0.604	0.529 0.583	0.648 0.195	1.230 0.936

From Figure 1 we see cum$\sqrt{f}$ performs best 18 times, but Durbin is best 14 times, and Ekman is best 9 times. Cum$\sqrt{f}$ peforms well over the range of r, Durbin does well with smaller values of r, and Ekman does well with larger values of r. Figure 2 presents a graphic display of the worst of the best for the regions were the three techniques cum$\sqrt{f}$, Durbin, and Ekman perform well. Figure 2 presents the technique with the largest ratio from Table 6. The trends exhibited here indicate that cum$\sqrt{f}$ can give larger variances for negative r, Durbin for moderate r, and Ekman across the range of r.

Dalenius [3] suggested the approximation $V_L/V_{L-1} = (L-1)^2/L^2$ (for rectangular distributions) to quantify the gains caused by stratification. For L = 2, 3, 4, and 5 we get from the formula 0.250, 0.444, 0.562 and 0.640 respectivly. Table 6 presents the average gain by crop or livestock for each technique.

In general, the average gain is slightly more than that estimated by Dalenius. There is less gain in precision for the variables than with lower correlation (livestock in Ohio) with the stratification variable than with those with higher correlation (crops in Ohio). Comparing gains against the distribution of the stratifying variable we see that there are more gains (25 vs 15) for the unimodal distribution (Ohio). Defining any gain exceeding $(L-1)^2/L^2$ as significant, we see that for the unimodal distribution there are more significant gains (22 vs. 14). Finally, examining significant gains by correlation and by technique we see that cum$\sqrt{f}$ does best for the higher correlations (crop in Ohio and Kansas), Durbin performs about the same for crop and livestock and the remaining three produce more significant gains for the lower correlation (livestock in Ohio and Kansas).

Table 6: The average gain V_L/V_{L-1} for crops and livestock by straitification technique (Ohio/Kansas)

		crops	livestock
cum f	2	.188 .231	.248 .220
	3	.388 .440	.388 .464
	4	.546 .554	.610 .486
	5	.616 .627	.747 .634
Durbin	2	.202 .231	.248 .220
	3	.401 .419	.425 .418
	4	.541 .584	.550 .540
	5	.608 .598	.623 .574
Ekman	2	.176 .231	.287 .220
	3	.470 .441	.422 .464
	4	.522 .607	.545 .616
	5	.679 .597	.679 .536
EAO	2	.263 .256	.158 .226
	3	.424 .422	.496 .502
	4	.559 .508	.606 .512
	5	.602 .624	.662 .653

Table 7: (Con't)

		crops	livestock
Sethi	2	.236 / .228	.292 / .216
	3	.402 / .457	.482 / .486
	4	.593 / .558	.628 / .512
	5	.640 / .620	.747 / .652

Summary

This study compared five approximate techniques for stratification in an agricultural setting. The comparisons were based on area sampling units from two States. The stratification variable (percent of land under cultivation) was different from the variables to be estimated (corn, wheat, soybeans, cattle and hogs).

The rules divided themselves into two groups based on stratum boundary values, cum$\sqrt{f}$, Durbin, Ekman, and Equal Aggregate Output, Sethi. Comparisons were based on variances obtained using the current SRS stratification. Cum$\sqrt{f}$, Durbin, and Ekman performed well for variables either with negative correlations or moderate positive correlations with the stratification variable. All five rules were comparable for small positive correlations.

Using Dalenius' approximation, $(L-1)^2/L^2$, for gains due to increasing the number of strata it was found that the most significant gains were produced when the stratification variable was unimodal. Ekman, Equal Aggregate Output, and Sethi had more significant gains for variables not highly correlated with the stratification variable, gains and cum$\sqrt{f}$ produced more significant gaims with the higher correlated variables. It was found that the approximation $(L-1)^2/L^2$ was an overestimate of the gains due to increasing the number of strata (concurring with the results in [2]).

References

[1] Anderson, D.W., Kish, L., Cornell, R.G. "Quantifying Gains From Stratification for Optimum and Approximately Optimum Strata Using a Bivariate Normal Model", Tech. Report 4, Department of Biostatistics, The University of Michigan, (1975), Ann Arbor, Michigan 48104

[2] Cochran, W.G. "Comparison of Methods for Determining Stratum Boundaries", Bulletin of the International Statistical Institute, 38(2), Tokyo (1961), pp. 345-358.

[3] Dalenius, T. Sampling in Sweden, Chapter 8, Almquist and Wiksell, Stockholm (1957).

[4] Dalenius, T. and Hodges, J.L., Jr. "Minimum Variance Stratification", Journal of the American Statistical Association, 54 (1959), pp. 88-101

[5] Hansen, M.H., Hurwitz, W.N., Madow, W.G. Sample Survey Methods and Theory, John Wiley and Sons, Inc. 1953.

[6] Hess, I., Sethi, V.K., and Balakrishnan, T.R. "Stratification: A Practical Investigaion", Journal of the American Statistical Association, 61, (1966) pp. 74-90.

[7] Kpedekpo, G.M.K. "Recent Advances on Some Aspects of Stratified Sample Design. A Review of the Literature", Metrika, 20 (1073), pp. 54-64.

[8] Mahalanobis, P.C. "Some Aspects of the Design of Sample Surveys. Sankhya 12 (1952) pp. 1-7.

number of strata										
5	E	D	D	S	C	EAO	C	D	C	C
4	CD	D	CD	S	E	EAO	E	CD	E	E
3	D	C	D	EAO	C	EAO	C	D	C	C
2	CDE	D	CDE	EAO	EAOS	S	C	CDE	CE	C
		-.356	-.216	.028	.077	.118	.240	.554	.609	.614	.652	r

Figure 1: Best technique (smallest ratio from Table 6) vs. Correlation coefficient (ordered by increasing magnitude). C=cum$\sqrt{f}$, D=Durbin, E+Ekman, EAO=EAO, S=Sethi.

number of strata						
5	C	C	E	E	E	D
4	E	C	D	E	D	D
3	CE	D	D	CE	D	D
2	CDE	C	E	CDE	D	E
		-.356	-.216	.554	.609	.614	.652	r

Figure 2: Worst technique (largest ratio from Table 6) vs. correlation coefficient over ranges where cum$\sqrt{f}$, Durbin and Ekman perform well.

A STRATEGY FOR THE ANALYSIS OF MULTIPLE AREA STUDIES

Daniel H. Freeman, Jr., Robert W. Makuch, and Jan A. J. Stolwijk, Yale University School of Medicine

Many studies are designed in such a way that detailed information on individuals is collected by the use of a personal interview or examination. This process is then repeated in a variety of different geographic areas. The objective of such studies is to assess the variation of personal social and/or health attributes across a variety of environmental or ecological conditions. Thereby appropriate associations among the study variables may be evaluated. Studies of this type may be used to generate hypotheses concerning causal relationships and to support a variety of policy decisions. It should be recognized that such studies cannot actually test causal hypotheses, but replication of the results provides reassurance to the investigators about the phenomena under study. For these reasons it is important that statisticians examine the methodological issues associated with integration of individual measurement data and area wide aggregate data.

An aspect of this problem which is not widely discussed is the appropriate methodology when the individual measurements are categorical and the aggregate measurements are continuous in nature. For example, the individual measurements may be smoking status, sex, area of residence and size of urban area and the area wide measurements may be concentration of pollutants. In Table 1 we have precisely this type of data where three different pollutants are measured in micrograms per stere or cubit meter. The details concerning the collection and alternative analyses of the data are found in other sources (Berman(1976) and U.S. Environmental Protection Agency (1974)).

Table 1. Prevalence of chronic bronchitis per 100 population and sample size by smoking status, gender, area of study, and pollutant exposure.

Smoking Status and Gender		Area of study: Western Metropolitan				Western Non-Metropolitan				Eastern Metropolitan				Eastern Non-Metropolitan		
Never Smoked	Female	2.3	2.0	4.7	5.2	1.4	0.5	1.1	3.6	1.5	2.0	7.5	4.9	2.1	4.0	3.5
	n	755	755	772	667	440	207	94	337	333	197	411	529	384	202	344
	Male	3.0	3.6	2.3	6.8	1.1	0.0	4.9	4.9	2.0	4.6	18.0	14.2	2.8	5.2	5.2
	n	396	367	350	265	273	87	41	102	100	174	384	499	214	97	115
Formerly Smoked	Female	5.3	4.0	7.0	7.1	3.0	4.6	0.0	1.8	3.9	3.8	9.0	4.5	1.4	7.8	0.0
	n	75	101	114	84	131	66	27	112	102	144	233	226	140	64	61
	Male	2.6	3.4	5.4	6.0	0.0	5.3	0.0	4.7	5.0	13.9	18.0	18.7	5.7	8.2	5.6
	n	230	177	241	133	244	113	58	127	101	144	222	198	212	98	90
Currently Smokes	Female	17.1	14.7	15.3	22.2	8.7	14.2	13.7	13.8	11.8	13.9	19.8	16.6	7.1	7.0	9.8
	n	214	286	295	212	218	205	95	376	315	267	535	607	128	281	183
	Male	19.9	18.6	20.1	26.8	12.4	20.9	20.0	18.4	19.0	13.9	21.3	22.1	15.0	15.2	17.6
	n	272	311	354	209	209	187	85	250	260	216	492	526	132	287	159
Pollutant		Average Annual Concentration in micrograms per cubic meter (μ g/s)														
Sulphur Dioxide (SO_2)		10.	18.	32.	92.	10.	26.	67.	177.	374.	30.	174.	247.	13.	14.	4.
Total Suspended Particulates (PA)		88.	84.	50.	70.	50.	45.	115.	65.	102.	41.	84.	108.	38.	63.	48.
Suspended Sulphates		3.7	4.7	8.6	15.0	3.3	4.9	7.3	7.2	11.3	10.0	8.6	14.8	5.8	7.8	6.8

Inevitably, the comparability of the measurement across areas is of major concern and there is always the risk of confounding among the variables of interest. Typically, the individual measurements are thought of as blocking variables and the analysis focuses on some response of interest as if it were a continuous and normally distributed random variable. Moreover, it is assumed that the errors or residuals of such a model have essentially constant variance. A final assumption which is almost always made is that individual measurements are made on members of a simple random of some operationally defined population.

D.R. Cox (1970: pp. 16-18) notes that if the underlying probabilities associated with the response of interest lie between 0.2 and 0.8 then the usual least squares analysis will not in general be misleading. However, for data such as in the example this is clearly inappropriate since the observed prevalences range between 0 and 26.8 per cent. Moreover, for groups such as females who do not now smoke the prevalences are strictly less than 10.0 per cent. Cox notes several other difficulties with ordinary least squares analysis:

1. The method of estimation cannot be fully efficient.
2. The predicted values must be restricted to lie between 0 and 1.
3. It is not reasonable to extrapolate the regression equations outside the range of observation because of the obvious approximation being used in the linear equations.

Cox observed that each of these objections may be dealt with for binary variables through use of the logistic transformation. However, he does not examine the problem of what to do when the response is polytomous and ordinal or when the sample is actually based on a complex probability sample.

When the latter two issues are important a more general methodology and strategy of analysis is required. The strategy of Koch, Freeman, and Freeman (1975) (KFF) provides the appropriate framework for analysis. It is based on an elaboration of the method of Grizzle, Starmer, and Koch (1969) (GSK). It has been employed in a previous analysis of multiple area studies by Makuch and Freeman(1976) using the data of Heneley, Jain and Wells (1976). An aspect of the strategy known as modularization (Freeman, Freeman, and Brock (1977)) is appropriate for the analysis of data sets such as in the example. It is important to note that while the data set at hand is binary and thus lends itself to the logistic transformation, this is <u>not</u> a necessary condition for the analysis. If the original data were available it would be possible to use the original scaling (to 7) or alternatively either a ridit or probit scaling of the responses. Rather than re-iterate previously published material the data will be used to illustrate a strategy for the analysis of multiple area studies.

As noted earlier the data are divided in 6 sub-populations according to smoking status and gender. There are a total of 15 areas where these sub-populations were examined and data were collected on the atmospheric concentration of sulphur dioxide, suspended particulates and suspended sulphates. The areas may be broken into two regions and two urban classes, (West, East) and Metropolitan, Other). Notice this characterization leads to an unbalanced design. The logit of prevalance rates for each of six sub-populations was fit to the linear model shown in Table 2a.

This model is relatively straight-forward and "b" corresponds to a "base-line prevalence of chronic bronchitis" in eastern non-metropolitan populations. "R" is the change in bronchitis rates found in the West while "U" is the metropolitan effect. Notice that these two are treated as additive effects on the logistic scale. A significant interaction would have been equivalent to confounding in the data. Alternatively it would mean that at most two pollutants could be examined. The fourth parameter i the effect due to SO_2. There was no evidence of an interaction between SO_2 and either region or urbanity. The remaining four terms correspond to regional effects of particulates (PA) and sulphates (SU). Again there was no evidence of pollutant by urbanity interaction. Using the weighted least squares algorithm, KFF and GSK, leads to parameter estimates which are

1. Fully efficient for large samples (GSK),
2. Can incorporate either the simple or complex random sample design (KFF),
3. Computationally straight forward,
4. Robust against heteroscedasticity (GSK),
5. Available on any scale involving linearizable functions (KFF).

The resulting test statistics are shown for each sub-population or module in <u>Table 2b.</u> The test of fit of the model is non-significant in each module. The region effect is significant in 5 modules, its interaction in 3. Overall there is a significant pollution effect in four modules. This may be broken into its components; showing SO_2 in only one module, sulphates in four, and particulates in two. Moreover the separate regional sulphate and particulate effects are clearly necessary. One may then interpret these tests or more appropriately indices of significance by considering the corresponding estimates shown in <u>Table 2c.</u>

The effects indicate an increase in bronchitis if the estimate is positive. The region effect is generally small but dramatically reduced bronchitis among Western males who have never smoked. The persons in metropolitan areas have elevated rates. SO_2 has relatively little effect. Where the particulate effect is significant it is negative in the West and positive in the East. Conversely sulphates are positive in the West and negative in the East.

The next step in the analysis is to combine the effects across the modules. This was done following the algorithm of Freeman, Freeman and Brock (1977). It is entirely comparable to backwards elimination in regression analysis. This results in the model shown in Tables 3a to 3c. Based on the fit statistic it is evident that the model is quite acceptable. All of the parameters are nominally significant at the 0.05 level. However, if one adjusts the degrees of freedom to reflect the appropriate variation space (shown under total) only WPA becomes non-significant. Interpretations of the parameters are shown in Table 3b and the corresponding module parameter estimates are shown in Table 3c. Either these or the estimates in Table 3a. may be used to generate the approximate response surfaces.

Briefly the analysis indicates that in the West there is no sex differences among non-smokers. The urban dweller generally has an increased prevalence. There is no evidence in these data of an effect due to SO_2. In the West particulates have a small negative effect among non-smokers but sulphates clearly increase bronchitis for all groups. The eastern picture is basically reversed.

Thus in the East it appears that particulates are associated with increased bronchitis prevalence in all groups, and sulphates have an unexplained negative correlation with bronchitis.

ACKNOWLEDGEMENTS

The authors wish to thank Mrs. Bea Zinn and Miss Joann Raccio for their administrative support.

REFERENCES

Grizzle, S.E., Starmer, C.F., and Koch, G.G.(1969). Analysis of categorical data by linear models. Biometrics 25, 489-504.

Henley, N.S., Jain, S.C., and Wells, H.D. (1976). Relative importance of program input and environmental constraints to family planning programs in Haryana, India. Presented at the 1976 Annual Meeting of the Population Association of America (in Montreal).

Makuch, R.W., and Freeman, D.H. (1976). A multiple logit analysis of a family planning system. Soc. Stat. Sect. Proc. Am. Stat. Assoc. 572-577.

Cox, D.R. (1970). The Analysis of Binary Data. London: Methuen Co. LTD.

Koch, G.G., Freeman, D.H., and Freeman, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. Int. Stat. Rev. 43, 59-78.

Berman, M.D. (1976). The impact of sulphur dioxide pollution on chronic respiratory disease. Unpublished manuscript.

U. S. Environmental Protection Agency (1974). Health Consequences of Sulphur Oxides, Research Triangle Park, N.C.: National Environmental Research Center.

Freemen, D.H., Freeman, J.L., Brock, D.B. (1977). Modularization for the analysis of interactions in complex sample survey data. Proceedings of the 41st Session of the International Statistical Institute. To appear.

Table 2a. Model used within each gender - smoking module

$$X_w b = \begin{bmatrix} 1 & 1 & 1 & 10. & 88. & 0 & 3.7 & 0 \\ 1 & 1 & 1 & 18. & 84. & 0 & 4.7 & 0 \\ 1 & 1 & 1 & 32. & 50. & 0 & 8.6 & 0 \\ 1 & 1 & 1 & 92. & 70. & 0 & 15.0 & 0 \\ 1 & 1 & 0 & 10. & 50. & 0 & 3.3 & 0 \\ 1 & 1 & 0 & 26. & 45. & 0 & 4.9 & 0 \\ 1 & 1 & 0 & 67. & 115. & 0 & 7.3 & 0 \\ 1 & 1 & 0 & 177. & 65. & 0 & 7.2 & 0 \\ 1 & 1 & 0 & 374. & 102. & 0 & 11.3 & 0 \\ 1 & 0 & 1 & 30. & 0 & 41. & 0 & 10.0 \\ 1 & 0 & 1 & 174. & 0 & 84. & 0 & 8.6 \\ 1 & 0 & 1 & 247. & 0 & 108. & 0 & 14.8 \\ 1 & 0 & 0 & 13. & 0 & 38. & 0 & 5.8 \\ 1 & 0 & 0 & 14. & 0 & 63. & 0 & 7.8 \\ 1 & 0 & 0 & 4. & 0 & 48. & 0 & 6.8 \end{bmatrix} \begin{bmatrix} b \\ \\ R \\ \\ U \\ \\ SO_2 \\ \\ WPA \\ \\ EPA \\ \\ WSU \\ \\ ESU \end{bmatrix}$$

Table 2b. Analysis of variation within modules, Q-statistics

Source	df	Never Smoked		Once Smoked		Now Smokes	
		Female	Male	Female	Male	Female	Male
Model	7	47.00*	131.90*	13.46	80.81*	39.77*	20.56*
Region Total	3	19.52*	38.03*	7.96*	33.45*	7.47	7.85*
Interaction	2	13.47*	16.60*	7.76*	2.34	4.04	5.08
Urban Total	1	2.96	5.27*	2.12	7.64*	19.14*	1.13
Pollution Total	5	25.83*	30.20*	7.98	13.94*	9.04	12.41*
SO_2 Total	1	0.03	0.31	0.19	4.99*	0.80	0.40
Particulate Total	2	12.10*	7.97*	4.83	4.61	2.79	5.15
Interaction	1	10.21*	2.08	4.35*	2.22	2.78	3.99*
Sulphate Total	2	13.70*	16.79*	7.26*	0.06	3.55	6.60*
Interaction	1	12.18*	16.41*	7.01*	0.06	3.53	4.33*
Within Module Error	7	8.24	8.11	6.63	6.78	13.06	6.78
Total Variation	14	55.24*	140.01*	20.08	87.59*	52.84*	27.35*
Percent Explained		85.1	94.1	67.0	92.3	75.3	75.2

*Statistic excedes 95-th percentile of corresponding χ^2 distribution

Table 2c. Within module parameter estimates and estimated standard errors.

Module Label	Estimates and Standard Errors					
	Never Smoked		Once Smoked		Now Smokes	
	Female	Male	Female	Male	Female	Male
b	-3.51*	-3.36*	-3.10*	-2.42*	-2.26*	-2.03*
SE	0.48	0.43	0.70	0.41	0.26	0.25
R: Present if west	0.10	-1.88*	-0.32	-0.18	0.16	0.19
SE	0.65	0.83	0.93	0.71	0.34	0.31
U: Present if metro.	0.41	0.57*	0.47	0.63*	0.55*	0.12
SE	0.24	0.25	0.32	0.23	0.12	0.11
SO_2 (10^{-8} g/s)	0.03	0.11	-0.10	0.39*	0.07	-0.04
SE	0.15	0.20	0.22	0.17	0.08	0.07
Western Particulates	-1.31*	1.14	0.30	-1.89*	-0.19	0.12
(10^{-8} g/s) SE	0.62	0.86	1.08	0.89	0.35	0.31
Eastern Particulates	2.50*	2.93	2.49	-0.47	0.66	1.14*
(10^{-8} g/s) SE	1.20	1.08	1.31	0.83	0.50	0.51
Western Sulphates	6.86*	7.26*	5.26	0.94	1.65	4.26*
(10^{-8} g/s) SE	2.69	3.48	5.34	4.51	2.04	1.87
Eastern Sulphates	-17.66*	-16.35*	-18.88*	-0.59	-5.17	-3.27
(10^{-8} g/s) se	6.41	4.51	7.73	4.98	3.16	3.24

*Ratio of estimate squared to variance exceeds 95-th percentile of χ^2 distribution, df = 1.

Table 3a. Final across module analysis

Analysis of Variation Source/Label	Estimate	Standard Error	Degrees of Freedom Total	Net	Q
Model	———	———	47	8	915.66
b	-3.64	0.13	—	—	——
S	1.21	0.13	16	1	91.60
G	0.69	0.06	8	1	128.09
U	0.50	0.07	6	1	59.38
WPA ($\times 10^{-8}$ g/s)	-0.53	0.19	6	1	7.85
EPA ($\times 10^{-8}$ g/s)	0.94	0.12	6	1	57.80
WSU ($\times 10^{-8}$ g/s)	4.65	0.79	6	1	35.03
ESU ($\times 10^{-8}$ g/s)	-17.55	2.33	6	1	56.52
SESU($\times 10^{-8}$ g/s)	13.26	1.75	6	1	57.56
Error	Final Reduction			8	12.25
	Backwards Elimination			19	23.73
	Initial Model			12	5.19
	Within Modules			42	49.60
	Total Error			81	90.77
Total				89	1006.43

Table 3b. Interpretation of final model parameters

Label	Coefficient(s)	Interpretation of effect on bronchitis
b	1	baseline logit - prevalence of chronic bronchitis for western females who are non-metropolitan and have never smoked.
S	1	person now smokes
	0	otherwise
G	1	Eastern male or smoking western male
	0	otherwise
U	1	Metropolitan person other than smoking males
	0	otherwise
WPA	(Particulates x 10^{-8} g/s)	for western non-smokers
	0	othewise
EPA	(Particulates x 10^{-8} g/s) x 3	for eastern never smoked or female ex-smokers
	x 1	for eastern smoker or male ex-smoker
	x 0	otherwise
WSU	(Sulphates x 10^{-8} g/s)	if western person
	0	otherwise
ESU	(Sulphates x 10^{-8} g/s)	if eastern person except male ex-smokers
	0	otherwise
SESU	(Sulphates x 10^{-8} g/s)	if eastern smoker
	0	otherwise

Table 3c. Fitted within module parameter estimates and estimated standard errors based on final model.

	Estimates and Standard Errors					
Module Label	Never Smoked		Once Smoked		Now Smokes	
	Female	Male	Female	Male	Female	Male
b	-3.64	-2.95	-3.64	-2.95	-2.43	-1.73
SE	0.13	0.12	0.13	0.12	0.08	0.07
R: Present if west	0	-0.69	0	-0.69	0	0
SE	0	0.06	0	0.06	0	0
U: Present if metro.	0.50	0.50	0.50	0.50	0.50	0
SE	0.07	0.07	0.07	0.07	0.07	0
SO_2 (10^{-8} g/s)	0	0	0	0	0	0
SE	0	0	0	0	0	0
Western Particulates	-0.53	-0.53	-0.53	-0.53	0	0
(10^{-8} g/s)SE	0.19	0.19	0.19	0.19	0	0
Eastern Particulates	2.81	2.81	2.81	0.94	0.94	0.94
(10^{-8} g/s)SE	0.37	0.37	0.37	0.12	0.12	0.12
Western Sulphates	4.65	4.65	4.65	4.65	4.65	4.65
(10^{-8} g/s)SE	0.79	0.79	0.79	0.79	0.79	0.79
Eastern Sulphates	-17.55	-17.55	-17.55	0	-4.28	-4.28
(10^{-8} g/s)SE	2.33	2.33	2.33	0	1.09	1.09

COMMENTS ON THE ANOVA STRATEGY FOR THE COMPUTATION OF INTRACLASS RELIABILITY

Rob Selvage, Washington State University

The ANOVA strategy for the computation of intraclass reliability coefficients is well established (Hoyt, 1941; Ebel, 1951; Winer, 1971). But inherent to the name of the approach, the analysis of variance assumptions are being overlooked if not ignored.

ANOVA requires normality and homoscedasticity of error variances for the cell distributions For cases with less than severe deviations from these assumptions, conventional data transformation can be applied to the data. There is sufficient recovery of the assumptions to warrant using ANOVA for computing the reliability coefficients in such cases. For example, given the data at Table I, ANOVA intraclass reliability was .42 before data were squared to induce the needed assumptions. The squared data yield a .46 coefficient.

TABLE I

RATINGS OF FOUR ITEMS BY FIVE JUDGES
RATINGS NEARLY NORMAL

	JUDGES				
Item	I	II	III	IV	V
A	5	4	3	4	5
B	3	5	3	3	5
C	4	3	2	2	2
D	1	1	3	3	2

However, given a severely deviate data such as at Table II, the ANOVA strategy proves less than fruitful. The ANOVA coefficient for Table II data is .055; squareing this data yields only modest improvement with a .17 coefficient.

TABLE II

RATINGS OF FOUR ITEMS BY FIVE JUDGES
RATINGS SEVERELY NONNORMAL

	JUDGES				
Item	I	II	III	IV	V
A	2	2	3	2	2
B	2	2	2	2	2
C	2	2	2	2	1
D	1	2	2	2	2

Granted, inducing normality on this data set reflected an improvement in the magnitude of the coefficient, but observe the consistency of the scores in Table II. Should not the intraclass reliability coefficient be much larger than .17? Obviously, yes.

These two examples demonstrate (1) in cases with less than severe deviations from the ANOVA assumptions, conventional transformations can be applied with moderate success, and (2) in cases with severe deviations from the ANOVA assumptions, the strategy is not markedly improved by attempting to induce normality and the strategy falls short of reflecting consistency.

In the behavioral and social sciences research literature, ANOVA computed intraclass reliability is commonplace. It is frequently found in cases of judges' rating items such as in Tables I and II. The rating scales for this purpose are notorious for having restricted ranges on the values judges may use to rate items. The ratings of one item may easily be all nearly equal, such as the case in Table II. The ANOVA strategy fails to compute coefficients which reflect the consistency of ratings when ANOVA assumptions are grossly violated as in the case of virtually equal ratings (Table II). ANOVA requires a substantial nonzero between item variance to obtain a significant coefficient. Data sets such as Table II cannot produce this needed between item variance.

An alternate technique for the computation of intraclass reliability (Finn, 1970) is a ratio of observed variance to expected variance subtracted from one. For example, for a five point scale used by five judges to rate four items, each point in the scale is expected to be used four times. Thus, the expected variance is 2.0, via using

$$\sigma^2 = \frac{\Sigma(X-\mu)^2}{N}.$$

Applying

$$r = 1 - \frac{\text{observed variance}}{\text{expected variance}}$$

to the data set at Table I, r = .538 and to Table II, r = .925 coefficient. Certainly, this strategy more accurately reflected the consistency observed in Table II.

In following the above discussion, note that the underlying distribution is discrete rectangular. Likewise, if the data were assumed distributed continuous rectangular (uniform), the expected value of the variance would be computed by $\sigma^2 = \frac{(b-a)^2}{12}$, according to Hogg and Craig (1971) where b = 5 and a = 1. The expected variance is 1.333 rather than 2.0. Applying this to Tables I and II, yields r = .305 and r = .89 respectively.

For the rather common data set at Table I, four different coefficients are computed thus far. These are (1) ANOVA, r = .42 with original data, (2) ANOVA, r = .46 with normality induced, (3) Finn, r = .538 with discrete rectangular distribution and (4) r = .305 with continuous rectangular distribution assumed. Coefficients for Table II data are more unsettling. They are (1) ANOVA, r = .055, (2) ANOVA with normality induced, r =

.17, (3) Finn, r = .925 and (4) Finn with uniform distribution assumption, r = .89. Which assumptions and strategy should a researcher choose?

Most researchers would dispute the plausibility of judges equally likely utility of all possible points of a 5-point scale. They would argue that for such a scale the scores are more likely to be normally distributed about the middle score. Thus, the underlying distribution is normal and the data must be analyzed accordingly.

Others would argue that though the judges use only 5 points on the scale, these points are only representative of possible values along the continuum from one to five. The whole numbers are used simply to expedite the rating procedure. Thus, provided scores are considered equally likely, this supports the notion that the underlying distribution is continuous rectangular (uniform). Likewise, if the ratings are considered to have a central tendency, the underlying distribution is normal.

It appears that the decision of which assumptions and underlying distributions best fit the data is most critical in determining the intraclass reliability coefficient. Care must be taken to avoid a misapplication of a strategy to a particular data set. Such as is the case of the ANOVA being applied to the Table II data. It is important to emphasize, that once the assumptions are made, the subsequent coefficient should be reported and possibilities should not be juggled to obtain the most desirable one.

REFERENCES

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Finn, R.H. A note on estimating the reliability of categorical data. Educational and Psychological Measurement, 1970, 30, 71-76.

Hogg, R.V., and Craig, A.T. Introduction to Mathematical Statistics (3rd ed.), New York: Macmillan, 1971.

Hoyt, C. Test reliability estimated by analysis of variance. Psychometrika, 1941, 6, 153-160.

Winer, B.J. Statistical Principles in Experimental Design (2nd ed.), New York: McGraw-Hill, 1971.

PERFORMANCE OF THE TRADITIONAL F TESTS IN SPLIT-PLOT DESIGNS UNDER COVARIANCE HETEROGENEITY

Huynh Huynh, University of South Carolina

and

Leonard S. Feldt, University of Iowa

1. INTRODUCTION

In split-plot analyses of variance the traditional F tests for the treatment and interaction effects demand that the population covariance matrices exhibit a specific structure (Huynh and Feldt, 1970). When this requirement is not fulfilled, some distortion in the level of significance may be expected for these tests. Greenhouse and Geisser (1958), extending a result by Box (1954b), concluded that where the covariance matrices for the plots are equal, the traditional treatment and interaction mean square ratios (MSR) are approximately distributed as central F variates with reduced degrees of freedom. A correction factor, $\varepsilon \leq 1$, evaluated from the common population covariance matrix, may be used to ascertain the degree to which this matrix conforms to the required structure ($\varepsilon = 1$ for strict conformity). This observation, coupled with the simulation results of Collier et al. (1967), indicates that the traditional F tests in split-plot designs with identical covariance matrix will err on the liberal side, e.g., show a size that is larger than the nominal alpha.

In the present paper a theoretical solution is obtained for the problem of determining Type I error probabilities for the tests of the split-plot design. The problem is solved in its general form. That is, the sampling distributions of the mean square ratios for main effects and interaction are derived under any arbitrary set of covariance matrices for the main plots. This solution, coupled with formulas derived by Imhof (1962), makes it possible to determine the exact size of the traditional tests.

2. DISTRIBUTIONS OF THE RATIOS MSR_A AND MSR_{AB} IN THE SPLIT-PLOT DESIGN

Consider g independent k-component normal variates

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{ij} \cdots, x_{kj})',$$

$j = 1, \cdots g$

with mean vectors

$$\mu_j = (\mu_{1j}, \mu_{2j} \cdots, \mu_{ij}, \cdots, \mu_{kj})'$$

and non-singular covariance matrices Σ_j which need not be equal. Each of the k-components (first subscript) corresponds to a level of treatment category A; each of the g populations (second subscript) corresponds to a level of treatment category B (which is also referred to as "group" or "plot"). Thus, the measures under the levels of A are related; the measures under the levels of B are independent. For each population j a random sample of size n_j is drawn, whose members are denoted by

$$X_{js} = (x_{1js}, x_{2js}, \cdots, x_{ijs}, \cdots, x_{kjs})',$$

$s = 1, 2, \cdots, n_j$. The vector X_{js} can be conceptualized as the score vector of the sth member drawn at random from the jth plot.

Let $n = \sum_j n_j$ be the total number of cases (or observation vectors). The effect of the ith treatment of the A category and the interaction of the ith treatment with the jth plot are respectively $\mu_{i.} - \mu_{..}$ and $\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$. The dot (.) notation refers to weighted means. The null hypotheses of interest are

$$H_A: \mu_{i.} - \mu_{..} = 0 \text{ for all } i$$

$$H_{AB}: \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..} = 0 \text{ for all } i, j.$$

The sums of squares associated with the treatments, the interaction and the residual or within error are defined as

$$SS_A = \sum_{i=1}^{k} n(x_{i..} - x_{...})^2$$

$$SS_{AB} = \sum_{i=1}^{k} \sum_{j=1}^{g} n_j(x_{ij.} - x_{i..} - x_{.j.} + x_{...})^2$$

$$SS_{error(w)} = \sum_{j=1}^{g} \left[\sum_{i=1}^{k} \sum_{s=1}^{n_j} (x_{ijs} - x_{.js} - x_{ij.} + x_{.j.})^2 \right].$$

The mean square ratios normally used to test H_A and H_{AB} are

$$MSR_A = MS_A/MS_{error(w)} = (n - g)SS_A/SS_{error(w)}$$

$$MSR_{AB} = MS_{AB}/MS_{error(w)}$$

$$= (n - g)SS_{AB}/(g - 1)SS_{error(w)} .$$

To obtain the distribution for MSR_A and MSR_{AB}, let $D = I - \underline{1}\underline{1}'/k$ where I denotes an appropriate identity matrix and $\underline{1}$ is the vector having k components all equal to 1. It may be verified that

$$(1) \quad MSR_A = (n-g) \sum_{i=1}^{k-1} \nu_i \chi_i^2(1; \delta_i^2) / \sum_{j=1}^{g} \sum_{i=1}^{k-1} \lambda_{ji} \chi_{ji}^2 (n_j - 1)$$

where the ν_i's are the eigenvalues of $D \sum_{j=1}^{g} n_j \Sigma_j / n$, the λ_{ji}'s are those of matrices $D\Sigma_j$, and all of the chi-squares are independent. Moreover, the chi-squares in the numerator are central if and only if the hypothesis H_A is true.

A particular case of interest is represented by the situation in which all the covariance matrices Σ_j are equal to Σ. Then $\sum_{j=1}^{g} n_j \Sigma_j / n = \Sigma$ and $\lambda_{ji} = \nu_i$ for all j. Hence, for this case

$$(2) \quad MSR_A = (n-g) \sum_{i=1}^{k-1} \nu_i \chi_i^2(1; \delta_i^2) / \sum_{i=1}^{k-1} \nu_i \chi_i^2 (n-g)$$

Consider now two matrices. The first matrix, Σ^*, may be formed by g^2 submatrices. Those on the main "diagonal" are $\Sigma_1/n_1, \ldots, \Sigma_g/n_g$ and the others are all zero. The second matrix, G, is also formed by g^2 submatrices. Those on the "diagonal" are equal to $n_j(1-n_j/n)D$, $1 \leq j \leq g$. The submatrix on the ith "row" of the jth "column" is equal to $-n_j n_j D/n (1 \leq i \neq j \leq g)$. It may then be verified that

$$(3) \quad MSR_{AB} = \frac{n-g}{g-1} \sum_{i=1}^{(k-1)(g-1)} \gamma_i \chi_i^2 (1; \delta_i^2) \sum_{j=1}^{g} \sum_{i=1}^{k-1} \lambda_{ji} \chi_{ji}^2 (n_j - 1)$$

where the γ_i's are the positive eigenvalues of the matrix $\Sigma^* G$ and the chi-squares are independent. As before, the non-centrality parameters δ_i^2 are zero if and only if the hypothesis H_{AB} is true.

For the particular case in which all the covariance matrices Σ_j are equal to Σ, the positive eigenvalues of $\Sigma^* G$ are the eigenvalues (ν_i) of $D\Sigma$, each with order of multiplicity $(g - 1)$. Hence,

$$(4) \quad MSR_{AB} = (n-g) \sum_{i=1}^{k-1} \nu_i \chi_i^2(g-1; \delta_i^2) / (g-1) \sum_{i=1}^{k-1} \nu_i \chi_i^2 (n-g).$$

Formulas (1), (2), (3), and (4), coupled with computational techniques outlined in the next sections, make it possible to compute the probability that a mean square ratio exceeds the critical values of the traditional tests.

3. COMPUTING THE EIGENVALUES

The matrices whose eigenvalues govern the distribution of the mean square ratio are always of the form $(I - \underline{1}\underline{1}'/k)\Sigma$. Let $B = \Sigma^{-1}$ and $D = I - \underline{1}\underline{1}'/k$. Then $(I - \underline{1}\underline{1}'/k)\Sigma = DB^{-1}$. Here D and B are symmetric and B is positive definite. In the present study computation of the eigenvalues was performed via the IBM-supplied subroutine NROOT (1971). The obtained values are accurate up to probably the fifth decimal. This degree of accuracy is sufficient for the purposes under consideration here.

4. COMPUTING THE EXACT PROBABILITIES

The probabilities of the Type I error associated with the traditional tests of the split-plot design can always be written in the form $Pr(Q \geq 0)$ where $Q = \sum_{i=1}^{m} \alpha_i \chi_i^2(h_i)$, all the chi-squares being mutually independent. Imhof (1962) showed that

$$(5) \quad Pr(Q > 0) = 1/2 + \pi^{-1} \int_0^{\infty} \frac{\sin \theta(u)}{u\rho(u)} \, du$$

where $$\theta(u) = \sum_{i=1}^{m} [h_i \tan^{-1}(\alpha_i u)]/2$$

$$\rho(u) = \prod_{i=1}^{m} (1 + \alpha_i^2 u^2)^{h_i/4} .$$

He also showed that

$$\lim_{u \to 0} \sin\theta(u)/u\rho(u) = (\sum_{i=1}^{m} \alpha_i h_i)/2$$

and that $u\rho(u)$ increases monotonically toward infinity. This allows the numerical integration in (5) to be carried out only over the finite range $0 \leq u \leq U$. The upper limit U was set large enough so that the error due to the truncated interval of integration was sufficiently small. All the integrations were performed via the Gaussian quadrature with 32 points. In this scheme the integrating function was replaced by an appropriate polynomial of degree 63, and the integration was performed as if the function were the polynomial. This method of integration was carried out with the IBM subroutine DQG32 (1971). It was set in such a way

Table 1

Some Population Covariance
Matrices Used in the Study (k = 5)

Description	Elements*				
A, ε = .388	1.00				
Source: computer-simulated	.86	1.00			
	.96	.86	1.00		
	.64	.88	.66	1.00	
	.44	.77	.60	.91	1.00
B, ε = .420	1.00				
Source: computed from data in	.85	1.00			
Lindquist (1962, page 167)	.48	.32	1.00		
	.34	.47	.83	1.00	
	.83	.71	.88	.76	1.00
C, ε = .522	1.00				
Source: fictitious	.80	1.00			
	.60	.80	1.00		
	.40	.60	.80	1.00	
	.30	.40	.60	.80	1.00
D, ε = .752	1.00				
Source: Wechsler (1958, page	.81	1.00			
100, Table 20, Variables:	.74	.70	1.00		
Voc., Inf., Sim., BD, OA)	.53	.58	.52	1.00	
	.43	.45	.39	.61	1.00
E, ε = .831	1.00				
Source: Thurstone and	.62	1.00			
Thurstone (1938)	.62	.67	1.00		
	.54	.53	.62	1.00	
	.29	.38	.48	.62	1.00

*All correlations are rounded to the second decimal.

that all the reported probabilities were accurate up to the last tabulated decimal.

5. SITUATIONS CONSIDERED IN THE STUDY OF TYPE I ERROR

In the present study the number of treatments (A) was set at k = 5, and the number of main plots (B) at g = 3. The total number of sampling units was set at n = 18 and 33. It may be recalled that when the covariance matrices are equal, the distributions of the mean square ratios do not depend on the plot (group) sizes n_j per se, but only on their sum, n. It is interesting to note that when n increases indefinitely, each mean square ratio tends stochastically to a linear combination of chi-squares. Therefore, it should be expected that various probabilities associated with large values of n would not vary markedly.

To simplify the study, only covariance matrices with equal variances (1.0 in every case) were used in the study. Under this condition, the traditional tests are valid only when the covariances or correlations are equal. Five matrices with heterogeneous correlations were considered. These matrices had correction factors ε = .388, .420, .522, .752, and .831, respectively. They are displayed in Table 1. In other phases of the study symmetric matrices were employed. These matrices, designated S_ρ, had homogeneous variances of 1.0 and homogeneous correlations indicated by the subscript value. Thus, $S_{.30}$ represents a matrix with variances of 1.0 and all correlations equal to .30.

6. RESULTS FOR THE CASE OF EQUAL COVARIANCE MATRICES

The true probabilities of Type I error, computed as described earlier, are presented in Table 2. They suggest the following trends:

(a) The traditional tests always err on the liberal side, especially when ε and n are small, and α = 2.5 or 1 per cent. Increasing the sample size leads, in most cases, to a slight reduction in the actual probability of Type I error.

(b) Failure of the common covariance matrix to exhibit the required structure has less effect

Table 2

Exact Per Cents of Type I Error Associated with the Traditional Tests in the Split-Plot Design with Equal Σ

Matrix	ε	n	α(%) for Test of Treatment Effect				α(%) for Test of Interaction Effect			
			10	5	2.5	1	10	5	2.5	1
A	.388	18	14.80	10.46	7.54	5.00	16.97	12.08	8.85	5.80
		33	14.50	10.22	7.38	4.90	16.56	11.77	8.53	5.67
		∞	14.13	9.97	7.20	4.78	16.04	11.42	8.28	5.52
B	.420	18	14.79	10.19	7.17	4.60	16.65	11.60	8.21	5.29
		33	14.50	9.95	7.00	4.50	16.27	11.29	8.00	5.16
		∞	14.36	10.12	7.29	4.44	15.76	10.95	7.76	5.02
C	.552	18	13.02	8.40	5.55	3.29	14.40	9.35	6.18	3.64
		33	12.84	8.29	5.50	3.28	14.19	9.22	6.13	3.65
		∞	12.58	8.15	5.43	3.26	13.85	9.05	6.05	3.63
D	.752	18	11.40	6.60	3.91	2.01	12.10	7.04	4.16	2.11
		33	11.33	6.56	3.91	2.03	12.03	7.00	4.17	2.15
		∞	11.18	6.50	3.90	2.04	11.84	6.95	4.17	2.18
E	.831	18	10.86	6.02	3.40	1.64	11.33	6.30	3.56	1.70
		33	10.82	5.99	3.41	1.66	11.30	6.29	3.57	1.73
		∞	10.70	5.96	3.41	1.68	11.17	6.26	3.59	1.76

on the size of the test of H_A than on the size of the test of H_{AB}.

7. RESULTS FOR THE CASE OF UNEQUAL COVARIANCE MATRICES

Preliminary computation indicated that when equality of the covariance matrices does not hold, variation in the plot sizes and the range of the correlations play a major role. Therefore, this part of the study was subdivided into three phases. First, in order to assess the effect of unequal plot sizes, the covariance matrices were restricted to type S_ρ (for which $\varepsilon = 1$). Extreme cases were included to dramatize this effect. Next were considered matrices with wide ranges for the correlations. Finally, matrices with moderate ranges of correlations and different correction factors were considered.

(a) Effect of Unequal Plot Sizes. The data reported in Table 3 confirm the salutary effect of equal plot sizes for the test of interaction. Inequality of plot sizes has little effect on the test of treatment effects. However, variation in plot sizes may seriously invalidate the test of no interaction. The results for the test of H_A are consistent with those of the Box studies (1954a). Box found that, in the case of the completely randomized design, inequality of variance has little effect on the F test so long as the sample sizes are kept equal.

In view of these results, subsequent investigation was made only for the case of equal plot size.

(b) Effect of High Correlations. Exact probabilities of Type I error were also computed for experiments with matrices involving very high correlations. Matrices D, E, and $S_{.99}$ for the three plots was one such configuration. In these situations the probability of Type I error rose markedly above the nominal level, particularly for the test of interaction.

The advantage of the split-plot design over the factorial design depends on the size of the correlation between measures within plots. The higher the correlation, the

Table 3
Exact Per Cents of Type I Error Associated with the Traditional Tests in the Split-Plot Design with Unequal Σ_i: Effect of Unequal Plot Sizes

Matrix for Plot 1	Matrix for Plot 2	Matrix for Plot 3	Size for Plot 1	Size for Plot 2	Size for Plot 3	α(%) for Test of Treatment Effect 10	5	2.5	1	α(%) for Test of Interaction Effect 10	5	2.5	1
$S_{.10}$	$S_{.90}$	$S_{.90}$	11	11	11	10.78	5.61	2.95	1.28	13.91	8.60	5.41	2.99
			8	11	14	11.78	6.31	3.42	1.53	30.56	22.28	16.35	10.91
			11	14	8	10.78	5.61	2.95	1.28	13.91	8.60	5.42	2.98
			3	15	15	15.15	8.63	4.95	2.39	72.35	64.76	57.82	49.56
			29	2	2	8.56	4.10	1.98	0.75	0.00	0.00	0.00	0.00
$S_{.30}$	$S_{.50}$	$S_{.70}$	11	11	11	10.11	5.06	2.55	1.03	10.50	5.43	2.84	1.21
			8	11	14	10.36	5.22	2.64	1.07	15.87	9.00	5.12	2.43
			8	14	11	10.22	5.13	2.58	1.04	13.04	7.10	3.90	1.77
			3	15	15	10.62	5.38	2.74	1.12	22.94	24.27	8.86	4.70
			29	2	2	9.33	4.56	2.24	0.88	0.85	0.26	0.09	0.02
$S_{.40}$	$S_{.50}$	$S_{.60}$	11	11	11	10.05	5.02	2.51	1.01	10.16	5.11	2.58	1.05
			8	11	14	10.17	5.09	2.55	1.03	12.66	6.70	3.55	1.54
			14	11	8	9.95	4.95	2.47	0.99	8.05	3.84	1.85	0.71
			3	15	15	10.29	5.16	2.60	1.05	15.59	8.68	4.84	2.22
			29	2	2	9.62	4.74	2.35	0.93	3.18	1.24	0.49	0.15

smaller the residual error variance, and the greater is the power of the test. However, when the assumption about covariance matrices is not fulfilled (or only approximately so, as in the case of the matrices E with $\varepsilon = .831$ and D with $\varepsilon = .752$), high correlations may result in a much greater chance of Type I error than would be anticipated.

(c) Effect of Heterogeneity of the Correction Factors. The data reported in Table 4 reveal that departures from the nominal values of α become more serious as the correction factors ε decrease. The effect cannot be ignored when $\varepsilon < .75$. Deviations at $\alpha = 10$ or 5 per cent are not intolerably large when all of the $\varepsilon > .75$. Extremely heterogeneous covariance matrices (with ε in the neighborhood of .5 or .4) almost completely invalidate the traditional tests.

8. CONCLUDING REMARKS

Data are presented in this study describing the performance of the traditional F tests for the split-plot design when nonstandard conditions hold for the covariance matrices. In all situations under investigation, the test for interaction proved to be more vulnerable than the one for treatment effects, especially when the plot sizes are not equal. When heterogeneity of covariance matrices is suspected, or homogeneity appears to hold but $\varepsilon < .8$ for each matrix, multivariate procedures or approximate F tests should be considered. These give better control of Type I error (Arnold, 1973; Huynh and Feldt, 1976; Buynh, in press).

References

Arnold, S. F. (1973) Application of the theory of products of problems to certain patterned covariance matrices. Annals of Statistics, 1, 682-699.

Box, G. E. P. (1954a) Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effects of inequality of variance in the one-way classification. Annals of Mathematical Statistics. 25, 290-302.

Box, G. E. P. (1954b) Some theorems on quadratic forms applied in the study of analysis of variance problems. II. Effects of inequality of variance and of correlation between errors in the two-way classification. Annals of Mathematical Statistics. 25, 484-498.

Collier, R. O., F. B. Mandeville, G. K. & Hayes, T. F. (1967) Estimates of test sizes for several test procedures based on conventional variance ratios in the repeated measures design. Psychometrika, 32, 339-353.

Geisser, S. & Greenhouse, S. (1958) An extension of Box's results on the use of the F distribution in multivariate analysis. Annals of Mathematicsl Statistics, 29, 885-891.

Table 4

Exact Per Cents of Type I Error Associated with the Traditional Tests in the Split-Plot Design with Unequal Σ_i: Effect of Heterogeneity of the Correction Factors ε

Matrix and ε for Plot*			Size for Plot			α(%) for Test of Treatment Effect				α(%) for Test of Interaction Effect			
1	2	3	1	2	3	10	5	2.5	1	10	5	2.5	1
D	E	$S_{.50}$	6	6	6	10.52	5.58	3.00	1.34	10.97	5.92	3.23	1.47
.752	.831	1	11	11	11	10.47	5.53	2.98	1.34	10.93	5.88	3.22	1.48
D	E	E	6	6	6	10.93	6.10	3.48	1.70	11.52	6.48	3.70	1.80
.752	.831	.831	11	11	11	10.84	6.06	3.48	1.72	11.43	8.46	3.71	1.83
D	D	E	6	6	6	11.10	6.29	3.65	1.82	11.78	6.73	3.90	1.93
.752	.752	.831	11	11	11	11.00	6.25	3.64	1.84	11.67	6.69	3.91	1.97
B	C	C	6	6	6	12.09	7.23	4.43	2.38	13.74	8.62	5.52	3.12
.420	.522	.752	11	11	11	11.80	7.00	4.28	2.29	13.43	8.40	5.39	3.06
A	A	E	6	6	6	12.36	7.86	5.14	3.00	14.47	9.36	6.16	3.61
3.88	.388	.831	11	11	11	12.07	7.69	5.03	2.96	14.14	9.15	6.04	3.56
A	B	C	6	6	6	12.72	7.94	5.08	2.89	15.13	10.06	6.81	4.15
.388	.420	.522	11	11	11	12.26	7.57	4.81	2.72	14.65	9.68	6.55	4.00

*The average correlations are .76 for A, .69 for B, .61 for C, .58 for D and .54 for E.

Huynh, H. & Feldt, L. S. (1970) Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. Journal of American Statistical Association, 65. 1582-1589.

Huynh, H. & Feldt, L. S. (1976) Estimation of the Box correction for degrees of freedom from sample data in the randomized block and split-plot designs. Journal of Educational Statistics, 1, 69-82.

Huynh, H. (1978) Some approximate tests for repeated measurement designs. Psychometrika, in press.

Imhof, J. P. (1961) Computing the distribution of quadratic forms in normal variables. Biometrika, 48, 419-426.

IBM Application Program, System/360. (1971) Scientific subroutines package (360-CM-03X) Version III, Programmer's manual. White Plains, New York: IBM Corporation Technical Publication Department.

Lindquist, E. F. (1962) Design and analysis of experiments in psychology and education. Boston: Houghton Mifflin Co.

Thurstone, L. L. & Thurstone, T. F. (1938) Primary mental abilities, Psychometric Monograph, No. 1, University of Chicago Press.

Wechsler, D. (1958) The measurement and appraisal of adult intelligence. Baltimore: Williams and Wilkins.

MULTIVARIATE RATIO-TYPE ESTIMATORS

B. V. Sukhatme, Iowa State University
Lal Chand, J. N. K. V. V.

1. Introduction

Let a finite population consist of N distinct identifiable units U_i with values $x_{0i}, x_{1i}, \ldots, x_{\lambda i}$ $i = 1, 2, \ldots, N$ of the characteristics $X_0, X_1, \ldots, X_\lambda$. Consider the problem of estimating the population mean $\bar{x}_{ON} = \frac{1}{N}\sum_{i=1}^{N} x_{0i}$ when data on two or more auxiliary characteristics X_i $i=1, 2, \ldots, \lambda$ correlated with X_0 are available or can be obtained easily. In this situation, it is customary to use data on auxiliary characteristics to obtain ratio-type estimators of $\bar{x}_{ON}$. Several authors including Olkin [3], Raj [4], Rao and Mudholkar [5], Shukla [6], Singh [7] and [8], Smith [9] and Srivastava [10, 11, 12] have proposed ratio-type estimators utilizing data on several auxiliary variables. The estimators involve unknown weights which have to be estimated and assume knowledge of the population means of the auxiliary characteristics used. Clearly, none of the estimators proposed is satisfactory from the point of view of users and there is a need to investigate the matter further. The object of this paper is to present ratio-type estimators based on two or more auxiliary characteristics which do not involve unknown weights and at the most assume knowledge of the population mean of the auxiliary characteristic least correlated with X_0 along with appropriate expressions for bias and mean square error. Almost unbiased ratio-type estimators are also developed and a discussion is given concerning the efficiency of these estimators.

2. Multivariate Ratio-type Estimators

Let ρ_{0t} denote the correlation coefficient between X_0 and X_t. We shall assume that for $i < j$ $\rho_{0j} < \rho_{0i}$. We shall first consider the case when $\lambda = 2$ and assume three phase simple random sampling without replacement in which n_2 units are drawn from N in the first phase to observe X_2, a sub-sample of n_1 units is drawn from n_2 in the second phase to observe X_1 and a sub-sample of n_0 units is drawn from n_1 in the final phase to observe X_0. Let $\bar{x}_{tm}$ denote the sample mean based on m units corresponding to the characteristic X_t.

If $\bar{x}_{2N}$, the population mean of the characteristic X_2 is unknown, the ratio-type estimator of $\bar{x}_{ON}$ based on the use of two auxiliary variables X_1 and X_2 is defined as

$$t_{2d} = \frac{\bar{x}_{On_0}}{\bar{x}_{1n_0}} \frac{\bar{x}_{1n_1}}{\bar{x}_{2n_1}} \bar{x}_{2n_2} \qquad (2.1)$$

If $\bar{x}_{2N}$ is known, the ratio-type estimator of $\bar{x}_{ON}$ is defined as

$$t_2 = \frac{\bar{x}_{0n_0}}{\bar{x}_{1n_0}} \frac{\bar{x}_{1n_1}}{\bar{x}_{2n_1}} \bar{x}_{2N} \qquad (2.2)$$

The multivariate ratio-type estimator corresponding to λ auxiliary variables is now obvious. If $\bar{x}_{\lambda N}$ is not known, the estimator is defined as

$$t_{\lambda d} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda n_\lambda} \qquad (2.3)$$

If $\bar{x}_{\lambda N}$ is known, the estimator is defined as

$$t_\lambda = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda N} \qquad (2.4)$$

It is assumed that sampling is carried out in $(\lambda + 1)$ phases with simple random sampling without replacement in each of the phases and may be diagramatically described as follows.

$$N \xrightarrow{SRS} n_\lambda (X_\lambda) \xrightarrow{SRS} n_{\lambda-1} (X_{\lambda-1}) \cdots \cdots \xrightarrow{SRS} n_1 (X_1) \xrightarrow{SRS} n_0 (X_0)$$

where at a particular phase n_t denotes the sample size to be drawn at random from n_{t+1} and X_t denotes the characteristic to be observed on n_t units.

3. Bias and Mean Square Error of the Multivariate Ratio-type Estimators

Consider first the estimator $t_{\lambda d}$. By definition

$$\left.\begin{aligned} \text{Bias}\,(t_{\lambda d}) &= E\,(t_{\lambda d}) - \bar{x}_{ON} \\ \text{and MSE}\,(t_{\lambda d}) &= E(t_{\lambda d} - \bar{x}_{ON})^2 \end{aligned}\right\} \qquad (3.1)$$

It is not possible to obtain exact expressions for the bias and mean square error. However, expressing $t_{\lambda d}$ as a power series in powers of $\delta \bar{x}_{in_i} = \dfrac{\bar{x}_{in_i} - \bar{x}_{iN}}{\bar{x}_{iN}}$, ignoring terms of order higher than two and taking expectation term by term, we obtain

$$\text{Bias}_1\,(t_{\lambda d}) = \bar{x}_{ON} \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(C_{x_i}^2 - C_{x_i x_0}\right) \qquad (3.2)$$

and

$$\text{MSE}_1\,(t_{\lambda d}) = \bar{x}_{ON}^2 \left\{\left(\frac{1}{n_0} - \frac{1}{N}\right) C_{x_0}^2 - \sum_{i=1}^{\lambda}\left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(2C_{x_0 x_i} - C_{x_i}^2\right)\right\} \qquad (3.3)$$

where $$C_{x_i}^2 = \frac{S_{x_i}^2}{\bar{x}_{iN}^2}, \quad C_{x_i x_0} = \frac{S_{x_i x_0}}{\bar{x}_{iN}\bar{x}_{0N}} \qquad (3.4)$$

with $$S_{x_i}^2 = \sum_{1}^{N} (x_{it} - \bar{x}_{iN})^2/(N-1)$$

and $$S_{x_i x_0} = \sum_{1}^{N} (x_{it} - \bar{x}_{iN})(x_{0t} - \bar{x}_{0N})/(N-1) \qquad (3.5)$$

Following the procedure of David and Sukhatme [1], it can now be shown that

$$\left.\begin{aligned} |\,\text{Bias}\,(t_{\lambda d}) - \text{Bias}_1\,(t_{\lambda d})\,| &\le \frac{A_1}{n^2} \\ |\,\text{MSE}\,(t_{\lambda d}) - \text{MSE}_1\,(t_{\lambda d})\,| &\le \frac{A_2}{n^2} \end{aligned}\right\} \qquad (3.6)$$

where A_1 and A_2 are finite. It follows that (3.2) and (3.3) provide first order approximations to the bias and mean square error of the estimator $t_{\lambda d}$. In a similar manner, it can be shown that first order approximations to the bias and mean square error of t_λ are

$$\text{Bias}_1\,(t_\lambda) = \bar{x}_{0N}\left\{\sum_{i=1}^{\lambda-1}\left(\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(C_{x_i}^2 - C_{x_0 x_i}\right) + \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N}\right)\left(C_{x_\lambda}^2 - C_{x_0 x_\lambda}\right)\right\}$$

and

$$\text{MSE}_1\,(t_\lambda) = \bar{x}_{0N}^2\left\{\left(\frac{1}{n_0} - \frac{1}{N}\right) C_{x_0}^2 - \sum_{i=1}^{\lambda-1}\frac{1}{n_{i-1}} - \frac{1}{n_i}\right)\left(2C_{x_0 x_i} - C_{x_i}^2\right) - \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N}\right)\left(2C_{x_0 x_\lambda} - C_{x_\lambda}^2\right)\right\} \qquad (3.8)$$

Higher order approximations to the bias and mean square error have been obtained by Lal Chand [2]. However, the expressions are complicated and will not be presented.

If the population is assumed to be so large that finite correction factors can be ignored and is symmetrically distributed about its means, the expressions simplify considerably. In particular it can be shown that the second order approximations to the bias and mean square error for $\lambda = 2$ are given by

$$\text{Bias}_2\,(t_{2d}) = \text{Bias}_1\,(t_{2d})\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{C_{x_2}^2}{n_1}\right] \qquad (3.9)$$

$$\text{MSE}_2\,(t_{2d}) = \text{MSE}_1\,(t_{2d})\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{3C_{x_2}^2}{n_1} + \frac{3C_{x_2}^2}{n_2}\right] \qquad (3.10)$$

and

$$\text{Bias}_2\,(t_2) = \text{Bias}_1\,(t_2)\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{C_{x_2}^2}{n_1}\right] \qquad (3.11)$$

$$\text{MSE}_2\,(t_2) = \text{MSE}_1\,(t_2)\left[1 + \frac{3C_{x_1}^2}{n_0} + \frac{3C_{x_2}^2}{n_1}\right] \qquad (3.12)$$

where the first order approximations are obtained from the expressions (3.2), (3.3), (3.7) and (3.8) by taking $\lambda = 2$.

4. Almost Unbiased Multivariate Ratio-type Estimators

In this section, we shall present multivariate analogs of the ratio-type estimators presented in section 3 which are almost unbiased in the sense that the bias to the first order of approximation is zero. The estimators corresponding to $t_{\lambda d}$ and t_λ are

$$t_{\lambda dM} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda n_\lambda} \left[1 - \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) \left\{ \frac{s^2_{x_i}}{\bar{x}^2_{in_0}} - \frac{s_{x_i x_0}}{\bar{x}_{in_0} \bar{x}_{0n_0}} \right\} \right] \quad (4.1)$$

and

$$t_{\lambda M} = \prod_{i=1}^{\lambda} \left[\frac{\bar{x}_{i-1, n_{i-1}}}{\bar{x}_{i, n_{i-1}}} \right] \bar{x}_{\lambda N} \left[1 - \sum_{i=1}^{\lambda-1} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) \left\{ \frac{s^2_{x_i}}{\bar{x}^2_{in_0}} - \frac{s_{x_i x_0}}{\bar{x}_{in_0} \bar{x}_{0n_0}} \right\} - \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N} \right) \left\{ \frac{s^2_{x_\lambda}}{\bar{x}^2_{\lambda n_0}} - \frac{s_{x_\lambda x_0}}{\bar{x}_{\lambda n_0} \bar{x}_{0n_0}} \right\} \right] \quad (4.2)$$

Expressing $t_{\lambda dM}$ and $t_{\lambda M}$ as power series in powers of $\delta\bar{x}_{in_i}$, ignoring powers of order higher than two and taking expectation term by term, it can be verified that to the first order of approximation $t_{\lambda dM}$ and $t_{\lambda M}$ are almost unbiased estimators of $\bar{x}_{0N}$. Proceeding in a similar manner and evaluating their mean square errors, it can be seen that to the first order of approximation $t_{\lambda dM}$ and $t_{\lambda M}$ have the same mean square errors as $t_{\lambda d}$ and t_λ respectively. We have thus proved the following result

Theorem 4.1 The estimators $t_{\lambda dM}$ and $t_{\lambda M}$ are almost unbiased estimators of $\bar{X}_{0N}$.

Further, to the first order of approximation

$$MSE_1 (t_{\lambda dM}) = MSE_1 (t_{\lambda d})$$

and

$$MSE_1 (t_{\lambda M}) = MSE_1 (t_\lambda)$$

where $MSE_1 (t_{\lambda d})$ and $MSE_1 (t_\lambda)$ are given by (3.3) and (3.8) respectively.

5. Comparison of Estimators

For the purpose of comparison, we shall consider the mean square errors of the appropriate estimators to the first order of approximation only. Since $t_{\lambda d}$ and t_λ have the same mean square errors as $t_{\lambda dM}$ and $t_{\lambda M}$ to the first order of approximation, it is enough to consider $t_{\lambda dM}$ and $t_{\lambda M}$.

We have

$$V(\bar{x}_{0n_0}) - MSE_1 (t_{\lambda dM}) = \bar{x}^2_{0N} \sum_{i=1}^{\lambda} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) (2C_{x_0 x_i} - C^2_{x_i}) \quad (5.1)$$

and

$$V(\bar{x}_{0n_0}) - MSE_1 (t_{\lambda M}) = \bar{x}^2_{0N} \left\{ \sum_{i=1}^{\lambda-1} \left(\frac{1}{n_{i-1}} - \frac{1}{n_i} \right) (2C_{x_0 x_i} - C^2_{x_i}) + \left(\frac{1}{n_{\lambda-1}} - \frac{1}{N} \right) (2C_{x_0 x_\lambda} - C^2_{x_\lambda}) \right\} \quad (5.2)$$

It follows that if

$$\rho_{0i} > \frac{1}{2} \frac{C_{x_i}}{C_{x_0}} \quad \text{for } i=1, 2, \ldots, \lambda \quad (5.3)$$

then both the estimators $t_{\lambda dM}$ and $t_{\lambda M}$ will be more efficient than the simple mean estimator $\bar{x}_{0n_0}$ which does not use auxiliary data on any of the 0 variables.

Further, we have

$$MSE_1 (t_{\lambda-1\, dM}) - MSE_1 (t_{\lambda dM}) = \left(\frac{1}{n_{\lambda-1}} - \frac{1}{n_\lambda} \right) (2C_{x_0 x_\lambda} - C^2_{x_\lambda}) \quad (5.4)$$

It follows that if inequality (5.3) is true

$$MSE_1 (t_{\lambda dM}) < MSE_1 (t_{\lambda-1\, dM}) \quad (5.5)$$

for all values of λ.

It can also be seen that if inequality (5.3) is true, then

$$MSE_1 (t_{\lambda M}) < MSE_1 (t_{\lambda dM}) \quad (5.6)$$

Combining all these results, we have the following

Theorem 5.1 If

$$\rho_{0i} > \frac{1}{2} \frac{C_{x_i}}{C_{x_0}} \quad \text{for } i=1, 2, \ldots, \lambda, \text{ then}$$

$$MSE_1 (t_\lambda) < MSE_1 (t_{\lambda d}) < MSE_1 (t_{\lambda-1\, d}) \cdots$$

$$< MSE_1 (t_{1d}) < V(\bar{x}_{0n_0})$$

and

$$MSE_1 (t_{\lambda M}) < MSE_1 (t_{\lambda dM}) < MSE_1 (t_{\lambda-1\, dM}) \cdots$$

$$< MSE_1 (t_{1dM}) < V(\bar{x}_{0n_0})$$

Finally, we shall compare t_λ for $\lambda = 2$ with the ratio estimator

$\hat{\bar{x}}_{0N} = \bar{x}_{0n_0} \frac{\bar{x}_{2N}}{\bar{x}_{2n_0}}$. Then noting that

$$MSE_1(\hat{\bar{x}}_{0N}) = \left(\frac{1}{n_0} - \frac{1}{N}\right) \bar{x}_{0N}^2 \left[C_{x_0}^2 + C_{x_2}^2 - 2C_{x_0x_2}\right]$$

it can be seen that $MSE_1(t_2) < MSE_1(\hat{\bar{x}}_{0N})$ provided

$$\sum_1^N \left(x_{0i} - \frac{\bar{x}_{0N}}{\bar{x}_{2N}} x_{2i}\right)^2 > \sum_1^N \left(x_{0i} - \frac{\bar{x}_{0N}}{\bar{x}_{1N}} x_{1i}\right)^2$$

The above condition would be always true provided X_1 is a better auxiliary variable than X_2 for ratio method of estimation as assumed in this paper.

We have seen that the ratio-type estimator based on λ auxiliary variables is more efficient than the one based on $(\lambda-1)$ auxiliary variables provided

$$\rho_{0\lambda} > \frac{1}{2} \frac{C_{x_\lambda}}{C_{x_0}} .$$

Although, this result is of considerable value, what is more interesting is to know whether the reduction in variance is worth the extra cost required to observe the additional auxiliary variable. For the sake of simplicity, we shall consider the case $\lambda = 2$ and choose that estimator for which the mean square error is minimum when the total cost of collecting data cannot exceed a specified amount C_0.

Consider a simple cost function of the form

$$C = c_0 n_0 + c_1 b_1 + c_2 n_2 \qquad (5.7)$$

where c_i is the cost per unit of observing the characteristic X_i i=0, 1, 2. We shall now determine n_i such that $MSE_1(t_{2dM})$ is minimum subject to the condition that $C \leq C_0$. It can be seen that the optimal values of n_i to achieve this are given by

$$\sqrt{\frac{Q_4/c_0}{n_0}} = \sqrt{\frac{Q_5/c_1}{n_1}} = \sqrt{\frac{Q_3/c_2}{n_2}} = \frac{\sqrt{Q_4 c_0} + \sqrt{Q_5 c_1} + \sqrt{Q_3 c_2}}{C_0} \qquad (5.8)$$

where

$$Q_3 = 2C_{x_0x_2} - C_{x_2}^2$$

$$Q_4 = C_{x_0}^2 - 2C_{x_0x_1} + C_{x_1}^2 \qquad (5.9)$$

and

$$Q_5 = C_{x_2}^2 - 2C_{x_0x_2} + 2C_{x_0x_1} - C_{x_1}^2$$

For optimal choice of the n_i, the optimal mean square error of the estimator t_{2dM} is given by

$$\left[MSE_1(t_{2dM})\right]_{opt} = \frac{\left[\sqrt{Q_4 c_0} + \sqrt{Q_5 c_1} + \sqrt{Q_3 c_2}\right]^2 \bar{x}_{0N}^2}{C_0} \qquad (5.10)$$

In a similar manner, it can be seen that

$$\left[MSE_1(t_{1dM})\right]_{opt} = \frac{\left[\sqrt{Q_4 c_0} + \sqrt{Q_1 c_1}\right]^2 \bar{x}_{0N}^2}{C_0} \qquad (5.11)$$

where

$$Q_1 = 2C_{x_0x_1} - C_{x_1}^2 \qquad (5.12)$$

and

$$V(\bar{x}_{0n_0})_{opt} = \frac{C_{x_0}^2 \bar{x}_{0N}^2 c_0}{C_0} \qquad (5.13)$$

Comparing the mean square errors, it can be seen that

$$MSE_1(t_{2dM})_{opt} < MSE_1(t_{1dM})_{opt} < V(\bar{x}_{0n_0})_{opt}$$

$$\text{if } \frac{c_2}{c_1} < \frac{\left(\sqrt{Q_1} - \sqrt{Q_5}\right)^2}{Q_3} \text{ and } \frac{c_1}{c_0} < \frac{\left(C_{x_0} - \sqrt{Q_4}\right)^2}{Q_1} \qquad (5.14)$$

Since $t_{\lambda dM}$ and $t_{\lambda d}$ have the same mean square error to the first order of approximation, it follows that

$$MSE_1(t_{2d})_{opt} < MSE_1(t_{1d})_{opt} < V(\bar{x}_{0n_0})_{opt}$$

provided (5.14) is true.

6. Numerical Illustration

For the purpose of illustration, we shall consider the census data relating to 99 counties of Iowa. The three characteristics we shall consider are

X_0: Bushels of apples harvested in 1964

X_1: Apple trees of bearing age in 1964

X_2: Bushels of apples harvested in 1959

For this population, we have

$$\bar{x}_{0N} = .293458 \times 10^4 \quad \bar{x}_{1N} = .103182 \times 10^4$$

$$\bar{x}_{2N} = .365149 \times 10^4$$

$$\rho_{x_0x_1} = .93 \quad \rho_{x_0x_2} = .84 \quad \rho_{x_1x_2} = .77$$

$$C_{x_0}^2 = .402004 \times 10^1 \quad C_{x_1}^2 = .255280 \times 10^1$$

$$C_{x_2}^2 = .209379 \times 10^1$$

$$C_{x_0x_1} = .297075 \times 10^1 \quad C_{x_0x_2} = .244329 \times 10^1$$

$$C_{x_1x_2} = .177110 \times 10^1$$

For the purpose of comparing the different estimators, we shall assume that we have a large population with population parameters as given above. Further, we shall take

$n_0 = 30$ $n_1 = 60$ and $n_2 = 120$

The relevant results for comparing the different estimators are given in Table 1 below.

As is to be expected, the ratio-type estimator t_{2dM} based on two auxiliary variables is the most efficient of all the three estimators, the gain in efficiency over t_{1dM} based on one auxiliary variable being 40% while that over the mean estimator is 139%.

7. References

[1] David, I. P. and Sukhatme, B. V. 1974, On the bias and mean square error of the ratio estimator, Journal of the American Statistical Association, 69, 404-466.

[2] Lal Chand 1975, Some ratio-type estimators based on two or more auxiliary variables. Ph. D. thesis, Iowa State University, Ames, Iowa.

[3] Olkin, I. 1958, Multivariate ratio estimation for finite populations, Biometrika, 45, 154-165

[4] Raj, D. 1965, On a method of using multi-auxiliary information in sample surveys, Journal of the American Statistical Association, 60, 270-277.

[5] Rao, P. S. R. S. and Mudholkar, G. S. 1967, Generalized Multivariate estimator for the mean of finite populations, Journal of the American Statistical Association, 62, 1009-1012.

[6] Shukla, G. K. 1966, An alternative multivariate ratio estimate for finite population, Calcutta Statistical Association Bulletin, 15, 127-134.

[7] Singh, M. P. 1967, Multivariate product method of estimation for finite populations, Journal of the Indian Society of Agricultural Statistics, 19, 1-10.

[8] Singh, M. P. 1967, Ratio cum product method of estimation, Metrika, 12, 34, 42.

[9] Smith, T. M. F. 1966, Ratio of ratios and their applications, Journal of the Royal Statistical Society, Series A, 129, 531-533.

[10] Srivastava, S. K. 1965, An estimation of the mean of a finite population using several auxiliary variables, Journal of the Indian Statistical Association, 3, 189-194.

[11] Srivastava, S. K. 1967, An estimator using auxiliary information in sample surveys, Calcutta Statistical Association Bulletin, 16, 121-132.

[12] Srivastava, S. K. 1971, A generalized estimator for the mean of a finite population using multi-auxiliary information, Journal of the American Statistical Association, 66, 404-407.

Table 1

Estimator	Mean Square Error	Relative Efficiency w. r. t. $\bar{x}_{0n_0}$	Relative Efficiency w. r. t. t_{1dM}
$\bar{x}_{0n_0}$	$.115399 \times 10^7$	1	0.59
t_{1dM}	$.676577 \times 10^6$	1.70	1
t_{2dM}	$.481886 \times 10^6$	2.39	1.40

THE EFFECT OF AGGREGATION ON THE INTERPRETATION OF R^2

James E. Prather, Georgia State University

The use of aggregate data in regression analysis is pervasive in such fields of study as public policy, demography, political science, economics, and sociology. For several decades, a debate on the proper specification of aggregate models, so that inferences could be made about micro-level relationships from the macro-level estimators, has permeated literature in these fields. For most investigators, this question remains unresolved or insoluble, though there have been continuous refinements of techniques designed to mitigate aggregation problems (Irwin & Lichtman, 1976; Smith, 1977).

This paper does not focus upon macro to micro inference directly, rather it is concerned with the interpretation of the standard measure of goodness-of-fit for regression analysis--the multiple-correlation-squared (R^2). The importance of the R^2 as a test statistic is the rationale for exploring its interpretation when using macro-level data for analyses employing least squares regression. However, it is acknowledged that it is not possible to divorce substantive problems of model formation from the methodological questions concerning technique. Thus, a review of previous work on aggregate allows one to view the question holistically, rather than as solely a problem of calculation or reading a computer printout.

The previous works on analyzing grouped data can for heuristic purposes be divided into two separate development paths. The two perspectives can be illustrated by the seminal work of Robinson (1950) in sociology and of Prais and Aitchison (1954) in economics. As has been previously noted, Robinson's "ecological correlation" approach and the grouping in linear models approach of Prais and Aitchison complement each other. A review of the aggregation issue from these two perspectives will be presented in the next two sections.

The importance of the R^2 is that it is often employed as a measure of the power and amount of explanatory worth of a particular specification. Even though this paper does not focus on model building, the use of R^2 in model selection with aggregate data does warrant considering specification impact on R^2.

Analysis of Covariance Approach to Aggregation

The analysis of variance method is illustrated by partitioning the sum of squares about the mean for Y (the dependent variable) into "explained" sums of squares and residual sum of squares. Following the notation of Johnston (1972:192-207) a simple model is defined as

$$y = X\beta + u \tag{1}$$

Where the sample y is a column vector (n x 1) of micro-level observations composed of p sub-vectors --i.e., the groups. The independent variables are the X matrix (n x k) divided into p groups and the first column is all ones to allow a constant term, while β is a vector (k x 1) of the estimators. The vector u contains stochastic noise values where $E(u)=0$. To incorporate the possible effect of the p groups, then an expanded model is

$$y = D\alpha + X\beta + u, \tag{2}$$

which allows the p groups to have different constant terms, thus α is a vector of (p - 1) elements. The D' matrix is of dummy variables with order (Mp x [p-1]), where $M = \sum_{i=1}^{p} m_i$; is the sum of the number of observations in each p, for instance:

$$D' = \begin{matrix} 1\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0\ 1\ 1\ 1\ 1\ 0\ 0\ 0 \\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1\ 1 \end{matrix} \tag{3}$$

Remembering that D has p groups, with each p having m elements. To estimate (1) above, start with

$$y = X\hat{\beta} + s, \tag{4}$$

which can be estimated by

$$\hat{\beta} = (X'X)^{-1}X'y, \tag{5}$$

where s gives the least square residuals. An additional relationship may be derived as,

$$y'y = \hat{\beta}'X'y + s's \tag{6}$$

Returning to (2) above, the estimation of

$$y = D\hat{\hat{\alpha}} + X\hat{\hat{\beta}} + e \tag{7}$$

becomes

$$\begin{matrix} \hat{\hat{\alpha}} \\ \hat{\hat{\beta}} \end{matrix} = \begin{matrix} D'D & D'X \\ X'D & X'X \end{matrix}^{-1} \begin{matrix} D'y \\ X'y \end{matrix} \tag{8}$$

and from (6) above

$$y'y = \hat{\alpha}'D'y + \hat{\hat{\beta}}'X'y + e'e. \tag{9}$$

The e vector contains residuals for (7).

To calculate the R^2 for analysis of the covariance problem, it is necessary to define (Thiel, 1971, p. 176)

$$1-R^2 = \frac{e'e}{y'Ay} \tag{10}$$

where

$$A = I - \frac{1}{N} V V' \tag{11}$$

with V a vector n ones. The A matrix is to transform to deviations from the mean.

If a standard analysis of covariance were desired, the terms given in Table 1 would be the appropriate residual sum of squares to use for an F-test after converting by degrees of freedom to determine mean squares. However, our interest is in the R^2's that would be associated with the differing levels. The micro-level R^2 is for the "Total" formula. Compare this to the macro-level or aggregate R^2 which has an additional factor of 'D'y -- indicating that the value of $\hat{\alpha}$ vector would inflate the R^2 to the extent that it is related to y. When $\hat{\bar{\alpha}}$ is vector of zeros or near zeros it could be concluded that the grouping factor had no independent effect on the dependent variables and the between groups R^2 would equal the total R^2. To restate the above, if the grouping is random, then the between groups R^2 is an unbiased estimate of the total R^2 -- though not as efficient as the total R^2 estimate (Cramer, 1964). The experimental statistician would note the treatment (i.e., groups) had no significant effect. There are undoubtedly many investigators using aggregated data whose research would be much easier if the grouping was random. Grunfeld and Griliches (1960) noted the phenomenon of the higher R^2 that was often found with grouped data and referred to it as a "synchronization" effect. As a historical note, Gehkel and Bichel (1934), Thorndike (1939), and Yule and Kendall (1950) observed the same problem. At the time there was no clear explanation except the intuitive one that "grouping" on substantive factors caused this to happen. The formula in Table 1 clearly shows that what is happening is that the additional variance is accounted for by the grouping estimators. Thus, the gain in the R^2 is not due to better data but simply the contribution of the grouping scheme -- D -- and not to the variables of interest in aggregate analysis -- the X matrix.

The Generalized Least-Squares Approach to Aggregation

In this section, if we start with (1) of the previous section the grouping of observations into p groups and taking means yields (Johnston, pp. 228-241):

$$\bar{y} = \bar{X}\beta + \bar{u} \tag{12}$$

Then the ungrouped data are related to the aggregated in these forms,

$$\bar{y} = Gy \tag{13}$$

$$\bar{X} = GX \tag{14}$$

$$\bar{u} = Gu \tag{15}$$

with G as the grouping matrix of (m x n). The form of G is, for instance,

$$G = \begin{matrix} 1/1 & 1/1 & 0 & 0 & 0 & 0 \ldots 0 \\ 0 & 0 & 1/2 & 1/2 & 1/2 & 1/2 \ldots 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \ldots 1/p \end{matrix} \tag{16}$$

While $E(\bar{u}) = 0$ it is also noted that

$$E(\bar{u}\bar{u}') = \sigma^2 I \tag{17}$$

which means that the estimators will be unbiased but inefficient. However, it is the case that

$$E(\bar{u}\bar{u}') = \sigma^2 GG' \tag{18}$$

which is efficient. To estimated B, the generalized least squares is

$$b = [\bar{X}'(GG')^{-1}\bar{X}]^{-1}\bar{X}'(GG')^{-1}\bar{y} \tag{19}$$

and

$$var(b) = \sigma^2[\bar{X}'(GG')^{-1}\bar{X}]^{-1} \tag{20}$$

Here, generalized least squares overcomes the heteroscedastic problem (17) by inserting the grouping factor G in (18). The expression $(GG')^{-1}$ is actually a weighting matrix which contains the numbers in each group. Note that the generalized least squares estimates are not as efficient as the ungrouped ones.

The R^2 question may now be approached, when recalling from Table 1 that R^2 for equation (1) is

$$1 - R^2 = \frac{s's}{y'Ay} \tag{21}$$

by simple reexpression. But what about the R^2 for the groups values? The R^2 for equation (12) could be of the form

$$1-\bar{R}^2 = \frac{\bar{e}'(GG')^{-1}\bar{e}}{\bar{y}'(GG')^{-1}A\bar{y}} \tag{22}$$

where

$$\bar{e}'(GG')^{-1}\bar{e} = \bar{y}(GG')^{-1}\bar{y} - b'\bar{X}(GG')^{-1}\bar{y} \tag{23}$$

By definition the sums of squares may be partitioned:

$$y'Ay = \bar{y}'(GG')^{-1}\bar{A}\bar{y} + y^{*}'y^{*} \tag{24}$$

with

$$y^{*} = y - DGy \tag{25}$$

referring back to equation (3). Thus, it must be the case that

$$y'Ay \geq \bar{y}'(GG')^{-1}\bar{A}\bar{y} \tag{26}$$

and we can see that the reduction of the denominator for between groups sum of squares is again a function of D -- the relationship of the grouping factor with y. As the association of y with D increases, the between sum of squares decreases - i.e., $\bar{R}^2$ for between groups must increase.

An Example of the Effect of Aggregation on R^2

Of substantive interest in political sociology has been voter participation in the electoral process. In light of the traditional democratic norms concerning the importance of citizen participation, researchers have, through the years, focused on this problem. Though much of what is known about the factors influencing voter participation derives from survey, micro-level

data, there have been numerous occasions when aggregate data have been employed to investigate voting behavior (Alford and Lee, 1968). Studies using aggregate data have most often used correlational methods, seldom attempting to estimate regression coefficients. This example will illustrate grouping data by census blocks and tracts (a common procedure in macro-level voting studies) as it compares with ungrouped responses. Kim, Petrocik and Enokson (1975) treat the problems of analyzing voting with aggregate data where micro and macro data are combined and systematically measure the interaction.

The model of voter participation used in this example is drawn from the literature based on micro-level survey data. It is hoped that this will lessen the likelihood of misspecification and thus avoid that additional handicap. Ben-Sira (1977) has suggested a model based on a thorough review of the previous research on voting and notes that there has been shown to be a strong association between socio-economic status and voting. The trend is for individuals to have a higher propensity to vote, given a higher social status. The components of the model are presented in Table 2.

The data is from a one percent survey of Atlanta and suburban Fulton County conducted in 1976 and yielding over 7,000 respondents. The substantive model and this data provide a background to test the methodological problem of the effect of grouping on correlational measures such as the R^2. Incomplete data were accounted for by the mean substitution technique which does not bias the regression coefficients but does lower variance and efficiency. No missing data cases for the dependent variable were included.

The illustration is in the form of three regression analyses: one each of three levels of aggregation -- census blocks, tracts, and total respondents. The specification remained the same for each level at which the data were grouped. The model for the grouped data was that of equation (12) and was estimated as a special case of generalized least squares, weighted least squares. Due to the limitations of the Statistical Package for Social Sciences (SPSS) software, a dummy constant term had to be included in grouped equations along with the actual constant term, but this does not affect these examples. The ungrouped data was simply estimated by the model in equation (4). The R^2 in the micro-level specification was found to be .14, a modest coefficient but not unrespectable, given that voting was coded as a dicotomy with having voted in the previous five years as a "1" and not having voted as "0". The strongest variable was the years of schooling.

The grouping by census blocks resulted in 2867 blocks for 7018 individuals and, as shown in Table 2, the resulting R^2 was .18, indicating a modest increase from the total R^2 of .14. The partitioning of sum of squares is presented in Table 3 and indicates that the within groups R^2 is .12, suggesting that controlling for the effects of grouping by blocks has a slight but measurable impact on the specification. Additionally, this implies that grouping by census blocks could possibly be done for purposes of confidentiality or if there were need to reduce data components (see Feige & Watts, 1972).

As is shown in Table 2, the data grouped by census tracts were analyzed. The partitioning of the sum of squares is given in Table 3. The R^2 for total, and also for within census tracts, was .14, however, the between census tracts R^2 was found to be .60 -- a clear example of the effect of grouping on the R^2. Here the increase in R^2 was due to the association between census tracts (the D matrix), as a proxy measure of contextual factors, with the dependent variable (the y vector). In addition, it should be noted that while the regression estimators were not seriously affected by the grouping by tracts, the standard errors of estimators were inflated along with the standardized regression coefficients (β). Thus, the typical measures of the importance of the regression were altered to a considerable degree by aggregation effect. But to restate, for the researcher who is hoping to estimate the micro specification with aggregate data, the R^2 and standardized coefficients are to a large degree a product of the grouping effect itself rather than the substantive in dependent variables.

Summary and Discussion

The purpose of this paper has been to approach the problem of the effect of grouping on R^2 from the analysis of covariance approach, and relate it to the clustering approach of generalized least squares. While through the years there have been warnings against over-reliance on R^2 with grouped data, there continues to be statements such as:

> An additional motivation for using grouped data, however, is that even with sophisticated operational definitions of income and prices, these explanatory variables alone appear to "explain" only a small part of the variations in demand for specific goods and services in individual household data. Grouping observations by the independent variables considerably increases the "explanatory power" of the estimating equation. (Michael and Becker, 1973, pp. 379-380).

It is hoped that the above authors were not seriously claiming that grouping increased the substantive "explanatory power" of their specification. What, in all likelihood, occurred was an artifactual increase in R^2 that the grouping factor induced. It should be noted that there does exist the possibility of an actual "aggregation gain" when, for instance, the micro equation is misspecified and the grouping factor (the D matrix) helps correct the poorly specified micro model (see Irwin and Lichtman, 1976, pp. 423-433). A similar point has been made by Hanuschek, Jackson, and Kain (1974).

The generalized least squares perspective can also be an aid in investigating both temporal and spatial autocorrelation. The G matrix can be used to correct for such misspecifications as heteroscedasticity and autocorrelation. Granger and Newbold (1974) have cautioned that high R^2 may be generated by a misspecified temporal autocorrelation structure. Spatial autocorrelation can result in inefficiency of the estimates of cross sectional studies (Lebanon and Rosenthal, 1975; Cliff, Haggett, Ord, Bassett and Davies, 1975).

The researcher cannot expect the R^2 determined from grouped data to be a robust measure for use in evaluating models unless the grouping procedure is random with respect to the dependent variable. The R^2 "inflation problem" is actually a specification issue where methodology and technique are, at best, only partial factors in a more complete solution.

References

Alford, R.R. and E.C. Lee (1968) "Voting Turnout in American Cities." American Political Science Review, 62: 796-813.

Ben-Sira, Z.C. (1977) "A facet theoretical approach to voting behavior." Quality and Quantity, 11: 167-188.

Cliff, A.D., P. Haggett, J.K. Ord, K.A. Bassett, and R.B. Davies (1975) Elements of Spatial Structure: A Quantitative Approach. New York: Cambridge University Press.

Cramer, J.S. (1964) "Efficient grouping: regression and correlation in Engle Curve Analysis." Journal of the American Statistical Association, 59: 233-250.

Feige, E.L. and H.W. Watts (1972) "An investigation of the consequences of partial aggregation of micro-economic data." Econometrica 40: 343-360.

Gehkle, C. and K. Biehel (1934) "Certain effects of grouping the size of the correlation coefficient in census tract material." Journal of the American Statistical Association Supplement 29: 169-170.

Granger, C.W.J. and P. Newbold (1974) "Spurious regression in econometrics." Journal of Econometrics 2: 111-120.

Grunfeld, Y. and Z. Griliches (1960) "Is aggregation necessarily bad?" Review of Economics and Statistics 42: 1-13.

Hanushek, E.A., J.E. Jackson and J.F. Kain (1974) "Model specification, use of aggregate data, and the ecological correlation fallacy." Political Methodology 1: 87-106.

Irwin, L. and A.J. Lichtman (1976) "Across the great divide: inferring individual level behavior from aggregate data." Political Methodology 3: 411-439.

Johnston, J. (1972) Econometric Methods (2nd edition). New York: McGraw-Hill.

Kim, J.O., J.R. Petrocik, and S.N. Enokson (1975) "Voter turnout among the American States: systemic and individual components." American Political Science Review 69: 107-123.

Lebanon, A. and H. Rosenthal (1975) "Least squares estimation for models of cross-sectional correlation." Political Methodology 2: 221-244.

Michael, R.T. and G.S. Becker (1973) "On the new theory of consumer behavior." Swedish Journal of Economics 75: 378-396.

Prais, S.J. and J. Aitchison (1954) "The grouping of observations in regression analysis." Review of the International Statistical Institute 22: 1-22.

Robinson, W.S. (1950) "Ecological correlations and the behavior of individuals." American Sociological Review 15: 351-357.

Smith, K.W. (1977) "Another look at the clustering perspective on aggregation problems." Sociological Methods and Research 5: 289-315.

Theil, H. (1971) Principles of Econometrics. New York: Wiley.

Thorndike, E.L. (1939) "On the fallacy of inputing the correlations for groups to the individuals or smaller groups." American Journal of Psychology 52: 122-124.

Yule, G.U. and M.G. Kendall (1950) An Introduction to the Theory of Statistics. London: Charles Griffin.

Table 2

COMPARISON OF REGRESSIONS FOR MICRO-LEVEL DATA WITH CENSUS BLOCK AND TRACT AGGREGATIONS

Dependent Variable: Voted in Previous Five Years

	Micro-Level			Macro-Level					
				Census Blocks			Census Tracts		
Independent Variables	Estimator	Standard Error of Estimator		Estimator	Standard Error of Estimator		Estimator	Standard Error of Estimator	
Schooling (years)	.040	.0016	.32	.040	.0024	.35	.065	.010	.74
Age (10 year units)	.038	.0032	.15	.033	.0046	.13	.045	.022	.16
Income ($10,000 units)	.022	.0050	.052	.031	.0079	.074	.054	.034	.16
Race (White)	-.029	.011	-.033	-.034	.015	-.043	-.081	.033	-.19
Political Efficacy	-.0019	.0036	-.0058	.0029	.0053	.0293	-.023	.026	-.051
Public Interest	.0068	.00062	.13	.0062	.00096	.12	.0074	.0045	.13
Governmental Salience (high or low)	.019	.010	.021	.022	.016	.024	-.073	.066	-.077
Constant	.017			.011	.019	.0095	.22	.14	.11
Dummy Constant	*			-.012			-.15		
R^2	.14			.18			.60		
Standard Error of Estimate	.41			.34			.11		
N	7018			2867			137		

*Not needed in the micro-level specification

Table 1

ANALYSIS OF COVARIANCE APPROACH TO GROUPED DATA

Source of Variation	Residual Sum of Squares	$1-R^2$
Between	$e'e = y'y - \hat{\hat{\alpha}}'D'y - \hat{\hat{\beta}}'X'y$	$\frac{y'y - \hat{\hat{\alpha}}'D'y - \hat{\hat{\beta}}'X'y}{y'Ay}$
Within	$s's - e'e = \hat{\hat{\alpha}}'D'y + \hat{\hat{\beta}}'X'y - \hat{\beta}'X'y$	$\frac{\hat{\hat{\alpha}}'D'y + \hat{\hat{\beta}}'X'y - \hat{\beta}'X'y}{y'Ay}$
Total	$s's = y'y - \hat{\beta}'X'y$	$\frac{y'y - \hat{\beta}'X'y}{y'Ay}$

Table 3

EFFECT OF GROUPING BY CENSUS TRACTS AND BLOCKS ON SUMS OF SQUARES FOR VOTING MODEL

Source of Variation	Sums of Squares: Regression	Residual	Total	R^2
Between Census Tracts	2.39	1.59	3.97	.60
Within Census Tracts	189.56	1183.77	1373.34	.14
Total	191.95	1185.36	1377.31	.14
Between Census Blocks	71.82	333.72	405.54	.18
Within Census Blocks	119.82	851.64	971.77	.12
Total	191.94	1185.36	1377.31	.14

AN APPLICATION OF BALANCED REPEATED REPLICATION TO THE ESTIMATION OF VARIANCE COMPONENTS

Judy A. Bean, University of Iowa
George A. Schnack, National Center for Health Statistics

1. INTRODUCTION

For purposes of designing surveys, survey statisticians need to know the components of variation inherent in the stages of a sampling plan or be able to estimate them from previous surveys. This paper presents the results of applying the balanced repeated replication (BRR) technique to data collected in the Health Interview Survey in order to estimate the variance components of four statistics.

In recent years, the BRR method has been adopted for estimating variances of estimates from complex probability surveys but has not been employed for estimating variance components. In 1975 Casady (3) showed for the first time that the BRR method can be adapted to estimate the variance components of a linear estimator from a two-stage stratified design. When sampling without replacement at both stages, the BRR estimators of total variance and within variation are biased. The between variability is estimated by subtracting the within estimate from the estimate of total variance. Bean (2) has derived another version of the BRR technique that yields unbiased estimates of the within component for the same sample design. However, no one has investigated the use of the method for survey designs that are more complicated than a simple two-stage stratified one.

2. METHODOLOGY

2.1 The BRR Estimators

Before describing the methodology of this study, the BRR estimators of variance components for a simple design will be featured.

Let us consider a finite population of N primary units classified into L strata each containing N_i units (i = 1, 2, ..., L) with $N = \sum_{i=1}^{L} N_i$.

Each primary unit consists of M_{ij} elements. Denote by X_{ijk} the measurement of interest on the k^{th} element in the j^{th} primary unit of the i^{th} stratum and by

$$X_{ij.} = \sum_{k=1}^{M_{ij}} X_{ijk}, \quad X_{i..} = \sum_{j=1}^{N_i} X_{ij.},$$

and $X_{...} = \sum_{i=1}^{L} X_{i..}$ the population primary unit total, the population stratum total and the population total. A random sample of n_i units is drawn without replacement from the i^{th} stratum; within each selected primary unit, m_{ij} elements are selected randomly without replacement. The n_i's and m_{ij}'s are assumed to be even numbers. The customary unbiased estimator of the population total, X..., is

$$X' = \sum_{i=1}^{L} \frac{N_i}{n_i} \sum_{j=1}^{n_i} \frac{M_{ij}}{m_{ij}} \sum_{k=1}^{m_{ij}} X_{ijk} \qquad (1)$$

and the variance of X' is:

$$\sigma^2_{X'} = \sum_{i=1}^{L} N_i^2 (1 - f_i) n_i^{-1} S_i^2$$

$$+ \sum_{i=1}^{L} \sum_{j=1}^{N_i} N_i M_{ij}^2 (1 f_{ij}) n_i^{-1} m_{ij}^{-1} S_{ij}^2 \qquad (2)$$

where f_i = the first stage sampling fraction, f_{ij} = the second stage sampling fraction,

$$S_{ij}^2 = \sum_{k=1}^{M_{ij}} (X_{ijk} - \bar{X}_{ij.})^2 (M_{ij} - 1)^{-1}, \text{ and}$$

$$S_i^2 = \sum_{j=1}^{N_i} (X_{ij.} - \bar{X}_{i..})^2 (N_i - 1)^{-1}.$$ The first term on the right-hand side of the equation (2) is the variability between primary units. The second term is the variation among the elements within the primary units.

To obtain an estimate of the total variance for X' by the BRR procedure, the n_i sampled primary units are randomly split into two groups, each of size $n_i/2$. Next, using an orthogonal matrix (for more details see McCarthy (7)), A half-samples are created by randomly selecting one of the two groups of the primary units from each of the L strata. Utilizing only the data from each half-sample, A estimates of the population parameter are made. The BRR estimate of $\sigma^2_{x'}$ is

$$\hat{\sigma}^2_{x'} = \sum_{\alpha=1}^{A} (X'_\alpha - X')^2 A^{-1} \qquad (3)$$

To estimate the within component, denoted as σ^2_w, each of the primary units is considered to be a pseudostratum. Here, the m_{ij} sampled elements are randomly placed in one of two equal sized groups. A half-sample, thus, consists of choosing one of the two groups of elements from each of the n_i primary units. The data from a half-sample is subjected to the same estimation procedure as the data from the total sample, creating another estimate of X... By means of a second orthogonal pattern, B estimates of X' are produced. Then an estimate of the within component is:

$$\hat{\sigma}^2_w = \sum_{\beta=1}^{B} (X'_\beta - X')^2 B^{-1} \qquad (4)$$

2.2 Sample Data

The data for the study were those collected in the 1973 Health Interview Survey (HIS) of 120,493 civilian noninstitutional individuals. A description of the survey has been published by the National Center for Health Statistics (9) but

the sample design and estimation procedure used will be outlined to illustrate its intricacies.

The sampling plan of HIS is to select one primary unit which is either a county or group of counties of the United States from each of 376 strata with probability proportional to size. Some of the strata contain only one primary unit. The second stage units chosen are clusters of approximately 4 households. For each selected household, information concerning a person's perception of his/her health is gathered for each person residing at the household.

After these data are subjected to an extensive editing procedure, estimates of morbidity are produced using a complex estimation equation. The equation includes unequal weighting caused by unequal probabilities of selection, nonresponse adjustment and two ratio adjustments.

To recapitulate, features of the design are unequal probabilities of selection, stratification, clustering and strata containing only one unit. For estimation purposes, an adjustment for nonresponse and two ratio adjustments are performed.

2.3 Study Design

An underlying assumption of the BRR method is that at least two units are chosen from each stratum; however, for surveys not fulfilling this requirement, the practice is to pair primary units based on characteristics of the strata they represent. The sampled primary units in HIS from strata consisting of more than one unit were collapsed to form pseudostrata; strata consisting of one primary unit each were grouped together in a particular fashion to form an additional set of pseudostrata which will be called self-representing (SR) pseudostrata. A distinction is made between the two groups because the variation in the SR pseudostrata only reflects the within variability, not the between variation. This is taken into account when estimating the variance of estimates.

Even after the 376 strata are collapsed into pairs, there are still 160 pseudostrata which means a 160 x 160 orthogonal matrix is needed to estimate the variance of an estimate. The number of half-samples required for the BRR method equals the first multiple of 4 large as or larger than the number of pseudostrata. Since each primary unit is assumed to be a pseudostratum for estimating the within component, the size requirement for an orthogonal matrix here is greater than 160 x 160. Because the main objective of the investigation was to simply demonstrate that the BRR method can be applied, the decision was made to use only data from the South region. The reason for choosing this geographic location was that the South was the largest; it consists of data for 38,053 persons.

For clarification the steps involved in the preparation of the sample data for use by the BRR method are reviewed.

A. Estimation of total variance:

Here the SR primary units were grouped to form 10 pseudostrata; the remaining units were paired into an additional 61 strata. Within each stratum there must be two primary units. For the 10 SR pseudostrata, the clusters of households within each pseudostratum were randomly partitioned into two groups. The other 62 pseudostrata consisted of two primary units each. Thus, with a total of 71 pseudostrata, the size requirement for the orthogonal matrix is 72 x 71.

B. Estimator of within variance:

To use the BRR method here, the assumption of two units selected from each stratum must be met. First, each of the sampled primary units in the 61 non-SR pseudostrata was considered to be a pseudostratum resulting in 122 pseudostrata. Secondly, within each of these primary units the clusters of 4 households were randomly allocated into one of two groups. The partitioning of the 10 SR pseudostrata for step A was retained for this step. Because a 132 x 132 orthogonal matrix does not exist, a 136 x 132 matrix was employed.

2.5 Variance Estimators

For each statistic produced, its variability was estimated in two ways using the BRR method; these two versions are described by McCarthy (7). The variance estimators are:

$$\hat{\sigma}^2_{\theta''} = \sum_{\alpha=1}^{72} (\theta'_\alpha - \theta'')^2/72 \qquad (5)$$

and

$$c\hat{\sigma}^2_{\theta''} = \sum_{\alpha=1}^{72} (\theta^*_\alpha - \theta'')^2/72 \qquad (6)$$

where θ'' = the final nonresponse ratio adjusted estimate, θ'_α = the nonresponse ratio adjusted estimate secured from the α^{th} half-sample, and θ^*_α = the nonresponse ratio adjusted estimate secured from the α^{th} complement half-sample (the primary units not in the α^{th} half-sample).

A comment on the estimates produced from the half-samples is necessary. As mentioned earlier, three sets of adjustment factors are applied in order to take advantage of ratio estimation, poststratification and imputation for nonresponse. Therefore, the correct method for estimation is the calculation of these adjustment factors for each particular half-sample. This is straight forward but requires considerable work. Studies by Simmons and Baird (10) and Kish and Frankel (4,5) indicate that the adjustment factors based on the parent sample can be applied without the estimates being seriously biased. Contrarily, the results of investigations by Bean (1) and Lemeshow (6) conclude that the adjustment of factors should be computed for each specific half-sample. Due to costs and time for this feasibility study, the adjustment factors for the entire sample were applied to estimate within and total variance.

There were 132 pseudostrata (10 SR pseudostrata and 122 others) and no known 132 x 132 orthogonal matrix; thus, an orthogonal matrix, 136 x 132, was utilized in computing the BRR estimate of the within component of variation. The estimators are:

$$\hat{\sigma}^2_w = \sum_{\beta=1}^{136} (\hat{\theta}_\beta - \theta'')^2/136 \qquad (7)$$

and

$$c\hat{\sigma}^2_w = \sum_{\beta=1}^{136} (\tilde{\theta}_\beta - \theta'')^2/136 \qquad (8)$$

where $\hat{\theta}_\beta$ = the nonresponse ratio adjusted estimate produced from the β^{th} half-sample, and $\tilde{\theta}_\beta$ = the nonresponse ratio adjusted estimate produced from the β^{th} complement half-sample.

3. EMPIRICAL RESULTS

As stated previously using the Health Interview Survey data for the South region, the BRR technique was applied to produce estimates of total variance and within variation. McCarthy (8) shows that if the average of the half-sample estimates of the parameter is essentially the same as the total sample estimate of the parameter, the differential bias of the average and the estimator θ'' will be close to zero. This is important since the BRR estimate of variance will reflect that differential bias. A relationship between the variance of the mean of the half-sample estimates and the variance of θ'' is derived. From this McCarthy infers that when this differential bias is small the estimate of the variance of θ'' is "good".

For the data presented in this paper, Table 1 gives the mean of the half-sample estimates, the mean of the complement half-sample estimates and the total sample estimates. The means are close to the value of θ''; thus, the inference is that the BRR estimate of variance is "good". Besides this evidence, Bean (1) has demonstrated that the BRR method yields a satisfactory estimate of variance of a ratio estimator.

In calculating an estimate of within variability, half-sample estimates of the population parameters are computed. The mean of these half-sample estimates and the mean of their complement half-sample estimates are presented in Table 2 along with total sample estimates. These three estimates are almost identical, meaning the differential bias here is near zero. One may wish to argue that if this bias is close to zero the estimate of variance which in this situation is an estimate of the within component is a "good" estimate; however, such an argument is based on the fact that the type of relationship found for the total variance estimate must hold for the within component estimate. To date, there is no derivation of the relationship of the variances here so the results are to be interpreted cautiously. The conclusion is that the differential bias between the mean of the half-samples estimates/complement half-sample estimates and θ'' is negligible so the within component estimate is not inflated by the bias.

The estimates of variance using the BRR method are shown in Table 3. For example, 72.69% of the population living in the South saw a doctor last year. This estimate has a variance of 15.0×10^{-8}; the variance is partitioned into 3.0×10^{-8} from sampling the primary units and 11.9×10^{-8} from sampling within the primary units. For three variables, the within estimate is less than the total; both BRR methods give similar values. For the variable dental visits, the within estimate is larger than the estimate for total. Thus, the between estimate is negative which causes some embarrassment. Presently, no answer to the question of what can be done when this event happens is available. To assume the component is zero implies the primary units do not vary among themselves which is unlikely.

Perhaps a more meaningful statistic is displayed in Table 4. The numbers in the table give the percent contribution of each component. The within component contributes approximately 79% of the variability for the three variables number of restricted activity days, number of bed disability days and proportion of population seeing a physician. The variables represent aggregate estimators and a PQ type.

4. DISCUSSION

The results presented here are encouraging but a considerable amount of research remains. The reason for encouragement is that earlier studies performed by statisticians at the National Center for Health Statistics suggest that, for a typical statistic in HIS, the between PSU contribution to variance is in the range of 10% to 20%. For this study the between PSU component is about 20%. One concern about the findings is that the components are too similar. Later work, not given here, indicates that with a different pairing of the PSU's, more realistic component values are obtained. Therefore, an investigation of the effect of varying the pairing scheme may be necessary. We are presently preparing to do additional computations using other statistics for the full 376-PSU sample design in order to assess the problem.

One of the criticisms made of the BRR technique is that the estimates of variance components can not be computed using this method. However, with the work of Casady (3) and this feasibility study this criticism is no longer valid. The purpose of the investigation, to demonstrate that the BRR technique can be utilized to produce estimates of variance, has been accomplished. Whether or not these are the "best" estimates of the components cannot be answered. The limited evidence presented indicates that the estimates are reasonable. Not only are investigations comparing different methods for estimating variance components needed but further theoretical work must be done in order to estimate the variance within the strata. The estimates of stratum variances are the crucial values in designing other surveys.

REFERENCES

1. Bean, Judy A., (1975), "Distribution and Properties of Variance Estimators for Complex Multistage Probability Samples," Data Evaluation and Methods Research, PHS Publication No. 75-1339, Series 2, No. 65, Washington, D.C.: Government Printing Office.

2. Bean, Judy A., (1977), "Unbiased BRR Estimator of Within Component for Two Stage Stratified Design." University of Iowa (Internal Memorandum).

3. Casady, Robert J., (1975), "The Estimation of Variance Components Using Balanced Repeated Replication," Proceedings of the Social Statistics Section of the American Statistical Association, 352-357.

4. Kish, Leslie and Frankel, Martin R. (1968), "Balanced Repeated Replication for Analytical Statistics," Proceedings of the Social Statistics Section of the American Statistical Association, 2-10.

5. _____(1970), "Balanced Repeated Replication for Standard Errors", Journal of the American Statistical Association, 65, 1071-1094.

6. Lemeshow, Stanley (1976), "The Use of Unique Statistical Weights for Estimating Variances with the Balanced Half-Sample Technique," Proceedings of the Social Statistics Section of the American Statistical Association, 507-512.

7. McCarthy, Philip J. (1966), "Replication: An Approach to the Analysis of Data from Sample Surveys," Vital and Health Statistics, PHS Publication No. 1000, Series 2, No. 14, Washington, D.C.: Government Printing Office.

8. McCarthy, Philip J. (1969), "Pseudoreplication: Further Evaluation and Application of the Balanced Half-Sample Technique," Vital and Health Statistics, PHS Publication No. 1000, Series 2, No. 31, Washington, D.C: Government Printing Office.

9. National Health Survey (1958), "The Statistical Design of the Health Household-Interview," Health Statistics, PHS Publication No. 584-A2, Washington, D.C.: Government Printing Office.

10. Simmons, Walt R. and Baird, James T., Jr. (1968), "Pseudoreplication in the NCHS Health Examination Survey," Proceedings of the Social Statistics Section of the American Statistical Association, 19-30.

11. Tepping, Benjamin J. (1968), "The Estimation of Variance in Complex Surveys, "Proceedings of the Social Statistics Section of the American Statistical Association, 11-18.

Table 1. Comparison of the Estimate for the Total Sample with the Averages of the Half-Sample Estimates Used in Estimating Total Variance[a]

Variable	θ"	Averages	
		Half-Sample	Complement Half-Sample
Number of restricted activity days	$1{,}197.57 \times 10^6$	$1{,}201.56 \times 10^6$	$1{,}193.56 \times 10^6$
Number of bed disability days	479.18×10^6	480.25×10^6	478.09×10^6
Number of dental visits	81.24×10^6	80.68×10^6	81.81×10^6
Proportion of population seeing a physician	72.69×10^{-2}	72.73×10^{-2}	72.76×10^{-2}

[a]See the text for a description of these estimates.

Table 2. Comparison of the Estimate for the Total Sample with the Averages of the Half-Sample Estimates Used in Estimating Within Variation[a]

Variable	θ"	Averages	
		Half-Sample	Complement Half-Sample
Number of restricted activity days	$1{,}197.57 \times 10^6$	$1{,}196.29 \times 10^6$	$1{,}198.82 \times 10^6$
Number of bed disability days	479.18×10^6	477.48×10^6	480.86×10^6
Number of dental visits	81.24×10^6	81.24×10^6	81.24×10^6
Proportion of population seeing a physician	72.69×10^{-2}	72.67×10^{-2}	72.71×10^{-2}

[a]See the text for a description of these estimates.

Table 3. Balanced Repeated Replication Estimates of Total Variance and Components for Four Variables[a]

Variable	Variance Estimates					
	Half-Sample			Complement		
	Total	Between	Within	Total	Between	Within
Number of restricted activity days	$1{,}063.8 \times 10^{12}$	228.7×10^{12}	835.1×10^{12}	$1{,}063.7 \times 10^{12}$	228.6×10^{12}	835.1×10^{12}
Number of bed disability days	252.1×10^{12}	53.7×10^{12}	198.4×10^{12}	252.1×10^{12}	53.8×10^{12}	198.3×10^{12}
Number of dental visits	652.4×10^{10}		703.6×10^{10}	652.5×10^{10}		703.6×10^{10}
Proportion of population seeing a physician	150.0×10^{-7}	30.4×10^{-7}	119.6×10^{-7}	150.9×10^{-7}	31.8×10^{-7}	119.1×10^{-7}

[a]See the text for a description of these estimates. A blank indicates the estimate of variance was negative.

Table 4. The Percent Contribution of Each Component to the Total Variance

Variable	Contribution	
	Between	Within
Number of restricted days	21.5%	78.5%
Number of bed disability days	21.3%	78.7%
Number of dental visits[a]		
Proportion of population seeing a physician	20.3%	79.7%

[a]A blank indicates the percentage was either negative or over a hundred.

DISCUSSION: MISCELLANEOUS STATISTICAL ANALYSIS, SOCIAL STATISTICS CONTRIBUTED PAPERS SESSION XXV

Gary M. Shapiro, U. S. Bureau of the Census

I will discuss only two of the papers given, those by Sukhatme-Chand and by Kondo-Schnack.

I. SUKHATME-CHAND PAPER

The Sukhatme-Chand paper is a good, professional paper. I do question, however, how much practical value it has. Are there really many instances in which the results can be applied? The paper is primarily concerned with the following situation: Initially, a large sample is taken and one variable is observed; then a subsample is taken and a second variable is observed on this subsample; then a second subsample (a subset of the first subsample) is taken and the variable of real interest is observed. Do situations like this really occur?

In my personal experience, the auxiliary variables have usually not been estimated from samples, but rather have been subject to zero variance. This situation is considered in the paper as a special case and Theorem 5.1 in particular is useful. However, the formulae given for the bias and mean square error for this special case are trivial. Also, the cost discussion and example are not applicable to this situation.

II. KONDO-SCHNACK PAPER

One major advantage of the Keyfitz (or Taylor series linearized) approach [3], [4], [5] over the replication approach to variance estimation has been the inability of the replication variance approach to estimate the components of variance separately. Thus, a general elimination of this inability would significantly improve the value of replication variance estimation. In Casady [2] and Bean [1] (referenced in the Kondo-Schnack papers), it has been shown that components estimation is possible in the case of linear estimates. Linear estimates are of little interest, though, as much simpler methods of variance estimation than either replication or Keyfitz are acceptable. It appears to me, however, that the same procedure will work as well for nonlinear estimates as for linear. I believe Casady's results can be easily generalized for nonlinear estimates.

The Kondo-Schnack paper fully accomplishes its rather modest goal of illustrating the application of the previously developed theory of Casady [2] and Bean [1]. The replication procedure for estimating components is somewhat inconvenient in that completely separate sets of replications are needed for the within and total variance estimates, but this doesn't appear to have caused the authors any major problems. The empirical results are acceptable; some negative estimates of between PSU variance are obtained, but these would also be likely to occur if Keyfitz variance estimates had been made.

I have two minor criticisms of the paper. First, a reader of the paper is left with the impression that the National Center for Health Statistics (NCHS) designed and conducted the Health Interview Survey. In fact, the Bureau of the Census, under contract to NCHS, conducted the survey and was primarily responsible for the design. This is a common problem: Papers written by staff members of organizations who sponsor surveys quite frequently fail to properly acknowledge the organization which actually designed and conducted a survey. The second criticism relates to some incorrect numbers given in the paper. Health Interview Survey was redesigned in 1973, with 376 sample PSU's instead of the previous 357 sample PSU's, and with clusters averaging four households instead of the previous six.

REFERENCES

[1] Bean, Judy A. "Unbiased BRR Estimator of Within Component for Two-Stage Stratified Design," Internal University of Iowa memorandum, 1977.

[2] Casady, Robert J. "The Estimation of Variance Components Using Balanced Repeated Replication." Proceedings of the Social Statistics Section, American Statistical Association, 1975, pp. 352-357.

[3] Keyfitz, Nathan. "Estimates of Sampling Variance Where Two Units are Selected from Each Stratum." Journal of the American Statistical Association, Vol. 52, (Dec. 1957), pp. 503-510.

[4] Tepping, Benjamin J. "Variance Estimation in Complex Surveys." Proceedings of the Social Statistics Section, American Statistical Association, 1968, pp. 11-18.

[5] Woodruff, Ralph H. "A Simple Method of Approximating the Variance of a Complicated Estimate." Journal of the American Statistical Association, Vol. 66, No. 334, (June 1971), pp. 411-414.

A CENSORED NONPARAMETRIC TEST OF Ho: $G = 1-(1-F)^K$

Ronald Regal, SUNY Albany

To indicate why one might be interested in the null hypothesis Ho: $G = 1-(1-F)^K$, consider the data in Table 1 adapted from Restle and Davis (1962) and Davis and Restle (1963). First, 251 subjects were separated into M=163 single (individual) units and N=22 groups units with K=4 subjects in each group. Each of the M+N=185 units was given the same problem to solve. Table 1 gives approximate results for Restle and Davis' gold problem. For details on the relation of these data to the charts in Restle and Davis (1962), see Regal (1975). Table 1 gives the times in seconds of the m=71 individuals and n=17 groups who solved the problem before the time limit 0f 678 seconds. The 17 group times are identified by (G). The times of the M-m=92 and N-n=5 groups who had not solved by the time limit are considered to have been right censored by the time limit.

1. Solution Times

62 (G)	241	370	520
92	252	372	528
95	256	375	538
100(G)	259(G)	378	547
101	261	381	548(G)
111	280	383	558
130	290(G)	386	597
131(G)	295	388	611
135	300	390	615
140(G)	310	392(G)	618
141	313	395	620(G)
162	317	399	637
165	320	399(G)	639(G)
168	340	409	649
170	343	418	649(G)
181	346	440(G)	666
181(G)	348	452	667(G)
191	350	459	669
201	352	469	672
210(G)	355	481	675
221	358	489	677
229	361	509	677(G)

(G) = group time
92 individuals and 5 groups did not solve by the time limit of 678 seconds.

To introduce some notation, let F be the distribution function of the time required for a randomly chosen individual to solve working alone, and let G be the distribution function for the time needed by K=4 subjects working as a group. One could test Ho: F=G by a number of nonparametric methods which allow for censoring and ties such as we have in Table 1. A summary of some such methods is given by Gehan (1976).

However, even if one knew, say, that a group of size K=4 is expected to perform better than a single individual, one would still not know whether a problem should be solved sooner by four subjects working together or four subjects working separately. The true test of the effectiveness of the grouping comes in comparing the group solution time to the best (minimum) time of K independently working individuals. The null hypothesis that one is equally likely to receive a solution before any time t from a group of size K or from K independently working individuals is

$$H_0: \quad G(t) = 1 - (1 - F(t))^K.$$

Lorge and Solomon (1955) proposed such a model for group problem solving when one only observes the numbers of solving individuals and groups, m and n. Fienberg and Larntz (1971) gave methods for testing the Lorge-Solomon model given such data. The problem here is to develop nonparametric methods of analyzing and testing the Lorge-Solomon type model, H_0: $G = 1 - (1-F)^K$, with timed data containing right censoring and ties.

As first step, define

S_j = # of individual (single unit) solutions among the first j combined solutions.

For the data of Table 1 for example

$S_1 = 0$ $\quad S_{16} =$ unknown $\}$ tie with time of 181 seconds
$S_2 = 1$ $\quad S_{17} = 12$
$S_3 = 3$ $\quad S_{88} = 71$
$S_{15} = 11$ $\quad S_{185} = 163$

Results from Koul and Staudte (1972) can be used to give approximations to the distribution of S_j under H_0: $G = 1-(1-F)^K$. Define

$$V_j = j/(M + N)$$

and let V_j^* be the unique value in [0,1] such that

$$V_j = \lambda V_j^* + (1 - \lambda)[1 - (1 - V_j^*)^K]$$

where

$$\lambda = M/(M + N).$$

Then

$$E(S_j) \simeq M V_j^*$$

and for $i \leq j$

$$\mathrm{Cov}(S_i,S_j) \approx \left(\frac{MN}{M+N}\right)\frac{(1-V_j^*)^K}{\lambda+(1-\lambda)K(1-V_j^*)^{K-1}} \cdot$$

$$\frac{(1-\lambda)K^2V_i^*(1-V_i^*)^{K-1}+\lambda[1-(1-V_i^*)^K]}{\lambda+(1-\lambda)K(1-V_i^*)^{K-1}}$$

Suitably standardized and extended, the S_j process converges weakly to a normal stochastic process. For details and justifications see Regal (1975). As an example of the approximations consider S_{88} under the conditions of Table 1. Using $V_{88} = 88/185 = 0.4757$, the above results suggest approximations of 68.486 and 2.166 for the mean and variance of S_{88} compared to exact values of 68.498 and 2.174 found through recursive methods (Regal, 1975). Since $S_{88} = 71$ for Table 1, there are more than the expected number of individuals or too few groups compared to expectations under H_0. Hence at the 88th checkpoint the groups are doing worse than expected under the Lorge-Solomon model.

Similar comparisons of S_j to the expected value of S_j under Ho can be made at those values of j for which S_j is known. A graphical presentation of the deviations from the Lorge-Solomon model is provided by the plot of $S_j-E(S_j)$ as a function of j. Figure A gives such a plot for the data of Table 1. A possible interpretation of Figure A is that at the beginning the groups did nearly as well as independently working individuals, but as time went by, the grouping started impeding solution. Since the $\mathrm{Var}(S_j)$ is smaller for small j and large j than for intermediate j, one might wonder how much of the apparent peaking in Figure A can be explained by increased variability. Figure B shows a plot of the standardized variable $(S_j-E(S_j))/(\mathrm{Var}(S_j))^{\frac{1}{2}}$. Figure B lends itself to the same sort of interpretations as Figure A in this case.

Although Figures A and B suggest that the group performance falls short of the performance of an equal number of independently working individuals, we still need an overall test of the Lorge-Solomon model, Ho: $G = 1-(1-F)^K$. One possibility is the statistic

$$\frac{\Sigma(S_j-E(S_j))}{(\mathrm{Var}(\Sigma S_j))^{\frac{1}{2}}} \cdot$$

In the case of no censoring or ties this can be shown to be equivalent to the Wilcoxon rank sum or Mann-Whitney statistic, and results from Lehman (1953) can be used to give the exact mean and variance. See Regal (1975) for details, including a comparison of the normal approximation and the exact distribution.

With ties, including ties due to censoring, one possibility is to make inferences conditional on the observed pattern of ties and assign midvalues. For example in Table 1 there is a tie between an individual and a group for places number 16 and 17, and S_{16} is unknown. Giving S_{16} a value of $(S_{15} + S_{17})$ can be shown to be equivalent to using midranks in the Wilcoxon rank sum test. Graphically, the statistic $\Sigma(S_j-ES_j)$ with midvalues attempts to integrate Figure A extended out to $S_{185}-E(S_{185})\equiv 0$. Using $(S_{15} + S_{17})/2$ for S_{16}, the statistic ΣS_j involves S_{15} and S_{17} each multiplied by 1.50 in this case. The variance of ΣS_j would be figured accordingly, using the approximation given above for $\mathrm{Cov}(S_i,S_j)$. Since the inference is conditional on the pattern of ties in the data, ties between individuals or ties between groups would be treated similarly for variance calculations. For the data of Table 1 the resulting standardized score is 2.31 which corresponds to a 2-sided normal significance level of 0.021. Hence the Lorge-Solomon model would be rejected at the 5% level but not at the 1% level.

In summary, methods have been displayed for graphical presentation of deviations from the Lorge-Solomon type model for timed data and for testing the significance of these deviations. These methods allow for ties, including ties due to censoring by a single common time limit. More complicated forms of censoring can be handled along the line of Mantel (1966).

REFERENCES

Davis, J. H. and Restle, F. (1963). The analysis of problems and prediction of group problem solving. J. Abnorm. Soc. Psychol., 66, 103-116.

Fienberg, S. E. and Larntz, F. K. (1971). Some models for individual-group comparisons and group behavior. Psychometrika, 36, 349-368.

Gehan, E. A. (1976). Analyzing survival data by distribution-free methods. Presented at the Thirty-second Annual Princeton Conference on Applied Statistics.

Koul, H. L. and Staudte, R. G. (1972). Weak convergence of weighted empirical cumulatives based on ranks. Ann. Math. Statist., 43, 832-841.

Lehmann, E. L. (1953). The power of rank tests. Ann. Math. Statist., 24, 23-43.

Lorge, I. and Solomon, H. (1955). Two models of group behavior in the solution of eureka-type problems. Psychometrika, 20, 139-148.

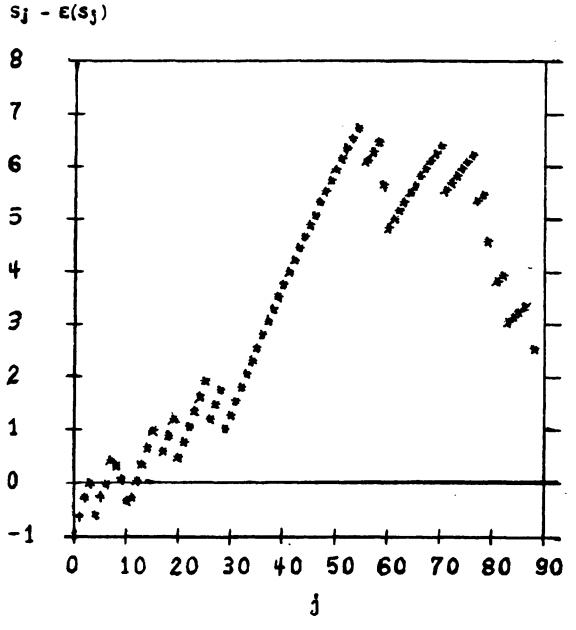
Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. Cancer Chemotherapy Reports, 50, 163-170.

Regal, R. (1975). Some parametric and

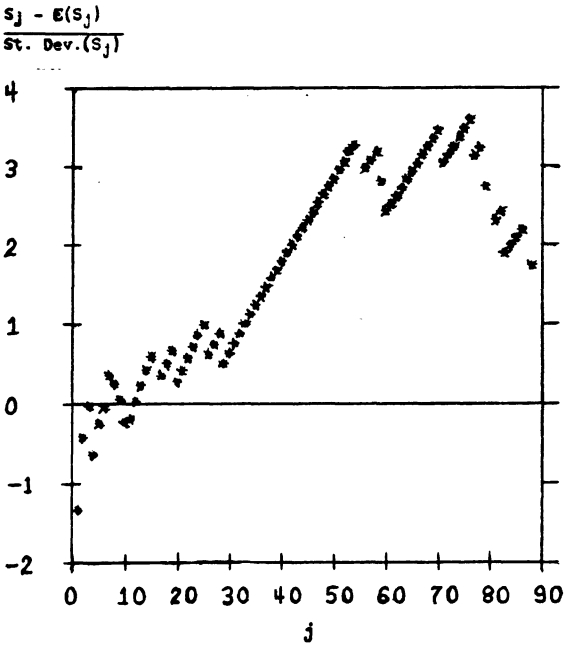
nonparametric methods for individual and group performance. Unpublished Ph.D. thesis, University of Minnesota.

Restle, F. and Davis, J. H. (1962). Success and speed of problem solving by individuals and groups. Psychol. Rev., 69, 520-526.

A. DEVIATIONS $S_j - E(S_j)$ FROM LORGE-SOLOMON MODEL



B. STANDARDIZED DEVIATIONS FROM LORGE-SOLOMON MODEL



ON NON-HIERARCHICAL LOG-LINEAR MODELS AND THE ITERATIVE PROPORTIONAL FITTING ALGORITHM[1]

Jay Magidson, Abt Associates Inc.

Abstract

Recent literature on log-linear models gives the impression that the Interative Proportional Fitting (IPF) algorithm yields maximum likelihood estimates only for hierarchical (not non-hierarchical) models. While it is true that hierarchical models are often more easily interpreted than non-hierarchical models, it is shown here that the IPF algorithm (and existing computer programs designed for hierarchical models) can be used to estimate any non-hierarchical model. This follows directly from the symmetry between qualitative/categorical indicator variables and appropriately defined "interaction variables." The general approach is illustrated here using data from the study of "The American Soldier," Stouffer et al. (1949). We also illustrate how a qualitative analogue to the R^2 in quantitative regression analysis can be used to partition "qualitative variance" in 14 logit models.

Introduction

In recent years, new statistical methods involving log-linear models have become available for analyzing the relationships among qualitative/categorical variables. The approaches recommended by Goodman (1970), Bishop (1969), Grizzle, Starmer and Koch (1969), Ku and Kullback (1974) and others differ in certain respects but they all formulate the same multiplicative analogue to the additive analysis of variance (ANOVA) model. The development of log-linear models has led to major advances in the statistical analysis of qualitative data.

Some of the recent literature in this area conveys the impression that in order to estimate non-hierarchical log-linear models (i.e., models which hypothesize some higher order interaction terms but which exclude certain lower order terms), one must use some algorithm other than the Iterative Proportional Fitting (IPF) algorithm. The Deming-Stephan (1940) IPF algorithm is recommended by Bishop (1969) and Goodman (1970) for estimating hierarchical models. The purpose of this paper is to point out that the IPF algorithm can also be used to estimate non-hierarchical models. In this paper we illustrate the general estimation approach and also recommend the use of a qualitative analogue to R^2.

For concreteness, we use the data from the study of "The American Soldier" by Stouffer et al. (1949) to motivate the discussion and illustrate the approach. We show how non-hierarchical models which can be transformed into hierarchical models as well as non-hierarchical models which cannot be transformed into hierarchical models can all be estimated using Goodman's ECTA computer program, a program designed to estimate hierarchical models using the IPF algorithm.

We will use data from 8,036 soldiers to predict (D) Camp Preference (North or South) based on knowledge of the three explanatory variables (A) Race (Black or White), (B) Region of Origin (North or South) and (C) Present Location (North or South). Although this paper is limited to a subset of log-linear models (the logit model) and involves only four variables which are all dichotomous, the logic presented here can easily be generalized to log-linear models other than logit models involving any number of polytomous (not necessarily dichotomous) variables (see Magidson, 1976).

We also present a comparison of the likelihood ratio chi-square, the goodness of fit chi-square and the correlation ratio for each of the logit models fitted to the data. The two chi-square statistics indicate how well a model fits the data while the correlation ratio η^2 measures the proportion of variance explained. Beginning with a model of complete independence where $\eta^2_{D.ABC} = 0$, the correlation ratio steadily increases to .35 for the saturated model while the corresponding chi-square values steadily decline indicating that the models which explain the most variance also fit the data best. This gives empirical support to the meaningfulness of η^2, a qualitative analogue to R^2 which is seldom reported for logit models.

Preliminary Analysis of the Data

Table 1 displays the data in the form of a 2-way table where the rows are associated with the explanatory variables Race, Region of Origin and Present Location and the columns refer to the levels (categories) of the dependent variable Camp Preference. The conditional proportions and conditional odds in favor of a northern (and a southern) camp preference are also given in this table.

Thus, for example we see that 91.5% of the 423 black northerners in northern camps prefer a northern camp while only 9.5% of the 960 white southerners in southern camps prefer a northern camp. An equivalent way of looking at these figures is in terms of the odds in favor of a northern camp. For black northerners in northern camps the odds are 387:36 (or 10.75:1) in favor of a northern camp preference while the corresponding odds for white southerners in southern camps is 91:869 (or 0.105:1).

The single best predictor of Camp Preference is (B) Region of Origin. This can be seen directly from Table 1 by noting that a higher proportion of northern-born soldiers prefers the north than southern-born soldiers in every case regardless of Race and Present Location (i.e., even the group of northern-born soldiers least likely to prefer the north, is more likely to prefer the north than any group of southern-born soldiers). Similarly, it is seen that the next best predictor is (C) Present Location while the

weakest predictor is (A) Race.

The Saturated and Unsaturated Logit Models

The saturated logit model for predicting Camp Preference as a function of (A) Race, (B) Region of Origin and (C) Present Location is

$$\log \Omega = \beta + \beta^{A}X^{A} + \beta^{B}X^{B} + \beta^{C}X^{C} + \beta^{AB}X^{AB} + \beta^{AC}X^{AC} + \beta^{BC}X^{BC} + \beta^{ABC}X^{ABC} \quad (1)$$

where Ω denotes the expected value of the conditional odds in favor of a northern camp preference and the X's are indicators associated with the explanatory variables. The X's are defined in Table 2. We will refer to X^{A}, X^{B}, and X^{C} as "main variables" and to the other X's as "interaction variables." We will also refer to the X's as vectors as displayed in Table 2.

Model 1 is a saturated or full rank model because the X-vectors (together with a vector of ones) form a basis for the entire 8-dimensional space. Thus, improved prediction is not possible by including additional variables into the model because any additional variables can be expressed as linear combinations of the X's and absorbed into model 1. The basis vectors are displayed in the form of a design matrix in Table 2. It can easily be verified that the basis is orthogonal (although the estimated β-parameters will not be orthogonal).

Unsaturated or restricted models can be formed from model 1 by omitting some of the X's (i.e., setting some of the β's to zero). Each unsaturated model therefore corresponds to a hypothesis that the vector of expected odds of preferring the north is located in the subspace spanned by the X-vectors included in the model. For example, the main-effects-only model hypothesizes that the vector of odds is located in the subspace spanned by X^{A}, X^{B} and X^{C} (and the constant vector). We will now distinguish between hierarchical hypotheses (models) and non-hierarchical hypotheses (models).

A model including one or more interaction vectors is said to be hierarchical if all lower order X-variables having the same superscripts are also included in the model. Thus, the model including X^{BC} is hierarchical if it also includes X^{B} and X^{C}, otherwise it is non-hierarchical. It follows that the saturated model is the only hierarchical model containing the X^{ABC} vector. A model which excludes all interaction vectors is also said to be hierarchical. Thus, the main-effects-only model is hierarchical, the model which includes only X^{A} (and the constant) is hierarchical and the total independence model which omits all of the X's is also hierarchical.

Any model which is not hierarchical is said to be non-hierarchical. Thus, for example, the model which omits all X-vectors except for X^{A}, X^{B} and X^{BC} is non-hierarchical because it excludes X^{C}.

Results from 14 Logit Models

Table 3 summarizes the results for 13 unsaturated logit models (and the saturated model). Models H_1-H_{10} are hierarchical models estimated earlier by Goodman (1972a) using the ECTA computer program. Models H_{11}-H_{13} are non-hierarchical models also estimated using ECTA. Model H_0 is the saturated model.

The main-effects-only model is designated as model H_2. There are 4 degrees of freedom associated with this model corresponding to the 4 interaction terms omitted. The large chi-square value is significant at well beyond the .01 level so we reject the main-effects-only model in favor of a model postulating interaction.

Model H_1 fits the data exceptionally well as indicated by a chi-square value of only 1.5 with 3 degrees of freedom. This parsimonious model hypothesizes only one interaction term, the (BC) Region of Origin/Present Location term. This model is accepted by Goodman (1972a) for this data. It states that black soldiers are about 2.1 times more likely to prefer the north than white soldiers having the same region of origin and the same present location. (Since there are no interaction terms associated with Race in model H_1, this number is constant over the four joint categories of Region of Origin/Present Location.)[2]

Table 4 compares the estimates of the parameters in model H_1 with those of the saturated model. The estimates are almost identical to two decimal places. Notice that the estimated parameters associated with the interaction variables are smaller in magnitude than those associated with the main variables. Also notice that these estimates are consistent with our preliminary analysis which concluded that (B) Region of Origin was the most important predictor, (C) Present Location was next in importance while (A) Race was the least important explanatory variable for the prediction of (D) Camp Preference as indicated by the correlation ratio. We discuss these correlation ratios in more detail in a later section.

The statistical significance of the BC term in model H_1 can be tested by subtracting the likelihood ratio chi-square for model H_1 from the likelihood ratio chi-square for the main-effects-only model H_2. This difference is asymptotically distributed as a chi-square statistic with one degree of freedom under the null hypothesis that the main-effects-only model is correct (i.e., the null hypothesis is that $\beta^{BC} = 0$ in model H_1). The number of degrees of freedom is the difference in degrees of freedom between these two models. This difference (24.96 - 1.45 = 23.51) is highly significant so we reject the null hypothesis (model H_2) and accept model H_1.

The significance of A in model H_1 can be similarly tested by subtracting the chi-square value for model H_1 from the chi-square value for

the hierarchical model H_3. Similarly, the significance of B and C can be tested using the non-hierarchical models H_{11} and H_{12} respectively. All parameter estimates in model H_1 are statistically significant at well beyond the .01 level.

Model H_{13} is similar to model H_1. The only difference is that it includes the highest order interaction term ABC instead of the BC term. Model H_1 fits the data exceptionally well but model H_{13} does not fit well at all. In the next section we show that model H_{13} cannot be transformed into a hierarchical model by simple transformations while the other non-hierarchical models H_{11} and H_{12} can be so transformed. We also show how these three non-hierarchical models were all estimated using the ECTA computer program.

The proportion of variance explained by these 14 logit models is given in the rightmost column of Table 3. The correlation ratios are discussed in a later section.

The general approach is to convert any model to a main-effects-only model by viewing all variables as main variables whether they are in fact main variables or interaction variables. This will generally involve inputting a larger number of variables into ECTA than is really the case and some (or many) of the frequencies will be structural zeros.

For purposes of illustration, let us first consider the 4 models H_1, H_2, H_{11} and H_{12}. These models include only 4 of the X-variables in their formulation. They include the dependent variable D, and the X-variables, X^A, X^B, X^C and X^{BC}. The coercion approach to estimating these models is to input a 5-way table of frequencies rather than a 4-way table despite the fact that there are really only the four dichotomous variables A, B, C and D. Table 10 displays the 32 frequencies input for these models, 16 of which are structural zeros.[3]

Table 11 gives the marginal tables which are fit for each of these models based on the inputted frequencies of Table 10. The {BCD} table is the 2x2 table which crossclassifies the D dichotomy with the X^{BC} dichotomy. It is different from the 2x2x2 {BCD} table which crossclassifies the three dichotomies B, C and D.

Model H_1 can be estimated based on the inputted frequencies given in Table 10 by specifying that the {B } 3-way table be fit instead of specifying that the three 2-way tables {BD} , { CD} ,{ BCD} be fit. The fact that these alternative specifications are equivalent is shown in Magidson (1976).

Model H_{13} can be estimated in a similar fashion by inputting the 32 frequencies given in Table 12. Or all 5 unsaturated models can be estimated from a single set of frequencies if the 64 frequencies (with 48 structural zeros) corresponding to the 6-way table formed by the X-variables X^A, X^B, X^C, X^{BC}, X^{ABC} and D are input. Taking this logic to the extreme, any model can be estimated based on an 8-way table which also includes the X^{AB} and X^{AC} terms.

Thus, we have shown how any non-hierarchical model can be estimated using ECTA, a program designed for hierarchical models. For occasional estimation of non-hierarchical models, the ECTA program should suffice. For extensive estimation of non-hierarchical models, ECTA can easily be modified to include an option so that one need not input any structural zeros. In any case, it is the IPF algorithm which can be used to calculate ML estimates for the expected frequencies under any hierarchical or non-hierarchical model of the kind usually considered.[4]

References

Birch, M.W., "Maximum Likelihood in Three-Way Contingency Tables," Journal of the Royal Statistical Society, Series B, 25 (1963), 220-233.

Bishop, Y.M.M., "Full Contingency Tables, Logits and Split Contingency Tables," Biometrics, 25 (1969), 383-400.

Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W., Discrete Multivariate Analysis: Theory and Practice, Cambridge, Mass.: MIT Press, 1975.

Bloomfield, P., "Linear Transformation for Multivariate Binary Data," Biometrics, 30 (1974), 609-617.

Deming, W.E. and Stephan, F.F., "On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Tables Are Known," Annals of Mathematical Statistics, 11 (1940), 427-444.

Evers, M. and Namboodiri, N.K., "Weighted Least Squares Versus Maximum Likelihood Estimates of Non-Hierarchical Models," paper presented at the annual meeting of the American Statistical Association, August 1976.

Gini, C.W., "Variability and Mutability," contribution to the study of statistical distributions and relations, Studi Economico - Guiridici della R. Universita de Cagliari, 1912.

Goodman, L.A., "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classification," Journal of the American Statistical Association, 65 (1970), 225-256.

Goodman, L.A., "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications," Technometrics, 13 (1971), 33-61.

Goodman, L.A., "A Modified Multiple Regression Approach to the Analysis of Dichotomous Variables," American Sociological Review, 37 (1972a), 28-46.

Goodman, L.A., "A General Model for the Analysis of Surveys," American Journal of Sociology, 77 (1972b), 1035-1086.

Goodman, L.A., "Causal Analysis from Panel Studies and Other Kinds of Surveys," American Journal of Sociology, 78, No. 5 (1973), 1135-1191.

Goodman, L.A., "The Relationship Between Modified and the More Usual Multiple Regression Approach to the Analysis of Dichotomous Variables," Sociological Methodology, 1974.

Goodman, L.A. and Kruskal, W.H., "Measures of Association for Cross Classifications," Journal of the American Statistical Association, 49 (1954), 732-764.

Grizzle, J.E., Starmer, C.F. and Koch, G.G., "Analysis of Categorical Data by Linear Models," Biometrics, 25 (1969), 489-504.

Haberman, S., book review of Bishop, Feinberg and Holland, Discrete Multivariate Analysis: Theory and Practice in Annals of Statistics, 4, No. 4 (1976), 817-820.

Ku, H.H. and Kullback, S., "Loglinear Models in Contingency Table Analysis," The American Statistician, 28, No. 4 (1974), 115-122.

Light, R.J. and Margolin, B.H., "An Analysis of Variance for Categorical Data," Journal of the American Statistical Association, 66 (1971), 534-544.

Magidson, J., "Qualitative Variables and Simultaneous Equation Econometric Models," Discussion Paper No. 142, Center for Mathematical Studies, Northwestern University, Evanston, Ill., 1975.

Magidson, J., "The Multivariate Analysis of Qualitative Variance: Analyzing the Probability of an Event as a Function of Observable Variables," unpublished doctoral thesis, Northwestern University, June 1976.

Nelder, J.A., "Hypothesis Testing in Linear Models (letters to the editor)," The American Statistician, 30, No. 2 (1976), 103.

Stouffer, S.A., Suchman, E.A., Devinney, L.C., Star, S.A. and Williams, R.M., Jr., The American Soldier: Adjustment During Army Life: Studies in Social Psychology in World War II, Vol. 1, Princeton, N.J.: Princeton University Press, 1949.

Swan, J., Magidson, J. and Berk, R., "Estimation of Non-Hierarchical Log-Linear Models by Circulation, Subversion and Coercion," 1976.

Footnotes

[1] This is an abbreviated version of the original paper, prepared especially for these proceedings. Copies of the complete paper are available upon request from the author at Abt Associates Inc., 55 Wheeler Street, Cambridge, Massachusetts 02138.

[2] The number 2.1 is twice the estimate of β^A expressed in units of odds. By taking the logarithm of 2.1 we convert back to the logit formulation where the parameters are expressed in logarithms of odds. (See Goodman, 1972b.)

[3] ECTA has an option which allows the user to specify which frequencies correspond to structural zeros.

[4] We can conceive of models formed by other kinds of restrictions of course, but these other models are beyond the scope of this paper.

Table 1 Cross-classification of Soldiers with Respect to 4 Dichotomized Variables: (A) Race, (B) Region of Origin, (C) Location of Present Camp, and (D) Preference as to Camp Location

(B) Region of Origin	(C) Location of Present Camp	(A) Race	Number of Soldiers Preferring Camp: In North Freq.	Prob.	Odds	In South Freq.	Prob.	Odds	Total
North	North	Black	387	.915	10.750	36	.085	0.093	423
North	North	White	955	.855	5.895	162	.145	0.170	1117
North	South	Black	876	.778	3.504	250	.222	0.285	1126
North	South	White	874	.632	1.714	510	.368	0.584	1384
South	North	Black	383	.587	1.419	270	.413	0.705	653
South	North	White	104	.371	0.591	176	.629	1.692	280
South	South	Black	381	.182	0.223	1712	.818	4.493	2093
South	South	White	91	.095	0.105	869	.905	9.549	960
			4051	.504	1.017	3985	.496	0.984	8036

Table 3 The Results from Fourteen Logit Models for the Prediction of (D) Location Preference Based On the Explanatory Variables (A) Race, (B) Region of Origin and (C) Present Location

Model	Explanatory Variables Included in the Model	Degrees of Freedom	Likelihood Ratio Chi-Square	Goodness of Fit Chi-Square	Proportion of Variance $\hat{\eta}_{D.ABC}$ Explained
H_0	ALL (A,B,C,AB,AC,BC,ABC)	0	0	0	.350
H_9	A,B,C,AB,BC	2	0.68	0.69	.349
H_8	A,B,C,AC,BC	2	1.32	1.34	.349
H_1	A,B,C,BC	3	1.45	1.46	.349
H_{10}	A,B,C,AB,AC	2	17.29	18.73	.347
H_{13}	A,B,C,ABC	3	24.80	25.48	.345
H_2	A,B,C	4	24.96	25.73	.345
H_3	B,C,BC	4	152.65	147.59	.336
H_4	B,C	5	186.36	180.26	.329
H_{12}	A,B,BC	4	674.78	675.74	.285
H_6	A,B	5	695.01	727.16	.282
H_{11}	A,C,BC	4	1604.57	1905.35	.176
H_5	A,C	5	2286.83	2187.71	.099
H_7	NONE	7	3111.47	2812.64	0

Table 2 The Orthogonal Basis Vectors for the Saturated Logit Model

i	j	k	constant	X^A	X^B	X^C	X^{AB}	X^{AC}	X^{BC}	X^{ABC}
1	1	1	+1	+1	+1	+1	+1	+1	+1	+1
1	1	2	+1	+1	+1	-1	+1	-1	-1	-1
1	2	1	+1	+1	-1	+1	-1	+1	-1	-1
1	2	2	+1	+1	-1	-1	-1	-1	+1	+1
2	1	1	+1	-1	+1	+1	-1	-1	+1	-1
2	1	2	+1	-1	+1	-1	-1	+1	-1	+1
2	2	1	+1	-1	-1	+1	+1	-1	-1	+1
2	2	2	+1	-1	-1	-1	+1	+1	+1	-1

Table 11 Four Logit Models and the Marginal Tables Fit in Order to Estimate These Models by the Coercion Approach Using the Input Data from Table 10

Model	Explanatory Variables Included	Marginal Tables Fit
H_2	A,B,C	{ABC},{AD},{BD},{CD}
H_1	A,B,C,BC	{ABC},{AD},{BD},{CD},{$\overline{BCD}$} or {ABC},{AD},{BCD}
H_{11}	A,C,BC	{ABC},{AD},{CD},{$\overline{BCD}$}
H_{12}	A,B,BC	{ABC},{AD},{BD},{$\overline{BCD}$}

Table 10 The Frequencies Input to ECTA to Estimate Models H_1, H_2, H_{11} and H_{12} by the Coercion Approach

X^A	X^B	X^C	X^{BC}	North	South
1	1	1	1	387	36
1	1	1	-1	0	0
1	1	-1	1	0	0
1	1	-1	-1	876	250
1	-1	1	1	0	0
1	-1	1	-1	383	270
1	-1	-1	1	381	1,712
1	-1	-1	-1	0	0
-1	1	1	1	955	162
-1	1	1	-1	0	0
-1	1	-1	1	0	0
-1	1	-1	-1	874	510
-1	-1	1	1	0	0
-1	-1	-1	-1	104	176
-1	-1	-1	1	91	869
-1	-1	1	-1	0	0

Table 12 The Frequencies Input to ECTA to Estimate Model H_{13} by the Coercion Approach

X^A	X^B	X^C	X^{ABC}	North	South
1	1	1	1	387	36
1	1	1	-1	0	0
1	1	-1	1	0	0
1	1	-1	-1	876	250
1	-1	1	1	0	0
1	-1	1	-1	383	270
1	-1	-1	1	381	1,712
1	-1	-1	-1	0	0
-1	1	1	1	0	0
-1	1	1	-1	955	162
-1	1	-1	1	874	510
-1	1	-1	-1	0	0
-1	-1	1	1	104	176
-1	-1	-1	-1	0	0
-1	-1	-1	1	0	0
-1	-1	1	-1	91	869

A MONTE CARLO STUDY OF MODEL SELECTION PROCEDURES FOR THE ANALYSIS OF CATEGORICAL DATA

Elizabeth Martin Fischer, University of North Carolina

Introduction

Researchers and social scientists frequently confront data analysis situations for which existing theory provides little or no guidance concerning either the determinants of the dependent variable of interest or the nature of the relationships among variables. In such situations, researchers must rely upon a combination of intuition, previous empirical studies, theory, and exploratory data analysis in order to select an appropriate subset of explanatory variables and a model which adequately describes relationships among them. The present paper is concerned with techniques which may be used in the exploratory, model-building stage of research to analyze multi dimensional contingency tables.

We begin with a brief overview of model selection procedures for contingency table data. Four procedures are examined in detail; they are:

1. Stepwise backward elimination of parameters from a saturated model;
2. Stepwise backward elimination of parameters from a homogeneous baseline model;
3. Stepwise forward selection from a homogeneous baseline model;
4. Direct estimation, in which terms are eliminated from a saturated model based upon tests of significance of standardized parameter estimates.

The procedures are evaluated using Monte Carlo simulation techniques. We first specify a true model characterizing a hypothetical population and then analyze repeated samples generated from the hypothetical population. Because the true model is known, the results permit comparison of selection procedures. The following questions are considered:

1. How adequate are the various model selection techniques? That is, how likely is each to lead to the selection of the "true" model, when the true model is known?
2. What are appropriate criteria for an acceptable model? In particular, how well should a model fit in order to be considered acceptable?
3. How much confidence should be placed in the results of applying selection procedures when samples are small? Does the adequacy of the techniques depend upon the nature of the underlying population model?

Overview of Model Selection Procedures

In developing a model to describe a set of data, the researcher first identifies a set of variables for inclusion in the analysis and then specifies the model or equation relating the variables. We assume here that an appropriate set of variables has been identified, and consider techniques to identify a good fitting, parsimonious model to account for the data. A variety of statistical procedures to select log linear models have been developed to assist the researcher in the decision. (See e.g. Goodman 1971, 1973, Birch 1964, Bishop, Fienberg, and Holland 1975, and Brown 1976.) The search procedures vary in several key ways.

Stepwise versus simultaneous tests of parameters may be employed to eliminate or add terms to a model. Stepwise selection requires a test for each parameter to be included in or deleted from a model, while simultaneous procedures test multiple effect parameters simultaneously, and thus require fewer tests to select a final model. Goodman (1973) suggests that simultaneous tests may be employed as an initial screening procedure to eliminate some models from consideration before applying stepwise procedures. As with linear regression, forward or backward stepwise procedures may be employed. Forward selection involves stepwise addition of effect parameters to a model according to some criterion of statistical importance, while backward elimination "prunes" a saturated model by sequential deletion of parameters whose estimated values are statistically insignificant. Goodman (1971, p. 45) cites Draper and Smith (1966) to suggest that backward elimination is superior to forward selection, but provides no evidence concerning their relative performance.

Different methods rely upon different statistical criteria for adding or deleting effect parameters. Goodman (1971) advocates the use of the difference chi-square test statistic, which is the difference in chi-square values for two models, one including the parameter and one excluding it. A statistically significant difference between the two models implies that the effect is significant and must be included in the final model. Higgins and Koch (1977) rely upon chi-square divided by its degrees of freedom to assess the magnitude of an effect parameter. Goodman (1971) also advocates significance tests of the standardized parameter estimates as a criterion for inclusion in a model. Benedetti and Brown (1976) suggest that with large samples selection should not be based upon statistical significance, and advocate the selection of a model which explains a certain fraction of the lack of fit of a baseline model.

In all of these cases, the researcher must also determine the α-level to be used as the criterion of acceptance or rejection of parameters and models. Most applications rely upon conventional α levels of .05 or .10, but there is no evidence that these levels produce optimal results for any or all of the procedures. (In the context of linear regression, α-levels of .10 or .05 do not produce optimal results for all selection procedures.) Perhaps for this reason Goodman (1971) cautions against strict interpretations of significance levels associated with models, suggesting that they should be used as a simple way of taking account of degrees of freedom in assessing the relative goodness of fit of different models.

There have been few studies which assess the adequacy of different techniques and decision rules for selecting log linear models. One exception is Benedetti and Brown (1976), who use real world contingency table data to evaluate forward selection, backward elimination, direct

estimation, and other procedures. Using as a criterion of success the selection of a model which cannot be significantly improved by adding parameters, and from which parameters cannot be dropped, they find that forward selection, backward selection, or a combination of the two performs most adequately. They recommend against the use of simultaneous tests to screen models from consideration, because they find that prior screening led to the exclusion of relevant parameters from the selected model. In addition, they recommend against the use of the difference chi-square when samples are large, arguing that such tests will always be statistically significant with large samples, and that a more appropriate criterion is selection of a model which explains a certain fraction of the lack of fit of a baseline model.

Although the Benedetti and Brown study is useful, the conclusions which can be drawn from two sets of analyses are limited. In addition, because the study is based upon analysis of real data, it is not known which (if any) of the selection procedures arrived at the true population model. Analyses based upon artificial data with known properties provide a more systematic basis for comparing different procedures. There have been a number of such simulation studies of procedures for selecting linear models. Although the findings may not be generalized directly to the log linear case, they are relevant to the issues raised here. The findings suggest that the performance of different linear regression model selection procedures depends in a complex fashion upon the data analysis situation. Dempster, Schatzoff, and Wermuth (1977) find that the performance of different selection procedures is affected by collinearity and multicollinearity among independent variables, centrality in the original model, and the pattern of true regression coefficients. In addition, the choice of significance level has an inconsistent effect upon the accuracy of stepwise selection procedures which rely upon significance tests as decision criteria. Based upon a simulation analysis, Kennedy and Bancroft (1971) recommend sequential deletion (using an α-level of .25) over forward selection. They further find that no single α-level is universally superior for all combinations of parameter values. Finally, Bendel and Afifi (1977) find that the relative performance of different stopping rules in forward stepwise regression depends upon sample size and the number of effect parameters. They recommend that the α-level used with backward elimination be half that used with forward selection.

The complexity of model selection in the linear regression context suggests the importance of evaluating procedures for selecting log linear models. The present paper offers preliminary findings of a Monte Carlo investigation of selecting models to describe multidimensional contingency tables. Two factors which affect the performance of algorithms for selecting linear regression models (the characteristics of the true population model and the size of the sample) are systematically varied. The results suggest that the selection procedures generally perform well, although the adequacy of different procedures depends upon the data analysis situation.

Method

The present analysis attempts to replicate typical analysis problems by simulating a variety of data analysis situations. The characteristics of the true model are varied, and the size of the sample is varied from 50 to 8000. It should be noted at the outset, however, that all of the simulations are based upon four-way cross-classifications of dichotomous variables, and that all models are relatively simple. We further assume that:

1. One true population model gives rise to the observed data.
2. All relevant variables and no irrelevant variables are included in the model.
3. Although the form of the equation is unknown, the correct model is hierarchical. That is, inclusion of a higher order term necessarily results in inclusion of lower order terms involving the same variables. (E.g. if the three-way interaction pertaining to variables A, B, and C is included, then all two-way and one-way effects involving A, B, and C are also included.)

Model selection procedures are compared according to how accurately they identify the correct form of the population model. Thus, a "true" model is one which includes all relevant effect parameters, and excludes all irrelevant effect parameters. The rationale for this broad definition is that in the exploratory stages of research the presence or absence of an effect is very often of primary interest; it is this aspect of specification error that is the focus of the present study.

The true models used to generate the simulated data are reported in Table 1. The only difference between the five hypothetical models is the magnitude of the three-way interaction pertaining to variables A, B, and C, which varies from .00 in Model 1 to .55 in Model 5. Based upon the true population parameters, the multinomial distributions underlying each of the five models are calculated and used to generate random samples via computer algorithm. Fifty replications are generated for each combination of model type and sample size. The selection algorithms are then applied to each random sample to identify models which describe the data.

The selection procedures compared here are discussed by Goodman (1971, 1973), Bishop, Fienberg, and Holland (1975), Benedetti and Brown (1976) and others.[1] In the first set of analyses, a significance level of $\alpha = .05$ is used for all steps in the selection methods. The methods are:

1. Direct estimation. Goodman (1971, 1973) advocates the use of direct estimation as a guide to further stepwise selection of models; here it is applied as a procedure to select a final model. Under the null hypothesis of no effect, standardized parameter estimates are distributed normally and may be tested directly for statistical significance. Standardized parameter estimates in the saturated model are tested (using a critical value of 1.96) and non-significant parameters are deleted, unless deletion would result in a non-hierarchical model. Zero cells in the multiway table are replaced by 1/2 prior to estimating the

parameters of the saturated model. (This practice conforms to Goodman's recommendation in 1964 (see p. 633) but not his later recommendation to add 1/2 to all cells.)

Table 1. Effect parameters for simulated samples

λ effect pertaining to:	Non-standardized λ effects for Model: 1	2	3	4	5
A	.00	.00	.00	.00	.00
B	.00	.00	.00	.00	.00
C	.00	.00	.00	.00	.00
D	.00	.00	.00	.00	.00
AB	.25	.25	.25	.25	.25
AC	.25	.25	.25	.25	.25
AD	.25	.25	.25	.25	.25
BC	.25	.25	.25	.25	.25
BD	.25	.25	.25	.25	.25
CD	.25	.25	.25	.25	.25
ABC	.00	.10	.25	.40	.55
ABD	.00	.00	.00	.00	.00
BCD	.00	.00	.00	.00	.00
ABCD	.00	.00	.00	.00	.00

2. Backward elimination. Models are selected by deleting parameters in a stepwise fashion from a model. The decision to delete a parameter is based upon the statistical significance of the difference chi-square values[2] comparing two models which differ by the presence or absence of the parameter in question. The backward elimination algorithm first tests parameters of the highest order, and proceeds in systematic fashion to tests of lower order terms; at each stage, the parameter associated with the minimum p-value for the difference chi-square is deleted. That is, the term which contributes least to the overall goodness-of-fit of the model is eliminated. Deletion of parameters stops when further deletion would result in a statistically significant loss of fit, or when the goodness of fit of the overall model falls below the specified rejection level.

The model from which parameters are deleted may be either:

a) The saturated model including all main and interaction effects, or
b) A homogeneous baseline model which is selected as an initial best-fitting model.

Homogeneous models which include terms of uniform order (i.e. for an n-way table, models which include all terms of order 1, 2, . . ., k, k + 1, . . ., n) are sequentially compared using the difference chi-square to assess differences in goodness of fit. If the highest order (n-way) interaction is statistically significant, selection is terminated and the saturated model is selected (without further stepwise testing) as the final model. Otherwise, the baseline model is chosen as the model of order k, where k is the lowest order model (1) which fits acceptably ($p > .05$), (2) which fits significantly better than the model of order k - 1, and (3) which is not improved by the addition of all terms of order k + 1. Terms are then deleted in stepwise fashion from the k^{th} order model.[3]

3. Stepwise forward selection. Models are built by adding parameters in a sequential fashion to a baseline model, beginning with lower order terms and proceeding in systematic fashion to higher order terms. The algorithm is analogous to backward elimination, except that at each stage the parameter which most improves the goodness of fit is added. (That is, the term associated with the highest p-value for the difference chi-square is added.) The baseline model is selected using the procedure described for (2b), except that the baseline model is of order k - 1. Addition of parameters stops when no further addition results in a statistically significant improvement in goodness of fit.

Results

Results are found in Table 2; each entry represents the proportion of 50 replications for which the true model is selected using different selection strategies when sample size and the hypothetical population model are varied.

When averaged over samples of varying size and different population models, the data suggest relatively small overall differences in the success of the four model selection strategies. The proportion of correct selections in 2000 replications varies from .64 for backward elimination from the saturated model to .53 for direct estimation. However, the relative and absolute performance of different selection procedures varies according to the data analysis situation.

Not surprisingly, the probability of selecting the correct model is considerably greater when samples are large than when they are small, regardless of which selection procedure is used. However, sample size has a greater effect upon the accuracy of some procedures than others. Direct estimation performs very poorly, and worse than any of the stepwise procedures, when samples are small. For $n \leq 250$, the proportion of correct selections using direct estimation is .14, while the stepwise procedures select the true model for an average of .40 of the replications in samples of the same size. When samples are large ($n \geq 500$), average differences in accuracy among selection procedures are small; the proportion of correct selections is about .77 for all techniques. Sample size has a nonmonotonic effect upon the accuracy of selection procedures. All four techniques are most likely to select true models for sample sizes of 2000 or 4000, with accuracy declining somewhat in larger samples.

The nature of the true population model affects the likelihood that a correct selection will be made using any of the search procedures. When the true model contains a small three-way interaction effect (ABC = .10 in Model 2), no search procedure reliably selects the correct model unless the sample size is 2000 or larger. This finding suggests that if a small effect is theoretically or practically important, a relatively large sample is required to detect it reliably using the procedures examined here. In contrast, when the true model includes a large interaction term (ABC = .40 and .55 for Models 4 and 5, respectively) the stepwise procedures select the true model for over half of the replications even in samples of size 50.

The relative advantage of different model selection strategies also depends upon the true population model. Comparison of results for models 2-5 indicates that the larger the ABC interaction term, the more likely the three stepwise

Table 2. Selection of "true" model using four selection strategies and varying sample size and the characteristics of the true model.

	Sample Size								
	50	100	250	500	1000	2000	4000	8000	Total
Model 1 (ABC = .00)									
Backward deletion from saturated model	.00	.00	.36	.56	.72	.82	.78	.72	.50
Backward deletion from baseline model	.00	.02	.46	.82	.90	.90	.98	.82	.61
Forward selection from baseline model	.02	.04	.40	.56	.72	.84	.78	.72	.51
Direct estimation	.00	.04	.32	.64	.80	.78	.80	.68	.51
Model 2 (ABC = .10)									
Backward deletion from saturated model	.04	.08	.24	.38	.40	.82	.86	.78	.45
Backward deletion from baseline model	.04	.02	.04	.18	.32	.72	.86	.78	.37
Forward selection from baseline model	.04	.02	.16	.38	.38	.72	.86	.78	.42
Direct estimation	.00	.00	.22	.35	.40	.74	.86	.80	.42
Model 3 (ABC = .25)									
Backward deletion from saturated model	.16	.36	.82	.76	.84	.96	.86	.86	.70
Backward deletion from baseline model	.10	.20	.72	.76	.84	.96	.86	.86	.66
Forward selection from baseline model	.10	.20	.80	.78	.84	.96	.88	.86	.68
Direct estimation	.00	.03	.55	.80	.84	.94	.82	.84	.60
Model 4 (ABC = .40)									
Backward deletion from saturated model	.62	.72	.76	.84	.76	.80	.84	.74	.76
Backward deletion from baseline model	.52	.50	.76	.84	.76	.80	.84	.74	.72
Forward selection from baseline model	.56	.50	.78	.84	.76	.80	.84	.76	.73
Direct estimation	.00	.07	.40	.84	.82	.82	.88	.78	.58
Model 5 (ABC = .55)									
Backward deletion from saturated model	.60	.84	.86	.80	.82	.86	.84	.72	.79
Backward deletion from baseline model	.52	.84	.86	.80	.82	.86	.84	.72	.78
Forward selection from baseline model	.54	.86	.86	.80	.82	.86	.84	.74	.79
Direct estimation	.00	.04	.40	.63	.88	.86	.90	.76	.56
Total									
Backward deletion from saturated model	.28	.40	.61	.67	.71	.85	.84	.76	.64
Backward deletion from baseline model	.24	.32	.57	.68	.73	.85	.88	.78	.63
Forward selection from baseline model	.25	.32	.60	.67	.70	.84	.84	.77	.63
Direct estimation	.00	.04	.38	.65	.75	.83	.85	.77	.53

procedures are to select the true model. In contrast, the accuracy of the direct estimation procedure is relatively unaffected by the nature of the population model. Consequently, the relative superiority of the stepwise selection procedures over direct estimation increases as the size of the ABC interaction term is increased.

Differences among the three stepwise selection procedures are relatively small. Backward elimination from a homogeneous baseline model performs better than other procedures when the underlying population model is homogeneous, as is Model 1, which includes all two-way interaction terms. Backward elimination from a baseline model appears to be less sensitive to the presence of small higher order interaction terms than the other stepwise procedures; this is a disadvantage when the true model includes small higher order terms (as is the case for Model 2) but is an advantage when such terms merely represent "noise" (as is the case for Model 1).

All four procedures for selecting a log linear model to describe a multiway contingency table rely upon criteria of statistical significance to accept or reject individual parameters and models. There has been little investigation of an appropriate α-level to select log linear models, although as noted the optimal significance level for selecting linear models generally differs according to procedure, and is usually higher than a conventional α of .05. Some attention was therefore given to the questions of an appropriate level of significance to be used as the criterion of rejection, and at what stage in

the selection process the stopping rule should be invoked.

Table 3 presents the results of varying the α level used to select models which describe Model 3 samples, employing backward elimination from a homogeneous baseline model. The results are somewhat surprising. It was thought that use of a less stringent α level (e.g. .10) might improve the accuracy of model selection procedures when samples are small, but this proved not to be the case. Instead, a more stringent α of .01 yielded improved accuracy when averaged over all sample sizes; the small decrease in accuracy for samples of size 100 or less is more than balanced by improvements in accuracy for larger samples. An α-level of .01 implies that a parameter is included in a model only if it is statistically significant at the .01 level. More controversially, an α level of .01 implies that a model is rejected only if it is associated with a probability of .01 or less. It is counterintuitive that a strategy which accepts models which fit so poorly by conventional standards is nevertheless most likely to lead to selection of the true model, especially when samples are large. Similar results are found for forward selection, and hold as well when tested using Model 1 samples. Although further investigation of appropriate α levels is warranted, these results tentatively suggest that an α-level of .01 may produce better results than α-levels of .05 or .10 when samples are 250 cases or larger.

Table 3. Proportion of correct selections for different levels of α.

Sample size	α = .01	.05	.10	.25	.50
50	.06	.10	.10	.00	.00
100	.10	.20	.22	.08	.02
250	.76	.72	.72	.24	.02
500	.90	.76	.66	.38	.02
1000	.92	.84	.66	.26	.04
2000	.98	.96	.41	.42	.14
4000	.98	.86	.68	.28	.06
8000	.94	.86	.72	.38	.06
Total	.71	.66	.52	.26	.04

A second issue is the question of when in the search process stopping rules should be invoked. The stepwise selection procedures used here terminate search when overall goodness of fit of the selected model falls below the specified rejection level. The criterion that the selected model must be associated with a probability of .05 or greater is applied not only to the final selection, but to all intermediate models in the stepwise selection process. This is particularly problematic for backward elimination if, for example, lower order models fit well although higher order models fit poorly. This may occur when deletion of higher order terms adds degrees of freedom but does not much reduce goodness of fit. If the stopping rule is invoked at an intermediate stage, search will terminate before good-fitting, lower order models are tested. It is possible that the search process would be improved if the criterion for overall goodness of fit of a model is applied only to the final model which results from the search.

A possible example of premature termination of the search process occurs when the highest order (four-way) interaction term is statistically significant. In this case, all four selection procedures terminate search and select the saturated model as the final model. For the 2000 samples analyzed in Table 2, the four-way interaction is significant at the .05 level in 118 samples (or for .06 of the replications, which is slightly greater than the chance expectation under the null hypothesis). Of course, in all cases the four-way term represents random variation, since it is included in none of the true models. In addition, in many cases lower order models fit the data well. If the search for a good-fitting model is continued ignoring the significant four-way interaction term, the true model is selected (and fits acceptably) in 48 of the 118 samples. Thus, for the models considered here, the likelihood of selecting a true model is marginally improved if stopping rules are not employed to terminate search.

Discussion

When averaged over sample sizes and model types, the results indicate relatively small overall differences in the performance of difference selection strategies for the data analysis situations simulated here. The principal finding is that direct estimation performs worse than any of the stepwise procedures, due mainly to its relatively poor performance when samples are small and the true model includes a large interaction term. Thus, the results of the simulation suggest that if the researcher has identified the appropriate set of variables to analyze, if the population model is hierarchical and relatively simple, and if one of the stepwise procedures is used, the true model may be selected with probability between about .25 and .90, depending upon the size of the sample. The results suggest that these model selection procedures (particularly direct estimation) should generally not be applied to very small samples ($n < 250$). However, even direct estimation performs quite well for large samples, suggesting that Goodman (1971) may be too cautious in his recommendation against the use of standardized λ's as a simple guide to the selection of models.

When samples are large, the simple strategies analyzed here perform relatively well and about equally. This suggests that there may be little need for the complex, multidirectional selection strategies such as those proposed by Goodman (1971, 1973) and Benedetti and Brown (1976). Of course, it must be emphasized that four-way tables characterized by relatively simple hierarchical models have been simulated; the results reported here may not generalize to larger tables or more complex situations. Nevertheless, an emphasis upon the development of many alternative methods for selecting models may be misplaced. Instead, it may be more appropriate to focus attention upon other important issues concerning the selection of descriptive models for categorical data. One such issue is the neglected problem of how variables should be selected for inclusion in an analysis. Koch and his students have recently developed criteria for the selection of variables (see e.g. Higgins and Koch, 1977) although the adequacy of such procedures has not been evaluated.

Finally, the results reported here are germaine to two points made by Benedetti and Brown (1976). The recommend against the selection of a homogeneous baseline model prior to application of stepwise procedures, because it may lead to the exclusion of relevant parameters. Comparison of the results obtained by backward elimination from a saturated versus homogeneous baseline model indicates that when models are not screened, backward deletion is somewhat more sensitive to the presence of interaction effects (i.e. in Models 2-5) but is also more likely to detect interaction where there is none (i.e. in Model 1). Thus, neither method is superior in all data analysis situations. Benedetti and Brown (1976) also argue that for large samples the difference chi-square should not be used as the criterion for inclusion of terms, and that a more appropriate test would be based upon explained variance. Although the two decision rules are not compared here, the results do not indicate that the difference chi-square is an inappropriate statistical criterion for model selection. The performance of the search procedure may be improved, especially for large samples, by using an α level of .01 as the criterion for acceptance or rejection of models and parameters.

Footnotes

[1] Stepwise model selection is carried out by a computer program (MAT) developed at the University of Chicago and modified at the University of Michigan, the University of North Carolina and Duke University.

[2] The difference chi-square for a model M_1 and a model M_2 containing additional effect parameter(s) is calculated as $\chi^2_1 - \chi^2_2$, with degrees of freedom $df_1 - df_2$. A difference chi-square value associated with $p \leq .05$ indicates that M_2 fits significantly better than M_1, and the term(s) in M_2 should therefore be retained. If $p > .05$ the term(s) in M_2 do not make a significant contribution to goodness of fit and may be deleted.

[3] This procedure differs somewhat from that described by Bishop, Fienberg, and Holland (1975, p. 157-8). Backward elimination is not confined to intervening models which include all terms of order k - 1 and some or all terms of order k, but may delete terms of order k - 2, etc. However, if terms of order k are statistically significant and must be included in a model, then terms of order k - 2 will not be tested or deleted using the stepwise algorithm employed here.

Acknowledgments

This research was supported by NSF Grant #SOC76-02484. I would like to thank Miltiades Damanakis for assistance in computing.

References

Bendell, R.B. and A.A. Afifi. 1977. "Comparison of stopping rules in forward 'stepwise' regression." Journal of the American Statistical Association 72 (357):46-53.

Benedetti, J.K. and M.B. Brown. 1976. "Selecting log-linear models: A comparison of strategies." Paper presented at the Annual Meeting of the American Statistical Association, Boston, Massachusetts, August 1976.

Birch, M.W. 1964. The detection of partial association, II: the general case. Journal of the Royal Statistical Society Series B27: 11-24.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland. 1975. Discrete Multivariate Analysis. Cambridge, Mass: The MIT Press.

Brown, M.B. 1976. Screening effects in multidimensional contingency tables. Applied Statistics 25:37-46.

Dempster, A.P., M. Schatzoff, and N. Wermuth. 1977. "A simulation study of alternatives to ordinary least squares." Journal of the American Statistical Association. 72 (357): 77-91.

Draper, N. and H. Smith. 1966. Applied Regression Analysis. New York: John Wiley.

Goodman, L.A. 1964. "Interactions in multidimensional contingency tables." Annals of Mathematical Statistics 35:632-646.

_____. 1971. "The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications." Technometrics 13 (1):33-61.

_____. 1973. "Guided and unguided methods for the selection of models for a set of T multidimensional contingency tables." Journal of the American Statistical Association 68 (341):165-175.

Higgins, J.E. and G.G. Koch. 1977. "Variable selection and generalized chi-square analysis of categorical data applied to a large cross-sectional occupational health survey." International Statistical Review 45:51-62.

Kennedy, W.F. and T.A. Bancroft. 1971. "Model building for prediction based upon repeated significance tests." The Annals of Mathematical Statistics 42 (4):1273-1284.

GUIDE FOR ANALYZING A FOUR DIMENSIONAL DESIGN (ONE BETWEEN AND THREE WITHIN SUBJECTS) VIA MULTIPLE REGRESSION TECHNIQUE*

Betty B. Carlton Harry R. Barker
University of Alabama

Problem

In recent years there has been increasing use in the behavioral sciences of an analysis of variance design utilizing three within subject and one between subject effects. Recognizing that multiple regression procedure is gradually becoming a preferred way of doing analysis of variance, it seemed appropriate to fit the complex design into a multiple regression framework. The purpose of this paper is to set up a stepwise procedure for doing a four-way analysis of variance (one between, three within subjects) using the multiple regression technique.

Method

The multiple regression utilized for the problem was Corr03 (Barker and Barker, 1977). This program involves input and output features of particular interest to the behavioral scientist; however, most computer programs for regression analysis are satisfactory computer programs. The variables were coded according to the effect coding method described by Overall and Spiegel (1969). A series of multiple regression runs were executed. From the model referred to as the full model, the total sum of squares was obtained. Reduced models then provided the sum of squares for everything except a main or interaction effect. By subtracting each reduced model in turn from the full model the sum of squares related to the omitted effect was found. In many designs not involving repeated measures the residual sum of squares from the full model is the appropriate error term for calculating F ratios. In the design of interest in this paper, however, the residual error term must be broken up into seven different error terms.

The hypothetical example of a four dimensional study used for this paper follows a dissertation from the Department of Psychology of The University of Alabama. The A effect is a between subjects organismic variable consisting of Alcoholics versus Social Drinkers. All remaining effects are within subjects; i.e., the subjects each receive all remaining three treatments. The B effect is a stimulus male or female counselor, the C effect is a positive or negative stimulus scene, and the D effect is a familiar or an unfamiliar stimulus person. The B, C, and D effects are related to scenes shown to the subjects (Alcoholics or Social Drinkers) on a TV monitor. The reaction of the subjects under each of the possible combinations was measured on a single scale providing eight repeated measures for each subject (see figure 1). To simplify the model for the purpose of this paper very simple measures were provided for the scores and the number of subjects per group was limited to two.

Calculations were done first by hand using the classic analysis of variance approach. Multiple regression models were then set up to provide sums of squares corresponding to those that were hand calculated and to develop the procedure for determining correct error terms.

Results

Table 1 illustrates how the variables were coded. Note that the blank spaces in coding columns were interpreted by the computer as zeros, as was intended, and zeros may be added for clarification of the coding procedure if desired.

Table 2 exemplifies how the coded variables were designated for the multiple regression runs to calculate the sums of squares necessary for the analysis of variance computations. The variables used in Table 2 are numbered according to the designated numbers in Table 1. Variable 1, for instance, refers to the A dimension, variable 2 to B, variable 11 to AXBXC, etc.

The first model regresses the dependent variable (#32) onto all main effects and treatment interactions. This model is designated a full model. The value of multiple R2 represents the proportion of the variance of the dependent variable which is associated with all treatments. Models 3 through 17 represent reduced models; that is, the dependent variable is regressed onto all treatment effects except the one treatment under consideration. The difference in multiple R2 for the full and reduced models represents the proportion of variance of the dependent variable associated only with the treatment under consideration. The remaining models are of particular interest in this paper because they enable computation of the numerous error terms required for the F ratios.

Paper presented at The American Statistical Association, August 18, 1977(Chicago, Ill.).

Information provided by the multiple regression models was then used to calculate the F ratios as shown in Table 3.

The A effect in this design is a between subjects effect. As can be seen from the tables a reduced model (Model 2) provided the between subjects error term directly by regression onto the subject variable. For all other main effects and treatment interactions, an error term which involves an interaction with both the treatment and the subjects is required. For the AXB interaction, for instance, an AXBXSubject interaction was needed for the error term; AXC was tested against an AXCXSubjects; AXD, against an AXDXSubjects. Again referring to the tables it can be seen that these subject interactions were coded, regressed onto directly, and used as the error term for the appropriate interaction.

Summary

A procedure has been presented which enables a frequently used analysis of variance design to be accomplished by a multiple regression computer program. The design involves one between subject effect and three within subject effects.

References

Barker, H. R. and Barker, B. M. Behavioral Sciences Statistics Program Library. University of Alabama, 2nd rev. edition, 1977.

Overall, J.E., and Spiegel, D.K. Concerning Least Squares Analysis of Experimental Data. Psychological Bulletin, 1969, 72, 311-322.

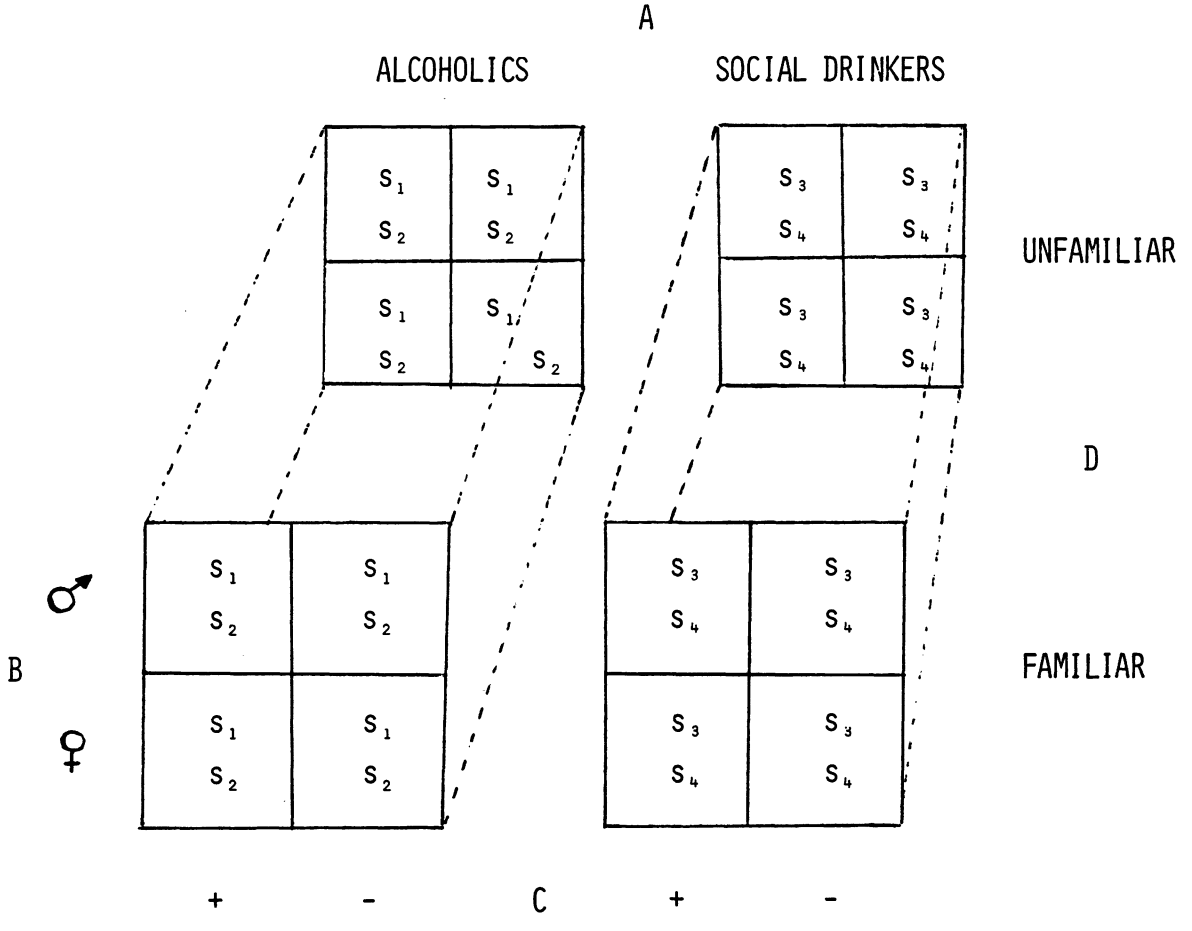


FIGURE 1
FOUR DIMENSIONAL DESIGN

TABLE 1

CODING VARIABLES FOR USE IN MULTIPLE REGRESSION

VARIABLE NUMBER 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
A	B	C	D	A X B	A X C	A X D	B X C	B X D	C X D	A X B X C	A X B X D	A X C X D	B X C X D	A X B X C X D	S_1	S_3	A X B X S_1	A X B X S_3	A X C X S_1	A X C X S_3	A X D X S_1	A X D X S_3	A X B X C X S_1	A X B X C X S_3	A X B X D X S_1	A X B X D X S_3	A X C X D X S_1	A X C X D X S_3	A X B X C X D X S_3	A X B X C X D X S_3	DV
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1		1		1		1		1		1		1		1		1
1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	1		-1		1		1		-1		-1		1		-1		3
1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	1		1		-1		1		-1		1		-1		-1		2
1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	1		-1		-1		1		1		-1		-1		1		4
1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	1		1		1		-1		1		-1		-1		-1		5
1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	1		-1		1		-1		-1		1		-1		1		7
1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	1		1		-1		-1		-1		-1		1		1		6
1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	1		-1		-1		-1		1		1		1		-1		8
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	-1		-1		-1		-1		-1		-1		-1		-1		1
1	-1	1	1	-1	1	1	-1	-1	1	-1	-1	1	-1	-1	-1		1		-1		-1		1		1		-1		1		3
1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	-1	-1	-1		-1		1		-1		1		-1		1		1		3
1	-1	-1	1	-1	-1	1	1	-1	-1	1	-1	-1	1	1	-1		1		1		-1		-1		1		1		-1		7
1	1	1	-1	1	1	-1	1	-1	-1	1	-1	-1	-1	-1	-1		-1		-1		1		-1		1		1		1		3
1	-1	1	-1	-1	1	-1	-1	1	-1	-1	1	-1	1	1	-1		1		-1		1		1		-1		1		-1		6
1	1	-1	-1	1	-1	-1	-1	-1	1	-1	-1	1	1	1	-1		-1		1		1		1		1		-1		-1		9
1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1		1		1		1		-1		-1		-1		1		4
-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1		1		-1		-1		-1		-1		-1		-1		-1	1
-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1		1		1		-1		-1		1		1		-1		1	7
-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1		1		-1		1		-1		1		-1		1		1	4
-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1		1		1		1		-1		-1		1		1		-1	7
-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1		1		-1		-1		1		-1		1		1		1	3
-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1		1		1		-1		1		1		-1		1		1	7
-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1		1		-1		1		1		1		-1		1		-1	4
-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1		1		1		1		1		-1		-1		-1		1	9
-1	1	1	1	-1	-1	-1	1	1	1	-1	-1	-1	1	-1		-1		1		1		1		1		1		1		1	2
-1	-1	1	1	1	-1	-1	-1	-1	1	1	1	-1	-1	1		-1		-1		1		1		-1		-1		1		-1	5
-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1		-1		1		-1		1		-1		1		-1		-1	5
-1	-1	-1	1	1	1	-1	1	-1	-1	-1	1	1	1	-1		-1		-1		-1		1		1		-1		-1		1	9
-1	1	1	-1	-1	-1	1	1	-1	-1	-1	1	1	-1	1		-1		1		1		-1		1		-1		-1		-1	1
-1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	1	-1		-1		-1		1		-1		-1		1		-1		1	6
-1	1	-1	-1	-1	1	1	-1	-1	1	1	1	-1	1	-1		-1		1		-1		-1		-1		-1		1		1	4
-1	-1	-1	-1	1	1	1	1	1	1	-1	-1	-1	-1	1		-1		-1		-1		-1		1		1		1		-1	7

TABLE 2

VARIABLES USED IN MULTIPLE REGRESSION PROGRAMS

	DEP. VAR. #	INDEPENDENT VARIABLE NUMBERS (VARIABLE NUMBERS FROM TABLE 1)
MODEL 1	32	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 2	32	16 17
MODEL 3	32	2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 4	32	1 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 5	32	1 2 4 5 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 6	32	1 2 3 5 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 7	32	1 2 3 4 6 7 8 9 10 11 12 13 14 15 16 17
MODEL 8	32	1 2 3 4 5 7 8 9 10 11 12 13 14 15 16 17
MODEL 9	32	1 2 3 4 5 6 8 9 10 11 12 13 14 15 16 17
MODEL 10	32	1 2 3 4 5 6 7 9 10 11 12 13 14 15 16 17
MODEL 11	32	1 2 3 4 5 6 7 8 10 11 12 13 14 15 16 17
MODEL 12	32	1 2 3 4 5 6 7 8 9 11 12 13 14 15 16 17
MODEL 13	32	1 2 3 4 5 6 7 8 9 10 12 13 14 15 16 17
MODEL 14	32	1 2 3 4 5 6 7 8 9 10 11 13 14 15 16 17
MODEL 15	32	1 2 3 4 5 6 7 8 9 10 11 12 14 15 16 17
MODEL 16	32	1 2 3 4 5 6 7 8 9 10 11 12 13 15 16 17
MODEL 17	32	1 2 3 4 5 6 7 8 9 10 11 12 13 14 16 17
MODEL 18	32	18 19
MODEL 19	32	20 21
MODEL 20	32	22 23
MODEL 21	32	24 25
MODEL 22	32	26 27
MODEL 23	32	28 29
MODEL 24	32	30 31

TABLE 3

MODELS FOR SUBTRACTION PROCESS

MODELS	TREATMENT	ERROR TERM	MODELS
1-3	A	BETW.S	2
1-4	B	ABS	18
1-5	C	ACS	19
1-6	D	ADS	20
1-7	AB	ABS	18
1-8	AC	ACS	19
1-9	AD	ADS	20
1-10	BC	ABCS	21
1-11	BD	ABDS	22
1-12	CD	ACDS	23
1-13	ABC	ABCS	21
1-14	ABD	ABDS	22
1-15	ACD	ACDS	23
1-16	BCD	ABCDS	24
1-17	ABCD	ABCDS	24

ESTIMATION OF RARE EVENTS BY SIMPLE CLUSTER SAMPLING WITH MULTIPLICITY

PAUL S. LEVY
University of Illinois at the Medical Center
School of Public Health

1. Introduction

The use of sample surveys with multiplicity has been advocated by Sirken (1970) for estimating the number of demographic events (e.g. births, deaths) occurring in a particular time period. Since Sirken's original article, the theory of multiplicity estimation (also called network sampling) has been extended to stratified random sampling (Sirken, 1972; Levy, 1977), estimation of proportions (Sirken and Levy, 1974) and estimation of response errors (Nathan, 1976). In addition, sample surveys with multiplicity have been used in a wide variety of applications (Sirken, 1972; Sirken and Levy, 1974; Sirken et al., 1975; Nathan et al., 1977). In this report, the theory of multiplicity estimation is extended to simple cluster sampling, and an unbiased estimator is proposed for estimating the total number of events under this type of sampling design.

A survey with multiplicity is one in which an element (e.g. birth, death, individual having some attribute, etc.) may be linked to more than one enumeration unit by an algorithm or counting rule. For example, a counting rule in a survey with multiplicity might link a birth to the households of the grandparents as well as to the parents' household whereas a conventional counting rule would link the birth only to the household of the parents.

2. Development of the Estimator:

2.1 Population Parameters

Let us suppose that a population contains N enumeration units (e.u.'s) grouped into M primary sampling units (PSU's) with PSU i containing N_i e.u.'s; i = 1, ..., M, and that a counting rule links Y events labeled $I_1, \ldots, I_Y$ to enumeration units according to an indicator variable, $\delta_{\alpha ij}$ given by

$$\delta_{\alpha ij} = \begin{cases} 1 & \text{if event } I_\alpha \text{ is linked to e.u. } j \text{ in PSU } i \text{ by the counting rule} \\ 0 & \text{otherwise} \end{cases}$$

where
$\alpha = 1, \ldots, Y$, $i = 1, \ldots, M$ and $j = 1, \ldots, N_i$.

For any counting rule, the following parameters can be defined which characterize the network linking the enumeration units to the elements:

$$s_{\alpha i} = \sum_{j=1}^{N} \delta_{\alpha ij} ;$$

$$s_\alpha = \sum_{i=1}^{M} s_{\alpha i} ;$$

$$t_{\alpha i} = \begin{cases} 1 & \text{if } s_{\alpha i} > 0 \\ 0 & \text{if } s_{\alpha i} = 0 \end{cases}$$

$$t_\alpha = \sum_{i=1}^{M} t_{\alpha i}$$

The parameter, $S_{\alpha i}$, denotes the number of enumeration units in a particular PSU (i) that are linked to a particular element, I_α, whereas s_α denotes the total number of enumeration units linked to an element by a counting rule and is referred to as the multiplicity of the element with respect to the counting rule. Clearly, for conventional counting rule, each s_α would be equal to unity. The parameter, t_α, denotes the number of PSU's in which a particular element, I_α, is linked to one or more enumeration units.

Let $(z_{\alpha i} : \alpha = 1, \ldots, Y; i = 1, \ldots, M; t_{\alpha i} = 1)$ be any set of weights defined for all (α, i) such that $t_{\alpha i} = 1$ with the property:

$$\sum_{i=1}^{M} z_{\alpha i} \, s_{\alpha i} = 1 \qquad \alpha = 1, \ldots, Y$$

We then define the following parameters for each PSU.

$$\lambda'_{ij} = \sum_{\alpha=1}^{Y} z_{\alpha i} \, \delta_{\alpha ij} , \quad j = 1, \ldots, N_i$$

$$Y_i^* = \sum_{j=1}^{N_i} \lambda'_{ij} \; ; \quad E_i = \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \, s_{\alpha i} / Y_i^*$$

The parameter, λ'_{ij}, represents the basic summary information obtained from enumeration units concerning elements, while the weights $\{z_{\alpha i}\}$ are functions of the particular network linking enumeration to elements and are chosen to make estimates of Y unbiased. Some possible choices of $z_{\alpha i}$ might be $1/s_\alpha$ or $1/(s_{\alpha i} t_\alpha)$ for those counting rules which link elements to enumeration units in more than one PSU. For counting rules which link elements to enumeration units in only one PSU, the $z_{\alpha i}$ might be set equal to $1/s_{\alpha i}$, and for conventional counting rules, the $z_{\alpha i}$ would be equal to 1. The E_i are generalizations of parameters found by Sirken and Levy (1974) to be involved in the variances of

estimates obtained from multiplicity surveys, while the parameters, Y_i^*, although not necessarily integer valued, could be interpreted as being the "effective number" of elements linked to PSU i by the enumeration rule. It can be shown that $\sum_{i=1}^{M} Y_i^* \equiv Y$

Using nomenclature similar to that in Hansen, Hurwitz, and Madow (1953), we define for the variable, λ'_{ij}, the between PSU variance

$$S_{1Y}^{2*} = \sum_{i=1}^{M} (Y_i^* - \bar{Y})^2 \; / \; (M - 1)$$

the within PSU variance,

$$S_{2Y}^{2*} = \frac{1}{\bar{N}} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (\lambda'_{ij} - \bar{Y}_i^*)^2$$

and the intra-class correlation coefficient

$$\delta = \left(\frac{M-1}{M} S_{1Y}^{2*} - \bar{N} S_{2Y}^{2*} \right) / \left(\frac{M-1}{M} S_{1Y}^{2*} + \bar{N} (\bar{N}-1) S_{2Y}^{2*} \right)$$

where

$$\bar{Y} = \sum_{i=1}^{M} Y_i^* / M = Y/M \text{ (since } \sum_{i=1}^{M} Y_i^* = Y\text{)} ,$$

and

$$\bar{N} = \sum_{i=1}^{M} N_i / M .$$

With these definitions, the following theorems can be proved.

Theorem 1.

The variance, S_{1Y}^{2*} among PSU's with respect to λ'_{ij} is equal to the expression given by:

$$S_{1Y}^{2*} = \left[M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'} \right] / (M-1).$$

where

$$E_k = \sum_{\alpha} \sum_{i} z_{\alpha i}^2 \; s_{\alpha i}^2 \; / \; Y$$

and

$$v_{\alpha\alpha'} = \sum_{i} z_{\alpha i} \; z_{\alpha' i} \; s_{\alpha i} \; s_{\alpha' i} \; ; \; \alpha' \neq \alpha.$$

PROOF

$$\sum_{i=1}^{M} (Y_i^* - \bar{Y})^2 = \sum_{i=1}^{M} (Y_i^*)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \left(\sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} z_{\alpha i} \; \delta_{\alpha ij} \right)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \left(\sum_{\alpha=1}^{Y} z_{\alpha i} \; s_{\alpha i} \right)^2 - M \bar{Y}^2$$

$$= \sum_{i=1}^{M} \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \; s_{\alpha i}^2 + \sum_{i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} s_{\alpha i} s_{\alpha' i} - M \bar{Y}^2$$

$$= M \bar{Y} \; (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

q.e.d.

Theorem 2.

The within PSU variance with respect to λ'_{ij} is given by

$$S_{2Y*}^2 = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} \bar{Y}_i^* \; (E_i - \bar{Y}_i^*) + \frac{1}{N} \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'}$$

where $v_{i\alpha\alpha'} = \sum_{i=1}^{N_i} z_{\alpha i} \; z_{\alpha' i} \; \delta_{\alpha ij} \; \delta_{\alpha' ij}$ for $\alpha' \neq \alpha$

PROOF

$$\sum_{i=1}^{M} \frac{N_i}{N_i - 1} \sum_{j=1}^{N_i} (\lambda'_{ij} - \bar{Y}_i^*)^2$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left(\sum_{j=1}^{N_i} (\lambda'_{ij})^2 - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left(\sum_{j=1}^{N_i} \left(\sum_{\alpha=1}^{Y} z_{\alpha i} \; \delta_{\alpha ij} \right)^2 - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left(\sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} z_{\alpha i}^2 \delta_{\alpha ij}^2 + \sum_{j=1}^{N_i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} \delta_{\alpha ij} \delta_{\alpha' ij} - N_i (\bar{Y}_i^*)^2 \right).$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left(\sum_{\alpha=1}^{Y} z_{\alpha i}^2 s_{\alpha i} + \sum_{i=1}^{N_i} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} z_{\alpha i} z_{\alpha' i} \delta_{\alpha i j} \delta_{\alpha' i j} - N_i (\bar{Y}_i^*)^2 \right)$$

$$= \sum_{i=1}^{M} \frac{N_i}{N_i - 1} \left(N_i \bar{Y}_i^* (E_i - \bar{Y}_i^*) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'} \right)$$

q.e.d.

Corollary 1.

The intra class correlation coefficient, δ, is given by

$$\delta = \frac{A - B}{A + (N-1)B}$$

where

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} \frac{N_L^2}{N_i - 1} \bar{Y}_i^* (E_i - \bar{Y}_i^*) + \sum_{i=1}^{M} \frac{N_L^2}{N_i - 1} \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{i\alpha\alpha'}$$

Proof follows from Theorems 1 and 2 and the definition of δ.

q.e.d.

Corollary 2.

If the assumption is made that an enumeration unit is linked to no more than one element then $v_{i\alpha\alpha'} = 0$ for all i, α, and α'; S_{2Y}^{2*} is given by

$$S_{2Y}^{2*} = \frac{1}{N} \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} (E_i - \bar{Y}_i^*)$$

and

$$\delta = \frac{A - B}{A + (N-1)B}$$

where

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\alpha=1}^{Y} \sum_{\substack{\alpha'=1 \\ \alpha' \neq \alpha}}^{Y} v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} \frac{N_i^2}{N_i - 1} \bar{Y}_i^* (E_i - \bar{Y}_i^*)$$

2.2 Estimation of Y, the Total Number of Events from the Sample

Let us assume that the sample design used to estimate the number of events, Y, in the population is a simple two stage cluster sample as defined by Hansen, Hurwitz, and Madow (1953). In other words, a simple random sample of m PSU's is taken from the M PSU's in the population, and within each sample PSU (i), a simple random sample of n_i enumeration units is taken from the N_i enumeration units in the PSU, with the second stage sampling fraction, n_i/N_i, the same for each PSU.

If (for convenience) the sample PSU's are labelled 1, ..., m and the sample enumeration units within each sample PSU are denoted i_1, ..., i_{n_i} where i = 1, ..., m, then the estimator Y' of Y as given by

$$Y' = \frac{M}{m} \sum_{i=1}^{m} \frac{N_i}{n_i} \sum_{j=1}^{n_i} \lambda'_{ij} \quad (1)$$

is an unbiased estimator of Y as shown below in Theorem 3.

Theorem 3.

The estimator, Y' of Y as defined in equation (1) is an unbiased estimator of Y.

PROOF

For a given sample PSU, i, the expected value over all possible second stage samples of $\sum_{j=1}^{n_i} \lambda'_{ij}$ is given by

$$E\left(\sum_{j=1}^{n_i} \lambda'_{ij}\right) = \frac{n_i}{N_i} \sum_{j=1}^{N_i} \lambda'_{ij} = \frac{n_i}{N_i} Y_i^*$$

Thus, the expected value of Y' over all possible samples is given by

$$E(Y') = \frac{M}{m} \underset{\text{all } i}{E} \left(\sum_{i=1}^{M} \frac{N_i}{n_i} \underset{\text{all } j}{E} \left(\sum_{j=1}^{n_i} \lambda_{ij} \right) \right)$$

$$= \frac{M}{m} \frac{m}{M} \sum_{i=1}^{M} Y_i^*.$$

$$= \sum_{i=1}^{M} Y_i^*.$$

$$= Y$$

q.e.d.

3. Some Relationships Involving δ When all $v_{i\alpha\alpha'} = 0$

When $v_{i\alpha\alpha'} = 0$ for all α, α' and for all i, then the intraclass correlation coefficient, δ, is given by

$$\delta = (A - B) \;/\; (A + (\bar{N} - 1)\, B)$$

where (assuming $N_i / (N_i - 1) \doteq 1$)

$$A = M \bar{Y} (E_k - \bar{Y}) + \sum_{\substack{\alpha \\ \alpha' \neq \alpha}} \sum v_{\alpha\alpha'}$$

and

$$B = \sum_{i=1}^{M} N_i \bar{Y}_i^* (E_i - \bar{Y}_i^*) .$$

Clearly $A \geq 0$ and $B \geq 0$ since they are quadratic forms. It can be shown that since $B > 0$

$$\frac{d\,\delta}{d\,A} = \frac{B}{[A + (\bar{N} - 1)\, B]^2} \geq 0 ,$$

and therefore, δ varies directly with A.

On the other hand, since $A > 0$, then

$$\frac{d\,\delta}{d\,B} = \frac{-\bar{N} A}{[A + (\bar{N} - 1)\, B]^2} \leq 0$$

and hence δ varies inversely with B.

Let us examine δ for the set of weights $z_{\alpha i} = 1/(t_\alpha s_{\alpha i})$; $\alpha = 1, \ldots, Y$; $i = 1, \ldots, M$. For this set of weights, we have:

$$E_i = \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha^2 s_{\alpha i}} \;/\; \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha} ,$$

$$\bar{Y}_i^* = \frac{1}{N-1} \sum_{\alpha=1}^{Y} \frac{t_{\alpha i}}{t_\alpha} ,$$

$$E_k = \sum_{\alpha=1}^{Y} \frac{1}{t_\alpha} \;/\; Y$$

and

$$\sum_{\substack{\alpha \\ \alpha' \neq \alpha}} \sum_{\alpha'} v_{\alpha\alpha' i} = \sum_{\alpha} \sum_{\alpha'} \sum_{i} \frac{t_{\alpha i}\, t_{\alpha' i}}{t_\alpha\, t_{\alpha'}}$$

If the multiplicities, $s_{\alpha i}$, are increased without increasing the t_α or $t_{\alpha i}$, then the E_i would decrease which would cause a decrease in B since the $\bar{Y}_i^*$ would be unaffected. Since, also, the $v_{\alpha\alpha'}$, and E_k would not be affected by change in the $s_{\alpha i}$, it follows that A would not be affected. Hence, <u>increase</u> in $s_{\alpha i}$ would result in an <u>increase</u> in the intra-class correlation coefficient, δ.

REFERENCES

1. Sirken, M.G. (1970): "Household Surveys with Multiplicity" <u>Journal of the American Statistical Association</u> <u>65</u>, 257 - 266.

2. Sirken, M.G. (1972): "Stratified Sample Surveys with Multiplicity" <u>Journal of the American Statistical Association</u> <u>67</u>, 224 - 227.

3. Levy, P. S. (1978): "Optimum Allocation in Stratified Random Network Sampling for Estimating the Prevalence of Attributes in Rare Populations" To appear in <u>Journal of the American Statistical Association, December, 1977.</u>

4. Sirken, M.G. and Levy, P.S. (1974): "Multiplicity Estimation of Proportions Based on Ratios of Random Variables" <u>Journal of the American Statistical Association</u> <u>69</u>, 68 - 73.

5. Sirken, M. G., Indurfurth, G. P., Burnham, C. E., and Danchik, K. M. (1975): "Household Sample Surveys of Diabetes: Design Effects of Counting Rules" <u>American Statistical Association, Proceedings of the Social Statistics Section</u>, 659 - 663.

6. Nathan, G., Schmelz, U. O. and Kenvin, J. (1977): "Multiplicity Study of Marriages and Births in Israel" <u>Vital and Health Statistics, Series 2</u> No. 70, NCHS, Rockville, Maryland.

7. Nathan, G. (1976): "An Empirical Study of Response and Sampling Errors for Multiplicity Estimates with Different Counting Rules" <u>Journal of the American Statistical Association 71,</u> 808 - 815.

ASYMPTOTIC DISTRIBUTION OF AGREEMENT SCORE (KAPPA) WITH FIXED MARGINAL TOTALS

Y. S. Lin, Booz, Allen & Hamilton, Inc.

I. INTRODUCTION

One particular form of contingency table (ordered I x I table) gives rise to a special problem of statistical interest - measurement of agreement. Suppose two raters independently categorize items or responses among the same set of nominal categories, and we wish to develop a measure of agreement for these raters. This problem can be viewed as one of measuring the reliability between two raters. Goodman and Kruskal (1954) suggested that for the situation when each of the r raters independently assigns N responses (one to each of the N objects) among I categories a measure of agreement, adjusted for chance, among r raters is needed.

Many coefficients of relative agreement measure have been proposed within the last two decades. The more widely used agreement coefficient has been the one called Kappa that was suggested by Cohen (1960) and others. Kappa coefficient for a two rater is defined as:

$$\kappa = (\theta_1 - \theta_2)/(1-\theta_2) \qquad (1)$$

where $\theta_1 = \sum_i P_{ii}$ and $\theta_2 = \sum_i P_{i.}\, P_{.i}$,

P_{ij} = true proportion that an object is assigned by rater 1 to category i and in category j by rater 2 .

Let X_{ij} be the number of objects assigned to (i,j) cell in the ordered I x I contingency table and $N = \sum_i \sum_j X_{ij}$. The maximum likelihood estimator for κ under the multinomial sampling situation is:

$$\hat{\kappa} = (\hat{\theta}_1 - \hat{\theta}_2)/(1-\hat{\theta}_2) \qquad (2)$$

where $\hat{\theta}_1 = \sum_i X_{ii}/N$ and $\hat{\theta}_2 = \sum_i X_{i.}\, X_{.i}/N^2$.

II. ASYMPTOTIC DISTRIBUTION OF $\hat{\kappa}$ WITH FIXED MARGINAL TOTALS

The asymptotic variance of $\hat{\kappa}$ as given by some authors (e.g., Cohen {1960, 1968}, Fleiss {1971}, Marx and Light {1973}) is of the form:

$$\mathrm{Var}(\hat{\kappa}) = \begin{cases} \dfrac{\theta_1(1-\theta_1)}{N(1-\theta_2)^2} & \text{for non-null case} \\ \dfrac{\theta_2}{N(1-\theta_2)} & \text{for null case.} \end{cases} \qquad (3)$$

It was later shown (by Fleiss, Cohen & Everitt {1969} and Bishop, Fienberg & Holland {1975}, etc.) that the expressions in (3) are not correct for sampling situations without fixed marginal totals under both null (raters are independent) and non-null cases. One might think that the asymptotic variances given in (3) are appropriate for the situation with fixed marginal totals. In this paper, we obtain the conditional (on both marginals) asymptotic variances for $\hat{\kappa}$ for both null and non-null cases and compare them to that of (3).

Because of the computational difficulty, we use the simplest case of two rating categories (I=2). Let P_{ij} (i,j = 1,2) be the probabilities for (i,j) cell and X_{ij} denote the (i,j) cell counts obtained in the experiment. Assume $X_{1.}$, $X_{2.}$, $X_{.1}$ and $X_{.2}$ are fixed with $X_{1.} + X_{2.} = N$. Using the $\hat{\kappa}$ as defined in (2) to obtain the variance of $\hat{\kappa}$, we need to get the variance of $(X_{11} + X_{22})$ conditional on the marginal totals: $X_{1.}$ and $X_{.1}$. The conditional distribution of X_{11} given the marginal totals is obtained by Harkness and Katz {1964} to be the "extended hypergeometric distribution".

$$f(X_{11} \mid X_{1.}, X_{.1}) = g(X_{1.}, X_{.1}, t) \binom{X_{1.}}{X_{11}} \binom{X_{2.}}{X_{.1}-X_{11}} t^{X_{11}} \qquad (4)$$

where

$$g(X_{1.}, X_{.1}, t) = \left[\sum_a \binom{X_{1.}}{a} \binom{X_{2.}}{X_{.1} - a} t^a \right]^{-1}$$

and $t = P_{11}P_{22}/P_{12}P_{21}$; with $0 < t < \infty$. In the general non-null situation, we can replace P_{11} by $\lambda P_{1.}P_{.1}$, P_{12} by $P_{1.}(1-\lambda P_{.1})$, P_{21} by $P_{.1}(1-\lambda P_{1.})$ and P_{22} by $1-P_{1.}-P_{.1}+\lambda P_{1.}P_{.1}$, where

$$\mathrm{Max}\left(0, \frac{P_{1.} + P_{.1} - 1}{P_{1.}\, P_{.1}}\right) < \lambda < \mathrm{Min}\left(\frac{1}{P_{1.}}, \frac{1}{P_{.1}}\right),$$

and $t = \dfrac{\lambda\,(1 - P_{1.} - P_{.1} + \lambda\, P_{1.}\, P_{.1})}{(1 - \lambda\, P_{1.})\,(1 - \lambda\, P_{.1})}$.

Under the null hypothesis of independence where $P_{ij} = P_{i.}P_{.j}$ (i,j = 1,2), then t = 1 (i.e., λ=1) and the expression (4) reduces to the ordinary hypergeometric distribution

$$f(X_{11} \mid X_{1.}, X_{.1}) = \frac{\binom{X_{1.}}{X_{11}} \binom{N - X_{.1}}{X_{.1} - X_{11}}}{\binom{N}{X_{.1}}} \qquad (5)$$

The conditional distribution given in (5) is generally used to perform the exact test of independence for a 2x2 contingency table with small samples. For the large sample case (as $N \to \infty$) with $X_{1.}/N \to P_{1.}$ and $X_{.1}/N \to P_{.1}$, Harkness and Katz {1964} obtain the asymptotic mean and variance of X_{11} as

$$E(X_{11} \mid X_{1.}, X_{.1}) = X_{1.}Q = \lambda^* \frac{X_{1.}X_{.1}}{N},$$

$$\text{Var}\,(X_{11} \mid X_{1.}, X_{.1}) = \left(\sum_{i,\,j\,=\,1}^{2} \frac{1}{P^*_{ij}} \right)^{-1}, \qquad (6)$$

where

$$Q = \frac{-d + \left[d^2 + 4X_{1.}X_{.1}t\,(1-t) \right]^{\frac{1}{2}}}{2\,(1-t)\,X_{1.}}$$

$$d = N - (X_{1.} + X_{.1})\,(1-t)$$

$$\lambda^* = \frac{NQ}{X_{.1}},$$

and $\{P_{ij}^*\}$ are $\{P_{ij}\}$ expressed in terms of λ^*, $X_{.1}/N$ in place of λ, $P_{i.}$ and $P_{.j}$. In the null case, $t = 1$, (i.e., $\lambda^* = 1$) then

$$E_0\,(X_{11} \mid X_{1.}, X_{.1}) = \frac{X_{1.}X_{.1}}{N}$$

and

$$\text{Var}_0\,(X_{11} \mid X_{1.}, X_{.1}) = \frac{X_{1.}X_{.1}X_{2.}X_{.2}}{N^2\,(N-1)}. \qquad (7)$$

Another way of obtaining the conditional variance of X_{11} given the marginal totals $X_{1.}$ and $X_{.1}$ under the null case is to use a lemma due to Hinkley (1974).

Lemma (Hinkley)

If $S = S(X)$ is complete minimal sufficient for such that $E_X\,(a\,(X);\,\theta) = b\,(\theta)$
and $E_S\,(c\,(S);\,\theta) = b\,(\theta)$,
then $E_{X|S}\,(a\,(x) \mid S) = c\,(s)$. (8)

Using the lemma above, we need to show that

$$E_{X|S}\,(X_{11} - E_{11})^2 = \frac{X_{1.}X_{.1}X_{2.}X_{.2}}{N^2\,(N-1)},$$

where

$$E_{11} = E_0\,(X_{11} \mid X_{1.}, X_{.1}) = \frac{X_{1.}X_{.1}}{N}.$$

Let $Q = \left(X_{11} - \frac{X_{1.}X_{.1}}{N} \right)^2$.

and $E\,(Q \mid S) = E\,(X_{11}^2 \mid S) - \frac{X_{1.}X_{.1}}{N}^2$,

then $E(X_{11}^2) = NP_{1.}P_{.1} + N\,(N-1)\,P_{1.}^2P_{.1}^2 = b\,(\theta)$. Now we need to construct $c\,(s)$ such that (8) is satisfied. This is accomplished by first showing that

$$E\,(X_{1.}^2X_{.1}) = N^2\,(N-1)\,P_{11}P_{1.} + N^2\,P_{11},$$

$$E\,(X_{1.}X_{.1}^2) = N^2\,(N-1)\,P_{11}P_{.1} + N^2P_{11},$$

and

$$E(X_{1.}^2X_{.1}^2) = N^2\,(N-1)^2P_{11}^2 + N^2\,(N-1)P_{11}(P_{1.}+P_{.1}) + N^2P_{11}.$$

Hence

$$c(s) = (X_{1.}^2X_{.1}^2 - X_{1.}X_{.1}^2 - X_{1.}^2X_{.1} + NX_{1.}X_{.1})/(N(N-1)).$$

thus,

$$E\,(c\,(s)\,;\,\theta) = b\,(\theta).$$

Substituting $c\,(s)$ into $E\,(Q \mid S)$ above, we get:

$$E\,(Q \mid S) = \frac{X_{1.}^2X_{.1}^2}{N}\left(\frac{1}{N-1} - \frac{1}{N} \right) + \frac{NX_{1.}X_{.1} - X_{1.}X_{.1} - X_{1.}X_{.1}^2}{N\,(N-1)} = \frac{X_{1.}X_{.1}X_{2.}X_{.2}}{N^2\,(N-1)},$$

which is the conditional variance of X_{11} under null case as given in (7). Since

$$\text{Var}\,(X_{11}+X_{22} \mid X_{1.}, X_{.1}) = \text{Var}\,(X_{11} + (N - X_{1.} - X_{.1} + X_{11}) \mid X_{1.}, X_{.1}) = \text{Var}\,(2X_{11} \mid X_{1.}, X_{.1}) = 4\,\text{Var}\,(X_{11} \mid X_{1.}, X_{.1})$$

we get the asymptotic variance of $\hat{\kappa}$, given $X_{1.}$ and $X_{.1}$, as

$$\text{Var}\ (\hat{\kappa}\,|X_{1.},X_{.1}) = \frac{4\text{Var}\,(X_{11}|X_{1.},X_{.1})}{N^2\left(1-\sum_i \frac{X_{i.}X_{.i}}{N^2}\right)} \quad (9)$$

Where var $(X_{11}|X_{1.},X_{.1})$ are given in either (7) or (6) depending upon whether or not the hypothesis of independence is assumed. Thus, for the null case, the asymptotic variance of $\hat{\kappa}$ given the marginal totals is

$$\text{Var}_0(\hat{\kappa}|X_{1.},X_{.1}) = \frac{4\,X_{1.}X_{.1}X_{2.}X_{.2}}{N^4(N-1)\left(1-\sum_i \frac{X_{i.}X_{.i}}{N^2}\right)} \quad (10)$$

$$= \frac{4\,P_{1.}P_{.1}P_{2.}P_{.2}}{(N-1)\left(1-\sum_i \frac{X_{i.}X_{.i}}{N^2}\right)} .$$

Comparing (10) to the second expression in (3), we see that the asymptotic variance for null case as given by some authors referred to earlier is incorrect even for the conditional situation.

Applying (6) to (9), we obtain the non-null asymptotic variance of $\hat{\kappa}$ with fixed marginals to be

$$\text{Var}\ (\hat{\kappa}\,X_{1.},X_{.1}) = \frac{4}{N^2\left(\sum_{i,j}\frac{1}{P_{ij}^*}\right)\left(1-\sum_i \frac{X_{i.}X_{.i}}{N^2}\right)}, \quad (11)$$

where P_{ij}^* are defined as before.

This conditional asymptotic variance of $\hat{\kappa}$ for the non-null case, as given in (10), is also different from the first expression of (3). Thus, we concluded that the asymptotic variances as given in (3) are not correct either for the unconditional or for the conditional cases.

REFERENCES

Goodman, L.A. and Kruskal, W.H. (1954), "Measures of Association for Cross Classifications", JASA, 49, 732-64.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales", Education and Psychological Measurement, 20, No. 1, 37-46.

Cohen, J. (1968), Weighted Kappa: Nominal Scale Agreement With Provision For Scaled Disagreement or Partial Credit", Psychological Bulletin, 70, No. 4,213-20.

Fleiss, J.L. (1971), "Measuring Nominal Scale Agreement Among Many Raters",Psychological Bulletin, 76, No. 5, 378-82.

Light, R.J. (1971), "Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternations", Psychological Bulletin, 76, No.5, 365-77.

Marx, T.J. and Light, R.J. (1973), "A Many Observer Agreement Measure for Qualitative Response Data", Mimeograph, Laboratory of Human Development, Harvard Graduate School of Education, Cambridge, Massachusetts.

Fleiss, J.L. Cohen, J. and Everitt, B.S. (1969) "Large Sample Standard Errors for Kappa and Weighted Kappa", Psychological Bulletin, 72, No.5, 323-27.

Bishop, Y.M.M. Fienberg, S.E. and Holland, P.W. (1975), Discrete Multivariate Analysis - Theory and Practices, MIT Press, Cambridge Massachusetts.

Harkness, W.L. and Katz, L. (1964), "Comparison of the Power Functions For The Test of Independence In 2x2 Contingency Tables", Annal of Math. Stat., 1115-27.

Hinkley, D. (1974), "On Expectation Conditional On A Sufficient Statistic", University of Minnesota School of Statistics, Technical Report No. 237.

THE EFFECT OF THE 55 MPH SPEED LIMIT ON MOTOR VEHICLE DEATH RATES IN MARYLAND

Jeremiah J. German, Towson State University

The use of the roadway is of obvious economic concern. The "oil crisis" late in 1973 resulted in the reduction of maximum speed limits in 1974 for the purpose of conserving gasoline. This reduction in speed limits has had other consequences. Goods transported on high speed roadways have incurred higher transportation costs. This reduced speed limit has increased costs to consumers directly as well in the greater length of time consumed in personal trips. Unfortunately for public policy considerations, costs have not been publicized as much as the presumed gains of reduced gasoline consumption and the serendipitous benefit of lower motor vehicle death rates. Any study of the economic impact of the lowered speed limits must include the lives saved as well as reduced gas consumption on the benefit side. This paper examines the available evidence on the effectiveness of the reduced speed limit on motor vehicle deaths in Maryland.

The Federal-Aid Highway Amendments of 1974 voted by Congress in December of 1973 established a national speed limit of 55 MPH to take effect on January 4, 1974. Some states had lowered their speed limits in the last weeks of 1973. For convenience, the entire year of 1973 is represented in the period prior to adoption of a national 55 MPH speed limit while the years 1974 and 1975 are represented as the years in which the 55 MPH speed limit was in existence.

The absolute number of motor vehicle deaths in the U.S. as well as in Maryland has been increasing in an irregular pattern since the end of World War II. In this study, the examination of the absolute number of motor vehicle deaths in Maryland was limited to a five year period; a three year period prior to and a two year period following the imposition of the 55 MPH speed limit.

Primary data in the form of several thousand motor vehicle death records were classified as to place of occurence. Table I classifies motor vehicle deaths which took place on low speed roads and those which took place on roads having speed limits of 60 MPH or greater prior to the imposition of the 55 MPH speed limit.

Since 1973, the absolute number of deaths has declined in low speed as well as in previously high speed roads. Between 1973 and 1974, the absolute decline in total deaths in Maryland was 88, of which 22 occurred on roads which had been high speed. During 1974, deaths declined 21% on previously high speed roads as compared to a decline of only 10% on low speed roads. Before attributing this relatively greater decline in deaths on former high speed roads to the lowering of speed limits, an examination of the change that took place in the preceding year, 1973, a period with the speed limits unchanged, shows an increase in deaths for low speed roads and an absolute decrease of 23 deaths, or a decline of 19%, on high speed roads.

Change in road usage between high speed and low speed roads were not available. However, one would expect that a reduction in the speed limit would reduce some of the advantages of using high speed roads relative to low speed roads and would result in some incremental shift to the use of low speed roads. Therefore, comparisons of the absolute number of deaths before and after the change in speed limits would lead to an overestimate of the reduction in deaths attributable to the lowered speed limit.

TABLE 1

MARYLAND MOTOR VEHICLE DEATHS 1971 - 1975

Year	Total Deaths	Deaths on Routes $<$ 60 MPH Speed Limit	Deaths on Routes $\geq$ 60 MPH Speed Limit Prior to 1974
1971	795	683	112
1972	815	686	129
1973	822	715	106
1974	734	650	84
1975	688	627	61

Sources: Maryland Department of Transportation, Division of Traffic.

Unlike the picture presented by the statistics on absolute number of deaths, motor vehicle death rates have been declining in a consistent pattern. The use of the death rate, death per 100,000,000 motor vehicle miles travelled, avoids the problem of the impact of changes in the amount of travel resulting from periodic shortages and sharply rising prices of gasoline. Death rates for the U.S. and Maryland for a five year period preceding the change in speed limits and the two following years are shown in Table 2. The decline in the death rate for the U.S. from 1973 to 1974 of .66 (deaths per 100,000,000 MVM), a decline of approximately 16%, is three times the average annual decline in the preceding four years.

The National Highway Traffic Safety Administration in the U.S. Department of Transportation indicates that lower speed limits are the largest single factor in the reduction of deaths and account for half of the decline in number of deaths, about 8%. The Statistical Division of the National Safety Council has reported on the sharp decline in the number of deaths for the first four months of 1974 in the U.S. The number of deaths declined by 24%, of which 11% is attributed to the reduction in speed. The

TABLE 2

U.S. AND MARYLAND MOTOR VEHICLE DEATH RATES 1969 - 1975
(Deaths per 100,000,000 Motor Vehicle Miles)

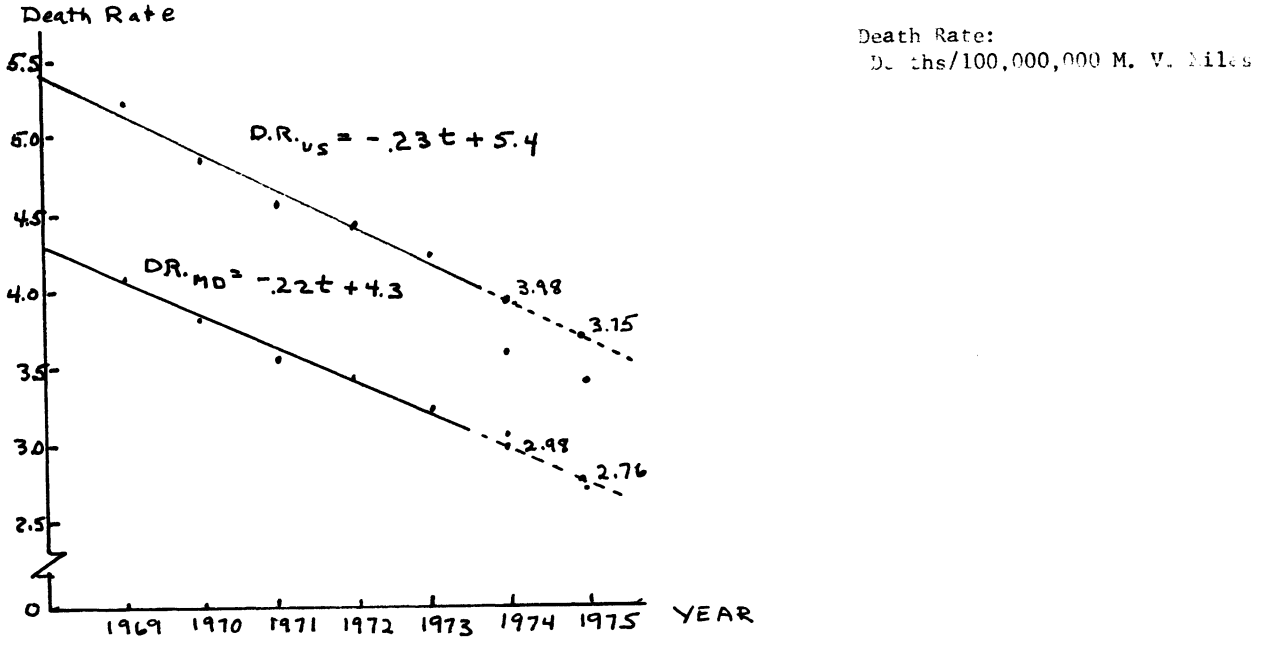
Year	U.S. Death Rate	Maryland Death Rate
1969	5.21	4.10
1970	4.88	3.84
1971	4.58	3.58
1972	4.44	3.45
1973	4.26	3.22
1974	3.60	3.08
1975	3.45	2.73

Sources: U.S. Department of Transportation, Federal Highway Administration, Highway Statistics

National Safety Council, Accident Facts, Annual Issues

FIGURE 1

TRENDS IN MOTOR VEHICLE DEATH RATES IN THE U.S. AND MARYLAND FOR THE FIVE YEAR PERIOD, 1969-1973



method used in the NSC analysis is to partial out each factor that affects motor vehicle fatalities, such as speed reduction, reduction in travel, change of road used and increased use of safety belts. To measure the role of reduced speed, nationally, sample information from previous state studies was used. This information is combined with other accident data to establish the roles of various factors in reducing the death rate.

The pattern of declining death rates for the U.S. and Maryland prior to the imposition of the 55 MPH speed limit is shown in Figure 1. Linear least squares trend lines were fitted to death rates in the U.S. and Maryland for the five year period 1969 to 1973 preceding the 55 MPH speed limit. The correlation coefficients for the U.S. and Maryland fitted lines were .98 and .97 respectively. The general downward trend has been attributed to the continuing improvements in the roadway, safer automobiles and improved driving habits. The 20% lower death rate for Maryland as compared to the U.S. reflects the higher than average degree of urbanization which is associated with lower death rates. It is, however, remarkable that both trend lines have the same approximate slope. However, since the institution of the 55 MPH speed limit, the U.S. death rate has dropped below trend, whereas Maryland's death rate was slightly above trend for 1974 and slightly below in 1975.

Based on trend, the U.S. expected death rate was 3.98 as compared to actual rate of 3.60 in 1974. If we attribute this difference to the lowered speed limits, then this 10% decline below the expected compares favorably with the 11% decline determined by the National Safety Council and U.S. Department of Transportation estimated decline in the death rate of 9%. Using the trend line as above, we estimate the 1975 effect of the lowered speed limits for the U.S. to be 8%.

Examining the trend line for Maryland, we find the actual death rate for 1974 slightly above trend, implying no effect or slight negative effect of the speed limit on death rates. For 1975, however, the actual death rate is about 3% below the expected death rate. Overall it appears that the reduced speed limit has had no appreciable effect on the death rate in Maryland.

A feasible hypothesis is that the differences in decline in death rates among the states is related to the proportion of miles travelled on high speed roads (subject to the reduced speed limit) to total miles travelled. In the West, where a relatively smaller proportion of each state is urbanized, there are generally higher proportions of high speed miles travelled. The reduction in the speed limit will have a greater effect on death rates in those states as compared to states like Maryland that are highly urbanized.

Data on the proportion of miles travelled on high speed roads by state was not available for testing the above hypothesis. However, population density was considered as a proxy, i.e. population density was assumedto be inversely related to the proportion of miles travelled on high speed roads to total miles travelled. Table 3 shows the results of two tests. The first test sorted the 48 contiguous states into the 24 most dense states and the 24 least dense states. The average decline in the death rate between the two year period prior to (1972 and 1973) and the two year period after (1974 and 1975) the imposition of the 55 MPH speed limit was computed for each state and the average decline for the twenty four least dense states was compared to the average decline for the twenty four most dense states. The 24 least dense states show a one-third greater decline than the 24 most dense states.

If we apply the same test to the extremes of the

Table 3

AVERAGE DECLINE IN DEATH RATE BETWEEN 1972-73 AND 1974-75
(States Grouped by Population Density)

	Decline in Death Rate
24 most dense states	.73
24 least dense states	1.05
8 most dense states	.42
8 least dense states	1.37

Sources: National Safety Council, Accident Facts, Annual Issues

U.S. Bureau of the Census, Statistical Abstract of the United States, 1976

distribution, taking the eight least and eight most dense states, the contrast is even more striking with the least dense states showing a three times greater decline. In addition, the Spearman rank correlation coefficient (r_s=.53, p=.001) computed from the above state data strongly supports the relationship between the change in death rates and population density.

These analyses are consistent with the assertion that the reduced speed limits have differential effect with respect to the density of the state. Maryland, a highly dense state with relatively small proportion of miles travelled on high speed roads, shows the expected lower impact of the reduced speed limit.

An Application of Logistic Regression to Occupational Health Data

James E. Higgins and Gary G. Koch, University of North Carolina at Chapel Hill

1. Introduction

The Occupational Safety and Health Act of 1970 focused national attention on the problems of occupationally related diseases and helped to energize a flurry of activity investigating the location, cause, and prevention of these diseases. Because of the relatively recent widespread interest in studying occupational diseases and the expense associated with such studies, most of the epidemiological investigations have relied on essentially cross-sectional surveys of disease prevalence.

In the cotton textile industry, investigators have concentrated on byssinosis or "brown lung" disease. The existence of a dose-response relationship between extended exposure to respirable cotton dust and the chest tightness syndrome of byssinosis has been well documented in studies by Marchant, et al. (1973) and Martin and Higgins (1976), among others. Of further interest are the effects on byssinosis prevalence of other variables such as length of exposure, smoking habits, sex, and race. Recently, Higgins and Koch (1977) offered a variable selection scheme to reduce the number of independent variables before applying weighted least squares methodology to analyze byssinosis prevalence in a large data set.

This paper applies logistic regression for the analysis and operates with the complete set of independent variables. The method employs the simultaneous implementation of maximum likelihood and weighted least squares estimation procedures in a way which emphasizes their respective strengths.

2. Data

The data for analysis were drawn from a 1973 survey of pulmonary function among employees of a large cotton textile company (Martin and Higgins, 1976). Byssinosis was classified at two levels, complaint of byssinosis symptoms and no complaint, and the responses were observed among seventy-two sub-populations of employees defined by:

- Dustiness of work area (W): workplace 1 (most dusty), workplace 2 (less dusty), workplace 3 (least dusty);
- Smoking habit (Sm): smoker or non-smoker at the time of the survey;
- Length of employment (E): <10 years (1), 10-20 years (2), and ≥20 years (3);
- Sex (Sx): male or female;
- Race (R): white or other races.

Since each of the 5419 employees under study spent their entire period of employment in only one of the three workplace classifications, this categorical variable was considered to be a reasonable measure of their relative degree of dust exposure.

3. Analysis

The observed data are given in Table 1. There are seven sub-populations in which no employees were observed, and twenty-seven of the remaining sixty-five sub-populations had no complaints of byssinosis.

The Functional Asymptotic Regression Methodology (FARM) given by Koch, Imrey, Freeman, and Tolley (1977) can be used to model the sixty-five sub-populations. FARM is a class of two-stage procedures for categorical data analysis which obtains efficient parameter estimates and consistent covariance estimates from some underlying first stage model and employs weighted least squares (WLS) methods to examine these at a second stage.

For a first stage model, assume that π_{i1}, the probability that an individual in the i-th sub-population has a complaint of byssinosis, can be adequately represented by the logistic function

$$\pi_{i1} = (1+\exp(-x_i'\beta))^{-1} = \exp(x_i'\beta)/(1+\exp(x_i'\beta)) ,$$

where $i = 1, 2, \ldots, 65$, x_i is a $1\times t$ "design vector" and β is a $t\times 1$ vector of model parameters. Since it is assumed that $\pi_{i1} + \pi_{i2} = 1$, where π_{i2} is the probability that an individual in the i-th sub-population does not have a complaint of byssinosis, we have that $\pi_{i2} = (1+\exp(x_i'\beta))^{-1}$ and

$$\log_e(\pi_{i1}/\pi_{i2}) = x_i'\beta \quad . \qquad (1)$$

At the first stage, assuming that the sub-populations are independent and follow the binomial distribution, the log-likelihood for the observed table is

$$\sum_{i=1}^{65} \log_e(n_i!/n_{i1}!\, n_{i2}!) + \sum_{i=1}^{65} n_{i1}\, x_i'\beta - \sum_{i=1}^{65} n_i(1+\exp(x_i'\beta)) \quad , \qquad (2)$$

where n_{ij} represents the number of employees observed in the i-th sub-population with byssinosis complaint j and $n_i = n_{i1} + n_{i2}$. For a given set of design vectors x_i, the expression (2) can be maximized by successive approximation numerical methods like those given in Kaplan and Elston (1972) to calculate maximum likelihood estimators (MLE) $\hat{\beta}$ for β. These estimators, in turn, may be converted to a corresponding predicted frequency vector and analyzed by an extension of the WLS approach of Grizzle, Starmer, and Koch (1969), which provides a consistent estimator (based on the inverse of the Fisher information matrix) of the covariance matrix $V_{\hat{\beta}}$. Computer software for the WLS analysis is provided by the program GENCAT (Landis, Stanish, Freeman, and Koch, 1976).

Alternatively, if the design vectors x_i used in the direct maximization are of an appropriate form, the MLE can be generated by Iterative Proportional Fitting (IPF) of hierarchical models to marginal tables which are sufficient for model parameters. (IPF is discussed in detail in Bishop, Fienberg, and Holland (1975), as well as elsewhere, and computer software is available through the program ECTA (1974).) In particular, the $65\times t$ design matrix X provided by $X' = (x_1, x_2, \ldots, x_{65})$ must be such that when transformed by an appropriate linear transformation, it is hierarchical with respect to the set of byssinosis complaint responses together with the independent variables

which define the sub-populations (for details see Koch, et al., 1977).

Thus, regardless of whether β is estimated by direct maximization of expression (2) or IPF, logit functions of the form of expression (1) of the predicted byssinosis proportions $\hat{\pi}_{ij}$, instead of the observed proportions p_{ij}, can be operated on by WLS computational algorithms and consistent estimators for the covariance matrices of $\hat{\beta}$ and the π_{i1} can be determined for use in subsequent FARM analyses.

Based on prior experience with the data, a first stage analysis is formulated in terms of a six module main effect model. The six modules are formed by the six combinations of workplace and smoking levels. Within each module, main effect designs including a module mean, two employment effects, and single sex and race effects are constructed so that the overall design X_1 contains 30 parameters. The MLE predicted frequencies given in Table 1 were actually obtained by direct maximization of the log-likelihood expression (2). However, the predicted frequencies could be obtained by IPF, with a slight modification to the standard ECTA program, by fitting the Employment vs. Sex vs. Race, Employment vs. Byssinosis, Sex vs. Byssinosis, and Race vs. Byssinosis marginal configurations for each module and using zero starting values for null sub-populations. The logits of the predicted frequencies are then analyzed in a second stage using WLS and FARM chi-square test statistics.

The design is reduced to a three module main effect model, with the modules formed by the three workplace levels. Each of the modules is a main effect design including a module mean, two employment effects, and single smoking, sex, and race effects so that the overall design X_2 employs 18 parameters. Using FARM test statistics, design X_2 can be further reduced to an 8 parameter complete main effect design X_3 with an overall mean, two effects each for workplace and employment, and single effects for smoking, sex, and race.

Alternatively, new MLE predicted frequencies can be generated based on X_2 or X_3 and logits formed from them can be analyzed using FARM test statistics. Table 2 displays three sets of parameter estimates and FARM test statistics for design X_3 that result from using MLE predicted frequencies from the 30 parameter design X_1, the 18 parameter design X_2, and the 8 parameter design X_3. The X_3 parameter estimates resulting from using X_1 and X_2 MLE predicted frequencies are obtained from consistent estimators based on FARM methodology, while the estimates resulting from using X_3 MLE predicted frequencies are MLE. Test statistics for X_3 based on each of the three sets of MLE predicted frequencies indicate that sex and race can be dropped from the model and that workplace and employment can each be adequately represented by single effects. Further, parameter estimates for smoking and employment are roughly equal so that a single parameter can be formed to represent a smoking-employment effect.

The final model indicated, X_4, is a 3 parameter main effect design with an overall mean, a workplace effect, and a combined smoking-employment effect. The parametrization and parameter estimates for X_4 are displayed in Table 3 for three sets of MLE-predicted frequencies and predicted byssinosis prevalences are given in columns 11 and 12 of Table 4 for X_4 reduced from the 8 parameter MLE design X_3.

An alternative 3 parameter design X_5 is given in Table 4 with corresponding parameter estimates. Model X_5, a refinement of the module designs X_1 and X_2, estimates workplace 2 and 3 byssinosis logits by an overall mean while workplace 1 logits are estimated with the addition of a combined smoking-employment effect. Predicted byssinosis prevalences for X_5 reduced from the X_2 MLE predicted frequencies are given in columns 9 and 10 of Table 4 along with predicted prevalences for design X_5 using observed prevalences (without log transform) of byssinosis complaints from an 8 sub-population table formed by collapsing the original 72 sub-population table into 2 workplace levels, 2 employment levels, and 2 smoking levels. A more complete documentation of the analysis stages is given in Higgins and Koch (1977a).

4. Discussion

The two-stage approach taken here to log-linear model analysis represents one method of dealing with a large, complete contingency table that is complicated by numerous cell frequencies that are small or zero. A previous approach by Higgins and Koch (1977) avoided this complication by eliminating some independent variables through variable selection and further increasing cell frequencies by collapsing to permit WLS analysis on the linear prevalence scale since cell frequencies were of adequate size (i.e., ≥ 5). On the other hand, the numerous small cell sizes in the complete table may invalidate the inferential procedures of the WLS methodology since they depend on the multinormality of the observed cell proportions. In this regard, if hierarchical log-linear models are considered appropriate, as is the case here, IPF may be preferable inasmuch as the asymptotic theory for the MLE depends on the multinormality of selected marginal configurations.

The initial 30 parameter model X_1 has problems with small frequency counts in some of the marginal tables required for generating MLE predicted frequencies. Consequently, the statistical validity of all the results based on design X_1 MLE predicted frequencies may not be ensured but the results are of interest as a procedure for identifying "unimportant" sources of variation for elimination from the model. However, the two sets of MLE predicted frequencies which are obtained on the basis of X_2 and X_3 can reasonably be presumed as appropriate (in terms of marginal cell frequencies) for ensuring statistical validity. Nonetheless, parameter estimates and corresponding standard errors based on all three sets of MLE predicted frequencies are quite similar at the various stages of model reduction (see Tables 2 and 3).

Finally, the relative merits of designs X_4 and X_5 need to be considered before choosing one as a final model. Since no interaction is detected among the variables workplace, employment and smoking, model X_4 supports the choice of a log-linear model based on dose-response considerations, if one is willing to make certain assumptions about the nature of the data taken from this cross-sectional survey (as they pertain to the longitudinal etiology of the disease, for which further discus-

sion is given in Higgins and Koch, 1977a). With model X_4, the conceptual "dose" is an additive function of the pertinent main effect parameters for the respective sub-populations, and the parameters can be interpreted as measures of relative risk that are associated with the specific effects of one of the occupational disease environment variables after controlling for the others. On the other hand, design X_5 can be interpreted as considering the combined smoking-employment effect at workplaces 2 and 3 to be medically insignificant, although statistically significant, when compared to the effect at workplace 1. Thus, model X_5 indicates that the effects of smoking and length of employment need only be considered important for employees at workplace 1.

ACKNOWLEDGMENTS

This research was in part supported through a Joint Statistical Agreement with Burroughs Wellcome Company. The authors would like to thank Jean Harrison and Jean McKinney for their conscientious typing of this manuscript.

REFERENCES

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W. Discrete Multivariate Analysis (M.I.T. Press, 1975).

ECTA, University of Chicago (1974).

Grizzle, J.E., Starmer, C.F., Koch, G.G. Biometrics 25 (1969)489-504.

Higgins, J.E., Koch, G.G. Internat. Statistical Review 45 (1977) 51-62.

Higgins, J.E., Koch, G.G., in preparation (1977a).

Kaplan, E.B., Elston, R.C., U.N.C. Mimeo Series No. 823 (1973).

Koch, G.G., Imrey P.B., Freeman, J.L., Tolley, H.D., Proceedings of the 9-th I.B.C. (1976).

Landis, J.R., Stanish, W.M., Freeman, J.L., Koch, G.G., Comp. Prog. in Biomed. 6 (1976) 196-231.

Martin, E.F., Higgins, J.E., J. Occup. Med. 8 (1976) 455-462.

Merchant, J.A., et al., J. Occup. Med. 15 (1973) 212-221.

TABLE 1. CONTINGENCY TABLES BASED ON OBSERVED AND LOG-LINEAR MODEL PREDICTED (MLE FOR DESIGN $\underset{\sim}{X}_1$) FREQUENCIES

				Observed Frequencies				MLE Log-Linear Model Predicted Frequencies for Design $\underset{\sim}{X}_1$			
				Smokers byssinosis		Non-Smokers byssinosis		Smokers byssinosis		Non-Smokers byssinosis	
W	E	Sx	R	Yes	No	Yes	No	Yes	No	Yes	No
1	1	M	W	3	37	0	16	5.41680	34.58320	0.63328	15.36672
1	1	M	OR	25	139	6	75	22.84520	141.15480	5.46345	75.53655
1	1	F	W	0	5	0	4	0.29220	4.70780	0.07644	3.92356
1	1	F	OR	2	22	1	24	1.44648	22.55352	0.82675	24.17325
1	2	M	W	8	21	2	8	6.82747	22.17253	1.14710	8.85290
1	2	M	OR	8	30	1	9	9.17244	28.82756	1.85290	8.14710
1	2	F	W	0	0	0	0	0.00000	0.00000	0.00000	0.00000
1	2	F	OR	0	0	0	0	0.00000	0.00000	0.00000	0.00000
1	3	M	W	31	77	5	47	29.33496	78.66504	5.04660	46.95340
1	3	M	OR	10	31	3	15	11.40374	29.59626	2.85678	15.14322
1	3	F	W	0	1	0	2	0.12876	0.87124	0.09674	1.90326
1	3	F	OR	0	1	0	0	0.13248	0.86752	0.00000	0.00000
2	1	M	W	0	74	0	35	0.18870	73.81130	0.41300	34.58700
2	1	M	OR	0	88	1	47	0.22088	87.77912	0.84335	47.15664
2	1	F	W	1	93	1	54	1.01896	92.98104	0.76065	54.23935
2	1	F	OR	2	145	3	142	1.57143	145.42857	2.98265	142.01735
2	2	M	W	1	50	1	16	0.47736	50.52264	0.28781	16.71219
2	2	M	OR	0	5	0	0	0.04615	4.95385	0.00000	0.00000
2	2	F	W	1	33	0	30	1.32294	32.67706	0.59460	29.40540
2	2	F	OR	0	4	0	4	0.15364	3.84636	0.11760	3.88240
2	3	M	W	1	141	0	39	1.05932	140.94068	0.43875	38.56125
2	3	M	OR	0	1	0	1	0.00736	0.99261	0.01676	0.98324
2	3	F	W	3	91	3	187	2.93374	91.06626	2.50610	187.49390
2	3	F	OR	0	0	0	2	0.00000	0.00000	0.03924	1.96076
3	1	M	W	2	258	0	134	3.19800	256.80200	0.89646	133.10345
3	1	M	OR	3	242	1	122	2.20255	242.79745	1.11192	121.88808
3	1	F	W	3	180	2	169	2.73036	180.26964	1.46547	169.53453
3	1	F	OR	3	260	4	301	2.86933	260.13067	3.52885	301.47115
3	2	M	W	1	187	0	58	1.70328	186.29672	0.30392	57.69608
3	2	M	OR	0	33	0	7	0.21813	32.78187	0.04956	6.95044
3	2	F	W	2	94	1	90	1.05504	94.94496	0.61061	90.38939
3	2	F	OR	0	3	0	4	0.02409	2.97591	0.03628	3.96372
3	3	M	W	12	495	3	182	10.02846	496.97154	1.40415	183.59585
3	3	M	OR	0	45	0	23	0.65160	44.34840	0.23575	22.76425
3	3	F	W	3	176	2	340	4.28705	174.71295	3.32082	338.67918
3	3	F	OR	0	2	0	3	0.03510	1.96490	0.03933	2.96067

TABLE 2

PARAMETER ESTIMATES AND CHI-SQUARE TEST STATISTICS (Q) FOR MAIN EFFECT DESIGN X_{3} WITH THREE SETS OF PREDICTED (MLE) FREQUENCIES

1. Parameterization of X_{3} 65x8

Source of Variation	Estimated Incremental Parameter	Indicator Variable
mean	b_1	x_1 = 1 always
main effect: Workplace	b_2, b_3	x_2 = {1 high dust; -1 moderate dust; 0 low dust}, x_3 = {1 high dust; 0 moderate dust; -1 low dust}
main effect: Smoking	b_4	x_4 = {1 smoker; -1 non-smoker}
main effect: Employment (years)	b_5, b_6	x_5 = {1 <10; -1 10 to 20; 0 ≥20}, x_6 = {1 <10; 0 10 to 20; -1 ≥20}
main effect: Sex	b_7	x_7 = {1 male; -1 female}
main effect: Race	b_8	x_8 = {1 white; -1 other races}

2. Parameter estimates and corresponding standard errors

Frequencies Predicted by Design	b_1	Workplace b_2	Workplace b_3	Smoking b_4	Employment b_5	Employment b_6	Sex b_7	Race b_8
X_{1}	-3.362	0.689	1.013	0.304	-0.129	-0.286	-0.065	-0.062
(30 parameter)	(0.120)	(0.194)	(0.139)	(0.099)	(0.157)	(0.130)	(0.127)	(0.103)
X_{2}	-3.395	0.756	0.998	0.312	-0.141	-0.292	-0.093	-0.055
(18 parameter)	(0.120)	(0.193)	(0.139)	(0.099)	(0.157)	(0.130)	(0.128)	(0.102)
X_{3}	-3.477	0.810	0.960	0.321	-0.125	-0.314	-0.062	-0.058
(8 parameter)	(0.124)	(0.179)	(0.138)	(0.097)	(0.154)	(0.128)	(0.114)	(0.104)

Q_R(22 D.F.) = 17.36 for X_{1} reduced to X_{3}; Q_R(10 D.F.) = 12.41 for X_{2} reduced to X_{3}

Q_R: WLS chi-square reduction goodness of fit statistic

3. Chi-square statistics (Q) for design X_{3} effects

		Q for frequencies predicted by design		
		X_{1}	X_{2}	X_{3}
Effect	D.F.	(30 parameter)	(18 parameter)	(8 parameter)
Workplace	2	143.44**	150.25**	177.99**
Smoking	1	9.40**	9.92**	10.88**
Employment	2	11.13**	12.03**	12.47**
Sex	1	0.26	0.53	0.29
Race	1	0.36	0.29	0.32

** significant at the 0.01 level

TABLE 3

PARAMETER ESTIMATES FOR MAIN EFFECT DESIGN $\underset{\sim}{X}_4$ AND REDUCED MODULE DESIGN $\underset{\sim}{X}_5$ WITH THREE SETS OF PREDICTED MLE FREQUENCIES

1. Parameterization of $\underset{65\times 3}{\underset{\sim}{X}_4}$ and $\underset{65\times 3}{\underset{\sim}{X}_5}$

	Design $\underset{\sim}{X}_4$		Design $\underset{\sim}{X}_5$	
Source of Variation	Estimated Incremental Parameter	Indicator Variable	Estimated Incremental Parameter	Indicator Variable
mean	b_1	x_1 = 1 always	b_1'	x_1 = 1 always
Workplace effect	b_2	$x_2 = \begin{cases} 1 & \text{high dust} \\ 0 & \text{moderate \& low dust} \end{cases}$	b_2'	$x_2 = \begin{cases} 1 & \text{high dust} \\ 0 & \text{moderate \& low dust} \end{cases}$
Smoking-Employment effect	b_3	$x_3 = \begin{cases} 2 & \text{smoker employed} \geq 10 \text{ years} \\ 1 & \text{smoker employed} < 10 \text{ years or non-smoker employed} \geq 10 \text{ years} \\ 0 & \text{non-smoker employed} < 10 \text{ years} \end{cases}$	b_3'	$x_3 = \begin{cases} 2 & \text{smoker employed} \geq 10 \text{ years at workplace 1} \\ 1 & \text{smoker employed} < 10 \text{ years or non-smoker employed} \geq 10 \text{ years at workplace 1} \\ 0 & \text{other employees} \end{cases}$

2. Parameter estimates and corresponding standard errors for $\underset{\sim}{X}_4$ and $\underset{\sim}{X}_5$

	Design $\underset{\sim}{X}_4$			Design $\underset{\sim}{X}_5$		
Frequencies Predicted by Design		Workplace	Smoking-Employment		Workplace	Smoking-Employment
	b_1	b_2	b_3	b_1'	b_2'	b_3'
$\underset{\sim}{X}_1$	-4.939	2.586	0.572	-4.253	1.474	0.873
(30 parameter)	(0.196)	(0.172)	(0.122)	(0.130)	(0.298)	(0.173)
$\underset{\sim}{X}_2$	-5.001	2.614	0.593	-4.290	1.499	0.878
(18 parameter)	(0.201)	(0.172)	(0.128)	(0.130)	(0.300)	(0.174)
$\underset{\sim}{X}_3$	-5.109	2.662	0.619	--	--	--
(8 parameter)	(0.214)	(0.169)	(0.126)	--	--	--

Q_R^*(27 D.F.) = 19.80 for $\underset{\sim}{X}_1$ reduced to $\underset{\sim}{X}_4$

Q_R(15 D.F.) = 14.54 for $\underset{\sim}{X}_2$ reduced to $\underset{\sim}{X}_4$

Q_R(5.D.F.) = 1.44 for $\underset{\sim}{X}_3$ reduced to $\underset{\sim}{X}_4$

Q_R(27 D.F.) = 16.02 for $\underset{\sim}{X}_1$ reduced to $\underset{\sim}{X}_5$

Q_R(15 D.F.) = 10.43 for $\underset{\sim}{X}_2$ reduced to $\underset{\sim}{X}_5$

* Q_R: WLS chi-square reduction goodness of fit statistic

TABLE 4

OBSERVED, LINEAR, AND LOG-LINEAR MODEL PREDICTED BYSSINOSIS PREVALENCES WITH CORRESPONDING STANDARD ERRORS

						Predicted Byssinosis Prevalences					
						Reduced Module Design X_5				Main Effect Design X_4	
				Observed Byssinosis Prevalence (Estimated s.e. x 10^3)		WLS Linear Collapsed Observed Table (Estimated s.e. x 10^3)		FARM Log-Linear Reduced From X_2 MLE (Estimated s.e. x 10^3)		FARM Log-Linear Reduced From X_3 MLE (Estimated s.e. x 10^3)	
W	E	Sx	R	Smokers	Non-Smokers	Smokers	Non-Smokers	Smokers	Non-Smokers	Smokers	Non-Smokers
1	1	M	W	0.075(42)	0.000	0.143(13)	0.045(19)	0.129(15)	0.058(15)	0.139(14)	0.080(15)
1	1	M	OR	0.152(28)	0.074(29)	0.143(13)	0.045(19)	0.129(15)	0.058(15)	0.139(14)	0.080(15)
1	1	F	W	0.000	0.000	0.143(13)	0.045(19)	0.129(15)	0.058(15)	0.139(14)	0.080(15)
1	1	F	OR	0.083(56)	0.040(39)	0.143(13)	0.045(19)	0.129(15)	0.058(15)	0.139(14)	0.080(15)
1	2	M	W	0.276(83)	0.200(126)	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	2	M	OR	0.211(66)	0.100(95)	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	2	F	W	*	*	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	2	F	OR	*	*	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	3	M	W	0.287(44)	0.096(41)	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	3	M	OR	0.244(67)	0.167(88)	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	3	F	W	0.000	0.000	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
1	3	F	OR	0.000	*	0.240(24)	0.143(13)	0.262(29)	0.129(15)	0.230(24)	0.139(14)
2	1	M	W	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
2	1	M	OR	0.000	0.021(21)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
2	1	F	W	0.011(11)	0.018(18)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
2	1	F	OR	0.014(09)	0.021(12)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
2	2	M	W	0.020(19)	0.059(57)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	2	M	OR	0.000	*	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	2	F	W	0.029(29)	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	2	F	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	3	M	W	0.007(07)	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	3	M	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	3	F	W	0.032(18)	0.016(09)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
2	3	F	OR	*	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	1	M	W	0.008(05)	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
3	1	M	OR	0.012(07)	0.008(08)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
3	1	F	W	0.016(09)	0.012(08)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
3	1	F	OR	0.011(06)	0.013(07)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.011(02)	0.006(01)
3	2	M	W	0.005(05)	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	2	M	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	2	F	W	0.021(15)	0.011(11)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	3	F	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	3	M	W	0.024(07)	0.016(09)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	3	M	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	3	F	W	0.017(10)	0.006(04)	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)
3	3	F	OR	0.000	0.000	0.012(02)	0.012(02)	0.014(02)	0.014(02)	0.020(03)	0.011(02)

* No employees were observed in this sub-population

HOW COST EFFECTIVE IS PSRO AT A LARGE UNIVERSITY HOSPITAL?

Emma L. Frazier, Mead Johnson & Company
M. C. Miller, and M. C. Westphal, Medical University of South Carolina

There has been an increasing awareness of the rising costs of health care in the United States. In 1950, 4.6 percent of the Gross National Product (GNP) was being spent for health care; by 1970, health care expenditures rose to 8.6 percent of the GNP or $139 billion (Rice, 1977).

In an attempt to control the costs of hospitalization and improve the quality of care rendered to patients covered by Medicare and Medicaid, the Federal Government has imposed requirements for peer review on hospitals caring for these individuals. Professional Standards Review Organization (PSRO) was created as a part of the 1972 Social Security Amendments to determine whether:

a) . . . services . . . are or were medically necessary;

b) the quality of services meets professionally recognized standards of health care;

c) . . . services . . . (could be) provided on an outpatient basis or more economically in an inpatient health care facility of a different type (Public Law 92-603).

The PSRO system is composed of several interrelated activities. Two of these mechanisms are the concurrent admission and continued stay reviews. The admission and continued stay reviews examine the patient's admission to and continued stay in the medical facility to determine if the admission and extended length of stay are medically necessary.

In November of 1975, a "fully delegated" PSRO became operational at the Medical University Hospital of South Carolina in Charleston, South Carolina. Two of the fundamental review mechanisms, the admission and continued stay reviews, are performed under the auspices of the Utilization Review Committee at the Medical University Hospital of South Carolina. The Utilization Review Committee (URC) is composed of sixteen physicians and four non-medical personnel at the Medical University Hospital. This Committee is serviced by four trained and experienced non-physician review coordinators who perform the initial and periodical review of medical records.

Within the first working day after the patient's admission to the hospital, an admission review must be performed to determine the medical necessity of the admission. If the admission is deemed medically necessary, an initial length of stay based on diagnosis-specified criteria established by Southern region norms is assigned. Before the end of this assigned length of stay, the need for an extended length of stay must be approved. The patient's medical record is periodically reviewed until the patient is discharged to certify the need for an extended length of stay based on PSRO disgnosis-specific norms and criteria.

If the coordinator questions the necessity of an admission or an extended length of stay, a physician of the Utilization Review Committee is consulted. Whenever the Utilization Review Committee or Subcommittee (composed of four physicians) finds an admission or extended length of stay to be medically necessary, the fiscal intermediary attending physician and the hospital administrator are notified in writing. If the inpatient admission or extended length of stay is denied by the physician advisors of the Utilization Review Committee, the hospital is reimbursed only for approved inpatient days that the patient has stayed in the hospital.

Some studies have reported that PSRO is cost effective in reducing the lengths of stay, admissions and/or costs of hospitalization (Brain, 1973; Flashner, 1973). Flashner et al (1973) reported a reduction of approximately $9 million in hospital reimbursement succeeding the initiation of PSRO review procedures. Most of these studies, basically performed in the private hospital sector, were criticized for failure to account for causal factors and for weaknesses in methodology (Davidson et al, 1973).

Unlike most private hospitals where studies on the effectiveness of PSRO have been performed, the Medical University Hospital (MUH) is a teaching hospital serviced by over three hundred staff physicians. The hospital generally serves as a referral center for the tri-county area which includes Charleston, Dorchester, and Berkeley counties. The medical staff at MUH believes that optimal care is already being rendered to the population that they serve in the shortest time possible and that PSRO is not effective in reducing hospital reimbursement by third party payers. Still, the question remains to be answered: Is PSRO effective in reducing the costs of medical care services rendered to Medicare and Medicaid patients?

METHODS OF PROCEDURE

Cost

Costs included: the hourly salaries of the review coordinators multiplied by hours spent by each coordinator on PSRO review and review related duties; hourly salaries of members of the Utilization Review Committee calculated on the basis of a 40-hour week and multipled by the estimated number of hours spent in Utilization Review Committee meetings and other review activities; fringe benefits (15.95 percent

of salaries) and overhead cost (53 percent of salaries). All salary costs for coordinators and members of the Utilization Review Committee were borne by the Medical University Hosptial.

Effectiveness

Effectiveness was determined by direct and indirect measures. Direct measures of effectiveness were the number of admissions, extended lengths of stays, and hospital services denied by the Utilization Review Committee. The changes in pertinent hospital utilization variables over several time periods served as indirect measurements of the effectiveness of PSRO reveiw mechanisms.

Indirect evidence of the effectiveness of PSRO was sought by comparing the average lengths of stay (ALOS), average cost per patient (ACOS) and average number of admissions (ANAD) over several time periods. For each of the dependent variables, ACOS, ALOS and ANAD, a three-way analysis of variance was performed analyzing these factors: a) type of patient (Medicare-Medicaid (reviewed) vs nonMedicare-Medicaid (not reviewd)). b) month of year (January-July, i.e., the first fully operational seven-month period following the advent of PSRO). c) advent of PSRO (1975 (before PSRO), 1976 (after PSRO) and 1977 (1 year after the advent of PSRO)). The basic design therefore, followed a 2x7x3 factorial experiment.

Succeeding the analyses described above, an analysis of covariance was performed for the ACOS and ALOS adjusting for the number of Medicare and Medicaid admissions to the Medical University Hospital. It has been suggested by Flashner et al (1973) that if too many Medicaid patients are admitted to the hospital, the hospital population will contain more individuals with mild illness.

RESULTS

Table 1 show the costs of the PSRO review procedure for January-July of 1976. The total cost was $28,938 which annualized to $49,608.

TABLE I.

COST OF PSRO REVIEW FOR JANUARY-JULY (1976)

SALARIES	HOURS	COSTS
Review Coordinators	3705	15,380
Physicians	98	1,037
Other Committee Members	100	711
		17,128
Fringe Benefits (15.95%)		2,732
Overhead (53%)		9,078
TOTAL COST		28,938
Total Cost Annualized		49,608

No direct evidence of PSRO effectiveness could be found since there were not any denials of admission, extended lengths of stays or services rendered during the seven-month study period.

The analysis of variance of the indirect measures of PSRO effectiveness (ACOS, ALOS and ANAD) showed no time-of-the-year effect and a significant interaction between the advent of PSRO and the proportion of Medicare vs non-Medicaid patients admitted to the hospital.

The data were collapsed over months and Duncan's New Multiple Range tests were performed for the ACOS, ALOS and ANAD for the groups: Before PSRO Not Reviewed (BPNR), After PSRO Not Reviewed (APNR), Before PSRO Reviewed (BPR), After PSRO Reviewed (APR), After one year of PSRO Not Reviewed (A1NR), and After one year of PSRO Reviewed (A1R). The ranked and underscored homogeneous means are shown in Table 2.

TABLE 2.

RESULTS OF DUNCAN"S MULTIPLE RANGE TESTS FOR ACOS, ALOS AND ANAD

ACOS:	BPNR	APNR	APR	BPR	A1NR	A1R
	1128	1461	1642	1756	1935	2098
ALOS:	APNR	BPNR	A1R	APR	A1NR	BPR
	7.4	7.4	7.5	7.6	7.6	9.1
ANAD:	BPR	APR	A1R	A1NR	APNR	BPNR
	365	498	547	1055	1068	1100

The analyses of covariance for the ACOS and ALOS adjusting for the number of Medicare and Medicaid admissions showed that: a) the variation in the number of Medicare-Medicaid admissions accounted for a significant portion of the variation observed in the ALOS and ACOS, b) the adjusted mean costs of stay were significantly different before and after the initation of PSRO, and c) no significant differences existed between the average length of stay for the study periods.

Even though the analysis showed significant differences between the average cost per patient over the study period, the question of how much of these differences were due to inflation remained to be answered. To examine this issue we assumed that the increases in hospital costs for non-Medicare-Medicaid patients were due to the inflation rate, while the changes in costs for Medicare-Medicaid patients resulted from both inflation and PSRO review. When the inflation rate of the not reviewed patients is used to correct the ASOS of the PSRO reviewed group, there is a significant cost savings of $261 per reviewed patient over the seven-month period.

Extrapolating to the 1976 Medicare-Medicaid patient population of 1976 patients we find a saving of 1.5 million dollars associated with a PSRO review process which cost $49,608. Obviously the PSRO review is cost-effective in this teaching hospital. The mechanism through which the ACOS was reduced is not clear since there was not a significant reduction in the ALOS for the reviewed population when corrected for the increased number of Medicare-Medicaid admissions.

REFERENCES

Brain, E. (1973), "Foundation for medical care control of hospital utilization: CHAP - a PSRO prototype", New England Journal of Medicine, 288, 878.

Davidson, S.M., Worker, R.C., and Klein, D.N. (1973), "Professional standards review organizations: A critique", JAMA 226, 1106.

Flashner, B.A., Reed, S., Coburn, R.W., and Fine, P.R. (1973), "Professional standards review organizations. Analysis of their development and implementation based on a preliminary review of the hospital Admission and Surveillance Program in Illinois", JAMA, 223, 1473

Rice, Dorothy B. (1977), "The role of statistics in the development of health care policy", The American Statistician 31, 101.

United States Statistics at Large. 92nd Congress, 2nd Session, 1972 (1973), Washington, D.C.: U.S. Government Printing Office, 86, 429.

CLUSTERING RANK PREFERENCE DATA

James Beckett III, Bowling Green State University

INTRODUCTION

Situations for which rank preference data are appropriate are numerous. Problems involving N judges ranking k objects are common; the analysis of said problems being handled in straightforward fashion via the well-known Friedman χ_r^2 test. Large values of χ_r^2 (or equivalently Kendall's coefficient of concordance $W = \frac{\chi_r^2}{N(k-1)}$) indicate that the group of judges is basically in agreement on some consensus rank ordering of the objects. If χ_r^2 is not significant, we state that we have not found enough evidence to indicate that the ranks were not assigned randomly, i.e., no apparent difference in objects. After a significant χ_r^2, multiple comparisons [Miller (1966)] should be performed to find out which objects are judged different.

If judges can be a priori grouped into subgroups according to one or more classification factors, a more complete analysis is obtained through the use of ANACONDA (Analysis of Concordance) [Beckett & Schucany (1975)]. The concept of ANACONDA is based on partitioning the total agreement into the agreement (or disagreement) between and within the subgroups. The agreement between two subgroups of judges is measured by the statistic $\mathcal{L}$, which can be expressed as the inner product of the two rank sum vector $\underline{S}$ and $\underline{T}$, i.e., $\mathcal{L} = \underline{S}'\underline{T} = \sum_{j=1}^{k} S_j T_j$, where the elements of $\underline{S}$ and $\underline{T}$ are $S_j = \sum_{i=1}^{m} R_{ij}$, $j = 1,2,\ldots,k$ and $T_j = \sum_{i=1}^{n} R'_{ij}$, $j = 1,2,\ldots, k$ where $R_{ij}(R'_{ij})$ represents the rank given the j^{th} object by the i^{th} judge in group one (two). The small sample distribution has been tabulated [Schucany & Frawley (1973)] while the asymptotic distribution of $\mathcal{L}$ is normal for large m, n, and k. The linear scaling of $\mathcal{L}$, $\mathcal{W} = \frac{12\mathcal{L} - 3mnk(k+1)^2}{mn(k^3-k)}$ is often useful as a generalized coefficient of concordance such that $-1 \leq \mathcal{W} \leq 1$. Also it has shown that $\mathcal{W} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} \rho_{ij}}{mn}$ where ρ_{ij} is Spearman's ρ between the i^{th} judge in group one and the j^{th} judge in group two, i.e., $\mathcal{W}$ is the average Spearman ρ between the two groups. Note that $\mathcal{W} = -1$ indicates disagreement between groups along with agreement within each group.

MOTIVATION

The underlying principle of cluster analysis is quite simple: identify subsets of individuals that tend to be relatively similar and group them together. There are two major steps common to the many methods used to cluster individuals: 1) computing quantitative indices of multivariate similarity between all pairs of individuals and 2) analyzing similarity matrices to identify homogeneous subgroups. Suppose N judges are ranking k objects and that we wish to cluster these judges on the basis of their preferences for the objects. We consider the N X N similarity matrix

$$\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & 1 & \cdots & \rho_{2N} \\ \vdots & \vdots & & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & 1 \end{pmatrix}$$

, where ρ_{ij} is the Spearman rank order correlation between the i^{th} and j^{th} judges. It is desirable to maximize the within cluster similarity and minimize the between cluster similarity. The minimization of $\mathcal{W}$ accomplishes both of these goals simultaneously.

The clustering procedures proposed herein should be considered as a logical third step in a comprehensive analysis of rank data following the Friedman χ^2 and multiple comparisons (if necessary). Clusters with interpretable or physical meaning might also indicate breakdowns of judges into subgroups such that an ANACONDA analysis might be illuminating for this data set or subsequent similar problems.

Suppose we have 6 judges ranking 3 objects A, B, and C in the following fashion.

	A	B	C
J1	1	2	3
J2	3	2	1
J3	1	2	3
J4	3	2	1
J5	1	2	3
J6	3	2	1

The obtained value of χ_r^2 is 0 indicating no agreement. However it is apparent the agreement of J2, J4, and J6 has been "cancelled" by the agreement (on the opposite ordering) of J1, J3, and J5. A conclusion of no agreement is clearly not appropriate if, for example, J1, J3, and J5 are women, while J2, J4, and J6 are men. In such a situation the subgroups should be considered separately; indeed, the value $\mathcal{W}$ for the male-female breakdown is -1 indicating agreement within each subgroup but on opposite orderings. We

examine the 6 x 6 similarity matrix as previously defined:

$$\begin{array}{c} \\ J1 \\ J2 \\ J3 \\ J4 \\ J5 \\ J6 \end{array} \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 & -1 & 1 \end{pmatrix}$$

Relabeling the judges 1,3,5,2,4,6 provides the clearer rearrangement of the similarity matrix below:

$$\begin{array}{c} \\ J1 \\ J3 \\ J5 \\ J2 \\ J4 \\ J6 \end{array} \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & -1 \\ \hline -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \\ -1 & -1 & -1 & 1 & 1 & 1 \end{array}\right)$$

The general idea herein is an extension of the above idea—a simple rearranging (relabeling) of the similarity matrix such that the elements in the upper right (lower left equivalently) corner of the matrix are small (ideally close to -1). The average of the elements in this block equals $\mathcal{W}$, the generalized coefficient of concordance between two groups of judges. $\mathcal{W}$ is (small) large if there is (dis)agreement between groups along with agreement within each group. Thus choosing members of clusters to minimize $\mathcal{W}$ will simultaneously maximize within cluster similarity and minimize betwen cluster similarity.

SPECIAL CONSIDERATIONS PARTICULAR TO RANK DATA

When one of the objects is clearly superior (or inferior) to the other k-1 objects, we must be wary of the high power of the Friedman test. Although it is proper that the Friedman test should reject, in considering the situation where we have k treatments of which one is a control (obviously inferior in e.g., agricultural or pharmaceutical studies) which has been added merely for reference, perhaps we should question our choice of objects if we are seeking a measure of agreement. The extreme high power of χ^2_r in this situation (one superior object) has been demonstrated by Beckett (1975). For example with a group of 6 judges ranking 5 objects with each judge recognizing the first object as clearly superior, the smallest value of χ^2_r attainable is 15 which itself is highly significant. Obviously clearly superior (or inferior) items are of no value in our clustering scheme and in fact their presence may mask some important inter-relationships between other objects and the potential subgroups. In such cases these "non-informative" objects should be ignored for purposes of clustering.

One could perform multiple comparisons (with small α) to separate or throw off objects clearly superior (or inferior) with significantly large or small rank totals. The remaining objects in the middle can be considered as the discriminating or "critical items". Regardless of the reduction (if any) of the objects to the critical subset, after the clusters are determined, for each cluster a cluster average rank profile should be presented based on all objects.

A usual problem in standard cluster analysis procedures is that larger problems quickly become too big for the computer. Here due to the data being in ranks, the data can be reduced to (k! + no. of unique tied rank orderings) since there are k! possible permutations of the ranks 1 to k. Disregarding ties, 1000 or more judges ranking 5 products can be reduced to at most 5! = 120 rank orderings each with a certain multiplicity.

PROCEDURES

With a divisive clustering algorithm we seek to divide the judges into two sub-groups or clusters. A feasible starting point might be to search for two pairs of judges who are as diametrically opposed as possible as measured by the smallest $\mathcal{W}$ (hopefully -1) obtained and use these pairs as the cluster nuclei to which judges will be added. Another approach which would be especially useful with a large number of judges would be to choose as the first cluster nucleus the observed consensus rank ordering of all N judges and to choose as the second cluster nucleus the conjugate rank ordering (opposite to the first cluster nucleus). Aside from being quicker, the latter approach would yield clusters representing the majority opinion (1st) and dissenting or minority opinion (2nd) as well as make the procedure less dependent on the order in which the data are read in.

After the two cluster nuclei are chosen, judges are added to clusters sequentially in such a way that $\mathcal{W}$ is minimized at each step. A stopping rule could be chosen (such as $\mathcal{W} \leq c$, for some chosen $c \leq 0$) or all judges could be forced into one of the two clusters. By stopping when $\mathcal{W}$ rises to some negative stopping value we would wind up with two clusters plus possibly some unclustered judges in the middle - these judges in the middle could be considered as making up a third cluster.

An agglomerative approach can be begun essentially by hand. The possible rank orderings can be grouped into classes. For example with k=4, Class 1 is chosen, say (1,2,3,4); then Class 2 contains those rank orders that can be obtained by one permutation of adjacent objects, i.e., {(2,1,3,4), (1,3,2,4), (1,2,4,3)}. Class 3 is obtained by two permutations of adjacent objects with reference to the Class 1 order or by trying one additional permutation referring to the rank orders in Class 2. For 4 objects we will have 7 classes; generally there are $\frac{k(k-1)}{2} + 1$ classes. The rank correlation between Class 1 and any one of rank orderings in Class 2 is .8 (generally $1 - \frac{12}{k(k^2-1)}$); the rank correlation between any two members within

class 2 is at least .40 $(1 - \frac{36}{n(n^2-1)})$. As long as we restrict our two clusters from having members from classes above and below the median class, W will remain below 0. This indicates that our cluster algorithm can be further streamlined by immediately adding to the cluster nuclei [Class 1 and Class $(\frac{k(k-1)}{2} + 1)$] those judges with rank preference orderings belonging to Class 2 and Class $(\frac{k(k-1)}{2})$, respectively.

EXAMPLES AND APPLICATIONS

Example 1 [Hollander & Wolfe, p. 140]. The data in Table 1 were obtained by Woodward (1970). Woodward, shortstop of the 1970 Cincinnati Reds National League baseball team, considered three methods of rounding first base. The best method is defined to be the one that, on the average, minimizes the time to reach second base.

TABLE 1. Rounding First Base Times

	Methods		
Players	Round Out	Narrow Angle	Wide Angle
1	5.40(1)	5.50(2)	5.55(3)
2	5.85(3)	5.70(1)	5.75(2)
3	5.20(1)	5.60(3)	5.50(2)
4	5.55(3)	5.50(2)	5.40(1)
5	5.90(3)	5.85(2)	5.70(1)
6	5.45(1)	5.55(2)	5.60(3)
7	5.40(2.5)	5.40(2.5)	5.35(1)
8	5.45(2)	5.50(3)	5.35(1)
9	5.25(3)	5.15(2)	5.00(1)
10	5.85(3)	5.80(2)	5.70(1)
11	5.25(3)	5.20(2)	5.10(1)
12	5.65(3)	5.55(2)	5.45(1)
13	5.60(3)	5.35(1)	5.45(2)
14	5.05(3)	5.00(2)	4.95(1)
15	5.50(2.5)	5.50(2.5)	5.40(1)
16	5.45(1)	5.55(3)	5.50(2)
17	5.55(2.5)	5.55(2.5)	5.35(1)
18	5.45(1)	5.50(2)	5.55(3)
19	5.50(3)	5.45(2)	5.25(1)
20	5.65(3)	5.60(2)	5.40(1)
21	5.70(3)	5.65(2)	5.55(1)
22	6.30(2.5)	6.30(2.5)	6.25(1)
	$R_1 = 53$	$R_2 = 47$	$R_3 = 32$

The value of χ_r^2 (adjusted for ties) is 11.1 which is significant at the .005 level. Hence we conclude the methods are not all the same with respect to speed. Multiple comparison of methods indicates Method 3 differs significantly from method 1 at the .01 experimentwise error rate. "Some" would continue and claim without statistical justification that method 3 is best. Regardless, the assumption of no block-treatment interaction (a fundamental assumption which is often overlooked) may be of greater concern here - Is one method best for all (types of) players? A quick perusal of the data shows players 1,6, and 18 performing opposite to the majority of the players. Perhaps method 1 really is best for these players due to some physical characteristics that they possess. Our cluster analysis provides the following result: Cluster 1: Players 2,4,5, 7-15,17,19-22; Cluster 2: Players 1,3,7,16,18 with $W = -.665$ highly significant and indicative of disagreement between the two groups. However, this disagreement has been manufactured and is meaningful only if the clusters are interpretable.

Example 2, [Gibbons, p. 353]. In a collaborative study of dry milk powders, six different types A to F are tested in each of seven different laboratories, and ranked in order of decreasing quality, that is, 1 = best, 6 = poorest. The results shown below are from Bliss (1967, p. 339).

TABLE 2.

	Rank for Powder					
Lab	A	B	C	D	E	F
1	2	3	6	1	5	4
2	2	1	3	4	5	6
3	1	2	3	5	4	6
4	2	3	1	5	6	4
5	4	1.5	1.5	6	3	5
6	1	3	4	5	2	6
7	2	4	1	5	6	3

Here W may be used as a check for an outlier (hospital 1). Employing hospital 1 as a singleton sub-group or cluster we obtain $W = -.1$ which is not significant. Had it been significant, an investigation of what makes hospital 1 significantly different from the others may have been profitable. However, not enough evidence is present to conclude all hospitals should not be considered as one group. In this situation W turns out to be a linear multiple of Page's L (1963).

SUMMARY AND COMMENTS

The informal procedures outlined herein should be useful in many of the problems for which a Friedman analysis is appropriate. Specifically, ANACONDA may be helpful in identifying agreement between and within subgroups of judges. Interpretable clusters may indicate future breakdowns or sub-groupings of judges as well as point out potential outliers and violations of the no block-treatment interaction assumption.

REFERENCES

Beckett, J. (1975). Some properties and applications of a statistic for analyzing concordance of rankings of groups of judges. Ph.D. dissertation, Southern Methodist University.

Beckett, J. and Schucany, W.R. (1975). ANACONDA: Analysis of concordance of g groups of judges. Proceedings of the Social Statistics Section of the American Statistical Assn., 311-313. (Presented at national ASA meeting in Atlanta, Aug. 75).

Bliss, C.I. (1967). Statistics in Biology. McGraw-Hill Book Co., New York.

Hollander and Wolfe (1973). Nonparametric Statistical Methods. John Wiley & Sons.

Gibbons, J.D. (1976). Nonparametric Methods for Quantitative Analysis. Holt, Rinehart, and Winston.

Miller, R.G. (1966). Simultaneous Statistical Inference. New York: McGraw-Hill Book Co.

Page, E.B. (1963). "Ordered hypothesis for multiple treatments: a significance test for linear ranks," Journal of the American Statistical Association, 58, 216-230.

Schucany, W.R. and Beckett, J. (1976). Analysis of multiple sets of incomplete rankings. Communications in Statistics, 5, 1327-1334. (Special issue: Recent theory and applications of nonparametric statistics.)

Schucany, W.R. and Frawley, W.H. (1973). "A rank test for two group concordance," Psychometrika, 38, 249-258.

Woodward, W.F. (1970). A comparison of base running methods in baseball. M.S. thesis, Florida State University.

SEQUENTIAL TESTS FOR THE COEFFICIENT OF CORRELATION EXACT WALD REGIONS, OPERATING CHARACTERISTIC AND AVERAGE SAMPLE NUMBER

Don B. Campbell, Vidya Taneja, Western Illinois University
Leo A. Aroian, Union College and University

ABSTRACT

It is shown how exact Wald regions for the sequential probability ratio test for the coefficient of correlation $\rho = \rho_0$ versus $\rho = \rho_1$ may be found, and also how to determine the operating characteristic function, OC, and the average sample number ASN, by Monte Carlo techniques. A two decision example, and a three decision example $\rho = \rho_0$ versus $\rho = \rho_1$ and $\rho = \rho_2$ are included.

1. Introduction

Let $\{x_{1i}, x_{2i}\}$ be pairs of observations given by a normal bivariate distribution with unknown parameters μ_1, μ_2, σ_1, σ_2, and ρ. We determine exact regions for the Wald sequential test: H_0, $\rho = \rho_0$ versus H_1, $\rho = \rho_1$, $\rho_1 > \rho_0$. The values of the operating characteristic function, OC, and the average sample number ASN, are approxomated by Monte Carlo methods. We expect to obtain exact results in the future. So far as the authors know, no exact Wald regions for the coefficient of correlation have been determined. A FORTRAN program is available so any desired regions may be constructed very quickly.

2. Description of the Test

Let

$$\bar{x}_{1n} = \sum_{i=1}^{n} x_{1i}/n, \quad \bar{x}_{2n} = \sum_{i=1}^{n} x_{2i}/n,$$

$$s_{1n}^2 = \sum_{i=1}^{n} (x_{1i}-x_{1n})^2/n,$$

$$s_{2n}^2 = \sum_{i=1}^{n} (x_{2i}-\bar{x}_{2n})^2/n,$$

$$r_n = \sum_{i=1}^{n} (x_{1i}-\bar{x}_{1n})(x_{2i}-\bar{x}_{2n})/n s_{1n} s_{2n},$$

define the two sample means, the sample variances, and the sample coefficient of correlation respectively after observation n is taken. We test the hypothesis H_0: $\rho = \rho_0$, versus H_1: $\rho = \rho_1$, $\rho_1 > \rho_0$. The Wald sequential probability ratio test limits are given by $r_n(u)$, the upper limit for r_n, and $r_n(\ell)$ the lower limit for r_n. As soon as $r_n \le r_n(\ell)$ accept $\rho = \rho_0$, and as soon as $r_n \ge r_n(u)$ accept $\rho = \rho_1$. The values of $r_n(u)$ and $r_n(\ell)$ are determined as follows.

First $Z_n(r_n)$ must be found:

$$Z_2(r_2) = \ln(\pi - 2\sin^{-1}\rho_1) - \ln(\pi - 2\sin^{-1}\rho_0), \quad \text{if } r_2 = -1,$$

$$= \ln(\pi + 2\sin^{-1}\rho_1) - \ln(\pi + 2\sin^{-1}\rho_0), \quad \text{if } r_2 = 1. \tag{2.1}$$

$$Z_n(r_n) = .5(n-1)(\ln(1-\rho_1^2) - \ln(1-\rho_0^2)) - (n-1.5)(\ln(1-\rho_1 r_n) - \ln(1-\rho_0 r_n)) + \ln F(.5,.5,n-.5;\ .5(1+\rho_1 r_n)) - \ln F(.5,.5,n-.5;\ .5(1+\rho_0 r_n)), \text{if } n_2 > 2.$$

Note that $F(v,v',v";z) = \sum_{j=0}^{\infty} \frac{\Gamma(v+j)\Gamma(v'+j)\Gamma(v")}{\Gamma(v)\Gamma(v')\Gamma(v"+j)} (z^j/j!)$ is the hypergeometric function. Let $b=\ln(\beta/(1-\alpha))$ and $a=\ln(1-\beta)/\alpha$. If $n = 2$ and $r_2 = -1$, and $Z_2(-1) \le b$, accept $\rho = \rho_0$; if $n = 2$ and $r_2 = 1$, and $Z_2(1) \ge a$, accept $\rho = \rho_1$. If $n \ge 3$, r_n is computed from $Z_n(r_n) = b$ and $Z_n(r_n) = a$ and $\rho = \rho_0$ or $\rho = \rho_1$ is accepted depending on whether $r_n \le r_n(\ell)$, or $r_n \ge r_n(u)$ where $r_n(\ell)$ and $r_n(u)$ are solutions of $Z_n(r_n(\ell)) = b$, and $Z_n(r_n(u)) = a$.

This test and all the preceding results are due to B.K. Ghosh (1970). As n becomes large, the following approximations to $r_n(\ell)$ and $r_n(u)$ are useful:

$$(2.2)\quad r_n(u) \text{ or } r_n(\ell) = (W-1)/(W\rho_1 - \rho_0),$$

$$(2.3)\quad W = \{(1-\rho_0^2)/(1-\rho_1^2)\}^{\frac{1}{2}} \exp(\omega/n)$$

$$(2.4)\quad W = \{(1-\rho_0^2)/(1-\rho_1^2)\}^{\frac{1}{2}+\frac{1}{4}n} \exp(\omega/n-1.5)$$

Note ω is a dummy variable to be replaced by b or a. Formula (2.3) is correct to $0(n^{-1})$ and formula (2.4) is correct to $0(n^{-2})$. If b or a is used in the exponentials in formulas (2.3) or (2.4) for ω, and if the resulting W's are substituted in formula (2.2), then $r_n(\ell)$ or $r_n(u)$ is determined. If $n \to \infty$ we obtain

$$(2.5)\quad r_\infty = \{((1-\rho_0^2)(1-\rho_1^2))^{\frac{1}{2}} - 1\}/\{(\rho_1((1-\rho_0^2)/(1-\rho_1^2))^{\frac{1}{2}} - \rho_0\}$$

Equations (2.2), (2.3) and (2.4) are reformulations of those of Ghosh (1970), page 324, and are somewhat simpler to calculate. If $\rho_1 < \rho_0$, a simple interchange of ρ_0 and ρ_1 is used in (2.1) with corresponding changes in α and β.

3. Calculation of the Regions

$Z_n(r_n(\ell)) = b$ and $Z_n(r_n(u)) = a$ are solved by trial and error and repeated linear interpolation. The solutions are nearly correct to four decimal places throughout, but occasionally the fourth

decimal may be in error by as much as ± 2. A computer program is available from Don Campbell. The programming and calculation of the tables were efficiently handled by Sheri Butler. The approximate formulas for $r_n(\ell)$ and $r_n(u)$ (2.2) and (2.4) may be used to extent the tables.

4. Monte Carlo

In all cases 1000 values of the coefficient of correlations were calculated at the beginning of each run and these were continued until they went into the rejection region or acceptance region or were truncated at the truncation point, where they were placed in the acceptance or rejection region by the use of r_∞ given by (2.5). If $r_\infty \leq r(u)$ then H_1 was chosen, otherwise H_o. We generate two unit normal variates (Y_1, Y_2) with correlation coefficient as follows. First generate two independent unit normal variates U_1 and U_2 by the Box-Muller formulas

$$U_1 = (-2 \log R_1)^{\frac{1}{2}} \sin 2\pi R_2$$

$$U_2 = (-2 \log R_1)^{\frac{1}{2}} \cos 2\pi R_2$$

where R_1 and R_2 are random (0,1) variates. Then

$$Y_1 = U_1 \text{ and } Y_2 = \rho U_1 + (1-\rho^2)^{\frac{1}{2}} U_2,$$

as given by Wold (1948).

The direct method was used throughout these simulations. At each trial, the number of acceptances for H_o, H_1 and the number continuing into the next trial are found. Thus, at each exit point the number of items for each value of ρ has been determined and the distribution at this point may be found using the values of $\rho = -1$, $\rho_o - \Delta$, ρ_o, $\rho_o + \Delta$, $\rho_o + 2\Delta$, $\rho_o + 3\Delta$, ρ_1, $\rho_1 + \Delta$, and 1, where $\Delta = .25(\rho_1 - \rho_o)$. From this distribution an estimate of ρ may be made and also approximate confidence limits may be found provided the Monte Carlo trials are sufficiently extensive.

The direct method not only provides the OC and ASN but gives the DSN, decisive sample number distribution, and the conditional distribution at each point.

5. Example, $\alpha = \beta = .10$, $\rho_o = 0$, $\rho_1 = .25$

As an example, we choose $\alpha = \beta = .10$, $\rho_o = 0$, $\rho_1 = .25$. We give the region, the point of truncation m_o, the OC and the ASN, the conditional distribution at one point, the estimate for ρ after a sequential decision has been reached, and the sample size N for the corresponding fixed size test with the same α and β. Corresponding results for a variety of other cases are given in another paper, Taneja et al (1977). The region is given in Table 1. The truncation point is $m_o = 124$, the fixed size sample is N=104. The OC and ASN are given in Table 2. Usually the truncation point m_o was chosen as 1.2 times the fixed size sample N.

The actual value of α is $\alpha^1 = .111$ instead of .10 while the actual value of β is $\beta^1 = .105$ instead of the planned .10. The Monte Carlo trials, 1000, are too few to estimate α if the trials were continued to infinity. The greatest value of the ASN is 75.81, so the fixed size sample test is 73% efficient and only 55% efficient at $\rho_o = 0$ and $\rho_1 = .25$. This shows the real savings in observations needed to reach a decision.

Replications of the Monte Carlo trials show that the OC may be off as much as one unit in the second decimal place, and the ASN by one unit in the second digit at $\rho = \rho_o$ and $\rho = \rho_1$. Elsewhere the errors are somewhat larger.

Suppose the test terminates with acceptance at observation 25, what is the estimate of ρ using the mean value of the conditional distribution and approximate confidence limits for ρ based on this result? From the Monte Carlo trials at decision point 25 (not given here) we have the results:

$\rho = -1$,	$-.0625$,	0,	.0625	.125
0	10	13	8	7
0	.244	.317	.195	.171

(continued)

$\rho = .1875$	.25,	.3125,	1
1	2	0	0
.024	.049	0	0

where the first line are the values of ρ, the second line the number of times out of the 1000 trials that the test was terminated at 25, and the last line the estimated probabilities.

The mean value of this estimated distribution is the estimated value of ρ, .035, based on the 41 exits at this point.

6. Theory of the Three Decision Test

For three decision test we choose H_o $\rho = \rho_o$ versus a two-sided alternative H_1, $\rho = \rho_1$, $\rho_1 > \rho_o - \Delta$, and $\rho = \rho_2$, $\rho_2 < \rho_o + \Delta$.

The operating characteristic function is given piecewise:

$OC_o(\rho)$ = probability of accepting H_o, that is, $\rho = \rho_o$,

$OC_1(\rho)$ = probability of accepting $\rho = \rho_1$,

$OC_2(\rho)$ = probability of accepting $\rho = \rho_2$

and

$OC_1 + OC_o + OC_2 = 1$.

We choose α as follows:

$OC_o(\rho_o | \rho = \rho_o) = 1-2\alpha$,

$OC_1(\rho_1 | \rho = \rho_1) = 1-\alpha$, $OC_2(\rho_2 | \rho = \rho_2) = 1-\alpha$.

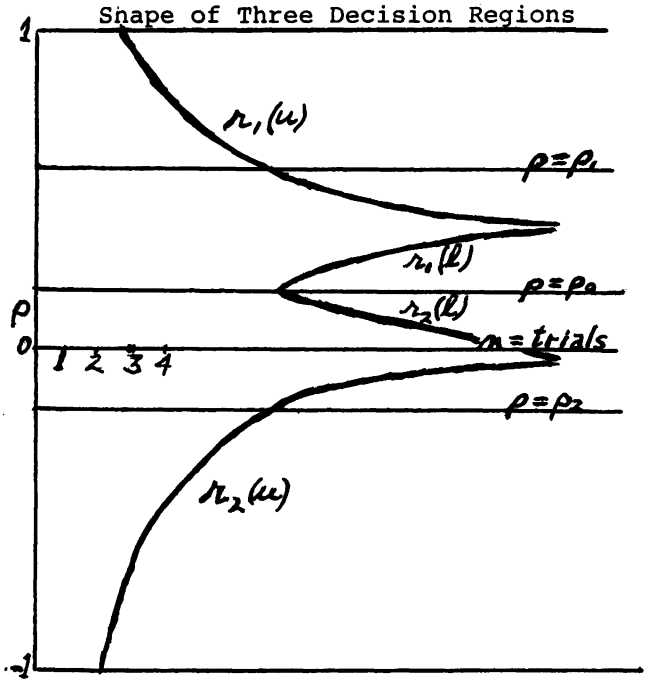
Thus, if α is chosen as .10, the probability of rejecting $\rho = \rho_o$ when $\rho = \rho_o$ is .20, while if either $\rho = \rho_1$ or ρ_2 the probability of rejecting $\rho = \rho_1$ or ρ_2 is .10 when $\rho = \rho_1$ or $\rho = \rho_2$ is true.

The region for the three decision test is determined by combining two two decision regions. First a two decision region is found for $\rho = \rho_o$ versus $\rho = \rho_1$ with $\alpha = \beta$, $\rho_1 = \rho_o + \Delta$, $\Delta > 0$, since $\rho_1 > \rho_o$ in formula (2.4). This gives us $r_1(u)$ and $r_1(\ell)$ boundaries. Next, a two decision region is found with $\rho_1^1 = -(\rho_o - \Delta)$, and $\rho_o^1 = -\rho_o$, since $\rho_1^1 > \rho_o^1$ in formulas (2.1)-(2.5), replacing ρ_1 and ρ_o by ρ_1^1 and ρ_o^1. This gives the values of $r_2(u)$ and $r_2(\ell)$. The two regions are combined. The values of $r_i(u)$ are unchanged, but the values of $r_i(\ell)$ are deleted until they intersect on the line $\rho = \rho_o$.

Intuitively we may expect the two regions to be symmetric about the line $\rho = \rho_o$. This happens if $\rho_o = 0$, but not otherwise since the distribution of r is only symmetric if $\rho = 0$. If we choose any two decision plan with $\alpha = \beta$ then the three decision regions will have approximately $OC(\rho = \rho_o) = 1-2\alpha$, $OC(\rho=\rho_1) = 1-\alpha = OC(\rho=\rho_2)$. If we choose a two decision plan with $\alpha = .5\beta$, then approximately $OC(\rho=\rho_o) = OC(\rho=\rho_1) = OC(\rho=\rho_2) = (1-\alpha)$. We may, of course, combine the two two-decision regions in such a way $OC(\rho=\rho_o) = 1-\alpha_1$, $OC(\rho=\rho_1) = 1-\alpha_2$ and $OC(\rho=\rho_2) = 1-\alpha_3$. The three decision regions always have the following shape, Figure 1:

FIGURE 1

Shape of Three Decision Regions



These three decision regions are essentially generalized Barnard regions. They are also similar to Wald-Sobel regions except no decisions are possible until one of the boundaries is reached.

8. Example of a Three Decision Test

We take the two decision test with $\alpha = \beta = .10$, $\rho_o = 0$ and $\rho_1 = .25$. We rotate the region about $\rho = 0$ to obtain the region below $\rho = 0$. The values of $r_i(u)$ are now plus and minus, the values of the middle section $r_i(\ell)$ start at trial 70 as $\pm.0017$, trial 71 $\pm.0030$, trial 72 $\pm.0053$, . . ., and at trial 124 $\pm.0574$. Thus, if the value of r exceeds $r_1(u)$ make the decision $\rho = \rho_1 = .25$, if r lies between $\pm r_i(\ell)$ conclude $\rho = \rho_o = 0$, or if $r < r_2(u)$ conclude $\rho = \rho_2 = -.25$.

The results for the OC and ASN are given in Table 3. The actual values of the OC are very close to the planned values .90, .80, .90 at $\rho = -.25$, 0, .25, namely .901, .797, and .890. The OC_{-1} and OC_1 are symmetric to each other about zero. The actual values mirror this and serve as a check on the Monte Carlo trials. The fixed size test takes 104 observations, so the efficiency of it varies from 11% to 73% as compared to the sequential test.

TABLE 1.

Values of $r_n(\ell)$ and $r_n(u)$ $\rho_o = 0$, $\rho_1 = .25$, $\alpha = \beta = .10$.

n	$r_n(\ell)$	$r_n(u)$
1 - 10	(No decision possible)	
11	-.8707	.9307
12	-.7658	.8620
13	-.6807	.8043
14	-.6104	.7548
15	-.5514	.7120
16	-.5011	.6748
17	-.4577	.6420
18	-.4199	.6129
19	-.3867	.5868
20	-.3573	.5634
21	-.3310	.5424
22	-.3075	.5232
23	-.2863	.5056
24	-.2671	.4897
25	-.2495	.4749
26	-.2334	.4614
27	-.2187	.4488
28	-.2051	.4372
29	-.1926	.4264
30	-.1809	.4162
31	-.1701	.4068
32	-.1600	.3981
33	-.1506	.3897
34	-.1417	.3819
35	-.1335	.3745
36	-.1257	.3676
37	-.1184	.3610
38	-.1114	.3549
39	-.1049	.3489
40	-.0987	.3433

n	$r_n(\ell)$	$r_n(u)$
41	-.0928	.3381
42	-.0872	.3330
43	-.0819	.3282
44	-.0769	.3235
45	-.0721	.3190
46	-.0675	.3148
47	-.0632	.3109
48	-.0590	.3070
49	-.0549	.3033
50	-.0511	.2998
52	-.0439	.2931
54	-.0373	.2869
56	-.0311	.2811
58	-.0254	.2758
60	-.0201	.2707
62	-.0106	.2661
64	-.0106	.2617
66	-.0062	.2577
68	-.0021	.2538
70	.0017	.2502
72	.0053	.2467
74	.0087	.2434
76	.0119	.2403
78	.0150	.2375
80	.0178	.2346
82	.0206	.2320
84	.0233	.2295
86	.0257	.2271
88	.0281	.2249
90	.0304	.2227
92	.0325	.2206
94	.0346	.2186
96	.0365	.2166
98	.0385	.2148
100	.0403	.2130
102	.0420	.2114
104	.0437	.2097
106	.0453	.2082
108	.0468	.2067
110	.0484	.2052

Table 1 (Continued)

n	$r_n(\ell)$	$r_n(u)$
112	.0498	.2038
114	.0512	.2025
116	.0525	.2011
118	.0538	.1999
120	.0551	.1987
122	.0563	.1975
124	.0574	.1963

REFERENCES

Aroian, L.A. Applications of the direct method in sequential analysis, Technometrics, 1976, 18, 301-306.

Box, G.E.P. and Muller, M.E., A note on the generation of random normal deviates, Annals of Mathematical Statistics, 1958, 29, 610-611.

Ghosh, B.K., *Sequential tests of statistical hypotheses*, Reading, Mass.: Addison-Wesley, 1970.

Taneja, V., Campbell, D. and Aroian, L.A., Tables of the regions, operating characteristic function, and average sample number for Wald sequential tests of the coefficient of correlation, submitted to Psychometrika, 1977.

Wold, H.O., *Random normal deviates.* Tracts for Computers, No. 25, Cambridge, England: Cambridge University Press, 1948.

TABLE 2.
OC and ASN, $\rho_o = 0$, $\rho_1 = .25$, $\alpha = \beta = .10$

ρ	-1	-.0625	0	.0625	.125	.1875	.25	.3125	1
OC	1	.957	.889	.749	.487	.272	.105	.036	0
ASN	11	48.49	57.10	68.78	75.81	69.30	57.27	45.70	11

TABLE 3.
OC and ASN, Three Decision Test, $\rho_2 = -.25$, $\rho_o = .0$, $\rho_1 = .25$

ρ	-1.	-.3125	-.25	-1875	-.125	-.0625	0	.0625	.125	.1875	.25	.3125	1.
OC_{-1}	1.	.967	.901	.722	.475	.247	.092	.035	.007	.003	.003	.002	0
OC_o	0	.033	.097	.271	.506	.718	.797	.724	.527	.277	.107	.027	0
OC_1	0	.000	.002	.007	.019	.035	.111	.241	.466	.720	.890	.971	1
ASN	9	46.50	57.30	69.44	76.27	74.56	75.76	74.81	76.50	69.06	56.08	45.94	9.

THE VALIDATION OF MODELS BASED ON LATENT STRUCTURES

C. Mitchell Dayton and George B. Macready, University of Maryland

A. INTRODUCTION

Many research areas share a common methodological concern with fitting pre-established dependency structures to data gathered from human subjects under conditions which introduce error into the measurement processes. Data are summarized, often, as a series of "successes" or "failures", while a theoretical model postulates some kind of sequential dependency among the tasks. The purpose of this paper is to summarize a class of probabilistic models which is useful for analyzing data purported to reflect hierarchic structures. The historical antecedents for the models discussed in this paper stem from the work of Lazarfeld and Henry (1968). Recent advances in estimation and hypothesis testing are due to Proctor (1970), Murray (1971), Goodman (1974, 1975, 1976), Dayton and Macready (1976), and Macready and Dayton (1977a). For a more complete overview of the theory underlying these models and for applications to real data sets, the above references, as well as Macready and Dayton (1977b), may be consulted.

B. THE GENERAL MODEL AND SOME SPECIAL CASES

It is assumed that all respondents (subjects) can be, in theory, identified with a set of "latent classes" which represent the levels of an a priori hierarchic structure. Furthermore, this presentation is limited to dichotomous response data; that is, we assume K distinct tasks, each of which can be scored 0,1 for a sample of n respondents (such 0,1 scoring may result from a true point variable, or from artificial dichotomization of a continuous variable). For convenience, let $\underline{u}_s$ be an observed response vector with elements 0,1 and let $\underline{v}_j$ be one of q theoretical vectors corresponding to the latent classes in the hierarchic structure. The basic concept of the general probabilistic model is that the observed vectors arise from the theoretical vectors due to response errors which obey a law of local independence. Using the notation $P(\cdot)$ for probabilities, the model is:

$$(1)\quad P(\underline{u}_s) = \sum_{j=1}^{q} P(\underline{u}_s \mid \underline{v}_j)\cdot\theta_j$$

$$(2)\quad P(\underline{u}_s \mid \underline{v}_j) = \prod_{i=1}^{K} \alpha_i^{a_{ijs}}(1-\alpha_i)^{b_{ijs}}\beta_i^{c_{ijs}}(1-\beta_i)^{d_{ijs}}$$

where the parameters are:

- θ_j - the true proportion of respondents which falls in the j^{th} latent class
- α_i - the probability of an intrusion error on task i
- β_i - the probability of an omission error on task i
- $a_{ijs}, b_{ijs}, c_{ijs}, d_{ijs}$ - numerical coefficients which are 0,1 and which relate the observed vector to the theoretical vector (methods for determining these coefficients are shown in connection with special cases of the model which are described below)

Note that $P(\underline{u}_s \mid \underline{v}_j)$ is the conditional probability that the observed vector, $\underline{u}_s$, arises from the j^{th} latent class through the occurrence of appropriate intrusion and/or omission errors. Such conditional probabilities are, in turn, generated by a product of probabilities associated with the individual tasks. The association of such (unconditional) probabilities with the tasks is equivalent to assuming a condition of local independence in the sense that a respondent's behavior is independent (without memory) across tasks.

Although the general model as presented in equations (1) and (2) can be fitted to data under certain circumstances, most of the applications which have been pursued to date have centered about simplified forms of the model. In Section C., we present some special cases so that the form of the models and the notation utilized are made clear. Two classes of models are distinguished: Extreme Groups Models and Hierarchic Models. In the cases of Extreme Groups, there are only two theoretical vectors - one corresponding to complete "failure", $\underline{v}_1$ = (0 0 0 ... 0), and one corresponding to complete "success", $\underline{v}_2$ = (1 1 1 ... 1). All other observed vectors must arise by intrusion or omission errors. Two special cases of the Extreme Groups Model are:

Case 1 - each task has an unique intrusion and omission error component (α_i and β_i);

Case 2 - all intrusion occurs at a constant rate (α) and all omission at a constant rate (β).

Within the class of Hierarchic Models, we include all linear and non-linear (e.g., convergent or divergent) hierarchies of arbitrary complexity. Four special cases of the Hierarchic Model are distinguished:

Case 1 - α_i and β_i as for the Extreme Groups Model, above;

Case 2 - separate error rates (α_i) per task, but intrusion and omission occur at this same rate for a given task (that is, Case 1 with $\alpha_i = \beta_i$);

Case 3 - intrusion (α) and omission (β) constant across tasks as in Case 2 of the Extreme Groups Model, above;

Case 4 - a single error rate for intrusion and omission across all tasks (i.e., Case 3 with $\alpha = \beta$).

It is evident that many other special cases can be defined by appropriate restrictions (or generalizations) with respect to the operation of errors. However, the cases referenced above have been studied both theoretically and practically. Ordinarily, the hierarchic structure (set of latent classes) can be specified in detail on the basis of a priori considerations (e.g., a Guttman scale implies a linear hierarchy), but the way in

which error probabilities enter into the model is more open to speculation. Thus, the various cases distinguished for Hierarchic Models permit some flexibility in fitting real data sets.

C. PARAMETRIZATION OF THE MODELS

(1) Case 1 of the Extreme Groups Model - For this special case, the only a priori vectors are $\underline{v}_1' = (0\ 0\ 0\ \ldots\ 0)$ and $\underline{v}_2' = (1\ 1\ 1\ \ldots\ 1)$ and the probabilistic model in equations (1) and (2) simplifies considerably. Thus,

$$P(\underline{u}_s) = P(\underline{u}_s \mid \underline{v}_1)\cdot\theta_1 + P(\underline{u}_s \mid \underline{v}_2)\cdot\theta_2$$

Further, let the elements in $\underline{u}_s$ be denoted x_{is}, so that $\underline{u}_s' = (x_{1s}, x_{2s}, \ldots, x_{Ks})$; then,

$$P(\underline{u}_s \mid \underline{v}_1) = \prod_{i=1}^{K} \alpha_i^{x_{is}}(1-\alpha_i)^{1-x_{is}}$$

$$P(\underline{u}_s \mid \underline{v}_2) = \prod_{i=1}^{K} \beta_i^{1-x_{is}}(1-\beta_i)^{x_{is}}$$

The topic of estimating parametric values from real data sets is discussed in Section D., below; since, in general, the various patterns of observed score vectors may have different probabilities of occurring under the models, it is apparent that estimation must be based on the total set of 2^K observed score vectors. With the exception of Case 2 of the current model, this requirement to have data summarized for each of the 2^K possible observed score vectors holds for all of the special cases considered in this paper.

(2) Case 2 of the Extreme Groups Model - If we restrict the intrusion errors, α_i, to a constant value (α) across the K tasks and the omission errors, β_i, to a constant value (β) across the K tasks, Case 2 is obtained. For this case, the probabilistic model takes on an especially simple form since the number of errors necessary to account for the observed score vectors is a function of the total "score" $X = \sum_{i=1}^{K} x_i$ associated with such a vector. Thus,

$$P(X \mid \underline{v}_1) = ({}_KC_X)\alpha^X(1-\alpha)^{K-X}$$

$$P(X \mid \underline{v}_2) = ({}_KC_X)\beta^{K-X}(1-\beta)^X$$

where ${}_KC_X$ is the combinations operator. Note that each of these conditional probabilities is of the form of a binomial and the model becomes, in effect, the mixture of two binomial processes with binomial parameters α and β, and with mixture, θ_1. Since all observed score vectors which yield the same score, X, have the same probability of occurring under this model, data can be analyzed from scores alone. That is, unlike Case 1 where the 2^K patterns are needed, the data can be summarized as K+1 score frequencies. Of course, this simplification makes the model less flexible with respect to representing real data sets.

(3) Case 1 of the Hierarchic Model - This model is summarized in equations (1) and (2) without simplification. Unfortunately, for arbitrary hierarchies (including linear forms), it does not seem to be possible to obtain estimates for all of the parameters simultaneously from real data sets by conventional estimation procedures (maximum likelihood). For this reason, the model as embodied in Case 1 is non-identifiable and we must turn our attention to the restricted cases in order to arrive at practical solutions.

(4) Case 2 of the Hierarchic Model - For this case, the intrusion and omission error rates are restricted to be equal for a given task, but each task has an unique error parameter (α_i). Thus, the model in equation (2) becomes:

$$P(\underline{u}_s \mid \underline{v}_j) = \prod_{i=1}^{K} \alpha_i^{a_{ijs}}(1-\alpha_i)^{1-a_{ijs}}$$

where a_{ijs} is 0,1 and determined as follows: let the i^{th} element in $\underline{u}_s$ be x_{is} and the i^{th} element in $\underline{v}_j$ be t_{ij}. Then,

$$a_{ijs} = \begin{cases} 0 & \text{if } x_{is} - t_{ij} = 0 \\ 1 & \text{otherwise} \end{cases}$$

In effect, whenever corresponding elements in the observed and theoretical vectors fail to match, a_{ijs} is given the value 1 and this introduces the error parameter into the model for this task. Otherwise, the value $1-\alpha_i$ enters and this is the probability of not making an error for the i^{th} task.

(5) Case 3 of the Hierarchic Model - In dealing with hierarchic structures, we have had the most experience in applying this case since the number of parameters which must be estimated remains reasonably small even for fairly large numbers of tasks. The notion of separate intrusion and omission error rates is retained, but these rates are assumed to be constant (or homogeneous) across tasks. For notational purposes, let α be this constant intrusion error rate and β be the constant omission error rate. Then, the model in equation (2) becomes:

$$P(\underline{u}_s \mid \underline{v}_j) = \alpha^{a_{js}}(1-\alpha)^{b_{js}}\beta^{c_{js}}(1-\beta)^{d_{js}}$$

where the coefficients a_{js}, b_{js}, c_{js}, and d_{js} are determined from the following rules based on the elements x_{is} of $\underline{u}_s$ and the elements t_{ij} of $\underline{v}_j$:

a_{js} is the number of times $t_{ij} = 0$ when $x_{is} = 1$ (number of intrusions)

b_{js} is the number of times $t_{ij} = 0$ when $x_{is} = 0$ (number of non-intrusions)

c_{js} is the number of times $t_{ij} = 1$ when $x_{is} = 0$ (number of omissions)

d_{js} is the number of times $t_{ij} = 1$ when $x_{is} = 1$ (number of non-omissions)

D. ESTIMATION OF PARAMETERS

Within certain broad limits of identifiability, parameter estimates can be obtained by means of computer-based algorithms for both types of Extreme Groups Model and for cases 2, 3, and 4 of the Hierarchic Model. The method of estimation which is employed ordinarily is that of maximum likelihood. Unfortunately, there are no simple, algebraic formulae which can be derived for these models since they are non-linear in the parameters. Nevertheless, computerized procedures can locate the maximum likelihood estimate if initial guessed values for all parameters are used and, then, iteratively improved until they converge on the appropriate values. Programs developed by us have been based on Fisher's method of "scoring" and, in general, the final solution does not depend upon good choices for initial guessed values (i.e., the algorithm is relatively insensitive to starting values). However, "boundary problems" arise with some regularity, and the programs have options which will force the final solution to take on acceptable values (i.e., all the θ_j, α_i, and β_i are restricted to the interval 0,1)[1].

For a total of n respondents, the likelihood for the sample is:

$$(3)\quad L = \prod_{s=1}^{n} P(\underline{u}_s) = \prod_{s=1}^{n} \sum_{j=1}^{q} P(\underline{u}_s / \underline{v}_j) \cdot \theta_j$$

and the general method of solution involves solving the system of partial derivatives:

$$\partial \mathrm{Log}_e L / \partial \theta_j = 0 \quad , \quad j = 1,\ldots,q-1$$

$$\partial \mathrm{Log}_e L / \partial \alpha_i = 0 \quad , \quad i = 1,\ldots, K$$

$$\partial \mathrm{Log}_e L / \partial \beta_i = 0 \quad , \quad i = 1,\ldots, K$$

with suitable restrictions placed on the α_i and β_i to provide non-singularity (identifiability) for the system. Computation of the derivatives is greatly simplified if Fisher's method of scoring (Rao, 1965) is used and solution of the system can be pursued iteratively by the method of Newton-Raphson. An important by-product of this approach is that the matrix of partial second derivatives provides a basis for estimating large-sample sampling variances of the parameter estimates (i.e., the inverse of this matrix, with signs changed, contains asymptotic variances and covariances for the estimates when maximum likelihood estimates are substituted in the second partial derivatives). Further discussion of the conditions for identifiability and problems concerning boundaries for the estimates is presented in Dayton and Macready (1976).

E. ASSESSING GOODNESS OF FIT

Given the computerized estimation procedures which are available, it is possible to derive maximum likelihood estimates of the parameters if there are sufficient degrees of freedom once all parameters are specified and if the system of equations based on (3) is identifiable. However, the estimates do not necessarily provide good fit to the observed data. An assessment of goodness of fit can be made in several ways, but the simplest procedure is to utilize the maximum likelihood estimates of $P(\underline{u}_s)$ for each of the 2^K types of observed score vectors and, then, to apply an ordinary (Pearson) chi-square goodness-of-fit test based on observed and expected frequencies for these 2^K types. An alternative method which yields generally comparable values for the test statistic is the likelihood ratio chi-square test which, in effect, compares the expected frequencies (generated as for the Pearson case) with those arising under an unrestricted multinomial model. Degrees of freedom for both types of test are computed as $2^K - m - 1$, where m is the number of independent parameters estimated under the probabilistic model (i.e., for Case 1 of the Extreme Groups Model $m = 2K + 1$, while for Case 2, $m = 3$; for Hierarchic Models, under Case 2, $m = K + q - 1$, under Case 3, $m = q + 1$, and under Case 4, $m = q$).

In addition to assessing how well a given model fits an observed set of data, it is possible to compare the differential fit of alternate models under certain circumstances. A general rule is that the model with fewer parameters must be derivable, in theory, by a process of parameter restriction from the model which has the greater number of parameters. For example, Cases 1 and 2 of the Extreme Groups Model can be compared for differential fit since Case 2 can be derived from Case 1 by the restrictions $\alpha_i = \alpha$, $\beta_i = \beta$ for $i = 1,\ldots,K$. However, Cases 2 and 3 of the Hierarchic Model cannot be compared since neither case can be obtained from the other by a single set of restrictions; note, nevertheless, that Case 4 can be derived from either Case 2 or Case 3 and can be compared with either of these. An appropriate test statistic for comparing the relative fits of two models which meet the preceding conditions can be based on the difference between their respective goodness-of-fit chi-square values (based on either the Pearson or likelihood ratio approach) with degrees of freedom equal to the difference in degrees of freedom from these same two tests.

FOOTNOTE

[1]The authors will make available at no cost a Users Manual and single-copy listings of FORTRAN programs for all cases discussed above, with the exception of Case 1 of the Hierarchic Model. Written requests should be sent to the Department of Measurement & Statistics, College of Education, University of Maryland, College Park, Md. 20742

REFERENCES

Dayton, C. M. & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189 - 204.

Goodman, L. A. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

_____ The analysis of systems of qualitative variables when some of the variables are unobservable. Part I - a modified latent structure approach. American Journal of Sociology, 1975, 79, 1179-1259.

_____ A new model for scaling response patterns: an application of the quasi-independence concept. Journal of the American Statistical Association, 1976, 70, 755-768,

Lazarfeld, P. F. & Henry, H. W. Latent Structure Analysis. Boston: Houghton-Mifflin, 1968.

Macready, G. B. & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977a, 2, 99-120.

_____ Statistical comparisons among hierarchies based on latent structure models. Paper read the the Annual Meeting of the American Educational Research Association, New York, 1977b

Murray, J. R. Statistical models for qualitative data with classification errors. Ph.D. Thesis, University of Chicago, 1971.

Proctor, C. H. A probabilistic formulation and statistical analysis for Guttman scaling. Psychometrika, 1970, 35, 73-78.

Rao, C. R. Linear Statistical Inference and its Applications. New York: Wiley, 1965.

GLOSSARY OF NONSAMPLING ERROR TERMS
AN ILLUSTRATION OF A SEMANTIC PROBLEM IN STATISTICS

Richard E. Deighton, U.S. Postal Service
James R. Poland, U.S. Postal Service
Joel R. Stubbs, Internal Revenue Service
Robert D. Tortora, Department of Agriculture

PREFACE

This glossary has been prepared for the OMB Federal Committee on Statistical Methodology by the Subcommittee on Nonsampling Errors. During subcommittee meetings it became obvious that the terminology on nonsampling errors left something to be desired. Consequently, the subcommittee decided to develop this glossary. A search of current literature for acceptable definitions highlighted the prevailing problem instead of providing a solution to it.

Prior to examining the contents of this glossary, it is important to understand what the purpose of the glossary is and what it is not. The purpose of the glossary is to highlight a semantic problem represented by the fact that:

- The same term is sometimes used with different meanings; and
- The same phenomenon is sometimes called by more than one term.

The glossary contributes towards the task of developing a standardized terminology. However, it is beyond the scope of the present OMB Subcommittee on Nonsampling Errors to pursue a task of this magnitude.

This glossary is not intended to be used as a dictionary for nonsampling error terms. The authors did not conduct an exhaustive search of the literature, nor did they attempt to select or specify a preferred definition for any term. The reference to a subcommittee document (O1 in the bibliography) is included solely for the convenience of the reader and does not exist as a separate document.

Terms in the glossary are listed alphabetically. The bibliography at the end contains the references which are used to obtain definitions for the terms. Codes for the references are alphanumeric (i.e., A1, A2, A3, etc. designate the first, second, and third references associated with the first letter of the author's name). For each term defined in the glossary, one or more references are given. For example, on Page 1, the term "ACCURACY" is found in three references, viz., B2, K1, S1.

If a definition is taken from a text book, the page number of the referenced text is included at the end of the definition. (See BIAS, CONSTANT - Reference H2, Page 17.) Quotation marks are used to identify when a definition was copied verbatim from a reference. The absence of quotation marks implies that the definition was paraphrased or that the definition was taken out of context. When a definition was taken out of context, an editors' note was often added to make the reader aware that the definition can be applied to a more general subject. The editors' note is enclosed in brackets [ ] (e.g., see BOUNDED RECALL). A "Comment" indicating similar terms which appear in the glossary has been added at the end of each relevant term.

Members of the Glossary Task Force wish to express their sincere appreciation to Professor Tore Dalenius for the criticism, suggestions, and reference material which he provided during the preparation of this glossary. Also, the members of the task force wish to thank all other members of the Subcommittee on Nonsampling Errors for their comments and suggestions on the numerous drafts of the glossary.

ACCURACY
- B2 "The quality of a survey result that is measured by the difference between the survey figure and the value being estimated. The true value is seldom known, although it can be approximated in some instances." p. 48.
- K1 "Accuracy in the general statistical sense denotes the closeness of computations or estimates to the exact or true values. In a more specialized sense the word also occurs as meaning (a) in relation to an estimator, unbiassedness; (b) in relation to the reciprocal of the standard error, the precision (q.v.). Neither usage can be recommended."
- S1 "Closeness to the true value."
  Comment - See CORRECT VALUE, SURVEY VALUE, and TRUE VALUE.

ALLOCATION
- O1 The process of assigning values to units in the nonresponse group of a survey according to the characteristics that have been observed for the response group or by any other imputation procedure.

AUDIT
- O1 The process of applying more extensive methods of measurement to a subsample during the scheduled conduct of a survey in order to determine the effect of nonsampling errors.
  Comment - See POST-AUDIT.

BIAS
- B2 "The difference between the expected value of an estimator and the value that would be obtained from all the population elements with no corresponding errors of measurement being made. This true value is what we are trying to estimate." p. 48.
- H2 The difference between the expected value of the estimator and the true value being estimated. Whenever the bias is 0, the estimator is said to be unbiased. p. 17.
- K1 "Generally, an effect which deprives a statistical result of representativeness by

systematically distorting it, as distinct from a random error which may distort on any one occasion but balances out on the average."
Comment - See SYSTEMATIC ERROR.

BIAS, CONSTANT
C1 That component of the total bias in a survey estimator that affects all of the units alike. p. 389.

BOUNDED RECALL
D4 "An interview where the respondent is reminded of what he reported in an earlier interview and is then asked only to report on any new events that occurred subsequent to the bounding interview."
N1 A method of interview that is designed to prevent shifting in time of expenditures reported by respondents.
[Editor's Note - Definition was given for expenditures but may apply to other characteristics.]
S2 "Bounded recall procedures involve a series of interviews with the same panel of respondents. At the beginning of the bounded interview, which is the second or later interview, the respondent is told about the expenditures reported during the previous interview, and is then asked about additional expenditures made since then. The interviewer also checks the new expenditures reported with previous expenditures to make sure that no duplication has occurred." p. 83.
[Editor's Note - Definition was given for expenditures but may apply to other characteristics.]

BOUNDING
B2 "Prevention of erroneous shifts of the timing of events by having the enumerator or respondent supply at the start of the interview (or in a mail survey) a record of events reported in the previous interview." p. 48.

CLASSIFICATION ERRORS
H2 Errors caused by conceptual problems and misinterpretations in the application of classification systems to survey data. p. 84.

CLASSIFICATION ERROR RATE
O1 The proportion of responses that have been incorrectly classified in a survey.

CODING
W1 "Coding is a technical procedure for converting verbal information into numbers or other symbols which can be more easily counted and tabulated." p. 234.

CODING ERROR
B3 Errors that occur during the coding of sample data.
O1 The assignment of an incorrect code to a survey response.

COMPILING ERRORS
H2 Errors introduced in operations on the original observations such as editing, coding, punching, tabulating and transcribing. p. 84.

COMPLETE COVERAGE
M2 "A survey (or census) should be called complete if virtually all of the units in the population under study are covered." p. 54.
Comment - See COVERAGE ERROR and INCOMPLETE COVERAGE.

COMPLETENESS RATE
H6 The completeness rate is the percentage of interviews in which the required information is given by the respondent. This rate reflects in part interviewer effectiveness in the interview. However, it is not independent of the response rate. A low response rate may imply that the respondents interviewed are more likely to be cooperative than is the case with a high response rate. In a sense, the interviewer with a low response rate can be thought of as disposing of his uncooperative sample members at the door and interviewing only the relatively cooperative ones, therefore obtaining a higher completeness rate. p. 13.

CONDITIONING EFFECT
B2 "The effect on responses resulting from the previous collection of data from the same respondents in recurring surveys." p. 48.

CONTENT ERROR
B2 "Errors of observation or objective measurement, of recording, of imputation, or of other processing results in associating a wrong value of the characteristic with a specified unit. (Coverage errors are excluded from this definition.)" p. 48.

CORRECT VALUE
C1 The value obtained for a unit that is measured without error. p. 374.
Comment - See ACCURACY, SURVEY VALUE, and TRUE VALUE.

COVERAGE ERROR
B2 "The error in an estimate that results from (1) failure to include in the frame all units belonging to the defined population; failure to include specified units in the conduct of the survey (undercoverage) and (2) inclusion of some units erroneously either because of a defective frame or because of inclusion of unspecified units or inclusion of specified units more than once, in the actual survey (overcoverage)." p. 48.
Comment - See NONCOVERAGE, OVERCOVERAGE and UNDERCOVERAGE.

DEFINED GOAL
B2 "The approximation to the true value that would be obtained if the survey were carried out using the specified frame, the method of measurement for the specified characteristic and the method of summarizing the measurements as required in the survey plan." p. 48.
H4 "Specifications actually set forth for the statistical survey, if carried out precisely and rigorously, would yield the defined goals."
Comment - See EXPECTED VALUE, IDEAL GOAL and SPECIFICATION.

DEFINITIONAL ERRORS
H2 Errors that occur in surveys whenever the definitions of the characteristics for which data are to be collected are not pertinent to the purposes of the survey or are not clear to the respondents. p. 83.

EDITING

P1 Identifying potential problems is the first objective of the editing process. We also classify as editing, operations performed on the record information that are designed to conform it to the desired format or units. Filling a blank on the basis of redundant information on the form is editing. Similarly, when a respondent indicates that he has reported in pounds whereas reporting in tons was requested, we would regard the conversion of his figure to the specified unit as an editing correction. We are also inclined to classify as editing the supplying of missing totals where the component detail has been reported.

W1 "Editing is a preliminary step in which responses are inspected, corrected and sometimes precoded according to a fixed set of rules." p. 234.
Comment - See EDITING CHANGE and IMPUTATION.

EDITING CHANGE

P1 A code that is inserted on a form as a result of an editing process. For example, where a woman is coded as the "wife" of the head of the household and the field for marital status is blank, the code for "married" may be inserted as an editing change in this case.
Comment - See EDITING and IMPUTATION.

ERROR

K1 "In general, a mistake or error in the colloquial sense. There may, for example, be a gross error or avoidable mistake; an error of reference, when data concerning one phenomenon are attributed to another; copying errors; an error of interpretation.
"In a more limited sense the word 'error' is used in statistics to denote the difference between an occurring value and its 'true' or 'expected' value. There is here no imputation of mistake on the part of a human agent; the deviation is a chance effect. In this sense we have, for example, errors of observation (q.v.), errors in equations (q.v.), errors of the first and second kinds (q.v.) in testing hypotheses, and the error band (q.v.) surrounding an estimate; and also the normal curve of errors itself."

ERROR PLANTING

D1 A method of control where a set of errors is introduced into the material being subjected to control. If the control were perfect, all the planted errors would be detected. In practice, only a fraction is detected. This fraction may obviously be used as a measure of the performance of the control operation. p. 153.

EQUAL COMPLETE COVERAGE

D3 "The equal complete coverage is by definition the result that would be obtained from investigation of all the sampling units in the frame (segments of area, business establishments, accounts, manufactured articles) by the same fieldworkers or inspectors, using the same definitions and procedures, and exercising the same care as they exercised on the sample, and at about the same period of time. The concept of the equal complete coverage is fundamental to the use of samples. The adjective equal signifies that the same methods must be used for the equal complete coverage as for the sample. Every sample is a selected portion of the sampling units in the frame; hence A SAMPLE IS A SELECTED PORTION OF RESULTS OF THE EQUAL COMPLETE COVERAGE."
Comment - See COVERAGE ERROR.

EXPECTED VALUE

H4 The hypothetical averages from the conceived replicates of the survey all conducted under the same essential conditions.
Comment - See DEFINED GOAL, IDEAL GOAL, and SPECIFICATION.

FOLLOW-UP

D4 "A procedure whereby those members of a selected sample for whom a response is not obtained by one data collection strategy (e.g., telephone or mail) are contacted by the same or another data collection strategy in order to increase response rate. It can also be used to designate repeated surveys among a panel of respondents."

FRAME

C1 A list of the units which make up the population. p. 7.

K2 "Physical lists and procedures that can account for all the sampling units without the physical effort of actually listing them." p. 53.

U1 The frame consists of previously available descriptions of the objects or material related to the physical field in the form of maps, lists, directories, etc., from which sampling units may be constructed and a set of sampling units selected; and also information on communications, transport, etc., which may be of value in improving the design for the choice of sampling units, and in the formation of strata, etc. p. 7.
Comment - See SAMPLED POPULATION and TARGET POPULATION.

GROSS DIFFERENCE

B4 The number of cases that are classified differently in the initial survey or census and its replication. p. 2.

IDEAL GOAL

H4 The set of statistics that would have been produced had all of the requirements been precisely defined and rigorously met constitutes the ideal goal of the statistical survey.
Comment - See DEFINED GOAL, EXPECTED VALUE, and SPECIFICATION.

IMPUTATION

P1 The process of developing estimates for missing or inconsistent data in a survey. Data obtained from other units in the survey are usually used in developing the estimate.
Example: An editing test classifies an age as wrong when a man is reported as 6 years of age, and also as head of the household, with a wife age 35 and a child age 10. A more rational figure than the 6 is supplied by some procedure such as using the same age difference between husband and wife as

appeared in the preceding household of similar type.
Comment - See EDITING and EDITING CHANGE.

INCOMPLETE COVERAGE
M2 A survey (or census) should be called incomplete if a substantial number of the units in the population under study are arbitrarily excluded. p. 54.
Comment - See COMPLETE COVERAGE and COVERAGE ERROR.

INDEX OF INCONSISTENCY
B1 The proportion of the total variance of a characteristic that is accounted for by the response variance.

INTERVIEWER BIAS
K1 Bias in the responses which is the direct result of the action of the interviewer.

INTERVIEWER ERROR
O1 Errors in the responses obtained in a survey that are due to actions of the interviewer.

INTERVIEWER VARIANCE
B2 "That component of the nonsampling variance which is due to the different ways in which different interviewers elicit or record responses." p. 48.

ITEM NONRESPONSE
B5 "The type of nonresponse in which some questions, but not all, are answered for a particular unit." p. 914.
O1 The type of nonresponse in which a question is missed for an interviewed unit.

LIMITS OF ERROR
D3 The limits of error are the maximum overestimate and the maximum underestimate from the combination of the sampling and the nonsampling errors.

MEAN SQUARE ERROR
K1 "The second moment of a set of observations about some arbitrary origin. If that origin is the mean of the observations, the mean-square deviation is equivalent to the variance (q.v.)."
S1 "Mean value over trials of the square of the response error. It may be expressed as the sum of variances, covariances and the square of the response bias."
O1 The variance of the estimate plus the bias squared.

MEASUREMENT ERROR
B2 "(1) As applied to individual units of analysis, measurement error means the difference between the observed or imputed value and the true value. (2) As applied to an estimate, measurement error means the difference between the estimate and the true value, thus including all sampling as well as nonsampling errors." p. 48.

MEMORY ERROR
M2 Errors associated with the recall of answers to questions about the past. p. 45.

NET DIFFERENCE
B4 The net difference of a tabulated figure for a given class is the difference between the total for the class obtained in the reinterview (or appropriate records) and the original surveys. p. 3.

NONCONTACT
O1 A type of nonresponse in which the interviewer has not been able to contact the respondent.

NONCOVERAGE
C1 "Failure to locate or to visit some units in the sample." p. 360.
K2 "Refers to the negative error of failure to include elements that would properly belong in the sample." p. 529.
Comment - See COVERAGE ERROR, OVERCOVERAGE, and UNDERCOVERAGE.

NONINTERVIEW
O1 The type of nonresponse in which no information is available from occupied sample units for such reasons as: not at home, refusals, incapacity and lost schedules.

NONINTERVIEW ADJUSTMENT
O1 A method of adjusting the weights for interviewed units in a survey to the extent needed to account for occupied sample units for which no information was obtained.

NONOBSERVATION ERROR
K2 "Failure to obtain data from parts of the survey population which results from two sources: noncoverage and nonresponse." p. 527.

NONRESPONDENT
M2 Those persons in a sample from whom information has not been obtained. p. 172.

NONRESPONSE
B2 "The failure to elicit responses for units of analysis in a population or sample because of various reasons such as absence from home, failure to return questionnaires, refusals, omission of one or more entries in a form, vacant houses, etc." p. 50.
C1 "We shall use the term nonresponse to refer to the failure to measure some of the units in the selected sample." p. 355.
K1 "In sample surveys, the failure to obtain information from a designated individual for any reason (death, absence, refusal to reply) is often called a nonresponse and the proportion of such individuals of the sample aimed at is called the nonresponse rate. It would be better, however, to call this a 'failure' rate or a 'non-achievement' rate and to confine 'nonresponse' to those cases where the individual concerned is contacted but refuses to reply or is unable to do so for reasons such as deafness or illness."
K2 "Nonresponse refers to many sources of failure to obtain observations (responses, measurements) on some elements selected and designated for the sample." p. 532.

NONRESPONSE RATE
D4 "The complement of response rate. The numerator is those eligible respondents selected in a sample for whom information is not obtained because of refusals, not found at home, unavailable by reason of illness, incompetence, language difficulty, etc. The denominator is the total number of eligible respondents initially selected for the sample." p. 46.

NONSAMPLING ERROR
B2 "The error in an estimate arising at any stage in a survey from such sources as varying interpretation of questions by enumerators, unwillingness or inability of respondents to give correct answers, nonresponse, improper coverage, and other sources exclusive of sampling error. This

definition includes all components of the Mean Square Error (MSE) except sampling variance." p. 50.

K1 "An error in sample estimates which cannot be attributed to sampling fluctuations. Such errors may arise from many different sources such as defects in the frame, faulty demarcation of sample-units, defects in the selection of sample-units, mistakes in the collection of data (due to personal variations or misunderstandings or bias or negligence or dishonesty on the part of the investigator or of the interviewee), mistakes at the stage of the processing of the data, etc.
"The term 'response error' is sometimes used for mistakes in the collection of data and would not, strictly speaking, cover errors due to nonresponse. The use of the word 'bias' in the place of error, e.g. 'response bias' is not uncommon. The term 'ascertainment error' (Mahalanobis) is preferable as it would include errors due to nonresponse and also cases of collection of data by methods other than interviewing, e.g. direct physical observation of fields for crop estimates."
Comment - See OBSERVATIONAL ERROR and RESPONSE ERROR.

NOT AT HOME

C1 "Persons who reside at home but are temporarily away from the house." p. 360.

OBSERVATIONAL ERROR

K1 "This term ought to mean an error of observation but sometimes occurs as meaning a response error."

K2 "Errors which are caused by obtaining and recording observations incorrectly." p. 520.
Comment - See NONSAMPLING ERROR and RESPONSE ERROR.

OVERCOVERAGE

K2 "Positive errors which occur due to the inclusion in the sample of elements that do not belong there." p. 529.
Comment - See COVERAGE ERROR, NONCOVERAGE, and UNDERCOVERAGE.

POST-AUDIT

O1 The process of applying more extensive methods of measurement to a subsample after the scheduled conduct of a survey in order to determine the effect of nonsampling errors.
Comment - See AUDIT.

PRECISION

B2 "The quality of a sample result that is measured by the difference between the sample result and the result which would be obtained if a complete count were taken using the same survey procedures. Same as reliability. Usually defined by stating the sampling error." p. 50.

C1 Refers to the size of deviations from the mean obtained by repeated application of the sampling procedure. p. 16.

H2 "The difference between a sample result and the result of a complete count taken under the same conditions ... or the reliability." p. 10.

H3 A measure of how close the set of possible sample estimates for a particular sample design may be expected to come to the value being estimated. p. 7.

K1 "In exact usage precision is distinguished from accuracy. The latter refers to closeness of an observation to the quantity intended to be observed. Precision is a quality associated with a class of measurements and refers to the way in which repeated observations conform to themselves; and in a somewhat narrower sense refers to the dispersion of the observations, or some measure of it, whether or not the mean value around which the dispersion is measured approximates to the 'true' value. In general, the precision of an estimator varies with the square root of the number of observations upon which it is based."

PREFERRED TECHNIQUE

D3 "Any result, whatever it be, is the result of applying some set of operations. Although there is no true value, we do have the liberty to define and to accept a specified set of operations as preferred, and the results thereof as a master standard (so-called by Harold F. Dodge). Thus, there may be, by agreement of the experts in the subject-matter, for any desired property of the material, a preferred survey-technique."
Comment - See WORKING TECHNIQUE.

PROCESS CONTROL

B2 A statistical quality control technique where frequent small samples are taken and evaluated to control clerical operations. p. 8.

QUALITY CHECK

M2 "An intensive study of a small sample (relative to the size of the survey) where every effort is made to attain the highest level of accuracy possible." p. 396.

QUALITY CONTROL

B2 "Observation and procedure used in any operation of a survey in order to prevent or reduce the effect of nonsampling errors." p. 50.

K1 "A method of controlling the quality of a manufactured product which is produced in large numbers. It aims at tracing and eliminating systematic variations in quality, or reducing them to an acceptable level, leaving the remaining variation to chance. The process is then said to be statistically under control."

RECALL

N1 A method of obtaining information by means of an interview in which the respondent is required to remember past events. A common application is the recall of consumer expenditures.
Comment - See BOUNDED RECALL and UNBOUNDED RECALL.

RECALL ERRORS

H2 "Many questions in surveys refer to happenings or conditions in the past, and there is a problem in both remembering the event and of associating it with the correct time period." p. 84.

RECALL LOSSES

N1 Omissions of expenditures due to forgetting of items.

[Editor's Note - Definition was given for expenditures but may apply to other characteristics.]

RECALL PERIOD

N1 Refers to the period of time for which the respondent's report of expenditures is to be utilized.
[Editor's Note - Definition was given for expenditures but may apply to other characteristics.]

RECORD CHECK

B2 "A study in which data on individual units obtained by one method of data collection are checked against data for the same units from available records obtained by a different method of data collection (for example, comparison of ages reported in census with information from birth certificates)." p. 50.

REFUSAL RATE

K1 "In the sampling of human populations, the proportions of individuals who, though successfully contacted, refuse to give the information sought. The proportion is usually (and preferably) calculated by dividing the number of refusals by the total number of the sample which it was originally desired to achieve."

RELEVANCE

H4 "Standards of relevance are concerned with the difference between the ideal goal of a survey and the statistics called for by the survey specifications."
Comment - See RELEVANCE ERROR.

RELEVANCE ERROR

O1 The difference between the ideal goal of a survey and the statistics called for by the survey specifications.
Comment - See RELEVANCE.

RELIABILITY

M1 The confidence that can be assigned to a conclusion of a probabilistic nature.
[Editor's Note - Translation taken from Crespo, see Reference S1.]

M2 "The extent that repeat measurements made by a scale or test under constant conditions will give the same result (assuming no change in the basic characteristic - e.g. attitude - being measured)." p. 353.

S1 "The degree of confidence in terms of probability associated with conclusions based on a random experiment."
Comment - See VALIDITY.

RESPONSE BIAS

B3 The difference between the average of the averages of the responses over a large number of independent repetitiions of the census and the unknown average that could be measured if the census were accomplished under ideal conditions and without error. p. 1.

S1 "Difference between average reported value over trials and true values. It is combined bias as algebraic sum of all bias terms representing diverse source of biases."

RESPONSE DEVIATION

B3 "The difference between the response recorded for a person on a particular trial and the average of the responses over all trials for the same person." p. 2.

S1 "Difference between individual reported value and the average over hypothetical trials under the same general conditions."

RESPONSE ERROR

B2 "That part of the nonsampling error which is due to the failure of the respondent to report the correct value (respondent error) or the interviewer to record the value correctly (interviewer error). It includes both the consistent response bias and the variable errors of responses which tend to balance out." p. 50.

S1 "Difference between reported and true value."
Comment - See NONSAMPLING ERROR and OBSERVATIONAL ERROR.

RESPONSE RATE

D4 "The percentage of an eligible sample for whom information is obtained. For an interview survey the numerator of the formula is the number of interviews. The denominator is the total sample size minus non-eligible respondents; that is, minus those not meeting the criteria for a potential respondent as defined for that particular study."

H6 "The percentage of times an interviewer obtains interviews at sample addresses where contacts are made, i.e.,

$$\frac{\text{Number of interviews}}{\text{Number of contacts}}.$$

Since a contact must be either an interview or a refusal, the response rate is also equal to 1 --

$$\frac{\text{Number of refusals}}{\text{Number of contacts}}.$$" p. 13.

W1 The response rate is the proportion of the eligible respondents in the sample who were successfully interviewed. For example, the denominator may be the total number of occupied dwellings, and the numerator may be the number of completed interviews. p. 294.

RESPONSE VARIANCE

B2 "That part of the response error which tends to balance out over repeated trials or over a large number of interviewers." p. 50.

B3 "The variance among the trial means over a large number of trials." p. 2.

D4 "The response variance of a survey estimator is the sum of the simple response variance and the correlated response variance."

RESPONSE VARIANCE, CORRELATED

D4 "The correlated response variance is the contribution to the total variance arising from non-zero correlations (in the sense of the distribution of measurement errors) between the response of sample units."

H1 The contribution to the total response variance from the correlations among response deviations.

RESPONSE VARIANCE, UNCORRELATED (SIMPLE)

D4 "The sample response variance contribution to the total variance arises from the variability of each survey response about

its own expected value. In terms of a simple random sampling design, the simple response variance is the population mean of the variance of each population unit."

H1 "The variance of the individual response deviations over all possible trials."

H5 "The basic trial-to-trial variability in response, averaged over the elements in the population." p. 116.

S1 "Variance of the reported value over trials."

ROTATION BIAS

O1 A type of bias that occurs in panel surveys which consist of repeated interviews on the same units. Although these surveys are designed so that the estimates of a characteristic are expected to be nearly the same for each panel in the survey, this expectation has not been realized. For example, an estimate from a panel that is in the survey for the first time may differ significantly from estimates from the panels that have been in the survey longer.

Z1 "The downward tendency in the value of the characteristics reported if the observation of the same units is continued over a longer period of time. For example, it was found in expenditure surveys that the average expenditure per item per person is usually higher in the first week of the survey than in the second or the third." p. 203.

SAMPLE DESIGN

H3 A procedure that consists of a sampling plan and method of investigation. p. 7.

K1 "The usage is not uniform as regards the precise meaning of this and similar terms like 'sample plan,' 'survey design,' 'sampling plan' or 'sampling design.' These cover one or more parts constituting the entire planning of a (sample) survey inclusive of processing, etc. The term 'sampling plan' may be restricted to mean all steps taken in selecting the sample; the term 'sample design' may cover in addition the method of estimation; and 'survey design' may cover also other aspects of the survey, e.g. choice and training of interviewers, tabulation plans, etc. 'Sample design' is sometimes used in a clearly defined sense, with reference to a given frame, as the set of rules or specifications for the drawing of a sample in an unequivocal manner."
Comment - See SURVEY DESIGN.

SAMPLE VERIFICATION

H2 A quality control procedure for keeping certain clerical errors at a satisfactory level. p. 618.

SAMPLED POPULATION

C1 "The population to be sampled." p. 6.
Comment - See FRAME and TARGET POPULATION.

SAMPLING BIAS

B2 "That part of the difference between the expected value of the sample estimator and the true value of the characteristic which results from the sampling procedure, the estimating procedure, or their combination." p. 50.

SAMPLING ERROR (OF ESTIMATOR)

B2 "That part of the error of an estimator which is due to the fact that the estimator is obtained from a sample rather than a 100 percent enumeration using the same procedures. The sampling error has an expected frequency distribution for repeated samples, and the sampling error is described by stating a multiple of the standard deviation of this distribution." p. 50.

K1 "That part of the difference between a population value and an estimator thereof, derived from a random sample, which is due to the fact that only a sample of values is observed; as distinct from errors due to imperfect selection, bias in response or estimation, errors of observation and recording, etc. The totality of sampling errors in all possible samples of the same size generates the sampling distribution of the statistic which is being used to estimate the parent value."

SAMPLING VARIANCE

D4 "The sampling variance is that contribution to the total variance arising from the random selection of a sample, rather than a complete enumeration, from the population." p. 45.

H1 The component of the total variance of the survey that represents the contribution due to sampling.

SPECIFICATION

H4 Detailed description of the collection, compilation and presentation of the survey data.
Comment - See DEFINED GOAL, EXPECTED VALUE and IDEAL GOAL.

SPECIFICATION ERRORS

M3 Errors at the planning stage because (i) data specification is inadequate and inconsistent with respect to the objectives of the survey; (ii) omission or duplication of units, incomplete units or faulty enumeration methods and (iii) inaccurate or inappropriate methods of interview. p. 451.

STANDARD ERROR OF ESTIMATE

B2 "This term refers to the sampling error, calculated as the square root of the variance of the estimator." p. 50.

K1 "An expression for the standard deviation of the observed values about a regression line, i.e. an estimator of the variation likely to be encountered in making predictions from the regression equation. For example, in simple linear regression of y on x the standard error of estimate of y is given by $\sigma_y (1 - r^2)^{\frac{1}{2}}$ where σ_y^2 is the variance of y and r is the correlation between y and x."

STATISTICAL AUDIT (CONTROL)

D2 A procedure to detect the existence of errors that are made in carrying out the fieldwork, the interviewing, the coding, the computations, and other work. p. 71.

SURVEY DESIGN

H3 "By the survey design will be meant the sample design together with the questionnaire and the method of obtaining the information from the sample, or, more

generally, the method of measurement. Thus, the survey design includes the plans for all the parts of the survey except the statement of the objectives. It includes:
(a) The questionnaire,
(b) Decision on method of observation or interview,
(c) Sample design,
(d) Choice and training of interviewers,
(e) Assignments of interviewers,
(f) Decisions on treatment of noninterviews,
(g) Estimation equations,
(h) Processing of questionnaires,
(i) Preparation of tables,
(j) Studies of precision and accuracy of information,

as well as instructions and methods followed for carrying through these operations." p. 8.
Comment - See SAMPLE DESIGN.

SURVEY VALUE
B5 "A value obtained in a complete survey which is intended to be the 'true' value, but which may not be the same because the 'true' data cannot be collected, the population cannot be defined exactly, or there are uncontrollable biases in the process of collecting and assembling the data. For example, age may be poorly reported if someone other than the person involved responds; sometimes the person himself does not know his age." p. 913.
Comment - See ACCURACY, CORRECT VALUE, and TRUE VALUE.

SYSTEMATIC ERROR
K1 "As opposed to a random error, an error which is in some sense biassed, that is to say, has a distribution with mean (or some equally acceptable measure of location) not at zero."
Comment - See BIAS.

TABULATION ERRORS
M3 Errors occurring during the tabulation stage of survey procedures. p. 451.

TARGET POPULATION
C1 "The population about which information is wanted." p. 6.
Comment - See FRAME and SAMPLED POPULATION.

TELESCOPING
R1 The tendency of the respondent to allocate an event to a period other than the reference period (also called border bias). p. 211.
S2 "A telescoping error occurs when the respondent misremembers the duration of an event. While one might imagine that errors would be randomly distributed around the true duration, the errors are primarily in the direction of remembering an event as having occurred more recently than it did. This is due to the respondent's wish to perform the task required of him. When in doubt, the respondent prefers to give too much information rather than too little." p. 69.

TEMPORARILY ABSENT
O1 A sampling unit for which a respondent cannot be contacted during the survey period.

TRUE VALUE
B5 "An idealized concept of a quantity which is to be measured; in some cases it can be achieved, but in others there is disagreement as to the definition of the quantity. Illustrations are the number of persons who are 'unemployed,' and the 'dollar value of farm sales.'
"In most surveys an approximation to the 'true' value is used, defined in such a way that one would expect to be able to measure it provided there were sufficent time, money, knowledge of techniques, etc., and no errors in the reporting, collection, and processing of the data." p. 913.
D2 A population value determined by a specified set of operations that one preferred, and the results thereof as a master standard. p. 62.
S1 That result which would be obtained with perfect measuring instruments and without committing any error of any type both in collecting the primary data and in carrying out mathematical operations.
Comment - See ACCURACY, CORRECT VALUE and SURVEY VALUE.

UNBOUNDED RECALL
N1 "Ordinary type of recall, where respondents are asked for expenditures made since a given date and no control is exercised over the possibility that respondents may erroneously shift some of their expenditures reports into or out of the recall period." [Editor's Note - Definition was given for expenditures but may apply to other characteristics.]
Comment - See BOUNDED RECALL and RECALL.

UNDERCOVERAGE
B2 "The error in an estimate that results from failure to include in the frame all units belonging to the defined population." p. 48.
O1 A type of nonsampling error that results from either failure to include all appropriate sampling units in the frame or failure to include some of the units that are already on the frame.
Comment - See COVERAGE ERROR, OVERCOVERAGE and NONCOVERAGE.

VALIDITY
D4 "A valid measure is one that measures what it claims to and not something else. Validity is a continuous concept so most measures fall between total validity and total nonvalidity. A totally valid measure is one without bias."

VARIANCE, INTERACTION TERM
D4 "The interaction contribution to the total variance of estimate is that component arising from a non-zero covariance between measurement error and sampling error."

VERIFICATION, DEPENDENT
B6 A method of verifying coding quality in which high level clerks review the work of production coders and determine whether or not the codes assigned are correct.

VERIFICATION, INDEPENDENT
B6 A method of verifying coding quality in which two or more independent codings of items are conducted for an identical sample

of persons and then the coding results are matched.

WORKING TECHNIQUE

D3 "Unfortunately, it often happens that the preferred technique, usable on a labora tory-scale, is too expensive to apply in a full-scale survey, or it may be objectionable otherwise. Experts in the subject-matter must then supply also a working technique. Thus, the preferred technique by which to define a person's age might be to compute the difference in time between today and the date shown on his birth-certificate. But some people don't have birth-certificates at all, and few people have them handy. Moreover, some people would not be happy with an interviewer who asked for birth-certificates. The Passport Division can ask for birth-certificates, but interviewers may only ask the person how old he is, and record the result. This would be the working technique by which to measure age." Comment - See PREFERRED TECHNIQUE.

Bibliography

B1. Bershad, Max A. (1964), The Index of Inconsistency for L-fold Classification System, Unpublished Bureau of the Census Memorandum, Washington, D.C.

B2. Bureau of the Census, Course on Nonsampling Errors, Lectures 1-9, International Statistics Program Center, Washington, D.C.

B3. Bureau of the Census (1963), Evaluation and Research Program of the U.S. Censuses of Population and Housing, 1960: Background, Procedures, and Forms, Series ER60, No. 1, Washington, D.C.

B4. Bureau of the Census (1968), The Current Population Survey Reinterview Program, January 1961 through December 1966, Technical Paper No. 19, Washington, D.C.

B5. Bureau of the Census (1976), The Statistical Abstract of the United States, Washington, D.C.

B6. Bureau of the Census (1965), United States Census of Population and Housing 1960: Quality Control of Preparatory Operations, Microfilming and Coding, Washington, D.C.

C1. Cochran, W. G. (1963), Sampling Techniques, J. Wiley and Sons, Inc., New York.

D1. Dalenius, Tore (1974), Ends and Means of Total Survey Design, University of Stockholm, Stockholm.

D2. Deming, W. Edwards (1960), Sample Design in Business Research, J. Wiley and Sons, Inc., New York.

D3. Deming, W. Edwards (1960), Uncertainties in Statistical Data, and Their Relation to the Design and Management of Statistical Surveys and Experiments, Bulletin of the International Statistical Institute, Tokyo.

D4. Department of Health, Education and Welfare (1975), Advances in Health Survey Research Methods, Research Proceeding Series, DHEW Publication No. (HRS) 77-3154, Washington, D.C.

H1. Hansen, M. H., Hurwitz, W. N., and Bershad, M. A. (1960), Measurement Errors in Censuses and Surveys, Bulletin of the International Statistical Institute, Tokyo.

H2. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), Sample Survey Methods and Theory, Volume I, Methods and Applications, J. Wiley and Sons, Inc., New York.

H3. Hansen, M. H., Hurwitz, W. N., and Madow, W. G. (1953), Sample Survey Methods and Theory, Volume II, Theory, J. Wiley and Sons, Inc., New York.

H4. Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1967), Standardization of Procedures for the Evaluation of Data: Measurement Errors and Statistical Standards in the Bureau of the Census, Bulletin of the International Statistical Institute, Sydney.

H5. Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1964), The Estimation and Interpretation of Gross Differences and the Simple Response Variance, Contributions to Statistics, Statistical Publishing Society, Calcutta.

H6. Hauck, Mathew and Steinkamp, Stanley (1964), Survey Reliability and Interviewer Competence, Studies in Consumer Savings, No. 4, Bureau of Economic and Business Research, University of Illinois, Urbana.

K1. Kendall, M. G. and Buckland, W. R. (1960), A Dictionary of Statistical Terms, Oliver and Boyd, Edinburgh.

K2. Kish, Leslie (1965), Survey Sampling, J. Wiley and Sons, Inc., New York.

M1. Morice, E. (1968), Dictionnaire de Statistique, Paris.

M2. Moser, C. A. and Kalton, G. (1971), Survey Methods in Social Investigation, Second Edition, Basic Books, Inc., New York.

M3. Murthy, M. (1967), Sampling Theory and Methods, Statistical Publishing Society, Calcutta.

N1. Neter, J. and Waksberg, J. (1965), Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study, Technical Paper No. 11, Bureau of the Census, Washington, D.C.

O1. OMB Subcommittee on Nonsampling Errors (1976), Washington, D.C.

P1. Pritzker, L., Ogus, J., and Hansen, M. H. (1965), Computer Editing Methods - Some Applications and Results, Bulletin of the International Statistical Institute, Belgrade.

R1. Raj, D. (1972), The Design of Sample Surveys, McGraw-Hill Book Company, New York.

S1. Sanchez-Crespo, J. L. (1975), Notes on the Accuracy, Precision and Reliability of Statistical Data, Bulletin of the International Statistical Institute, Warsaw.

S2. Sudman, Seymour and Bradburn, Norman (1974), Response Effects in Surveys, Aldine Publishing Company, Chicago.

U1. United Nations (1964), Recommendations for the Preparation of Sample Survey Reports (Provisional Issue), Series C, No. 1, Rev. 2, New York.

W1. Warwick, Donald P. and Lininger, Charles A. (1975), The Sample Survey: Theory and Practice, McGraw-Hill Book Company, New York.

Z1. Zarkovich, S.S. (1966), Quality of Statistical Data, Food and Agriculture Organization of the United Nations, Rome.

AN ANALYSIS OF COVARIANCE MODEL FOR INTERRUPTED TIME SERIES EXPERIMENTS

Bradley E. Huitema, Western Michigan University

I. Introduction

The time series quasi experiment is often a useful design in cases where randomization is impossible but data can be collected across time. If there are n_1 observation points before an intervention to the time series process and n_2 observation points after the intervention, there is generally interest in analyzing pre-post changes in the process. Among the methods of analysis available for this interrupted time series design are those suggested by Box and Tiao (1965 and 1975), Glass, Willson and Gottman (1975) and Jones, Crowell and Kapuniai (1969).

A basic problem with this design is that events concomitant with the planned intervention must be considered as alternative explanations of the change. One method of dealing with this interpretation problem is to employ one or more control series in which the intervention is not applied. If the change in the control series is not the same as the change in the basic series, evidence for the effect of the intervention is strengthened.

A method of analyzing change in the basic series which is free, in a linear sense, of the change in one or more concomitant series is described in this paper. The procedure involves

(1) regressing the basic time series on the concomitant series for the preintervention data,

(2) fitting a first order autoregressive (Markov) model to the residuals of (1) and

(3) testing differences in the postintervention phase between observed points and points predicted from information contained in (a) the concomitant series and (b) autoregression in the residuals of the fitted regression.

Certain aspects of this procedure are extensions of the Jones model.

II. The Model

The proposed model for the time series process is

$$Y_T = \mu_Y + \beta_{1\cdot 2,3,\cdots,m}(X_{1,T} - \mu_{X_1}) + \beta_{2\cdot 1,3,\cdots,m}(X_{2,T} - \mu_{X_2})\cdots + \beta_{m\cdot 1,2,\cdots,m-1}(X_{m,T} - \mu_{X_m}) + \alpha\left[Y_{T-1} - (Y|X_1, X_2, \cdots, X_m)_{T-1}\right] + \varepsilon_T$$

where

Y_T is the dependent variable score at time T which is any of the equally spaced observation points,

μ_Y is the process mean for the basic (dependent variable) series,

μ_{X_1} through μ_{X_m}, are the means of the concomitant series (i.e., covariates) one through m,

$\beta_{1\cdot 2,3\cdots m}$ through $\beta_{m\cdot 1,2,\cdots,m-1}$ are the partial regression coefficients obtained from regressing Y_T on covariates $X_{1,T}, X_{2,T}, \cdots, X_{m,T}$

$X_{1,T}, X_{2,T}, \cdots, X_{m,T}$ are scores on covariates $1,2,\cdots,m$ measured at time T. These scores are obtained from any available set of concomitant time series and may be in the form of continuous scores or dummy values which indicate the presence or absence of a condition.

α is the first order autoregression parameter relating the residuals $\left[Y_{T-1} - (Y|X_1, Y_2, \cdots, X_m)_{T-1}\right]$ and $\left[Y_T - (Y|X_1, X_2, \cdots X_m)_T\right]$ and

ε_T is the error which is NID(o,σ^2).

III. Estimation

The parameters of the model are estimated as follows:

μ_Y is estimated using

$$\hat{\mu} = \frac{1}{n}\sum_{T=1}^{n_1} Y_T$$

where n_1 is the number of preintervention observations,

μ_{X_1} through μ_{X_m} are estimated using

$$\hat{\mu}_{X_i} = \frac{1}{n_1}\sum_{T=1}^{n_1} X_i,$$

the partial regression coefficients are estimated using ordinary least squares on the n_1 preintervention points and the first order autoregression parameter is estimated using

$$\hat{\alpha} = \frac{SP_1}{SS_c} = \frac{\sum_{T=2}^{n_1}\left[Y_{T-1} - (Y|X_1, X_2, \cdots, X_m)_{T-1}\right]\left[Y_T - (Y|X_1, X_2, \cdots X_m)_T\right]}{\left(\frac{n_1-1}{n_1}\right)\sum_{T=1}^{n_1}\left[Y_T - (Y|X_1, X_2, \cdots X_m)_T\right]^2}$$

where SP_1 is the lag one sum of products of the residuals of the regression and SS_c is the zero lag sum of squares corrected by a term that allows for the difference between the number of observations that are associated with the sum of products and sum of squares.

IV. Testing for Intervention Effects

Two related tests are suggested for testing for change in the time series following the intervention.

A. Testing for Postintervention Change at Individual Postintervention Points

The test statistic for evaluating the change in the time series at a post intervention point specified *a priori* is

$$t = \frac{Y_T - \hat{Y}_T}{\sqrt{\left[\frac{SS_c\ (1-R^2 - \hat{\alpha}(1-R^2))}{n_1-m-3}\right]\left[1 + \underline{c}_T'(\underline{X}'\underline{X})^{-1}\underline{c}_T\right]}}$$

where

$\hat{Y}$ is the predicted postintervention value based on the fitted model,

R^2 is the coefficient of multiple determination based on the fitted model,

$\underline{c}_T$ is the unity augmented column vector of covariate scores measured at time T, i.e.,

$$\underline{c}_T = \begin{bmatrix} 1 \\ X_{1,T} \\ X_{2,T} \\ \cdot \\ \cdot \\ \cdot \\ X_{m,T} \end{bmatrix},$$

$\underline{X}$ is the unity augmented covariate score matrix based on the n_1 preintervention data points, i.e.,

$$\underline{X} = \begin{bmatrix} 1 & X_{1,1} & X_{2,1} & \cdot & \cdot & \cdot & X_{m,1} \\ 1 & X_{1,2} & X_{2,2} & & & & X_{m,2} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & X_{1,n_1} & X_{2,n_1} & \cdot & \cdot & \cdot & X_{m,n_1} \end{bmatrix}$$

The test statistic t is compared with the critical value of the conventional t statistic based on n_1-m-3 degrees of freedom.

B. Testing for Overall Change in the Whole Postintervention Series

If interest lies in evaluating the intervention points, the following approximate test is suggested

$$z = \frac{\sum_{n_1+1}^{n_1+n_2} t_T}{\sqrt{n_2\left[(n_1-m-3)/(n_1-m-5)\right]}}.$$

If the first order autoregressive model fits the residuals of the regression, the individual t tests will be approximately independent and the test statistic z will be approximately a standard normal variable.

V. Example

Drunkenness arrest data from two Michigan counties are plotted in Figures A and B. A program that was expected to have an effect on the arrests in the first county (Kalamazoo) is the intervention that occurs after week 39. A comparison of the data from the experimental county with the covariate data from the control county (Calhoun) which was not exposed to the program, reveals a somewhat disturbing pattern. The arrests appear to drop for both the experimental and control counties. In order to evaluate whether or not the postintervention change is significant for the experimental county after controlling for change in the control county, we apply the tests of Section IV.

Individual tests:

Week	t
40	-1.48
41	-1.41
42	- .63
43	-1.45
44	-1.11
45	-1.74
46	- .63
47	-1.25

Overall test:

$$z = \frac{-9.70}{\sqrt{8(35/33)}} = -3.33.$$

None of the individual t tests are significant using $\alpha = .05$, but the observed values associated with these tests are all less than the predicted values. This is indicated by the negative signs associated with the t values. As would be expected, when the combined information from these individual t values is employed in the overall test, the conclusion is that significant postintervention change, beyond that which is found in the control county, took place in the experimental county.

References

Box, G. E. P., and Tiao, G. C. A change in level of nonstationary time series. Biometrika, 1965, 52, 181-192.

Box, G. E. P., and Tiao, G. C. Intervention analysis with applications to economic and environmental problems. Journal of the American Statistical Association, 1975, 70, 70-79.

Glass, G. V., Willson, V. L., and Gottman, J. M. Design and analysis of time-series experiments. Boulder: Colorado Associated University Press, 1975.

Jones, R. H., Crowell, D. H., and Kapuniai, L. E. Change detection model for serially correlated data. Psychological Bulletin, 1969, 71, 352-358.

Figure A: Kalamazoo County Weekly Total Arrests For Drunkenness

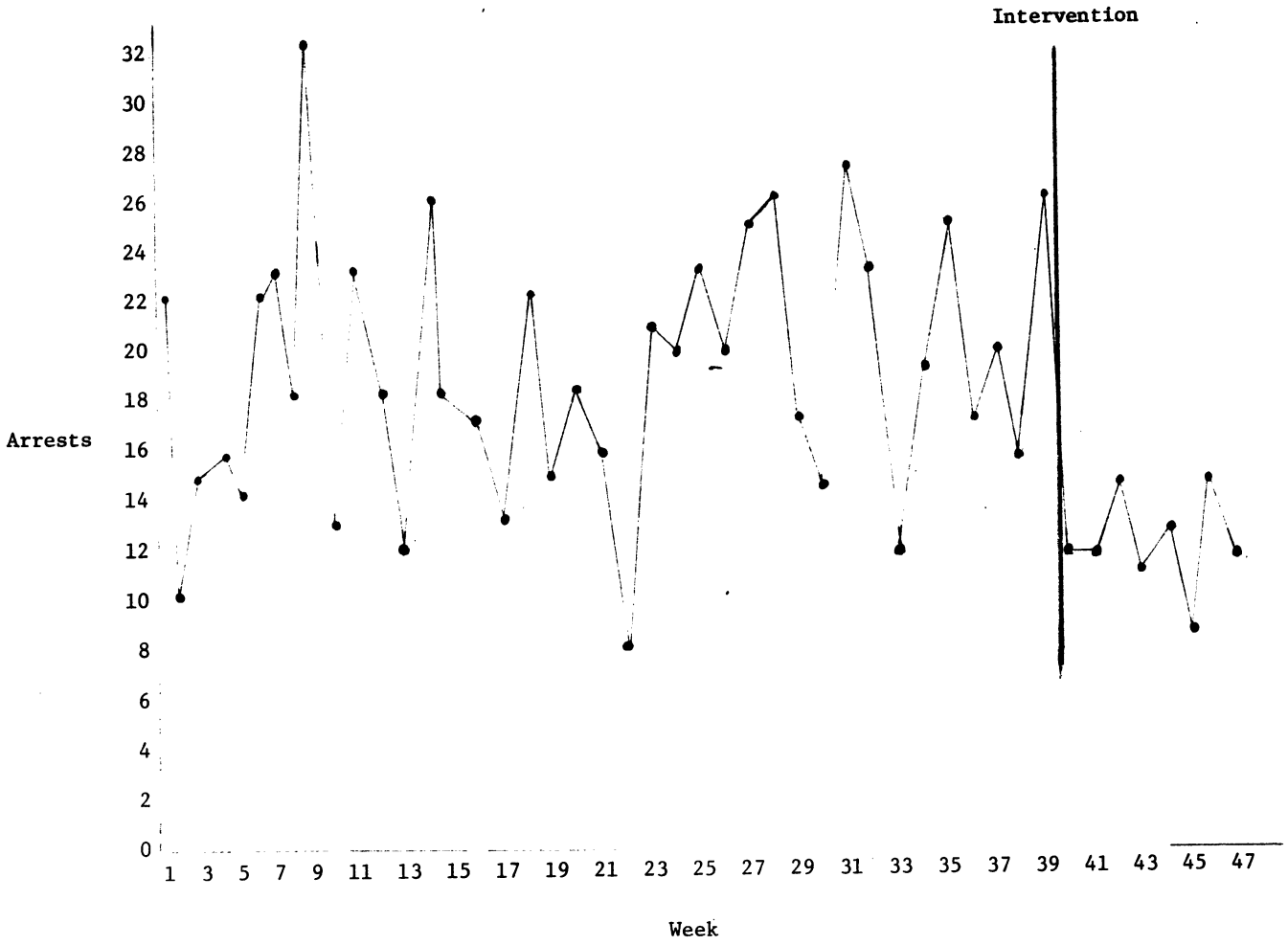
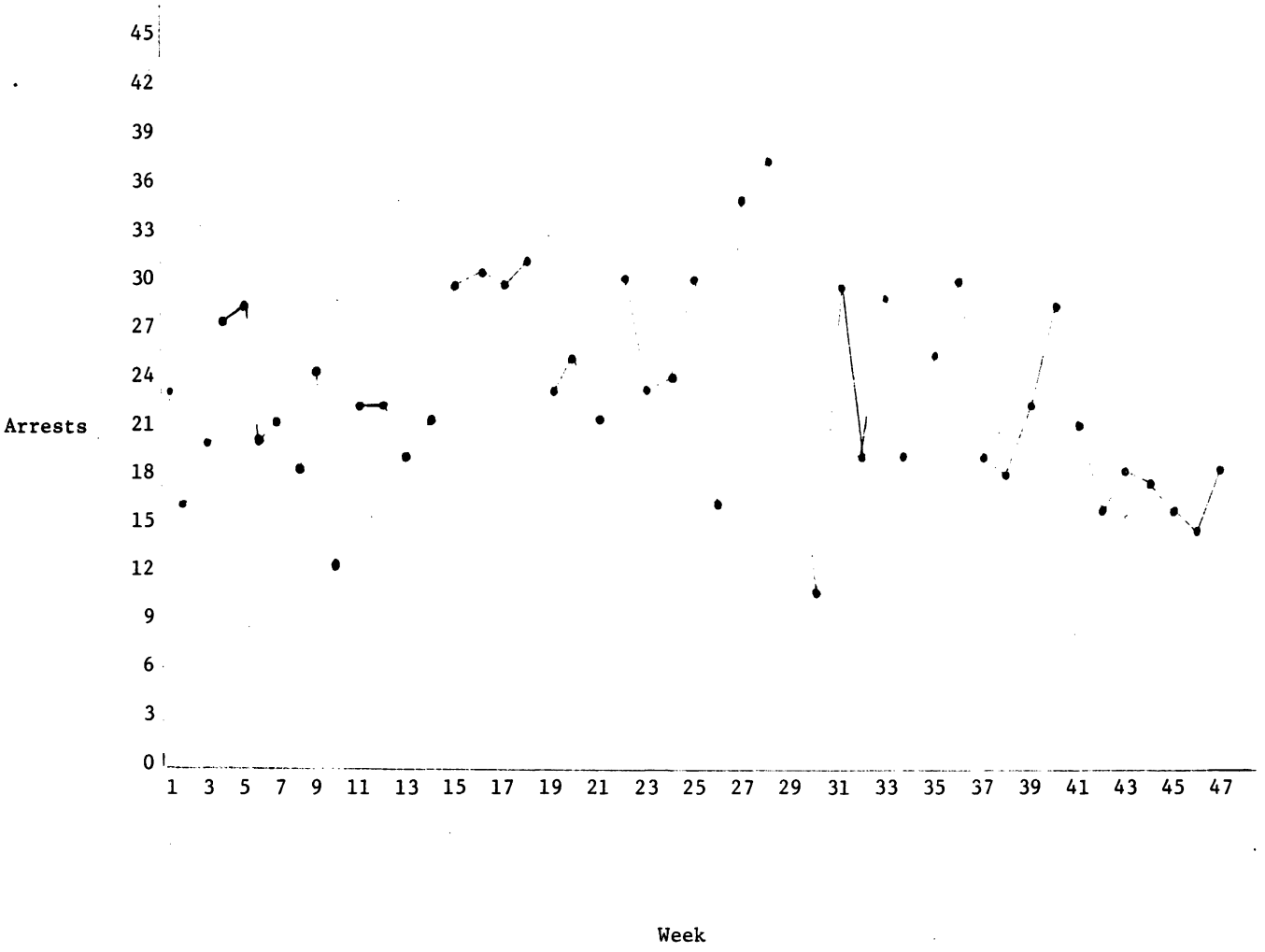


Figure B: Calhoun County Weekly Total Arrests For Drunkenness



ON THE DEVELOPMENT OF A DIVORCE INTENSITY FUNCTION

John B. Keats, Louisiana Tech University
John E. Cole, Western Electric Company

The nature of divorce as a function of marriage duration is virtually unassessed. The purpose of this paper is to examine divorce statistics in much the same way that one might look at mortality. Just as the intensity of mortality is varying at each moment of age, the intensity of divorce is varying at each moment of marriage. Therefore, part of this paper deals with measuring this instantaneous variation. After these estimates of divorce intensity were found for each interval of marriage, a smooth curve was fitted. This curve, a hazard function for divorce, was then used to attain conditional probabilities of divorce for duration of marriage intervals. This paper is believed to be the first attempt to develop and use a divorce intensity function. Nineteen Seventy-One is used as the base year for the study as it was the most recent year for which the necessary vital statistics data were available.

As the method of the present study is similar to a method used to obtain mid-interval estimates of the "force of mortality", this method will be reviewed. A typical mid-interval estimate of the mortality intensity or "force of mortality", μ_i, for a population of N_i people of exact age x_i subject to death in the interval $(x_i, x_i + n_i)$ for a given year is given by:

$$M_i = D_i/\{n_i(N_i - D_i) + a_i n_i D_i\} \qquad \{1\}$$

where M_i is the age specific death rate, D_i is the number of deaths in the interval, n_i is the interval width, and a_i is the fraction of the interval $(x_i, x_i + n_i)$ before death occurs. The denominator of {1} is usually estimated by the midyear population, P_i, which is obtained from the Bureau of the Census (2). Thus, the age-specific death rate is the ratio of the deaths in an interval to the average number of individuals exposed to the risk of death.

A mid-interval estimate of divorce intensity, directly analogous to the age-specific death rate was developed for a population of N_i couples whose marriages endured x_i years. These couples are subject to divorce in the interval $(x_i, x_i + n_i)$, where x_i is the number of years married and n_i is the width of the ith interval.

The estimate is given by:

$$D_i = T_i \,/\, \{.5(N_i' + N_i' - T_i - L_i)\} \qquad \{2\}$$

where D_i is the duration of marriage-specific divorce rate, T_i is the number of divorces in 1971 among couples married $x_i - x_{i+1}$ years as of 1971, L_i is the number of marriages $x_i - x_{i+1}$ years prior to 1971 ended by death of a spouse during 1971, N_i' is $N_i - E_i - M_i$ which is the number of couples married $x_i - x_{i+1}$ years prior to 1971, exposed to the risk of divorce at the beginning of 1971, N_i is the number of couples married $x_i - x_{i+1}$ years prior to 1971, E_i is the number of marriages $x_i - x_{i+1}$ years prior to 1971 ended by divorce prior to 1971 and M_i is the number of marriages $x_i - x_{i+1}$ years prior to 1971 ended by death of a spouse prior to 1971. Since $N_i' - T_i - L_i$ represents the number of cohort couples still married by the end of 1971, D_i is the ratio of divorces in a duration of marriage interval to the average number of couples exposed to the risk of divorce. Thus, for each duration of marriage interval, $D_i = \hat{\delta}_i$. $\hat{\delta}_i$ represents the divorce intensity value after $x_i + n_i/2$ years of marriage.

T_i was obtained by multiplying the number of divorces granted in 1971, 764,000, by percentages given in Table 2-4 (3). These percentages were based on the divorce registration area which consisted of samples taken in 29 states. N_i represents the number of marriages $x_i - x_{i+1}$ years prior to 1971, for i = 1,2, ..., 10, x_i and x_{i+1} differ by one year, and for i = 11, 12, ..., 14, x_i and x_{i+1} differ by five years. Values of N_i, i = 2,3, ..., 10, were developed by considering the 12 months of 1971 in which a divorce could have occurred. Examination of recent data indicated that the number of divorces in each month is rather constant and consequently divorces in each of the 12 months were considered equally likely. A couple divorced in any month in 1971 after x_i years of marriage could have been married x_i years and zero months to x_i years and 11 months prior to the month of their divorce. Thus, considering these time intervals for each of the 12 months, there were nine 23 month intervals $x_i - x_{i+1}$ years prior to 1971 over which couples exposed to the risk of divorce in 1971 could have been married. For example, couples married 3-4 years in 1971 could have been married in any of the months of the interval (February, 1967, December 1968). A weighting scheme for each of the 23 months was developed based on the number of times each month was a possible marriage month. N_2 through N_{10} were determined using the weights with marriage data for each of the 23 months. The weights for the 23 months sum to 12. 1970 and 1971 marriages were averaged to obtain N_1, marriages 0-1 years prior to 1971.

N_{11}, N_{12}, N_{13}, and N_{14} represent marriages over a five year period. There are 71 possible marriage months involved for each of these N_i values. For example, the months for N_{11} (10 years 0 months - 14 years, 11 months of marriage) are February 1956-December 1961. The numerators of the weights ranged from one to 12 and each denominator was 12, so that the sum of the weights was 60. No marriage by month data was available prior to 1949. For months in years prior to 1949, the 1949 data was used. For N_{15}, the number of couples married 30-45 years prior to 1971, the marriages in the years 1926-1940 were added. Examination of data at the time where divorces were recorded in the interval 40-45 years of marriage revealed significantly large enough numbers to allow for the possibility of divorce up to 45 years of marriage. The number of divorces beyond 45 years was considered negligible.

E_i, divorces prior to 1971 among couples married $x_i - x_{i+1}$ years as of 1971 was obtained by applying percent divorces by duration of marriage in the registration area from 1926-1970 to the total number of divorces for each of these years. It was assumed that a divorce within the first 6 months of marriage was impossible. To demonstrate a typical calculation, E_4 is used. Couples married 3-4 years prior to 1971 were married in either 1967 or 1968. These couples

could have been divorced in 1967 or 1968 before their marriage had endured one year. There were 30,334 divorces in 1967 and 29,764 divorces in 1968 among couples married less than one year. A weighting scheme, too detailed to describe here, yielded 29,899 as the appropriate figure for the divorces within one year of marriage. There were 54,954 divorces in 1969 among couples married 1-2 years. There were 66,552 divorces in 1970 among couples married 2-3 years. Therefore, E_4 =29,899 + 54,954 + 66,552 = 151,405.

To obtain M_i values, a force of mortality function of the form $\mu_x=ax^b\exp(cx)$ developed by Keats and Como (1) and based on 1970 data was used to obtain $q(t_1,t_2) = 1-\exp(-\int_0^{t_2} {}_x dx)/\exp(-\int_0^{t_1} \mu_x dx)$. {3}

$q(t_1,t_2)$ represents the probability of death in an interval given survival prior to the interval. For each i, the N_i value was used with Table 1 to determine the approximate number of brides and grooms in each age category.

Table 1: Age at time of marriage based on averages of years 1962-1971. From (3):

Age	Percent Brides	Percent Grooms
15 - 20	36	14
20 - 25	36	46
25 - 30	10	16
30 - 35	5	7
35 - 45	6	8
45 - 65	6	7
≤ 65	1	2

{3} was then used with these values on a year to year basis from the year of marriage until 1971. {3} was first applied to the males for a one year period. This yielded the number of married males in each of seven categories dying within one year of marriage. Table 2 was used to identify and remove from the female population, the resulting widows in each age category. {3} was then applied to the females for a one year period and the number deceased in each age category was identified. Use of Table 2 then removed the resulting widowers from each age category. One year was added to the ages of the survivors in each age category and the process was repeated through 1970. The corresponding figure for 1971, L_i, was developed by extending the procedure one additional year.

Table 3 presents values of the statistics T_i, N_i, E_i, M_i, N_i', L_i, and D_i for fifteen marriage duration intervals. The D_i values are mid-interval estimates of δ_i, the divorce intensity value. These fifteen D_i values were plotted against duration of marriage mid-intervals ($x_i + n_i/2$) and an attempt was made to fit them with a continuous curve. Several functions were applied to these fifteen points, the best of which was the form

$$\delta_x = ax\exp(b\sqrt{x}). \qquad \{4\}$$

a and b were estimated by least squares methodology after applying the natural logarithm to both sides of the equation. The resulting values were: a = .08722864 and b = 1.17031155. The curve of {4} was an excellent fit to the data as R^2 =.9433. Figure 1 presents this curve and the 15 mid-interval estimates of the divorce intensity.

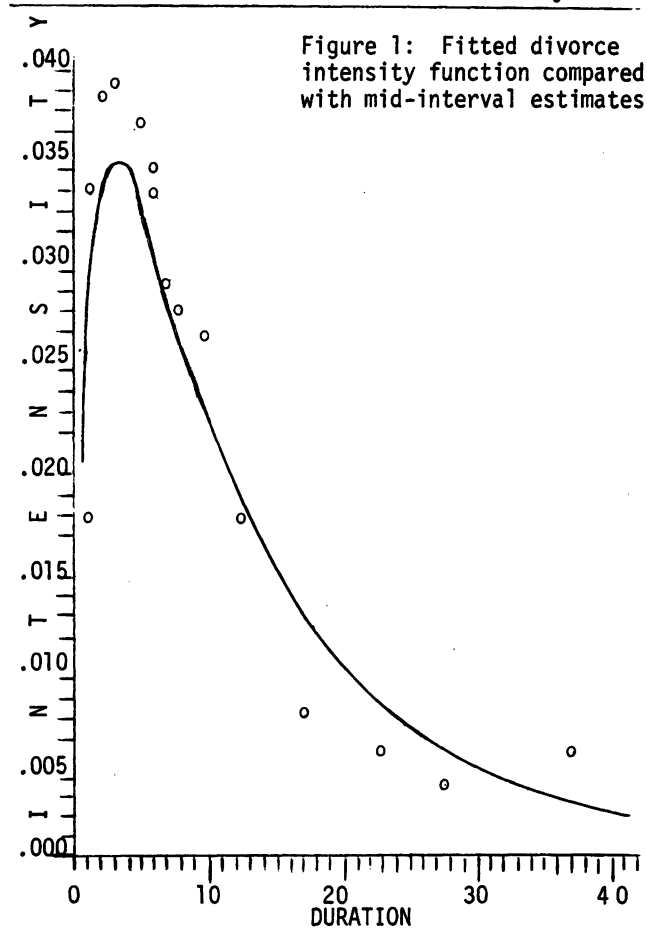


Figure 1: Fitted divorce intensity function compared with mid-interval estimates.

Table 2: From (3), 1971:

Groom's/Bride's Age at Time of Marriage	Percent Brides (Grooms) in Each Age Group						
	15-19	20-24	25-29	30-34	35-44	45-64	≥65
15-19	86.46(37.47)	12.72(54.38)	.60(6.48)	.15(1.07)	.07(.48)	0(.11)	0(.01)
20-24	39.41(4.67)	55.21(64.53)	4.34(22.98)	.71(5.15)	.29(2.24)	.04(.43)	0(0)
25-29	12.80(.81)	53.59(18.73)	25.51(40.37)	5.82(21.93)	2.06(14.84)	.22(3.22)	0(.10)
30-34	5.15(.44)	29.28(6.52)	33.79(19.65)	20.24(28.07)	10.23(33.76)	1.31(11.28)	0(.28)
35-44	2.12(.19)	11.68(2.25)	20.97(5.83)	22.32(11.85)	34.43(43.50)	8.41(34.94)	.07(1.44)
45-64	.50(0)	2.30(.28)	4.67(.64)	7.64(1.56)	28.38(11.01)	54.38(69.26)	2.13(17.18)
≥65	.19(0)	.14(0)	.62(0)	.75(0)	4.66(.56)	52.60(16.84)	41.04(82.66)

Table 3:

i	x_i	x_{i+1}	T_i	N_i	E_i	M_i	N_i'	L_i	D_i
1	0	1	35,144	2,174,642	7,080	6,875	2,162,035	14,314	.01644
2	1	2	67,232	2,162,998	33,826	14,238	2,114,934	15,062	.03242
3	2	3	72,580	2,112,331	92,691	28,615	1,991,025	15,563	.03729
4	3	4	67,232	2,006,415	151,405	41,963	1,813,047	15,636	.03795
5	4	5	57,300	1,887,581	199,612	54,188	1,633,781	15,550	.03588
6	5	6	48,896	1,823,164	238,147	67,358	1,517,659	15,861	.03293
7	6	7	45,076	1,755,331	259,827	80,123	1,415,381	16,118	.03256
8	7	8	36,672	1,712,248	277,431	93,878	1,340,939	16,579	.02791
9	8	9	31,324	1,609,081	282,756	103,802	1,222,523	16,408	.02614
10	9	10	28,268	1,552,582	290,980	115,990	1,145,612	16,660	.02517
11	10	15	103,140	7,556,332	447,585	818,250	6,290,497	90,905	.01665
12	15	20	68,760	7,660,747	1,091,589	1,300,757	5,268,401	127,322	.00665
13	20	25	53,480	8,979,748	1,426,867	2,187,817	5,365,064	133,026	.00507
14	25	30	27,504	8,295,415	1,879,514	2,680,387	3,735,514	156,015	.00377
15	30	45	19,864	18,865,575	4,678,455	10,064,267	4,122,853	544,484	.00517
			762,472	70,154,190	11,357,765	17,658,508	41,139,265	1,209,503	

Let $q(x_i, x_j) = \Pr\{\text{divorce } (x_i, x_j) \mid \text{no divorce } (.5, x_i)\} = 1 - \exp\left(-\int_{.5}^{x_j} \delta_x dx\right) / \exp\left(-\int_{.5}^{x_i} \delta_x dx\right)$ {5}

Since the integration required in {5} cannot be performed directly, the infinite series representation of e^x was employed; thus:

$$\exp\left(-\int_{.5}^{x_m} ax \exp(b\sqrt{x})dx\right) = \exp\left\{-a\sum_{K=0}^{\infty} \frac{b^k(x_m^{K/2+2} - .5^{K/2+2})}{K!\,(K/2+2)}\right\} \quad \{6\}$$

A computer program revealed that for each of the 15 x_i values, the sum in {6} converged after 50 terms ($e < 1 \times 10^{-11}$). This program also evaluated {5} for the 15 intervals shown in Table 3. The observed $q(x_i, x_j)$ values of Table 4 below were obtained using

$$q(x_i, x_j) = T_i/(N_i'' - E_i' - M_i') \quad \{7\}$$

where T_i is the number of divorces in 1971 among couples married $x_i - x_i$ years as of 1971, N_i'' is the number of couples married x_i years prior to 1971, E_i' is the number of marriages x_i years prior to 1971 ended by divorce prior to 1971, and M_i' is the number of marriages x_i years prior to 1971 ended by death of a spouse prior to 1971.

Table 4 provides a measure of the efficiency of equation {5} as a predictor of divorce. It is seen that the values obtained using equation {5} ($q(x_i, x_j)$ predicted) differ only slightly from the values obtained from {7} ($q(x_i, x_j)$ observed) with the exception of q(20,25). Equation {5} and the methodology of {6} may be used to obtain conditional probabilities of divorce not only for the x_i and x_j values of Table 4, but for intervals of any survival point and length.

Table 4: Predicted and observed values of $q(x_i, x_j)$ for selected intervals.

Interval x_i	Interval x_j	$q(x_i, x_j)$ Predicted	$q(x_i, x_j)$ Observed
0.5	1.0	0.011691	0.016255
1.0	2.0	0.030405	0.031789
2.0	3.0	0.033575	0.036454
4.0	5.0	0.032233	0.035072
6.0	7.0	0.028284	0.031847
8.0	9.0	0.024157	0.025622
10.0	15.0	0.083955	0.092281
15.0	20.0	0.055905	0.068918
20.0	25.0	0.037651	0.061111
25.0	30.0	0.025754	0.026938
30.0	45.0	0.039021	0.032978

The validity of the statistics used in this study may be examined by calculating the divorce rate per 1,000 married women from the figures of Table 3 and comparing it with the 1971 United States Office of Vital Statistics published figure. From Table 3, the number of divorces per 1,000 married women is 764,000/41,139.265 = 18 which is reasonably close to the published figure of 16.

Table 3 may also be used to estimate the percentage of marriages ending in divorce. There were 18,865,575 couples married 30-45 years prior to 1971. Among these couples, there were 4,678,455 divorces prior to 1971 plus an additional 19,864 during 1971. Although divorce data beyond 1971 was unavailable, we may project to future years assuming the 1971 number of divorces (19,864) among the 30-45 years married group will be constant throughout the period 1972-1985. The study period ends in 1985 as the couples married 30 years in 1971 approach 45 years of marriage in 1985. Furthermore, for each ensuing year, one fewer duration of marriage age is to be

tallied in projecting future divorces, i.e., each year removes another group from the count, since this group has exceeded the 45 year marriage duration, and by assumption, no divorces are possible beyond this period. Assuming a uniform distribution of divorce throughout the 15 year interval, 14/15 of 19,864 couples were divorced in 1972, 13/15 of 19,864 in 1973, ..., 1/15 of 19,864 will be divorced in 1985. In this manner, projected divorces totaled 139,048. Adding the 4,678,455 divorces prior to 1971, and the 19,864 in 1971, a total of 4,837,367 divorces have resulted. Dividing by 18,865,575, it is estimated that 25.64% of all marriages end in divorce.

Although not critical to the aforementioned results, there were a number of minor assumptions which were obviously incorrect. The divorce data was obtained assuming that for the entire divorce population, the percent divorces in any interval was the same as for the divorce registration area (sample). The Keats-Como mortality function was general (it included both men and women) and was based on the year 1970. The number of couples whose marriage ended by death would have been more accurate if separate male and female mortality functions had been used and if they were changed for each year in the study (1926-1971). The Groom-Bride and Bride-Groom age at marriage matrices (Table 2) were applied over the 45 year span of the study, yet these figures were based on 1971. For more accuracy, different matrices should have been used each year.

REFERENCES

(1) Keats, J.B., and Como, B., On Fitting a Force of Mortality Function, Paper presented at the annual meeting of the Louisiana Academy of Sciences, Southern University, Baton Rouge, Louisiana, (February 1976).

(2) United States Department of Commerce, Bureau of the Census, Current Population Reports, Series P-25.

(3) United States Office of Vital Statistics, Vital Statistics of the United States, Vol. III, Government Printing Office, Washington, D. C.

THE PROBABILITY OF ONE SPOUSE OUTLIVING THE OTHER BY SEX AND AGE

S. Mitra, Emory University

1. The Problem

The probability measure of the event that one of the spouses will survive a given interval of time while the other will not, depends (a) on their ages at the beginning of the interval as well as (b) on the associated probabilities of surviving and dying in the interval by the respective spouses. The problem is a particular case of the more general example dealing with multiple decrement tables (Jordan, 1967). However, this special case has not received much attention since Depoid (1938) studied the probabilities of a marriage being terminated by the death of the husband or of the wife after a given number of years of marriage. He also mentioned about the eventual probabilities of becoming a widow or a widower.

It may be mentioned that the eventual probability of becoming a widow or a widower can indeed be regarded as an asymptote that is reached as the interval over which such a probability is calculated, is gradually enlarged. In this paper, however, an attempt has been made to directly derive the eventual probabilities of any one of the spouses outliving the other for specific age combinations, that apply either at the time of marriage or at any time thereafter. Needless to say, the dissolution of marriage through separation or divorce is not relevant for the problem defined in this manner.

2. Derivation of the Probability Function

For reasons of operational simplicity in the derivation of these results, the survivorship probabilities of the two sexes are regarded as independent of one another. These probabilities are usually obtained from the respective life tables, so that mortality differentials by marital statuses, if any, are also ignored. Under these conditions, the probability that both of the spouses will survive a given interval of time can be obtained as the product of the survivorship probabilities of the two sexes corresponding to their respective ages and the length of the interval. The probabilities of one or both of them dying in that interval can also be easily obtained.

Relaxing the restriction of a specific interval of time, let the probability that a husband a years old will outlive his wife aged b, be denoted by P(a,b). Disregarding the possibility of their dying at the same instant of time, the complementary probability, namely

$$Q(a,b) = 1 - P(a,b) \qquad (1)$$

provides the probability measure of the same wife's outliving her husband. The probability of their jointly surviving a period of x years can be expressed as

$$C(a+x,b+x) = \frac{\ell_m(a+x)\ell_f(b+x)}{\ell_m(a)\ell_f(b)} \qquad (2)$$

in which $\ell_m(a+x)$ and $\ell_f(b+x)$ are the male and the female probabilities of survival from birth to ages a+x and b+x respectively which coincide with the end of the time interval while $\ell_m(a)$ and $\ell_f(b)$ are the corresponding probabilities at the beginning of the interval.

Next, the probability that the female spouse will die at age b+x, leaving her male partner a widower, can be written as

$$C(a+x,b+x)\mu_f(b+x)dx \qquad (3)$$

in which $\mu_f(b+x)$ is the force of mortality at age b+x for the females, that is,

$$\mu_f(b+x) = -\frac{d}{dx}\ell_f(b+x)/\ell_f(b+x) \qquad (4)$$

Therefore, the eventual probability of the male's outliving the female spouse can be obtained from

$$P(a,b) = \int_0^{\alpha(a,b)} C(a+x)\mu_f(b+x)dx \qquad (5)$$

where $\alpha(a,b)$, the upper limit of the integral, depends on the values of a,b as well as on the life spans of the two sexes. For all practical purposes, however, the values of $\alpha(a,b)$ may be left unspecified.

Because of (2), P(a,b) can be alternatively expressed as

$$P(a,b) = \frac{\int_0^{\alpha(a,b)} \ell_m(a+x)\,\ell_f(b+x)\mu_f(b+x)dx}{\ell_m(a)\ell_f(b)} \qquad (6)$$

and the complementary probability as

$$Q(a,b) = \frac{\int_0^{\alpha(a,b)} \ell_m(a+x)\ell_f(b+x)\mu_m(a+x)dx}{\ell_m(a)\ell_f(b)} \qquad (7)$$

It is easy to see that (6) and (7) satisfy (1) as they should. This is because

$$-\frac{d}{dx}[\ell_m(a+x)\ell_f(b+x)] =$$

$$\ell_m(a+x)\ell_f(b+x)[\mu_m(a+x) + \mu_f(b+x)] \qquad (8)$$

and thus the sum of the integrals in the numerators of (6) and (7) simplifies to $\ell_m(a)\,\ell_f(b)$, their common denominator.

Another expression for P(a,b) may be derived by noting the equality

$$P(a,b) = \frac{P(a,b)}{P(a,b)+Q(a,b)} \qquad (9)$$

so that, a combination of (6) and (7) results in the expression

$$P(a,b) = \frac{\int_0^{\alpha(a,b)} \ell_m(a+x)\,\ell_f(b+x)\,\mu_f(b+x)\,dx}{\int_0^{\alpha(a,b)} \ell_m(a+x)\,\ell_f(b+x)\,[\mu_m(a+x)+\mu_f(b+x)]\,dx} \quad (10)$$

3. Approximate Algebraic Solution of P(a,b)

What follows next is the description of a method suggested for the reduction of the ratio of the integrals in (10). First, it is acknowledged that, in general, the force of mortality can be regarded as a reasonably smooth and a monotonically increasing function of age in the age interval that excludes the childhood years. Also known is the fact that in such an interval, the function can very well be approximated by the Gompertz curve, namely,

$$\mu(x) = BC^x \quad (11)$$

The limitations of the model, found for the most part in the old ages, are known to be true primarily with respect to mortality experiences observed during an interval of time, rather than those that are applicable to generation mortality (Spiegelman, 1969). Consequently, the model can be expected to provide a fair approximation of the mortality experiences, especially in those countries, where the patterns of mortality exhibit little or only minor changes over time.

It is also known that the force of mortality is not affected much by variations in C at the younger ages, and therefore it is generally approximated by an average of its values observed at higher ages. Table 1 shows the values of C for the two sexes, obtained from successive five year age intervals (n=5) beginning from age 30, for the 1973 U.S. Life Tables (Vital Statistics of the U.S. 1973, DHEW), as

$$C^n = \frac{\int_0^n \mu(y+n+x)\,dx}{\int_0^n \mu(y+x)\,dx} = \frac{\ln[\ell(y+n)/\ell(y+2n)]}{\ln[\ell(y)/\ell(y+n)]} \quad (12)$$

Interestingly enough, the parameter C shows only minor variations by age and sex. This is quite logical in the sense that, under normal conditions, the patterns of mortality of the two sexes cannot be unrelated with one another. For all practical purposes therefore, C can be assumed as constant for the two sexes. In that event, it is possible to write

$$\mu_m(a+x) = K(a,b)\mu_f(b+x) \quad (13)$$

where

$$K(a,b) = (B_m/B_f)C^{a-b} \quad (14)$$

in which B_m and B_f are the values of the parameter B in (11) for the males and the females respectively. Substituting (13) in (10) and simplifying, the equation

$$P(a,b) = \frac{1}{1+K(a,b)} = \frac{1}{1+(B_m/B_f)C^{a-b}} \quad (15)$$

is obtained. It is of considerable interest to note from (15) that, in a given population, the probability of losing a spouse depends primarily on the age difference of the two spouses, but not on their specific ages. From a mathematical point of view, this surprising finding follows from the reasonable assumption of the Gompertz law of mortality with an additional (but reasonable nonetheless) restriction that for any age the forces of mortality for the two sexes remain proportional to one another. It is easy to see that minor deviations from these assumptions will not drastically affect the aforementioned results.

Also of interest to note is that the conclusions drawn about the behavior of the eventual probability measures apply also to the same calculated over a shorter time interval. In other words, the probability, say, $P_t(a,b)$, that the wife will be the first to die within the next t years, conditional to at least one of them dying in that interval, will be the same as shown in (15). This is so because changing the upper limits of the integrals in (10) from $\alpha(a,b)$ to some t, has no effect on its value when Gompertz law of mortality is assumed. The unconditional probabilities (Depoid, 1938), as mentioned earlier, will increase with t and approach the eventual probability as the limiting value.

4. Empirical Estimates of K(a,b)

As shown in (15) K(a,b) consists of two factors, namely, B_m/B_f and C^{a-b}, in which the former can be expressed as the ratio of the forces of mortality of the corresponding sexes when both spouses are of the same age. That is to say,

$$\frac{B_m}{B_f} = \frac{\mu_m(x)}{\mu_f(x)} \quad (16)$$

Since, in practice, the ratio will show some variation by age, an estimate of the same can be obtained as

$$\frac{B_m}{B_f} = \frac{\int \mu_m(x)\,dx}{\int \mu_f(x)\,dx} \quad (17)$$

where the limits of the integrals may be set at convenience. From practical considerations, the lower limit of the integral may be set at age 30 whereas the upper limit may be determined by the lower boundary, say α, of the last interval (open ended) of the life tables. In that case (17) can be simplified as

$$\frac{B_m}{B_f} = \frac{\ln[\ell_m(30)/\ell_m(\alpha)]}{\ln[\ell_f(30)/\ell_f(\alpha)]} \quad (18)$$

Similarly, the parameter C which can be expressed either as

$$C^n = \frac{\mu_m(x+n)}{\mu_m(x)} \quad (19)$$

or as

$$C^n = \frac{\mu_f(x+n)}{\mu_f(x)} \tag{20}$$

will also show some variation by age and sex. Accordingly, an estimate of C may be generated from

$$C^n = \frac{1}{2}\left[\frac{\ell n[\ell_m(30+n)/\ell_m(\alpha)]}{\ell n[\ell_m(30)/\ell_m(\alpha-n)]} + \frac{\ell n[\ell_f(30+n)/\ell_f(\alpha)]}{\ell n[\ell_f(30)/\ell_f(\alpha-n)]}\right] \tag{21}$$

The values of K(a,b) for different combinations of a and b can then be obtained from (14) through appropriate substitutions. These, for the 1973 U.S. Life Tables (n=5 and α=85), are shown in Table 2.

5. Solution of P(a,b) by Numerical Methods

The integrals appearing in (10) can of course be evaluated by numerical methods which, as such, will be free from assumptions about the forces of mortality made earlier. First, it may be noted that from the definition of the force of mortality given in (4), it is possible to write

$$\ell_f(b+x)\mu_f(b+x) = -d\ell_f(b+x) \tag{22}$$

so that the numerator of (10) can be rewritten as

$$-\int_0^{\alpha(a,b)} \ell_m(a+x)\,d\ell_f(b+x) \tag{23}$$

In general, the derivative of $\ell(x)$ can be assumed as constant over a small age interval and therefore, for such an interval of length n (usually no greater than 5) years,

$$-\frac{d}{dx}\ell_f(b+x) = \frac{\ell_f(b+x) - \ell_f(b+x+n)}{n} = \frac{{}_nd_f(b+x)}{n} \tag{24}$$

is the average annual female deaths in the age interval b+x to b+x+n. Therefore, (23) simplifies into

$$H_m(a,b) = \frac{1}{n}\sum_{i=0} {}_nL_m(a+in)\,{}_nd_f(b+x) \tag{25}$$

where

$${}_nL_m(a+in) = \int_{in}^{(i+1)n} \ell_m(a+x)\,dx \tag{26}$$

is the size of the stationary population in the age interval a+in to a+(i+1)n. Unfortunately, the upper limit of i needed for the evaluation of (25) is not known as the life table functions are generally not available beyond some age α (80 or 85). Consequently, assumptions have to be made about the contributions of the terms in (25) beyond α. It appears though, that when a and b are both sufficiently smaller than α, P(a,b), which because of (25) and a similar definition of $H_f(a,b)$ reduces to

$$P(a,b) = \frac{H_m(a,b)}{H_m(a,b)+H_f(a,b)} \tag{27}$$

after appropriate substitutions in (9), can be assumed to remain unaffected when the H functions are evaluated over the age interval (a,α) for $a\geq b$ or over (b,α) when $a<b$. Although, it is not immediately apparent, such an assumption leads to the mathematical equality

$$P(a,b) = P(\alpha,\alpha-a+b) \tag{28}$$

for $a\geq b$, and

$$P(a,b) = P(\alpha-a+b,\alpha) \tag{29}$$

otherwise. These equalities can be established by first distinguishing the H functions over the reduced interval as H^α. In that case, for a>b, (27) can be written as

$$P(a,b) = \frac{H_m^\alpha(a,b) + H_m(\alpha,\alpha-a+b)}{H_m^\alpha(a,b)+H_f^\alpha(a,b)+H_m(\alpha,\alpha-a+b)+H_f(\alpha,\alpha-a+b)} \tag{30}$$

Thus, the assumption of the equality

$$P(a,b) = \frac{H_m^\alpha(a,b)}{H_m^\alpha(a,b)+H_f^\alpha(a,b)} \tag{31}$$

produces the other equality, namely,

$$P(a,b) = \frac{H_m(\alpha,\alpha-a+b)}{H_m(\alpha-a+b)+H_f(\alpha-a+b)} \tag{32}$$

which also equals $P(\alpha,\alpha-a+b)$ because of (27). Similarly, (29) can be established for a<b. From logical considerations, it may be added that (28) and (29) further imply (but do not mathematically require) that

$$P(a,b) = P(a+h,b+h) \tag{33}$$

for all h. This is so, because like P(a,b), the probability P(a+h,b+h) can also be estimated from $H^\alpha(a+h,b+h)$, at least for small h, in which case

$$P(a+h,b+h) = P(\alpha,\alpha-a+b) \text{ for } a>b \tag{34}$$

Because of (32), therefore, the equality proposed in (33) should also hold.

It is obvious that if P(a,b) is estimated from (32) for a given age combination a and b of the spouses and the same is set equal to $P(\alpha,\alpha-a+b)$, the substitution of the latter in (6), where a and b are changed respectively to a+h and b+h, will generate the estimate of P(a+h,b+h) as

$$P(a+h,b+h) = \frac{H_m^\alpha(a+h,b+h)+\ell_m(\alpha)\,\ell_f(\alpha-a+b)P(\alpha,\alpha-a+b)}{\ell_m(a+h)\,\ell_f(b+h)} \tag{35}$$

Needless to say, the mathematical equality of P(a,b) and P(a+h,b+h) does not follow from such a procedure, however, as mentioned earlier, the difference between these two probabilities should be negligible.

The reader must have noted the equivalence of

the end results generated by the present method with those based on the Gompertz law of mortality. Surprising as the results may be, the eventual probability of becoming a widow or a widower (and similarly the conditional probability over any time interval), seems to be determined by the age difference of the spouses and not by their actual ages.

6. Application on the U.S. Data and Discussion of Results

The values of P(a,b) have been calculated from the 1973 U.S. Life Tables, on the basis of the two methods presented in this paper. These are shown in Table 3, in which the age difference between the two spouses has an arbitrarily chosen range of -10 to 10 years. It may be recalled that the principal parameters of the Gompertz model were estimated from the age interval (30, 85), from which K(a,b) values were obtained and shown in Table 2. Substitutions of these values in (15) provide estimates of P(a,b) which are shown in cols (2-3) of Table 3. Next, the $H^{\alpha}(a,b)$ functions, required for the method based on numerical integration, are obtained for those combinations of a and b, such that the

$$\text{minimum } (a,b) = 15 \qquad (36)$$

The choice of age 15 is based on a reasonable minimum of the observed ages of marriage, and in this way the difference between α and the minimum (a,b) is maximized in order to strengthen the assumption resulting in (31) and (32). From the same life tables, these H functions are then obtained for α=85, for substitutions in (31) to generate the estimates of P(a,b) for different age combinations of the two spouses.

As expected, the values of P(a+h,b+h) are found to be virtually invariant with respect to h, and instead of reproducing all such values, only the minimum and the maximum values have been shown in cols (4-7) of Table 3 for the integral values of $|a-b| \leq 10$.

The closeness of the estimates generated by the two different methods mutually reinforce the validity of the separate assumptions on which these are based. According to the tabled values, the probability of becoming a widow is at least twice as large than that of becoming a widower when the husband is two to three years older than the wife. The differential risks of losing a spouse for this currently normative age difference is worth noting. The two probabilities become equal when the husband is about seven years younger than the wife, a figure which is slightly less than the difference between the life expectancies of the two sexes. It will be interesting to see how these probability measures compare with those generated from other life tables.

7. Acknowledgement

I am grateful to Fr. Joseph C. O'Hara and Mr. S.N. Banerjee, both of the Department of Sociology and Anthropology for developing the computer program and for preparing the tables presented in the paper.

8. References

Depoid, Pierre

1938 "Tables nouvelles relatives a la population francaise", Bulletin de la statistique generale de la France 27 (II), 269-324

Jordan, C.W., Jr.

1967 Life Contingencies, The Society of Actuaries, Chicago

Speigelman, M.

1969 Introduction to Demography, Cambridge, Mass.

USDHEW

1976 Vital Statistics of the United States, 1973, Vol. II, Section 5, Life Tables

TABLE 1. Estimated values of C of the Gompertz model $\mu(x) = BC^x$ for U.S. by sex for the year 1973 beginning age 30 (Source: 1973 U.S. Life Tables)

Age	30	35	40	45	50	55	60	65	70	75
C(male)	1.06	1.08	1.10	1.09	1.10	1.09	1.08	1.08	1.09	1.07
C(female)	1.09	1.09	1.09	1.08	1.09	1.08	1.08	1.11	1.11	1.09

TABLE 2. Estimated values of $K(a,b)=(B_m/B_f)C^{a-b}$ for different values of husband wife age differentials a-b (Source: 1973 U.S. Life Tables)

$\lvert a-b \rvert$	0	1	2	3	4	5	6	7	8	9	10
K(a,b) $a>b$	1.71	1.86	2.02	2.20	2.40	2.61	2.84	3.09	3.36	3.66	3.99
K(a,b) $a\leq b$	1.71	1.57	1.44	1.32	1.22	1.12	1.02	0.94	0.87	0.80	0.73

TABLE 3. Probability P(a,b) of husband a years old, outliving the wife aged b years, estimated by the methods based on (1) Gompertz law of mortality and (2) numerical integration (Source: 1973 U.S. Life Tables)

P(a,b) by Gompertz law			Optimum values of P(a,b) by Numerical Integration			
			$a\geq b$		$a\leq b$	
$\lvert a-b \rvert$	$a>b$	$a<b$	Minimum	Maximum	Minimum	Maximum
(1)	(2)	(3)	(4)	(5)	(6)	(7)
0	.370	.370	.345	.382	.345	.382
1	.350	.390	.327	.361	.364	.398
2	.331	.410	.308	.341	.383	.415
3	.312	.431	.291	.321	.403	.433
4	.295	.452	.273	.301	.422	.452
5	.277	.473	.256	.282	.443	.472
6	.261	.494	.240	.263	.463	.492
7	.244	.515	.224	.245	.484	.512
8	.229	.536	.209	.227	.504	.532
9	.215	.557	.196	.211	.525	.551
10	.201	.578	.183	.195	.545	.571

AN EMPIRICAL COMPARISON OF THE SIMPLE INFLATION, SYNTHETIC AND COMPOSITE ESTIMATORS FOR SMALL AREA STATISTICS

Wesley L. Schaible, Dwight B. Brock, and George A. Schnack
National Center for Health Statistics

I. Introduction

Large samples such as those of the Current Population Survey (CPS) and Health Interview Survey (HIS) have been designed to provide national and regional estimates. As useful as such statistics are, there is considerable demand for additional estimates for smaller geographic areas, for example, States and counties. One way to meet this demand is to redesign the survey, but this can be both expensive and time consuming. Depending upon resources and objectives, other approaches, although they may produce biased estimates, deserve consideration. Several biased estimators were considered in 1968 in the publication, Synthetic State Estimates of Disability. The authors stated that the sample size (and design) of HIS was inadequate to make State estimates by conventional procedures and suggested that a synthetic estimator be used.

This estimator has since received considerable attention. Levy (1971) used mortality data to compute average relative errors of synthetic estimates for States. Gonzalez and Waksberg (1973) calculated mean square errors averaged over all small areas and compared synthetic and direct estimates for selected Standard Metropolitan Statistical Areas. Gonzalez and Hoza (1975) investigated errors of synthetic estimates using unemployment data for counties from the CPS and the 1970 Census. Namekata, Levy and O'Rourke (1975) investigated synthetic State estimates of work loss disability in a similar manner. Schaible, Brock and Schnack (1977) compared the average squared errors of synthetic and direct estimates of unemployment rates for county groups in Texas.

It is the purpose of this paper to compare the synthetic estimator, a simple direct estimator and a composite estimator, which is a weighted function of the other two. To provide information regarding the performance of the three estimators each was used with 1970 HIS data to produce estimates of unemployment rates for 25 HIS primary sampling units (PSU's) in Texas. Comparable parameter values were obtained from the U.S. Bureau of the Census (1972), General Social and Economic Characteristics. A similar procedure was followed in estimating the percent of the population completing college for each of the fifty States and the District of Columbia. Three years of HIS data were combined (1969-1971), and comparable population values were obtained from the 1970 Census Public Use Sample Tapes. The State values obtained from the one-in-one hundred sample on these tapes were treated as population values for comparison with estimates.

Traditionally the estimator used to produce estimates reflects the design of the sample from which the data were collected. Even though this is not entirely true of the estimators considered in this paper a few remarks about the HIS design will be useful as background for the comparisons presented. For more complete details on this design, see NCHS (1958). The HIS uses a multistage probability design which permits continuous sampling of households from the civilian, noninstitutionalized population of the United States. The first stage of the 1970 design consisted of a sample of 357 primary sampling units (PSU's) chosen from approximately 1,900 geographically defined units covering the 50 States and the District of Columbia. A PSU is defined as a county, a group of contiguous counties or a Standard Metropolitan Statistical Area. Within each PSU, Census enumeration districts are ordered geographically and divided into small clusters of households. A systematic sample of clusters is then selected. The 1970 HIS sample was composed of some 37,000 households, or a total of about 116,000 individuals.

II. Estimators

The synthetic estimator generally has a negligible variance but often a nonnegligible bias that can only be estimated under special conditions (Schaible, 1975). This non-quantifiable bias is a serious problem and is one reason the synthetic estimator has been used only in special situations. The justification for using this estimator is based on the assumption that the characteristic being estimated is correlated with certain demographic characteristics of the population. The first step in constructing a synthetic estimate is to create a cross-classification of demographic cells in such a way that the local area population in each cell is known. The synthetic estimate for a local area is then formed by weighting a larger area estimate of the health characteristic for each demographic cell by the proportion of the local area population in that cell and then summing over all cells.

For a more precise definition of the synthetic estimator let $y_{d\alpha i}$ denote the observation of interest on the ith individual in the αth demographic cell in the dth local area. Here $i=1,2,\ldots n_{d\alpha}$, the number of sample units in the dth local area and αth cell, $d=1,\ldots,D$, the total number of local areas, and $\alpha=1,2,\ldots k$, the number of α-cells. Also, let $N_{d\alpha}$ represent the number of people in the population in area d and cell α. The sample mean of the αth demographic cell for the larger area is then

$$\bar{y}_{.\alpha} = \sum_{d=1}^{D} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i} / n_{.\alpha} ,$$

and the simple synthetic estimator for local area d is

$$\bar{y}'_{d.} = \sum_{\alpha=1}^{k} N_{d\alpha} \bar{y}_{.\alpha} / N_{d.} . \qquad (1)$$

Two synthetic estimators are used here, one when the small areas investigated are States and a slightly different one when the small areas are county groups. The estimator used to produce estimates for States is described above, except for the addition of appropriate sampling weights and a ratio-adjustment. This ratio-adjustment forces the weighted sum of the individual State synthetic estimates in a geographic region to be consistent with the usual HIS probability estimate for that region. There is evidence to suggest that when estimating for States, the synthetic estimator with this adjustment has smaller average squared error than without the adjustment (Schaible et al, 1976). The α-cells for State synthetic estimates in this paper were defined to be the 64 cells created by cross-classifying the following variables:

1. Color: white; other

2. Sex: male; female

3. Age: under 17 years; 17-44 years; 45-64 years; 65 years and over

4. Family size: fewer than 5 members; 5 members or more

5. Industry of head of family: Standard Industrial Classifications: (1) forestry and fisheries, agriculture, construction, mining, and manufacturing; (2) all other industries.

For these cells 1970 State populations are available from the Census Public Use Tapes, and reliable national estimates are available from HIS. However, for county estimates, where the larger area was defined as the Southern Region, the HIS sample sizes in some cells are small. In this case, 8 cells were defined by the age and sex groups above. County populations were available for these cells in the Bureau of the Census (1971), General Population Characteristics. The synthetic estimator used for county group estimates did not contain a ratio adjustment.

If data from a sample designed to make estimates for a large area are to be used to make estimates for a small area and there are no sample units in the small area of interest, then obviously conventional estimation methods cannot be used and a synthetic approach must be considered. However, when estimating for a small area that contains sample units the possibility of using conventional estimators should not be ignored. It is evident that at some point, as the number of sample units in an area increases, a conventional estimator becomes more desirable than a synthetic one. This is true whether or not the sample was designed to produce estimates for small areas. Thus a second, more direct approach, is to use conventional estimators with the sample units that fall in the small area of interest. This approach, while not new, has received little attention in the literature, but it would seem to have potential for areas where sample sizes are reasonably large. For example, in California 96 percent of the population reside in the primary sampling units surveyed by HIS and the total HIS sample size exceeds 10,000 persons each year. In situations such as this, one suspects that a conventional direct estimator might be more appropriate than a synthetic one.

The simplest of the conventional estimators is the unweighted sample mean or simple inflation estimator, which for local area d may be written as

$$\bar{y}_{d.} = \sum_{\alpha=1}^{k} \sum_{i=1}^{n_{d\alpha}} y_{d\alpha i} / n_{d.} \quad . \qquad (2)$$

The simple inflation estimator is by far the most widely used of the three considered here. Its simplicity is appealing and with appropriate sample design it is unbiased and its variance can be estimated. However, when used to estimate for small areas from samples designed for large areas (as are the HIS and CPS) the conventional sampling theory model yields little information about the properties of this estimator. For this reason alternative estimators have been proposed.

The idea of a composite estimator is not new; it was discussed in the 1968 publication cited above. It was also mentioned there that a desirable feature of such an estimator would be that the synthetic component receive more weight when the State sample size was small and the direct component receive more weight when the sample size was large. Royall (1973) in a discussion of papers by Gonzalez (1973) and Ericksen (1973), also suggested that a choice between direct and synthetic approaches need not be made but that "... a combination of the two is better than either taken alone." Also, as related by Gonzalez and Hoza (1975), "In a seminar given at the Bureau of the Census in March 1975, William G. Madow suggested a combination of synthetic estimates and observed values for the primary sampling units included in the CPS." Investigations into the basis for and properties of the composite and other related estimators are presently taking place (Royall, 1977, Schaible, 1977).

One rather obvious approach to arrive at a specific composite estimator is to weight each component by the inverse of its squared error and then normalize so the sum of the resulting weights is unity. Empirical comparisons of the errors of various direct and synthetic estimators for States and county groups led to a specific formulation of such a composite estimator. Given a design assume the expected squared error of the simple inflation estimator is of the form $b/n_{d.}$, and that of the synthetic estimator is b', where b and b' are constants. Then if each component estimator is weighted by the inverse of its expected error, the following composite estimator results:

$$\bar{y}''_{d.} = (c_d)\,\bar{y}_{d.} + (1-c_d)\,\bar{y}'_{d.} \quad , \qquad (3)$$

where $c_d = n_{d.}/(n_{d.} + b/b')$. The quantity b/b' is the small area sample size at which the expected errors of the two component estimators are equal.

III. Results

Figures 1, 2 and 3 show the plots of esti-

mated unemployment rates versus the actual rates obtained from the 1970 Census for the three estimators considered. The vertical distance from a point to the 45 degree line shows the magnitude of the error of the estimate represented by that point. The average squared error (Table 1) is simply the average over the 25 county groups of the squares of the differences between the estimates and the corresponding Census values. The correlations (Table 2) given are Pearson's product moment correlation coefficients.

The average squared error in estimated unemployment rates produced by the simple inflation estimator is large, 6.85 percentage points. However, it should be noted that the 1970 HIS sample sizes of the civilian labor force in these county groups are generally small. In 18 of the 25 county groups the number of sample people in the civilian labor force is less than 90. As would be expected, large errors occur in county groups where the sample sizes are small. Actual unemployment rates range from 2.2 to 6.6 percent, while the simple inflation estimates range from 0.0 to 11.6 percent. The correlation coefficient between simple inflation estimates and actual values is .52.

The plot of actual and synthetic unemployment rates is shown in Figure 2. The average squared error of the synthetic estimates is 1.27, much smaller than that of the simple inflation estimator. However, the correlation coefficient of estimates and actual values is only .08. The synthetic estimates cluster around 3.5 percent and range from 3.2 to 3.8 percent. This clustering near the value of the larger area mean is a common characteristic of the synthetic estimator. This is at least partially due to the fact that the variables used to define the α-cells are often not sufficiently correlated with the item being estimated to yield a good estimate for a given small area. When this is true, the magnitude of the bias for a given small area will increase with the difference between the small area parameter and the parameter of the larger area used to produce estimates. These results suggest that the synthetic estimator may be a poor choice if one is interested in either estimating levels of those areas with extreme values or comparing levels between small areas.

The composite estimator, by weighting the simple inflation estimate less heavily when the sample size is small, tends to reduce the large errors of the simple inflation estimates in those areas with small sample sizes; and by weighting the simple inflation estimate more heavily when the sample size is large, tends to reduce the large errors of the synthetic estimator when the actual value of the small area is very different from that of the large area. The plot of the composite estimates is shown in Figure 3. The average squared error is .92, less than that of either the simple inflation or synthetic estimator. The correlation coefficient is .51, essentially the same as that of the simple inflation estimator.

Figures 4, 5, and 6 show the plots of State estimates of percent of the population completing college versus the percent obtained from the 1970 Census for each of the three estimators. Average squared errors are shown in Table 1 and correlation coefficients in Table 2.

States of course have much larger HIS sample sizes than county groups, and this is reflected in the difference between the plots of simple inflation estimates in Figures 1 and 4. Also, as in Figure 1, the large deviations in Figure 4 are generally those of estimates for States with relatively small sample sizes. The average squared error of the simple inflation estimates of the percent completing college is 1.81 and the correlation coefficient between estimate and actual value is .69.

The synthetic estimates for the percent completing college (Figure 5) are more closely clustered around the 45 degree line than the county group synthetic estimates of the unemployment rate (Figure 2). This might be partially due to differences in the predictability of characteristics of States and counties and/or the difference in the number of cells used in the synthetic estimator and the variable estimated. The average squared error of the synthetic estimates of percent completing college is 1.76, essentially the same as that of the simple inflation estimator. The correlation coefficient is .45. The majority of the difference between the correlation coefficients of the simple inflation and synthetic estimators is explained by the point representing the District of Columbia (actual percent 11.2). The observation that the synthetic estimator often does not do well in estimating for certain areas including the District of Columbia has been made before (personal communication, Levy, Gonzalez).

The average squared error of the composite estimates of the percent completing college is 1.09 and the correlation .72. As in the previous example, the composite estimator yields a smaller average squared error than either of the component estimators and also produces a correlation as good as the better of the two component estimators.

IV. Summary

In estimating both the unemployment rates for county groups in Texas and the percent of the population completing college for States the composite estimator has an average squared error approximately 30 percent less than that of the synthetic estimator. With both variables the synthetic estimator has smaller average squared error than does the simple inflation estimator, the other component of the composite estimator. Also, when estimates are correlated with actual values the composite estimator has correlation coefficients as large as those of the simple inflation estimator which are larger than those of the synthetic estimator.

There are, of course, other ways to define weights for the composite estimator. In fact, preliminary results with these and other data indicate that other weighting schemes can produce further reductions in average squared error and further increases in the correlation with actual values. Preliminary results also indicate that the composite estimator is remarkably robust against poor estimates of the unknown quantity b/b'.

The above is only a small empirical study of the performance of three estimators under rather restricted circumstances. However, these results are encouraging, and investigations of the properties of composite estimators are continuing.

References

Ericksen, Eugene P. (1973): "Recent Developments in Estimation for Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 37-41.

Gonzalez, Maria E. (1973): "Use and Evaluation of Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 33-36.

Gonzalez, Maria E. and Waksberg, Joseph E. (1973): "Estimation of the Error of Synthetic Estimates." Presented at the first meeting of the International Association of Survey Statisticians, Vienna, Austria.

Gonzalez, Maria E. and Hoza, Christine (1975): "Small Area Estimation of Unemployment." Proceedings of the American Statistical Association, Social Statistics Section, pp. 437-443.

Levy, Paul S. (1971): "The Use of Mortality Data in Evaluating Synthetic Estimates." Proceedings of the American Statistical Association, Social Statistics Section, pp. 328-331.

Namekata, Tsukasa, Levy, Paul S., and O'Rourke, Thomas W. (1975): "Synthetic Estimates of Work Loss Disability for Each State and the District of Columbia." Public Health Reports, 90, pp. 532-538.

National Center for Health Statistics (1958): "Statistical Design of the Health Household Interview Survey." Health Statistics, Series A-2. Publication No. 584-A2. Public Health Service, Washington, D.C.

Royall, Richard M. (1973): "Discussion of two papers on Recent Developments in Estimation of Local Areas." Proceedings of the American Statistical Association, Social Statistics Section, pp. 43-44.

Royall, Richard M. (1977): "Statistical Theory of Small Area Estimates - Use of Prediction Models." Unpublished report prepared under contract from the National Center for Health Statistics.

Schaible, Wesley L. (1975): "A Comparison of the Mean Square Errors of the Postratified, Synthetic and Modified Synthetic Estimators." Unpublished report, Office of Statistical Research, National Center for Health Statistics.

Schaible, W.L., Casady, B., Schnack, G.A. and Brock, D.B. (1976): "An Empirical Comparison of Some Conventional and Synthetic Estimators for Small Areas." Draft report, National Center for Health Statistics.

Schaible, Wesley L., Brock, Dwight B. and Schnack, George A. (1977): "An Empirical Comparison of Two Estimators for Small Areas." Presented at the Second Annual Data Use Conference of the National Center for Health Statistics. Dallas, Texas.

Schaible, Wesley L. (1977): "Notes on Composite Estimators for Small Areas." Unpublished memoranda, Office of Statistical Research, National Center for Health Statistics.

U.S. Bureau of the Census (1971): General Population Characteristics-Texas, PC(1)-C45, U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1972): General Social and Economic Characteristics-Texas, PC(1)-C45, U.S. Government Printing Office, Washington, D.C.

Acknowledgements

The authors would like to thank Barry Peyton and Eugene Diggs of the Office of Statistical Research programming staff for their efforts in computing the estimates and providing the graphical presentations for this paper. Also, the authors thank Kay Barrett and Anita Powell for typing the manuscript.

TABLE 1. Average Squared Errors of the Simple Inflation, Synthetic and Composite Estimators for Two Variables, Health Interview Survey, 1970.

Estimator	Variable	
	Unemployment Rate	Percent Completing College
Simple Inflation	6.85	1.81
Synthetic	1.27	1.76
Composite	.92	1.09

TABLE 2. Correlation Coefficients between Estimate and Actual Value for Three Estimators and Two Variables, Health Interview Survey, 1970.

Estimator	Variable	
	Unemployment Rate	Percent Completing College
Simple Inflation	.52	.69
Synthetic	.08	.45
Composite	.51	.72

FIGURE 1. Unemployment Rates, Simple Inflation Estimates and Actual Values for 25 County Groups in Texas, 1970

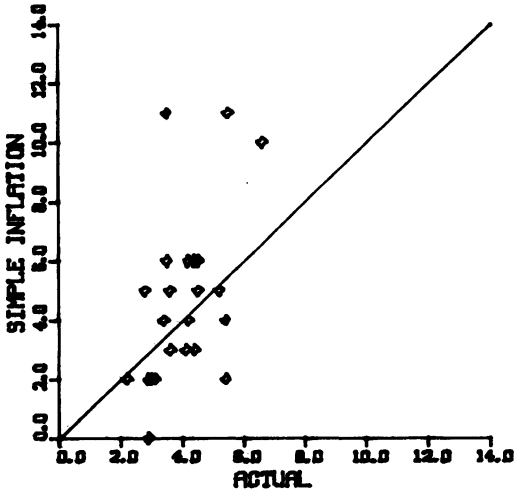


FIGURE 2. Unemployment Rates, Synthetic Estimates and Actual Values for 25 County Groups in Texas, 1970

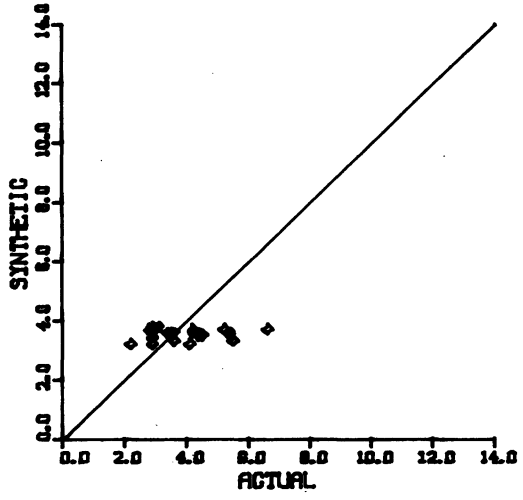


FIGURE 3. Unemployment Rates, Composite Estimates and Actual Values for 25 County Groups in Texas, 1970

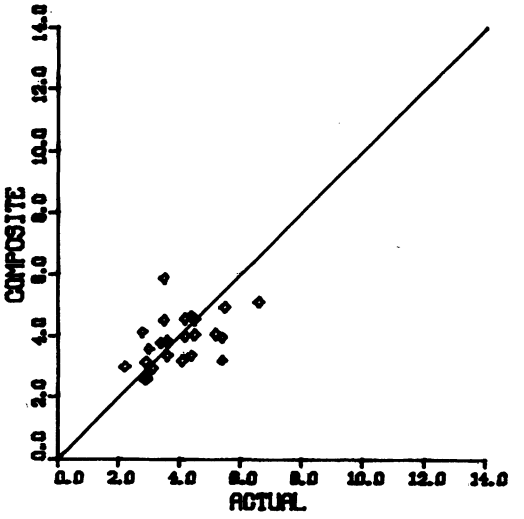


FIGURE 4. Percent of the Population Who Have Completed College, Simple Inflation Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971

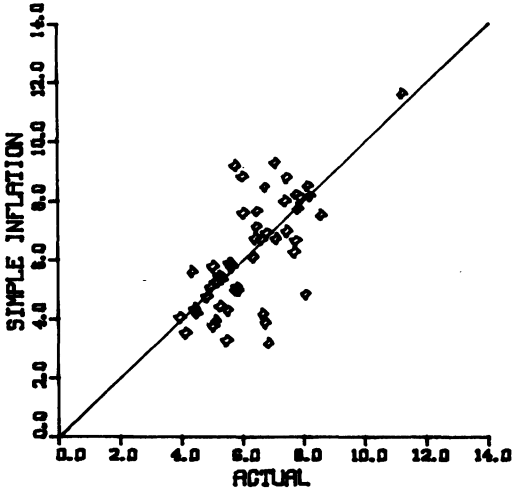


FIGURE 5. Percent of the Population Who Have Completed College, Synthetic Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971

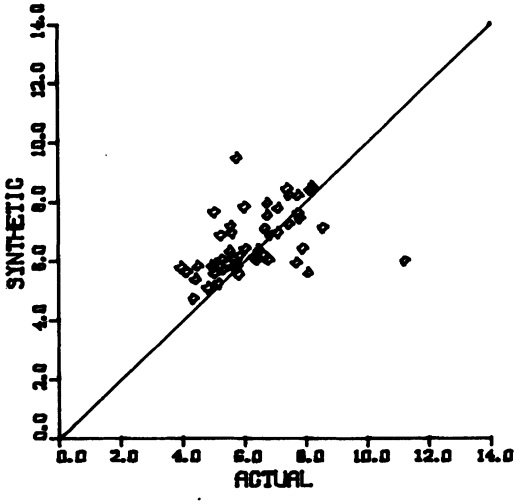
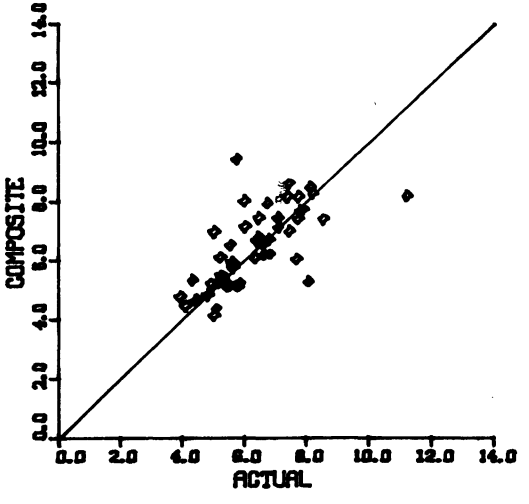


FIGURE 6. Percent of the Population Who Have Completed College, Composite Estimates and Actual Values for Fifty States and the District of Columbia, 1969-1971



APPLICATION OF MONTE CARLO SIMULATION TO RIDGE REGRESSION

Wen-Fu P. Shih, Florida Atlantic University
Sun-Fu Shih, University of Florida, IFAS, AREC

ABSTRACT

Ridge regression has been introduced to solve the multicollinearity in multiple linear regression. The performance of approximate ridge estimators (AOPT) and Optimume ridge coefficients (OPT) are compared with ordinary least square (OLS) estimators by using the Monte Carlo simulation technique. The results indicated that the correlation coefficient r between two independent variables is an important factor to choose the method. For instance, when the r value is less than 0.5, the OLS performs as good as AOPT and OPT methods. When the r lies around 0.6 to 0.8, OPT is the best technique among those three methods, and AOPT is better than OLS. For those data with high correlations such as 0.9, both OPT and AOPT are all good methods to use. In general conclusion, the performance of OPT is better than AOPT, and the AOPT is better than OLS in terms of minimizing the mean square error of estimation in regression analysis to solve the multicollinearity among the independent variables.

INTRODUCTION

Multiple linear regression technique has been extensively used by the fields of engineering, science, technology, economics and social science for data analysis. But, the estimation of regression coefficients can present problems when multicollinearity (highly correlated independent variables) exists among variables. For discussion of problems of multicollinearity, see Althauser, 1971; Blalock, 1963, 1944; Christ, 1966; Gordon, 1963; Johnston, 1972; Rockwell, 1975.

The problems of multicollinearity have been solved by one statistical technique called ridge regression which was introduced by Hoerl and Kennard (1970a, 1970b) and applied by others (Deegan, 1975; Dempster, Schartzoff and Wermuth, 1977; McDonald and Schwing, 1973; McDonald and Galarneau, 1973; Vinod, 1975; etc.) These authors showed that by adding a small non-negative constant k to the diagonal of the correlation matrix of independent variables to substantially reduce error variance and thereby control for the general instability of ordinary least square (OLS) estimates.

The question remains, however, of the appropriate amount of bias to introduce as the ridge analysis increment. In recent papers Hoerl, et.al. (1975) have suggested an approximation to the optimum value k (AOPT) so that ridge regression produces a smaller square error than OLS. Shih and Kasarda (1977) have proposed a method for selecting the optimal k (OPT) for ridge analysis in terms of minimizing the mean square error of estimation.

The purpose of this paper is based on the Monte Carlo simulation technique to compare the performance of ordinary least square (OLS) with AOPT and OPT ridge regressions by simulating the different patterns of regression coefficients with different degrees of collinearity or multicollinearity among independent variables.

METHODOLOGY DESCRIPTION

Ridge Regression and Optimization

Consider the standard model for multiple linear regression:

$$Y = X\beta + \varepsilon \tag{1}$$

where Y is a nx1 vector of observations on a dependent variable, X is a nxp matrix of nonstochastic regressors with rank p, β is a px1 vector of unknown regression coefficients, and ε is a nx1 vector of unknown disturbances. Assuming $E(\varepsilon)=0$, and $E(\varepsilon\varepsilon')=\sigma^2 I_n$, the ordinary least square estimator (OLS) of β is

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{2}$$

$$\text{with } Var(\hat{\beta}) = \sigma^2(X'X)^{-1} \tag{3}$$

where $\hat{\beta}$ is an unbiased estimator of β and has the minimum variance within the class of unbiased estimators (Goldberger, 1964). As we noted, if the X's are highly collinear, the variance of β tends to become large, and little confidence can be placed in $\hat{\beta}$ as an estimator of β. By adding positive constant k to each of the diagonal elements of X'X one can reduce the variance of the regression estimate, but at the expense of some bias. The resulting is the ridge estimator

$$\beta^* = (X'X + kI)^{-1}X'Y \tag{4}$$

where k is a positive scalar, and β^* is a biased estimator of β with

$$Var(\beta^*) = (X'X+kI)^{-1}X'X(X'X+kI)^{-1}\sigma^2 \tag{5}$$

β^* is equal to $\hat{\beta}$ when k equals zero.

As have been shown in Hoerl and Kennard (1970 a, b), the mean square error of ridge estimator β^* can be written as

$$MSE(\beta^*) = \sum_i VAR(\beta_i^*) + Bias^2(\beta^*) \tag{6}$$

where

$$\sum_i VAR(\beta_i^*) = \sigma^2 T_r(Z) \tag{7}$$

with

$$Z = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{8}$$

and

$$Bias^2(\beta^*) = [E(\beta^*-\beta)]'[E(\beta^*-\beta)] \tag{9}$$

In equation 6, the total variance decreases as k increases, while the square bias increases with k. Based on these monotonic properties and existence of minimum point which also shown by Hoerl and Kennard (1970a, b), Shih and Kasarda (1977) have also shown that by computerized iteration procedures one can locate the optimum point k, which minimizes a consisten estimator of MSE (β^*), namely $mse(\beta^*(k))$. Where

$$mse(\beta^*(k)) = \hat{\sigma}^2 T_r(Z) + (\beta^*(k)-\hat{\beta})'(\beta^*(k)-\hat{\beta}) \tag{10}$$

where

$$\hat{\sigma}^2 = (Y'Y - \hat{\beta}X'Y)/(n-p) \tag{11}$$

Let V(k) denote an estimator of total variance of $\beta^*(k)$ and BS(k) denote an estimator of the bias square, then equation (10) becomes

$$mse(\beta^*(k)) = V(k) + BS(k) \tag{12}$$

where

$$V(k) = \hat{\sigma}^2 T_r(Z),$$

$$BS(k) = (\beta^*(k)-\hat{\beta})'(\beta^*(k)-\hat{\beta}),$$

$$V(0) = \hat{\sigma}^2 T_r(X'X)^{-1}, \quad \text{and}$$

$S(0) = 0$ for $k \geq 0$

For given $k_1 > k_2 \geq 0$, we know that

$$V(k_2) \geq V(k_1) \quad (13)$$

and

$$S(k_2) \leq S(k_1) \quad (14)$$

The existence of a minimum point shows that $mse[\beta^*(k-c)] > mse[\beta^*(k)]$

$$< mse[\beta^*(k+c)] \quad (15)$$

for a small constant c and leads to the conclusion that k is a point which gives the minimum $mse(\beta^*)$.

The iteration procedures to obtain this point can be summarized as follows:

(a) Read input data and desired tolerance of accuracy.

(b) Compute $\hat{\sigma}^2$ and $\hat{\beta}$ from equations 11 and 2, respectively.

(c) Initiate the $k = 0$ and an increment $\Delta k = 0.1$.

(d) Compute V(0) and BS(0).

(e) Let a new variable $k1 = k + \Delta k$.

(f) Compute V(k1) and BS(k1).

(g) Check the relationship between BS(k1) - BS(k) and V(k) - V(k1).

(h) If BS(k1) - BS(k) < V(k) - V(k1), then let k = k1, V(k) = V(k1), BS(k) = BS(k1), and the procedures of e, f, and g are repeated.

(i) If BS(k1) - BS(k) $\geq$ V(k) - V(k1), then check if the Δk is less than a desired tolerance. If the answer is no, the Δk is replaced by $\Delta k/10$ and the procedures of e, f, and g are repeated. If the answer is yes, the iteration procedures are complete, and k1 is the optimal value.

The above procedures have been converted to a computer program with Fortran IV language.

Hoerl, et. al. (1975) also suggested that an approximation method to obtain the optimal value k_a (AOPT) can be expressed as follows:

$$k_a = P\hat{\sigma}^2/\hat{\beta}'\hat{\beta} \quad (16)$$

By simulation technique they showed that k_a can produce a smaller average square error than OLS, the distribution of squared errors for the regression coefficients has a smaller variance than does that for OLS, and the probability that the ridge regression produces a smaller square error than OLS is greater than 0.50.

Monte Carlo Simulations

Applying simulation techniques to examine the performance of ridge regression has been found in many recent publications (McDonald and Galarnean, 1973; Hoerl, et.al., 1975; and Dempster, et.al., 1977). The basic Monte Carlo simulation used in this study is described as follows. The observations x_{ij} are generated based on the following simulation generator:

$$x_{ij} = (1-\alpha^2)^{0.5} Z_{ij} + \alpha Z_{i(p+1)}; \quad (17)$$
$$i = 1,\ldots,n;\ j = 1,\ldots,p.$$

where, n is the number of observations for each explanatory variable; p is the number of independent variables; Z_{ij} are independent standard normal pseudo-random numbers and α is specified so that the correlation between any two independent variables is given by α^2. The X's are then standardized so that X'X is in correlation form. A true regression coefficient β is chosen as a normalized eigenvector corresponding to the largest eigenvalues of the X'X matrix. Newhouse and Oman (1971) have noted that $MSE(\beta^*)$ is minimized when β is such eigenvector subject to the constraint that $||\beta|| = 1$ when k is fixed.

Observations on the dependent variable are determined by

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad (18)$$

where x's are computed from equation 17, ε_i are independent normal $(0, \sigma^2)$ psuedo-random numbers and β_0 is taken to be zero. The variables are standardized so that X'Y represents the vector of correlations of the dependent variable with each independent variable.

Based on the Central Limit Theory, the major error involved in Monte Carlo simulation is a statistical sample error which is proportional to the $(1/\sqrt{N})$, where N is the total number of trials (Shih and Hamrick, 1974). In other words, one must increase the sample size by a factor of 4 in order to halve the possible error. Therefore, an additional set of samples W must be generated to increase the accuracy of simulation.

Comparisons of Different Methods

The observations generated by Monte Carlo simulation are used to perform ordinary least square (OLS) estimators, optimum ridge coefficients (OPT), and approximate ridge estimators (AOPT). The symbols $\hat{\beta}$, $\beta^*(k)$, and $\beta^{**}(k)$ represent the standardized coefficients of OLS, OPT and AOPT, respectively. Those standardized coefficients are then transformed back to the original coefficients. The constant terms are then computed as:

$$\hat{\beta}_0 = \bar{y} - \sum_{j=1}^{p} \hat{\beta}_i \bar{x}_j \quad \text{for OLS} \quad (19)$$

$$\beta_0^*(k) = \bar{y} - \sum_{j=1}^{p} \beta_j^*(k)\bar{x}_j \quad \text{for OPT} \quad (20)$$

and

$$\beta^{**}(k) = \bar{y} - \sum_{j=1}^{p} \beta_j^{**}(k)\bar{x}_j \quad \text{for AOPT} \quad (21)$$

where

$$\bar{y} = (1/n)\left(\sum_{i=1}^{n} y_i\right)$$

$$\bar{x}_j = (1/n)\left(\sum_{i=1}^{n} x_{ij}\right)$$

The total mean square errors are computed as

$$L = L(0) = \sum_{i=0}^{p} (\beta_i - \hat{\beta}_i)^2 \quad \text{for OLS} \quad (22)$$

$$L^* = L[\beta^*(k)] = \sum_{i=0}^{p} [\beta_i - \beta_i^*(k)]^2 \quad \text{for OPT} \quad (23)$$

$$L^{**} = L[\beta^{**}(k)] = \sum_{i=0}^{p} [\beta_i - \beta_i^{**}(k)]^2 \quad \text{for AOPT} \quad (24)$$

As noted above β_0 equals zero.

In order to find the regression coefficients which can produce a smaller mean square error than the corresponding least squares estimator a measure of the improvement can be obtained by

$$M^* = E[L^*(k)]/E[L(0)] \quad (25)$$

and

$$M^{**} = E[L^{**}(k)]/E[L(0)] \quad (26)$$

where E[L(k)] is the average sum of mean square error of specific ridge estimator.

$$E[L^*(k)] = (1/W) \sum_{w=1}^{W} L_w^* \quad (27)$$

$$E[L^{**}(k)] = (1/W) \sum_{w=1}^{W} L_w^{**} \quad (28)$$

and

$$E[L(0)] = (1/W) \sum_{w=1}^{W} L_w \quad (29)$$

where L_w, L_w^*, and L_w^{**} are total mean square error of sample w for OLS, OPT, and AOPT, respectively. W is the number of sample sets as indicated in the section of Monte Carlo simulation. The smaller value of M* and M** implies that the method used has a better solution.

EXAMPLE OF APPLICATIONS

The values of n = 100; p = 3; α = 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99; and σ = 0.01, 0.21, 0.41, 0.61, 0.81, and 1.01 are used in this study. The coefficients of β corresponding to each α value are given in Table 1.

TABLE 1: Values of α and Coefficient Vectors β Used in Simulation

α	β_1	β_2	β_3
0.60	.575	.573	.583
0.70	.576	.574	.582
0.80	.576	.575	.581
0.90	.577	.576	.580
0.95	.577	.577	.579
0.99	.577	.577	.578

Combining the six coefficients of α with six standard deviations of σ, thirty-six sets of data are generated. The values of standardized coefficients and total mean square are computed based on equations 19, 20, 21, 22, 23, and 24.

As mentioned in previous sections of Monte Carlo simulation, the accuracy of estimations can be improved by increasing the sample size so that additional 50 samples of observations with n = 100 and p = 3 are generated to each of the 36 different sets of data. The independent variables and true coefficient β are unchanged, while the random error term is varied, so that the dependent variable is changed. The average of optimum k for OPT and AOPT; and average mean square for OLS, OPT and AOPT in these 50 samples are computed. The results are also listed in Table 2. The measure of M* and M** used to compare the different methods are computed based on equations 25 and 26. The results are also listed in Table 2.

RESULTS AND DISCUSSIONS

As the example given in previous sections the results of the performance of MSE in each method of OLS, OPT and AOPT with α = 0.6, 0.7, 0.8, 0.9, 0.95, and 0.99 corresponding to the six error terms σ were plotted on Figures 1, 2, 3, 4, 5, and 6, respectively. The following conclusions can be drawn.

First, when the correlation coefficient between independent variables are less than 0.5 (i.e. the cases of α = 0.6 and 0.7 as shown in Figure 1 and 2), the MSE of OLS is close to the ridge estimators of OPT and AOPT. The deviation of result between OPT and AOPT is also negligible. As Table 2 shows, the value of M* and M** are close to 1 when the r value equal 0.36 and 0.49. This implies that the OLS method is as good as OPT and AOPT methods when the correlation coefficient is less than 0.5.

Second, when the correlation coefficient exist between 0.6 and 0.8 (i.e. the cases of α = 0.8 and 0.9 plotted on Figures 3 and 4), the MSE of OLS is much larger than OPT and AOPT, and OPT is much smaller than AOPT. As Table 2 shows, the value of M* is much less than M**. These imply that the ridge regression analysis is required when correlation coefficients exist between 0.6 and 0.8, and the OPT method is much better than AOPT method.

Third, when the correlation coefficient is greater than 0.9 (i.e. the cases of α = 0.95 and 0.99 as shown in Figures 5 and 6), the MSE of OPT is much larger than both OPT and AOPT methods and the performance of the two ridge type estimators gave an approximate same solution. As Table 2 shows, the value of M* is similar to the M** when α equals 0.95 and 0.99. These imply that the ridge analysis is required when the correlations is greater than 0.9 and both OPT and AOPT methods give a similar solution.

Fourth, all M* and M** in Table 2 are decreasing while error terms σ are increasing. This implies that higher the error term σ the better ridge estimator is performed. For instance, M* equals .972 when σ = 0.01 and equals .379 when σ = 1.01 for the data with α = .95. This means that the MSE of OPT is 97% of OLS when error is 0.01, but it is only 38% of the OLS when error becomes 1.01.

Fifth, as Table 2 shows, the value of M* is much smaller than M** and M** is smaller than or equal to one. This concludes that the performance of OPT is better than AOPT, and the AOPT is better than OLS in terms of minimizing the mean square error of estimation in regression analysis to solve the multicollinearity among the independent variables.

SUMMARY AND CONCLUSIONS

The estimation of regression coefficients in multiple linear regression can present problems when multicollinearity exists among independent variables. This type of problem has been solved by one of the statistical methods called ridge regression. This technique shows that by adding a non-negative constant "k" to the diagonal of correlation matrix it is possible to substantially reduce error variance of estimators. The methods of optimum ridge coefficients (OPT) and approximate ridge estimators (AOPT) are used in this study to compare the performance of each technique with the ordinary least square (OLS) estimators.

The Monte Carlo simulations are used to generate the observations of independent variables. The simulations are performed based on different correlation coefficients and error terms. Comparisons of the AOPT and OPT methods are made with OLS technique. The results indicated that when correlation between two independent variables is less than 0.5, the OLS performs as good as ridge regression, i.e., the multicollinearity problem does not exist in this case. But, when correlation lies aroun 0.6 and 0.8, OPT method is considered the best among those three methods, and AOPT method is better than OLS method. For those data with high correlation such as 0.9,

TABLE 2: Comparisons of the Simulation Results of AOPT and OPT Methods with OLS Method in Different Correlation Coefficient and Standard Deviations.

α	Correl. Coeffi. r	Stand. Devia. σ	OLS L(0)	OPT		AOPT		M*	M**
				L(B*(k))	Ave. k	L(B**(k))	Ave. k		
0.60	0.36	0.01	0.000	0.000	0.000	0.000	0.000	1.000	1.000
		0.21	0.026	0.026	0.106	0.025	0.009	1.000	0.962
		0.41	0.081	0.068	0.212	0.072	0.034	0.834	0.891
		0.61	0.224	0.164	0.273	0.183	0.071	0.733	0.816
		0.81	0.271	0.193	0.375	0.204	0.122	0.694	0.754
		1.01	0.596	0.413	0.384	0.440	0.184	0.693	0.739
0.70	0.49	0.01	0.000	0.000	0.000	0.000	0.000	1.000	1.000
		0.21	0.031	0.029	0.122	0.029	0.010	0.954	0.948
		0.41	0.096	0.072	0.227	0.082	0.037	0.746	0.856
		0.61	0.264	0.174	0.277	0.205	0.075	0.658	0.774
		0.81	0.319	0.200	0.376	0.222	0.129	0.627	0.693
		1.01	0.703	0.428	0.377	0.476	0.188	0.609	0.677
0.80	0.64	0.01	0.000	0.000	0.000	0.000	0.000	1.000	1.000
		0.21	0.041	0.033	0.136	0.038	0.011	0.805	0.927
		0.41	0.127	0.074	0.224	0.101	0.039	0.583	0.798
		0.61	0.349	0.189	0.276	0.242	0.079	0.541	0.693
		0.81	0.418	0.219	0.375	0.255	0.135	0.524	0.610
		1.01	0.926	0.404	0.353	0.548	0.186	0.501	0.592
0.90	0.81	0.01	0.000	0.000	0.001	0.002	0.000	1.032	1.000
		0.21	0.073	0.033	0.156	0.063	0.012	0.452	0.863
		0.41	0.273	0.133	0.200	0.190	0.042	0.487	0.693
		0.61	0.445	0.158	0.257	0.234	0.081	0.355	0.524
		0.81	0.872	0.368	0.324	0.437	0.137	0.422	0.501
		1.01	1.301	0.509	0.357	0.582	0.185	0.391	0.447
0.95	0.91	0.01	0.000	0.000	0.010	0.000	0.000	0.972	1.000
		0.21	0.136	0.055	0.122	0.106	0.013	0.402	0.780
		0.41	0.415	0.095	0.169	0.247	0.040	0.230	0.523
		0.61	1.147	0.437	0.177	0.542	0.072	0.381	0.472
		0.81	1.352	0.434	0.250	0.507	0.120	0.321	0.375
		1.01	2.987	1.133	0.194	1.201	0.128	0.379	0.402
0.99	0.98	0.01	0.002	0.001	0.031	0.002	0.000	0.795	0.800
		0.21	0.641	0.231	0.086	0.332	0.010	0.357	0.517
		0.41	1.960	0.495	0.100	0.606	0.026	0.253	0.309
		0.61	5.454	1.968	0.079	1.950	0.036	0.361	0.357
		0.81	6.418	1.488	0.130	1.797	0.066	0.226	0.281
		1.01	14.092	4.971	0.104	4.763	0.053	0.352	0.338

both OPT and AOPT methods are good techniques to solve the multicollinearity in linear regression models.

REFERENCES

Althauser, R.P. (1971). "Multicollinearity and Non-Additive Regression Models," pp. 453-472 in H.M. Blalock, ed., Causal Models in the Social Sciences," Chicago: Aldine.

Blalock, H.M., Jr. (1963). "Correlated Independent Variables: The Problem of Multicollinearity." Social Forces 42:233-237.

_____. (1964). Causal Inferences in Nonexperimental Research. Chapel Hill: University of North Carolina Press.

Christ, C.F. (1966). Econometric Models and Methods. New York: Wiley.

Deegan, J., Jr. (1975). "The Process of Political Development.: Sociological Methods and Research 3:384-415.

Dempster, A.P., M. Schatzoff, and N. Wermuth. (1977). "A Simulation Study of Alternative to Ordinary Least Square." Journal of the American Statistical Association 72:77-90.

Goldberger, A.S. (1964). Econometric Theory. New York: Wiley.

Gordon, R.A. (1968). "Issues in Multiple Regression." American Journal of Sociology 73:592-616.

Hoerl, A.E. and R.W. Kennard (1970a). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." Technometrics 12:55-68.

Hoerl, A.E. and R.W. Kennard (1970b). "Ridge Regression: Applications to Nonorthogonal Problems." Technometrics 12:69-82.

Hoerl, A.E., R.W. Kennard, and K.F. Baldwin (1975). "Ridge Regression: Some Simulations." Communications in Statistics 4(2):105-123.

Johnston, J. (1972). Econometric Methods, 2nd ed. New York: McGraw-Hill.

McDonald, G.C. and D.I. Galarneau (1973). "A Monte Carlo Evaluation of Some Ridge Type Estimators," Publication GMR-1322-B, General Motors Research Laboratories, Warren, Michigan.

McDonald, G.C. and R.C. Richard (1973). "Instabilities of Regression Estimates Relating Air Pollution to Mortality." Technometrics 15, 15:463-481.

Newhouse, J.P. and S.D. Oman (1971). "An Evaluation of Ridge Estimators," Rand Report No. R-716-PR:1-28.

Rockwell, R.C. (1975). "Assessment of Multicollinearity." Sociological Methods and Research 3:308-320.

Shih, S.F. and R.L. Hamrick (1974). "A Technique Used to Determine Random Point Position." Water Resource Bulletin 10(5):884-898.

Shih, W.F.P. and J.D. Kasarda (1977). "Optimal Bias in Ridge Regression Approach to Multicollinearity." Sociological Methods & Research, 5:461-470.

Vinod, H.D., (1976). "Application of New Ridge Regression Method to A Study of Bell System Scale Economics." Journal of the American Statistical Association 71:835-841.

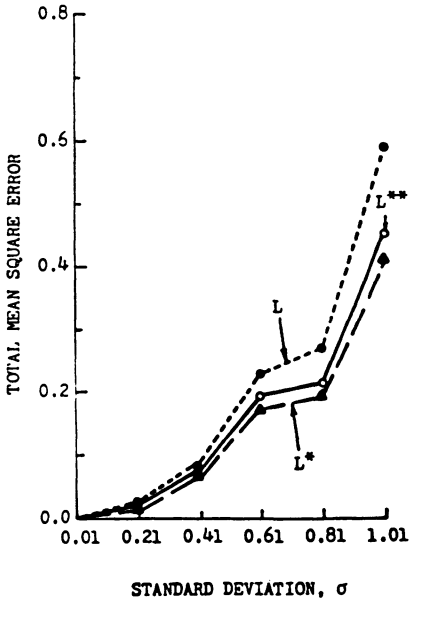


Fig. 1. Total mean square error related to standard deviation, $\alpha = 0.6$.

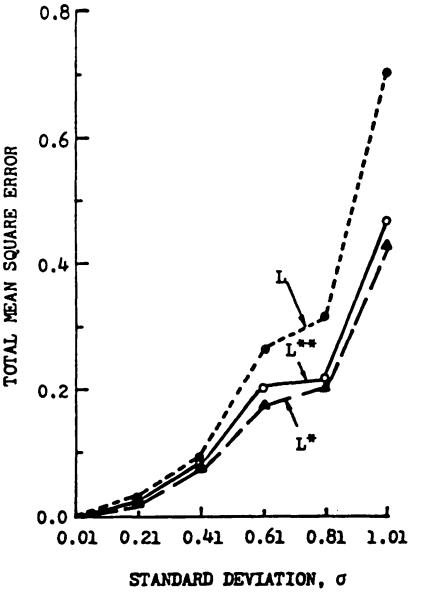


Fig. 2. Total mean square error related to standard deviation, $\alpha = 0.7$.

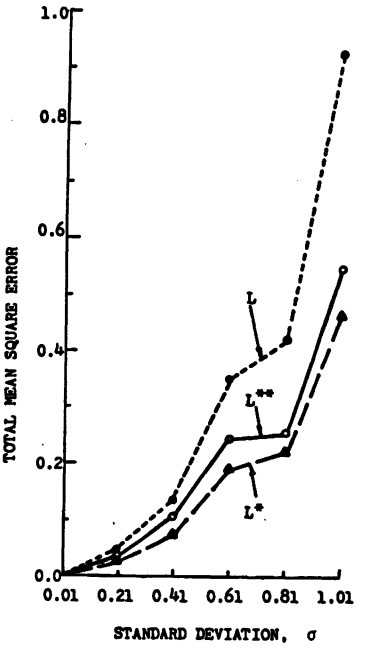


Fig. 3. Total mean square error related to standard deviation, α = 0.8.

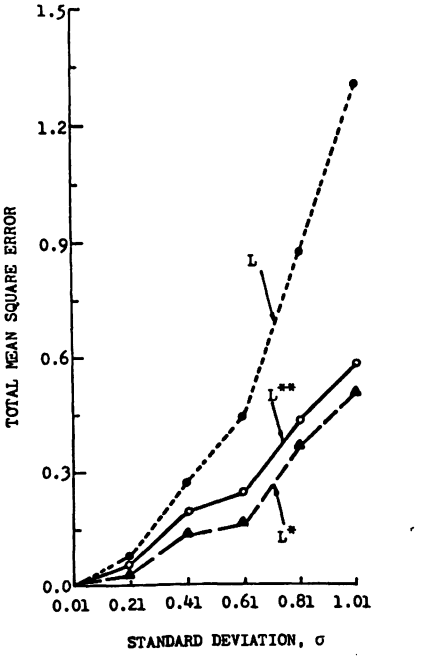


Fig. 4. Total mean square error related to standard deviation, α = 0.9.

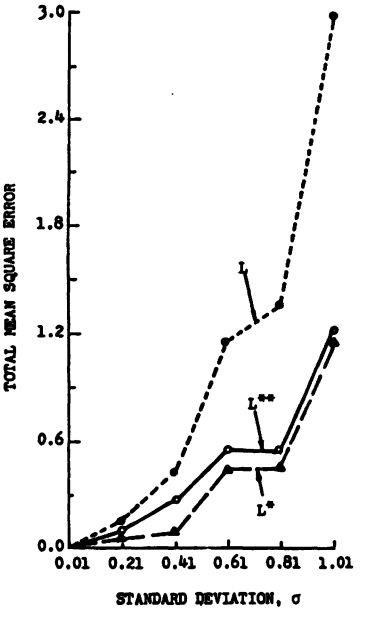


Fig. 5. Total mean square error related to standard deviation, α = 0.95.

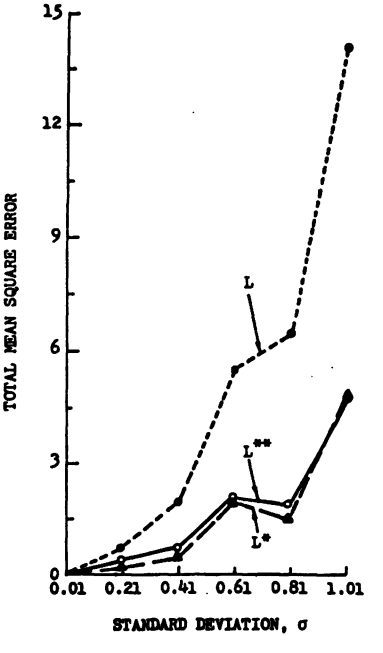


Fig. 6. Total mean square error related to standard deviation, α = 0.99.

EVALUATING BEHAVIORS IN NATURALISTIC SETTINGS: ISSUES OF RELIABILITY, VALIDITY, AND OBSERVER BIAS

Sara S. Sparrow, Yale University
Domenic V. Cicchetti, V.A. Hospital, West Haven, Connecticut
JoAnn Robinson, Cornell University

Introduction

The purpose of this paper is to discuss the issues of reliability, experimenter bias, and validity as they apply to various aspects of the behavior of moderately, severely, and profoundly retarded, institutionalized children. As noted recently by Longabaugh (1977), only a few published studies have even considered these crucial variables. Instead, the observations of single experimenters have usually been assumed to represent a valid assessment of a wide range of behavior occurring in a variety of naturalistic settings. It is our contention that this viewpoint is inconsistent with the reports of broad bodies of literature in both medicine and the behavioral sciences. Specifically, a number of clinical investigators have pointed to the fact that observer variability is an essentially ubiquitous phenomenon spanning multiple and diverse areas of medicine (e.g., Cicchetti, 1977: Cicchetti & Conn, 1976; Etter, Dunn, Kammer, Osmond, & Reese, 1960; and Koran, 1975a and 1975b). In fact, it is rather common for *inter*observer disagreement in medical diagnoses to range between about 25-30%, and for *intra*observer variation to reach proportions of 15-20%. To mention a second major area of clinical investigation, Helzer and associates (1977) report much higher overall agreement in the assessment of neuropsychiatric diagnosis than previous studies. Nonetheless, the extent of interobserver disagreement on *specific* categories of illness showed considerable variability (e.g., 29% disagreement in diagnosing for alcoholism). The same general results occur in the field of mental retardation. For example, Balthazar reports generally acceptable levels of interrater agreement in assessing various behaviors of mentally retarded children. Yet in one reliability study, independent observers agreed only between 42% and 69% in the rating of 16 of 64 areas of behavioral assessment (Balthazar, 1973).

Sources of the Data

This report focuses upon the behavior of mentally retarded children as it occurs and develops in naturalistic institutionalized settings. The data derive more generally from a longitudinal investigation of the effects of a sensorimotor patterning treatment program on the behavior of mentally retarded children in residence at the Seaside Regional Center in New London, Connecticut. Specific sources of data, as displayed in Figure 1, are based upon: (1) Levels of cognitive, psychomotor, social, and self-control behavior, as measured by the Behavior Rating Inventory for the Retarded (BRIR), due to Sparrow and Cicchetti (1977); (2) Results of standardized IQ tests, such as the Catell and the Stanford Binet; (3) Results of performance on unstandardized tests, constructed by the senior author (which included assessment of motor and language development); and (4) Direct behavioral observations (assessing levels of affect, communication, activity, and play). Since the data arising from the first three sources will appear in future publications, this report will be based mainly upon data derived from direct observations of the behavior of mentally retarded children.

Analyses of the Data

In a recent study (Cicchetti, 1977) the issues of reliability, bias, and validity were discussed in the context of medical investigations. This report will focus upon these issues in the field of mental retardation.

Reliability

When we speak of observer reliability, we are concerned with the extent to which independently derived measurements or judgments agree or are interchangeable one with the other. Reliability can be assessed either between two or more independent observations of the same phenomenon (*inter*observer reliability) or within the same observer (*intra*observer reliability). Further, with respect to qualitative data, we can speak in terms of either *overall* agreement or *specific* agreement. Thus, in our research, we assessed interobserver reliability in rating the *communication* level of a group of 49 mentally retarded children, on a six category ordinal scale, as one of the following: (1) none: (2) prespeech sounds only: (3) gestures or sounds; (4) talking to self; (5) noncommunicative speech; or (6) echolalic speech. Using ordinal weighting systems developed by Cicchetti (1976) with the weighted kappa statistic due to Cohen and colleagues (e.g., Cohen, 1968; and Fleiss, Cohen & Everitt, 1969), we assessed both overall observer agreement as well as interobserver *specific* agreement for each of the six categories of the scale. The formulae for the specific agreement indices were recently developed by Cicchetti, Fontana, and Noel Dowds (1977) and are available upon request. The results in Table 1 show that the overall level of agreement is extremely high, even when corrected for the amount of agreement expected by chance alone. Thus, we obtained 95.69% observer agreement (PO); as against 76.56% expected by chance (PC). The level of chance-corrected agreement or kappa (PO-PC)/(1-PC) was .82, with +1 representing perfect chance-corrected agreement. It is interesting to note that the indices of specific rater agreement are also quite respectable, with PO values ranging between 91.67% and 100% agreement, and chance-corrected or specific kappa values ranging between .66 and 1.00. (The value of .66, it should be noted, is consistent with data presented by Koran (1975a; 1975b) for a wide range of clinical judgments, across many diverse fields of medical diagnosis.)

A second variable, level of play, was independently observed in 30 mentally retarded children, and was assessed by a four category ordinal scale as one of the following: (1) does not play at all; (2) plays with non-toy object(s); (3) uses toy(s) inappropriately; or (4) uses toy(s) as intended. These data are given in Table 2 and once again show very high levels of overall agreement, specific agreement and chance-corrected agreement. Thus, PO and overall kappa values are 97.33% and .92, respectively, while specific agreement indices (SO) range between 95.00% and 100%. Chance-corrected specific agreement levels range between .78 and 1.00.

A third variable which independent raters observed was affect, scored as 1 = no or negative affect: and 2 = positive affect. Thirty-nine children were available for assessment on this variable. Results in Table 3 showed that, consistent with the data for communication and play, overall interobserver agreement was high (PO = 94.87%; overall kappa = .72; SO (negative affect) and SO (positive affect) were 97.14% and 75%, with chance-corrected agreement being .72 in each case).

Finally, we assessed levels of interobserver agreement for level of physical activity which could be rated as: 1 = sleeping or no movement; 2 = prone with some movement: 3 = sitting in wheelchair: 4 = sitting or kneeling; 5 = standing; 6 = crawling or creeping; 7 = walking; and 8 = running. As for each of the other variables, overall levels, as given in Table 4, were very high (PO = 99.22%; Overall kappa = .93; SO values ranged between 92.31% and 100%; and chance-corrected specific kappa values ranged between .73 and 1.00).

Observer Bias

The question of observer bias is one of the extent to which one observer evaluates a given phenomenon systematically differently than other observers who have independently assessed the same phenomenon. Thus, to the extent that agreement is very high, and disagreements tend to occur in an essentially random pattern, observer bias does not occur. However, when it does occur it suggests that the observers are not always using the same frames of reference to make the same judgments. As noted by a number of investigators, Longabaugh (1977); Johnson and Bolstad (1973); and Reid (1970), even well-trained observers whose reliability has not been assessed periodically may become biased with respect to their judgments. This phenomenon is referred to as either *observer drift* or *instrument decay* (Campbell & Stanley, 1966). As an example of how pervasive the phenomenon can become, one study reports a drop from 70% to 51% in the extent of observer agreement levels as a function of observer drift or instrument decay (Reid, 1970).

With respect to our longitudinal investigation of mentally retarded children, we made periodic checks upon the reliability of our rater pairs but fortunately found essentially no levels of observer drift which were of clinical concern. At least three plausible reasons why we did not observe a phenomenon which others doing naturalistic observations did indeed experience include the following: (1) The observers were very carefully trained in the use of our rating techniques (both standard and nonstandard): (2) The behaviors we rated were very carefully defined into nonoverlapping categories of classification: and (3) The reported levels of instrument decay cited in the literature were based upon observations of nonretarded samples whose range of expression of behavior tends, in the main, to be more varied, less stereotypic, and hence less clearly delineated than the subjects we refer to in our research.

Validity

The phenomenon of validity is one of answering the question: Does our measuring instrument indeed measure what it purports to measure? There are many different types of validity measures, and these have been discussed by numerous authors, including the following: Balthazar and English (1969); Bechtoldt (1959); Cicchetti (1977): Cronbach (1960 and 1971): French and Michael (1966); Greenwood and Perry (1968); Guion (1974); Nihira (1976); and Nihira, Foster, Shellhaas, and Leland (1974). Some of the more familiar types of validity assessment reported in the literature include: (1) content validity; (2) criterion related validity; (3) construct validity; and (4) factorial validity. The paper by Guion (1974) is an excellent reference for a detailed and comprehensive description of the first three types of validity assessment. Most of these were utilized recently by Sparrow and Cicchetti (1977) in their assessment of the validity of the aforementioned Behavior Rating Inventory for the Retarded (BRIR). As one method of assessment, we used factorial validity, a technique utilized by several investigators in the field of mental retardation (e.g., Balthazar & English, 1969: and Nihira, Foster, Shellhaas, & Leland, 1974). As a result of that experience, we strongly recommend that investigators contemplating using this form of validity assessment heed the following advice (which has *not* as a rule been reported in the mental retardation literature): (1) It is preferable to have *a priori* "factors" against which to compare the empirically derived factors. (2) It is wise to use more than one type of major factor analytic technique (e.g., principal components, principal factors, each with orthogonal and oblique rotations). The purpose of this suggestion is to determine the extent to which different techniques might affect the particular factors obtained (following upon the advice of Frane & Hill, 1975). It was our experience, for example, that a principal components, oblique rotation solution produced fewer BRIR items which overlapped two or more factors than did other types of analyses. (3) It is important to report the percentage of overall variance accounted for by the factor analysis.

In summary, this paper has attempted to discuss the issues of observer reliability, observer bias (or drift), and validity in the context of the behavior of institutionalized retarded

children. Although the high levels of reliability achieved in our sample are somewhat at a variance with the assessment of clinical phenomena based upon nonretarded samples, the central issues discussed here appear to have a broad range of applicability to behavioral science, medicine, and other fields of clinical investigation.

As a final note, computer programs for assessing levels of observer reliability and bias are available upon request.

References

Balthazar, E.E. *The Balthazar scales of adaptive behavior. Section II. The scales of social adaptation (BSAB-II).* Palo Alto, California: Consulting Psychologists Press, 1973.

Balthazar, E.E. & English, G.E. A factorial study of unstructured ward behaviors. *American Journal of Mental Deficiency*, 1969, *74*, 353-360.

Bechtoldt, H.P. Construct validity: A critique. *American Psychologist*, 1959, *14*, 619-629.

Campbell, D.T. & Stanley, J.C. *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally, 1966.

Cicchetti, D.V. Assessing inter-rater reliability for rating scales: Resolving some basic issues. *British Journal of Psychiatry*, 1976, *129*, 452-456.

Cicchetti, D.V. Assessing observer and method variability in medicine. To appear in *Connecticut Medicine*, 1977 (by invitation).

Cicchetti, D.V. & Conn, H.O. A statistical analysis of reviewer agreement and bias in evaluating medical abstracts. *Yale Journal of Biology and Medicine*, 1976, *49*, 373-383.

Cicchetti, D.V., Fontana, A.F., & Dowds, B. Noel. Assessing specific category reliabilities for rating scales in behavioral research. Paper presented at meeting of the American Psychological Association, San Francisco, California, August 1977.

Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Cronbach, L.J. Need for critical evaluation of tests. In L.J. Cronbach (2nd ed.) *Essentials of Psychological Testing.* New York: Harper & Row, 1960, pp. 96-125.

Cronbach, L.J. Test validation. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Etter, L.E., Dunn, J.P., Kammer, A.G., Osmond, L.H., & Reese, L.C. Gastroduodenal X-ray diagnosis: A comparison of radiographic technics and interpretations. *Radiology*, 1960, *74*, 766-770.

Fleiss, J.L., Cohen, J., & Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Frane, J.W. & Hill, M.A. Annotated computer output for factor analysis: A supplement to the writeup for computer program BMDP4M. In W.J. Dixon (Ed.) *BMDP Biomedical Computer Programs.* Los Angeles: University of California Press, 1975.

French, J.W. & Michael, W.B. (and the joint committee of the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education). *Standards for Educational and Psychological Tests and Manuals.* Washington, D.C.: American Psychological Association, 1966, pp. 1-40.

Greenwood, D. & Perry, R. Use of the Adaptive Behavior Checklist as a means of determining unit placement in a facility for the retarded. A paper presented at the meeting of the Rocky Mountain Psychological Association, Denver, Colorado, May 1968.

Guion, R.M. Open a new window: Validities and values in psychological measurement. *American Psychologist*, 1974, *29*, 287-296.

Helzer, J.E., Clayton, P.J., Pambakian, R., Reich, T., Woodruff, R.A., & Reveley, M.A. Reliability of psychiatric diagnosis. II. The test/retest reliability of diagnostic classification. *Archives of General Psychiatry*, 1977, *34*, 136-141.

Johnson, S.M. & Bolstad, O.D. Methodological issues in naturalistic observation: Some problems and solutions for field research. In L.A. Hamberlynch, L.C. Handy, & E.J. Mash (Eds.), *International Conference on Behavior Modification, Behavior Change.* Champaign, Illinois: Research Press, 1973.

Koran, L.M. The reliability of clinical methods, data and judgments. *The New England Journal of Medicine*, 1975, *293*, 642-646 (First of Two Parts).

Koran, L.M. The reliability of clinical methods, data and judgments. *The New England Journal of Medicine*, 1975, *293*, 695-701 (Second of Two Parts).

Longabaugh, R. The systematic observation of behavior in naturalistic settings. Manuscript, in preparation, 1977.

Nihira, K. Dimensions of adaptive behavior in institutionalized mentally retarded children and adults: Developmental perspective. *American Journal of Mental Deficiency*, 1976, *81*, 215-226.

Nihira, K., Foster, R., Shellhaas, M., & Leland, H. *AAMD Adaptive Behavior Scale*, 1974 revision. Washington, D.C.: American Association on Mental Deficiency, 1974.

Reid, J.B. Reliability assessment of observation data: A possible methodological problem. *Child Development*, 1970, *41*, 1143-1150.

Sparrow, S.S. & Cicchetti, D.V. The behavior rating inventory for the retarded (BRIR): A scale applicable to moderate, severe, and profound retardation. To appear in *American Journal of Mental Deficiency*, (January) 1978.

FIGURE 1

BEHAVIORAL ASSESSMENT OF MENTALLY RETARDED CHILDREN

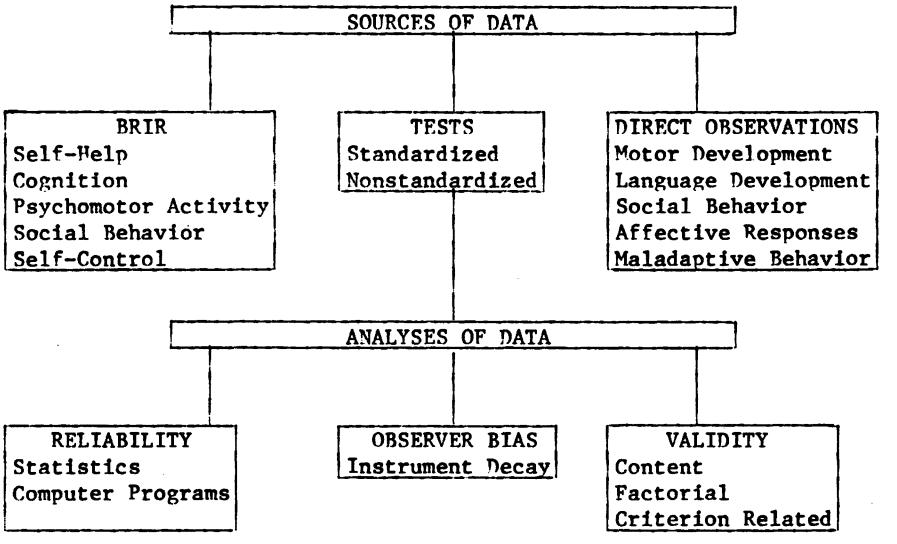


TABLE 1

OBSERVER AGREEMENT IN RATING THE HIGHEST LEVEL OF COMMUNICATION OF SERIOUSLY RETARDED CHILDREN

Category of Communication	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) None	.51	.9733	.8118	.86
(2) Prespeech Only	.31	.9407	.8254	.66
(3) Gestures or Sounds	.06	.9259	.7120	.74
(4) Talking to Self	.08	.9167	.5420	.82
(5) Noncommunicative Speech	.02	1.0000	.3447	1.00
(6) Echolalic Speech	.02	1.0000	.1882	1.00
Entire Scale	1.00	.9569	.7656	.82

TABLE 2

OBSERVER AGREEMENT IN RATING THE HIGHEST LEVEL OF PLAY ACTIVITY OF SERIOUSLY RETARDED CHILDREN

Category of Play	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) No Play	.13	1.0000	.3567	1.00
(2) Play with Non-Toys	.28	.9882	.7035	.96
(3) Inappropriate Play with Toys	.27	.9500	.7725	.78
(4) Appropriate Play with Toys	.32	.9684	.6435	.91
Entire Scale	1.00	.9733	.6567	.92

TABLE 3

OBSERVER AGREEMENT IN RATING AFFECT LEVELS OF SERIOUSLY RETARDED CHILDREN

Category of Affect	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) None or Negative	.10	.7500	.1026	.72
(2) Positive	.90	.9714	.8974	.72
Entire Scale	1.00	.9487	.8159	.72

TABLE 4

OBSERVER AGREEMENT IN RATING HIGHEST LEVEL OF ACTIVITY OF SERIOUSLY RETARDED CHILDREN

Category of Activity	Average Rater Frequency of Application	Index of Observer Agreement		
		Obtained	Expected	Chance-Corrected
(1) Sleeping or No Movement	.02	1.0000	.4301	1.00
(2) Prone with Some Movement	.00	NA[1]	NA	NA
(3) Sitting in Wheelchair	.04	1.0000	.7975	1.00
(4) Sitting or Kneeling	.34	.9930	.9096	.92
(5) Standing	.41	.9923	.9356	.88
(6) Crawling or Creeping	.18	.9915	.8587	.94
(7) Walking	.01	.9231	.7174	.73
(8) Running	.00	NA	NA	NA
Entire Scale	1.00	.9922	.8946	.93

[1]Note. NA denotes not applicable

SAMPLE-SIZE DETERMINATION AND INFERENTIAL STATISTICAL TECHNIQUE

Vernon W. Stone and Russell W. Irvine, Georgia State University

INTRODUCTION

More frequently it is the rule, rather than the exception, for a researcher to report the use of a given sample size without indicating the basis for the determination thereof. Just as we expect to be apprised of the research design and the statistical design, we should expect sufficient information with respect to the sampling design, of which the determination of sample size is a fundamental dimension.

This presentation may be considered an initial taxonomical effort of sample-size formulae. However, the formulae presented, of course, constitute nothing more than a sample of research situations.

Part I, "Selected Sample-Size Computational Approaches," is predicated upon a priori data, affording, for the most part, the direct sample-size determinations. Included therein are formulae for the estimation of the population mean, the estimation of the population proportion, determination of sample size for t-test, determination of sample size for X^2, and determination of sample size for the F-test. Some direct and indirect duplication of formulae are to be observed. Numerical values may be substituted for the purpose of demonstrating the means of attaining the desired applicatory results.

Part II, "Selected Sample-Size Tabular Approaches," is predicated upon a posteriori data, affording, for the most part, the indirect sample-size determinations. Included therein are formulae for the estimation of the population mean, the estimation of the population proportion, the determination of sample size for t-test, the determination of sample size for X^2, the determination of sample size for the F-test, and the determination of sample size for r, including correlation coefficients based upon the continuous interval scale.

It must be recognized that the Type I and II Errors may be either one-tailed or two-tailed, that is, one side or two sides. The Type I Error refers to alpha (α), the probability of erroneously rejecting the null hypothesis. Accordingly, one minus alpha ($1-\alpha$) represents the probability of not rejecting the null hypothesis, that is, not making a Type I Error. The Type II Error refers to beta (β), the probability of erroneously accepting the null hypothesis. Accordingly, one minus beta ($1-\beta$) represents the probability of not accepting the null hypothesis, that is, not making a Type II Error. The standard normal deviates (z_α and z_β, as well as t_α and t_β), are employed, respectively, for indicating the probabilities of the two errors. The use of t values, however, usually requires iterative stabilization.

It should be noted that more and more emphasis is being placed upon the use of the power function ($1-\beta$) in hypothesis testing. Traditionally, there has been a focus on the significance criterion (α), while ignoring power ($1-\beta$). Some statisticians, of course, are of the convincing opinion that the consideration of power results in an abundance of liberality, for significance is more likely to be uncovered on the basis of the amount of effort put forth rather than on the basis of an empirically pragmatic meaningfulness. On the other hand, reputable statisticians maintain that the traditional approach--focusing on only alpha--results in a desirable degree of conservatism. While one may tend to align himself or herself definitively on the side of one of the positions, it must be remembered that sampling is served by the inclusion of the power consideration, for such does result in the determination and use of a larger sample size.

The finite population correction (fpc) is basically employed when n represents 5 percent or more of the population. When the percent is less than 5 percent, the effect on the sample size is negligible. In this regard, however, a defendable position is that fpc can profitably be applied in connection with all finite populations, thereby elevating the finite population to an infinite population.

A sample size is, at best, a tentative, operative estimate. Accordingly, it must be recognized and understood that various and sundry approaches to sample size are, can be, and should be employed. The basic consideration in this regard is the rationale for the use, justification, and defense thereof. The formulae relating, for example, to the estimation of the population mean and the population proportion are cases in point. The former focuses on measurement (the amount on a continuous basis), and the latter focuses on counting (the number on a discontinuous, or discrete, basis). To the extent that the population estimations can be justified--whether or not such estimations are actually effected--one of the foregoing estimation formulae may be in order, assuming, of course, random selection from a finite population. Hence, the basic consideration should not be a parametric estimation in fact; it should be a parametric estimation in theory.

Finally, it must be emphasized that a researcher is not, ordinarily, cognizant of the specific statistical technique(s) which will ultimately be employed. Since n must be known in order to collect the desired data, n must, usually, be known prior to the decision with respect to statistical techniques in the data analysis. That is to say, a given set of data can and will lend itself to the use of optional statistical techniques. Therefore, a means of determining initial sample size--prior to the decision to use the t-test, for example--must be available.

I. SELECTED SAMPLE-SIZE COMPUTATIONAL APPROACHES

A. ESTIMATION OF THE POPULATION MEAN

1. Type I Error without fpc

$$n = \left(\frac{z\sigma}{E}\right)^2$$

2. Type I Error with fpc

$$n = \frac{N(z\sigma)^2}{(NE^2) + (z\sigma)^2}$$

3. Type I and II Errors without fpc

$$n = \left(\frac{[z\alpha + z\beta]\sigma}{E}\right)^2$$

4. Type I and II Errors with fpc

$$n = \frac{N([z\alpha + z\beta]\sigma)^2}{(NE^2) + ([z\alpha + z\beta])^2}$$

5. Type I Error without fpc (requiring iteration for stability)

$$n = \left(\frac{t\sigma}{E}\right)^2$$

B. ESTIMATION OF THE POPULATION PROPORTION

1. Type I Error without fpc

$$n = \left(\frac{z^2 pq}{E^2}\right)$$

2. Type I Error with fpc

$$n = \frac{N(z^2 pq)}{(NE^2) + (z^2 pq)}$$

3. Type I and II Errors without fpc

$$n = \left(\frac{[z\alpha + z\beta]^2 pq}{E^2}\right)$$

4. Type I and II Errors with fpc

$$n = \frac{N([z\alpha + z\beta]^2 pq)}{(NE^2) + ([z\alpha + z\beta]^2 pq)}$$

5. Sokal and Rohlf (15, 17); cf. I-D-5

$$n = \frac{\text{Tabular Value (Table 1)}}{\text{(Angular Transformation Squared)}} = \frac{\alpha/1-\beta}{\delta^2}$$

C. DETERMINATION OF SAMPLE SIZE FOR t-TEST

1. Lacey (8) -- derived from t-test formula

$$n = \frac{2(s^2 t^2)}{D^2}$$

2. Walker and Lev (19) -- $\underline{d}$ represents the departure from hypothesis which it is desired to detect

$$n = \left(\frac{\sigma}{d}(z\alpha + z\beta)\right)^2 = \frac{\sigma^2}{d^2}(z\alpha + z\beta)^2$$

3. Walker and Lev (19) -- using equal-size samples and determining $\underline{n}$ for each sample

$$n_1 = \frac{\sigma_1^2 + \sigma_1\sigma_2}{d^2}(z\alpha + z\beta)^2$$

$$n_2 = \frac{\sigma_2^2 + \sigma_1\sigma_2}{d^2}(z\alpha + z\beta)^2$$

4. Hadley (5) -- Type I and II Errors -- $\underline{n}$ is sample size for one sample; $\underline{2n}$ is the sample size for both

$$n = 2\sigma^2(z\alpha + z\beta)^2 / \mu_1$$

5. Dixon and Massey (2) -- derived from $\underline{z}$ formula -- Type I Error

$$n = \frac{2(sz)^2}{D^2}$$

6. Dixon and Massey (2) -- derived from $\underline{z}$ formula -- Type I and II Errors

$$n = \frac{2([z\alpha + z\beta]s)^2}{D^2}$$

7. Marascuilo (10) -- Type I Error

$$n = \frac{2(st)^2}{D^2}$$

8. Marascuilo (10) -- Type I and II Errors

$$n = \frac{(t\alpha + t\beta)^2(2s^2)}{D^2}$$

9. Sokal and Rohlf (15, 17) -- Type I and II Errors without fpc (cf. I-D-5)

$$n = \frac{\text{Tabular Value (Table 1)}}{\text{(Angular Transformation Squared)}} = \frac{\alpha/1-\beta}{\delta^2}$$

D. DETERMINATION OF SAMPLE SIZE FOR X^2

1. Lacey (8) -- anticipated observed percentages vs. theoretical percentages

$$X^2 = \frac{(.8n - .5n)^2}{.5n} + \frac{(.2n - .5n)^2}{.5n} = .36n$$

$n = 6.635$ (1df, 1% level) / $.36n$

$= 18.43$ (19)

2. Lacey (8) -- anticipated observed percentages vs. theoretical percentages for 2 X 2 Control and Experimental Groups

	Failed	Passed	Totals
Control	.159n	.841n	n
	(.113n)	(.887n)	
Exp'l	.067n	.933n	n
	(.113n)	(.887n)	
	.226n	1.774n	2n

$$\chi^2 = \frac{(.046n)^2}{.113n} + \frac{(.046n)^2}{.113n} + \frac{(.046n)^2}{.887n} + \frac{(.046n)^2}{.887n} = .042222n$$

$n = 3.841$ (1 df, 5% level) / $.042222n = 90.97$ (91)

3. Lacey (8) -- assuming results of dichotomous survey with 10% maximum fiducial range

Owners	Non-owners	Total
.7n	.3n	n

$$\chi^2 = \frac{(.05n)^2}{.7n} + \frac{(.05n)^2}{.3n} = .0119n$$

$n = 6.635$ (1 df, 1% level) / $.0119n = 557.56$ (558)

4. Krejcie, et al. (7) and NEA (16) -- Cf. I-B-2

$$n = \chi^2 N pq \div d^2(N-1) + \chi^2 pq$$

5. Sokal and Rohlf (15, 17) -- Although the following is employed for detecting a true difference between two given percentages, the approach is applicable to natural and artificial dichotomies or rows and/or columns. The rationale is predicated upon the noncentrality parameter discerned from Table 2.

Assume that alpha is 5% and power is .80; moreover, that p_1 is 0.65 and p_2 is 0.55. By means of Rohlf and Sokal's Table K for angular transformation, the two proportions are converted to arcsines, angles, in degrees, whose sines correspond to the values given.

The value from Table 2 is 12,884.8; and delta square is $(53.73 - 47.87)^2 = 5.86^2 = 34.3396$

$$n = \frac{12,884.8}{34.3396} = 375.21 \quad (376)$$

2n (for two samples) = 752

N.B.: (1) When one of the percentages is theoretical, divide by two delta square ($2\delta^2$).
(2) When this overall approach yields a sample size of $n < 20$, the estimated n should be increased by the value of one (1).

E. DETERMINATION OF SAMPLE SIZE FOR F-TEST

1. Sokal (17) -- the basic formula uses t values for taking Type I and II Errors into account.

In studying four populations by means of ANOVA, the number of items from each population can be determined. The appropriate formula is

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left\{t_{\alpha[\nu]} + t_{2(1-P)[\nu]}\right\}^2$$

where n = number of replications

σ = True standard deviation

δ = the smallest true difference which it is desired to detect (N.B.: It is necessary to know only the ratio of α to δ, not their actual values)

ν = degrees of freedom of the sample standard deviation ($\sqrt{MS_{within}}$) with a groups and n replications per group

α = significance level (such as 0.05)

P = desired probability that a difference will be found to be significant (if it is as small as delta (δ))

$t_{\alpha[\nu]}, t_{2(1-P)[\nu]}$ values from a two-tailed t-table with degrees of freedom and corresponding to probabilities of α and $2(1 - P)$, respectively

Iterative Solution: Iterate to stability when necessary.

The ratio is given as 6/5. The initial n is 20. Then, ν is $(4(20 - 1)) = 4 \times 19 = 76$.

Substituted values on the basis of an n of 20 are:

$$n \geq 2\left(\frac{6}{5}\right)^2 \left\{t_{.01[76]} + t_{2(1-0.80)[76]}\right\}^2$$
$$= 2\left(\frac{6}{5}\right)^2 [2.64 + 0.847]^2 = 2(1.44)\,12.16 = 35.0$$

Next, try an n of 35. Substituted values are:

$$n \geq 2(1.44)[2.61 + 0.845]^2 = 2.88\,(11.94)$$
$$= 34.4\ (35)$$

Hence, 35 replications per population (a total of 140) are required for the four populations.

II. SELECTED SAMPLE-SIZE TABULAR APPROACHES

A. ESTIMATION OF THE POPULATION MEAN

1. Welkowitz, et al. (20) -- Gamma (γ), the effect size of the population, is determined by $\mu_1 - \mu_0 / \sigma$. Delta (δ), a function of n, $\gamma\sqrt{n}$, is read from Table 1.

$$n = \left(\frac{\delta}{\gamma}\right)^2$$

B. ESTIMATION OF THE POPULATION PROPORTION

1. Welkowitz, et al. (20) -- Gamma (γ), the population effect size, is determined by $p_1 - p_0 / \sqrt{p_0(1-p_0)}$. Delta ($\delta$), a function of n, $\gamma\sqrt{n}$, is read from Table 1.

$$n = \left(\frac{\delta}{\gamma}\right)^2$$

C. DETERMINATION OF SAMPLE SIZE FOR t-TEST

1. Dixon and Massey (2) -- for one-sample case -- d is read from the authors' table.

$$d = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

2. Dixon and Massey (2) -- for two-sample case -- d is read from authors' table.

$$d = \frac{\mu_1 - \mu_2}{\sigma\sqrt{(1/n_1) + (1/n_2)}}$$

3. Welkowitz, et al. (20) -- Gamma (γ), the population effect size, is determined by $\mu_1 - \mu_2 / \sigma$. Delta (δ), a function of n, $\gamma\sqrt{\frac{n}{2}}$, is read from Table 1. The resultant n is for each sample size; 2n (equal n's) is required for the computation.

$$n = \left(\frac{\delta}{\gamma}\right)^2 \qquad 2n = \left(\frac{\delta}{\gamma}\right)^2$$

4. Welkowitz, et al. (20) -- when the two sample sizes have unequal n's

$$n = \frac{2 n_1 n_2}{n_1 + n_2}$$

5. Dixon and Massey (2) -- for collected and analyzed paired data ($\sigma_1 = \sigma_2$) -- n provides the number of pairs of observations. Read d from the authors' table.

$$d = (\mu_1 - \mu_2) / \sqrt{(\sigma_1^2 + \sigma_2^2)/n}$$

6. Dixon and Massey (2) -- for deviation of μ from μ_0 -- one population mean -- one sample size -- d is read from the authors' table.

$$d = \frac{\mu - \mu_0}{\sigma/\sqrt{n}}$$

7. Dixon and Massey (2) -- differences in two population means -- two-sample cases -- d is read from the authors' table.

$$d = \frac{\mu_1 - \mu_2}{\sigma\sqrt{(1/n_1) + (1/n_2)}}$$

8. Guenther (3) -- Delta (δ) read from Owen's table (13).

$$\delta = (\mu_1 - \mu_2) / \sigma \left(\frac{1}{n_1} + \frac{1}{n_2}\right)^{\frac{1}{2}}$$

D. DETERMINATION OF SAMPLE SIZE FOR X^2

1. Sokal and Rohlf (15, 17) -- Cf. I-D-5, which is also applicable.

$$n \geq 2\left(\frac{\sigma}{\delta}\right)^2 \left\{t_{\alpha[\nu]} + t_{2(1-P)[\nu]}\right\}^2$$

E. DETERMINATION OF SAMPLE SIZE FOR F-TEST

1. Daniel (1) -- Phi (ϕ), a noncentrality parameter is read, converted, and interpreted on the basis of, for the most part, the Pearson and Hartley charts (14).

$$\phi' = \sqrt{\sum_{j=1}^{k} \alpha^2 j / k} \Big/ \sigma_e / \sqrt{n_j}$$

2. Guenther (3) -- One-way ANOVA.

$$\phi = \left[\frac{n}{r}\sum_{j=1}^{r} (\mu_j - \mu)^2\right]^{\frac{1}{2}} \Big/ \sigma$$

3. Guenther (3) -- Formula for an indirect determination of n.

$$n = \left[\sigma^2 \Big/ \sum_{j=1}^{r} (\mu_j - \mu)^2\right] r\phi^2$$

4. Guenther (3) -- Randomized Complete Blocks.

$$\phi = \left[\frac{n}{r}\sum_{j=1}^{r} (\mu_j - \mu)^2\right]^{\frac{1}{2}} \Big/ \sigma$$

5. Kirk (6) -- Basic formula.

$$\phi = \sqrt{\sum_{1}^{k} (\mu_j - \mu)^2 / k} \Big/ \sigma_\varepsilon / \sqrt{n}$$

6. Dixon and Massey (2) -- Basic formula.

$$\phi = \sqrt{\sum_{i=1}^{k} (\mu_i - \bar{\mu})^2 / k} \Big/ \sqrt{\sigma^2 / n}$$

7. Winer (21) -- The noncentrality parameter is read, converted, and interpreted on the basis of the Tiku tables (18).

$$\phi = \sqrt{\frac{n\Sigma\tau_j^2}{k\sigma_e^2}}$$

F. DETERMINATION OF SAMPLE SIZE FOR r

1. Welkowitz, et al. (20) -- Gamma (γ), the population effect size, $\gamma = \rho_1$, is determined by ρ_1, the correlation coefficient. Delta (δ), $\gamma\sqrt{n-1} = \rho_1\sqrt{n-1}$, is read from Table 1.

$$n = \left(\frac{\delta}{\gamma}\right)^2 + 1 = \left(\frac{\delta}{\rho_1}\right) + 1$$

Table 1
A Function of Significance Criterion (α) and Power ($1-\beta$)

	One-tailed test (α)			
	.05	.025	.01	.005
	Two-tailed test (α)			
Power ($1-\beta$)	.10	.05	.02	.01
.25	0.97	1.29	1.65	1.90
.50	1.64	1.96	2.33	2.58
.60	1.90	2.21	2.58	2.83
.67	2.08	2.39	2.76	3.01
.70	2.17	2.48	2.85	3.10
.75	2.32	2.63	3.00	3.25
.80	2.49	2.80	3.17	3.42
.85	2.68	3.00	3.36	3.61
.90	2.93	3.24	3.61	3.86
.95	3.29	3.60	3.97	4.22
.99	3.97	4.29	4.65	4.90
.999	4.37	5.05	5.42	5.67

Table 2
Alpha and Power (Sokal and Rohlf (15, 17))

Power ($1-\beta$)	α .1	.05	.01	.001
.50	4,442.2	6,306.4	8,883.7	10,891.5
.80	10,150.2	12,884.8	16,474.3	19,171.6
.90	14,059.3	17,249.8	21,368.5	24,426.2
.99	25,890.0	30,161.4	35,536.7	39,450.1

REFERENCES

1. Daniel, Wayne W. INTRODUCTORY STATISTICS WITH APPLICATIONS. Boston: Houghton Mifflin Company, 1977.

2. Dixon, Wilfrid J. and Frank J. Massey, Jr. INTRODUCTION TO STATISTICAL ANALYSIS, 3/e. N.Y.: McGraw-Hill Book Company, 1969.

3. Guenther, William C., "Power and Sample Size Determination When the Alternative Hypotheses Are Given in Quantiles," THE AMERICAN STATISTICIAN, Vol. 31, No. 3, August, 1977, pp. 117-18.

4. Guilford, J.P. and Benjamin Fruchter. FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION, 5/e. N.Y.: McGraw-Hill Book Company, 1973.

5. Hadley, G. INTRODUCTION TO PROBABILITY AND STATISTICAL DECISION THEORY. San Francisco: Holden-Day, Inc., 1967.

6. Kirk, R.E. EXPERIMENTAL DESIGN: PROCEDURES FOR THE BEHAVIORAL SCIENCES. Belmont, Calif.: Brooks/Cole Publishing Company, 1968.

7. Krejcie, Robert V. and Daryle W. Morgan, "Determining Sample Size for Research Activities," EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 1970, 30, pp. 607-10.

8. Lacey, Oliver L. STATISTICAL METHODS IN EXPERIMENTATION: AN INTRODUCTION. N.Y.: The Macmillan Company, 1953.

9. Lehmer, E., "Inverse Table of Probabilities of Errors of the Second Kind," ANNALS OF MATHEMATICAL STATISTICS, 15, 1944, pp. 388-98.

10. Marascuilo, Leonard A. STATISTICAL METHODS FOR BEHAVIORAL SCIENCE RESEARCH. N.Y.: McGraw-Hill Book Company, 1971.

11. National Bureau of Standards. TABLES OF POWER POINTS OF ANALYSIS OF VARIANCE TESTS (Non-Central F Tables), pre-publication copy. Washington, D.C.: National Bureau of Standards, undated.

12. Odeh, Robert E. and Martin Fox. SAMPLE SIZE CHOICE. N.Y.: Marcel Dekker, Inc., 1975.

13. Owen, D.B., "The Power of Student's t-Test," JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 60, 1965, pp. 320-33.

14. Pearson, E.S. and H.O. Hartley, "Charts of the Power Function of Analysis of Variance Tests, Derived from the Non-Central F-Distribution," BIOMETRIKA, 38, 1951, pp. 112-30.

15. Rohlf, F. James and Robert R. Sokal. STATISTICAL TABLES. San Francisco: W.H. Freeman and Company, 1969.

16. "Small Sample Techniques," THE NEA RESEARCH BULLETIN, 28, December, 1960, pp. 99-104.

17. Sokal, Robert F. and F. James Rohlf. BIOMETRY: THE PRINCIPLES AND PRACTICE OF STATISTICS IN BIOLOGICAL RESEARCH. San Francisco: W.H. Freeman and Company, 1969.

18. Tiku, M.L., "Noncentrality F Distribution" (Abbreviated from "Tests of the Power of the F Test," JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 1967, 62, pp. 525-39.

19. Walker, Helen M. and Joseph Lev. STATISTICAL INFERENCE. N.Y.: Henry HOlt and Company, Inc., 1953.

20. WElkowitz, Joan, Robert B. Ewen, and Jacob Cohen. INTRODUCTORY STATISTICS FOR THE BEHAVIORAL SCIENCES. N.Y.: Academic Press, 1971.

21. Winer, B.J. STATISTICAL PRINCIPLES IN EXPERIMENTAL DESIGN, 2/e. N.Y.: MCGraw-Hill Book Company, 1971.

PLANS FOR THE 1980 CENSUS

David L. Kaplan, U.S. Bureau of the Census

The roundtable discussion ranged over many aspects of the plans and preparations for the 1980 Census of Population and Housing, which will be conducted as of April 1, 1980. There were two elements of immediate interest to the participants. One was the newly introduced bill--H.R. 8871--which would make major changes in the format of the 1980 census; the discussion was necessarily limited since copies of this lengthy and complex bill had been available for less than a week.

The second was the recent announcement that the 1980 census dress rehearsal program would include censuses of the Richmond, Va. area (Richmond City and Chesterfield and Henrico Counties) and two small counties in Colorado (LaPlata and Montezuma) in April 1978; and a census of a portion of New York City (that part of Manhattan below Houston Street) in September 1978. The purpose of the dress rehearsal program is to use the planned final materials and procedures in locations which simulate various conditions the Bureau will face in counting everyone in the U.S. in 1980. After the dress rehearsal, only those materials and procedures which do not appear satisfactory for 1980 will be revised. This is different from the Bureau's test censuses in which alternative methods and questionnaires were tried out in a number of areas across the country during the last few years.

The discussion then turned to changes in subject content, touching upon such matters as the elimination of the word "head" from the question on household relationship; the problems in developing satisfactory questions on race and ethnic origin; the inclusion of a question on total income for all persons (instead of the traditional limitation to a sample) because of the need for more reliable data for very small places to use in revenue sharing and other government program allocation formulas; the continuing issue of how to measure housing quality; and the expansion of questions on housing shelter costs to include homeowners as well as renters.

Also mentioned was the fact that the Bureau was considering reducing the size of the sample from 20 percent to approximately 17 percent (i.e., from 1 in 5 to 1 in 6). During the discussion on subject content, a question was raised about the occupational classification system the Bureau expects to use in 1980 and information was provided to the particular participant by the Census Bureau after the meeting.

It was pointed out that the Bureau's major concern for the 1980 census is how to improve population coverage in the face of apparently increasing public apathy or even hostility. The coverage problem is especially great among minorities as suggested by the 1970 experience where the estimated rate of omission for blacks was approximately four times the rate for whites. Overall, the omission rate was estimated at 2.5 percent. The Bureau has undertaken a broad-scale and costly program to improve procedures and public communications, with special emphasis on minority groups. The dropoff in questionnaire mail-return rates experienced in recent pretest censuses was discussed, from the viewpoint of its impact on costs and time since more enumerator work is required in the followup of nonrespondents, as well as from the viewpoint of the dropoff being a potential indicator of public disinterest in cooperating in the census.

Mention was made of the fact that the first mid-decade population census is scheduled for 1985, in accordance with the law passed in late 1976. In approving this legislation, the Congress intentionally avoided establishing by law the scope and content of the mid-decade census. Rather, the legislation is flexible so that as the time approaches for detailed plans to be drawn, the Bureau of the Census may take into account data needs as seen for 1985. The Congress clearly intended that the law does not require that the mid-decade census duplicate the decennial census, provided that certain basic objectives are met--updates of characteristics along with population totals, particularly for the distribution of Federal funds to State and local governments, and the administration of Federal program benefits to various segments of the population.

Finally, the participants were informed that the Bureau has begun publication of an informal quarterly newsletter--entitled 1980 Census Update--which is available without charge to anyone interested in keeping informed on the progress of the 1980 census.

ANALYSIS OF TRENDS IN MARRIAGE AND DIVORCE

Paul C. Glick, U.S. Bureau of the Census

Changing marital lifestyles. Social changes have had mixed effects on the quality of family life in the United States. Far fewer married couples are living in poverty today than a decade ago, but because of the sharp increase in separation, divorce, and unwed motherhood far more of the families are single-parent families, 40 percent of which are in poverty. Far more women have acquired sufficient training to become active participants in the labor force, but a growing proportion of the economically independent women are now postponing marriage or have dissolved their marriages by divorce.

The burden of childbearing and childrearing has been diminishing as the birth rate has fallen, but one in every five children are growing up without the benefit of having two parents in the home--and one in every three are not living with both of their natural parents who are in their first marriage.

Most of the people who become divorced eventually remarry--about four of every five--but the transition period between marriages is generally very stressful. Many of those who remarry are far more satisfied in their remarriage than they were in their first marriage, but close to half of those who remarry after divorce become divorced once again.

Variations in marriage and divorce. Differences in divorce by educational level can be analyzed to special advantage if the age group 35 to 54 (or some other intermediate age group) is featured. These persons are old enough so that few additional marriages and divorces will occur in the highly educated segment who generally marry rather late, and yet the group excludes elderly persons, most of whose marriage experiences occurred several decades ago.

Some of the conclusions about differences in the level of divorce by education include the following for those 35 to 54: (1) the long-standing inverse relation between educational level and proportion divorced is now disappearing. In particular, the level of divorce for college-educated men (as a whole) has risen to the level of those with no college education (as a whole), whereas it was much lower in 1960; (2) the highest proportion divorced continues to be found among women with graduate school training (9.4 percent in 1976, up from 7.5 percent in 1973), whereas the lowest proportion among women is still that for women with exactly 4 years of college (6.5 percent in 1976, up from only 4.4 percent in 1970); (3) the greater likelihood of men than women to remarry is particularly evident in the following fact: the proportion of persons 35 to 54 at the graduate school level who remarried after divorce is half again as high for men (72 percent) as for women (48 percent); (4) projections of the proportion of persons who may eventually end their first marriage in divorce vary substantially according to educational level. The results of a study based on the Census Bureau's Current Population Survey for June 1975 show that, for both men and women born in the 1940's, the life-time proportion of persons who will end their first marriage in divorce is expected to be highest (about one-half) for those with an incomplete college education. The lowest proportion for men is that for those with graduate school training, and the lowest for women is that for those with exactly 4 years of college (about three-tenths); and (5) currently divorced women included twice as large a proportion as the still-married women who remained childless (13 percent versus 6.6 percent) and twice as large a proportion with only one child (18 percent versus 9 percent).

Among women in 1970 whose first marriage occurred in 1965 to 1969, the proportion with a premarital birth was nearly twice as high for those divorced by 1970 as for those still in their first marriage (17.1 percent versus 9.6 percent). Also, the divorced had a larger proportion who had a child conceived before marriage but born after marriage (24 percent versus 21 percent). Moreover, according to 1975 data, among divorced women who remarried, 11 percent of all their children were born in between divorce and remarriage. For Blacks this proportion was 14 percent, and for women with family income below $10,000 it was also 14 percent.

Living arrangements. Data for 1976 for the 2.8 million currently divorced men show that nearly half of them (46 percent) live in an apartment or house all alone. Another 20 percent of the divorced men live in with relatives (probably usually their parents) and 14 percent maintain a home of their own with some relatives present (but not always their own children). A small proportion of the remainder (5 percent of all divorced men) maintain a home that they share with only one other person, an unrelated woman; this figure is 8 percent for divorced men under 35 years of age.

Among the 4.4 million divorced women, one-half (52 percent) maintain their own apartment or house with relatives present, often one or more children. Another 27 percent live entirely alone, and 1 percent share their own house or apartment with only one other person, an unrelated man (2 percent for those under 30).

Some figures for comparison: 4 percent of men in their late twenties who had never married share their living quarters with an unrelated woman, as do 1.4 percent of single women of that age; and 2 percent of widowers and 1 percent of widows were reported in 1976 as sharing their house or apartment with an unrelated person of the opposite sex.

More information on divorce and remarriage classified by number and age of children and by age of women at the time of divorce may be found in a new Census Bureau publication entitled, "Marriage, Divorce, Widowhood, and Remarriage by Family Characteristics."

IMPROVING THE VALIDITY OF SURVEY DATA

Participants: Charles Cannell, Irene Hess, Martha Banks, Dan Freeman, Maria Gonzalez, Jean Jenkins Michael Lamphier, Eli Marks, Margaret Martin, Harold Nisselson, Bill Williams

To narrow the range of the topic, participants were invited to suggest aspects in which they had special interests. This led to suggestions as numerous and varied as the backgrounds of the participants.

Discussion began with responses to inquiries about the ASA pilot project on the assessment of survey practices and the problems of frame building to be faced in conducting a nationwide study. The directors have proposed a combination of methods in developing frames: construct lists of sponsors and surveys done for each; construct lists of survey-taking organizations and surveys conducted by each. Political polling and market research appear to be the most difficult areas for frame building. It was suggested that the universe of political polling might be limited to polls related to a particular election or class of elections (e.g., national presidential). Consideration of the problems encountered in constructing a frame of establishments conducting market research surveys led to a discussion of the importance of a clear, working definition not only of establishment but of any other term having a determinative role.

Nonresponse in relation to household personal interview surveys soon dominated the conversation. The frequently mentioned components of nonresponse and problems in dealing with them were reviewed. How nonresponse is defined and calculated has broad interpretations that vary with survey organizations. Nor was there expressed agreement among participants. It was suggested that in the case of quota samples, there should be a reporting of the number of households contacted in order to fill the assigned quotas. Otherwise, when data from such surveys are archived and distributed, analysts have interpretive difficulties. It was pointed out that one way to avoid this problem was to use probability samples.

Refusals may occur for many reasons. Interviewers concerned about their personal safety may refuse to go into some areas. Householders fearing strangers may not respond to a knock at the door. Entrances to apartment buildings may be barred by locked doors or by doormen. The sponsorship of a survey or the subject area may be grounds for refusals. The questionnaire design or the length of the interview may result in a partial refusal. We lack an understanding of why respondents refuse. It was strongly suggested that this issue should be investigated.

Over time there are changing ideas and changing perceptions of what is an acceptable approach to data collection. To illustrate, rather than asking respondents direct questions about voting in a past election, we might obtain names and addresses and go directly to voting records to determine who did or did not vote.

Little attention was given to sampling error as a factor contributing to the validity of survey data. However, there was a request that organizations distributing data sets and analytical programs also include programs for a proper calculation of sampling variability when data are not derived from simple random samples.

Exclusive of nonresponse, there remains a broad area of nonsampling errors that may have important effects on survey data. It is generally agreed that attention focuses largely on nonresponse because it is highly visible. Less visible are the effects of questionnaire design, question wording, interviewer error or bias, interviewer training and response error or bias. A high response rate does not guarantee high quality of data. What have we gained in pursuing a reluctant respondent until he agrees to grant an interview if his responses are irresponsible? It is not clear nor was there agreement on which effect should have more attention: nonresponse or response errors. More research is needed in these areas.

STATISTICS FOR HEALTH PLANNING

Dorothy P. Rice, National Center for Health Statistics

The National Health Planning and Resources Development Act (P.L. 93-641) was signed into law in January 1975. It was designed to create and support the capability for health planning to assure that needed health services are available, accessible and of high quality, but at the same time that there is not a costly, duplicative proliferation of services.

The nation has been divided into 212 Health Service Areas, and there is a Health Systems Agency (HSA) for each area which is responsible for area-wide planning. The agencies' functions include assembly and analysis of data, review of proposed new health services, reduction of unnecessary duplication of services and promotion of better services, and (in time) review of the appropriateness of existing services. The HSAs advise the State on Certificates of Need for new services.

The HSAs must develop Health Systems Plans for their area, which are statements of the goals for health investments in the community. These goals must be specific and quantitative wherever possible.

The P.L. 93-641 placed considerable emphasis on the acquisition and use of health statistics to analyze the health systems' strengths and weaknesses and determine the need for new services and identify areas which may have a surfeit of facilities and services. The goals and objectives of the planning process, evidenced in the Health Systems Plans, must be derived from the thoughtful analysis and interpretation of empirical data.

The HSAs must assemble and analyze the data for their area on health status, use and effect of the health care delivery systems, health resources, health financing and the environmental and occupa-

tional exposure factors affecting immediate and long-term health conditions.

In acquiring the data, agencies must tap existing sources of information and coordinate with the Cooperative Health Statistics System (CHSS), as well as PSROs, State, county and city health departments, other planning bodies, etc. Clearly, those involved in health statistics, can play a key role in making certain that data are available to planners to meet their challenging responsibilities.

Planning agencies are being advised to develop a population-based approach to data acquisition and planning. They are expected to build their information resources in a manner in which they can link events (births, deaths, discharges, etc.) to a defined population, such as by using a geocode (census designation or zip code).

Vital statistics are especially important to planners. The number and rates of births and fertility rates are indicators of the age distribution of the population (which affects the need for health services), as well as significant direct indicators of need for specific health services.

Questions concerning environmental and occupational safety and health are of increasing interest to planners. There are few sources of data on risk factors, morbidity or injury, particularly from reliable sources which are linked to a defined population. However, studies of mortality cross-classified by area of residence and occupation and business or industry could help to identify these jobs and geographic areas which have disproportionately high rates of death especially among the younger workers.

It was generally agreed among round table participants that mortality data tend not to be the best measures of the "health" of a population, at least in an advanced industrial society. The most important untapped source of information from vital records for planning purposes is information on patient utilization patterns of health services, especially hospitals.

The CHSS will be a major means of meeting the needs of the HSAs. The CHSS will help mold the current fragmented data collection activities throughout the country into a cohesive system that will produce comparable data in the detail required for most users.

The National Center for Health Statistics (NCHS) has the responsibility for developing the CHSS. When CHSS is fully developed, each State will have the capability to ensure availability at the local level of the same types of data that have in the past been available only at the national level. The CHSS, administered by NCHS, and authorized through P.L. 93-353 is an effort to build a health data system which will serve as the basis for effective planning at all levels of government in all areas of the country.

The NCHS and Bureau of Health Planning Resources Development (BHPRD) have an agreement and workplan to develop the data activities to meet the needs of the planning enterprise. NCHS has developed a source book on current national data that provides information to guide staff as to where data on health status, health resources, and health utilization are currently available.

NCHS is also developing and distributing "Statistical Notes for Health Planners" that are providing the methodology to HSAs for use of existing data available from Federal programs in an easily accessible and easily updated format. These "notes" will add to the library of statistical information to each HSA.

NCHS has a firm commitment to combine the best efforts of health statisticians and health planners toward the development and uses of a coordinated statistical support capability which will allow the best possible planning and resource allocation to take place in the health care delivery system.

MEASUREMENT OF DISABILITY: ROUNDTABLE DISCUSSION

Aaron Krute, Social Security Administration

The session focused on the measurement of work disability in the adult population by household surveys. The following definition of disability was used to assure a common frame of reference. Disability is the result of functional limitations arising from a mental or physical condition(s) interacting with a host of other factors such as age, work history, education, family situation, etc. to leave an individual incapable of adequately performing his/her generally accepted social role, e.g., working, keeping house, or going to school.

This definition, itself, highlights several significant difficulties of measurement. First, knowledge of the underlying disease or condition is not sufficient. More important are the residual physiological, anatomical or mental losses or abnormalities, i.e., impairments, that result. Impairments contribute to disability through the nature and extent of the functional limitations they cause. For example, loss of muscle strength may lead to an inability to lift, while the loss of a limb may mean an inability to walk. Identification and quantification of such limitations are very important for measurement.

Second, not every impairment results in a disability. Identical impairments with the same degree of severity may even result in different levels of disability. Thus, muscle weakness is much more limiting to a laborer than to an accounting clerk, while the opposite is true for good eyesight or manual dexterity (fingering).

Third, disability is determined by the interaction of limitations in function with situational and environmental factors. In the case of

work disability, the latter factors include work requirements, employer attitudes and practices and general labor market conditions. So measurement of disability depends on the observation of many socioeconomic and attitudinal variables.

Fourth, the difficulty is increased because disability represents a continuum where the demarcation between disability and no disability is unstable. In other words, disability is a changing state dependent on shifts in the many precipitating factors. Even when functional limitations are stable, a changing labor market can still change the state of work disability.

In this context, the basic measurement problem identified by the discussants was distinguishing inability to work as a result of some impairment from inability to work for other reasons. The concomitant practical problem was seen to be the kind of proxy measure, that can be developed by household survey, to substitute for the "ideal"--medical examinations and clinical and vocational assessments by teams of experts.

Next came a review of the current state of the art as revealed by four (4) major sources of recent survey data on the extent and nature of disability in the U. S. Discussion emphasized the fact that all of the surveys reviewed measure disability based on respondents' self-assessments of their own situations. Generally, respondents are asked a series of direct questions about how their health or physical condition affects their work activities. They are also asked if they have any of a specified list of chronic conditions or impairments, and which of them is primarily responsible for their work limitation.

Other measures based on functional capacity limitations were also described. Scales have been constructed to measure performance in activities of daily living (feeding, grooming, etc.), in use of public transportation, in ability to move about the community, and in work activities (lifting, stooping, reaching, etc.). Composite measures of the severity of functional limitation using these scales have been constructed. Such indexes provide a scale of functional level ranging from no limitations through dependency.

Figures from the 1972 Social Security Survey of Health and Work Characteristics and other SSA records were used to illustrate the lack of precision of present survey measures. According to the 1972 survey, there were 15.6 million adults aged 20-64 with some work disability in 1972, including 7.7 million who could not work at all. At the same time, only about 2 million persons were receiving social security disability insurance benefits (SSDI). Persons who apply for SSDI benefits can be presumed to consider themselves disabled. Yet, between 40 and 50 percent of such claims are disallowed for lack of severity.

Results are similar with the functional limitation index. About one-third of severely disabled respondents in the 1972 SSA survey had either no loss of function or a minor loss. At the other extreme, more than a third of those with dependency problems did not consider themselves severely disabled. Using regression analysis, functional limitations alone explained only 13 percent of the variance in severe disability among males and 8 percent among women.

Other investigators have reported similar findings. Nagi at Ohio State used figures from another survey of the disabled in 1972. Eight independent variables--including scaled assessments of physical and emotional performance, two health indices, plus age, sex, race and education--were regressed against work disability as the dependent variable. These regressions explained only 38 percent of the variance in work disability.

Several reasons were advanced for these results. Functional limitations used in the various indices may be inaccurate since they are also based on self-reporting. The nature and effect of disabling conditions, of functional limitations, and of relevant socioeconomic variables may not be specified completely or precisely enough. For example, indices might provide better measures if limitations were specified so that they could be matched against the requirements of various jobs. Finally, it was suggested that the form of the model used--an additive linear function--was wrong. It fails to take into account the fact that some functional limitations are cumulative while others may duplicate each other.

The Roundtable concluded with a brief look at some possible future directions for improved measurement. These included--

1. Improving the specification of functional limitations, including matching them to job requirements.
2. Refining survey instruments to provide better attitudinal and motivational information as well as more information about dimensions of chronic disease besides those that result in limitations in physical movement.
3. Constructing a disability index by measuring the "distance" between nonbeneficiary disabled persons in the population and SSDI beneficiaries with regard to a profile of characteristics.
4. Applying methodology being developed by Moshe Nordheim in Israel and Gerda Fillenbum at Duke University. In the former, teams of expert raters interview a sample of the population and assign an overall disability rating. A disability index, $y = f(x_1,\ldots,x_p)$ is constructed statistically. The survey data are used to estimate the parameters of the function. In the latter a panel of experts takes a set of characteristics, in the abstract, and maps alternative profiles of these characteristics to a set of numbers which represent degree of disability. Survey data are used to suggest meaningful profiles for mapping.

NEW DEVELOPMENTS IN EDUCATIONAL STATISTICS
Summary or Roundtable Discussion

Abbott L. Ferriss, Emory University

Participants, named at the end of this summary, represented a wide range of interests: the evaluation of local school-district educational programs, the projection of the supply and stock of nurses in the future, the adequacy of a state-wide educational statistics program, the development of a research program to serve a variety of planning and administrative needs, and finding ways to improve the Federal statistics program in education. While all of these interests were not equally served by the discussion, each contributed to the exchange of information and ideas.

The demographic "stock and flow" model (1) was briefly presented and the attempt to assess its application to U.S. data on education was discussed(2). Examples of the use of the Life Table model in producing various educational statistics -- educational expectancy, years of education remaining for a given age, etc. -- were cited as examples of the application of other demographic models to educational data (3). References to demographic models were provided the discussants (4).

Since several of the group were interested in longitudinal studies, the methodological work underway by the Committee on Methodology of Longitudinal Research of the Social Science Research Council was cited, and information was provided the discussants on access to these developments. This included a copy of the "Longitudinal Methods Bibliography", assembled by the Committee (5).

Of particular interest were the six projects supported by the National Institute of Education, each focusing upon various issues in longitudinal analysis: analysis of qualitative educational data, methodological problems in educational research, "developmental-educational considerations", evaluation research from a "general systems perspective," and others. The source of additional information on these developments was provided to the group (6).

Of particular interest was the Nesselroade-Baltes project, "Developmental-Educational Considerations in Longitudinal Research Methodology," which is developing a technical manual of methods of designing and analyzing data from longitudinal designs in the social sciences. The manual will consist of eight or ten chapters, each prepared by a different author, or authors, especially for the manual. The volume is expected to contribute significantly to the methodology of longitudinal analysis (7).

There was some comment, also, on data from the National Assessment of Educational Progress, the release of NAEP data tapes, and related matters (8), and the release of data from the National Longitudinal Study of the High School Class of 1972.

Since the present status and immediate future plans for much of the work in progress by the Federal government on education, particularly the studies of the National Center for Educational Statistics, are contained in Part 2 of the most recent (1977) issue of The Condition of Education, the discussants were referred to that volume, and a copy provided them (9).

Representatives from the U.S. Bureau of the Census and the Statistical Policies Division, U.S. Office of Management and Budget, present for the discussion, reviewed the prospects for new or improved developments in educational data from the standpoint of their agencies.

Participants in the discussion: Robert J. Cruise, Andrews University (Mich.); Carole Perlman, Chicago Board of Education and Roosevelt University; Aleda Roth, American Nurses Association, Kansas City, Mo.; Larry Suter, U.S. Bureau of the Census; Amanda Kautz, Hawaii State Manpower Commission; Khazan Agrawal, Chicago Board of Education; Kathy Wallman, U.S. Office of Management & Budget; Jack P. Kornfeld, ITT Research Institute, Chicago; Samuel T. Mayo, Loyola University, Chicago; Abbott L. Ferriss, Emory University, Discussion Leader.

References

(1) United Nations, Department of Economic and Social Affairs, Towards a System of Social and Demographic Statistics, Series F, No. 18, New York: United Nations, 1975.

(2) "A Test of Demographic Accounting Methods with U.S. Data on Education," Abbott L. Ferriss, Principal Investigator, National Science Foundation Grant SOC76-17387 to Emory University.

(3) Abbott L. Ferriss, "Trends in the School Enrollment of Females and Consequences," paper presented at the annual meeting of the Population Association of America, St. Louis, April 21, 1977, MS.

(4) "References on Demographic Models Applied to Educational Data" (Ditto).

(5) "Longitudinal Methods Bibliography: Papers Received by the Committee on Methodology of Longitudinal Research, Social Science Research Council," June 1977, available from SSRC, 650 Third Ave., New York, N.Y., 10016 (mimeo).

(6) "List of Projects on Longitudinal Models Supported by National Institute of Education," (Ditto); additional information available from Carlyle E. Maw, National Institute of Education, Dept. HEW, Washington, D.C.,20208.

(7) "Overview of Project, Developmental-Educational Considerations in Longitudinal Research Methodology," (NIE-C-74-0127), Principal Investigator John R. Nesselroade and Paul B. Baltes, The Pennsylvania State University.

(8) "User Tapes," National Assessment of Educational Progress, Education Commission of the States, 1860 Lincoln Street, Suite 700, Denver, Colorado, 80295.

(9) Shirley A. Radcliffe, The Condition of Education (1977 edition), part 2 (NCES 77-400) Washington: U.S. Government Printing Office, 1977.

DISCUSSION
MEASUREMENT OF PUBLIC WELFARE STATUS

Mitsuo Ono, U.S. Department of Health, Education, and Welfare

BACKGROUND: The macrohousehold income structure can be divided into five social systems: (1) employment, (2) social insurance, (3) welfare, (4) capital income, and (5) inter-intra household transfers. (Reference 1) The tax system encompasses all of these components. The discussion attempted to identify work needed to improve the public welfare statistical system covering such public assistance programs as the Aid to Families with Dependent Children (AFDC), Medical Assistance (Medicaid), Supplemental Security Income (SSI), Food Stamps (FS), certain social services (SS) categories, and the Work Incentives (WIN) programs. Compared with other components, e.g., the employment system, the public assistance statistical system can be significantly improved.

Detailed program descriptions are found in reference 2. According to reference 3, these programs involved total expenditures of $54 billion in FY 1976. About 66 percent of the total were Federal transfers, with the rest coming from State and local governments. Approximately 25 million beneficiaries participated in one or more of these programs.

We need to know the operating characteristics of these programs to understand statistical reporting problems. Some of these are: (1) Assistance programs are fragmented. Coordination efforts to reduce overlaps are difficult to implement effectively. (2) Since programs involve Federal, State, and local government participation, management becomes complicated because of competing priorities generated from legislative and administrative initiatives. (3) States' administrative structures for collecting and reporting data vary, e.g., State-administered versus county-administered operations. (4) Wide variations exist among States in channeling program funds, e.g., some States operate mostly through public agencies while others use contractors. (5) Priorities on information needs are always evolving because of legislative and administrative mandates. (6) Data processing capabilities of State agencies vary widely. Financial and grant award processing are given higher priorities than statistical reporting. (7) Although some States have privacy laws, others are still developing such legislation. (8) Because of complexity of program operations, the ideal integration of financial, cost, and performance data for planning and managerial purposes is not practical. These complex institutional arrangements, the lack of adequate analytical models (probably due to paucity of integrated data), the lack of adequate resources and difficult coordination and administrative problems encountered in producing data are important analytical considerations.

The writer believes that the production or supply side in generating data on public welfare assistance should have higher attention than the demand side on data needs. Thus, States need help in establishing computerized sample data files to generate adequate State data. National data could be consolidated from such sample State data. A project is currently under way in the State of Texas to test this concept. (Reference 4) In addition, better information on target eligible populations is required from general purpose sample surveys on households.

Other priorities include establishment of strong Federal-State-local government statistical cooperative systems, development of State confidentiality laws, formulation of minimum data sets, and standardization of data elements used by State agencies.

Finally, we need to develop a public assistance transaction accounts system which can trace the flow of transfer payments between and among different public welfare assistance program category populations, with appropriate accounting for multiple beneficiaries. (Reference 5) This social accounts system could also include social progress indicators.

DISCUSSION: Items discussed can be divided into four major headings. The first dealt with the need for better coordination and interchange of information among users and producers of general purpose household surveys and censuses, which provide data used to estimate low-income households and welfare programs' eligible population. These sources cover the Decennial Censuses, the Current Population Survey, the Survey of Income and Education, the Consumer Expenditures Survey, etc. Participants expressed the need for better documentation of User Manuals especially for public use samples, for more interchange of ideas between users and producers in forums such as CPS Workshops to take wider account of users' needs and problems, etc. It was also noted that

DHEW is currently testing a proposed Survey of Income and Program Participation which should provide data presently not included in the Current Population Survey.

The second covered the need for more accuracy of data obtained from household surveys and censuses, especially on income data. In this regard, it was noted that the Social Security Administration, the Internal Revenue Service, and the Census Bureau are jointly cooperating in evaluation projects aimed to obtain results which could be used to reduce not only survey response errors but also improve adjustments for nonresponses. These studies use administrative records and household survey data. New techniques derived from these projects will be valuable in improving future surveys and censuses, especially on collecting income data.

The third area of discussion dealt with the lack of adequate guidelines regarding the meaning and scope of confidentiality. There appears to be a need to differentiate situations where confidentiality rules can be used with some flexibility. This calls for clearer definitions.

The fourth topic covered work needed to develop and expand the use of sample microdata files for public welfare assistance statistical reporting and analyses in States which have capabilities of doing so. The basic approach used in the Texas demonstration project outlined in reference 4 appears to be promising.

Other areas of discussion touched on the need to obtain better small-area data from general purpose surveys and censuses for local government administrative use and the impact of the current OMB directive to reduce reporting burdens of Federal reports.

NOTE: Participants agreed that the discussion was made more interesting and useful because of the diverse background of discussants. A suggestion was made that, if possible, participants should review background papers before the meeting. As an alternative, it was suggested that participants be queried beforehand on topics/questions they would like to discuss and this listing be distributed before the meeting. The background paper used for this meeting can be obtained from the writer, address: OPRE, OHDS, DHEW, Room 2614, Switzer Building, 330 C Street, S. W., Washington, D. C. 20201.

REFERENCES:

1. Toward an Effective Income Support System: Problems, Prospects and Choices by M. C. Barth, G. C. Carcogno, and J. L. Palmer; also Overview Paper by I. Garfinkel, Institute for Research on Poverty, University of Wisconsin - Madison, 1974.

2. Studies in Public Welfare, Handbook on Public Income Transfer Programs, Paper No. 20, Joint Economic Committee Print, 93rd Congress, 2d Session, 1974.

3. "Social Welfare Expenditure, FY 1976" by A. M. Skolnik and S. R. Dales, Social Security Bulletin, January 1977. See also Social Welfare Expenditures Under Public Programs in the United States, 1929-1966, by I. C. Merriam and A. M. Skolnik, Social Security Administration, 1968.

4. "Development of Computerized Welfare Recipient Statistical Reporting System" by G. Higgins, R. Becker, and M. Ono. Presented at 1977 Annual Meeting of American Statistical Assoiiation

5. The Measurement of Economic and Social Performance, Milton Moss, Editor, Studies in Income and Wealth, Volume Thirty-Eight. National Bureau of Economic Research, New York, 1973. See also "Social Accounting for Transfer" by Robert Lampman, Reprint No. 143, Institute for Research on Poverty Reprint Series.

CURRENT NATIONAL FERTILITY SURVEYS

W.F. Pratt - National Center for Health Statistics

A great number of recent, current and projected national surveys have developed in many countries under the aegis of the World Fertility Survey. These are vey largely modelled on KAP studies and earlier national studies undertaken in a few developed countries. In the United States specifically, the major current national studies in the area of fertility are the 1975 National Fertility Study (based on a follow-back to once-married, currently married women in the 1970 NFS and a supplemental sample of women married in the intervening years), the Johns Hopkins studies of teenage pregnancy (1971 and 1976) and the National Survey of Family Growth (NSFG) 1973 and 1976.

The presentation and discussion focused largely on the NSFG. Described as a lineal descendant of the earlier NFS and GAF studies going back to 1955, the NSFG is a new data system in the National Center for Health Statistics. Field work for the first two cycles of the survey was done in 1973 and 1976, respectively. In order to exploit the data of these first two cycles as fully as possible, and to expand the coverage to include all women 15-44 years, regardless of marital status, Cycle III has been postponed to 1980.

The NSFG is a household survey based on personal interviews with an area probability sample of women 15 through 44 years of age, who have children of their own in the household or have ever been married, and who reside in the conterminous U.S. Completed interviews in the first two cycles were 9,797 and 8,611, respectively. The topics of the interviews included a detailed marital history, a complete pregnancy history with dates, outcomes, and various characteristics of each pregnancy, a pregnancy planning history with information on the "wantedness" of each pregnancy and details on

the specific contraceptive methods used in the three years preceding the interview, ability to bear children in the future and the intentions and expectations of couples regarding future births and future use of contraception, and examination of preferences for the number and sex of children, information on family planning services received including services to increase the chances of childbearing, and general social and demographic characteristics.

NSFG data will be published by the NCHS in Advance Data releases and in Series 23 of the Vital and Health Statistics Reports. Advance reports from Cycle I on contraceptive utilization, wanted and unwanted births, birth expectations and pregnant workers have been published, to be followed in the fall of 1977 and through 1978 by detailed reports on a wide range of topics such as trends in contraceptive utilization, the realization of family size goals, underlying preferences for family composition, employment before and after childbirth, trends in unwanted fertility, family planning services, use-effectiveness of contraception, short-term birth projections, and socio-economic differentials in expected family size. Advance data from Cycle II are expected to begin in the summer of 1978, followed by detailed reports throughout 1979. A public use tape for Cycle I has been made available by NCHS and a similar tape for Cycle II is anticipated for December 1978.

It was agreed that inclusion of single women in the NSFG was an important step because of their contribution to current levels of abortion and illegitimate conceptions, the growing interest in family planning services to young women and because their behavior and expectation about marriage and childbearing play a major part in the birth rates of the next few years. It was noted that, while single women with children of their own in the household were already in the NSFG, they comprise a very selective group of sexually active singles. The possible difficulties in obtaining reliable data on unmarried minors was considered. The need for parental consent, for instance, would add to the costs and possibly effect response rates adversely.

The need for better abortion data was emphasized. It was pointed out that responses to direct inquiries on abortions seemed to be improving though still short of complete candor. The randomized response technique, though yielding estimates vary substantially greater than those based on reports by abortion providers, left too many points of doubt to be a satisfactory procedure. Asking about the use of specific abortion techniques rather than the general and possibly loaded term "induced abortion" was suggested.

The increasing frequency of cohabiting couples suggests possible institutional changes in marriage that should be monitored through a survey like the NSFG. The survey presently includes "informal marriages" provided this information is volunteered in response to questions on "relationship to head" and "marital status." More direct questions might be developed for monitoring the frequency of these unions, though difficulties in obtaining reliable retrospective accounts of these unions were acknowledged.

The desirability of obtaining more family background characteristics was examined. Background characteristics of the couple, as presently asked, comprise the largest single section of the interview. Expansion of these items would probably be at the cost of information on one or more dependent variables. It was recognized, however, that the traditional background characteristics of couples were explaining less and less of the variation in fertility behavior. While the explanatory power of alternative characteristics need study, it was questioned whether one should disrupt valuable times series data in a large scale national survey to experiement with new items whose discriminant value was largely unknown.