BALANCED REPEATED REPLICATIONS FOR ANALYTICAL STATISTICS*

Leslie Kish and Martin Frankel, The University of Michigan

1. Summary

Balanced repeated replications (BRR) is a general method for computing standard errors. It has wide utility when specific mathematical methods are lacking, and especially for analytical statistics based on complex samples, where clustering destroys the independence of observations. We present results of methods we used since 1957 to measure standard errors of regression coefficients for several multivariate techniques. The basic design of the several samples comprised two primary selections (PS) per stratum.

Each replication was a half-sample, created by selecting one PS per stratum. The variance of the estimated coefficient b_* is measured by $(b_j-b_*)^2$ where b_j is the same estimator based on a half-sample. To increase the precision of the variance, we select repeated replications and compute the mean of the variance, $\Sigma(b_j-b_*)^2/J$. Balanced repeated replications reduce the number of repetitions needed. We obtained practically all available precision from 47 strata with 48 BRR. Though proofs are complete only for linear statistics, we offer rationale and results to indicate that BRR provides needed estimates of errors for non-linear statistics.

The ratios $\sqrt{deff}$ of actual to srs standard errors are investigated for several statistics in 5 sets of empirical studies. In each study average values of $\sqrt{deff}$ exceed 1.00, and range from 1.05 for widespread samples to 1.35 to more clustered samples.

2. Analytical statistics from complex samples

The development of standard statistical literature has been based on the assumption of samples of independent observations, which greatly facilitates obtaining interesting theoretical results. On the other hand this assumption of unrestricted random sampling is violated by the designs of most survey samples. Practical, economic designs often use clusters of sample elements, which induce strong correlations among them. These correlations have serious effects on statistics based on complex samples. They also pose formidable theoretical obstacles. It would be difficult to unravel the effects of some complex designs on the distribution of even one specific analytical statistic; it is even less reasonable to expect separate derivations for each statistic for all major designs. Hence, we need badly general methods for getting around those obstacles.

To clarify our discussion, we may think specifically of the coefficients of multiple linear regression which can typify the broader class of analytical statistics we need to estimate. The derivations of their distributions imply independence of observations (not of the variables); specific proofs of asymptotic theories, as well as laws of large numbers and central limit theorems, generally also assume independence. Yet we do conjecture that regression coefficients (and other statistics) based on large complex samples also tend in probability to approach the corresponding parameters; and that the approach is merely slowed by the correlation between elements of the same clusters. We are not alone in acting on such conjectures, and on the belief that proofs will come [Kish 1965, 2.8C; Tharakan, 1968]. After all, researchers have used data from complex samples to compute regression coefficients and other analytical statistics.

However, we should not expect that variances for regression coefficients, or other statistics, based on assumptions of independence will be valid. Rather we should expect that these will tend to underestimate the variance, as they have been shown to do for means and differences of means. We may view differences between means as the simplest form of analytical statistics, and we had abundant evidence about the design effects on their standard errors. We used these results as bases for conjectures about effects on other analytical statistics [Kish 1957; Kish 1965, 14.3]. These conjectures are born out by the results of the investigations here presented.

What alternatives have we? Research and researchers need analytical statistics, and cannot wait for the possible development of the extremely complex distribution theory necessary for complex samples. On the contrary, it is often impractical to use unrestricted random samples to conform to the distribution theory available for analytical statistics [Kish 1957, and 1965, 14.2]. Simple interpenetrating samples often will not serve because of the conflict they raise between desired stratification and adequate degrees of freedom [Jones 1956, Kish 1965, 4.4].

Many samples are highly stratified, with clustered selections from the strata. A model of two independent primary selections from each stratum is probably the most basic design that conforms adequately, if not perfectly, the actual design of many actual survey samples. Our investigations and our discussion are based on that model, and on balanced repeated replications (BRR) for obtaining estimates of standard errors, described next.

Before passing on to it we should mention a possible alternative in extensions of Taylor approximations, sometimes called "propagation of variances" or the "delta method." Tepping [1968] has recently called attention to its possibilities in complex samples, as have Deming [1960, 390 ff] and Kish [1965, 585]; it was used for standard errors of double ratios and index numbers [Kish 1968]. Note also a recent theoretical investigation [Brillinger and Tukey, 1964]. We reserve judgment until we see how useful, general,

* Supported by Grant GS-777 from the National Science Foundation

practical and robust it will be proven on data. Meanwhile, we suspect that BRR will fare better in most situations involving complex analytical statistics such as regression coefficients.

3. Balanced repeated replications

We propose this descriptive name for a method we used in a series of investigations since 1957, but especially since 1964. The method may be summarized in a few steps.

A. A replication (half-sample) is created by selecting at random one of the two primary selections (replicates) from each stratum. The replication reproduces accurately the complex design of the entire sample. From the j^{th} replication the desired statistics (b_{ij}) are computed. For example, compute the regression $y = b_{1j}x_1 + b_{2j}x_2 + b_{3j}x_3$ from the replication with the same estimation process used to compute $y = \hat{b}_1x_1 + \hat{b}_2x_2 + \hat{b}_3x_3$ from the entire sample. Then $(b_{1j}-\hat{b}_1)^2$ estimates the variance $\hat{b}_1$, the statistic computed from the entire sample. The variances of $\hat{b}_2$ and $\hat{b}_3$ are similarly estimated by $(b_{2j}-\hat{b}_2)^2$ and $(b_{3j}-\hat{b}_3)^2$.

Instead of $(b_{1j}-\hat{b}_1)$ we can use $(b_{1j}-b'_{1j})/2$, half of the difference between the selected replication and its complementary half-sample. The estimate $(b_{1j}-\hat{b}_1)$ is cheaper to compute, and our research shows small differences between results with the two methods. (See section 10).

B. Repeated replications are needed, because the estimates from (A), based on a single replication and a single degree of freedom, are extremely variable; almost uselessly so in most cases. But we can repeat the process (A) by drawing new replications to obtain new estimates b_{1j} and $(b_{1j}-\hat{b}_1)^2$. From k repetitions we compute the average of the k variance estimates $\sum_j(b_{1j}-\hat{b}_1)^2/k$. Because each of the k values is an estimate of the variance, so is their mean value. The precision of this average increases with the number of repetitions, but only slowly if the repeated replications are selected at random, without any connecting design. From H strata, full precision can be obtained from the 2^{H-1} possible ways of forming half-samples.

C. Balanced repeated replications reduce drastically the number of repetitions needed. For example, most of the precision from our 47 strata can be obtained from 48 balanced repetitions. This can be managed on large computers in many situations. The precision it yields is moderately adequate for usual needs: the coefficient of variation of the standard error is about $\sqrt{1/2(47)}$, or about 0.10.

D. However, a coefficient of variation of 10 percent is rather high when we want to distinguish design effects as low as 1.05 or 1.10 from 1.00. Hence we prefer to search for stability by investigating averages based on groups of statistics. A useful and common technique is to compute design effects, or deff [Kish, 1965, 812], the ratio of actual complex variances to their simple random variances. We have been averaging the values of $\sqrt{deff}$, the ratios of the standard errors, because these are less subject to extreme values. The differences between the average of deff and $\sqrt{deff}$ are not great in our practice, and the theoretical grounds for preferring one or the other not clear to us.

In computing, tabulating and accumulating design effects we have four aims: to check and improve the specific estimates of standard errors when these are highly variable; to estimate standard errors when only their simple random estimates are available; to appraise and understand sources of sampling variations; and to design better samples. We present later the results of our empirical investigations largely in terms of effects $\sqrt{deff}$, because these are more meaningful to the reader than a set of specific standard errors would be.

Simple random variances of most statistics can often be computed cheaply, based on analytical formulas already built into standard computing programs and on the entire sample; they are usually subject to much smaller sampling variations than are the actual variances computed with BRR. When analytical variances are not available, simple random variances can be computed also from simple random splits of the sample.

* * * * *

Because of its wide applicability and its essential simplicity, the method belongs perhaps to the class of "jack-knife" methods. We expect that these first applications to analytical statistics will be followed by many others, probably with modifications; we suggest some [Kish and Frankel, 1968]. That article contains results of several investigations we made to check the soundness of BRR, all very reassuring. Sections 10 and 11 contain some basic theoretical foundation and further reference.

We are aware of course that our methods, as it often happens in statistics, have run beyond rigorous mathematical foundations in places, especially in part D above. Our methods and our investigations have been stimulated by the needs of empirical research using regressions and other analytical statistics on data from complex samples. These methods are preferable to the available alternatives listed in section 2, and specifically to the universal usage of standard error estimates based on simple random assumption [Kish, 1957].

Repeated replications were first used for standard errors by the U. S. Census Bureau, as noted by Deming [1956]. At the Survey Research Center we soon began a series of computations for regression coefficients, and other analytical statistics, also introducing balancing into the repetitions. Our early programs were written by Irene Hess, Edwin Dean, and Kathleen Goode; since 1964 John Sonquist and K. S. Srikantan have contributed. Meanwhile, Margaret Gurney designed balanced replications for the U. S. Census Bureau [1963]. Walt R. Simmons used the method for surveys of the National Center for Health Statistics, and interested P. J. McCarthy, who developed an optimal method of balancing [1966, 1968]. A more efficient, flexible, and widely applicable set of programs for IBM 360 systems has been developed in 1968 by Martin Frankel and Neal Van Eck of the Survey Research Center, together with Carl Bixby and David Seigle of Interface, Inc. Ann Arbor; this work is supported by a grant from the National Center for Health Statistics.

4. Summary of empirical results

The results taken together constitute reassuring rewards for patient and difficult work since our first efforts in 1957, but especially between 1964 and 1967, when we completed the five projects described in section 5-9. These gave us some security for the following tentative conclusions. They also encouraged us to further research with better computing programs we completed in 1968.

First, the results reassure researchers who have been using data from our national samples for multivariate analyses: the standard errors computed by machine programs, based on srs assumptions, were not gross underestimates. Before these results one could reasonably fear that the design effects could be as high as they were for the standard errors of means, or conceivably even higher. (Effects of $\sqrt{deff} = 1.40$ on the standard errors of means were not rare.) For example, the average effect on standard errors was 1.17 for means, but only 1.06 for the regression coefficients from data described in section 5.

Second, design effects on standard errors of regression coefficients were shown to be estimable and of appreciable magnitudes. For example, an increase of $1.06^2 = 1.12$ in the variance corresponds to a similar decrease in the effective size of the sample. Ignoring it means that instead of error rates of 1.96 or 2.58 one is using levels of 1.85 or 2.43; hence, instead of five or one percent errors, one has 6.5 or 1.5 percent. Furthermore, the low effects of 1.06 and 1.10 were found in widespread samples; we expect greater effects in samples with greater concentrations.

These two conclusions contradict extreme positions. One extreme is the widespread wishful thinking that the design effects on the standard errors of analytical statistics, which had been generally neglected, will prove to be negligible. On the other extreme lurks the fear that those effects may eventually turn out to be as large as the large effects often found for the standard error of means; or perhaps larger. Though these extremes will remain mathematically possible for future data, their likelihood is greatly reduced by the consistency of the large and varied body of empirical data here exhibited. They show design effects consistently between the two extremes, somewhat closer to the lower. Since the design effect for means is more available, this empirical rule should be more helpful than assuming either extreme. Furthermore, the design effects for analytical statistics seem to resemble those for differences between pairs of means, also more available. For example, the mean factor of 1.06 for regression coefficients in section 5 resembles the factor of 1.07 found for differences of means for similar data.

Third, the similarity of effects on regression coefficients and on differences between means can be used. The data seem to show that, as we conjectured, the effects in regression tend to resemble the average ratios we can find more easily for differences of means for similar data. Standard errors for means are easier to compute, and we have considerable amounts of results for them. We can lean gently on those results to buttress our meager grounds for inference for regression coefficients and other analytical statistics.

Fourth, the factor $\sqrt{deff}$ seemed reasonably stable for all the coefficients in these data, across variables and equations. This is useful, because: a) individual standard errors for each variable in each equation would be expensive to compute; b) they would be subject to high variability; c) they would be difficult to present in the results [Kish, 1965, 14.1 - 14.2]. Instead for example, for the results in section 5, we rely on the factor 1.06 to adjust srs estimates of standard errors; or on $1.06^2 = 1.12$ to obtain "effective" sample sizes. Such averaging of standard errors is hazardous; of course, we investigated other methods of averaging computed standard errors; see sections 7 and 9.

Fifth, the design effect should not be assumed to be either small or uniform for all survey results. Our conjectures are to the contrary. When design effects are high for means they tend to be high for differences of means; this tendency is similar for other analytical statistics. The results in section 7 conform well to tnese conjectures. Differences are also found for different statistics from the same data. For example, simple correlations, and partial coefficients may have different effects than the regression coefficients.

We hope to deepen our understanding of the sources and magnitudes of these effects with research, empirical and theoretical, by others and by us. Our present phase is reminiscent of the emergence about 20 years ago of empirical results about design effects on means and aggregates.

5. Regression coefficients for a set of economic variables.

These computations, completed in 1964, were the first large scale set of results on regression coefficients. Standard errors were computed for 20 regression equations in which 7 predictor variables appear in different combinations of 1, 2, 3, or 4 at a time (see Table 5.1). They form the core of a study on Private Pensions and Individual Savings by Katona [1965, see especially Table 30].

The data represent 1,853 interviews obtained from members of the "crucial group," defined as "Complete families (husband and wife living together) with the head in the labor force and aged 35 to 64 with a family income of three thousand dollars or more." [Katona, 1965, p.8]. They constitute a subclass from a sample of 4,700 family units, selected with equal probability on three national surveys conducted in June 1962, January 1963, and June 1963. They came from the national sample of 74 primary sampling areas: the 12 largest metropolitan areas (self-representing), plus 62 other primary areas (other-representing) [Kish and Hess, 1965]. The latter were "collapsed" into 31 strata; from the former the primary units (tracts and blocks) were "combined" into 16 strata. Thus 47 strata of roughly equal size were created for the computations, in close accord with the sample design. In each of the 47 strata a pair of replicates formed the basis for the BRR computations with 48 repetitions.

The mean value of the effects $\sqrt{deff}$ for 60 regression coefficient was 1.0616. The separate values are shown in Table 5.1. When should we use the separate values and when the overall mean? Analysis of variance shows that the small differences between predictors are significant, but those between equations are not. We suggest a useful strategy: compute srs standard error and multiply by the $\sqrt{deff}$ for the predictor averaged over all equations.

Table 5.1 Effects $\sqrt{deff}$ for Standard Errors of 7 Predictors in 20 Regression Equations

Equation	1	2	3	4	5	6	7	Mean
1		1.002						1.002
2	1.030	1.006						1.018
3	1.023	1.115		1.062				1.067
4	1.172	1.011			1.016			1.066
5	1.026	1.027				0.956		1.003
6	1.090	0.970					1.194	1.085
7	1.017	1.078		1.071		0.892		1.015
8	1.070	0.978				0.929	1.149	1.032
9		1.057						1.057
10	1.141	1.048						1.095
11	1.217	0.939		1.165				1.107
12	0.997	1.109			0.966			1.024
13	1.150	1.004				1.415		1.190
14	1.083	1.077					0.986	1.049
15	1.217	0.939		1.165				1.107
16	1.109	1.052				1.096	0.971	1.057
17	1.178	1.001				1.000		1.060
18	1.186	1.068	1.015			1.165		1.109
19	0.925	0.942	0.959			1.122		0.987
20	1.258	0.985	1.029			1.082		1.089
Mean	1.105	1.020	1.001	1.116	0.991	1.073	1.075	1.062

6. Regression of voting on 4 attitude scales.

These data derive from 1,111 voters interviewed in October and again in November of 1964, in as many households selected in the 74 primary areas of the SRC national sample. A regression equation related the party receiving the vote of the respondent (as he stated it in the November interview) to 4 predictors, each of these an attitude scale of 9 points obtained in the October interviews. The sample design and the construction of 48 repetitions based on 47 computing strata were similar to those described in section 5.

Mean values of the effects $\sqrt{deff}$ on standard errors, in Table 6.1, are based on the averages of the BRR values obtained from $(b'_i-\hat{b})^2$ and $(b_i-\hat{b})^2$, the deviations of the half sample and its complement from the overall statistic $\hat{b}$. We also computed the BRR values of $(b_i-b'_i)^2/4$, and the effects from these are given in parentheses. These values are always lower, but only by a factor of 0.999 to 0.995. These small differences are reassuring and meaningful, as will be noted in section 11.

Table 6.1 Effects $\sqrt{deff}$ on Standard Errors of Regression Statistics

i	mean $\bar{x}_i$	b_i	r_{yxi}	$r_{yx_i \cdot x_j x_k x_l}$
1	1.215 (1.214)	0.958 (0.954)	1.013 (1.011)	0.935 (0.931)
2	1.018 (1.017)	1.069 (1.065)	1.104 (1.102)	1.089 (1.086)
3	1.168 (1.167)	1.136 (1.133)	1.298 (1.296)	1.198 (1.196)
4	1.021 (1.020)	0.897 (0.892)	0.970 (0.968)	0.941 (0.936)
Y	1.108 (1.107)	... ...	... ...	... ...
Mean	1.106 (1.105)	1.015 (1.011)	1.096 (1.094)	1.041 (1.037)

The mean effect 1.106 for standard errors of the 5 means is not surprising, though perhaps somewhat on the low side; 1.2 or 1.3 would be more in accord with other data of this type. This study, the smallest of the 5 we present, may be the best place to remind ourselves of the variability of our data; the coefficient of variation of these values is probably over 10 percent, $\sqrt{1/2(47)}$.

The mean effect for the multiple regression coefficient is 1.015, and this may also be on the low side. The mean effect is 1.096 for the first order correlations of the 4 predictors with the predictand, and 1.041 for the partial coefficients. The effect was 1.143 (1.138) for the multiple correlation coefficient, and 1.150 (1.146) for the adjusted coefficient.

There are 6 first order correlation coefficients between the 4 predictors, and the mean of the effects was 1.221 (1.219). (The 6 values were 1.212, 1.030, 1.320, 1.111, 1.265, 1.390.) This value is greater than we expected, and greater than the 1.106 for the means. We have no explanation for this case, and it does not hold in other projects.

To investigate the possible effects of nonnormality we repeated these computations after making a Fisher's z transformation of the coefficients. The differences for all the separate effects were reassuringly small, and the mean effect was the same to 3 decimal points.

We may add here that in our first set of BRR computations in 1957, for regression coefficients for an equation of political attitudes, we found a low mean effect, similar to the 1.015 reported here [Stokes, 1958].

7. 16 Regressions of physiological measurements of the NCHS

For 3,091 males in a national health examination survey, age, height, and weight were used as predictors of 16 physiological predictand variables, each in a separate equation. The sample came from clusters of four per segment, clustered in 42 primary areas [Simmons and Baird, 1968].

The average effect $\sqrt{deff}$ on the standard errors of means, correlations and regression coefficients is given in Table 7.1, col. (2). Individual values of $\sqrt{deff}$ for means, regression coefficients and multiple correlation coefficients are given in Table 7.2. These effects are larger than in our other results, but do not contradict them; on the contrary, this is in line with our conjectures: we expected large effects because of the large primary clusters, and because the effects on the means were large. We believe that the large effects here are not due to the nature of the variables, but to the large clusters of the sample, and possibly to the clustering of measurement errors by the survey teams.

Table 7.1 Effects of Design in 16 Regressions from 3 Predictors.

Each entry in (2) and (3) is the mean of a number in (1) values of $\sqrt{deff}$.

Statistical Type	(1)	(2)	(3)	(4) =(2)÷(3)
Ratio Means	18	1.7998	1.7549	1.0256
Simple Correlations	51	1.2616	1.2802	0.9855
Partial Correlations	48	1.3995	1.3487	1.0377
Multiple R	16	1.4653	1.4217	1.0307
Regression Coefs.	48	1.2948	1.2668	1.0221

Table 7.2 Effects, $\sqrt{deff}$, on Standard Errors of Means, Regression Statistics, and Multiple R's.

Equation	Means (Y_i)	$b(X_1)$	$b(X_2)$	$b(X_3)$	$R_{Mult.}$
1	2.022	1.126	1.169	.89	1.214
2	1.398	1.443	1.630	1.298	.942
3	1.194	1.343	1.522	1.290	.816
4	0.937	1.482	1.267	.986	1.298
5	2.093	1.160	1.441	1.514	1.543
6	1.371	1.359	1.246	1.607	1.866
7	2.023	1.219	1.426	1.564	1.701
8	1.893	1.281	1.233	1.309	1.666
9	1.885	1.361	1.465	1.110	2.036
10	1.404	1.330	1.045	1.486	1.640
11	2.470	1.550	1.374	1.323	1.836
12	1.523	0.935	1.270	1.316	1.496
13	1.833	1.178	0.857	1.147	1.175
14	2.227	1.391	1.220	1.328	2.012
15	1.204	1.453	1.355	1.165	.942
16	3.319	1.106	1.251	1.325	1.261
Mean	1.800	1.295	1.298	1.2914	1.465

These computations gave us an opportunity to investigate an issue related to the estimation scheme employed by the National Center for Health Statistics, which uses 12 age-sex categories for post-stratification. Since the post-stratification scheme is a function of the sample, the estimate based on the half-samples (i.e., the b_i and b'_i's) should be computed using post-stratification weights based on the particular half or complement half sample. Following this procedure we computed estimates of variance using a set of 16 half samples with no complements; values of $\sqrt{deff}$ for these results given in col. (2). Because the reweighting of each half sample is costly, we also computed the less costly variance estimates with weights assigned in the post-stratification of the total sample. Column (3) presents the average $\sqrt{deff}$ computed this way; for these estimates we used 24 replications with complementary half samples. Column (4) gives the ratio of the two methods; the simpler method (no reweighting of half-samples) does in general appear to underestimate the design effect. However, this degree of underestimation may perhaps be tolerated to reduce the cost of computing variances. Furthermore, to the extent that the factor of underestimation is found to be stable at 1.02 or 1.03, it may perhaps be used to adjust the cheaper estimate.

Analysis of variance of the 16 x 3 values in Table 7.2 of $\sqrt{deff}$ for the 48 regression coefficients showed no significance either for the 3 predictors or for the 16 equations. Hence for the standard error of an individual regression coefficient the best strategy may be to use 1.29 times the srs estimate of the standard error.

8. Dummy variables in regression equations

The data represents the sum of three national households samples conducted in August 1962, November 1962, and November 1963. Over 1300 interviews from each survey yielded a total of 3990, from the national sample of 74 primary areas of the Survey Research Center, similar to that in section 5. The samples are described by Lansing and Mueller in The Geographic Mobility of Labor [1967, pp 8-9, 349-358]. Again 48 balanced repetitions were used for 47 strata.

The multivariate analysis used a technique of "dummy variables" to represent nonmetric and nonscaleable predictors, and to overcome nonlinearity in other predictors. Essentially each category of every variable in the regression equation receives a value of 1 for members of the category, and 0 for nonmembers [Suits, 1957]. A standard program for the IBM 7090 (now rewritten for IBM 360) computed estimates of regression coefficients and their standard errors, as well as related statistics [Lansing and Mueller, 1967, pp 47-53, 397-417].

The program's formulas for standard errors were based, as usual, on srs assumptions. The aims of our BRR computations were to compute a set of standard errors which followed closely the complexities of the design, and especially to compute ratios of $\sqrt{deff}$ of the former to the latter. These computations were confined, due to the limitations of the program and the budget, to

6 regression equations, with a total of 64 predictor categories. See Tables 8.1. The mean value of the 64 values of $\sqrt{deff}$ is 1.10. There seems to be a fair amount of variation to be investigated later.

Table 8.1 Effects $\sqrt{deff}$ on Standard Errors of Regression Coefficient for Selected Predictor Classes in 6 "Dummy Variable" Regression Equations.

Selected Classes of Predictor Variables	Regression Equations a1	a2	b1	b2	c1	c2
College, Grad. or some	1.12	1.19	0.86	0.88	1.14	1.20
High School, Grad or some	0.86	0.80	1.02	0.75	...	...
Professional or technical	1.37	1.54	1.15	1.26	...	...
Other white collar	1.29	1.14	1.32	0.92	...	...
Blue collar	0.93	0.95	1.14	1.02	1.30	0.90
Family income > $10,000	1.10	1.44	0.98	0.86	...	...
Family income < $3,000	0.76	0.86	1.14	1.37	...	...
Financial reserves > $1,000	1.29	1.08	...	...	...	...
Financial reserves none	0.97	0.81	...	...	...	...
Unemployment, usual	1.06	1.37	1.28	0.97	...	...
Negro	0.80	0.60	0.93	0.82	0.93	0.80
Home, own or buying	1.03	1.37	0.86	0.71	1.22	1.10
Relatives, all live away	2.06	1.30	1.47	1.26	...	...
Relatives, most live away	1.15	1.20	1.07	1.04	...	...
Friends, all live away	...	...	1.46	1.23	...	...
Friends, most live away	...	...	1.45	1.07	...	...
	1.127	1.118	1.152	1.011	1.145	1.002

To conserve computational costs we selected 16 predictor classes from 9 predictor variables, as they appeared to be relevant. The 6 selected regressions equations are defined by their predictand variables, and by the subclasses on which they are based, as follows:

a) Predictand: Moved in period of one year after the study
   Subclass 1: Age under 30, n=306
   2: Age 35 and over, n=927

b) Subclass of sample: age under 35, n=979
   1) Predictand: Mobility last five years
   2) Predictand: Plans to move in next year

c) Subclass of sample: Age 35 and over, n=2991
   1) Predictand: Mobility last five years
   2) Predictand: Plans to move in next year

9. Statistics from multiple classification analysis (MCA)

A sample of 2214 family heads were interviewed in January and February 1965 in the 74 primary areas of the Center's national sample. One method for multivariate analysis consisted of an MCA equation to relate a "receptivity index" to 6 predictor variables comprising altogether 43 predictor classes; see Productive Americans by Morgan, Sirageldin, and Baerwaldt[1966, pp 360-378, 208-233].

MCA is a multivariate technique used for nonmetric data and to circumvent nonlinearity of the variables. Through iteration it obtains a least square solution for an equation relating the predictand variable to a linear expression of all the predictor classes. It is being utilized increasingly in survey research, where polytomized variables are most common. [Andrews, Morgan, Sonquist, 1967; Hess and Pillai, 1960; Kempthorne, 1952; Hill, 1959; Blau and Duncan, 1967, pp 128-140].

The wide utility and utilization of the model was the prime reason for our interest in the method. It also offered a new challenge for BRR techniques, because analytical expressions are lacking for standard errors of MCA coefficients even under srs assumptions.

Iterative methods for this large matrix were costly for our programs in 1966; our 1968 program will facilitate future computations. Hence we confined our computations to 12 partially balanced repetitions of half sample estimates of standard errors based on the complex sample design. We also wanted to compare these srs estimates in order to compute design effects; to do this we also computed 12 repetitions based on simple random splits of the sample.

One set of outputs of MCA analysis is a set of adjusted deviations; deviations for the predictand variable between the overall mean and the mean for each predictor class, after adjustment for all other variables in the equation. This deviation can also be compared to the raw unadjusted deviation for the same predictor class, thus noting the combined effect of the other variables.

For each variable MCA also yields a beta coefficient that indicates the relative explanatory value of each predictor variable; this is related to the adjusted means for all classes of that variable. Each beta may be compared to an eta for the same variable; its square is the correlation ratio which indicates the proportion of the total variance attributable to the unadjusted means of all classes of the variable.

Table 9.1 presents the computed standard errors for the eta and beta coefficients. The latter are important and seem rather stable in the neighborhood of 0.022. Both the complex and the simple random computations were based on 12 repetitions each. The latter were necessary because formulas are not available. The small number of repetitions we could afford makes these values unreliable; from the 12 strata, we expect roughly a coefficient of variation of $\sqrt{1/2(12)} = 0.2$ for the computed values. The mean values are $\sqrt{deff} = 1.222$ for the 6 values of ste(beta) and $\sqrt{deff} = 1.347$ for the corresponding ste(eta) values. We expect to check these values which appear to be higher than we expected.

Table 9.1 Standard Errors for eta and beta Coefficients in Multiple Classification Analysis. Predictand Is A Receptivity Index.

Predictor Variable	eta	ste (eta)	beta	ste (beta)
Education of Head	.5032	.0186	.1993	.0223
Age of Head	.4098	.0245	.1233	.0215
Total Family Income	.5676	.0165	.3309	.0258
Social Participation	.4296	.0258	.1586	.0232
Achievement Orient.	.3476	.0185	.1201	.0220
Sex & Marital Status	.2884	.0284	.0970	.0209

For the standard errors of the deviations a generalized table was deemed useful because of its simplicity, and because of the high variability of the 43 individual computed values. We conjectured that ste(d) = $a/\sqrt{n}$, where $\underline{a}$ is a constant and $\underline{n}$ the size of the predictor class, may be a fair approximation. For the 43 pairs of values of ste(d) and $\underline{n}$ we fitted a least square line to a log ste(d) = log a−0.5 log n. We thus obtained a value of a = 1.815 for the adjusted deviations. Similar computations for the unadjusted deviations gave a' = 2.441. The generalized table 9.2 for standard errors of deviations is given for relevant values of subclass size $\underline{n}$. The hyperbola $a/\sqrt{n}$ appears to fit well the values of ste(d) plotted against $\sqrt{n}$; similarly for the hyperbola $a'/\sqrt{n}$.

Table 9.2 Average Standard Errors of Deviations, Adjusted and Unadjusted, As Fitted to Two Curves $a/\sqrt{n}$.

n	25	50	75	100	150	200	300	400	500	600	1640
Adjusted $1.815/\sqrt{n}$	.363	.257	.210	.182	.148	.128	.105	.091	.081	.074	.045
Unadjusted $2.441/\sqrt{n}$	.488	.345	.282	.244	.199	.173	.141	.122	.109	.100	.060

It is likely that with more data, precision, and research better approximations will be found. Although we could not discern distinct patterns for diverse predictors, they probably exist. Furthermore, the size of the sample should depend not only on $\underline{n}$ but also on other parameters which allow the design effect to vary with $\underline{n}$, rather than to fix it with the constant $\underline{a}$.

For our results the mean design effect was small compared to errors in measuring it; systematic variations were not detected and can be neglected. For each of the 43 deviations we took the ratio of the two computed values: the actual clustered to the srs values (each value the mean of 12 repetitions). The mean of these $\sqrt{\text{deff}}$ ratios was 1.105 for the unadjusted and 1.025 for the adjusted deviations.

To investigate sources of variability we also fitted least square lines logarithmically to ste(d) = $a/\sqrt{n}$ to estimate values of $\underline{a}$ and $\underline{a}'$ for srs estimates of ste(d), as described above for complex estimates. The estimates of $\underline{a}$ and $\underline{a}'$ are compared in Table 9.3. That table also has values of a reasonable and simple model for the curve $a/\sqrt{n}$; for unadjusted deviations this is conjectured to be $s/\sqrt{n}$ and for the unadjusted deviations $\sqrt{1-R^2}\,s/\sqrt{n}$. The comparison with the computed values is reasonable. We may expect a reduction from the model which, using $\underline{s}$ from the entire sample, assumes random grouping to the mean srs values for meaningful groups.

Table 9.3 Values of the Constants (a) for Three Assumptions for the Curves $a/\sqrt{n}$

	for models	for replications	
		srs	complex
Adjusted a	$s/\sqrt{n}$ = 1.769	1.725	1.815
Unadjusted a'	$\sqrt{1-R^2}\,s/\sqrt{n}$ = 2.387	2.310	2.441

10. Properties of BRR estimates

BRR techniques for estimating variances yield useful approximations for a wide variety of statistics. We present below strong justifications, we believe, for their use. We make some simplifying assumptions necessitated by the present state of the theory. A few are made here merely to keep the exposition brief. Some of these will be weakened in a fuller exposition that will also present some elaboration, and especially the results of investigations into reliability of the technique [Kish and Frankel, 1968].

For simplicity we assume here two primary selections per stratum, selected entirely independently, hence with replacement. Within each of the H strata the two primary selections are replicates of the same selection process for representing the stratum. Then from the sample S we select at random one replicate from each stratum to constitute the replication H_i. The other replicates from each stratum constitute the associated complement C_i. Note that H_i and C_i constitute two replications of the same selection process. Furthermore S also represents the same selection process but doubled in every respect.

The sample estimating function $\underline{f}$ applied to the entire sample yields the estimate $\hat{b} = f(S)$. The same function applied to replication H_i yields $b_i = f(H_i)$; applied to C_i it yields $b'_i = f(C_i)$. Our goal here is to estimate $\text{Var}(\hat{b})$ by using b_i and b'_i; and to improve this estimate by repetitions of the process to constitute b_i and b'_i; etc. These may be viewed as samples from the 2^{H-1} replications that may be drawn from S.

It may help the reader if he can refer to a list of terms:

B is the population value being estimated, and we neglect differences from some true values B_{true} due to measurement errors, nonresponse, etc.

$\hat{b}$ is the statistic used by the researcher for estimating B, and variance $\text{Var}(\hat{b})$ needs to be estimated; we neglect here the possible existence of some better estimator b^*.

$\bar{b}_i=(b_i+b'_i)/2$ is the mean of a replication and its complement.

$b^{(k)}=\Sigma_k b_i/k$ and $b'^{(k)}=\Sigma_k b'_i/k$ are the means of $\underline{k}$ replications and their complements, and $\bar{b}^{(k)}=(b^{(k)}+b'^{(k)})/2 = \Sigma_k \bar{b}_i/k$. The $\bar{b}^{(t)}$ would be $\bar{b}^{(k)}$ based on all of the 2^{H-1} possible replications.

$b^{(k*)}$, $b'^{(k*)}$, and $\bar{b}^{(k*)}$ denote the above based on a balanced repeated replication.

The "linear case", when the estimation function is linear in the replicate values, has notable simplicities. In the linear case $\hat{b} = \bar{b}_i$, and thus $\hat{b} = \bar{b}^{(k)}$. But this does not hold for statistics in general which are nonlinear, in which our interests lie. We are chiefly interested in making estimates about $Var(\hat{b})$ from averages of $var(\bar{b}_i)$ for the nonlinear cases. We also want to relate $Var(\bar{b}^{(k)})$ and $Var(\bar{b}^{(k*)})$ to $Var(\hat{b})$.

We shall denote by $E(b_i)$ the expectation of $b_i = f(H_i)$ for a specific replication H_i over all possible samples with a specified sample design combining a selection process and estimation function. Because of symmetries we have equal expectation for all $\underline{i}$ and $\underline{j}$ and their means:

$$E(b_i) = E(b'_i) = E(\bar{b}_i) = E(\bar{b}_j) = E(\bar{b}^{(k)}) = E(\bar{b}^{(k*)}) = E(\bar{b}^{(t)}). \qquad (10.1)$$

However, $E(\hat{b}) \neq E(\bar{b}^{(t)})$ etc. in general. Furthermore

$$Var(\bar{b}_i) = E\left[\frac{(b_i-b'_i)^2}{4}\right] = E(b_i-\bar{b}_i)^2 = E(b'_i-\bar{b}_i)^2, \text{ and} \qquad (10.2)$$

$$Var(\bar{b}_i) = Var(\bar{b}_j) \text{ for all } \underline{i}, \underline{j}. \qquad (10.3)$$

The expression $(b_i-b'_i)^2/4$ estimates the variance of $\bar{b}_i$. Although $\bar{b}_i \neq \hat{b}$ for the nonlinear case, the difference is probably often slight. However, the estimate is extremely unstable and we must obtain more precision with repetitions. If we repeat the process $\underline{k}$ times we can get the mean variance estimate $\overline{var}_k(\bar{b}_i) = \Sigma(b_i-b'_i)^2/4k$. Since the variance of all replications has the same expectation, we have

$$E[\overline{var}_k(\bar{b}_i)] = Var(\bar{b}_i). \qquad (10.4)$$

Furthermore

$$Var(\bar{b}^{(k)}) = k^{-2}\ [\sum_i Var(\bar{b}_i) + \sum_{i \neq j} Cov(\bar{b}_i,\bar{b}_j)]$$

$$= k^{-2}\ [k\ Var(\bar{b}_i) + k(k-1)\ \rho_{\bar{b}_i,\bar{b}_j}\ Var(\bar{b}_i)]$$

$$= Var(\bar{b}_i)\ \left[1 - [(1-\rho_{b_i,b_j})(k-1)/k]\right] \qquad (10.5)$$

where $\rho_{\bar{b}_i,\bar{b}_j}$ denotes the correlation between $\bar{b}_i$ and $\bar{b}_j$ values. For the linear case $\rho_{\bar{b}_i,\bar{b}_j} = 1$, because all $\bar{b}_i$ are equal. In our investigations we have found for orthogonally balanced replications that $\rho_{\bar{b}_i,\bar{b}_j} \geq 0.98$, when the b's are regression coefficients.

For the linear case McCarthy [1966, 1968], has shown that orthogonally balanced patterns of repeated replication produce estimates of the variance equal to the estimate that would be produced much more laboriously from all 2^H possible replications. In our investigations on nonlinear statistics we used his orthogonal balancing because we believe it to be useful. We have made investigations to relate estimates of $Var(\bar{b}_i)$ and $Var(\bar{b}^{(k*)})$ to $Var(\hat{b})$, and $\bar{b}_i$ and $\bar{b}^{(k*)}$ to $\hat{b}$. The empirical investigations were reassuring and will be presented later [Kish and Frankel, 1968]. In summary, our estimates of $Var(\bar{b}_i)$ overestimate $Var(\bar{b}^{(k)})$, and $\bar{b}^{(k)}$ is generally very close to $\hat{b}$.

We note here a justification for the use of $Var(\bar{b}_i)$ to estimate $Var(\hat{b})$ based on design effects. The half-samples H_i and C_i are constituted to preserve the full complexity of the selection process of the entire sample S. The estimation function $\underline{f}$ is the same for $\hat{b}$ as for b_i and b'_i.

From Deff $(\bar{b}_i)$ = Deff $(\hat{b})$ we have

$$\frac{Var(\hat{b})}{\sigma^2_{\hat{b}}} = \frac{Var(\bar{b}_i)}{\sigma^2_{\bar{b}_i}} \text{ and } \frac{Var(\hat{b})}{Var(\bar{b}_i)} = \frac{\sigma^2_{\hat{b}}}{\sigma^2_{\bar{b}_i}}.$$

$\sigma^2_{\hat{b}}$ and $\sigma^2_{\bar{b}_i}$ are the variances of $\hat{b}$ and $\bar{b}_i$ under the assumption that S is a simple random sample of elements with replacement.

From nonreplicated methods we estimated $\sigma^2_{\hat{b}}$, $\sigma^2_{\bar{b}_i}$ and used their ratio as an estimate of the ratio of $Var(\hat{b})$ to $Var(\bar{b}_i)$. For the regression coefficient described in 4-9 we find that the mean value of this ratio across different regression coefficients was 1.000 with standard error (among statistics) of less than .01.

11. Approximations for $var(b^{(k)})$ and estimates of MSE from BRR

When dealing with estimation functions for linear cases we can take advantage [McCarthy 1966, 1968] of the identity between $\hat{b}$ and $\bar{b}_i$ to use the simpler computational form.

$$\overline{var}_k(\bar{b}_i) = \Sigma_k(b_i-\hat{b})^2/k = \Sigma_k(b'_i-\hat{b})^2/k = \Sigma_k(b_i-b'_i)^2/4k, \qquad (11.1)$$

since $\bar{b}_i = \hat{b}$ and $(b_i-\hat{b}) = (b'_i-\hat{b})$. Here $\underline{k}$ refers to the number of computations needed to obtain BRR. Only $\underline{k}$ computations are needed to obtain the b_i values, because the b'_i values yield no new information. However, for nonlinear cases the strict equality does not hold and computing both b_i and b'_i does yield more information and improve the variance estimate. Note that

$$(b_i-\hat{b})^2 = (b_i-\bar{b}_i + \bar{b}_i-\hat{b})^2 = (b_i-\bar{b}_i)^2$$

$$+ (\bar{b}_i-\hat{b})^2 + 2(b_i-\bar{b}_i)(\bar{b}_i-\hat{b})$$

$$=(b_i-b'_i)^2/4 + e_i^2 + e_i\ (b_i-b'_i),$$

if we remember that $(b_i-b'_i) = 2(b_i-\bar{b}_i)$, where $\bar{b}_i = (b_i+b'_i)/2$ and define $e_i = (\bar{b}_i-\hat{b})$. Then averaging over $\underline{k}$ computations we have

$$\frac{1}{k}\sum_i(b_i-\hat{b})^2 - \overline{\text{var}}_k(\bar{b}_i) = \frac{1}{k}\sum_i e_i^2 + \frac{1}{k}\Sigma e_i(b_i-b'_i), \text{ and}$$

$$\frac{1}{k}\Sigma(b'_i-\hat{b})^2 - \overline{\text{var}}_k(\bar{b}_i) = \frac{1}{k}\Sigma e_i^2 - \frac{1}{k}\Sigma e_i(b_i-b'_i). \quad (11.2)$$

In the linear case both terms vanish. In our investigations the first term is always very small. Although the second terms are not large, they raise the question of strategy for averaging the two sets of variances:

$$[\sum_i(b_i-\hat{b})^2 + (b'_i-\hat{b})^2]/2k = \overline{\text{var}}_k(\bar{b}_i) + \sum_i e_i^2/k. \quad (11.3)$$

If $\hat{b}$ estimates B and if its bias is inversely related to the number of primary selections $\underline{m}$ for a specified design, then for some constant $\underline{a}$:

$$B-E(\hat{b}) = a/2m \text{ and } B-E(\bar{b}_i) = B-E(\bar{b}^k) = a/m$$

and $[E(\hat{b}-\bar{b})]^2 = (a/2m)^2 = \text{Bias}^2(\hat{b})$.

Since $E(\frac{1}{k}\Sigma e_i^2) = E[(\bar{b}_i-\hat{b})^2] \geq [E(\bar{b}_i-\hat{b})]^2$,

the use of $[\Sigma(b_i-\hat{b})^2 + (b'_i-\hat{b})^2]/2k$ will yield a conservative estimate of the mean square error $\hat{b}$.

REFERENCES

Andrews, F., Morgan, J., and Sonquist, J. [1967], Multiple Classification Analysis, Ann Arbor: Institute for Social Research

Blau, P. M. and Duncan, O. D. [1967], The American Occupational Structure, New York: John Wiley and Sons.

Brillinger, D. R., and Tukey, J. W. [1964], Asymptotic Variances, Moments, Cumulants, and Other Average Values, Princeton University: Memorandum.

Cramer, H. [1946], Mathematical Methods of Statistics, Princeton, N. J.: Princeton University Press.

Deming, W. E. [1956], "On simplification of sample design through replication with equal probabilities and without stages," JASA (51), 24-53.

Gurney, M. [1962], "The variance of the replication method for estimating variances from the CPS design," unpublished memorandum, U. S. Bureau of the Census.

Hess, I., and Pillai, R. K. [1960], Multiple Classification Analysis, Ann Arbor: Survey Research Center, Dittoed memorandum.

Hill, T. P. [1959], "An analysis of the distribution of wages and salaries in Great Britain," Econometrics, 27, 355-381.

Katona, G. [1965], Private Pensions and Individual Savings, Ann Arbor: Institute for Social Research, Monograph No. 40.

Kempthorne, O. [1952], The Design and Analysis of Experiments, New York: John Wiley and Sons (p.95).

Kendall, M. G., and Stuart A. [1958], The Advanced Theory of Statistics, Vol. I, London: Griffin and Co.

Kish, L. [1957], "Confidence intervals in complex samples," American Sociological Review (22), 154-165.

Kish, L., and Hess, I. [1959], "On variances of ratios and their differences in multistage samples," JASA (54), 416-446.

Kish, L., and Hess, I. [1965], The Survey Research Center's National Sample of Dwellings, Ann Arbor: Institute for Social Research.

Kish, L. [1965], Survey Sampling, New York: John Wiley and Sons.

Kish, L. [1968], "Standard errors for indexes from complex samples," JASA, (63), 512-529.

Kish, L., and Frankel, M. [1968], "Balanced repeated replications for analytical statistics," submitted to the JASA.

Lansing, J. B., and Mueller, E. [1967], The Geographic Mobility of Labor, Ann Arbor: Institute for Social Research.

McCarthy, P. J. [1966], Replication: An Approach to the Analysis of Data from Complex Surveys, Washington: Public Health Service, Series 2, No. 14.

McCarthy, P. J. [1968], "Pseudo-replication, half samples," for Symposium on Foundations of Survey Sampling, University of North Carolina, Chapel Hill.

Morgan, J. N., Sirageldin, I. A., and Baerwaldt, N. [1966], Productive Americans, Ann Arbor: Institute for Social Research.

Simmons, W. R., and Baird, J. [1968], "Use of pseudo-replication in the NCHS health examination survey," Proceedings of the Social Statistics Section of ASA.

Stokes, D. E. [1966], "Some dynamic elements of contests for the presidency," American Political Science Review, (60), 19-28.

Suits, D. B. [1957], "Dummy variables in regression equations," JASA, (55), 548-551.

Tharakan, T. C. [1968], "Inference based on complex samples from finite population, I," unfinished manuscript, Institute for Social Research, University of Michigan.

Tepping, B. J. [1968], "The estimation of variance in complex surveys," Proceedings of the Social Statistics Section of ASA.

U. S. Bureau of the Census [1963], The Current Population Survey: A Report on Methodology, Technical Paper No. 7, Washington: Superintendent of Documents.