A GENERAL SYSTEM FOR THE EMPIRICAL EVALUATION OF STATISTICAL METHODS FOR DATA FROM COMPLEX SURVEYS

Myron J. Katzoff, Gretchen K. Jones and Lester R. Curtin
National Center for Health Statistics
3700 East-West Highway, Hyattsville, Maryland

1. INTRODUCTION

This paper describes some of the initial work on a system for the empirical study of statistical methods that can be used for investigating both survey design and analytic methods for complex health surveys. The general system involves the creation of a finite population universe and computer software which will permit repeated application of sampling procedures or analytic techniques to simulate the effect of a particular survey design. Such data-based empirical investigations have been carried out in the past by, among others, Frankel (1971) and Bean (1974). Because the major research focus will be health surveys, the finite population will consist of all the sample records from the National Health Interview Survey (NHIS) for 1979, 1980 and 1981.

Once a general system has been developed, survey design issues which can be investigated include the selection of variables for the construction of strata, the clustering criteria for the construction of strata, and the effectiveness of substratification or special sampling procedures for small, hard-to-interview or special emphasis population groups. Estimation issues, closely related to the design issues, that can be examined include consideration of alternatives for truncating weighting adjustments to improve the finite population estimates, considerations of the effectiveness of unit and item nonresponse adjustment procedures in relation to the objectives of an analysis, and the general statistical considerations related to selection of procedures that produce design-based estimates with minimum mean-square-errors.

Initially, areas of analytic concern will focus on the effect of survey design on some standard types of analysis for NHIS data: contingency table and subpopulation-mean comparisons, regression analysis, and generalized linear model analysis. Although much has been written about these subjects, it is still necessary to evaluate the efficacy of various analytical proposals for the particular surveys designs used by NCHS.

This paper represents a summary of current NCHS activities. The effort to date consists of development of the finite population, investigation of a clustering algorithm for forming PSU strata, and the intial considerations in the selection of PSU's for simulated survey designs.

2. DEVELOPMENT OF FINITE POPULATION

Data from the 1979, 1980, and 1981 National Health Interview Survey (NHIS) were pooled to form the sampling universe. The design of the survey for those years was an essentially self-weighting design. As such, it was necessary to use several years of data to yield a sufficient number of records for small population subgroups, such as Blacks and Hispanics.

The combination of three years of NHIS data gave a universe of approximately 320,000 persons, initially partitioned into the 376 PSU's of the 1979-81 NHIS design, 30,000 segments (compact clusters of households), and 100,000 households. The sample design for the NHIS for 1979-81 is summarized in Kovar and Poe (1985). However, a new survey design for NHIS was implemented beginning in 1985 (Massey, et al., 1989) and the NHIS will again be redesigned after the 1990 Census. The number of PSU's for the 1990's is subject to investigation. In order to investigate design options, the strata on the current file must be reformed. Thus, the inital work on the simulation system, described below, is oriented towards strata construction.

Because of the size of the population file, it was recognized that the simulation system should use some type of random access. Random access markedly decreases the proecessing time to select samples from the universe. The file was initally developed on the Model 204 data base and described in Jones and Sloss (1986). However, a further consideration was to make the system available to other researchers. This implies the need for mainframe portability. At this time, the population file exists on tape and software developement will proceed with the requirement that random access and other applications software must be farily portable.

The NHIS has a large number of descriptive and analytic variables. To reduce the size of the data file, a selection of variables was required for the simulated finite population. This was complicated somewhat because the variables for chronic conditions are different for each one-sixth subsample of the NHIS. The file must contain an indicator for which of the one-sixth subsamples (the "Chronic Condition Subsample" code) as well as the condition code itself. A list of variables included is shown in table 1. The chronic condition subsamples are shown in table 2.

3. BASIC SURVEY DESIGN REQUIREMENTS

For this purpose we have constrained our simulated surveys to resemble a current NHIS in some, but not all, respects. In particular:

(1) Each survey design implementation provides for a large number of Primary Sampling Units (PSU) to be chosen with certainty, the so-called self-representing (SR) PSUs; and a small number of PSUs (e.g., two) to be chosen from each of a large number of strata with a few PSUs, the so-called non-self-representing (NSR) PSUs.

(2) The strata containing NSR-PSUs were created independently in each Census region by a

hierarchical clustering procedure (described below).

(3) The simulation employs a method for selecting two PSUs per stratum without replacement with probability proportional to total PSU population (Brewer's method).

(4) There are two stages of sample selection with the second stage being clusters of housing units (segments) chosen with equal probability.

For the simulated surveys, the following tasks must be performed:

(i) construction of PSU strata;
(ii) selection of PSUs within strata;
(iii) determination of sample size and definition of a procedure for allocation of sample size to strata and PSUs;
(iv) development of estimators and a sample-data weighting procedure;
(v) formulation of variance estimators; and
(vi) calculation of the empirical distributions of estimates and variances for some basic health characteristics.

To date, we have worked on only the first three tasks. In the sections that follow, we discuss the techniques used in those tasks and we describe some of the outputs for simulations with the 1979-81 NHIS data.

4. CONSTRUCTION OF PSU STRATA

It may surprise no one familiar with survey sampling methodology that this task has so far been the most time-consuming. To facilitate the process of constructing strata, we first prepared tabulations by PSU of:

(1) total numbers of persons by race and sex for the age classes: 17-21,22-29, 30-39, 40-49,50-64,65+;
(2) mean household income;
(3) mean doctor visits per household; and
(4) mean short stay hospital days per household.

Direct inspection of these printouts suggested that a crude determinant of health status, as measured by (3) and (4), should be age distribution. Accordingly, we collapsed the age-sex-race distribution for each PSU on age and selected a technique for forming strata of NSR-PSUs by clustering them on these age distributions. We felt that 50-70 strata should be adequate for the 429 PSUs of our universe. The number of strata created for each region was determined by the proportion of 17+ pseudo-universe population in the region. The rough and ready rule for determining SR-NSR status was this: a PSU was designated SR when its population exceeded one-half the projected average stratum population for the region. Table 3 shows the results of the first step (creation of the SR strata) in the construction of PSU strata.

At least one clustering procedure using measures of dispersion or distance between the members of a set is now available in the Statistical Analysis System (SAS); for example, the one by Ward (1963), which would have been of special interest if it had been appropriate for frequency counts. Another "distance" method, which is not contained in SAS, is that by Friedman and Rubin (1967). However, that clustering procedure seems to require non-trivial modifications for simple applications; and, in an "asymptotic" sense, its form suggests the use of a chi-squared type of distance for proportions. This is a consequence of the facts that Friedman and Rubin work with Mahalanobis distance and the usual maximum likelihood estimates of multinomial proportions are asymptotically multinormal.

The latter observation led us to consider a computationally convenient asymptotic equivalent of the chi-squared criterion, the Kullback-Liebler (KL) divergence described in Kullback (1968). A brief statement of the KL-Divergence Clustering Algorithm is given below. The result of applying the algorithm in the form it is stated was that the population cutoff value was frequently exceeded by a substantial amount at stratum declaration in step (4) of this algorithm.

The consequences of this were that the last stratum formed had a population size significantly less than the others for the region; and sometimes fewer than the expected number of strata were formed. In subsequent executions of the algorithm, we applied a randomized rule for the inclusion of the last PSU: If Δ denotes the population for the last PSU before stratum declaration, S_E is the population cutoff value, and S_1 is stratum size immediately before stratum declaration, then include the last PSU in the stratum with probability $(S_E - S_1)/\Delta$; otherwise, exclude it. The reader may wish to note that this rule has the property that the expected stratum size will be S_E.

The reason for introducing the randomized rule was to create some stratification options which might not have the deficiencies mentioned by occasionally excluding the last PSU (or subcluster) at stratum declaration. We addressed this by aiming for uniformity in stratum sizes which we assessed by examining the sum of squared deviations from a desired stratum size. We chose the stratification which produced the smallest value for this criterion in 40 trials. Of course, if only a small number of strata are desired, it is possible to discard randomization and proceed in a completely deterministic manner but the computational burden becomes quite heavy when the required number of strata is not especially large (as, for example, in the South region where 12 NSR strata are desired).

<u>KL-Divergence Clustering Algorithm</u>

The KL-divergence clustering algorithm can be formulated in the following four steps:

(1) Let I denote the number of NSR-PSUs for the region under examination. We begin with I clusters, the PSUs themselves. The first operation is to compute for each of the distinct pairs of PSUs, the KL divergence number

$$J(1,2) = \sum_{1 \leq k \leq 6} (N_{1k}/N_1 - N_{2k}/N_2) \ln(N_{1k}N_2/N_{2k}N_1)$$

where N_{ik}, for $i=1,2$, denotes the number of persons in PSU i that belong to age category k; and $N_i=N_{i1}+N_{i2}+...+N_{i6}$.

(2) Find the pair of PSUs for which J(1,2) is a minimum, say the pair $\{i_0,j_0\}$ (with $i_0<j_0$). For this pair redefine $N_{i0,k}$ and N_{i0} by the equations

$$N_{i0,k}=N_{i0,k}+N_{j0,k} \qquad \text{for } k=1,2,...,6$$

$$N_{i0}=N_{i0}+N_{j0}$$

(3) Repeat the calculation of the KL Divergence given in (1) for the remaining pairs of clusters. Determine the pair of clusters $\{i_1,j_1\}$ for which J(1,2) is a minimum and redefine the quantities $N_{i1,k}$ and N_{i1} by

$$N_{i1,k}=N_{i1,k}+N_{j1,k}$$

$$N_{i1}=N_{i1}+N_{j1} \quad .$$

4) For the remaining clusters, calculate the KL divergence for each distinct pair and determine the pair of clusters for which J(1,2) is a minimum, say $\{i_2,j_2\}$. Compute

$$N_{i2,k}=N_{i2,k}+N_{j2,k}$$

$$N_{i2}=N_{i2}+N_{j2}$$

whenever the N_i for a cluster reaches the minimum value greater than or equal to the cutoff for the region, the NSR 17+ population divided by the number of NSR strata required, exclude that cluster from any further consideration. Such a cluster is to be declared a stratum. Step (4) is to be repeated until no more strata can be formed.

5. SELECTION OF PSUs

After forming the strata of NSR-PSUs, software was developed for selecting two PSUs from each stratum without replacement with probabilities P_i proportional to the stratum population contained by a PSU. We used Brewer's procedure which is discussed in Brewer and Hanif (1983) and Brewer (1975). According to Brewer's procedure, the first PSU is selected with working probabilities

$$C^{-1}P_i(1-P_i)/(1-2P_i) \qquad (1)$$

where C is just the sum of the quantities following C^{-1} in expression (1) and i indexes the PSUs in the stratum. If the first PSU chosen is the k-th, the second PSU is chosen from those remaining with probabilities $P_i/(1-P_k)$.

Application of Brewer's procedure yields inclusion probabilities $\pi_i=2P_i$ and joint selection probabilities of selecting PSUs i and j

$$\pi_{ij}=C^{-1}P_iP_j[(1-2P_i)^{-1}+(1-2P_j)^{-1}]$$

As Cochran (1977) pointed out, for the selection of two PSUs per stratum, the methods of Brewer, Rao and Durbin all produce the same π_i and π_{ij}.

6. SAMPLE SIZE, ALLOCATION AND SELECTION OF SEGMENTS

The sample size for simulations was determined so that there would be an absolute error not greater than 4% in the estimation of a proportion under simple random sampling if equal size samples were drawn from each region. Allowing for a design effect of as much as 2.00, we concluded that a sample size of 5000 persons would be adequate.

The sample size was spread proportionately across the SR-PSUs and NSR strata according to the pseudo-universe populations they contained. For each NSR stratum, the sample size was further proportionately allocated by the population of each of the two PSUs chosen.

The average number of persons per segment for each PSU was then used to convert number of sample persons to number of sample segments. If the average number of persons per segment is nearly constant, this allocation scheme produces sample weights which are approximately proportional to the inverses of the relative sizes of the PSUs chosen in each stratum. Thus, dispersion in the sample weights under this allocation scheme will be a reflection, largely, of the variation in the population sizes contained by the PSUs of each stratum. Contrast this situation with that of the weights of a self-weighting scheme in which the weights for a stratum would be constant and the sample size would be split equally between the two PSUs chosen for each stratum: heavily populated PSUs would be lightly sampled and lightly populated PSUs would be heavily sampled.

The procedure for selecting second stage units (segments) is method 1 in Sunter (1977). This is a sequential procedure for SRS sampling which can be implemented with a single pass of a sequential file. Therefore, it eliminates the need for a large amount of computer storage, an important consideration in any large-scale simulation. If M denotes the number of segments in a PSU and m, the number of segments to be chosen without replacement, then the first segment is chosen with probability m/M and, thereafter, segment j is chosen with probability $(m-m_j)/(M-j+1)$, where m_j is the number of units already selected prior to consideration of segment j. The process moves sequentially from one unit to the next and terminates when $m_j=m$.

A transparent feature of this sampling technique is that one can assign index values to the M units of a PSU and, then, generate randomly as many subsets of indices of size m in ascending or descending order as desired, place these on a file and rapidly calculate the usual sample statistics for a large number of second stage samples.

7. DISCUSSION

This paper presented a brief overview of the work to date in developing a general system to be used in empirical studies associated with complex surveys. At this time, the population universe has been completed, the clustering algorithm has been implemented, and one particular stratification has been generated. It is anticipated that future development will include the necessary software to implement a variety of stratifications and stage-wise selection mechanisms. This will allow rapid evaluation of survey design options.

In addition, the results of each simulated sample can be stored in order to examine the empirical distribution of specified statistics. This should be extremely helpful, not only from the standpoint of evaluating the performance of certain analytic methods in a variety of sample design contexts but also from the standpoint of selecting survey designs that are most appropriate for detailed analysis of the collected data.

REFERENCES

1. Bean, J.A. (1974). "Estimate and Sampling Variance in the Health Interview Survey." Vital and Health Statistics, Series 2, No. 38. National Center for Health Statistics, DHEW Pub. No. (HRA) 74-1288. Washington. U.S. Government Printing Office.

2. Brewer, K.R.W. (1975). "A Simple Procedure for Sampling πpswor". Australian Journal of Statistics, v.17, no.3, pp. 166-72.

3. Brewer, K.R.W. and Hanif, M. (1983). Sampling with Unequal Probabilities. Lecture Notes in Statistics, Springer-Verlag, New York.

4. Cochran, Wm. G. (1977). Sampling Techniques, 3rd Edition. Wiley, New York.

5. Frankel, M.R. (1971). "Inference from Survey Samples." Ann Arbor, Institute for Social Research, University of Michigan.

6. Friedman, H.P. and Rubin, J. (1967). "On Some Invariant Criteria for Grouping Data", Journal of the American Statistical Association, v.62, pp. 1159-1178.

7. Jones, G.K. and Sloss, R.W. (1986). "Use of a Data base Management System with Hierarchical Random Access for a Population Simulation Study." Presented at the International Biometrics Conference, Seattle, Washington.

8. Kovar, M.G. and Poe, G.S. (1985). "The National Health Interview Survey Design, 1973-84, and Procedures." Vital and Health Statistics, Series 1, No. 18. National Center for Health Statistics. DHHS Pub. No. (PHS) 85-1320. Public Health Service. Washington, D.C. U.S. Government Printing Office.

9. Kullback, S. (1968). Information Theory and Statistics. Wiley, New York.

10. Massey, J.T., Moore, T.F., Tadros, W., Parsons, V.L. (1989). "Sample Design and Estimation Procedures for the National Health Interview Survey, 1985-94." Vital and Health Statistics, Series 2 (to appear),

11. Ward, J.H. (1963). Hierarchial Grouping to Optimize and Objective Function", Journal of the American Statistical Association, v.58, pp. 236-244.

12. Sunter, A.B. (1977). "List Sesquential Sampling with Equal or Unequal Probabilities Without Replacement", Applied Statistics, v.26, pp. 261-268.

TABLE 1. DEMOGRAPHIC, DESIGN AND HEALTH VARIABLES ON THE FINITE POPULATION UNIVERSE FILES

DESIGN VARIABLES

Sequence Number
PSU
Data Year (79, 80, or 81)
Quarter
Segment Number
Household Number
Person Number
Interview Week
Pseudo PSU

DEMOGRAPHIC

Place of Residence (SMSA, non-SMSA)
Age
Race
Sex

SOCIO-ECONOMIC

Income
Education of Head
Education (adult only)
Telephone availability

HEALTH

Restricted Activity Days (2 weeks)
Doctor Visits (2 weeks)
Dental Visits (2 weeks)
At least one Doctor Visit (12 months)
Doctor Visits (12 months)
Short Stay Hospital Days (12 months)
Interval since last Doctor Visit
Height (17 + years only)
Weight (17 + years only)
Chronic Condition Subsample
Condition Code

TABLE 2. CHRONIC CONDITION SUBSAMPLES

Chronic Condition Sub Sample	Recode Used	Chronic Condition
(Selected Chronic Digestive Conditions)	163	Ulcer of Stomach Duodenum
(Selected Chronic Slan and Musculo-skeletal Contitions)	212	Arthritis, not elsewhere classified
(Chronic Conditions of the Genitourinary, Nervous, Endocrine, Metabolic and Blood Forming Systems and Other Selected Conditions)	090	Diabetes
(Selected Impairments)	X05 - X09	Hearing Impairments
(Selected Chronic Circulatory Conditions)	133	Hypertensive Disease not elseware classified
(Selective Chronic Respiratory Conditions)	151 - 152	Chronic Bronchitis or Emphysema

TABLE 3. RESULTS OF FORMING SELF-REPRESENTING PSUs

REGION	TOTAL NO. PSUS	NO. SR STRATA	NO. NSR STRATA	POP. SR STRATA	POP. NRS STRATA
Northeast	90	9	5	23221	29347
North Cen.	119	7	10	19241	42415
South	165	11	12	17212	57823
West	55	5	4	15509	28276