1. INTRODUCTION

This paper is to serve two purposes:

(1) to report on our progress to date in developing software for empirically studying the effects of complex survey designs on analyses of health survey data;

(2) to present some results from a first empirical study, in which we examine the properties of the BRR and Jackknife variance estimators; and

(3) to outline our approach to a second empirical study involving the comparison of vector means for subdomains.

In our earlier paper (Katzoff, et al 1988) we listed six major survey-design and analysis tasks that we shall expect the software we develop to perform:

(i) construction of primary sampling unit (PSU) strata;
(ii) selection of PSUs;
(iii) allocation of sample size;
(iv) generation of sample weights and first order estimates;
(v) computation of variances and their sample-based estimates; and
(vi) creation of empirical distributions of estimates, especially, estimates of variances, test criteria or other statistics of interest.

At the time of our previous paper, we had completed work on one set of options for tasks (i), (ii) and (iii) only. Much work to embed more investigative options in those portions of the software remains. (For example, we expect that further development of the software will provide for the examination of the consequences of applying a probability minimum replacement sampling technique (Chromy, 1986) and the unequal probability sampling scheme of Saxena, Singh and Srivastava (1988). However, we have given highest priority to developing at least one software option for (iv), (v) and (vi) because we felt that it was necessary to demonstrate our approach while interest remains high, to further the identification of problem areas and to permit development of software options in parallel with the conduct of some empirical studies.

2. DESIGN OF GENERAL SYSTEM FOR EMPIRICAL EVALUATIONS

To reflect the structure of current survey designs, we created a first design with a large number of strata consisting of individual counties, groups of contiguous counties and large metropolitan areas (SMSAs). Since these geographical units are chosen at the first stage of sample selection, we refer to them as primary sampling units (PSUs). Some PSUs are taken with probability one; they are called self-representing (SR) PSUs. PSUs chosen with probability less than one are called non-self-representing (NSR) PSUs. NSR PSUs have been aggregated into strata from each of which two PSUs are drawn for each simulation.

Before addressing the problems of computing sample weights and first order estimates, we considered addition of an option for a self-weighting design. For initial empirical work, we had originally only allowed for allocation of the sample in proportion to the PSU populations. For strata where PSUs were not chosen with certainty, this would result in a situation where the size of the sample drawn from a NSR PSU would depend upon which other PSU was included in the sample. It was easy to add this option and we felt it was important to do so since this would permit us comparisons with results from a common survey design practice. In general, we expect that sample allocation methodology for multistage designs for situations where the allocation of sample depends upon from which clusters the observational units are to be drawn will receive more attention in the future.

The sample weights are, of course, the inverses of the total probabilities of inclusion for each unit. They reflect a decision that, at the second stage of selection, clusters of housing units (called segments) were chosen with equal probabilities. To provide for the possibility of investigating some item and unit nonresponse patterns which depend upon membership in various subpopulations, sample weights are assigned to individual records.

Random Selection of PSUs. In the strata of NSR PSUs, the PSUs have been chosen with probabilities proportional to the amount of population in those geographical units according to Brewer's method (Brewer, 1985) as recommended by Pfefferman (1986).

In using Brewer's procedure the NSR PSU strata have been defined and the proportion of the stratum population in the I-th PSU, P, has been calculated for each value of the index I. Let the values of P be stored in a vector VECP of dimension D, equal to the number of NSR PSUs in the stratum.

These specifications result in a selection of two PSUs per stratum. The operations given in the next section must be repeated for each stratum:

(1) Calculate the entries of a vector VECP, of the same dimension as VECP, as follows:
   a. for each value of I=1,2,...,D, compute VECP(I)=P(1-P)/1-2P, (Note that P=VECP(I).
   b. calculate S, the sum of the coordinates of VECP.
   c. normalize the values of VECP(I) by dividing each value by S; in other words, replace the value of VECP(I) computed (1) a with VECP(I)/S. (2) Select a random number, RI, between zero and one and choose unit I if
I is the smallest index for which
R1<VECP1(1) + VECP1(2) + ... + VECP1(I).
Record and store the index of the unit
selected, IO say.

(3) Calculate the entries of a vector \text{VECP2}
also of dimension D as follows:
\begin{itemize}
  \item a. VECP2(1)=VECP(1) if I is not equal to IO;
  \item b. VECP2(1)=VECP(1) if I is equal to IO;
\end{itemize}

Record and store the index of this unit,
the second unit selected as the value of JO.

(5) Compute the joint inclusion probability
of units IO and JO as
\text{JOINT}=VECP1(10)*VECP2(10) + VECP1(10)*RATIO
where \text{RATIO}=VECP(10)/(1-VECP(10)).

Allocation of Sample and Selection of
Segments Under Proportional Allocation to PSUs by
Size. This section is concerned with the process
of distributing the sample across strata, PSUs,
and segments. For our purposes, a sample size of
5,000 was considered adequate.

(1) Begin by determining a sample size for each
stratum according to this formula:
\[ n(s) = \text{(population of stratum)} \times 5000 \]
\text{universe population
}

(2) Next, proportionately allocate the sample size
to the two PSUs selected for stratum s; that is,
compute
\[ n(1,s) = \frac{P_1}{P_1+P_2} \times n(s), \]
the sample size for the first PSU selected from
stratum s, and
\[ n(2,s) = \frac{P_2}{P_1+P_2} \times n(s), \]
the sample size for the second PSU selected from
stratum s, where \( P_1 \) and \( P_2 \) are the population
sizes for the first and second PSUs respectively.

(3) Finally, for \( i=1,2 \), determine the number of
segments to be selected from each PSU from this
expression:
\[ n'(i,s) = \frac{g_i(s)}{P_i} \times n(i,s) \]
with \( n'(i,s) \) rounded to the nearest integer, where
\text{g}_i(s) \text{ is the number of segments in the i-th PSU
chosen from stratum s.}

[When \text{g}_i(s) \text{ is a constant for all i and s].

It can be seen that the above allocation scheme
produces sample weights which are proportional
to the inverses of the relative sizes of the PSUs
chosen in each stratum only. In a self weighting
scheme, where the sample weights would be constant
throughout, \( n(1,s)=n(s)/2 \) if the constant is just
the inverse of the sample size divided by the
universe size.

Self-weighting selection scheme. The purpose
of this addition is to provide the user with the
option of examining a self-weighting design.
Using the earlier notation for SR PSUs \( n'(s) \) is
unchanged. For NSR PSUs, \( n'(s) = n(s)/2 \) and
\[ n'(i,s) = \frac{g_i(s)}{P_i} \times n(i,s). \]
Here we use the earlier definitions: (1) \( g_i(s) \)
equals the number of segments in the i-th PSU; and
(2) \( P_i \) is the population size for the i-th PSU.
Observe that since \( n'(i,s) \) here is not a function
of which \textit{pair} of PSUs were selected at the first
stage, you should always get the same value for
\( n'(i,s) \) for the i-th PSU. Therefore, the bias-
correction term discussed below should equal zero
for unbiased variance estimates when the self-
weighting design option is selected.

Selection of segments. The procedure for
selecting segments is method I for sequentially
selecting a SRS (Sunter, 1977). The procedure is
to be applied independently to each PSU.

Let \( M \) denote the number of segments in a PSU
and let \( m \) be the number of segments to be selected
at random (i.e., \( m=n'(i,s) \) for some i and s, in
the previous notation). For clarity suppose that
the \( M \) segments from which we are to draw a sample
have been serially assigned index values
1, 2, ..., \( M \). (In practice this is not necessary of
output as a selection matrix for the linear estimates,
in most cases of interest it should be easy to compile variance estimates with the help
of Taylor series linearization.

Design based variance. Under the design
currently programmed, the textbook formula
(Cochran, 1987 p. 301-2) for the NSR component
of total variance from NSR stratum s when the sample
sizes are nonrandom is given by
\[ V_{1,s} = \pi_{1,s} + \pi_{2,s} - \pi_{1,2,s} - \pi_{1,2,s} \]
\text{where
\begin{align*}
A &= \sum_{i=1,2} \sum_{s} n'(i,s) \pi_{i,s} \left[N'(i,s)-n'(i,s) \right] B_i \sum_{i=1,2} n'(i,s) \left[n'(i,s)-1 \right] \end{align*}

and
\begin{align*}
B_i &= \sum_{k=1} \left( \bar{Y}_{iks} - \bar{Y}_{i1s} \right)^2 n'(i,s) \left( \bar{Y}_{iks} - \bar{Y}_{i1s} \right) \end{align*}
Here, for sample PSU indices \( i=1,2 \), \( \bar{Y}_{iks} \)
is the weighted sum of the sample values for sample
segment k of sample PSU i drawn from NSR stratum s
and
\begin{align*}
\pi_{1,s} &= 2P_{1,s}, \text{ where } P_{1,s} \text{ is the proportion of}
\text{stratum s population that belongs to sample PSU 1 drawn from stratum s.}
\end{align*}
When the sample sizes at the second stage depend upon which PSUs are included in the sample, this bias correction term must be added to $V_{ls}$ to obtain unbiased variance estimates

$$\sum_{i=1}^{n'} \left[ n'i(s) \right]^2 \frac{2-\pi_{is}}{\left( n'i(s) - 1 \right) G}$$

where

$$G = \sum_{i=1}^{n'} \left( \bar{y}_{iks} - \bar{y}_{is} / n'i(s) \right) \left( \bar{y}_{iks} - \bar{y}_{is} / n'i(s) \right)'$$

and

$$n'i(s) = \sum_{j \in A(i)} (1 / \pi_{is}(i,j)) \frac{\pi_{ij,s}}{\pi_{is}}$$

In the last two formulas $n'_i(s)$ denotes the size of the sample drawn from PSU $i$ of stratum $s$ if the sample contains PSUs $i$ and $j$. $\pi_{is}$ and $\pi_{ij,s}$ are defined above as $\pi_{is}$. The component of total variance that arises from the SR PSUs is the sum over $s$, as $s$ runs over the SR-strata, of

$$V_{zs} = D \left( \frac{Y_{zs} - Y_s / n'_z}{n'_z} \right) \left( \frac{Y_{zs} - Y_s / n'_z}{n'_z} \right)'$$

where

$$D = \frac{n'(s)N'(s) - n'(s)}{N'(s)[n'(s) - 1]}$$

$n'(s)$ is the number of segments drawn from the $s$-th SR stratum and $N'(s)$ is the total number of segments contained by that stratum. $Y_{zs}$ and $Y_s$ are self-representing analogues of $\bar{y}_{iks}$ and $\bar{y}_{is}$ -- the subscript $i$ is unnecessary for the SR PSUs because they are taken with certainty and are, therefore, strata themselves.

In some empirical work it may be necessary to have the covariance matrix of the estimators and not just an estimate of that covariance matrix. The covariance matrix of the estimators is

$$V_s = \sum_{s \in \text{NSR}} Q_{zs} + \sum_{s \in \text{SR}} Q_{zs}$$

where, dropping the $s$ subscript to simplify the notation,

$$Q_1 = \sum_{i=1}^{n_1} \left( \frac{N_1}{n_1} \right)^2 \left( \frac{1}{N_1} \right) S_i^2 + \sum_{i=1}^{n_1} \left( \frac{N_1}{n_1} \right)^2 \left( \frac{1}{N_1} \right) S_i^2$$

and

$$Q_2 = N'(n' - n')^2 / (n')$$

where $N$ is the number of PSUs in the NSR stratum $p_1$, where (as previously described) $p_1$ is the PSU population expressed as a fraction of total stratum population. $N'$ or $N'_s$ is the number of segments in an SR-PSU or the $i$-th NSR-PSU of an SR-stratum respectively.

$n'$ is the sample size allocated to an SR stratum and $S_2$ or $S_i^2 = \left( \frac{N_1'}{n_1'} - 1 \right)^{-1} \sum_{k=1}^{n_1'} \left( y_{ik} - \bar{y}_{i} / N'_1 \right) \left( y_{ik} - \bar{y}_{i} / N'_1 \right)'$ for $y_{ik}$ equal to the $k$-th segment total and

$$Y_i = \sum_{k=1}^{n_i} y_{ik}$$

BRR and Jackknife Variance Estimates. The purpose of this first empirical study is to compare the distributional properties of the BRR and Jackknife estimators with those of the design-based unbiased estimates of variance. Our initial study of estimators of variance examined those for linear estimates of finite population totals. In so doing we hoped to avoid the confusion that can result from trying to deal with a lot of complexity in the very beginning. The variability and bias of the different variance estimators are of special interest to us.

Because the variance estimators are quadratic forms, we thought it would be useful to summarize the numerical results in two ways instead of just one: (1) computation of the first few central moments of the empirical distributions; and (2) finding a Satterthwaite approximation to a multiple of a chi-squared variable.

In producing the BRR and Jackknife estimates, we largely followed Rust (1985). To meet the requirements for the use of these procedures, the segments of the SR-PSUs were randomly assigned at each iteration of the simulation to one of two groups and the weighted sums of the variables for each group were formed. Similarly, for each NSR-PSU, a weighted sum of variables was calculated. For each stratum, $s$, we denote the sum for group $i$ or PSU $i$ by $Y_{is}$. (Note that for NSR strata, $Y_{is}$ is as previously defined.)

There are several ways of implementing the BRR variance estimation procedure. We now
describe the approach we used:
(1) compute the half-sample estimators of totals
\[ \bar{Y}_s = \frac{1}{60} \sum_{s=1}^{60} (2\delta_{s1,s} \bar{Y}_{1s} + 2\delta_{s2,s} \bar{Y}_{2s}) \]
where \( \delta_{s1,s} = 1 \), if PSU or group 1 of stratum \( s \) chosen
for the half-sample
\( \delta_{s1,s} = 0 \), otherwise
(2) put
\[ \hat{V} = \frac{1}{60} \sum_{s=1}^{60} (\bar{Y}_{1s} + \bar{Y}_{2s}) \]
and calculate the variance estimate
\[ \hat{V}_{\text{BBR}} = \frac{1}{60} \sum_{s=1}^{60} (\bar{Y}_s - \hat{Y})^2 \]
For the jackknife estimators we define jackknife totals
\[ \hat{V}_{(i,s)} = \frac{1}{60} \sum_{u \neq s} (\bar{Y}_{1u} + \bar{Y}_{2u}) + 20 \hat{Y}_{1s} \]
for \( j \) not equal to \( i \); \( i,j=1,2; \) and \( s=1,2, \ldots, 60 \).
The jackknife estimate of variance is then
\[ \hat{V}_j(Y) = \frac{1}{2} \sum_{s=1}^{60} (\bar{Y}_{(i,s)} - \hat{Y})^2 \]

4. COMPARISON OF SUBDOMAIN MEANS

As an illustrative case, we anticipate undertaking a detailed study of the comparison of vector means for short-stay hospital days in accordance with the age breakdown: 17-25, 26-32, 33-45, 46-58 and 58+, for white and nonwhite groups. The discussion of this section describes how we plan to use the basic linear estimates and their estimated and expected variances and covariances in that study. For now there is a limit on the dimensionality or size of vectors that can be considered in multivariate comparisons. This is due to our limiting the number of variables for which we can compute a variance-covariance matrix to twenty. However, we may later endeavor to circumvent this limitation somehow by using the observation that if \( X \) and \( Y \) are vector random variables, all the necessary variances and covariances for \( Z=(X',Y')' \) can be obtained from suitable operations with \( \text{Var}(X), \text{Var}(Y) \) and \( \text{Var}(X+Y) \).

In what follows we assume that algorithms for calculating the basic (or linear) estimates and their variances and covariances have been developed and are available for use without revision. The outputs of those algorithms are essential inputs for operations described here. In particular, it is assumed that we can now quantify the vector of basic estimates
\[ E'=(W_1, \ldots, W_5, PW_1, \ldots, PW_5, O_1, \ldots, O_5, PO_1, \ldots, PO_5) \]
and their covariance matrix, \( V \), where for \( i=1,2, \ldots, 5 \), \( W_i \) denotes the estimated number of short stay hospital days for white persons in age category \( i \); \( PW_i \) denotes the number of white persons in age category \( i \); \( O_i \) denotes the estimated number of short stay hospital days for nonwhite persons in age category \( i \); and \( PO_i \) denotes the estimated number of nonwhite persons in age category \( i \). We will have need of the 20x5 "multiplier" matrix for the estimates, \( M \), defined as four stacked diagonal matrices
\[
D[\begin{pmatrix} (PW_1)^{-1} & (PW_2)^{-1} & \ldots & (PW_5)^{-1} \\ (PW_1) & (PW_2)^{-1} & \ldots & (PW_5)^{-1} \\ (PW_1) & (PW_2) & \ldots & (PW_5) \\ (PW_1) & (PW_2) & \ldots & (PW_5)^{-1} \\ (PW_1) & (PW_2) & \ldots & (PW_5)^{-1} \end{pmatrix}]
\]
and to make use of the Taylor series linearization method. We will also have need of a second multiplier matrix \( M_o \), which has the same stacked-diagonal matrix structure as \( M \) but in which the estimates have been replaced with population values. It is important to note that \( M \) is calculated at each iterate of the simulation whereas \( M_o \) is calculated only once for a simulation. To assess our choices for nominal asymptotic significance levels, we need a noncentrality parameter which is to be computed only once for the simulation, which, in turn, entails additional one-time computations described here. For this purpose, define and quantify:
\[ m_0'=(W_{10}/PW_{10}, W_{20}/PW_{20}, \ldots, W_{50}/PW_{50}) \]
\[ m_0=(O_{10}/PO_{10}, \ldots, O_{50}/PO_{50}) \]
where the additional zero (second) subscript indicates a population value not an estimate or sample-based value. The covariance matrix of the estimators is calculated according to the procedure for obtaining $V_0$ given earlier. With these definitions the noncentrality parameter is

$$\delta^2 = \left( M_i - M_j \right)^{-1} \left( V_i - V_j \right).$$

For each iterate the approach is to quantify the vector of basic estimates $E$, the matrix $M$ and the estimated covariance matrix for the linear estimates, $V$. The next step is to quantify the vector

$$A' = (D_i, D_2, ..., D_s)$$

where $D_i = W_i/P_i$, $O_i/P_i$, and compute the value of the statistic $t^2 = A'(M'VM)^{-1}A$. Finally we form a histogram of the quantities $t^2$ and tabulate the 10%, 5%, 2.5% and 1% critical values; and compute the mean and variance of this empirical distribution of $t^2$ values. It should be informative to compute from an available SAS routine the 10%, 5%, 2.5% and 1% upper values for a chi-squared distribution with five degrees of freedom and with the noncentrality parameter $\delta^2$ defined above and compare these values to those obtained from the empirical distribution function.

REFERENCES


APPENDIX

Derivation of the Bias Correction Term

It is enough to consider the case of one stratum from which PSUs i and j are drawn by Brewer's method. The problem is appropriately viewed as finding a term that will yield unbiased estimate of variance when it is added to

$$Q_i = \pi_1 \pi_2 - \pi_1 \pi_2 \left( \frac{Y_i - Y_j}{\pi_1 \pi_2} \right)^2 + \pi_1^{-1} (N_i')^2 \left( \frac{1}{n_i'} - \frac{1}{N_i'} \right) S_i^2 + \pi_2^{-1} (N_j')^2 \left( \frac{1}{n_j'} - \frac{1}{N_j'} \right) S_j^2$$

where, dropping unnecessary subscripts, $\hat{Y}_i$ and $\hat{Y}_j$ are the unbiased estimators of $Y_i$ and $Y_j$ (defined at the end of section 2), $n_i'$ and $n_j'$ are functions of i and j, and $S_i^2$ and $S_j^2$ are the unbiased estimators of $S_i^2$ and $S_j^2$ (also defined at the end of section 2).

Using the expression for the variance of $\hat{Y}_i$ when $n_i'$ is fixed, we have that

$$E(Y_i^2 \mid \text{choice of PSUs}) = Y_i^2 + (N_i')^2 \left( \frac{n_i'}{N_i'} - \frac{1}{n_i'} \right) S_i^2$$

Since sampling within PSUs is independent, we also have that

$$E(Q_i \mid \text{choice of PSUs}) = \sum_{i=1}^{n_i'} \sum_{j>i}^{n_j'} t_i t_j \left( \frac{\pi_i \pi_j}{\pi_{ij}} \right) \left( \frac{Y_i - Y_j}{\pi_j} \right)^2 + \sum_{i=1}^{n_i'} \sum_{j>i}^{n_j'} t_i t_j (N_i')^2 \left( \frac{\pi_i}{n_i'} - (1-\pi_i)\pi_{ij} \right) \left( \frac{1}{n_i'} - \frac{1}{N_i'} \right) S_i^2$$

where

$$t_i = \begin{cases} 1, & \text{if PSU i is in the sample} \\ 0, & \text{if otherwise} \end{cases}$$

The bias correction term will now follow in the form

$$\sum_{k=1}^{N_i} (N_k'^2 \left( \frac{2 - \pi_k}{\pi_k} \right) (\zeta_k - \eta_k) S_k^2$$

if one now uses the facts that

$$\sum_{i=1}^{n_i} \pi_{ij} = \pi_i, \sum_{j=1}^{n_j} \pi_{ij} = \pi_j$$

and the definitions for $\eta_k$ and $\pi_k$. The specific expression shown in the text follows when the unweighted $Y_{ik}$ are replaced by the weighted values $\hat{Y}_{ik} = \frac{N_i' \hat{Y}_{ik}}{n_i', \hat{Y}_{ik}}$. 761