BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES
SRD Research Report Number: CENSUS/SRD/RR-86/14

TWO NEW VARIANCE ESTIMATION TECHNIQUES

by

Lawrence R. Ernst
Vicki J. Huggins
Diane E. Bailie
Statistical Research Division
Bureau of the Census
Room 3524, F.O.B. #3
Washington, D.C. 20233 U.S.A.

Recommended by: Paul P. Biemer

Report completed: May 28, 1986

Report issued: July 15, 1986 (Revised)

1. INTRODUCTION

The subjects of this paper are two relatively unrelated problems in variance estimation. Research into these problems was motivated by their applicability to the demographic surveys conducted by the Census Bureau, but their potential applications are more general. The first problem, which is the subject of Section 2, is the development of a methodology for pairing strata in one PSU per stratum designs, which minimizes the bias of the resulting variance estimator when using a collapsed stratum estimator of variance. The current designs of the Current Population Survey, the National Crime Survey and the American Housing Survey are examples of one PSU per stratum designs.

The second problem, which is the subject of Section 3, is the development of an alternative to the standard unbiased variance estimator for two PSU's per stratum, without replacement designs, that will have greater precision. The current design of the Survey of Income and Program Participation is essentially this type of design.

Since these two problems are quite different, discussion of them will take place separately in Sections 2 and 3.

2. OBTAINING A COLLAPSING THAT MINIMIZES THE BIAS OF THE COLLAPSED STRATUM VARIANCE ESTIMATOR

To obtain variance estimators for one PSU per stratum designs, a collapsed stratum variance estimator is generally employed, as explained in Wolter (1985). The first step in using such an estimator is the partitioning, or "collapsing", of the set of all strata into groups of two or more strata. Most commonly, each such group of strata consists of two actual strata, and the discussion in this section will be confined to this special case. The main purpose of this section will be to describe how the collapsing can be done in a fashion that in practice appears to be close to optimal in terms of minimizing the bias of the corresponding variance estimator.

We first present the collapsed stratum variance estimator, employing for the most part the notation of Wolter (1985). Consider a population total Y to be estimated by a linear estimator of the form $\hat{Y}_h = \sum_{h=1}^{L} \hat{Y}_h$, where L denotes the number of strata, which is assumed to be even, and $\hat{Y}_h$ is an unbiased estimator of the total in the h-th stratum. The collapsing results in G = L/2 groups of strata, with g1 and g2 denoting the two strata in the g-th group. The collapsed stratum variance estimator $\hat{V}(\hat{Y})$ of $V(\bar{Y})$, as given in Hansen, Hurwitz and Madow (1953), or Wolter (1985), reduces in the case of two strata per group to

$$\hat{V}(\hat{Y}) = \sum_{g=1}^{G} \left(\frac{2A_{g2}}{A_{g1} + A_{g2}} \hat{Y}_{g1} - \frac{2A_{g1}}{A_{g1} + A_{g2}} \hat{Y}_{g2} \right)^2, \qquad (2.1)$$

where A_{gh} is a known measure associated with stratum gh that tends to be well correlated with Y_{gh}. Commonly used values of A_{gh}, which will be discussed later in this section, include:

(i) 1 for all g,h, and

(ii) the population of the gh-th stratum from the most recent census.

We will simplify (2.1) by substituting

$$k_{g1} = \frac{2A_{g2}}{A_{g1} + A_{g2}}, \quad k_{g2} = \frac{2A_{g1}}{A_{g1} + A_{g2}},$$

which yields

$$\hat{V}(\hat{Y}) = \sum_g (k_{g1}\hat{Y}_{g1} - k_{g2}\hat{Y}_{g2})^2 \qquad (2.2)$$

Note that $k_{g1} + k_{g2} = 2$.

To obtain an expression for Bias $\hat{V}(\hat{Y})$, we observe that

$$\begin{aligned} E[\hat{V}(\hat{Y})] &= \sum_g (V(k_{g1}\hat{Y}_{g1} - k_{g2}\hat{Y}_{g2}) + [E(k_{g1}\hat{Y}_{g1} - k_{g2}\hat{Y}_{g2})]^2) \qquad (2.3) \\ &= \sum_g [(k_{g1}^2\sigma_{g1}^2 + k_{g2}^2\sigma_{g2}^2) + (k_{g1}Y_{g1} - k_{g2}Y_{g2})^2], \end{aligned}$$

where $\sigma_{gh}^2 = V(\hat{Y}_{gh})$. Since $V(\hat{Y}) = \sum_g (\sigma_{g1}^2 + \sigma_{g2}^2)$, it follows that

$$\text{Bias } \hat{V}(\hat{Y}) = \sum_g\sum_h (k_{gh}^2 - 1)\sigma_{gh}^2 + \sum_g (k_{g1}Y_{g1} - k_{g2}Y_{g2})^2 \qquad (2.4)$$

We observe the following about (2.4) in the two cases mentioned previously. In case (i), (2.4) reduces to

$$\text{Bias } \hat{V}(\hat{Y}) = \sum_g (Y_{g1} - Y_{g2})^2 \qquad (2.5)$$

since $k_{gh}=1$, while in case (ii) both terms of (2.4) are generally present. However, in case (ii) if A_{gh} and Y_{gh} are well correlated then the second term in (2.4) generally tends to be smaller than in case (i), and disappears altogether if A_{gh} is proportional to Y_{gh}. Also note that if $\sigma^2_{g1} = \sigma^2_{g2}$ for all g, then the first term in (2.4) is nonnegative since $k^2_{g1} + k^2_{g2} \geq 2$, but that in general it is possible for the first term of (2.4), and (2.4) itself to be negative, as is illustrated by examples in Hartley, Rao and Kiefer (1969).

In order to obtain a collapsing that minimizes (2.4), the value of (2.4) must be known for each possible pairing. If (2.4) only involves PSU or stratum totals then such information is assumed known at the time of the most recent census for any characteristic tabulated in the census. (Of course these values generally change between the time of the census and the time that the survey is conducted. This problem will be ignored for now, but returned to at the end of this section.) In case (i), only stratum totals are involved, so that the condition is met. In case (ii) there are several possible approaches. If A_{g1} is sufficiently close to A_{g2} for all g, then one might choose to ignore the first term of (2.4). If that is not acceptable, another possibility is to first rewrite (2.4) by replacing σ^2_{gh} by $\sigma^2_{ghw} + \sigma^2_{ghb}$ where σ^2_{ghw}, σ^2_{ghb} denote the within and between PSU variance respectively for the gh-th stratum. Then

$$\text{Bias } \hat{V}(\hat{Y}) = [\textstyle\sum_{g,h}(k^2_{gh}-1)\sigma^2_{ghw}] + [\sum_{g,h}(k^2_{gh}-1)\sigma^2_{ghb} + \sum_g (k_{g1}Y_{g1} - k_{g2}Y_{g2})^2]. \quad (2.6)$$

The terms within the second set of brackets in (2.6) meet the requirement of involving only PSU and stratum totals. However, census data alone cannot be used to obtain a value for the term within the first set of brackets, since σ^2_{ghw} depends on the particular within PSU sampling procedure employed. Instead, an estimator $\hat{\sigma}^2_{ghw}$ of σ^2_{ghw} could be obtained directly from the sample, and the estimator

$$\hat{\acute{V}}(\hat{Y}) = \hat{V}(\hat{Y}) + \sum_{g,h} (1-k^2_{gh})\hat{\sigma}^2_{ghw} \tag{2.7}$$

used in place of $\hat{V}(\hat{Y})$ to estimate $V(\hat{Y})$. If $\hat{\sigma}^2_{ghw}$ was an unbiased estimator of σ^2_{ghw}, then Bias $\hat{\acute{V}}(\hat{Y})$ would be the terms within the second set of brackets of (2.6). Although unbiased estimators of within PSU variance are not obtainable for the commonly used within PSU sampling procedures that employ systematic sampling, it may be possible to consider the bias of $\hat{\sigma}^2_{ghw}$ small enough to be ignored.

Whatever approach is chosen, it is assumed that for any collapsing, the contribution to the bias of the variance estimator from each pair of strata is known and nonnegative, and we turn to the key question of this section: Given the set of L strata, how should they be paired in order to minimize the bias of the variance estimator. In an attempt to answer this question, the problem will be formulated as a mathematical programming problem. First let the constants c_{ij}, $i<j$, $i,j=1,\dots,L$, denote the contribution to the bias of the variance estimator from the pair of strata i,j, if i and j are paired together. For example if the bias is given by (2.5), then $c_{ij}=(Y_i-Y_j)^2$. The total bias of the variance estimator corresponding to any collapsing would be

$$\sum_{\substack{i,j \\ i<j}}^{L} c_{ij}x_{ij} \ , \tag{2.8}$$

where

x_{ij} = 1, if strata i and j are paired together,

= 0, otherwise.

Then minimizing the bias of the variance estimator is equivalent to minimizing (2.8) subject to the constraints

$$x_{ij} = 0 \text{ or } 1 \text{ for all } i,j,\ i<j, \tag{2.9}$$

and that for each i exactly one member of the sequence

$$x_{1i}, x_{2i} \ldots, x_{(i-1)i}, x_{i(i+1)}, x_{i(i+2)}, \ldots, x_{iL}$$

is equal to 1, or equivalently,

$$\sum_{j=1}^{i-1} x_{ji} + \sum_{j=i+1}^{L} x_{ij} = 1,\ i=1,\ldots,L\ . \tag{2.10}$$

The problem defined by (2.8 - 2.10) is an integer programming problem. If L is sufficiently small, an optimal solution could be obtained by using any standard software for solving integer programming problems. Unfortunately, the solution time for such problems increases rapidly with increasing L, and if L is fairly large it would be impractical to solve the problem in this fashion.

It would be desirable if this integer programming problem could be transformed into a different form of mathematical programming problem that would be more efficient computationally. To this end, we define $c_{ij}=c_{ji}$ if $i>j$ and $c_{ii}=M$ for each i, where M is a suitably large constant, as will be explained. We then seek to minimize

$$\sum_{i,j} c_{ij}x_{ij}\ , \tag{2.11}$$

subject to the constraints

$$\sum_j x_{ij} = 1, \quad i=1,\ldots L, \tag{2.12}$$

$$\sum_i x_{ij} = 1, \quad j=1,\ldots L, \tag{2.13}$$

$$x_{ij} = 0 \text{ or } 1, \quad i,j = 1,\ldots L. \tag{2.14}$$

The problem (2.11 - 2.14) is an assignment problem. Software exists for solving assignment problems in reasonable time even for quite large L. The key question is whether an optimal solution to the assignment problem (2.11 - 2.14) leads to an optimal solution to the original integer programming problem (2.8 - 2.10). The answer would be yes if the following conditions were true for an optimal solution to this assignment problem:

$$x_{ii} = 0, \quad i = 1,\ldots,L \qquad ? \tag{2.15}$$

$$x_{ij} = x_{ji}, \; i,j = 1,\ldots,L \qquad ? \tag{2.16}$$

For, if these conditions were satisfied, then as a result of the symmetry in both the c_{ij}'s and x_{ij}'s, the subset of the optimal x_{ij}'s for the assignment problem for which $i<j$ would satisfy (2.10) and the corresponding value of (2.8) would be 1/2 the value of (2.11). Furthermore, the set x_{ij}, $i<j$ minimizes (2.8) subject to (2.9), (2.10), since if x'_{ij} , $i<j$, also satisfied (2.9), (2.10) and if we let $x'_{ij}=x'_{ji}$ for $i>j$, $x_{ii}=0$, then the entire set of x'_{ij}'s would satisfy (2.12 - 2.14) with

$$\sum_{\substack{i,j \\ i<j}} c_{ij}x'_{ij} = \frac{1}{2} \sum_{i,j} c_{ij}x'_{ij} \geq \frac{1}{2} \sum_{i,j} c_{ij}x_{ij} = \sum_{\substack{i,j \\ i<j}} c_{ij}x_{ij}.$$

Thus the value of (2.8) for x'_{ij}, $i<j$ is not less than (2.8) for

the set x_{ij}, $i<j$.

Now (2.15) always holds if M is set sufficiently large. For example, any $M > L\cdot\max\{c_{ij}: i<j\}$ would certainly suffice.

One might believe that (2.16) also always holds since $c_{ij}=c_{ji}$ for all i,j. However, this is false, as is established by the following counterexample. Let L=6 and take

$$c_{ij} = 0, \text{ if } i,j\leq 3 \text{ or } i,j\geq 4,\ i\neq j,$$
$$> 0, \text{ otherwise},$$

that is $c_{ij}=0$ for all elements of the array in the upper left or lower right quadrants of the array that are not on the main diagonal. Then the following set of x_{ij}'s satisfies (2.12 - 2.14) and yields a value of 0 for (2.11), and hence is an optimal solution to the assignment problem:

$$x_{12}=x_{23}=x_{31}=x_{45}=x_{56}=x_{64}=1,$$
$$x_{ij}=0 \text{ for all other } i,j.$$

Clearly this solution does not satisfy (2.16). Nor are there any other feasible solutions to this problem for which (2.16) holds and (2.11) is 0. To see why, observe that if a set of x_{ij}'s is a feasible solution to (2.11 - 2.14) and if $x_{12}=x_{21}=1$ then $x_{3j}=1$ for j=4,5 or 6 and hence (2.11) would be positive. Similiarly, any other feasible solution to this assignment problem for which $x_{ij}=x_{ji}=1$ for some i,j with $c_{ij}=0$, immediately forces $x_{i_1j_1}=1$ for some i_1,j_1 for which $c_{i_1j_1}>0$.

Although an optimal solution to (2.11 - 2.14) does not in general lead to an optimal solution to (2.8 - 2.10), it is believed that a nearly optimal solution can generally be obtained in an efficient manner by combining both of these problems as follows. First obtain an optimal solution to the assignment problem and let S_1 denote the set of strata i for which there exists a j satisfying $x_{ij}=x_{ji}=1$, while the set of all remaining strata is denoted by S_2. The pairing for the strata in S_1 is defined by this optimal solution to the assignment problem, that is, the i-th and j-th strata are paired if $x_{ij}=x_{ji}=1$. If S_2 is sufficiently small then the elements in it can be paired by obtaining an optimal solution to a problem like (2.8 - 2.10), but with S_2 now viewed as the set of all strata. If S_2 is still too large for this purpose, it can be partitioned into a collection of say t subsets $S_{21},\ldots,S_{2t}$, such that each such subset S_{2k} contains an even number of elements; each S_{2k} is small enough to efficiently obtain a solution to (2.8 - 2.10) with S_{2k} viewed as the set of all strata; and strata i and j are in the same S_{2k} if either $x_{ij}=1$ or $x_{ji}=1$ in the optimal solution to the assignment problem, provided this last requirement can be met without any of the S_{2k} becoming too large. (The rationale for grouping strata i,j for which either $x_{ij}=1$ or $x_{ji}=1$ in the same S_{2k} is that such a grouping tends to put together pairs of strata which would have a small contribution to the bias of the variance estimator.) The elements of S_{2k} are then paired by the optimal solution of (2.8 - 2.10) restricted to S_{2k}.

The procedure just described results in an optimal pairing of strata in S_1 and either an optimal pairing of strata in S_2 or, if S_2 is partitioned, an optimal pairing of strata in each of the

S_{2k}'s. However, it is not necessarily an optimal pairing for the entire set of L strata, since such a pairing may require that a stratum in one subset be paired with a stratum in another. Although it does not in general yield an optimal solution, it is believed that this approach would provide a good approximate solution in an efficient manner.

Remark: All of the preceding work has been with respect to a single characteristic Y. Since, as a practical matter, the same collapsing would generally be used for variance estimates for all characteristics, the collapsing criteria could be taken to be the minimization of the weighted average of the biases of the variance estimator for several key characteristics instead of the bias of a single characteristic, that is,

$$\sum_k W_k \text{ Bias } \hat{V}(\hat{Y}_k), \tag{2.17}$$

where $\hat{Y}_k$ is an unbiased estimator of Y_k. If all of these characteristics are considered of equal importance then W_k would be some value that would serve as a scaling factor. (One possible scaling factor is presented in the example below.) If some variables are more important than others, then W_k could be taken to be something greater than the corresponding scaling factor for the more important variables and less than the scaling factor for the less important variables.

Illustrative Example

The present design of the Current Population Survey (CPS) is used to illustrate this work. This survey has a one PSU per stratum design with 379 nonself-representing strata. (There are also self-representing strata that are not subject to a collapsed strata procedure since there is no between PSU variance for such strata). Because L is odd, one stratum was arbitrarily dropped for this illustration. After the remaining 378 strata are paired, the discarded stratum could then be grouped with one of the 189 pairs resulting in 188 pairs and one group of three strata. The pair that this strata is grouped with could be chosen to minimize the bias of the total collapsing by computing the bias for each of the 189 possible such groupings that could be created and choosing the grouping with smallest bias.

The comparison criterion is the value of (2.17) where the eight characteristics used were

Unemployed, Total	Civilian Labor Force, Total
Black	Black
Hispanic	Hispanic
Teenage (16-19)	Agriculture Employment

To obtain W_k, a random pairing was first selected and then for each k, W_k was taken to be (1/8)Bias $\hat{V}(\hat{Y}_k)$ corresponding to the random pairing. The minimization of the objective function with this W_k amounts to obtaining a particular pairing for which the average, over the eight characteristics, of the ratio of the bias for this pairing to the bias for the random pairing is minimized.

For each k, Bias $\hat{V}(\hat{Y}_k)$ was computed separately for the cases (i) and (ii), defined earlier, both for obtaining W_k and then for computing the objective function. For case (i), (2.5) was of course used in this computation while for case (ii), the second term only of (2.4) was used to obtain Bias $\hat{V}(\hat{Y}_k)$. 1980 census data was used for all computations. In case (i), the procedure resulted in sets S_1 and S_2 containing 316 and 62 strata respectively. S_2 was partitioned into 3 subsets. S_{21}, S_{22}, S_{23} consisting of 18, 20 and 24 strata. In case (ii), S_1 and S_2 contained 278 and 100 strata respectively. S_2 was partitioned in case (ii) into 4 sets of 26, 26, 24 and 24 strata. (The assignment problems were solved with software written by James Fagan, while Sperry's Functional Mathematical Programming System was used to solve the integer programming problems.)

The value of the objective function (2.8) corresponding to the final pairing obtained from this procedure for each case is presented in the first column of numbers in Table 2.1. The numbers in columns 2 - 4 provide an indication of the effectiveness of this procedure. Each value in the second column is 1/2 the corresponding minimum value of the assignment problem (2.11 - 2.14), which is a lower bound on the minimum value for the integer programming problem (2.8 - 2.10). The numbers in the third column are the values of (2.8) corresponding to a pairing by strata size, that is with the strata ordered in increasing size based on 1980 population, and the smallest stratum paired with the next smallest stratum, etc. The fourth column presents the values of (2.8) averaged over 10 random pairings independent of the random pairing used in computing the W_k's. The fact that this number is reasonably close to 1 in both cases provides an indication that results similar to these in this table would be expected if some other random pairing had been used to compute the W_k's. The table indicates that the procedure described in this section yields, for this set of data, a pairing with a bias reasonably close to optimal, and substantially below that obtained by either random pairings or pairings by strata size.

As previously noted, the biases of the variance estimator for any pairing change over time. Since a pairing for the current design of the CPS would be based on 1980 census data, but might be used for a time span that roughly averages 10 years away from 1980, it would be instructive to consider the bias of the variance estimator for the pairings used in constructing Table 2.1 with 1990 census data substituted for 1980 census data. Since 1990 data is not available, 1970 data was used instead on the assumption that the results from going backwards a decade would be indicative of the results going forwards a decade. The results are presented in Table 2.2. Columns 1, 3 and 4 of this table were obtained by using the same pairings as for the corresponding entries in Table 2.1, but with 1970 census data substituted for 1980 data. Column 2 was obtained by minimizing the assignment problem (2.11 - 2.14) with 1970 census data and multiplying the result by 1/2 to get a lower bound on the bias of the variance estimator for 1970. The table indicates that while the bias reduction deteriorates over time, as would be expected, it is still substantial for this set of data after 10 years.

Table 2.1 Bias Measure with 1980 Census Data

	Mathematical Programming Pairing	Lower Bound on Optimal Solution	Pairing by Strata Size	Average of 10 Random Pairings
A_{gh} = 1	.0620	.0462	.6455	1.0522
A_{gh} = POP	.0582	.0468	1.3987	1.0551

Table 2.2 Bias Measure with 1970 Census Data, but same Pairings as Table 2.1

	Mathematical Programming Pairing	Lower Bound on Optimal Solution	Pairing by Strata Size	Average of 10 Random Pairings
A_{gh} = 1	.1134	.0489	.6613	1.0499
A_{gh} = POP	.1874	.0428	1.3226	1.0545

3. AN UNBIASED ESTIMATOR OF VARIANCE WITH INCREASED PRECISION FOR MULTI-STAGE WITHOUT REPLACEMENT SAMPLING

The standard estimator of variance for $n(\geq 2)$ PSU's per stratum, multi-stage designs, with the PSU's chosen without replacement, as presented in Raj (1968), can itself have a large variance. In this section an alternative unbiased variance estimator is developed for the case n=2 that will in general have greater precision.

We proceed to explain this problem in detail. The variance estimator in Raj (1968) will first be presented. All expressions will be given in the special case of a single stratum, since the generalization to more than one stratum is routine.

The following notation will be used. Let π_i be the probability that the i-th PSU is in a sample of n PSU's out of N, and let π_{ij} be the probability that both the i-th and j-th PSU's are in sample. Let $\hat{Y}_i$ be an unbiased estimator of the i-th PSU total, Y_i, based on sampling at the second and subsequent stages, with $V(\hat{Y}_i|i) = \sigma_i^2$, and let $\hat{\sigma}_i^2$ denote an unbiased estimator of σ_i^2. Then an unbiased estimator of the population total Y is given by

$$\hat{Y} = \sum_{i=1}^{n} \frac{\hat{Y}_i}{\pi_i},$$

with

$$V(\hat{Y}) = \sum_{\substack{i,j \\ i<j}}^{N} (\pi_i \pi_j - \pi_{ij})\left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} \frac{\sigma_i^2}{\pi_i}, \qquad (3.1)$$

and an unbiased estimator $\hat{V}(\hat{Y})$ of $V(\hat{Y})$ given by

$$\hat{V}(\hat{Y}) = \sum_{\substack{i,j \\ i<j}}^{n} \left(\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\right)\left(\frac{\hat{Y}_i}{\pi_i}-\frac{\hat{Y}_j}{\pi_j}\right)^2 + \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i} . \tag{3.2}$$

The focus of this section will be on reducing the contribution to the variance of (3.2) that arises from the factor $(\pi_i\pi_j-\pi_{ij})/\pi_{ij}$, which from now on will be abbreviated by d_{ij}. Although d_{ij} is nonnegative for procedures to select PSU's such as the procedure of Brewer (1963) and Durbin (1967) for n=2 and its generalization for n>2 by Sampford (1967), d_{ij} in general does not have any upper bound and its variability can result in an undesirably large variance of (3.2).

To understand what can be done to alleviate this problem, first observe what each of the terms of (3.2) estimates. From the proof given in Raj (1968, Theorem 6.3) it follows that

$$E\left[\sum_{\substack{i,j \\ i<j}}^{n} d_{ij}\left(\frac{\hat{Y}_i}{\pi_i}-\frac{\hat{Y}_j}{\pi_j}\right)^2\right] = \sum_{\substack{i,j \\ i<j}}^{N} (\pi_i\pi_j-\pi_{ij})\left(\frac{Y_i}{\pi_i}-\frac{Y_j}{\pi_j}\right)^2 + \sum_{i=1}^{N} (1-\pi_i)\frac{\sigma_i^2}{\pi_i} ,$$

while

$$E\left(\sum_{i=1}^{n} \frac{\hat{\sigma}_i^2}{\pi_i}\right) = \sum_{i=1}^{N} \sigma_i^2 .$$

Thus the expected value of the first term in (3.2) is the entire first term in (3.1) (the between PSU variance) plus part of the second term (the within PSU variance), while the expected value of the second term in (3.2) is the remainder of the within PSU variance.

A superior alternative to estimating the between PSU variance by the first term in (3.2) does not appear to exist.

However, a general class of unbiased estimators of $V(\hat{Y})$ exists, which includes (3.2), from which a specific estimator can be chosen that reduces the variability associated with the estimation of the within PSU variance. This class of estimators has the form

$$\hat{V}_{k_{ij}}(\hat{Y}) = \sum_{\substack{i,j \\ i<j}}^{n} d_{ij} \left(\frac{\hat{Y}_i}{\pi_i} - \frac{\hat{Y}_j}{\pi_j}\right)^2 + \sum_{i,j}^{n} k_{ij} \frac{\hat{\sigma}_i^2}{\pi_i^2} , \tag{3.3}$$

where the k_{ij}'s are constants. (It is understood that i=j is excluded from the second summation in (3.3) and in all other expressions in this section.) Note that (3.2) is a special case of (3.3) with $k_{ij} = \pi_i/(n-1)$, and that in general $k_{ij} \neq k_{ji}$. Furthermore, in order for (3.3) to be an unbiased estimator of (3.1) restrictions must be placed on the k_{ij}'s. To establish what these restrictions are, note that the expected value of (3.3) conditioned on the specific set of sample PSU's is

$$\sum_{\substack{i,j \\ i<j}}^{n} d_{ij} \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i,j}^{n} (d_{ij} + k_{ij}) \frac{\sigma_i^2}{\pi_i^2} ; \tag{3.4}$$

consequently,

$$E[\hat{V}_{k_{ij}}(\hat{Y})] = \sum_{\substack{i,j \\ i<j}}^{N} (\pi_i\pi_j - \pi_{ij}) \left(\frac{Y_i}{\pi_i} - \frac{Y_j}{\pi_j}\right)^2 + \sum_{i,j}^{N} \pi_{ij}(d_{ij} + k_{ij}) \frac{\sigma_i^2}{\pi_i^2} . \tag{3.5}$$

Comparison of (3.5) with (3.1) shows that (3.3) is an unbiased estimator of $V(\hat{Y})$ if and only if

$$\sum_{j}^{N} \pi_{ij}(d_{ij} + k_{ij}) = \pi_i \ , \ i=1,\ldots N. \tag{3.6}$$

Furthermore, since by the proof in Raj (1968, Theorem 6.3)

$$\sum_{j}^{N} \pi_{ij} d_{ij} = \pi_i - \pi_i^2, \quad i=1,\ldots,N, \tag{3.7}$$

(3.6) can be rewritten in the alternate form

$$\sum_{j}^{N} \pi_{ij} k_{ij} = \pi_i^2, \quad i=1,\ldots,N . \tag{3.8}$$

Although (3.8) is clearly satisfied by $k_{ij} = \pi_i/(n-1)$, the $(d_{ij} + k_{ij})$'s can be quite variable for fixed i with this set of k_{ij}'s because of the variability in the d_{ij}'s. An alternate set of k_{ij}'s which clearly satisfies (3.6) and completely removes the variability of the $(d_{ij} + k_{ij})$'s is given by

$$k_{ij} = \frac{1}{n-1} - d_{ij}. \tag{3.9}$$

However, since d_{ij} can exceed $1/(n-1)$, (3.9) can be negative for some i,j's and negative estimates of variance can result. To avoid this possibility, a second set of constraints on the k_{ij}'s,

$$k_{ij} \geq 0, \quad i,j=1,\ldots,N, \quad i \neq j \tag{3.10}$$

is added to (3.6), and the set of k_{ij}'s defined by (3.9) will not be considered further in unmodified form.

One method of modifying (3.9) to satisfy (3.10) is to let

$$k_{ij} = \frac{1}{n-1} - d_{ij} \text{ if } d_{ij} < \frac{1}{n-1}, \tag{3.11}$$
$$= 0 \text{ otherwise.}$$

This method was suggested by Robert Fay of the Bureau of the Census (personal communication). However, this set of k_{ij}'s does not in general satisfy (3.6) and consequently yields a biased variance estimator.

From now on, we consider only the case where n=2 and present what is the major focus of this section, a set of k_{ij}'s which satisfies (3.6), (3.10) and which for each i minimizes the deviation of $d_{ij} + k_{ij}$ from 1 in the sense that for each $p > 1$,

$$E(|d_{ij}+k_{ij}-1|^p : i \text{ is in sample}) = \sum_{j}^{N} \frac{\pi_{ij}}{\pi_i} |d_{ij}+ k_{ij}- 1|^p \qquad (3.12)$$

is minimized subject to (3.6), (3.10). (The expectation in (3.12) is with respect to the other sample PSU j.) Deviations from 1 are considered because it follows from (3.6) that for fixed i this is the expected value of $d_{ij} + k_{ij}$ given that the i-th PSU is in sample. To define this set of k_{ij}'s for fixed i, we first relabel the sequence $d_{i1},\ldots,d_{i(i-1)},d_{i(i+1)},\ldots,d_{iN}$ to transform it into a nondecreasing sequence. Then let

$$a_{ij} = \frac{\pi_i - \sum_{t=j+1}^{N} \pi_{it} d_{it}}{\sum_{t=1}^{j} \pi_{it}}, \quad i,j=1,\ldots,N,\ i \neq j. \qquad (3.13)$$

Next, let m_i be the largest integer for which $d_{im_i} < a_{im_i}$, and finally, let

$$k_{ij} = a_{im_i} - d_{ij} \text{ if } j \le m_i, \qquad (3.14)$$
$$= 0 \text{ otherwise.}$$

Roughly, the motivation for (3.11) is that for each i, $d_{ij} + k_{ij}$ becomes a constant function of j except for those j which would require a negative k_{ij} to accomplish this. In fact, if $d_{ij} \le 1$ for all i,j, it can be shown that $a_{im_i} = 1$ for all i, $m_i = N$ for $i \neq N$, $m_N = N-1$, and (3.14) then reduces to (3.9) with $n=2$.

We proceed to establish that the k_{ij}'s satisfy the stated conditions, that is (3.6) and (3.10) are satisfied and (3.12) is minimized subject to these constraints. However, in order for m_i and hence (3.14) to be well defined it first must be shown that the set of j's for which $d_{ij} < a_{ij}$ is nonempty for each i. To prove this for $i \neq 1$, we establish that $d_{i1} < a_{i1}$, that is

$$\frac{\pi_i \pi_1 - \pi_{i1}}{\pi_{i1}} < \frac{\pi_i - \sum_{j=2}^{N} (\pi_i \pi_j - \pi_{ij})}{\pi_{i1}}.$$

This is equivalent to showing that $\sum_{j}^{N} (\pi_i \pi_j - \pi_{ij}) < \pi_i$, which follows immediately from (3.7). Similarly, for i=1 it can be shown that $d_{12} < a_{12}$.

Now (3.6) follows, since by (3.13), (3.14),

$$\sum_{j=1}^{N} \pi_{ij}(d_{ij} + k_{ij}) = a_{im_i}\left(\sum_{j=1}^{m_i} \pi_{ij}\right) + \sum_{j=m_i+1}^{N} \pi_{ij} d_{ij}$$

$$= \left(\pi_i - \sum_{j=m_i+1}^{N} \pi_{ij} d_{ij}\right) + \sum_{j=m_i+1}^{N} \pi_{ij} d_{ij} = \pi_i,$$

while (3.10) is immediate because

$$k_{ij} = a_{im_i} - d_{ij} \geq a_{im_i} - d_{im_i} \geq 0 \text{ for } j \leq m_i.$$

To show that (3.14) minimizes (3.12) subject to (3.6) and (3.10), again consider i as fixed and view the problem (3.12), (3.6), (3.10), as an optimization problem in the variables k_{ij}, j=1,...,N. Now an optimal solution to this problem may contain some j's for which $k_{ij} = 0$, this is the lower bound for the variables. Let $S = \{j: k_{ij} > 0\}$. A minimum value for the problem (3.12), (3.6), (3.10) conditioned on S, can be obtained by first minimizing

$$\sum_{j\varepsilon S} \frac{\pi_{ij}}{\pi_i} |d_{ij} + k_{ij} - 1|^p$$

subject to

$$\sum_{j\varepsilon S} \pi_{ij} k_{ij} = \pi_i^2 , \qquad (3.15)$$

which is equivalent to (3.8) and hence (3.6).
The method of Lagrange multipliers yields the unique set of k_{ij}'s for which

$$d_{ij} + k_{ij} \text{ are the same for all } j\varepsilon S \qquad (3.16)$$

and (3.15) is satisfied, as the only candidate for a minimum conditioned on S. Provided this set of k_{ij}'s also satisfies $k_{ij}>0$ for $j\varepsilon S$, this will be the conditional minimum, while if not, there will be no feasible solution to (3.12), (3.6), (3.10) corresponding to that S.

Let $S_0 = \{1,\ldots,m_i\} - \{i\}$. Corresponding to $S=S_0$, the set of k_{ij}'s satisfying (3.15) and (3.16) for $j\varepsilon S_0$, $k_{ij}=0$ otherwise, is precisely the set given by (3.14). Consequently, it remains to show that corresponding to any other S, the set of k_{ij}'s determined by (3.15) and (3.16) yields no feasible solution with lower value of (3.12). To do this we first consider the case $S_0 \subset S$ and let t be the largest element in S. Then, by (3.6),

$$(d_{it} + k_{it}) \sum_{j\varepsilon S} \pi_{ij} + \sum_{j\notin S} \pi_{ij} d_{ij} = \pi_i . \qquad (3.17)$$

Furthermore, since $d_{i(m_i+1)} \geq a_{i(m_i+1)}$ it follows that

$$d_{it} \sum_{j\varepsilon S} \pi_{ij} + \sum_{j\notin S} \pi_{ij} d_{ij} \geq d_{i(m_i+1)} \sum_{j=1}^{m_i+1} \pi_{ij} + \sum_{j=m_i+2}^{N} \pi_{ij} d_{ij} \geq \pi_i . \quad (3.18)$$

(If $m_i+1=i$, substitute m_i+2 for m_i+1 in the previous sentence.) Comparing (3.17) and (3.18) it follows that $k_{it} \leq 0$, which contradicts the definition of S. Consequently, there is no feasible solution to (3.12), (3.6), (3.10) if $S_0 \subset S$.

Now consider any S for which $S_0 \subset S$ does not hold. Let k^*_{ij} be the optimal solution conditioned on S, and choose $j_1 \varepsilon\ S_0 - S$, $j_2 \varepsilon S$. Then

$$k^*_{ij_1} = d_{ij_1} < d_{ij_2} + k^*_{ij_2} . \quad (3.19)$$

This is because, either $S \subset S_0$, in which case

$$d_{ij_1} < a_{im_i} \leq d_{ij_2} + k^*_{ij_2}$$

or $S-S_0 \neq \emptyset$, in which case for any $j_3 \varepsilon\ S-S_0$,

$$d_{ij_1} \leq d_{ij_3} < d_{ij_2} + k^*_{ij_2} .$$

Next consider the function

$$\frac{\pi_{ij_1}}{\pi_i} |d_{ij_1} + k_{ij_1} - 1|^p + \frac{\pi_{ij_2}}{\pi_i} |d_{ij_2} + k_{ij_2} - 1|^p , \quad (3.20)$$

subject to the constraint

$$\pi_{ij_1} k_{ij_1} + \pi_{ij_2} k_{ij_2} = \pi_{ij_2} k^*_{ij_2} . \quad (3.21)$$

By solving (3.21) for k_{ij_2} in terms of k_{ij_1} substituting the result in (3.20), one obtains a function of k_{ij_1} only. The first derivative of this function is negative when $k_{ij_1}=0$ because of (3.19), and the function is consequently decreasing for k_{ij_1} sufficiently small. Therefore, if $k^{**}_{ij_1}, k^{**}_{ij_2}$ satisfies (3.21), $k^{**}_{ij_1}$ is sufficiently small and positive, and $k^{**}_{ij}=k^{*}_{ij}$ for all j's other than j_1 and j_2, then the k^{**}_{ij}'s satisfy (3.6), (3.10) and yield a lower value for (3.12) than the k^{*}_{ij}'s. This shows that if $S \neq S_0$, then no set of k_{ij}'s for which $k_{ij}=0$ for $j \notin S$ will be an optimal solution to (3.12), (3.6), (3.10), and the set of k_{ij}'s defined by (3.12) must be optimal.

Illustrative Example

We will compare numerically our variance estimator, defined by (3.3), (3.14), with two other estimators previously described, the method given in Raj (1968) and defined by (3.2), and the estimator suggested by Fay and defined by (3.3), (3.11). These three variance estimators will be referred to as the conditional unbiased (CU), unconditional unbiased (UU), and conditional biased (CB) estimators respectively. ("Conditional" indicates that k_{ij} is conditioned on j.)

The survey used in the comparison was the original 1980 census based design for the Survey of Income and Program Participation (SIPP). (A sample cut took place before this design was ever implemented in which some sample PSU's were dropped, but this cut is not considered here.) There were 95 strata in that design from which two PSU's were selected without replacement. There were also 91 self-representing strata and eight nonself-representing strata from which one PSU was selected per stratum which will not be considered in this example.

The comparison criterion will be one component of the squared error of (3.3), namely the MSE of the second term in (3.4), which we denote by W, that is

$$W = (d_{ij} + k_{ij}) \frac{\sigma_i^2}{\pi_i^2} + (d_{ij} + k_{ji}) \frac{\sigma_j^2}{\pi_j^2},$$

where i and j are the sample PSU's. To simplify our computations, it will be assumed that σ_i^2 is proportional to π_i^2. Furthermore, since the comparison would not be affected by the constant of proportionality, we take $\sigma_i^2/\pi_i^2=1$ for all i, and thus W reduces to

$$W = 2d_{ij} + k_{ij} + k_{ji}.$$

Now from (3.6) it follows that

$$E(W) = \sum_{i}^{N} \pi_i = 2$$

for the CU and UU procedures, which is also the value for the second term in (3.2). For the CB procedure we have

$$E(W) = \sum_{i,j}^{N} \pi_{ij} \max \{d_{ij}, 1\}.$$

Furthermore, for all three procedures

$$V(W) = \sum_{\substack{i,j \\ i<j}}^{N} \pi_{ij}(2d_{ij} + k_{ij} + k_{ji})^2 - E(W)^2. \qquad (3.22)$$

In addition, for the CB procedure only

$$\text{Bias } W = E(W) - 2, \qquad (3.23)$$

while Bias W=0 for the other two procedures.

One modification of this work was necessary. In the actual selection of PSU's for SIPP, some small PSU's were combined to form a "rotation cluster" in 18 of the strata. In computing the joint probabilities, the cluster was initially treated as a single PSU. If the cluster was selected, then at any time during the life of the design one of the PSU's in the cluster would be in sample with probability proportional to size. (This was done because a new sample is chosen from the sample PSU's each year. For small PSU's there is not enough distinct ultimate sampling units to last the life of the design. PSU's in a cluster can be rotated in and out of sample to avoid this problem. See Alexander, Ernst and Haas (1982) for more details.) As a result of this procedure $\pi_{ij}=0$ if PSU's i and j are both in the rotation cluster, and unbiased estimators of variance are no longer possible. To obtain a class of estimators constructed with the goal of being approximately unbiased, the following modifications were made to (3.3) and (3.6). Let $T=\{(i,j): i \text{ or } j \text{ are not in the rotation cluster}\}$, $T_i=\{j:(i,j) \in T\}$,

$$f = \frac{\sum_{i,j}^{N} (\pi_i \pi_j - \pi_{ij})}{\sum_{(i,j) \epsilon T} (\pi_i \pi_j - \pi_{ij})}$$

and $d_{ij}^* = f d_{ij}$, $(i,j) \ \epsilon\ T$. Then modify, (3.3), (3.6), by substituting d_{ij}^* for d_{ij} in these expressions, and only summing over $j \epsilon T_i$. (The f factor is to compensate for the fact that the modified first term in (3.3) is a summation only over $(i,j) \ \epsilon\ T$.) These same substitutions in (3.11) and (3.14) are used to modify the CB and CU procedures. As for the UU procedure, $k_{ij} = \pi_i$ would not satisfy the modified (3.6), since

$$\sum_{j \epsilon T_i} \pi_{ij} d_{ij}^* \neq \pi_i - \pi_i^2$$

in general. Instead, take

$$k_{ij} = \frac{\pi_i - \sum_{j \epsilon T_i} \pi_{ij} d_{ij}^*}{\pi_i} \quad . \tag{3.24}$$

It should also be noted that for some i it is possible that $\sum_{j \epsilon T_i} \pi_{ij} d_{ij}^* > \pi_i$, in which case no nonnegative set of k_{ij}'s could satisfy the modified (3.6). In particular (3.24) would be negative and CU would not be defined since $d_{ij}^* > a_{ij}$ for all $j \epsilon T_i$. This problem arose in only 1 of the 95 strata under consideration in this illustration and that stratum was dropped from the example.

For each of the remaining 94 strata, V(W) was computed for all three methods and the resulting values summed over the 94 strata to obtain the first column of Table 3.1. Similarly, for the UB procedure, Bias W was computed for each stratum with the sum given in column 2 of this table. Finally, MSE, that is the sum of column 1 and the square of column 2, is presented for each of these three procedures in column 3.

TABLE 3.1

COMPARISON OF THREE VARIANCE ESTIMATORS ON SIPP DATA

Procedure	Variance	Bias	MSE
CU	11.6168	0	11.6168
UU	20.2374	0	20.2374
CB	8.2359	4.8941	32.1877

Thus for this particular design, MSE is smallest for the CU procedure.

REFERENCES

Alexander, C.H., Ernst, L.R., and Haas, M.E., (1982), "A System for Replacing Primary Sampling Units when the Units Have Been Exhausted," in Proceedings of the Survey Research Section, American Statistical Association, pp. 211-216.

Hansen, M.H. Hurwitz, W.N., and Madow, W.G., (1953) Sample Survey Methods and Theory Volume II, New York: John Wiley and Sons.

Hartley, H.O., Kao, J.N.K., and Kiefer, G., (1969), "Variance Estimation with One Unit per Stratum," Journal of the American Statistical Association, pp. 841-851.

Raj, Des (1968), Sampling Theory, New York: McGraw-Hill.

Wolter, Kirk M. (1985), Introduction to Variance Estimation, New York: Springer-Verlag.