Survey Methodology, June 1998
Vol. 24, No. 1, pp. 21-30
Statistics Canada

Estimation in Sample Surveys Using Frames With a Many-to-Many Structure

TERRI L. BYCZKOWSKI, MARTIN S. LEVY and DENNIS J. SWEENEY[1]

ABSTRACT

In sample surveys, the units contained in the sampling frame ideally have a one-to-one correspondence with the elements in the target population under study. In many cases, however, the frame has a many-to-many structure. That is, a unit in the frame may be associated with multiple target population elements and a target population element may be associated with multiple frame units. Such was the case in a building characteristics survey in which the frame was a list of street addresses, but the target population was commercial buildings. The frame was messy because a street address corresponded either to a single building, multiple buildings, or part of a building. In this paper, we develop estimators and formulas for their variances in both simple and stratified random sampling designs when the frame has a many-to-many structure.

KEY WORDS: Imperfect frames; Correspondence errors; Building characteristics survey; Weighting; Simple random sampling; Stratified random sampling.

1. INTRODUCTION

This research was motivated by a study that was conducted for a utility company to estimate various population characteristics of the commercial buildings located in their service area. Budgetary constraints prohibited the development of a list of commercial buildings using canvassing techniques. However, a sampling frame consisting of street addresses (*i.e.*, addresses at which a utility meter was located) was available. A drawback of this frame was that it had a many-to-many relationship with the target population of commercial buildings. That is, some units in the frame were associated with multiple target population elements, and some target population elements were associated with multiple frame units. In fact, several of the relationships between street addresses and commercial buildings were relatively complex.

An advantage of this frame, however, was that total annual electrical usage was available for each street address. This resulted in a variable upon which the frame of street addresses could be effectively stratified. One of the important characteristics to be measured was the total commercial square footage. Studies conducted in the United States have shown that energy consumption is associated with both building size and building activity. For example, consumption is higher for buildings used for health care or food sales, and lower for buildings used for religious worship or public assembly. Also, energy consumption is correlated with building size even if the activity of the building is not known, as was the case here (U.S. Department of Energy 1992).

There is a vast amount of literature dealing with imperfect sampling frames. Comprehensive summaries of this literature can be found in Kish (1965), Wright and Tsao (1983), and Lessler and Kalsbeek (1992). Another body of literature addresses multiplicity sampling in which the frame is constructed with a many-to-many structure by design. Here, frame imperfections are introduced in order to gather information more efficiently on rare occurrences in a population (Birnbaum and Sirken 1965, Sirken 1972a,b, and Casady and Sirken 1980). Hansen, Hurwitz and Madow (1953a,b) present an estimator for use with sampling frames that have a many-to-one structure; population elements are represented multiple times in the frame. This estimator has also been adopted for use by National Agricultural Statistics Service (NASS) surveys (Musser 1993) with respect to the many-to-one frame. Bandyopadhyay and Adhikari (1993) developed estimators for a ratio, population mean, and population total when an unknown amount of duplication is present in the frame. But, these estimators are restricted to the simple random sampling case and the many-to-one frame.

Two methods for estimating population characteristics using a frame with a many-to-many structure appear in the literature. First, the Horvitz-Thompson estimator (1952) provides unbiased estimates of population means and totals when varying probabilities of selection are present. Musser (1993) shows how to compute the correct inclusion probabilities for the population elements selected in simple random sampling from a many-to-one frame. However, Musser's method can be extended to obtain inclusion probabilities for population elements in a simple random sample from the many-to-many frame as well. Second, Lavallée (1995) adapted the Weight Share Method, applied to longitudinal surveys, to the use of frames with a many-to-many structure.

The purpose of this paper is to develop an alternative methodology for estimating population totals, counts, and

[1] Terri L. Byczkowski, Institute for Policy Research, Martin S. Levy and Dennis J. Sweeney, Department of Quantitative Analysis and Operations Management, University of Cincinnati, Cincinnati, OH 45221, U.S.A.

means when using sampling frames with a many-to-many structure under simple and stratified random sampling designs. Also, expressions for the variance of those estimators are derived. The results which we develop are not only of intrinsic interest, but expressions for the variance of the estimators are essential for the exploration of the effects of correspondence imperfections inherent in many-to-many sampling frames on the precision of these estimates.

In section 2 we present these estimates in the simple random sampling without replacement (SRSWOR) case. We also describe the sampling methodology under which these estimators are applicable, state a result on bias, and develop expressions for their variance.

In section 3 some of the results are extended to the case of stratified random sampling. In section 4 we develop conclusions, discuss limitations and make suggestions for future research.

2. MANY-TO-MANY FRAMES FOR SIMPLE RANDOM SAMPLING

It is useful to think of the relationship between the frame and the target population as a graph. The sampling units in the frame and the elements of the target population are the two sets of nodes; arcs link the sampling units to elements of the target population. These arcs reveal the structure of the relationship between the frame and the target population. Figure 2.1 shows an example of a frame and target population with a many-to-many relationship. There are 7 sampling units in the frame, 6 elements in the target population and 10 links (arcs) between the sampling units and the elements of the population. Thus, a graph with 13 nodes and 10 arcs represents this many-to-many structure. In this paper we assume that each population element is linked to the set of frame units by at least one arc and that each frame unit is linked to the set of population elements by at least one arc as well.

Let us fix some notation. We find it convenient to identify both frame units and population elements with their respective indices. Let $F = \{1, 2, ..., N\}$ denote the set of indices for N sampling units, and let $T = \{1, 2, ..., M\}$ denote the set of indices for the M target population elements. An arc can be represented as an ordered pair; the first element of which comes from F, and the second from T. A population element k in T is said to be represented by sampling unit j in F, if it is linked to it by an arc denoted (jk). This means that when j is in the sample there is a nonzero probability of collecting data from population element k. We will denote by y_k the measurement of interest on target population element k in T.

We now describe the sampling methodology under which the estimators developed herein are appropriate. Assume a SRSWOR of size n frame units is selected from F. The number of *population elements* included in the sample and measured, however, depends upon the nature of the association between the frame units and the population elements.

Under SRSWOR, one of four scenarios can occur when a frame unit is selected. In the first scenario, a frame unit corresponds to one and only one population element (a one-to-one structure). Here the surveyor would simply collect the information concerning the single population element corresponding to the selected frame unit (see frame unit 1 of Figure 2.1).

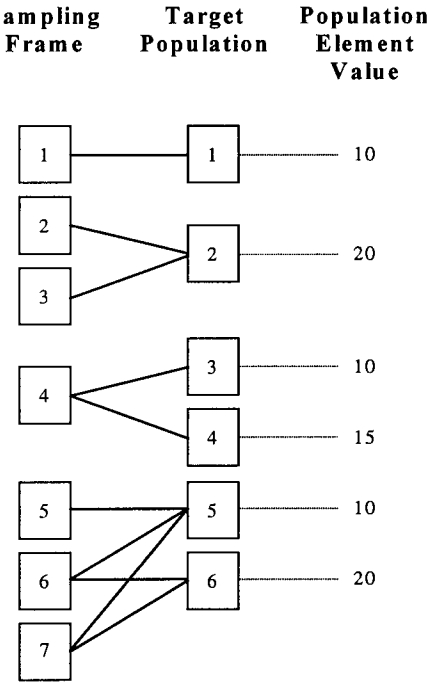


Figure 2.1. An example of the correspondence between the sampling frame and the target population

In the second scenario, several frame units correspond to one population element (a many-to-one structure). For example, in Figure 2.1, frame units 2 and 3 correspond to the single population element 2. In this case, if frame units 2 and/or 3 are included in the sample, information on population element 2 is collected. Thus, it is possible that population element 2 could appear in the sample, and as a record in the data set used to develop the estimates, up to two times.

In the third scenario, one frame unit corresponds to more than one population element (a one-to-many structure). For example, in Figure 2.1 frame unit 4 corresponds to population elements 3 and 4. Here, only one population element (3 or 4) is selected using a *randomization* independent of the choice of frame units. Economics dictated this policy because data collection entailed lengthy personal interviews conducted by individuals with technical backgrounds. In this paper we assume that these randomizations are conducted using equal probabilities. But, any probabilities could be used (*e.g.*, probability proportional to size) provided they are non-zero.

In the fourth scenario, a many-to-many structure exists. This is illustrated by frame units 5, 6 and 7 and population elements 5 and 6 in Figure 2.1. Since these complex cases are combinations of scenarios 2 and 3 above, the same sampling rules apply. For example, if frame unit 5 is selected, population element 5 is measured. If frame unit 6 is selected, only one of population elements 5 and 6 is randomly selected and measured.

2.1 Population Totals

2.1.1 Estimator for a Population Total

A many-to-many frame results in varying probabilities of selection. The estimators developed here involve a method of weighting, which is an extension of the estimator presented by Hansen *et al.* (1953a pp. 62-64). Their estimators and formulas for the variance of those estimators are restricted to the many-to-one frame structure. We extend those estimators to the many-to-many frame structure.

For a SRSWOR of size n, let $J_1, ..., J_n$ denote random variables such that $J_i = j$ if the i-th draw results in the selection of unit j from F. Hence $\Pr(J_i = j) = 1/N$ for j in F and $i = 1, ..., n$. Let $K_1, ..., K_n$ denote random variables such that $K_i = k$ if the i-th draw from F is followed by the selection of k from T. We can now think of drawing a random sample of arcs $\{(J_1K_1), ..., (J_nK_n)\}$ which has a joint probability distribution determined by both the SRSWOR sampling design and the subsequent randomization (if required) to choose an element in T. In particular, (J_iK_i) has marginal probability given by $\Pr\{(J_iK_i) = (jk)\} = (1/N)s_{jk}$, in which s_{jk} is the conditional probability given by, $s_{jk} = \Pr(K_i = k \mid J_i = j)$. That is, s_{jk} is the conditional probability of selecting population element k in T given that frame unit j in F is selected. These conditional probabilities will be referred to as arc probabilities and are illustrated for Figure 2.1 in Table 2.1.

Table 2.1
Arc Probabilities for Figure 2.1

Arc jk	1,1	2,2	3,2	4,3	4,4	5,5	6,5	6,6	7,5	7,6
s_{jk}	1	1	1	1/2	1/2	1	1/2	1/2	1/2	1/2

For k in T, let U_k denote the set of units in F that have arcs with a destination at k in T. Let $s_k = \sum_{j \in U_k} s_{jk}$. Using the language in Hansen *et al.* (1953a pp. 62-64) which motivated our development, we call s_k the *weight* for population element k in T. These weights for Figure 2.1 appear in Table 2.2.

Table 2.2
Calculation of the Population Element Weights (s_k) for Figure 2.1

k	1	2	3	4	5	6
(s_k)	1	2	1/2	1/2	2	1

Arc probabilities and weights are used to compute the marginal probabilities of the K_i, namely, $\Pr(K_i = k) = \sum_{j \in U_k} (1/N)s_{jk} = (1/N)s_k$, where k is in T, and $i = 1, ..., n$. Clearly, computing the arc probabilities is the key step in developing the correct weights for the data collected. It depends on properly ascertaining the graph structure for each sampling unit selected: a maximally connected (MC) subgraph. A connected subgraph is a subset of the nodes which are connected by a sequence of arcs. Maximal means that no node outside the subset is connected to a node belonging to the subset. There are 4 MC subgraphs in Figure 2.1. Each represents a different frame – population structure, namely, one-to-one, many-to-one, one-to-many, and many-to-many structure.

To develop the estimators it is not necessary to know the structure for the entire graph. It is only necessary to know the structure of the MC subgraphs to which *sampled* frame units belong.

We make the following observations about s_k and s_{jk}: (i) $s_k = W$ indicates that population element k has W times the probability of being selected on the i-th draw as that of a population element with a weight of one; (ii) $0 < s_k \le N$, $k = 1, ..., M$; (iii) $0 < s_{jk} \le 1$, $j \in U_k$ and $k = 1, ..., M$; (iv) with respect to the one-to-many frame structure, $s_{jk} = s_k$; (v) with respect to the many-to-one frame structure, $s_{jk} = 1$ for all k; and (vi) $\sum_{k=1}^{M} \sum_{j=1}^{N} s_{jk} = N$.

Now, let $x_1, ..., x_M$ denote the weighted values associated with the indices in T. That is, let $x_k = y_k/s_k$. Define random variables $x_{K_1}, ..., x_{K_n}$, associated with draws 1 through n from F, respectively, so that x_{K_i} takes the value x_k if $K_i = k$. Notice that we can write,

$$\mathrm{E}(x_{K_i}) = \sum_{k=1}^{M} x_k \Pr(K_i = k) = \frac{1}{N} \sum_{k=1}^{M} \frac{y_k}{s_k} s_k = \frac{Y}{N}, \qquad (2.1)$$

where $Y = \sum_{k=1}^{M} y_k$ is the true population total. We take as our estimator of the population total based upon a SRSWOR from a sampling frame with many-to-many structure,

$$\hat{Y} = \frac{N}{n} \sum_{i=1}^{n} x_{K_i}. \qquad (2.2)$$

Using (2.1) it follows that,

$$\mathrm{E}(\hat{Y}) = \mathrm{E}\left(\frac{N}{n} \sum_{i=1}^{n} x_{K_i} \right) = \frac{N}{n} \sum_{i=1}^{n} \mathrm{E}(x_{K_i}) = \frac{N}{n} n \frac{Y}{N} = Y.$$

We thus obtain,

Theorem 2-1: The estimator (2.2) for a population total used in SRSWOR is unbiased.

Using Figure 2.1, we now give a simple example of the use of this estimator. Suppose a simple random sample of four frame units was selected from the frame depicted in Figure 2.1 (2, 3, 4, and 7) which ultimately resulted in the selection of population elements 2, 4, and 5. The estimator of the population total,

$\hat{Y} = \frac{N}{n}\sum_{i=1}^{4} x_{K_i}$, has value $\frac{7}{4}\left[\frac{20}{2} + \frac{20}{2} + \frac{15}{(1/2)} + \frac{10}{2}\right] = \frac{385}{4}$.

The above estimator can also be used for a population count. We could estimate the size of the target population by letting $y_k = 1$ for all k. In addition, we could estimate the number of population elements that possess some characteristic by letting $y_k = 1$ for those population elements with the characteristic of interest and $y_k = 0$ for those without the characteristic.

2.1.2 Variance of the Estimator for a Population Total

First, some additional terminology and notation used in this section must be defined. Let P represent the set of all unordered pairs of arcs. We shall define an unordered pair of arcs as *inadmissible* if they cannot both be included in a sample. Formally let $Q = \{ j \text{ in } F$: more than one arc emerges from $j \}$. Then $R' = \{[jk, jk']: j \in Q$ and $k \neq k'\}$ is the set of unordered *inadmissible* pairs of arcs. Also, the set of unordered *admissible* pairs of arcs is the complementary set $R^* = P \setminus R'$.

To illustrate, consider Figure 2.1. The sampling methodology we employ requires that if frame unit 4 is selected, only one of population elements 3 and 4 can be included in the sample. Thus, {[4,3][4,4]} is an unordered inadmissible pair of arcs. The other unordered inadmissible pairs of arcs in Figure 2.1 are {[6,5][6,6]} and {[7,5][7,6]}. Thus, R' = {[4,3][4,4], [6,5][6,6], [7,5][7,6]}.

Theorem 2-2: The variance of the estimator (2.2) is,

$$V(\hat{Y}) = \frac{N}{n}\left[\sum_{k=1}^{M} \frac{y_k^2}{s_k} + 2\frac{(n-1)}{(N-1)}\sum \sum_{[jk,j'k']\in R^*} \left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'} s_{j'k'}}{s_{k'}}\right)\right] - Y^2, \quad (2.3)$$

where the double sum is over all unordered *admissible* pairs of arcs $[jk, j'k']$.

Proof:

$$V(\hat{Y}) = E\left[\left(\frac{N}{n}\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2$$

$$= \frac{N^2}{n^2} E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] - Y^2. \quad (2.4)$$

Now,

$$E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] = \sum_{i=1}^{n} E\left(x_{K_i}^2\right) + 2E\left(\sum\sum_{i<i'} \left(x_{K_i} x_{K_{i'}}\right)\right). \quad (2.5)$$

One can write

$$E\left(x_{K_i}^2\right) = \sum_{k=1}^{M} \left[x_k^2 \Pr(K_i = k)\right] = \sum_{k=1}^{M} \frac{y_k^2}{s_k^2}\frac{s_k}{N} = \frac{1}{N}\sum_{k=1}^{M} \frac{y_k^2}{s_k}. \quad (2.6)$$

As mentioned in Section 2.1, we can think of selecting a sample of arcs which ultimately leads to the selection of population elements. Each arc (jk) is associated with a value $x_k = y_k/s_k$ of the population element k at its destination. Thus, we can rewrite the double summation in (2.5) as a summation over admissible unordered pairs of arcs, R^*.

$$2E\left(\sum_{i'}\sum_{i<i'} \left(x_{K_i} x_{K_{i'}}\right)\right) =$$

$$2\binom{n}{2}\sum\sum_{[jk,j'k']\in R^*} \left[(x_k x_{k'}) \Pr(K_i = k, K_{i'} = k')\right]. \quad (2.7)$$

Now, by virtue of the independence of the randomization and the choice of frame units:

$$\Pr(\text{select}[jk, j'k'] \text{ in } R^*) = \Pr(\text{select}\{j, j'\} \text{ in } F)$$

$$\Pr(\text{select}[jk, j'k'] \text{ in } R^* \mid \text{select}\{j, j'\} \text{ in } F) = \frac{1}{\binom{N}{2}} s_{jk} s_{j'k'}.$$

Substituting into (2.7) results in,

$$n(n-1)\sum\sum_{[jk,j'k']\in R^*} \left[(x_k x_{k'})\frac{1}{\binom{N}{2}} s_{jk} s_{j'k'}\right] =$$

$$\frac{2n(n-1)}{N(N-1)}\sum\sum_{[jk,j'k']\in R^*} \left[\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'} s_{j'k'}}{s_{k'}}\right)\right]. \quad (2.8)$$

Now substituting (2.6) and (2.8) into (2.5) yields,

$$E\left[\left(\sum_{i=1}^{n} x_{K_i}\right)^2\right] = \frac{n}{N}\left[\sum_{k=1}^{M} \frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*} \left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'} s_{j'k'}}{s_{k'}}\right)\right]. \quad (2.9)$$

Finally substituting (2.9) into (2.4) gives the result (2.3).

Equation (2.3) is a generalization of the formula developed by Bandyopadhyay and Adhikari (1993) for the variance of the estimate of a population total in the case of the many-to-many frame structure. It can be shown that (2.3) reduces to their formula when the sampling frame is restricted to a many-to-one structure.

Corollary 2-1: An alternative form of the variance formula in **Theorem 2-2** is:

$$V(\hat{Y}) = \frac{N}{n}\left[\sum_{k=1}^{M}\frac{y_k^2}{s_k} + \frac{(n-1)}{(N-1)}\left(\left(\sum_{jk}\frac{y_k}{s_k}s_{jk}\right)^2 - \sum_{jk}\left(\frac{y_k}{s_k}s_{jk}\right)^2 - 2\sum\sum_{[jk,jk']\in R'}\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}\right)\right] - Y^2.$$

Proof:
Write,

$$\left(\sum_{jk}\frac{y_k s_{jk}}{s_k}\right)^2 = \sum_{jk}\left(\frac{y_k s_{jk}}{s_k}\right)^2 + 2\sum_{[jk,j'k']\in R^*}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}} + 2\sum_{[jk,jk']\in R'}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}.$$

It follows that:

$$\sum_{[jk,j'k']\in R^*}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}} = \frac{1}{2}\left(\sum_{jk}\frac{y_k s_{jk}}{s_k}\right)^2 - \frac{1}{2}\sum_{jk}\left(\frac{y_k s_{jk}}{s_k}\right)^2 - \sum_{[jk,jk']\in R'}\sum\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{jk'}}{s_{k'}}.$$

Substituting the above expression into (2.3) provides the result.

This formula is computationally simpler. Note that (2.3) requires that the term

$$\left(\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right)$$

be summed over all unordered admissible pairs of arcs (R^*), whereas this alternative formula only requires a summation over pairs of arcs that are inadmissible (R'). In most practical scenarios the number of admissible pairs of arcs will be far greater than the number of inadmissible pairs of arcs.

2.2 Population Means

2.2.1 Estimator for a Population Mean

The estimator for a population mean presented here extends the estimator presented by Hansen *et al.* (1953a) to the many-to-many frame structure.

Associated with the n draws from F, define random variables s_{K_i} and $z_{K_i} = 1/s_{K_i}$, so that s_{K_i} takes value s_k if $K_i = k$ for $i = 1, ..., n$ and $k = 1, ..., M$. The estimator for a population mean,

$$\bar{Y} = \frac{1}{M}\sum_{k=1}^{M} y_k,$$

when using SRSWOR and a many-to-many frame is:

$$\hat{\bar{Y}} = \frac{\sum_{i=1}^{n} x_{K_i}}{\sum_{i=1}^{n} z_{K_i}}. \qquad (2.10)$$

2.2.2 Mean Square Error (MSE) of the Estimator for a Population Mean

The estimator for a population mean is biased because it is a ratio estimator. But, it is well known that this bias becomes negligible for large samples and the bias is of order $1/n$ (Cochran 1977, p. 160).

Our approximation of the MSE requires a summation over R^{**}, the set of all *ordered admissible* pairs of arcs. Thus, if $[jk, j'k'] \in R^*$, then both $[jk, j'k'] \in R^{**}$ and $[j'k', jk] \in R^{**}$.

To approximate the mean square error of the estimator (2.10), we use

$$\mathrm{MSE}\left(\hat{\bar{Y}}\right) \approx \frac{M^2}{nN\left(\sum_{k=1}^{M}\frac{1}{s_k}\right)^2}\left[\left(\sum_{k=1}^{M}\frac{y_k^2}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*}\frac{y_k s_{jk}}{s_k}\frac{y_{k'}s_{j'k'}}{s_{k'}}\right) - 2\bar{Y}\left(\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^{**}}\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right)\right] + \bar{Y}^2\left(\sum_{k=1}^{M}\frac{1}{s_k} + \frac{2(n-1)}{(N-1)}\sum\sum_{[jk,j'k']\in R^*}\frac{s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right)\Bigg]. \qquad (2.11)$$

To justify this approximation let

$$\bar{x} = \frac{\sum_{i=1}^{n} x_{K_i}}{n},\ \bar{z} = \frac{\sum_{i=1}^{n} z_{K_i}}{n} \text{ and } \bar{Z} = \frac{\sum_{k=1}^{M}\frac{1}{s_k}}{M}.$$

Because $\hat{\bar{Y}}$ is a ratio of two estimates, the well known approximation for the mean square error (Cochran 1977, pp. 32-33) can be used:

$$\mathrm{MSE}(\hat{\bar{Y}}) = \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{z}}\right)^2 \approx \mathrm{E}\left(\frac{\bar{x} - \bar{Y}\bar{z}}{\bar{Z}}\right)^2 =$$

$$\frac{1}{\bar{Z}^2}\left[\mathrm{E}(\bar{x}^2) - 2\bar{Y}\mathrm{E}(\bar{z}\bar{x}) + \bar{Y}^2\mathrm{E}(\bar{z}^2)\right] =$$

$$\frac{M^2}{n^2\left(\sum_{j=1}^{M} z_j\right)^2}\left[\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\right)^2 - 2\bar{Y}\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\sum_{i=1}^{n} z_{K_i}\right) + \bar{Y}^2\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2\right]. \quad (2.12)$$

The first expectation in (2.12) is simply (2.9). Next, using (2.1) on the middle term in (2.12) results in

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\sum_{i=1}^{n} z_{K_i}\right) = \mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\frac{1}{s_{K_i}}\right) +$$

$$\mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i'}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) = \frac{n}{N}\sum_{k=1}^{M}\frac{y_k}{s_k} + \mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right).$$

Using (2.7) and (2.9) yields,

$$\mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) = n(n-1)\mathrm{E}\left(x_{K_i}\frac{1}{s_{K_{i'}}}\right) =$$

$$n(n-1)\sum_{[jk,j'k']\in R^{**}}\sum\left[\left(\frac{y_k}{s_k}\frac{1}{s_{k'}}\right)\Pr\left(x_{K_i} = \frac{y_k}{s_k}, \frac{1}{s_{K_{i'}}} = \frac{1}{s_{k'}}\right)\right] =$$

$$n(n-1)\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k}{s_k}\frac{1}{s_{k'}}\left(\frac{1}{N(N-1)}s_{jk}s_{j'k'}\right) =$$

$$\frac{n(n-1)}{N(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}.$$

Note that the double sum is over all admissible *ordered* pairs of arcs. Therefore,

$$\mathrm{E}\left(\sum_{i=1}^{n} x_{K_i}\frac{1}{s_{K_i}}\right) + \mathrm{E}\left(\sum_{\substack{i=1\\ i\neq i'}}^{n}\sum_{i'=1}^{n} x_{K_i}\frac{1}{s_{K_{i'}}}\right) =$$

$$\frac{n}{N}\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{n(n-1)}{N(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}} =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{y_k}{s_k} + \frac{(n-1)}{(N-1)}\sum_{[jk,j'k']\in R^{**}}\sum\frac{y_k s_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Finally, similar to (2.1),

$$\mathrm{E}\left(\sum_{i=1}^{n} z_{K_i}\right)^2 = \mathrm{E}\left(\sum_{i=1}^{n}\frac{1}{s_{K_i}}\right)^2 =$$

$$\frac{n}{N}\left(\sum_{k=1}^{M}\frac{1}{s_k} + \frac{2(n-1)}{(N-1)}\sum_{[jk,j'k']\in R^{*}}\sum\frac{S_{jk}}{s_k}\frac{s_{j'k'}}{s_{k'}}\right).$$

Substituting these expectations into equation (2.12) yields (2.11).

3. ESTIMATORS FOR MANY-TO-MANY FRAMES UNDER STRATIFIED RANDOM SAMPLING

3.1 Introduction

In this section we develop the estimators for a population count, mean, and total in the many-to-many frame case, when stratified random sampling is used. First, however, it is necessary to describe the sampling methodology under which these estimates are appropriate. Figure 3.1 provides an example that will be used throughout this section.

3.2 The Sampling Methodology

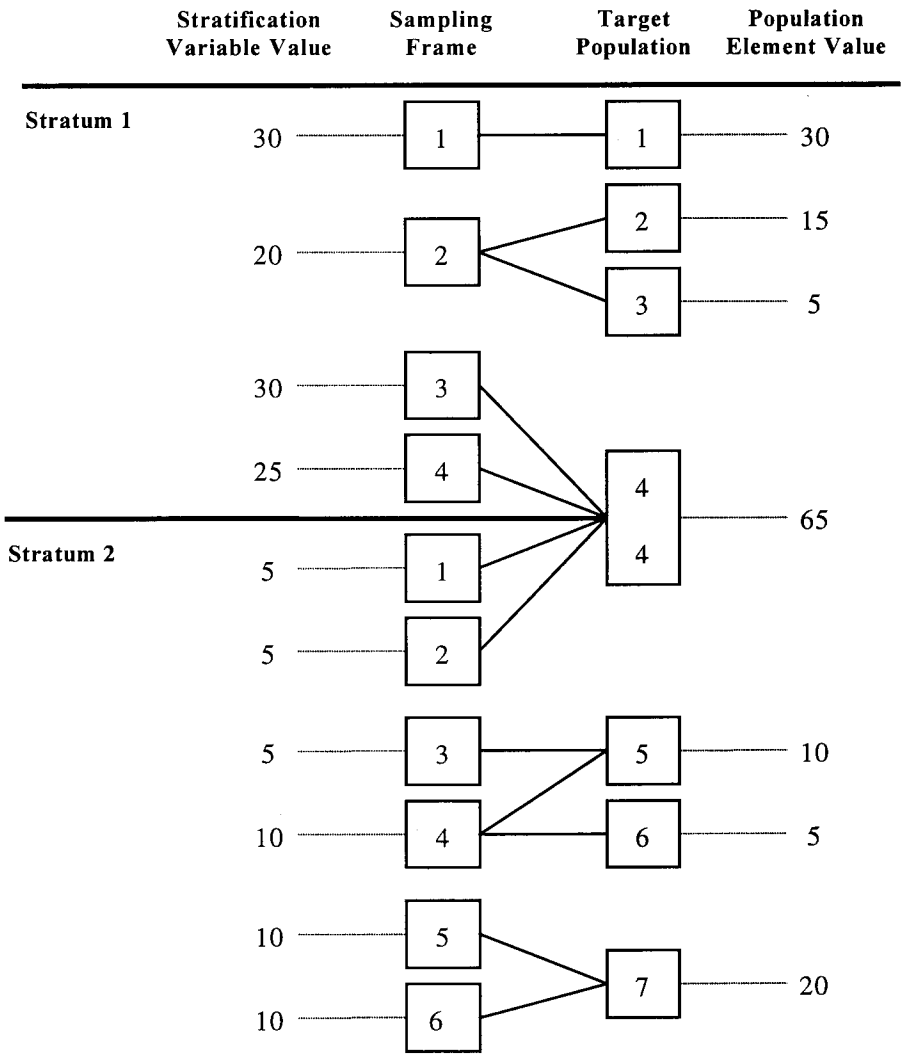
The same scenarios that were described in SRSWOR occur with respect to stratified random sampling. However, there are some additional problems that can arise in this case.

Consider the building characteristics study that motivated this research. Assume that the population element value in Figure 3.1 is the building size, and the stratification variable is electrical usage associated with the street address. Because the frame of street addresses had a many-to-many correspondence with the target population of commercial buildings, the following problems arose in addition to those mentioned in Section 2.1:

1. Mis-stratification: For example, frame unit (street address) 2 in stratum 1 appeared to be a large building because of the large electrical usage associated with it, and as a result, it was placed in the first stratum. The data collection revealed that the street address actually corresponded to two small buildings (population elements 2 and 3). In another example, frame units 5 and 6 in stratum 2 appeared to be two small buildings in the frame, and were placed in the second stratum. But, the corresponding population element 7 is one large building with two street addresses.
2. Crossover: For example, frame units 3 and 4 in stratum 1, and frame units 1 and 2 in stratum 2 each have a different street address and, as a result, appear in the frame to be two small and two large buildings. But, data

collection revealed that all four street addresses corresponded to only one building (*e.g.*, a strip mall). In this case, not only is mis-stratification a problem, but not all the frame units associated with a single building are included in the same strata. That is, one population element (*i.e.*, building) "crosses over" multiple strata.

In the next section we develop estimators for population totals and counts and show that these estimators are unbiased despite mis-stratification and crossover. As is usually the case, however, mis-stratification increases the variance of the estimates. Also, insofar as crossover induces mis-stratification, it too increases the variances of the estimates.



Note: Frame units were placed in stratum 1 if the value of the stratification variable was 20 or more. Otherwise, the frame units were placed in stratum 2.

Figure 3.1. An example of the correspondence between the frame and the target population in stratified random sampling

Table 3.1
Arc Probabilities for Figure 3.1

Arc hjk	1,1,1	1,2,2	1,2,3	1,3,4	1,4,4	2,1,4	2,2,4	2,3,5	2,4,5	2,4,6	2,5,7	2,6,7
s_{hjk}	1	1/2	1/2	1	1	1	1	1	1/2	1/2	1	1

3.3 Population Totals and Counts

3.3.1 Estimator for a Population Total

The estimator developed here involves a method of weighting which extends the estimator presented in Hansen *et al.* (1953a, pp. 62-64) to stratified random sampling when using a many-to-many frame.

Assume that F has been partitioned into L mutually exclusive and exhaustive strata $F_1, ..., F_L$ of size $N_1, ..., N_L$ respectively. Units in F_h will be denoted hj where $j = 1, ..., N_h$ and $h = 1, ..., L$. Also, assume that a stratified random sample (without replacement) of size $n = n_1 + ... + n_L$ has been drawn, where n_h is the sample size from F_h. Let $hJ_1, ..., hJ_{n_h}$ denote random variables such that $hJ_i = hj$ if the i-th draw from F_h results in the selection of hj. Let $hK_1, ..., hK_{n_h}$ denote random variables such that $hK_i = k$ if the i-th draw from F_h is followed by the selection of k from T. If hjk denotes the arc that originates at frame unit hj in F_h and terminates at k in T, the marginal probability of the random arc (hJ_i, hK_i) is given by,

$$\Pr\{(hJ_i, hK_i) = (hjk)\} = \frac{1}{N_h} s_{hjk},$$

in which $s_{hjk} = \Pr(hK_i = k \mid hJ_i = hj)$ is an arc probability. Note that s_{hjk} is the conditional probability of selecting population element k in T given that frame unit hj has been chosen. Assuming equal randomization probabilities, Table 3.1 shows the arc probabilities for Figure 3.1.

Let W_k denote the set of frame units hj in F that have arcs with a destination at k in T. For example, $W_4 = \{(1, 3), (1, 4), (2, 1), (2, 2)\}$. Also, define the population element weight $s_k = \sum_{hj \in W_k} S_{hjk}$.

Table 3.2 contains the weights (s_k) for all the population elements in Figure 3.1. The same observations concerning arc weights (s_{hjk}) and population element weights (s_k) made in section 2.3.1 apply here.

Table 3.2
Population Element Weights (s_k) for Figure 3.1

k	1	2	3	4	5	6	7
s_k	1	1/2	1/2	1+1+1+1=4	1+1/2=3/2	1/2	1+1=2

For each $h = 1, ..., L$ and $i = 1, ..., n_h$, let x_{hK_i} be random variables such that $x_{hK_i} = y_k / s_k$ if k in T is selected as a result of the selection of some hj in F_h.

The estimator of a population total for stratified random sampling, when using a sampling frame with a many-to-many structure is:

$$\hat{Y}_{st} = \sum_{h=1}^{L} \hat{Y}_h, \text{ where } \hat{Y}_h = \frac{N_h}{n_h} \sum_{i=1}^{n_h} x_{hK_i}. \qquad (3.1)$$

3.3.2 Variance of the Estimator for a Population Total

Prior to developing the variance of estimator (3.1), some additional terminology must be defined. Let q_{hk} denote the "stratum element weight". This additional weight is necessary because of the potential of crossover. Let U_{hk} denote the set of frame units in F_h that have arcs with a destination at population element k. For example, $U_{24} = \{(2, 1), (2, 2)\}$. Then define $q_{hk} = \sum_{hj \in U_{hk}} s_{hjk}$. To illustrate, recall in Figure 3.1 population element 4 is represented by two frame units in stratum 2, so $q_{24} = \sum_{2j \in U_{24}} s_{2j4} = 2$.

The weight q_{hk} plays the role of s_k when selection is restricted to F_h. In fact, $q_{hk} = s_k$ when there is no crossover. The probability of selecting any frame unit from F_h on step i out of n_h is $1/N_h$. But, the probability of selecting a population element k represented by a frame unit in F_h is $\Pr(hK_i = k) = q_{hk}/N_h$, for all $i = 1, ..., n_h$.

In order to develop the proof in this section, we introduce the term "apportioned stratum total" denoted by Y_h^*. In effect, the values of the population elements that are represented by frame units in multiple strata are apportioned among those strata. Let V_h denote the set of population elements associated with frame units in F_h. In our example $V_1 = \{1, 2, 3, 4\}$ and $V_2 = \{4, 5, 6, 7\}$. Let

$$Y_h^* = \sum_{k \in V_h} y_k q_{hk} / s_k$$

where y_k is the value of population element k, $k = 1, 2, ..., M$. When crossover is present, use of the weights q_{hk} and s_k apportion the measure y_k among the strata in which population element k is represented. We can think of the use of these weights as distributing the population element value among the strata depending upon the number of times the population element is represented in a stratum relative to the total number of times it is represented in the frame. For example in Figure 3.1 Y_1^* and Y_2^* are calculated as follows:

$$Y_1^* = \frac{30(1)}{1} + \frac{15(1/2)}{1/2} + \frac{5(1/2)}{1/2} + \frac{65(2)}{4} = 82.5$$

$$Y_2^* = \frac{65(2)}{4} + \frac{10(3/2)}{3/2} + \frac{5(1/2)}{1/2} + \frac{20(2)}{2} = 67.5.$$

Note that $\sum_{h=1}^{L} Y_h^* = Y$ whether or not crossover exists.

Theorem 3-1: The estimator for a population total (3.1) is unbiased.

Proof:

From (3.1),

$$\mathrm{E}\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} \frac{N_h}{n_h} \sum_{i=1}^{n_h} \mathrm{E}\left(x_{hK_i}\right). \qquad (3.2)$$

For each $i = 1, ..., n_h$,

$$\mathrm{E}\left(x_{hK_i}\right) = \sum_{k \in V_h} \frac{y_k}{s_k} \Pr(hK_i = k) =$$

$$\sum_{k \in V_h} \frac{y_k}{s_k} \frac{q_{hk}}{N_h} = \frac{1}{N_h} \sum_{k \in V_h} \frac{y_k q_{hk}}{s_k} = \frac{1}{N_h} Y_h^*. \qquad (3.3)$$

Substituting (3.3) into equation (3.2) yields $E(\hat{Y}_{st}) = Y$.

In the main result below we need the following notation. Let R_h^* and R_h' be the set of admissible and inadmissible unordered pairs of arcs originating in F_h, respectively. Definitions of the above are identical to the corresponding concepts for the SRSWOR case, but restricted now to strata.

Theorem 3-2: The variance of (3.1) is:

$$V\left(\hat{Y}_{st}\right) = \sum_{h=1}^{L} S_h^2, \tag{3.4}$$

where,

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k\in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{2(n_h-1)}{(N_h-1)} \times \sum_{[hjk,hj'k']\in R_h^*}\left(\frac{y_k s_{hjk}}{s_k}\frac{y_{k'} s_{hj'k'}}{s_{k'}}\right)\right] - \left(\sum_{k\in V_h}\frac{y_k q_{hk}}{s_k}\right)^2. \tag{3.5}$$

Proof: First write,

$$V\left(\hat{Y}_{st}\right) = E\left(\hat{Y}_{st}\right)^2 - Y^2 = E\left(\sum_{h=1}^{L}\hat{Y}_h^2\right) - \sum_{h=1}^{L}\left(Y_h^*\right)^2 + 2\left(E\left(\sum_{h<h'}\hat{Y}_h\hat{Y}_{h'}\right) - \sum_{h<h'}Y_h^*Y_{h'}^*\right). \tag{3.6}$$

The last two terms cancel because $\hat{Y}_h$ and $\hat{Y}_{h'}$ are independent. This follows since apportionment creates a new stratified population containing no crossover and samples chosen within different strata are independent. Thus, with

$$S_h^2 = E\left(\hat{Y}_h\right)^2 - \left(Y_h^*\right)^2, \quad V\left(\hat{Y}_{st}\right) = E\left(\sum_{h=1}^{L}\hat{Y}_h^2\right) - \sum_{h=1}^{L}\left(Y_h^*\right)^2 = \sum_{h=1}^{L}\left(E\left(\hat{Y}_h^2\right) - \left(Y_h^*\right)^2\right) = \sum_{h=1}^{L} S_h^2.$$

Now,

$$E\left(\hat{Y}_h\right)^2 = \frac{N_h^2}{n_h^2}E\left(\sum_{i=1}^{n_h} x_{hK_i}\right) = \frac{N_h^2}{n_h^2}\left(\sum_{i=1}^{n_h} E\left(x_{hK_i}\right)^2 + 2E\left(\sum_{i<i'} x_{hK_i}x_{hK_{i'}}\right)\right). \tag{3.7}$$

For each $i = 1, ..., n_h$,

$$E\left(x_{hK_i}\right)^2 = \sum_{k\in V_h}\left[\left(\frac{y_k}{s_k}\right)^2 \Pr(hK_i = k)\right] = \sum_{k\in V_h}\left[\left(\frac{y_k}{s_k}\right)^2\frac{q_{hk}}{N_h}\right]. \tag{3.8}$$

Then, using equation (2.7) and (2.8),

$$2E\left(\sum_{i>i'} x_{hK_i}x_{hK_{i'}}\right) = 2\binom{n_h}{2}E\left(x_{hK_i}x_{hK_{i'}}\right) =$$

$$n_h(n_h-1)\sum_{[hjk,hj'k']\in R_h^*}\frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\Pr(hK_i = k, hK_{i'} = k') =$$

$$n_h(n_h-1)\sum\sum_{[hjk,hj'k']\in R_h^*}\left[\left(\frac{y_k}{s_k}\frac{y_{k'}}{s_{k'}}\right)\binom{N_h}{2}^{-1}s_{jk}s_{j'k'}\right] =$$

$$\frac{2n_h(n_h-1)}{N_h(N_h-1)}\sum\sum_{[hjk,hj'k']\in R_h^*}\left[\frac{y_k s_{hjk}}{s_k}\frac{y_{k'}s_{hj'k'}}{s_{k'}}\right]. \tag{3.9}$$

Equation (3.5) now follows from (3.8), (3.9), and the definition of Y_h^*.

Using the method of Corollary 2-1, (3.5) can be simplified for computing purposes as follows:

$$S_h^2 = \frac{N_h}{n_h}\left[\sum_{k\in V_h} q_{hk}\left(\frac{y_k}{s_k}\right)^2 + \frac{(n_h-1)}{(N_h-1)}\left[\left(\sum_{hjk\in A_h}\frac{y_k s_{hjk}}{s_k}\right)^2 - \sum_{hjk\in A_h}\left(\frac{y_k s_{hjk}}{s_k}\right)^2 - 2\sum_{[hjk,hjk']\in R_h'}\left(\frac{y_k s_{hjk}}{s_k}\frac{y_{k'}s_{hjk'}}{s_{k'}}\right)\right]\right] - \left(\sum_{k\in V_h}\frac{y_k q_{hk}}{s_k}\right)^2,$$

where A_h denotes the set of arcs that originate at frame units in F_h.

3.4 Population Means

3.4.1 Estimator for a Population Mean

The estimator developed here for a population mean for stratified random sampling extends the estimator presented by Hansen *et al.* 1953a (pp. 62-64) to the case of a stratified random sample from a many-to-many frame.

The estimator for a population mean when using stratified random sampling and a many-to-many frame is:

$$\hat{\bar{Y}}_{st} = \sum_{h=1}^{L}\frac{N_h}{N}\hat{\bar{Y}}_h, \text{ where } \hat{\bar{Y}}_h = \frac{\sum_{i=1}^{n_h} x_{hK_i}}{\sum_{i=1}^{n_h}\frac{1}{s_{hK_i}}}. \tag{3.10}$$

As in the SRSWOR case, the estimator for a population mean is biased because it is a ratio estimator.

4. CONCLUSIONS

In this paper we have developed estimators for population totals, counts and means that are appropriate when the sampling frame has a many-to-many structure. We have focused on simple random sampling and stratified random sampling designs.

We used the method of weighting described in this paper in a study of commercial buildings for which a stratified random sample was employed. In this study, for which the sampling frame consisted of street addresses, interviewers recorded any additional street addresses that pertained to the selected building. It was then determined whether or not these additional street addresses were listed in the sampling frame, and whether or not they were connected to other population elements (commercial buildings). In more complex scenarios, the interviewers sometimes resorted to schematic sketches of the buildings and labelling all the pertinent addresses. This allowed us to determine the structure of all MC subgraphs in our sample and to develop the appropriate weights s_k.

In addition, we developed formulas for the variance of some of the estimators presented in this paper. It should be noted that these variance formulas are population parameters and do not translate readily into corresponding sample estimates. In fact, the authors are unaware of any optimal method for estimating the variances discussed in this paper. However, there are many computer intensive methods (balanced repeated replication, bootstrapping, *etc.*) for estimating variances in complex sample surveys (Wolter 1985). It should be emphasized that when using our estimators, each of these variance estimation schemes aims at a common target: the variance formulas we have developed.

Nevertheless, the usefulness of these variance formulas is in their application to the task of exploring the effects of frame imperfections, along with population characteristics, on the precision of estimation. Such an exploration, another future area of research, should result in recommendations and guidelines for the survey researcher on how to manage a frame with a many-to-many structure. That is, based upon frame and population characteristics, the survey researcher would be able to make strategic decisions concerning the options available: canvassing a population to remove correspondence imperfections, or using the estimators described herein.

Another area of future research is a comparison of the precision of our estimators to that of other estimators, such as the Horvitz-Thompson estimator. As noted in the introduction the Horvitz-Thompson estimator can be applied to sampling involving a many-to-many frame structure. An advantage of the Horvitz-Thompson estimator is that with properly identified first and second order inclusion probabilities, one can obtain both an estimate of a population characteristic and an unbiased estimate of its variance. In addition, the first order inclusion probabilities can be derived in a manner similar to Musser (1993) based only upon information from the MC subgraphs. However, these probabilities are very difficult to compute in a complex many-to-many frame structure such as ours. It is, however, relatively easy to calculate the necessary weights for our estimators.

ACKNOWLEDGMENTS

The authors thank the individuals at Cinergy Corporation for the opportunity to conduct the building characteristics survey which motivated this research. We also thank the three referees for their excellent suggestions that have led to a significant improvement of this paper.

REFERENCES

BANDYOPADHYAY, S., and ADHIKARI, A.K. (1993). Sampling from imperfect frames with unknown amount of duplication. *Survey Methodology*, 19, 193-197.

BIRNBAUM, Z.W., and SIRKEN, M.G. (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. *Vital and Health Statistics*, PHS Publication 1000, Ser. 2, *Data Evaluation and Methods Research*, no. 11. Hyattsville, MD: National Center for Health Statistics, Public Health Service, U.S. Department of Health and Human Services.

CASADY, R.J., and SIRKEN, M.G. (1980). A multiplicity estimator for multiple frame sampling. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 601-605.

COCHRAN, W.G. (1977). *Sampling Techniques*, (3rd ed.). New York: Wiley & Sons.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953a). *Sample Survey Methods and Theory 1, Methods and Applications*. New York: Wiley & Sons.

HANSEN, M.H., HURWITZ, W.N., and MADOW, W.G. (1953b). *Sample Survey Methods and Theory 2, Theory*. New York: Wiley & Sons.

HORVITZ, D.G., and THOMPSON, D.J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47, 663-685.

KISH, L. (1965). *Survey Sampling*. New York: Wiley & Sons.

LAVALLÉE, P. (1995). Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. *Survey Methodology*, 21, 25-32.

LESSLER, J.T., and KALSBEEK, W.D. (1992). *Nonsampling Error in Surveys*. New York: Wiley & Sons.

MUSSER, O. (1993). Unbiased estimation in the presence of frame duplication. *Proceedings of the International Conference on Establishment Surveys*, 889-892.

SIRKEN, M.G. (1972a). Household surveys with multiplicity. *Journal of the American Statistical Association*, 65, 257-266.

SIRKEN, M.G. (1972b). Stratified sample surveys with multiplicity. *Journal of the American Statistical Association*, 65, 224-227.

U.S. DEPARTMENT OF ENERGY, Energy Information Administration (1992). *Commercial Buildings Energy Consumption Survey*.

WOLTER, K.M. (1985). *Introduction to Variance Estimation*. New York: Springer-Verlag.

WRIGHT, T., and TSAO, H.J. (1983). A frame on frames: An annotated bibliography, (Ed., Tommy Wright). *Statistical Methods and the Improvement of Data Quality*, Orlando, Florida: Academic Press, 25-72.