SSC Annual Meeting, June 1996
Proceedings of the Survey Methods Section

ESTIMATION IN DUAL FRAME SURVEYS WITH COMPLEX DESIGNS

J.N.K. Rao and C.J. Skinner[1]

ABSTRACT

Dual frame surveys with at least one of the samples selected by a complex design are studied. A pseudo maximum likelihood estimator of a population total is proposed and its asymptotic properties are studied. A practical advantage of the proposed estimator is that it uses the same weights for all the variables, unlike some other estimators proposed in the literature. Alternative "single frame" estimators are also studied.

KEY WORDS: Overlapping frames; Pseudo maximum likelihood estimator; Single frame estimators.

RÉSUMÉ

Les auteurs étudient les enquêtes à double base de sondage pour lesquelles on prélève au moins un échantillon au moyen d'un plan d'échantillonnage complexe. Ils proposent un pseudo-estimateur du maximum de vraisemblance pour une population et en étudient les propriétés asymptotiques. Un avantage pratique de l'estimateur est qu'il applique les mêmes poids à toutes les variables, à l'inverse de quelques autres estimateurs qu'on retrouve dans la documentation. On examine aussi d'autres estimateurs à "base de sondage unique".

MOTS CLÉS: Chevauchement des bases de sondage; pseudo-estimateur du maximum de vraisemblance; estimateurs à base de sondage unique.

1. INTRODUCTION

In a dual frame survey, samples are drawn independently from two frames A and B. These frames may overlap and are assumed together to cover the population U of interest so that $U = A \cup B$. The data obtained from the two samples are combined to produce estimates of population totals or means.

Such surveys arise in a variety of settings. In a common example, one frame is complete, say $A = U$, but is expensive to sample, whereas the other frame B is incomplete but cheap to sample. Hartley (1962, 1974) discusses the advantages of sampling both frames in these circumstances to arrive at more efficient estimators compared to sampling from the complete frame only. Lepkowski and Groves (1986) describe an application where A is an address frame for which a sample of addresses is visited, and B is a telephone frame for which a sample of numbers is telephoned at lesser expense. Another example arises with rare populations. Here a screening sample from a general population address frame is combined with a sample from a much smaller list of individuals which may be incomplete but is thought likely to contain a high proportion of individuals in the rare population and is thus cheaper to sample (Kalton and Anderson 1986). Further examples are presented by Hartley (1974), Bankier (1986) and others.

For the case of simple random sampling from both frames, dual frame estimators of population totals and means have been proposed by Hartley (1962), Lund (1968), and Fuller and Burmeister (1972). Bankier (1986), and Skinner (1991) have also considered stratified random sampling. Only Fuller and Burmeister (1972) and Hartley (1974) seem to have considered the general case when at least one of the samples is selected by a complex design involving, for example, multi-stage sampling. Their estimators both consist of weighted combinations of domain estimates, but have the property that the weights depend on the variable of interest, y. This implies a need to recompute weights for every variable which will usually be operationally inconvenient in practice for statistical agencies conducting surveys with large numbers of variables. More importantly, such weights do not ensure consistency of figures when aggregated over variables,

[1] J.N.K. Rao, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario, Canada, K1S 5B6; C.J. Skinner, Department of Social Statistics, University of Southampton, Southampton, United Kingdom, S09 5NH.

unlike a single set of weights computed and used for all variables. Our primary aim in this article is to consider alternative estimators under complex designs where the same weights are used for all variables. Specifically, we propose a class of "pseudo" maximum likelihood (ML) estimators which have certain optimality properties under simple random sampling from both frames and which are likely to have reasonable efficiency under more complex designs used in practice. Properties of the pseudo ML estimator are studied. A consistent estimator of variance of the pseudo ML estimator is also given. Some alternative single frame estimators, based on the design induced by the two separate designs, are considered. Finally, results of a limited simulation study on the efficiency of the pseudo ML estimator are reported.

2. ESTIMATION PROBLEM

Following the classical notation of Hartley (1962), let $a=A \cap B^c$, $b=A^c \cap B$ and $ab=A \cap B$, where c denotes complement of a set. Thus, if N, N_A, N_B, N_a, N_b and N_{ab} denote the sizes of the sets U, A, B, a, b and ab respectively, then $N=N_a+N_b+N_{ab}, N_A=N_a+N_{ab}$ and $N_B=N_b+N_{ab}$. Further, let y_i be a value associated with the population unit i and define the totals $Y=\Sigma_U y_i, Y_a=\Sigma_a y_i, Y_b=\Sigma_b y_i$ and $Y_{ab} = \Sigma_{ab} y_i$. Note that

$$Y=Y_a+Y_b+Y_{ab}. \qquad (1)$$

Similarly, define the means $\mu = Y/N, \mu_a = Y_a / N_a$, $\mu_b = Y_b/N_b$ and $\mu_{ab}=Y_{ab} / N_{ab}$. The population total Y will be taken here as the parameter of primary interest.

Let s_A and s_B be samples drawn independently from A and B according to specified probability sampling designs $p'(s_A)$ and $p''(s_B)$ respectively. Suppose that y_i is observed for each unit in s_A and s_B. The estimation problem then is to use these data to construct a suitable estimator $\hat{Y}$ of Y and an estimator of its variance. This problem depends importantly on what is known about N_A, N_B and N_{ab}: Case 1 with N_A, N_B and N_{ab} known; Case 2 with N_A, N_B known but N_{ab} unknown; Case 3 with N_A, N_B and N_{ab} unknown. Case 1 may arise if frames A and B are available as lists of known length and if the overlap size N_{ab} can either be determined from these lists or is known; for example if frame A is complete then $N_{ab}=N_B$. Case 2 may arise if again A and B are lists of known length but reliable determination of overlap size is not possible because of practical difficulties in matching. Case 3 may arise if, for example, both A and B consist of clusters and lists are only available of clusters with measures of size. We focus on Case 2 in this article, but refer the reader to Skinner and Rao (1996) for cases 1 and 3.

3. ESTIMATORS OF HARTLEY AND FULLER AND BURMEISTER

Since $N_a, N_{ab}, Y_a, Y_{ab}, \mu_a$ and μ_{ab} are all characteristics of subpopulation A and since s_A is drawn from A by a conventional sampling design, standard estimators $\hat{N}_a', \hat{N}_{ab}', \hat{Y}_a', \hat{Y}_{ab}', \hat{\mu}_a'$ and $\hat{\mu}_{ab}'$ based on the data from s_A may be employed. For example, the following estimators may be used: $\hat{N}_a' = \Sigma_{s_a} w_i'$, $\hat{N}_{ab}'=\Sigma_{s'_{ab}} w_i', \hat{Y}_a'=\Sigma_{s_a} w_i' y_i, \hat{Y}_{ab}'=\Sigma_{s'_{ab}} w_i' y_i, \hat{\mu}_a' = \hat{Y}_a'/\hat{N}_a'$ and $\hat{\mu}_{ab}' = \hat{Y}_{ab}'/\hat{N}_{ab}'$, where $s_a = s_A \cap a, s_{ab}' = s_A \cap ab$ and $w_i'=N_A \pi_{Ai}^{-1} / \Sigma_{s_A} \pi_{Ai}^{-1}$ are the design weights based on the inclusion probabilities $\pi_{Ai}=Pr(i \in s_A)$. Analogous estimators based on s_B drawn from subpopulation B are denoted by $\hat{N}_b'' = \Sigma_{s_b} w_i''$, $\hat{N}_{ab}'' = \Sigma_{s''_{ab}} w_i''$, $\hat{Y}_b'' = \Sigma_{s_b} w_i'' y_i$, $\hat{Y}_{ab}'' = \Sigma_{s''_{ab}} w_i'' y_i$, $\hat{\mu}_b'' = \hat{Y}_b''/\hat{N}_b''$ and $\hat{\mu}_{ab}'' = \hat{Y}_{ab}'' / \hat{N}_{ab}''$, where $s_b = s_B \cap b$, $s_{ab}'' = s_B \cap ab$, $w_i'' = N_B \pi_{Bi}^{-1} \Sigma_{s_B} \pi_{Bi}^{-1}$ and $\pi_{Bi} = Pr(i \in s_B)$. A consequence of these definitions is that $\hat{N}_a'+\hat{N}_{ab}'=N_A$, $\hat{N}_b''+\hat{N}_{ab}''=N_B, 0 \le \hat{N}_{ab}' \le N_A$ and $0 \le \hat{N}_{ab}'' \le N_B$.

Using identity (1), Hartley (1974) proposed

$$\tilde{Y}_H = \hat{Y}_a' + \hat{Y}_b'' + \beta \hat{Y}_{ab}' + (1-\beta) \hat{Y}_{ab}''$$

as an estimator of Y, where β is chosen to minimize $\text{var}(\tilde{Y}_{FB})$. Typically the optimal value of β will be unknown and will need to be estimated from the sample data. As a consequence, the resulting estimator, $\hat{Y}_H$, will not be a simple linear combination of y values with weights the same for all y variables.

Fuller and Burmeister (1972) proposed

$$\tilde{Y}_{FB} = \hat{Y}_a' + \hat{Y}_b'' + \beta_1 \hat{Y}_{ab}' + (1-\beta_1)\hat{Y}_{ab}'' + \beta_2(\hat{N}_{ab}' - \hat{N}_{ab}'')$$

as an estimator of Y, where β_1 and β_2 are chosen to minimize $\mathrm{var}(\tilde{Y}_{FB})$. Again, on substitution of estimated values of β_1 and β_2, the resulting estimator, $\tilde{Y}_{FB}$, will not be a simple weighted combination of y values.

4. PSEUDO MAXIMUM LIKELIHOOD ESTIMATION

To motivate an alternative estimator that uses a single set of weights for all y variables, we first consider the special case when both s_A and s_B are selected by simple random sampling. Let n_A and n_B be the sizes of s_A and s_B respectively and again assume N_A and N_B are known. In this case, Fuller and Burmeister (1972) proposed

$$\hat{Y}_{srs} = (N_A - \hat{N}_{ab,srs})\,\hat{\mu}_{a,srs}' + (N_B - \hat{N}_{ab,srs})\,\hat{\mu}_{b,srs}'' + \hat{N}_{ab,srs}\,\hat{\mu}_{ab,srs}, \tag{2}$$

where $\hat{\mu}_{a,srs}' = \sum_{s_a} y_i / n_a$, $\hat{\mu}_{b,srs}'' = \sum_{s_b} y_i / n_b$, $\hat{\mu}_{ab,srs} = (n_{ab}' \hat{\mu}_{ab,srs}' + n_{ab}'' \hat{\mu}_{ab,srs}'') / (n_{ab}' + n_{ab}'')$ with $\hat{\mu}_{ab,srs}' = \sum_{s_{ab}'} y_i / n_{ab}'$, and $\hat{\mu}_{ab,srs}'' = \sum_{s_{ab}''} y_i / n_{ab}''$, and $n_a, n_b, n_{ab}', n_{ab}''$ are the sizes of s_a, s_b, s_{ab}' and s_{ab}'' respectively. Further, ignoring finite population corrections, $\hat{N}_{ab,srs}$ is the smallest root of the quadratic equation

$$(n_A + n_B)x^2 - (n_A N_B + n_B N_A + n_{ab}' N_A + n_{ab}'' N_B)x + (n_{ab}' + n_{ab}'')N_A N_B = 0. \tag{3}$$

Fuller and Burmeister (1972) show that $\hat{N}_{ab,srs}$ may be viewed as a maximum likelihood (ML) estimator of N_{ab}, where the score equation is given by (3). Moreover, Skinner (1991) provides an interpretation of $\hat{Y}_{srs}$ as an ML estimator of Y, provided the data on y are reduced to the totals of y over the sets s_a, s_b, s_{ab}' and s_{ab}''. Subject to this data reduction, $\hat{Y}_{srs}$ is therefore an asymptotically efficient estimator of Y under simple random sampling.

In the case of complex designs, $\hat{Y}_{srs}$ cannot, of course, be directly applied since it will in general be design-inconsistent for Y. Our idea is to modify $\hat{Y}_{srs}$ to achieve design-consistency under complex designs, while retaining the property of $\hat{Y}_{srs}$ that it is a linear weighted combination of the y_i. This idea is analogous to the notion of pseudo ML estimation in which a ML estimator under simple random sampling is modified to achieve consistent estimation under complex designs. The main advantages of a pseudo ML estimator are that it is design-consistent and typically has a simple form. Its main potential disadvantage is that it may not be asymptotically efficient although it may be hoped that any loss of efficiency will tend to be small in practice.

In order to obtain the pseudo ML estimator, first define $\hat{N}_{ab,srs}'$ and $\hat{N}_{ab,srs}''$ by $n_{ab}' = (n_A / N_A)\hat{N}_{ab,srs}'$ and $n_{ab}'' = (n_B / N_B)\hat{N}_{ab,srs}''$ respectively, where n_A and n_B are now arbitrary constants to be specified. Next replace each of the estimators $\hat{N}_{ab,srs}', \hat{N}_{ab,srs}'', \hat{\mu}_{ab,srs}'$, $\hat{\mu}_{ab,srs}'', \hat{\mu}_{a,srs}'$ and $\hat{\mu}_{b,srs}''$ in (2) by the corresponding estimators for complex designs given in Section 3. This gives our pseudo ML estimator of Y as

$$\hat{Y}_{PML} = (N_A - \hat{N}_{ab,PML})\hat{\mu}_a' + (N_B - \hat{N}_{ab,PML})\hat{\mu}_b'' + \hat{N}_{ab,PML}\hat{\mu}_{ab}, \tag{4}$$

where

$$\hat{\mu}_{ab} = \left[\frac{n_A}{N_A}\hat{N}_{ab}'\hat{\mu}_{ab}' + \frac{n_B}{N_B}\hat{N}_{ab}''\hat{\mu}_{ab}''\right] \Bigg/ \left[\frac{n_A}{N_A}\hat{N}_{ab}' + \frac{n_B}{N_B}\hat{N}_{ab}''\right].$$

Further, $\hat{N}_{ab,PML}$ is the smallest root of the quadratic equation

$$px^2 - qx + r = 0, \tag{5}$$

where $p = n_A + n_B, q = n_A N_B + n_B N_A + n_A \hat{N}_{ab}' + n_B \hat{N}_{ab}''$ and $r = n_A \hat{N}_{ab}' N_B + n_B \hat{N}_{ab}'' N_A$. Note that $q^2 - 4pr = (n_A N_B - n_B N_A - n_A \hat{N}_{ab}' + n_B \hat{N}_{ab}'')^2 + 4 n_A n_B (N_A - \hat{N}_{ab}')(N_B - \hat{N}_{ab}'')$, so the roots of (5) are always real, noting that $\hat{N}_{ab}' \le N_A$ and $\hat{N}_{ab}'' \le N_B$. Hence, $\hat{N}_{ab,PML}$ is well-defined.

5. ASYMPTOTIC PROPERTIES OF ESTIMATORS

We now present some asymptotic properties of the estimators $\hat{Y}_H, \hat{Y}_{FB}$ and $\hat{Y}_{PML}$. We refer the reader to Skinner and Rao (1996) for proofs and further details.

5.1 Asymptotic Variances

The estimators $\hat{Y}_H, \hat{Y}_{FB}$ and $\hat{Y}_{PML}$ are all design-consistent for Y. Further, if we choose optimal values β_1 and β_2 that minimize the asymptotic variance of $\hat{Y}_{FB}$, then we get

$$\text{avar}(\hat{Y}_{PML}) \geq \text{avar}(\hat{Y}_{FB,opt}),$$

where avar denotes asymptotic variance. Similarly, we have

$$\text{avar}(\hat{Y}_{H,opt}) \geq \text{avar}(\hat{Y}_{FB,opt}),$$

if one chooses optimal values β_1 and β_2 for $\hat{Y}_{FB}$ and optimal value β for $\hat{Y}_{HT}$. However, neither $\hat{Y}_{H(opt)}$ nor $\hat{Y}_{PML}$ is necessarily more efficient than the other, in general.

5.2 Choice of n_A and n_B

The PML estimator, $\hat{Y}_{PML}$, depends on n_A and n_B only via the ratio n_A/n_B. If n_A/n_B is chosen to minimize $\text{avar}(\hat{Y}_{PML})$, then the weights will depend on y, as in the case of $\hat{Y}_{FB,opt}$. To avoid this problem, we choose n_A/n_B to minimize $\text{avar}(\hat{N}_{ab,PML})$. Skinner and Rao (1996) have shown that this choice of n_A/n_B corresponds to taking (n_A, n_B) as $(\tilde{n}_A/d', \tilde{n}_B/d'')$, where $(\tilde{n}_A, \tilde{n}_B)$ denote the actual sizes of (s_A, s_B) and (d', d'') are the design effects of $\hat{N}'_{ab}$ and $\hat{N}''_{ab}$, *i.e.*, $d' = \text{avar}(\hat{N}'_{ab})/\text{avar}_{\text{srs}}(\hat{N}'_{ab})$ and $d'' = \text{avar}(\hat{N}''_{ab})/\text{avar}_{\text{srs}}(\hat{N}''_{ab})$, where avar_{srs} denotes asymptotic variance under simple random sampling. In practice, we replace d' and d'' by consistent estimators $\hat{d}'$ and $\hat{d}''$ or by values $\tilde{d}'$ and $\tilde{d}''$ available from past surveys. If $d' = d''$, then the simple choice $(n_A, n_B) = (\tilde{n}_A, \tilde{n}_B)$ gives the 'optimal ratio' n_A / n_B.

We denote the estimators of Y and N_{ab} based on the above 'optimal ratio' n_A / n_B as $\hat{Y}_{PML,opt}$ and $\hat{N}_{ab,PML,opt}$ respectively. It can be shown that the optimal Fuller-Burmeister estimator of N_{ab} has the same asymptotic variance as $\hat{N}_{ab,PML,opt}$.

If $\hat{d}'$ and $\hat{d}''$ are available, then we can also obtain an alternative simpler estimator of N_{ab} as

$$\tilde{N}_{ab,PML,opt} = \hat{\phi}\hat{N}'_{ab} + (1-\hat{\phi})\hat{N}''_{ab},$$

where $\hat{\phi} = n_A\hat{N}''_b/(n_A\hat{N}''_b + n_B\hat{N}'_a)$ with $n_A = \tilde{n}_A/\hat{d}'$ and $n_B = \tilde{n}_B/\hat{d}''$. This estimator is asymptotically equivalent to $\hat{N}_{ab,PML,opt}$. We denote the resulting estimator of Y as $\tilde{Y}_{PML,opt}$.

5.3 Efficiency of $\hat{Y}_{PML,opt}$

In Section 5.1 we noted that $\text{avar}(\hat{Y}_{PML}) \geq \text{avar}(\hat{Y}_{FB,opt})$ for any choice of (n_A, n_B) in $\hat{Y}_{PML}$. Hence, $\hat{Y}_{PML}$ never has smaller asymptotic variance than the optimal version of $\hat{Y}_{FB}$. We also know that under simple random sampling $\hat{Y}_{PML,opt}$ has an interpretation as a maximum likelihood estimator and hence may be expected to be asymptotically efficient. We now provide sufficient conditions for general designs under which $\hat{Y}_{PML,opt}$ has the same asymptotic variance as $\hat{Y}_{FB,opt}$. These conditions depend on the design effects of $\hat{\mu}'_{ab}$ and $\hat{\mu}''_{ab}$ which are defined in analogy with d' and d'' as $d'_\mu = \text{avar}(\hat{\mu}'_{ab})/\text{avar}_{\text{srs}}(\hat{\mu}'_{ab})$ and $d''_\mu = \text{avar}(\hat{\mu}''_{ab})/\text{avar}_{\text{srs}}(\hat{\mu}''_{ab})$.

We also define $\hat{\eta}_A = (\hat{\mu}'_a, \hat{\mu}'_{ab}, \hat{N}'_{ab}/N)^T$ and $\hat{\eta}_B = (\hat{\mu}''_b, \hat{\mu}''_{ab}, \hat{N}''_{ab}/N)^T$ with asymptotic covariance matrices Σ_A and Σ_B. The following theorem gives sufficient conditions for $\text{avar}(\hat{Y}_{PML,opt}) = \text{avar}(\hat{Y}_{FB,opt})$.

Theorem. If A1: Σ_A and Σ_B are both diagonal and A2: $d'_\mu/d''_\mu = d'/d''$, then $\text{avar}(\hat{Y}_{PML,opt}) = \text{avar}(\hat{Y}_{FB,opt})$.

Conditions A1 and A2 clearly hold under simple random sampling since $d' = d'' = d'_\mu = d''_\mu = 1$ and Σ_A and Σ_B are both diagonal. For general sampling designs either stratification or cluster sampling can lead to departures from conditions A1 and A2. If Σ_A and Σ_B do not differ greatly from diagonality and the design effects of different statistics are roughly proportional between frames, then the loss of efficiency of $\hat{Y}_{PML,opt}$ relative to $\hat{Y}_{FB,opt}$ will not be great. Conditions A1 and A2 suggest diagnostic checks based on the estimated covariance matrices $\hat{\Sigma}_A$ and $\hat{\Sigma}_B$ and the estimated design effects for judging the efficiency of $\hat{Y}_{PML,opt}$.

6. VARIANCE ESTIMATION

We now provide a consistent estimator of variance of $\hat{Y}_{PML}$ for specified choice of (n_A, n_B). Denote the usual estimators of variance of frame A estimator $\hat{Y}'_A = \Sigma_{s_A} w'_i y_i$ and frame B estimator $\hat{Y}''_B = \Sigma_{s_B} w''_i y_i$ as $v_A(y_i)$ and $v_B(y_i)$ respectively. Also, let $\theta = n_A N_B / (n_A N_B + n_B N_A)$, $\hat{\phi} = n_A\hat{N}''_b/(n_A\hat{N}''_b + n_B\hat{N}'_a)$ and $\hat{\lambda} = \hat{\mu}_{ab} - \hat{\mu}'_a - \hat{\mu}''_{ab}$. Further, denote $\hat{z}_{Ai} = y_i - \hat{\mu}'_a$ if $i \in s_a$ and $\hat{z}_{Ai} = \theta(y_i - \hat{\mu}_{ab}) + \hat{\lambda}\hat{\phi}$ if $i \in s'_{ab}$. Similarly, denote $\hat{z}_{Bi} = y_i - \hat{\mu}''_b$ if $i \in s_b$ and $\hat{z}_{Bi} = (1-\theta)(y_i - \hat{\mu}_{ab}) + \hat{\lambda}(1-\hat{\phi})$ if $i \in s''_{ab}$. A consistent estimator of variance of $\hat{Y}_{PML}$ is given by

$$\text{estvar}(\hat{Y}_{PML}) = v_A(\hat{z}_{Ai}) + v_B(\hat{z}_{Bi}). \tag{6}$$

Note that (6) can be computed from the single frame variance estimators, $v_A(y_i)$ and $v_B(y_i)$, by changing y_i to $\hat{z}_{Ai}$ in $v_A(y_i)$ and y_i to $\hat{z}_{Bi}$ in $v_B(y_i)$.

7. SINGLE FRAME ESTIMATION

All the estimators we have discussed so far might be termed two-stage estimators since first separate sets of estimators $(\hat{N}_a^{\prime}, \hat{N}_{ab}^{\prime}, \hat{\mu}_a^{\prime}, \hat{\mu}_{ab}^{\prime})$, and $(\hat{N}_b^{\prime\prime}, \hat{N}_{ab}^{\prime\prime}, \hat{\mu}_b^{\prime\prime}, \hat{\mu}_{ab}^{\prime\prime})$ are constructed from the two samples s_A and s_B and secondly these estimators are combined together to estimate Y. Such two-stage separation of the estimation process is not necessary, however, since the two sampling designs $p^{\prime}(s_A)$ and $p^{\prime\prime}(s_B)$ induce a well-defined design $p(s)$ on the set of samples $s=s_A \cup s_B$ in U. Thus, conventional estimators, which may be termed single-frame estimators may be constructed from $p(s)$. In particular, the Horvitz-Thompson estimator of Y may be used provided it is possible to determine the common units in samples s_A and s_B (Bankier, 1986). Kalton and Anderson (1986) proposed a simple estimator which does not require the identification of duplicate sample units. It is given by

$$\hat{Y}_S = \sum_{s_A} w_i y_i + \sum_{s_B} w_i y_i, \tag{7}$$

where $w_i = (\pi_{Ai} + \pi_{Bi})^{-1}$ and we define $\pi_{Ai}=0$ if $i \in b$ and $\pi_{Bi}=0$ if $i \in a$. The estimator (7) is approximately equal to the Horvitz-Thompson estimator when the π_{Ai} and π_{Bi} are small. It is unbiased with respect to the two-frame design.

An unbiased estimator of variance of $\hat{Y}_S$ with respect to the two-frame design is given by

$$\operatorname{estvar}(\hat{Y}_S) = v_A(\tilde{z}_{Ai}) + v_B(\tilde{z}_{Bi}). \tag{8}$$

with $\tilde{z}_{Ai} = \delta_{ai} y_i + (1-\delta_{ai}) y_i p_i$ and $\tilde{z}_{Bi} = \delta_{bi} y_i + (1-\delta_{bi}) y_i q_i$, where $p_i = \pi_{Ai}/(\pi_{Ai}+\pi_{Bi}), q_i = 1-p_i, \delta_{ai}=1$ if $i \in s_a^{\prime}, \delta_{ai}=0$ if $i \in s_{ab}^{\prime}, \delta_{bi}=1$ if $i \in s_b^{\prime\prime}$ and $\delta_{bi}=0$ if $i \in s_{ab}^{\prime\prime}$.

To implement the single-frame estimator $\hat{Y}_S$, it is necessary to determine the unit's inclusion probabilities both from the frame from which it is sampled and from the other frame. For surveys involving complex designs, such as stratified multistage sample from at least one of the frames, this may not be feasible in practice. This is a major practical limitation of single frame estimators.

When both sampling designs are self-weighting, *i.e.*, $\pi_{Ai} = \tilde{n}_A/N_A$ and $\pi_{Bi} = \tilde{n}_B/N_B$, then $\hat{Y}_S$ is a special case of $\hat{Y}_H$ with $\beta = \tilde{\theta}$, where $\tilde{\theta} = \tilde{n}_A N_B / (\tilde{n}_A N_B + \tilde{n}_B N_A)$. For general designs, it seems difficult to identify conditions when $\hat{Y}_S$ will be more (or less) efficient than the two-stage estimators. Theoretically, $\hat{Y}_S$ may be expected to perform relatively well when the overlap domain *ab* contains units for which just one of π_{Ai} or π_{Bi} is relatively very small, especially when the associated y_i values are extreme. This follows by noting that the weight min $(\pi_{Ai}^{-1}, \pi_{Bi}^{-1})$ will tend to make a larger contribution to the variance of the two-stage estimator than the weight $(\pi_{Ai}+\pi_{Bi})^{-1}$ to the variance of $\hat{Y}_S$. This point is illustrated in the following artificial example.

Example. Suppose both frames are complete with $N_{ab}=N_A=N_B=N=2000$ and $N_a=N_b=0$. The population is partitioned into two equal size strata with $i = 1, \dots, 1000$ in stratum 1 and $i = 1001, \dots, 2000$ in stratum 2. Suppose stratified sampling is employed in both frames with strata allocations (100, 1) and (1, 100) in s_A and s_B respectively so that $\pi_{Ai}=0.1$ for $i = 1, \dots, 1000$; $\pi_{Ai} = 0.001$ for $i = 1001, \dots, 2000$ and $\pi_{Bi} = 0.001$ for $i = 1, \dots, 1000$; $\pi_{Bi} = 0.1$ for $i = 1001, \dots, 2000$.

If the strata variances are equal, say S^2, then $\hat{Y}_{H,opt}, \hat{Y}_{FB,opt}, \hat{Y}_{PML}$ and $\hat{Y}_{PML,opt}$ all reduce by symmetry to $\hat{Y} = \frac{1}{2}(\Sigma_{s_A} y_i/\pi_{Ai} + \Sigma_{s_B} y_i/\pi_{Bi})$ with $\operatorname{var}(\hat{Y}) = 504{,}000 S^2$. In comparison, $\hat{Y}_S = (0.101)^{-1} (\Sigma_{s_A} y_i + \Sigma_{s_B} y_i)$ with $\operatorname{var}(\hat{Y}_S) = 17{,}841\ S^2$ so that $\hat{Y}_S$ achieves a 96% reduction in variance over the optimal two-stage estimators.

One disadvantage of the simple estimator $\hat{Y}_S$ is that its implied estimators of N_A and N_B are not equal to the known values N_A and N_B, unlike the implied estimators of $\hat{Y}_S$. One method of adjusting $\hat{Y}_S$ to achieve consistency with known N_A and N_B is to use raking ratio estimation (Bankier, 1986). Following the proof of Theorem 1 of Skinner (1991), it can be shown that the estimator of Y from the raking ratio procedure converges to

$$\hat{Y}_{RR} = (N_A - \hat{N}_{ab}^{RR})\hat{\mu}_a^{\prime} + (N_B - \hat{N}_{ab}^{RR})\hat{\mu}_b^{\prime\prime} + \hat{N}_{ab}^{RR}\hat{\mu}_{abS} \tag{9}$$

where $\hat{\mu}_{abS} = \hat{Y}_{abS}/\hat{N}_{abS}$

with $\hat{Y}_{abS} = \sum_{s_{ab}^{\prime}} w_i y_i + \sum_{s_{ab}^{\prime\prime}} w_i y_i, \hat{N}_{abS} = \sum_{s_{ab}^{\prime}} w_i + \sum_{s_{ab}^{\prime\prime}} w_i$.

Further, $\hat{N}_{ab}^{RR}$ is the smallest root of the quadratic equation

$$\hat{N}_{abS}x^2-[\hat{N}_{abS}(N_A+N_B)+\hat{N}_{aS}\hat{N}_{bS}]x+\hat{N}_{abS}N_AN_B=0,$$

where $\hat{N}_{aS}=\Sigma_{s_a}w_i$ and $\hat{N}_{bS}=\Sigma_{s_b}w_i$. Note that it is not necessary to perform actual raking, as in Bankier (1986), because $\hat{Y}_{RR}$ is expressed in a closed form. Also, $\hat{Y}_{RR}$ is a linear combination of the y_i so that it uses the same weights for all the variables, as in the case of $\hat{Y}_S$ and $\hat{Y}_{PML}$.

In the case of simple random sampling and equal sampling fractions, $n_A/N_A=n_B/N_B$, Skinner (1991) showed that $\hat{Y}_{RR}$ is considerably more efficient than $\hat{Y}_S$ and that the maximum possible gain in efficiency of $\hat{Y}_{FB,opt}$ relative to $\hat{Y}_{RR}$ is 3.7%. The efficiency of $\hat{Y}_{RR}$ for general complex designs remains to be explored.

8. SIMULATION RESULTS

Skinner and Rao (1996) conducted a limited simulation study on the relative efficiencies of PML estimators $\hat{Y}_{PML,opt}$ and $\tilde{Y}_{PML,opt}$, Hartley's estimator $\hat{Y}_H$, Fuller-Burmeister's estimator $\hat{Y}_{FB}$ and the single frame estimator $\hat{Y}_S$. Optimal values of β,β_1,β_2 and n_A/n_B estimated from the data were used in computing the estimators. Two-stage sampling with n sample clusters and m sample elements from each sample cluster ($\tilde{n}_A=nm$) was used in frame A and simple random sampling of $\tilde{n}_B$ elements in frame B. Repeated data sets from frames A and B were generated using models for specified parameter combinations and empirical mean squared errors (EMSE) of the estimators were calculated.

The dual frame estimators $\hat{Y}_{PML}$, $\tilde{Y}_{PML}$, $\hat{Y}_H$ and $\hat{Y}_{FB}$ performed similarly in terms of MSE. On the other hand, the single frame estimator $\hat{Y}_S$ displayed considerable increase in EMSE compared to the dual frame estimators when $N_a/N \leq N_b / N$ or $\tilde{n}_A$ is much larger than $\tilde{n}_B$. For example when $N_a/N=0.1, N_b/N=0.2$ and $\tilde{n}_A/\tilde{n}_B=2$, we have EMSE $(\hat{Y}_S)=5.66$ compared to EMSE $(\hat{Y}_{PML})=2.32$.

REFERENCES

Bankier, M.D. (1986). "Estimators based on several stratified samples with applications to multiple frame surveys," *Journal of the American Statistical Association*, 81, 1074-1079.

Fuller, W.A., and Burmeister, L.F. (1972). "Estimators for samples selected from two overlapping frames," in *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

Hartley, H.O. (1962). "Multiple frame surveys," in *Proceedings of the Social Statistics Section, American Statistical Association*, 203-206.

Hartley, H.O. (1974). "Multiple frame methodology and selected applications," *Sankhyā, Series* C, 36, 99-118.

Isaki, C.T., and Fuller, W.A. (1982). "Survey design under the regression superpopulation model." *Journal of the American Statistical Association*, 77, 89-96.

Kalton, G., and Anderson, D.W. (1986). "Sampling rare populations," *Journal of the Royal Statistical Society*, Series A, 149, 65-82.

Lepkowski, J.M., and Groves, R.M. (1986). "A mean squared error model for dual frame mixed model survey design," *Journal of the American Statistical Association*, 81, 930-937.

Lund, R.E. (1968). "Estimators in multiple frame surveys," in *Proceedings of the Social Statistics Section, American Statistical Association*, 282-288.

Skinner, C.J., (1991). "On the efficiency of raking ratio estimation for multiple frame surveys," *Journal of the American Statistical Association*, 86, 779-784.

Skinner, C.J., and Rao, J.N.K., (1996). "Estimation in dual frame surveys with complex designs," *Journal of the American Statistical Association*, 91, 349-356.