Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

Replication Variance Estimation for Two-Phase Stratified Sampling

Jae Kwang Kim, Alfredo Navarro & Wayne A Fuller

To cite this article: Jae Kwang Kim, Alfredo Navarro & Wayne A Fuller (2006) Replication Variance Estimation for Two-Phase Stratified Sampling, Journal of the American Statistical Association, 101:473, 312-320, DOI: 10.1198/016214505000000763

To link to this article: http://dx.doi.org/10.1198/016214505000000763

Replication Variance Estimation for Two-Phase Stratified Sampling

Jae Kwang KIM, Alfredo NAVARRO, and Wayne A. FULLER

In two-phase sampling, the second-phase sample is often a stratified sample based on the information observed in the first-phase sample. For the total of a population characteristic, either the double-expansion estimator or the reweighted expansion estimator can be used. Given a consistent first-phase replication variance estimator, we propose a consistent variance estimator that is applicable to both the double-expansion estimator and the reweighted expansion estimator. The proposed method can be extended to multiphase sampling.

KEY WORDS: Double-expansion estimator; Double sampling; Multiphase sampling; Reweighted expansion estimator.

1. INTRODUCTION

Two-phase sampling, also known as double sampling, can be a cost-effective technique in large-scale surveys. By selecting a large sample, observing cheap auxiliary variables, and properly incorporating the auxiliary variables into the second-phase sampling design, we can produce estimators with smaller variances than those based on a single-phase sampling design for the same cost. In one of the common procedures of two-phase sampling, the second-phase sample is selected using stratified sampling, where the strata are created on the basis of the first-phase observations.

Rao (1973) and Cochran (1977) gave formulas for variance estimation when the first phase is a simple random sample and the second phase is a stratified simple random sample. Kott (1990) derived a formula for variance estimation when the first phase is a stratified random sample and the second phase is a restratified simple random sample based on first-phase information. Rao and Shao (1992) proposed a jackknife variance estimation method in the context of hot-deck imputation where the response corresponds to a second phase with Poisson sampling in imputation cells. Yung and Rao (2000) extended the result of Rao and Shao to poststratification. Binder (1996) illustrated a "cookbook" approach for the two-phase ratio estimator. Binder, Babyak, Brodeur, Hidiroglou, and Jocelyn (2000) derived formulas for variance estimation for various estimators for two-phase restratified sampling. Fuller (1998) proposed a replicate variance estimation method for the two-phase regression estimator.

Among the methods cited, only the methods of Rao and Shao (1992) and Fuller (1998) are replication methods. One advantage of the replication method for variance estimation is its convenience for a multipurpose survey. That is, after we create the replication weights, we can directly apply the replication weights to estimate the variance for any variable.

Let the finite population be of size N, indexed from 1 to N, and let the finite population be partitioned into G groups, which we call the second-phase strata. The information about which group a unit belongs to is not obtained until the first-phase sample has been observed.

We consider the two-phase estimator in which the first-phase sample is used to define strata to be used for the second-phase sample. Let the parameter of interest be the population total $Y = \sum_{i=1}^{N} y_i$, where y_i is the study variable and N is assumed known. Suppose that we have a first-phase sample of size n. If we observe y_i on every element of the sample, then an unbiased estimator of Y is

$$\hat{Y}_1 = \sum_{i \in A_1} w_i y_i, \tag{1}$$

where $w_i = [\Pr(i \in A_1)]^{-1}$ and A_1 is the set of indices in the sample. Now, assume that instead of directly observing y_i for $i \in A_1$, we observe

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{iG}) \tag{2}$$

for all $i \in A_1$, where x_{ig} takes the value 1 if unit i belongs to the gth group and 0 otherwise. Assume that $\sum_{g=1}^{G} x_{ig} = 1$.

Let a subsample of total size r be selected from the first-phase sample and let A_2 be the set of indices for the second-phase sample. Let

$$w_i^* = \left[\Pr(i \in A_2 \mid i \in A_1)\right]^{-1}. \tag{3}$$

Let $n_{1g} = \sum_{i \in A_1} x_{ig}$ be the number of first-phase sample elements in group g and let $r_g = \sum_{i \in A_2} x_{ig}$ be the number of second-phase sample elements in group g. If the second-phase sample is selected by stratified simple random sampling with the groups as strata, then $w_i^* = r_g^{-1} n_{1g}$ for unit i with $x_{ig} = 1$.

Given the described two-phase sample, an unbiased estimator for the total of Y is

$$\hat{Y}_d = \sum_{i \in A_2} \alpha_{d,i} y_i, \tag{4}$$

where $\alpha_{d,i} = w_i w_i^*$. Kott and Stukel (1997) called the estimator in (4) the *double-expansion estimator* (DEE).

Another important estimator for the total of Y is

$$\hat{Y}_r = \sum_{g=1}^{G} \left(\sum_{i \in A_1} w_i x_{ig} \right) \frac{\sum_{i \in A_2} w_i x_{ig} y_i}{\sum_{i \in A_2} w_i x_{ig}} = \sum_{i \in A_2} \alpha_{r,i} y_i, \tag{5}$$

Jae Kwang Kim is Assistant Professor, Department of Applied Statistics, Yonsei University, Seoul 120-749, Korea (E-mail: *kimj@yonsei.ac.kr*). Alfredo Navarro is the ACS Branch Chief, Decennial Statistical Studies Division, Bureau of the Census, Washington, DC 20233 (E-mail: *alfredo.navarro@census.gov*). Wayne A. Fuller is Distinguished Professor Emeritus, Department of Statistics, Iowa State University, Ames, IA 50011 (E-mail: *waf@iastate.edu*). This research was supported in part by cooperative agreement 13-3AEU-0-80064 between Iowa State University, the U.S. National Agricultural Statistics Service, and the U.S. Bureau of the Census. Much of the research was conducted while the first author was a mathematical statistician at the U.S. Bureau of the Census. The authors thank the referees for comments and suggestions that improved the manuscript.

Journal of the American Statistical Association
March 2006, Vol. 101, No. 473, Theory and Methods
DOI 10.1198/016214505000000763

where

$$\alpha_{r,i} = \sum_{g=1}^{G} \left(\frac{\sum_{j \in A_1} w_j x_{jg}}{\sum_{j \in A_2} w_j x_{jg}} \right) w_i x_{ig}.$$

Kott and Stukel (1997) called the estimator in (5) the *reweighted expansion estimator* (REE).

Kott and Stukel (1997) examined possible replication methods for estimating the variance of the DEE and concluded that the jackknife methods that they considered cannot be used for this purpose. For the REE, the replication method proposed by Rao and Shao (1992) produces consistent variance estimates.

In the next section we discuss the asymptotic properties of the DEE and the REE. In Section 3 we give a replicate method for estimating the variance of the DEE and the REE. In Section 4 we extend the replication method to regression estimators and to multiphase stratified sampling. The variance estimation procedure was applied to the 2000 Census and Accuracy and Coverage Evaluation survey by Kim, Navarro, and Fuller (2000).

2. ASYMPTOTIC PROPERTIES

To derive the asymptotic properties of the estimators, we assume a sequence of samples and finite populations such as that described by Fuller (1975). Let $\{\zeta_n\}_{n=1}^{\infty}$ be a sequence of populations, each having $G_n \geq G_{n-1}$ groups of size X_{ng}, where the groups can cut across the first-phase strata. Associated with the ith element in the population is a vector, $\mathbf{x}_{ni} = (x_{ni1}, x_{ni2}, \ldots, x_{niG_n})$, of group indicators of dimension G_n, and the study variable y_{ni}. Let a sample of size n be selected from the nth population and assume that the population size N_n increases as n increases such that the limit of $N_n^{-1} n$ is a finite fraction, perhaps 0. Let $\mathcal{F}_n = \{(\mathbf{x}_{n1}, y_{n1}), (\mathbf{x}_{n2}, y_{n2}), \ldots, (\mathbf{x}_{nN_n}, y_{nN_n})\}$, $Y_{ng} = \sum_{i=1}^{N_n} x_{nig} y_{ni}$, and $Y_n = \sum_{i=1}^{N_n} y_{ni}$. Then $Y_n = \sum_{g=1}^{G_n} Y_{ng}$, because a unit belongs to one and only one group. Assume that the sequence of finite populations satisfies

$$N_n^{-1} \sum_{i=1}^{N_n} y_{ni}^{2+\tau} = O(1) \tag{6}$$

for some $\tau > 0$.

Because a general class of first-phase sampling designs is permitted, we directly specify the design properties of the estimators. Let A_{n1} be the set of indices for the first-phase sample selected by the first-phase sampling design from the nth finite population ζ_n. Let w_{ni} be the sampling weight of unit ni. Define

$$(\hat{X}_{ng1}, \hat{Y}_{ng1}) = \sum_{i \in A_{n1}} w_{ni} (x_{nig}, x_{nig} y_{ni}) \tag{7}$$

and

$$\bar{y}_{ng1} = \hat{X}_{ng1}^{-1} \hat{Y}_{ng1}, \tag{8}$$

where the subscript "1" emphasizes that the estimators are based on the first-phase sample. The estimator $\hat{X}_{ng1}$ is the estimated number of elements in group g, where the population number is X_{ng}. The estimator $\hat{Y}_{ng1}$ of the total of y for group g is not observed in a two-phase sample.

Assume that

$$E\{(\hat{X}_{ng1}, \hat{Y}_{ng1})' \mid \mathcal{F}_n\} = (X_{ng}, Y_{ng})' \tag{9}$$

and

$$\operatorname{var}\left\{ N_n^{-1} \sum_{g=1}^{G_n} (\hat{X}_{ng1}, \hat{Y}_{ng1})' \,\middle|\, \mathcal{F}_n \right\} = O(n^{-1}), \tag{10}$$

where the notation var{·} denotes the variance–covariance matrix when the argument is a vector variable.

Assume that a set of fixed probabilities π_{ng}, $g = 1, 2, \ldots, G_n$, is used to select a second-phase stratified random sample. Thus r_{ng} elements are selected from the n_{1ng} first-phase elements in group g without replacement with equal probability, where r_{ng} is the integer closest to $\pi_{ng} n_{1ng}$. We ignore this rounding error in the subsequent discussion. Let A_{n2} be the set of indices for the second-phase sample. Define the second-phase sample estimators

$$(\hat{X}_{ng2}, \hat{Y}_{ng2}) = \sum_{i \in A_{n2}} w_{ni} r_{ng}^{-1} n_{1ng} (x_{nig}, x_{nig} y_{ni})$$

and

$$\bar{y}_{ng2} = \begin{cases} \hat{X}_{ng2}^{-1} \hat{Y}_{ng2} & \text{if } r_{ng} > 0 \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where the subscript "2" emphasizes that the estimators are based on the second-phase sample.

The REE in (5) can be written as

$$\hat{Y}_{nr} = \sum_{g=1}^{G_n} \hat{X}_{ng1} \bar{y}_{ng2}, \tag{12}$$

and the DEE in (4) can be written as

$$\hat{Y}_{nd} = \sum_{g=1}^{G_n} \hat{Y}_{ng2}. \tag{13}$$

To formally define an estimator with finite moments, we assume that $r_{ng} \geq 1$ when $n_{1ng} \geq 1$.

The following theorem gives some asymptotic properties of the REE and DEE for a sequence of populations and samples in which the number of second-phase strata is permitted to increase. For fixed G_n, the variance formulas to appear in (21) and (23) correspond to (9.7.27) and (9.4.7) of Särndal, Swensson, and Wretman (1992).

Theorem 1. Let the sequence of finite populations and samples be as described earlier. Assume simple random sampling in each group at the second phase. Assume (6), (9), and (10). Let $\hat{Y}_{nr}$ be the REE defined in (12) and let $\hat{Y}_{nd}$ be the DEE defined in (13). Assume that

$$C_{xL} G_n^{-1} < N_n^{-1} X_{ng} < C_{xU} G_n^{-1} \quad \text{for all } n, \tag{14}$$

$$G_n < C_G n^{\lambda} \quad \text{for all } n, \tag{15}$$

$$C_{wL} \leq N_n^{-1} n w_{ni} \leq C_{wU} \quad \text{for all } n, \tag{16}$$

and

$$C_{\pi} < \pi_{ng} \leq 1 \quad \text{for all } g \text{ and all } n, \tag{17}$$

where $C_{xL}, C_{xU}, C_G, C_{wL}, C_{wU}$, and C_π are fixed positive constants, and that $0 \le \lambda < .5$. Also assume that

$$\operatorname{var}\{\hat{Y}_{n1} \mid \mathcal{F}_n\} < K_M \operatorname{var}\{\hat{Y}_{\mathrm{SRS},n1} \mid \mathcal{F}_n\}, \tag{18}$$

for a fixed K_M and for any y satisfying (6), where $\hat{Y}_{\mathrm{SRS},n1}$ is the estimator of Y_n based on a simple random sample of size n. Then the REE satisfies

$$E(\hat{Y}_{nr} \mid \mathcal{F}_n) = Y_n + o\big(n^{-1/2}N_n\big), \tag{19}$$

and the DEE is unbiased,

$$E(\hat{Y}_{nd} \mid \mathcal{F}_n) = Y_n. \tag{20}$$

The variance of the REE is

$$\begin{aligned}\operatorname{var}(\hat{Y}_{nr} \mid \mathcal{F}_n) &= \operatorname{var}(\hat{Y}_{n1} \mid \mathcal{F}_n) \\ &\quad + E\left\{\sum_{g=1}^{G_n} n_{1ng}^2\left(\frac{1}{r_{ng}} - \frac{1}{n_{1ng}}\right)\sigma_{nweg1}^2 \,\middle|\, \mathcal{F}_n\right\} \\ &\quad + o(n^{-1}N_n^2),\end{aligned} \tag{21}$$

where

$$\sigma_{nweg1}^2 = \begin{cases} (n_{1ng}-1)^{-1}\displaystyle\sum_{i\in A_{n1}} x_{nig}w_{ni}^2e_{nig}^2 & \text{if } n_{1ng} > 1 \\ 0 & \text{if } n_{1ng} \le 1, \end{cases} \tag{22}$$

$e_{nig} = y_{ni} - \bar{Y}_{ng}$, and $\bar{Y}_{ng} = X_{ng}^{-1}Y_{ng}$ is the population mean of y_{ni}'s in group g.

The variance of the DEE is

$$\begin{aligned}&\operatorname{var}(\hat{Y}_{nd} \mid \mathcal{F}_n) \\ &\quad = \operatorname{var}(\hat{Y}_{n1} \mid \mathcal{F}_n) \\ &\qquad + E\left\{\sum_{g=1}^{G_n} n_{1ng}^2\left(\frac{1}{r_{ng}} - \frac{1}{n_{1ng}}\right)\sigma_{nwyg1}^2 \,\middle|\, \mathcal{F}_n\right\},\end{aligned} \tag{23}$$

where

$$\sigma_{nwyg1}^2 = \begin{cases} (n_{1ng}-1)^{-1}\displaystyle\sum_{i\in A_{n1}}\left(x_{nig}w_{ni}y_{ni} - n_{1ng}^{-1}\sum_{j\in A_{n1}} x_{njg}w_{nj}y_{nj}\right)^2 \\ \qquad \text{if } n_{1ng} > 1 \\ 0 \quad \text{if } n_{1ng} \le 1. \end{cases}$$

For the proof see Appendix A.

Conditional on the first-phase sample, the DEE for a stratified second-phase sample is a Horvitz–Thompson-type estimator of the first-phase sample total of w_iy_i with weight $r_g^{-1}n_{1g}$, and thus is conditionally unbiased for $\hat{Y}_1$, conditional on A_{n1}. Conditional on the first-phase sample, the REE is a separate ratio estimator and is subject to ratio bias. By assumptions (14)–(17), the ratio bias is negligible in large samples.

The variance formulas (21) and (23) show that the variances of REE and DEE are no smaller than the variance of $\hat{Y}_1$. However, a two-phase sample may be cheaper, because of the fewer observations on the y variable. The variance formulas (21) and (23) also give some direction for stratification for the second-phase sampling. The variance of REE is minimized when the y_i's are the same within each group, whereas the variance of DEE is minimized when the weighted observations w_iy_i are the same within each group. Thus the REE will be more efficient than the DEE if the observations are relatively homogeneous within each group.

When the first-phase sampling weights are the same, as considered by Rao (1973) and Cochran (1977), then REE is equal to DEE and the ratio bias of REE is 0. Among the authors who have studied two-phase stratified sampling with unequal first-phase sampling weights, Särndal et al. (1992) focused on DEE-type estimation, whereas Kott and Stukel (1997) and Binder et al. (2000) focused on variance estimation for the REE.

3. REPLICATION VARIANCE ESTIMATION

We consider a replication method comprising the number of replicates, the replication factors, and the replication weights. Let the replicate variance estimator for the first-phase sample estimator $\hat{Y}_{n1}$ of (1) be written in the form

$$\hat{V}_{n1} = \sum_{k=1}^{L_n} c_{nk}\big(\hat{Y}_{n1}^{(k)} - \hat{Y}_{n1}\big)^2, \tag{24}$$

where $\hat{Y}_{n1}^{(k)}$ is the kth version of $\hat{Y}_{n1}$ based on the observations included in the kth replicate, L_n is the number of replications, and c_{nk} is a factor associated with replicate k determined by the replication method. The kth replicate for the complete first-phase sample estimator $\hat{Y}_{n1}$ be can be written in the form

$$\hat{Y}_{n1}^{(k)} = \sum_{i\in A_{n1}} w_{ni}^{(k)}y_{ni}, \tag{25}$$

where $w_{ni}^{(k)}$ denotes the replicate weight for the ith unit of the kth replication. Constructing the replicate variance estimator in (24) is possible if the sampling design is measurable (see Fay 1989). For example, consider a stratified random sample with $w_{hi} = n_h^{-1}N_h$ for unit i in stratum h. Then, the full-sample jackknife variance estimator is defined by the number of replicates $L_n = n$, the replication factor $c_{n,hk} = (1 - N_h^{-1}n_h)n_h^{-1}(n_h - 1)$ for the kth element in stratum h, and the replication weights

$$w_{hi}^{(sk)} = \begin{cases} 0 & \text{if } s = h \text{ and } k = i \\ (n_h-1)^{-1}N_h & \text{if } s = h \text{ and } k \ne i \\ n_h^{-1}N_h & \text{if } s \ne h. \end{cases}$$

Other commonly used replication methods, such as balanced half samples and the bootstrap, can also be written in the form (24).

Rao and Shao (1992) proposed an adjusted jackknife method for variance estimation in the context of hot-deck imputation. The imputation cell used in imputation corresponds to the second-phase stratum. The Rao–Shao jackknife replicate for the REE of (5) is

$$\hat{Y}_{nr}^{(k)} = \sum_{g=1}^{G_n}\left(\sum_{i\in A_{n1}} w_{ni}^{(k)}x_{nig}\right)\frac{\sum_{i\in A_{n2}} w_{ni}^{(k)}x_{nig}y_{ni}}{\sum_{i\in A_{n2}} w_{ni}^{(k)}x_{nig}}, \tag{26}$$

where the $w_{ni}^{(k)}$'s are the full-sample replicate weights of (25). The replicate variance estimator can be written as

$$\hat{V}_{nr} = \sum_{k=1}^{L_n} c_{nk}\big(\hat{Y}_{nr}^{(k)} - \hat{Y}_{nr}\big)^2, \tag{27}$$

where the c_{nk} are determined by L_n and the design. A complete set of jackknife replicates has a number of replicates equal to the number of first-phase primary sampling units. The replicate weights are applied to the second-phase units. Kott and Stukel (1997) suggested using the adjusted jackknife method defined by (26) and (27) to estimate the variance of the REE.

To estimate the variance of the DEE, we propose the replicate variance estimator

$$\hat{V}_{nd} = \sum_{k=1}^{L_n} c_{nk}\left(\hat{Y}_{nd}^{(k)} - \hat{Y}_{nd}\right)^2 \tag{28}$$

with replicates

$$\hat{Y}_{nd}^{(k)} = \sum_{g=1}^{G_n}\left(\sum_{i\in A_{n1}} w_{ni}^{(k)} w_{ni}^{-1} x_{nig}\right)\frac{\sum_{i\in A_{n2}} w_{ni}^{(k)} x_{nig} y_{ni}}{\sum_{i\in A_{n2}} w_{ni}^{(k)} w_{ni}^{-1} x_{nig}}. \tag{29}$$

The replicates (29) are motivated by the fact that DEE can be written as a special case of REE with weights equal to 1 and variables equal to $w_i y_i$.

We assume that the variance of a linear estimator of a total is a quadratic function of y and assume that

$$nN_n^{-2}\operatorname{var}\left(\sum_{i\in A_{n1}} w_{ni} y_{ni} \mid \mathcal{F}_n\right) = \sum_{i=1}^{N_n}\sum_{j=1}^{N_n} \Omega_{nij} y_{ni} y_{nj}, \tag{30}$$

where the coefficients Ω_{nij} satisfy

$$\sum_{i=1}^{N_n} |\Omega_{nij}| = O(N_n^{-1}). \tag{31}$$

Under simple random sampling, condition (31) is satisfied because

$$\Omega_{nij} = \begin{cases} N_n^{-1}(1 - N_n^{-1}n) & \text{if } i = j \\ -N_n^{-1}(N_n - 1)^{-1}(1 - N_n^{-1}n) & \text{if } i \neq j. \end{cases}$$

We establish the consistency of $\hat{V}_{nr}$ defined in (26) and (27) and the consistency of $\hat{V}_{nd}$ defined in (28) and (29) in the following theorem.

Theorem 2. Let the assumptions of Theorem 1 hold with the exceptions that (6) holds for $\tau \geq 2$ and that $0 \leq \lambda < 3^{-1}$ in (15). Assume (30) and (31). Let the first-phase sample be without replacement and the replication variance estimator for the complete sample be of the form (24). Assume that for any complete-sample estimator of a total, $\hat{\gamma}_n$, for a variable with fourth moments, the replicates satisfy

$$E\left\{\left[c_{nk}\left(\hat{\gamma}_n^{(k)} - \hat{\gamma}_n\right)^2\right]^2 \mid \mathcal{F}_n\right\} < K_\gamma L_n^{-2}\left[\operatorname{var}(\hat{\gamma}_n \mid \mathcal{F}_n)\right]^2 \tag{32}$$

for some constant K_γ, uniformly in n. Also assume that

$$c_{nk}^{-1} = O(L_n). \tag{33}$$

Let $\hat{V}(\hat{\theta}_n)$ be the first-phase sample replicate estimator of the variance of $\hat{\theta}_n = \sum_{i\in A_{n1}} w_{ni} y_{ni}$, and assume that

$$E\left\{\left[\frac{\hat{V}(\hat{\theta}_n)}{\operatorname{Var}(\hat{\theta}_n \mid \mathcal{F}_n)} - 1\right]^2 \mid \mathcal{F}_n\right\} = o(1) \tag{34}$$

for any y with bounded fourth moments. Then, the variance estimator defined in (27) with replicates (26) satisfies

$$\hat{V}_{nr} = \operatorname{var}(\hat{Y}_{nr} \mid \mathcal{F}_n) - \sum_{g=1}^{G_n} \pi_{ng}^{-1}(1 - \pi_{ng}) \sum_{i=1}^{N_n} x_{nig} e_{nig}^2 + o_p(n^{-1}N_n^2), \tag{35}$$

where $e_{nig} = y_{ni} - \bar{Y}_{ng}$. Also, the variance estimator defined in (28) with replicates (29) satisfies

$$\hat{V}_{nd} = \operatorname{var}(\hat{Y}_{nd} \mid \mathcal{F}_n) - \sum_{g=1}^{G_n} \pi_{ng}^{-1}(1 - \pi_{ng}) \sum_{i=1}^{N_n} x_{nig} \eta_{nig}^2 + o_p(n^{-1}N_n^2), \tag{36}$$

where $\eta_{nig} = y_{ni} - w_{ni}^{-1}(\sum_{i=1}^{N_n} x_{nig} w_{ni}^{-1})^{-1} \sum_{i=1}^{N_n} x_{nig} y_{ni}$.

For the proof see Appendix B.

By condition (32), no one of the squared deviates in the replication variance estimator dominates the others. Condition (34) states that the first-phase sample replication variance estimator is consistent. Conditions (32), (33), and (34) can be satisfied by many replication methods, including the jackknife, balanced half samples, and the bootstrap. Consistency results for the first-phase sample replication variance estimators have been discussed by, for example, Krewski and Rao (1981).

For fixed π_{ng}, the second term on the right side of equality (35) is small relative to the first term if all first-phase probabilities are small. If some first-phase probabilities are large, then an estimator of

$$\sum_{g=1}^{G_n} \pi_{ng}^{-1}(1 - \pi_{ng}) \sum_{i=1}^{N_n} x_{nig} e_{nig}^2$$

can be added to (27). An estimator is

$$\sum_{g=1}^{G_n} \pi_{ng}^{-2}(1 - \pi_{ng})(r_{ng} - 1)^{-1} r_{ng} \sum_{i\in A_{n2}} w_{ni} x_{nig} \hat{e}_{nig}^2, \tag{37}$$

where $\hat{e}_{nig} = y_{ni} - \bar{y}_{ng2}$ and $\bar{y}_{ng2}$ is defined in (11). For DEE variance estimation, the corresponding estimator is

$$\sum_{g=1}^{G_n} \pi_{ng}^{-2}(1 - \pi_{ng})(r_{ng} - 1)^{-1} r_{ng} \sum_{i\in A_{n2}} w_{ni} x_{nig} \hat{\eta}_{nig}^2, \tag{38}$$

where $\hat{\eta}_{nig} = y_{ni} - w_{ni}^{-1}(r_{ng}^{-1} \sum_{i\in A_{n2}} w_{ni} x_{nig} y_{ni})$. A replication version of the estimator (37) with r_n replicates is

$$\sum_{g=1}^{G_n} \sum_{i\in A_{n2}} c_{ngi}\left(\hat{Y}_{nr}^{(gi)} - \hat{Y}_{nr}\right)^2, \tag{39}$$

where $r_n = \sum_{g=1}^{G_n} r_{ng}$, $c_{ngi} = (r_{ng} - 1)^{-1} r_{ng} \pi_{ng}^{-2}(1 - \pi_{ng}) w_{ni} x_{nig}$ and $\hat{Y}_{nr}^{(gi)} = \hat{Y}_{nr} - (y_{ni} - \bar{y}_{ng2})$ for $x_{nig} = 1$. Expression (39) is algebraically equivalent to (37).

Remark 1. In Theorems 1 and 2, we assume stratified simple random sampling for the second phase. The proof of Theorem 2 rests on a proof for Poisson sampling. Hence the results also hold for Poisson sampling at the second phase.

Remark 2. Theorem 2 is stated and proven for without replacement sampling at the first phase. Under mild conditions, it can be shown that with first-phase replacement sampling,

$$\hat{V}_{nt} = \text{var}(\hat{Y}_{nt} \mid \mathcal{F}_n) + o_p(n^{-1}N_n^2)$$

for $t = r, d$.

Remark 3. Theorem 2 for Poisson sampling is an extension of the result of Rao and Shao (1992) for the REE in that we allow G_n to increase, but at a rate slower than $n^{1/3}$.

Remark 4. Using an argument similar to that of theorem 3.4 of Krewski and Rao (1981), it can be shown that the replication method can be applied to estimate the variance of a smooth function of several DEEs or of several REEs.

4. EXTENSIONS

4.1 Two-Phase Regression Estimator

To simplify the notation, we suppress the subscript n in what follows. The two-phase regression estimator can be written in the form

$$\hat{Y}_{t,\text{REG}} = \hat{\mathbf{T}}'_{c,1}\hat{\boldsymbol{\beta}}_2 := \sum_{i\in A_2} \alpha_i y_i, \tag{40}$$

where the notation $A := B$ denotes that B is defined to be equal to A, $\hat{\mathbf{T}}_{c,1}$ is the vector of estimated population totals estimated with the first-phase sample, $\hat{\boldsymbol{\beta}}_2$ is a vector of estimated regression coefficients estimated with the second-phase sample, and the α_i are functions of the sample but not of y. The estimator $\hat{\mathbf{T}}_{c,1}$ is the vector of realized first-phase sample sizes in the DEE and is the vector of estimated population sizes of the second-phase strata calculated from the first-phase sample for the REE.

For a control variable $\mathbf{c}_i$ of fixed dimension observed on the first-phase sample and a stratified second-phase sample, the estimated total is

$$\hat{\mathbf{T}}_{c,1} = \sum_{i\in A_1} w_i \mathbf{c}_i,$$

and the estimated regression coefficient is

$$\hat{\boldsymbol{\beta}}_2 = \left(\sum_{i\in A_2} w_i w_i^* \mathbf{c}_i \mathbf{c}_i'\right)^{-1} \sum_{i\in A_2} w_i w_i^* \mathbf{c}_i y_i,$$

where $w_i^* = r_g^{-1} n_{1g}$ for $x_{ig} = 1$. Then $\hat{\boldsymbol{\beta}}_2$ is a smooth function of two DEEs (a DEE for the total of $\mathbf{c}_i\mathbf{c}_i'$ and a DEE for the total of $\mathbf{c}_i y_i$), and by Remark 4, replicates can be used to construct a variance estimator for the two-phase regression estimator; that is, the kth replicate for $\hat{Y}_{t,\text{REG}}$ is

$$\hat{Y}_{t,\text{REG}}^{(k)} = \hat{\mathbf{T}}_{c,1}^{(k)\prime}\hat{\boldsymbol{\beta}}_2^{(k)} := \sum_{i\in A_2} \alpha_i^{(k)} y_i, \tag{41}$$

where

$$\hat{\mathbf{T}}_{c,1}^{(k)} = \sum_{i\in A_1} w_i^{(k)} \mathbf{c}_i,$$

$$\hat{\boldsymbol{\beta}}_2^{(k)} = \left(\sum_{i\in A_2} w_i^{(k)} w_i^{*(k)} \mathbf{c}_i \mathbf{c}_i'\right)^{-1} \sum_{i\in A_2} w_i^{(k)} w_i^{*(k)} \mathbf{c}_i y_i,$$

and $w_i^{*(k)} = (\sum_{i\in A_2} w_i^{(k)} w_i^{-1} x_{ig})^{-1} \sum_{i\in A_1} w_i^{(k)} w_i^{-1} x_{ig}$ for $x_{ig} = 1$. The resulting replication weights satisfy

$$\sum_{i\in A_2} \alpha_i^{(k)} \mathbf{c}_i = \sum_{i\in A_1} w_i^{(k)} \mathbf{c}_i \tag{42}$$

for each $k = 1, 2, \ldots, L$.

4.2 Three-Phase Sampling

Replication variance estimation can be extended to three-phase stratified sampling. Let a three-phase estimator be written in the form

$$\hat{Y}_3 = \hat{\mathbf{Z}}_2'\hat{\boldsymbol{\beta}}_3, \tag{43}$$

where $\hat{\mathbf{Z}}_2$ is the control total for certain characteristics, denoted by $\mathbf{z}_i$, calculated from the second-phase sample; $\hat{\boldsymbol{\beta}}_3$ is the estimated regression coefficients estimated with the third-phase sample; and A_3 is the set of indices for the third-phase sample. We assume that we can write

$$\begin{aligned} \hat{\mathbf{Z}}_2 &= \sum_{i\in A_2} \alpha_i \mathbf{z}_i, \\ \hat{\boldsymbol{\beta}}_3 &= \left(\sum_{i\in A_3} \alpha_i \alpha_i^* \mathbf{z}_i \mathbf{z}_i'\right)^{-1} \sum_{i\in A_3} \alpha_i \alpha_i^* \mathbf{z}_i y_i, \end{aligned} \tag{44}$$

where α_i is the two-phase sampling weight of unit i defined in (40) and α_i^* is the conditional sampling weight for third-phase sampling.

If the conditional third-phase sampling weight can be written as $\alpha_i^* = (\sum_{i\in A_3} I_{is})^{-1} \sum_{i\in A_2} I_{is}$ for unit i in the sth third-phase stratum, where I_{is} is the indicator function for the inclusion of unit i to the sth third-phase stratum, then the kth replicate for the three-phase estimator can be created as $\hat{Y}_3^{(k)} = \hat{\mathbf{Z}}_2^{(k)\prime}\hat{\boldsymbol{\beta}}_3^{(k)}$, where $\hat{\mathbf{Z}}_2^{(k)}$ and $\hat{\boldsymbol{\beta}}_3^{(k)}$ are constructed using $\alpha_i^{(k)}$ and $\alpha_i^{*(k)}$ instead of using α_i and α_i^* in (44), where $\alpha_i^{(k)}$ is as defined in (41) and $\alpha_i^{*(k)} = (\sum_{i\in A_3} \alpha_i^{(k)} \alpha_i^{-1} I_{is})^{-1} \sum_{i\in A_2} \alpha_i^{(k)} \alpha_i^{-1} I_{is}$ for unit i in the sth third-phase stratum.

APPENDIX A: PROOF OF THEOREM 1

First, we express the REE as

$$\hat{Y}_{nr} = \hat{Y}_{n1} + \sum_{g=1}^{G_n} \hat{X}_{ng1}(\bar{y}_{ng2} - \bar{y}_{ng1}), \tag{A.1}$$

where $\bar{y}_{ng1}$ and $\bar{y}_{ng2}$ are defined in (8) and (11). Write $y_{ni} = \sum_{g=1}^{G_n} x_{nig}(\bar{Y}_{ng} + e_{nig})$, where $\bar{Y}_{ng} = X_{ng}^{-1} Y_{ng}$ is the population mean of y for group g. Then (A.1) becomes

$$\hat{Y}_{nr} = \hat{Y}_{n1} + \sum_{g=1}^{G_n} (\hat{X}_{ng1}\hat{X}_{ng2}^{-1}\hat{T}_{eng2} - \hat{T}_{eng1}), \tag{A.2}$$

where

$$\begin{aligned} &(\hat{T}_{eng2}, \hat{T}_{eng1}, \hat{X}_{ng2}, \hat{X}_{ng1}) \\ &\qquad = \sum_{i\in A_{n1}} w_{ni} x_{nig} (\pi_{ng}^{-1} a_{ni} e_{nig}, e_{nig}, \pi_{ng}^{-1} a_{ni}, 1), \end{aligned}$$

π_{ng} is the second-phase selection probability in group g, and a_{ni} is the second-phase sample indicator.

By the unbiasedness assumption (9) and the definition of π_{ng},

$$E(\hat{T}_{eng1}, \hat{T}_{eng2}, \hat{X}_{ng1}, \hat{X}_{ng2} \mid \mathcal{F}_n) = (0, 0, X_{ng}, X_{ng}).$$

By assumptions (18) and (14),

$$\operatorname{var}\{(\hat{T}_{eng1}, \hat{X}_{ng1}) \mid \mathcal{F}_n\} = O(G_n^{-1} n^{-1} N_n^2). \tag{A.3}$$

Thus, using corollary 5.1.1.2 of Fuller (1996), we have

$$(\hat{T}_{eng1}, \hat{X}_{ng1}) = (0, X_{ng}) + O_p\big(G_n^{-1/2} n^{-1/2} N_n\big). \tag{A.4}$$

The variance of $\hat{T}_{eng2}$ can be decomposed into two parts,

$$\begin{aligned}\operatorname{var}(\hat{T}_{eng2} \mid \mathcal{F}_n) &= \operatorname{var}\{E(\hat{T}_{eng2} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) \mid \mathcal{F}_n\} \\ &\quad + E\{\operatorname{var}(\hat{T}_{eng2} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) \mid \mathcal{F}_n\},\end{aligned} \tag{A.5}$$

where $\mathbf{r} = (r_{n1}, r_{n2}, \ldots, r_{nG_n})$. Under simple random sampling in each group at the second phase,

$$E(\hat{T}_{eng2} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) = \hat{T}_{eng1} \tag{A.6}$$

and

$$\begin{aligned}&\operatorname{var}(\hat{T}_{eng2} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) \\ &\quad = \begin{cases} \pi_{ng}^{-1}(1-\pi_{ng}) n_{1ng}(n_{1ng}-1)^{-1} \\ \quad \times \left(\sum_{i \in A_{n1}} w_{ni}^2 x_{nig} e_{nig}^2 - n_{1ng}^{-1} \hat{T}_{eng1}^2 \right) & \text{if } n_{1ng} > 1 \\ 0 & \text{if } n_{1ng} \le 1. \end{cases}\end{aligned}$$

Thus,

$$\begin{aligned}\operatorname{var}(\hat{T}_{eng2} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) &\le 2\pi_{ng}^{-1}(1-\pi_{ng}) \sum_{i \in A_{n1}} w_{ni}^2 x_{nig} e_{nig}^2 \\ &= O_p(n^{-1} G_n^{-1} N_n^2),\end{aligned} \tag{A.7}$$

by (14) and (16). Thus, inserting (A.6) and (A.7) into (A.5), we have

$$\operatorname{var}\{(\hat{T}_{eng2}, \hat{X}_{ng2}) \mid \mathcal{F}_n\} = O(G_n^{-1} n^{-1} N_n^2). \tag{A.8}$$

By (14), $X_{ng}^{-1} = O(G_n N_n^{-1})$, and by a Taylor expansion,

$$\hat{X}_{ng1} \hat{X}_{ng2}^{-1} \hat{T}_{eng2} = \hat{T}_{eng2} + O_p(n^{-1} N_n). \tag{A.9}$$

Because the estimator is defined to have moments,

$$\begin{aligned}&\hat{Y}_{nr} - \hat{Y}_{n1} \\ &\quad = \sum_{g=1}^{G_n} \sum_{i \in A_{n1}} w_{ni} \left(\frac{a_{ni}}{\pi_{ng}} - 1 \right) x_{nig} e_{nig} + O_p(G_n n^{-1} N_n),\end{aligned} \tag{A.10}$$

and $O(G_n n^{-1} N_n) = o(n^{-1/2} N_n)$ by (15).

Define

$$\tilde{Y}_{nr} = \hat{Y}_{n1} + \sum_{g=1}^{G_n} \sum_{i \in A_{n1}} w_{ni} \left(\frac{a_{ni}}{\pi_{ng}} - 1 \right) x_{nig} e_{nig}.$$

By simple random sampling in each group at the second phase,

$$E(\tilde{Y}_{nr} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) = \hat{Y}_{n1} \tag{A.11}$$

and

$$\begin{aligned}&\operatorname{var}(\tilde{Y}_{nr} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) \\ &\quad = \sum_{g=1}^{G_n} \pi_{ng}^{-1}(1-\pi_{ng}) \frac{n_{1ng}}{n_{1ng}-1} \sum_{i \in A_{n1}} w_{ni}^2 x_{nig} e_{nig}^2 \\ &\qquad - \sum_{g=1}^{G_n} \pi_{ng}^{-1}(1-\pi_{ng})(n_{1ng}-1)^{-1} \left(\sum_{i \in A_{n1}} w_{ni} x_{nig} e_{nig} \right)^2\end{aligned} \tag{A.12}$$

for $n_{1ng} > 1$ and $\operatorname{var}(\tilde{Y}_{nr} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) = 0$ for $n_{1ng} \le 1$. The second term on the right side of equality (A.12) is $O_p(\sum_{g=1}^{G_n} n_{1ng}^{-1} N_n^2 G_n^{-1} \times n^{-1}) = o_p(n^{-1} N_n^2)$, by (14) and (15). Hence we have result (21).

The unbiasedness of the DEE follows directly from

$$\begin{aligned}E(\hat{Y}_{nd} \mid \mathcal{F}_n) &= E\{E(\hat{Y}_{nd} \mid \mathcal{F}_n, A_{n1}, \mathbf{r}) \mid \mathcal{F}_n\} \\ &= E(\hat{Y}_{n1} \mid \mathcal{F}_n).\end{aligned}$$

Result (20) follows because $\hat{Y}_{n1}$ is unbiased for Y_n, by (6).

To derive the variance of the DEE, we write

$$\hat{Y}_{nd} = \hat{Y}_{n1} + \sum_{g=1}^{G_n} \sum_{i \in A_{n1}} w_{ni} x_{nig} \left(\frac{a_{ni}}{\pi_{ng}} - 1 \right) y_{ni},$$

and, using the unbiasedness assumption, we have

$$\begin{aligned}&\operatorname{var}\{\hat{Y}_{nd} \mid \mathcal{F}_n\} \\ &\quad = \operatorname{var}\{\hat{Y}_{n1} \mid \mathcal{F}_n\} \\ &\qquad + E\left\{ \sum_{g=1}^{G_n} \operatorname{var}\left(\sum_{i \in A_{n1}} w_{ni} x_{nig} \frac{a_{ni}}{\pi_{ng}} y_{ni} \,\middle|\, \mathcal{F}_n, A_{n1}, \mathbf{r} \right) \,\middle|\, \mathcal{F}_n \right\}.\end{aligned}$$

By simple random sampling at phase two,

$$\operatorname{var}(\hat{Y}_{ng2} \mid \mathcal{F}_n, A_{n1}) = n_{1ng}^2 \left(\frac{1}{r_{ng}} - \frac{1}{n_{1ng}} \right) \sigma_{nwyg1}^2,$$

where

$$\hat{\sigma}_{nwyg1}^2 = (n_{1ng}-1)^{-1} \sum_{i \in A_{n1}} \left(x_{nig} w_{ni} y_{ni} - n_{1ng}^{-1} \sum_{j \in A_{n1}} x_{njg} w_{nj} y_{nj} \right)^2.$$

APPENDIX B: PROOF OF THEOREM 2

Either the REE or the DEE can be written as

$$\begin{aligned}\hat{Y}_{n,tp} &= \sum_{g=1}^{G_n} \Bigg[\sum_{i \in A_{n1}} w_{ni} x_{nig} q_{ni} \left(\sum_{i \in A_{n2}} \pi_{ng}^{-1} w_{ni} q_{ni} x_{nig} \right)^{-1} \\ &\qquad \times \sum_{i \in A_{n2}} \pi_{ng}^{-1} w_{ni} x_{nig} y_{ni} \Bigg] \\ &= \sum_{g=1}^{G_n} \hat{x}_{ng} \hat{z}_{ng}^{-1} \hat{u}_{ng},\end{aligned} \tag{B.1}$$

where

$$(\hat{x}_{ng}, \hat{z}_{ng}, \hat{u}_{ng}) = \sum_{i \in A_{n1}} w_{ni} x_{nig} (q_{ni}, \pi_{ng}^{-1} a_{ni} q_{ni}, \pi_{ng}^{-1} a_{ni} y_{ni})$$

and $q_{ni} = 1$ for the REE and $q_{ni} = w_{ni}^{-1} n^{-1} N_n$ for the DEE. Similarly, we can write the replicates

$$\hat{Y}_{n,tp}^{(k)} = \sum_{g=1}^{G_n} \hat{x}_{ng}^{(k)} \big(\hat{z}_{ng}^{(k)}\big)^{-1} \hat{u}_{ng}^{(k)}, \tag{B.2}$$

where

$$\big(\hat{x}_{ng}^{(k)}, \hat{z}_{ng}^{(k)}, \hat{u}_{ng}^{(k)}\big) = \sum_{i \in A_{n1}} w_{ni}^{(k)} x_{nig} (q_{ni}, \pi_{ng}^{-1} a_{ni} q_{ni}, \pi_{ng}^{-1} a_{ni} y_{ni}).$$

By the argument used for (A.3) and (A.8),

$$\operatorname{var}\{(\hat{x}_{ng}, \hat{z}_{ng}, \hat{u}_{ng}) \mid \mathcal{F}_n\} = O(G_n^{-1} n^{-1} N_n^2).$$

Thus, by assumption (32),

$$\begin{aligned}&c_{nk}^{1/2} \big(\hat{x}_{ng}^{(k)} - \hat{x}_{ng}, \hat{z}_{ng}^{(k)} - \hat{z}_{ng}, \hat{u}_{ng}^{(k)} - \hat{u}_{ng}\big) \\ &\qquad = O_p\big(L_n^{-1/2} G_n^{-1/2} n^{-1/2} N_n\big).\end{aligned} \tag{B.3}$$

By a Taylor expansion, using (B.3),

$$\begin{aligned} N_n^{-1}c_{nk}^{1/2}&\bigl[\hat{x}_{ng}^{(k)}\bigl(\hat{z}_{ng}^{(k)}\bigr)^{-1}\hat{u}_{ng}^{(k)}-\hat{x}_{ng}\hat{z}_{ng}^{-1}\hat{u}_{ng}\bigr] \\ &= N_n^{-1}c_{nk}^{1/2}\bigl[\hat{z}_{ng}^{-1}\hat{x}_{ng}\bigl(\hat{u}_{ng}^{(k)}-\hat{u}_{ng}\bigr)-\hat{z}_{ng}^{-2}\hat{x}_{ng}\hat{u}_{ng}\bigl(\hat{z}_{ng}^{(k)}-\hat{z}_{ng}\bigr) \\ &\qquad +\hat{z}_{ng}^{-1}\hat{u}_{ng}\bigl(\hat{x}_{ng}^{(k)}-\hat{x}_{ng}\bigr)\bigr]+N_n^{-1}c_{nk}^{1/2}R_{ng,k}, \end{aligned} \tag{B.4}$$

where

$$\begin{aligned} R_{ng,k} &= \tilde{z}_{ng}^{-1}\bigl(\hat{x}_{ng}^{(k)}-\hat{x}_{ng}\bigr)\bigl(\hat{u}_{ng}^{(k)}-\hat{u}_{ng}\bigr) \\ &\quad -\tilde{z}_{ng}^{-2}\tilde{u}_{ng}\bigl(\hat{x}_{ng}^{(k)}-\hat{x}_{ng}\bigr)\bigl(\hat{z}_{ng}^{(k)}-\hat{z}_{ng}\bigr) \\ &\quad -\tilde{z}_{ng}^{-2}\tilde{x}_{ng}\bigl(\hat{u}_{ng}^{(k)}-\hat{u}_{ng}\bigr)\bigl(\hat{z}_{ng}^{(k)}-\hat{z}_{ng}\bigr) \\ &\quad +\tilde{z}_{ng}^{-3}\tilde{x}_{ng}\tilde{u}_{ng}\bigl(\hat{z}_{ng}^{(k)}-\hat{z}_{ng}\bigr)^2, \end{aligned}$$

and $(\tilde{x}_{ng},\tilde{z}_{ng},\tilde{u}_{ng})$ is on the line segment joining $(\hat{x}_{ng},\hat{z}_{ng},\hat{u}_{ng})$ and $(\hat{x}_{ng}^{(k)},\hat{z}_{ng}^{(k)},\hat{u}_{ng}^{(k)})$. By (33), (B.3), and (15),

$$(\tilde{x}_{ng},\tilde{z}_{ng},\tilde{u}_{ng}) = (\hat{x}_{ng},\hat{z}_{ng},\hat{u}_{ng})\times[1+o_p(1)].$$

For example, by (14) and (32),

$$\begin{aligned} c_{nk}^{1/2}&\tilde{z}_{ng}^{-1}\bigl(\hat{x}_{ng}^{(k)}-\hat{x}_{ng}\bigr)\bigl(\hat{u}_{ng}^{(k)}-\hat{u}_{ng}\bigr) \\ &= c_{nk}^{-1/2}\tilde{z}_{ng}^{-1}\bigl[c_{nk}\bigl(\hat{x}_{ng}^{(k)}-\hat{x}_{ng}\bigr)\bigl(\hat{u}_{ng}^{(k)}-\hat{u}_{ng}\bigr)\bigr] \\ &= O_p\bigl(L_n^{-1/2}n^{-1}N_n\bigr) \end{aligned}$$

and

$$c_{nk}^{1/2}R_{ng,k} = O_p\bigl(L_n^{-1/2}n^{-1}N_n\bigr). \tag{B.5}$$

Also, by a Taylor expansion of $\bar{y}_{ng2}$ of (11) and applying (14), we have

$$\hat{z}_{ng}^{-1}\hat{x}_{ng} = 1+O_p\bigl(n^{-1/2}G_n^{1/2}\bigr) \tag{B.6}$$

and

$$\bar{y}_{ng2}-\ddot{Y}_{ng} = O_p\bigl(n^{-1/2}G_n^{1/2}\bigr), \tag{B.7}$$

where

$$\ddot{Y}_{ng} = \left(\sum_{i=1}^{N_n}x_{nig}q_{ni}\right)^{-1}\sum_{i=1}^{N_n}x_{nig}y_{ni}.$$

Thus, inserting (B.5)–(B.7) into (B.4), we have

$$\begin{aligned} c_{nk}^{1/2}&\bigl(\hat{Y}_{n,tp}^{(k)}-\hat{Y}_{n,tp}\bigr) \\ &= c_{nk}^{1/2}\sum_{g=1}^{G_n}\sum_{i\in A_{n1}}x_{nig}\bigl(w_{ni}^{(k)}-w_{ni}\bigr)[y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}] \\ &\quad +O_p\bigl(G_nn^{-1}L_n^{-1/2}N_n\bigr), \end{aligned} \tag{B.8}$$

where $\delta_{nig} = y_{ni}-q_{ni}\ddot{Y}_{ng}$. The remainder term is $O_p(G_nn^{-1}L_n^{-1/2}N_n)$ because of the existence of moments (see Fuller 1996, p. 304, ex. 21).

By assumption (32) and by (10), the main term of (B.8) is $O_p(n^{-1/2}L_n^{-1/2}N_n)$. It follows from (B.8) that

$$\begin{aligned} &\sum_{k=1}^{L_n}c_{nk}\bigl(\hat{Y}_{n,tp}^{(k)}-\hat{Y}_{n,tp}\bigr)^2 \\ &\quad =\sum_{k=1}^{L_n}c_{nk}\left\{\sum_{g=1}^{G_n}\sum_{i\in A_{n1}}x_{nig}\bigl(w_{ni}^{(k)}-w_{ni}\bigr)[y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}]\right\}^2 \\ &\qquad +O_p\bigl(G_nn^{-3/2}N_n^2\bigr). \end{aligned}$$

By (15), a term that is $O_p(G_nn^{-3/2}N_n^2)$ is $o_p(n^{-1}N_n^2)$.

Assume for now that the second-phase sampling is Bernoulli within each group (sometimes called "stratified Bernoulli sampling") with rates π_{ng}, $g=1,2,\ldots,G_n$. Let a_{ni}, $i=1,2,\ldots,N_n$, be random variables with $a_{ni}=1$ if unit i is selected for the second-phase sample and $a_{ni}=0$ if unit i is not selected. Conceptually, the second-phase sample indicator a_{ni} can be extended to the entire population. The extended definition of a_{ni} has been discussed by Fay (1991) and used by Rao and Shao (1992) and Shao and Steel (1999).

Fix $\mathbf{a}_n=(a_{n1},a_{n2},\ldots,a_{nN_n})$ and consider variance estimation for the population total of $[y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}]$ based on the sample, where the estimator of the total is

$$\begin{aligned} \tilde{Y}_{n,tp} &= \sum_{g=1}^{G_n}\sum_{i\in A_{n1}}w_{ni}x_{nig}[y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}] \\ &=: \sum_{i\in A_{n1}}w_{ni}\zeta_{ni}. \end{aligned} \tag{B.9}$$

By assumption, the full-sample variance estimator is consistent for any variable with fourth moments. Thus, because $y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}$ satisfies (6) with $\tau\ge 2$, by assumption (34), the replicate estimator of the variance of $\tilde{Y}_{n,tp}$ satisfies

$$\begin{aligned} \hat{V}&\{\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n\} \\ &= \operatorname{var}\left\{\sum_{g=1}^{G_n}\sum_{i\in A_{n1}}w_{ni}x_{nig}[y_{ni}+(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}]\,\middle|\,\mathbf{a}_n,\mathcal{F}_n\right\} \\ &\quad +o_p(n^{-1}N_n^2). \end{aligned} \tag{B.10}$$

The variance of $\tilde{Y}_{n,tp}$ can be expressed as

$$\begin{aligned} \operatorname{var}(\tilde{Y}_{n,tp}\mid\mathcal{F}_n) &= \operatorname{var}\bigl[E\{\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n\}\mid\mathcal{F}_n\bigr] \\ &\quad +E\bigl[\operatorname{var}\{\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n\}\mid\mathcal{F}_n\bigr]. \end{aligned} \tag{B.11}$$

We next show that $\hat{V}\{\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n\}$ of (B.10) is a consistent estimator of the last term of (B.11). For this, it suffices to demonstrate that

$$\operatorname{var}\bigl\{nN_n^{-2}\operatorname{var}(\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n)\mid\mathcal{F}_n\bigr\} = o(1). \tag{B.12}$$

Writing $u_{ni}=\sum_{g=1}^{G_n}x_{nig}(\pi_{ng}^{-1}a_{ni}-1)\delta_{nig}$ and $\hat{U}_{n1}=\sum_{i\in A_{n1}}w_{ni}\times u_{ni}$,

$$\begin{aligned} \operatorname{var}(\tilde{Y}_{n,tp}\mid\mathbf{a}_n,\mathcal{F}_n) &= \operatorname{var}(\hat{Y}_{n1}\mid\mathbf{a}_n,\mathcal{F}_n)+\operatorname{var}(\hat{U}_{n1}\mid\mathbf{a}_n,\mathcal{F}_n) \\ &\quad +2\operatorname{cov}(\hat{Y}_{n1},\hat{U}_{n1}\mid\mathbf{a}_n,\mathcal{F}_n). \end{aligned}$$

Using (30),

$$\begin{aligned} &\operatorname{var}\bigl\{nN_n^{-2}\operatorname{var}(\hat{U}_{n1}\mid\mathbf{a}_n,\mathcal{F}_n)\mid\mathcal{F}_n\bigr\} \\ &\quad =\sum_{i=1}^{N_n}\sum_{j=1}^{N_n}\sum_{k=1}^{N_n}\sum_{m=1}^{N_n}\Omega_{nij}\Omega_{nkm}\operatorname{cov}(u_{ni}u_{nj},u_{nk}u_{nm}\mid\mathcal{F}_n), \end{aligned} \tag{B.13}$$

where the covariances are with respect to the distribution of the a_{ni}'s. Because the a_{ni}'s are independent and $E(u_{ni}\mid\mathcal{F}_n)=0$, among the N_n^4 terms in the summation of (B.13), only those terms with $(i,j)=(k,m)$ or $(i,j)=(m,k)$ are nonzero. Thus the summation in (B.13) reduces to

$$\begin{aligned} &\sum_{i=1}^{N_n}\sum_{j=1}^{N_n}(\Omega_{nij}^2+\Omega_{nij}\Omega_{nji})\operatorname{var}(u_{ni}u_{nj}\mid\mathcal{F}_n) \\ &\qquad \le 2K_{nu1}\Bigl(\max_{i,j}|\Omega_{nij}|\Bigr)\left(\sum_{i=1}^{N_n}\sum_{j=1}^{N_n}|\Omega_{nij}|\right), \end{aligned} \tag{B.14}$$

where $K_{nu1} = \max_{i,j} \operatorname{var}(u_{ni}u_{nj} \mid \mathcal{F}_n)$. Because π_{ng}^{-1} is bounded and by (6) with $\tau \geq 2$, $K_{nu1} = O(1)$. By (31), $\max_{i,j} |\Omega_{nij}| = O(N_n^{-1})$ and

$$\operatorname{var}\{nN_n^{-2} \operatorname{var}(\hat{U}_{n1} \mid \mathbf{a}_n, \mathcal{F}_n) \mid \mathcal{F}_n\} = O(N_n^{-1}). \tag{B.15}$$

By (30),

$$\operatorname{var}\{nN_n^{-2} \operatorname{cov}(\hat{Y}_{n1}, \hat{U}_{n1} \mid \mathbf{a}_n, \mathcal{F}_n) \mid \mathcal{F}_n\} = \sum_{i=1}^{N_n}\sum_{j=1}^{N_n}\sum_{k=1}^{N_n}\sum_{m=1}^{N_n} \Omega_{nij}\Omega_{nkm}y_{nj}y_{nm} \operatorname{cov}(u_{ni}, u_{nk} \mid \mathcal{F}_n). \tag{B.16}$$

Because the a_{ni}'s are independent and $E(u_{ni} \mid \mathcal{F}_n) = 0$, the term in (B.16) reduces to

$$\sum_{i=1}^{N_n}\sum_{j=1}^{N_n}\sum_{m=1}^{N_n} \Omega_{nij}\Omega_{nim}y_{nj}y_{nm} \operatorname{var}(u_{ni} \mid \mathcal{F}_n) \leq K_{nu2}\sum_{i=1}^{N_n}\left(\sum_{j=1}^{N_n} \Omega_{nij}y_{nj}\right)^2,$$

where $K_{nu2} = \max_i \operatorname{var}(u_{ni} \mid \mathcal{F}_n)$. By (6), there exists K_y such that $y_{ni} \leq K_y N_n^{1/(2+\tau)}$ for all $i = 1, 2, \ldots, N_n$. Thus, by (31) we have

$$\operatorname{var}\{nN_n^{-2} \operatorname{cov}(\hat{Y}_{n1}, \hat{U}_{n1} \mid \mathbf{a}_n, \mathcal{F}_n) \mid \mathcal{F}_n\} = O(N_n^{-\tau/(2+\tau)}) = o(1). \tag{B.17}$$

Because $\operatorname{var}(\hat{Y}_{n1} \mid \mathbf{a}_n, \mathcal{F}_n)$ does not depend on $\mathbf{a}_n$, $\operatorname{var}\{\operatorname{var}(\hat{Y}_{n1} \mid \mathbf{a}_n, \mathcal{F}_n) \mid \mathcal{F}_n\} = 0$, and result (B.12) follows from (B.15) and (B.17).

Now,

$$E\{\tilde{Y}_{n,tp} - Y_N \mid \mathbf{a}_n, \mathcal{F}_n\} = \sum_{g=1}^{G_n}\sum_{i=1}^{N_n} x_{nig}(\pi_{ng}^{-1}a_{ni} - 1)\delta_{nig},$$

and the first term on the right side of the equality of (B.11) is

$$\operatorname{var}\big[E\{\tilde{Y}_{n,tp} \mid \mathbf{a}_n, \mathcal{F}_n\} \mid \mathcal{F}_n\big] = \sum_{g=1}^{G_n} \pi_{ng}^{-1}(1 - \pi_{ng})\sum_{i=1}^{N_n} x_{nig}\delta_{nig}^2. \tag{B.18}$$

Therefore, combining (B.10), (B.11), and (B.18), we have

$$\hat{V}\{\tilde{Y}_{n,tp} \mid \mathbf{a}_n, \mathcal{F}_n\} = \operatorname{var}(\tilde{Y}_{n,tp} \mid \mathcal{F}_n) - \sum_{g=1}^{G_n}\sum_{i=1}^{N_n} \pi_{ng}^{-1}(1 - \pi_{ng})x_{nig}\delta_{nig}^2 + o_p(n^{-1}N_n^2). \tag{B.19}$$

Because $\delta_{nig} = e_{nig} = y_{ni} - \bar{Y}_{ng}$ for REE, (35) follows for any type of Poisson sampling including stratified Bernoulli at the second phase.

By the definition of q_{ni} of (B.1), $\delta_{nig} = \eta_{nig} = y_{ni} - w_{ni}^{-1}n^{-1}N_n\ddot{Y}_{ng}$ for DEE. Substituting δ_{nig} for the DEE into the last expression of (B.10) and using

$$\sum_{i\in A_{n1}} w_{ni}\sum_{g=1}^{G_n}(\pi_{ng}^{-1}a_{ni} - 1)x_{nig}\eta_{nig} = \sum_{i\in A_{n1}} w_{ni}\sum_{g=1}^{G_n}(\pi_{ng}^{-1}a_{ni} - 1)x_{nig}y_{ni},$$

we have

$$\sum_{g=1}^{G_n}\sum_{i\in A_{n1}} w_{ni}x_{nig}[y_{ni} + (\pi_{ng}^{-1}a_{ni} - 1)\delta_{nig}] = \sum_{g=1}^{G_n}\sum_{i\in A_{n1}} w_{ni}x_{nig}\pi_{ng}^{-1}a_{ni}y_{ni} = \hat{Y}_{nd},$$

and (36) follows for any type of Poisson sampling at the second phase.

So far we have assumed that the second-phase sampling is Poisson. In Theorems 1 and 2, in contrast, it is stratified simple random sampling. Using the arguments of Hájek (1960), we can show that there exists a sequence of simple random sample stratum means that differ from the Bernoulli stratum means by a term that is an order in probability of $n_{1ng}^{-3/4}$. Let $\hat{\gamma}_{ng}$ be the difference between the Bernoulli stratum mean of δ_{nig} and the simple random sample mean. That is, letting a_{ni} and a_{ni}^* denote the Bernoulli sample indicator and the simple random sampling indicator of unit i,

$$\hat{\gamma}_{ng} = \frac{\sum_{i\in A_{n1}} w_{ni}x_{nig}a_{ni}\delta_{nig}}{\sum_{i\in A_{n1}} w_{ni}x_{nig}a_{ni}} - \frac{\sum_{i\in A_{n1}} w_{ni}x_{nig}a_{ni}^*\delta_{nig}}{\sum_{i\in A_{n1}} w_{ni}x_{nig}a_{ni}^*}.$$

Then, by the argument of Hájek (1960),

$$\operatorname{var}(\hat{\gamma}_{ng} \mid \mathcal{F}_n) = O(n_{1ng}^{-3/2}). \tag{B.20}$$

Thus, by (32) and (14),

$$c_{nk}^{1/2}(\hat{\gamma}_{ng}^{(k)} - \hat{\gamma}_{ng}) = O_p(L_n^{-1/2}n^{-3/4}G_n^{3/4}) \tag{B.21}$$

for

$$\hat{\gamma}_{ng}^{(k)} = \frac{\sum_{i\in A_{n1}} w_{ni}^{(k)}x_{nig}a_{ni}\delta_{nig}}{\sum_{i\in A_{n1}} w_{ni}^{(k)}x_{nig}a_{ni}} - \frac{\sum_{i\in A_{n1}} w_{ni}^{(k)}x_{nig}a_{ni}^*\delta_{nig}}{\sum_{i\in A_{n1}} w_{ni}^{(k)}x_{nig}a_{ni}^*}.$$

Define $\hat{z}_{ng}^*$ and $\hat{u}_{ng}^*$ to be $\hat{z}_{ng}$ and $\hat{u}_{ng}$ of (B.1), with a_{ni} replaced by a_{ni}^*. Also, define $\hat{z}_{ng}^{(k)*}$ and $\hat{u}_{ng}^{(k)*}$ using a_{ni}^* in expression (B.2), and define $\hat{Y}_{n,tp}^*$ and $\hat{Y}_{n,tp}^{(k)*}$ using a_{ni}^* in expressions (B.1) and (B.2). Then

$$\hat{Y}_{n,tp}^{(k)} - \tilde{Y}_{n,tp}^{(k)*} = \sum_{g=1}^{G_n} \hat{x}_{ng}^{(k)}\hat{\gamma}_{ng}^{(k)}$$

and

$$\hat{Y}_{n,tp} - \hat{Y}_{n,tp}^* = \sum_{g=1}^{G_n} \hat{x}_{ng}\hat{\gamma}_{ng}. \tag{B.22}$$

Thus,

$$c_{nk}^{1/2}(\hat{Y}_{n,tp}^{(k)} - \hat{Y}_{n,tp}) = c_{nk}^{1/2}(\hat{Y}_{n,tp}^{(k)*} - \hat{Y}_{n,tp}^*) + c_{nk}^{1/2}\sum_{g=1}^{G_n}(\hat{x}_{ng}^{(k)}\hat{\gamma}_{ng}^{(k)} - \hat{x}_{ng}\hat{\gamma}_{ng}).$$

By a Taylor expansion, using (B.3) and (B.21),

$$\begin{aligned} N_n^{-1}c_{nk}^{1/2}(\hat{x}_{ng}^{(k)}\hat{\gamma}_{ng}^{(k)} - \hat{x}_{ng}\hat{\gamma}_{ng}) &= \hat{\gamma}_{ng}\{N_n^{-1}c_{nk}^{1/2}(\hat{x}_{ng}^{(k)} - \hat{x}_{ng})\} \\ &\quad + N_n^{-1}\hat{x}_{ng}c_{nk}^{1/2}(\hat{\gamma}_{ng}^{(k)} - \hat{\gamma}_{ng}) + O_p(L_n^{-1}n^{-3/2}G_n^{3/2}) \\ &= O_p(L_n^{-1/2}n^{-3/4}G_n^{-1/4}). \end{aligned}$$

Therefore,

$$c_{nk}^{1/2}(\hat{Y}_{n,tp}^{(k)} - \hat{Y}_{n,tp}) = c_{nk}^{1/2}(\hat{Y}_{n,tp}^{(k)*} - \hat{Y}_{n,tp}^*) + O_p(L_n^{-1/2}n^{-3/4}G_n^{3/4}N_n),$$

and, by (15) with $\lambda < 3^{-1}$, a term that is $O_p(L_n^{-1/2}n^{-3/4}G_n^{3/4}N_n)$ is $o_p(L_n^{-1/2}n^{-1/2}N_n)$. It follows that the replication variance estimator satisfies

$$\sum_{k=1}^{L_n} c_{nk}(\hat{Y}_{n,tp}^{(k)} - \hat{Y}_{n,tp})^2 = \sum_{k=1}^{L_n} c_{nk}(\hat{Y}_{n,tp}^{(k)*} - \hat{Y}_{n,tp}^*)^2 + o_p(n^{-1}N_n^2). \tag{B.23}$$

Also, by (B.22) and using (14), (B.20), and (15) with $\lambda < 3^{-1}$,

$$\hat{Y}_{n,tp} - \hat{Y}^*_{n,tp} = o_p\big(n^{-1/2}N_n\big). \tag{B.24}$$

Therefore, by (B.19), (B.23), and (B.24), the consistency of the variance estimators under second-phase stratified simple random sampling is established.

[Received July 2004. Revised June 2005.]

REFERENCES

Binder, D. A. (1996), "Linearization Methods for Single Phase and Two-Phase Samples: A Cookbook Approach," *Survey Methodology*, 22, 17–22.

Binder, D. A., Babyak, C., Brodeur, M., Hidiroglou, M., and Jocelyn, W. (2000), "Variance Estimation for Two-Phase Stratified Sampling," *Canadian Journal of Statistics*, 28, 751–764.

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

Fay, R. E. (1989), "Theory and Application of Replicate Weighting for Variance Calculations," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 212–217.

——— (1991), "A Design-Based Perspective on Missing Data Variance," in *Proceedings of Bureau of the Census Annual Research Conference*, American Statistical Association, pp. 429–440.

Fuller, W. A. (1975), "Regression Analysis for Sample Survey," *Sankhyā*, Ser. C, 37, 117–132.

——— (1996), *Introduction to Statistical Time Series* (2nd ed.), New York: Wiley.

——— (1998), "Replication Variance Estimation for Two-Phase Samples," *Statistica Sinica*, 8, 1153–1164.

Hájek, J. (1960), "Limiting Distributions in Simple Random Sampling From a Finite Population," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5, 361–374.

Kim, J. K., Navarro, A., and Fuller, W. A. (2000), "Variance Estimation for 2000 Census Coverage Estimates," in *Proceedings of the Survey Research Methods Section*, American Statistical Association, pp. 515–520.

Kott, P. S. (1990), "Variance Estimation When a First-Phase Area Sample Is Restratified," *Survey Methodology*, 16, 99–103.

Kott, P. S., and Stukel, D. M. (1997), "Can the Jackknife Be Used With a Two-Phase Sample?" *Survey Methodology*, 23, 81–89.

Krewski, D., and Rao, J. N. K. (1981), "Inference From Stratified Samples: Properties of the Linearization, Jackknife and Balanced Repeated Replication Methods," *The Annals of Statistics*, 9, 1010–1019.

Rao, J. N. K. (1973), "On Double Sampling for Stratification and Analytical Surveys," *Biometrika*, 60, 125–133.

Rao, J. N. K., and Shao, J. (1992), "Jackknife Variance Estimation With Survey Data Under Hot Deck Imputation," *Biometrika*, 79, 811–822.

Särndal, C.-E., Swensson, B., and Wretman, J. (1992), *Model-Assisted Survey Sampling*, New York: Springer-Verlag.

Shao, J., and Steel, P. (1999), "Variance Estimation for Survey Data With Composite Imputation and Nonnegligible Sampling Fractions," *Journal of the American Statistical Association*, 94, 254–265.

Yung, W., and Rao, J. N. K. (2000), " Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information," *Journal of the American Statistical Association*, 95, 903–915.