Journal of Official Statistics, Vol. 27, No. 3, 2011, pp. 507–526

Random Group Variance Estimators for Survey Data with Random Hot Deck Imputation

Jun Shao[1] *and Qi Tang*[2]

Random hot deck imputation is often applied to survey data with nonresponse. One of the popular methods for variance estimation without nonresponse is the random group method, which has to be adjusted when it is applied to imputed data. One such kind of adjustment is reimputing nonrespondents in each random group. We show that the random group method with reimputation produces asymptotically unbiased and consistent variance estimators for estimated population totals. As a special case of our general result, the random group variance estimator for the case of no nonresponse is asymptotically unbiased and consistent, a result that has not been documented although the random group method is frequently used in applications. We also show how to apply a shortcut random group method, which reduces the computational complexity due to reimputation, and establish the asymptotic unbiasedness and consistency of the resulting variance estimators.

Key words: Hot deck imputation; nonresponse; variance estimation; random group; reimputation; shortcut.

1. Introduction

Nonresponse exists in most survey problems. Hot deck imputation is a very popular method to impute nonrespondents by respondents from the same variable (Kalton and Kasprzyk 1986; Rubin 1987). In this article, we focus on ignorable nonresponse, random hot deck imputation, and the most basic survey sampling design, the stratified probability proportional to size sampling design considered as a single stage sampling design or the first stage of a multi-stage sampling design.

Variance estimation is an important element in sample surveys. The random group method (Wolter 2007) is a popular replication method used in many economic surveys in agencies such as the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. Replication methods require more computation, but have the advantages of (1) requiring no separate theoretical derivations of a variance formula for each problem, which can be difficult or messy; (2) programming ease in complex situations; (3) using a unified recipe for various problems; and (4) to some degree, robustness against violations of models/assumptions.

[1] School of Finance and Statistics, East China Normal University, Shanghai, China, and Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A. Email: shao@stat.wisc.edu
[2] Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A. Email: qitang@stat.wisc.edu
Acknowledgments: The authors thank two referees for their helpful comments and suggestions. The research was partially supported by the National Science Foundation Grants SES-0705033 and DMS-1007454.

Although most imputation methods are designed so that treating imputed values as observed data and applying formulas for the case of no nonresponse leads to approximately valid survey estimators (such as estimators of population means or totals), treating imputed values as observed data and applying variance formulas for the case of no nonresponse produces substantial underestimation of variances when the proportion of nonrespondents is appreciable. Thus, various adjustment methods are proposed in the literature for imputed data. For replication methods, Rao and Shao (1992) proposed to apply an adjustment to every replicate to address the effect of nonresponse and imputation. Shao (2001) showed that the Rao-Shao adjustment is equivalent to reimputing every replicate, i.e., performing the same imputation procedure in every replicate using data in the replicate.

For replication methods such as the jackknife, balanced repeated replication, and bootstrap, the resulting variance estimators after reimputation have been shown to be asymptotically consistent (e.g., Rao and Shao 1992; Shao and Sitter 1996; Shao et al. 1998). However, the same result for the random group variance estimator with reimputation adjustment has not been established prior to our study. In fact, even the consistency of the random group variance estimator in the case of no nonresponse has not been documented. The first purpose of this article is to show the asymptotic unbiasedness and consistency of the random group variance estimator with reimputation for data with nonrespondents imputed by random hot deck. Of course, our result includes the case of no nonresponse as a special case.

Since reimputation has to be applied to every replicate, the computation of a replication variance estimator can be time-consuming and computer-intensive. Some shortcut replication methods have been considered to reduce the computational complexity (see Moore 2006; Thompson and Yung 2006; Haziza et al. 2010). These shortcut methods vary with the imputation method and/or the replication method. The second purpose of this article is to study the construction of a shortcut random group variance estimator that reduces the computational complexity due to reimputation and is still asymptotically unbiased and consistent.

The formulas for the random group variance estimator with reimputation and its shortcut version are introduced in Section 2. Asymptotic properties of variance estimators are studied in Section 3. In Section 4, the random group variance estimation method is applied to a data set for illustration. Proofs of the results are provided in the Appendix.

2. The Random Group Method and Its Shortcut

Let $\mathcal{P}$ be a finite population containing units indexed by i and S be a sample taken from $\mathcal{P}$ according to some sampling design. According to the sampling plan, survey weights w_i, $i \in S$, are constructed so that the Horvitz-Thompson type estimator $\hat{Y} = \sum_{i \in S} w_i y_i$ is unbiased (with respect to the repeated sampling) for the population total $Y = \sum_{i \in \mathcal{P}} y_i$, where y_i is a variable (item) of interest.

A replication method starts with the construction of K replicates. When there is no nonresponse, the data set (including the survey weight) is $\{(y_i, w_i),\ i \in S\}$. The kth replicate is then $\left\{\left(y_i, w_i^{(k)}\right),\ i \in S\right\}$, $k = 1, \ldots, K$. Note that only the survey weight w_i is changed to $w_i^{(k)}$. For the random group method (Wolter 2007), we randomly form K groups

of the same or nearly the same size. The union of the K groups is S and the kth replicate is $\{(y_i, g_i^{(k)} w_i), i \in S\}$, where

$$g_i^{(k)} = \begin{cases} K & \text{if } i \text{ is in group } k \\ 0 & \text{if } i \text{ is not in group } k \end{cases}$$

Let $\hat{Y}^{(k)} = \sum_{i \in S} g_i^{(k)} w_i y_i$, $k = 1, \ldots, K$. The random group variance estimator for $\hat{Y}$ is

$$v = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{Y}^{(k)} - \frac{1}{K} \sum_{j=1}^{K} \hat{Y}^{(j)} \right)^2$$

Forming random groups has to ensure that the original sampling design is reflected within the groups (Wolter 2007). If the original sampling design is one-stage stratified sampling with H strata, for example, then each group should contain all H strata. If cluster sampling (either one-stage or multi-stage) is used, then clusters should be considered as units in forming random groups and the random group variance estimator should be used when the number of clusters is large.

2.1. Imputation and Reimputation

Let $\mathcal{R} = \{i \in S, i \text{ is a respondent}\}$ and $\mathcal{N} = \{j \in S, j \text{ is a nonrespondent}\}$. For simplicity, we call y_i a respondent when $i \in \mathcal{R}$ and y_j a nonrespondent when $j \in \mathcal{N}$. When there are nonrespondents, we usually create $L \geq 1$ imputation cells such that the nonresponse probability in each imputation cell is nearly constant and then apply the random hot deck within each imputation cell. More specifically, a nonrespondent in an imputation cell is imputed by a respondent y_i in the same imputation cell selected with probability proportional to w_i. After imputation, treating imputed values as observed data leads to the following estimator of the population total Y:

$$\hat{Y}_I = \sum_{i \in \mathcal{R}} w_i y_i + \sum_{j \in \mathcal{N}} w_j \tilde{y}_j = \sum_{i \in \mathcal{R}} w_i y_i (1 + u_i) \tag{1}$$

where $\tilde{y}_j$ is the imputed value for a nonrespondent y_j, $u_i = \sum_{j \in \mathcal{N}} w_j d_{ij} / w_i$, $d_{ij} = 1$ if the imputed value $\tilde{y}_i = y_i$ and $d_{ij} = 0$ otherwise. Under the assumption that the nonresponse probability is constant within each imputation cell, $\hat{Y}_I$ is consistent and asymptotically normal (see, e.g., Rao and Shao 1992).

Treating imputed values as observed data and applying the formula for the case of no nonresponse leads to a naive variance estimator given by

$$v_I = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{Y}_I^{(k)} - \frac{1}{K} \sum_{j=1}^{K} \hat{Y}_I^{(j)} \right)^2 \tag{2}$$

with

$$\hat{Y}_I^{(k)} = \sum_{i \in \mathcal{R}} g_i^{(k)} w_i y_i + \sum_{j \in \mathcal{N}} g_j^{(k)} w_j \tilde{y}_j$$

However, v_I underestimates the variance of $\hat{Y}_I$, because it treats imputed values as observed data. The reimputation method can be described as follows. An imputed value in

an imputation cell and the kth group is treated as a nonrespondent and is reimputed by a respondent y_i in the same imputation cell and the kth group selected with probability proportional to $g_i^{(k)} w_i$. The resulting estimator of Y based on the kth group is

$$\hat{Y}_{RI}^{(k)} = \sum_{i \in \mathcal{R}} g_i^{(k)} w_i y_i + \sum_{j \in \mathcal{N}} g_j^{(k)} w_j \tilde{y}_j^{(k)}$$

where $\tilde{y}_j^{(k)}$ is the reimputed value. The key difference between $\hat{Y}_I^{(k)}$ and $\hat{Y}_{RI}^{(k)}$ is that the former uses the original imputed $\tilde{y}_j$ whereas the latter replaces $\tilde{y}_j$ by a reimputed value $\tilde{y}_j^{(k)}$ using the respondents in group k. The random group variance estimator with reimputation is given by

$$v_{RI} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{Y}_{RI}^{(k)} - \frac{1}{K} \sum_{j=1}^{K} \hat{Y}_{RI}^{(j)} \right)^2 \tag{3}$$

2.2. Shortcut

Since reimputation has to be carried out for every replicate, the computation of v_{RI} in (3) can be quite complicated. A shortcut proposed in Moore (2006) can be described as follows. Note that each sampled unit $i \in S$ is associated with a "group label" $g_i^{(k)}$ for forming the random groups. Instead of reimputing every replicate, we "impute" the group label $g_j^{(k)}$ associated with a nonrespondent y_j by $\tilde{g}_j^{(k)}$, the group label of the respondent used to impute y_j. Imputing the group label associated with a nonrespondent alters each replicate so that replicates have different sizes. This creates more variation among replicate estimators, which results in a variance estimator larger than the naive estimator v_I. The shortcut replicate estimator based on the kth group is

$$\hat{Y}_S^{(k)} = \sum_{i \in \mathcal{R}} g_i^{(k)} w_i y_i + \sum_{j \in \mathcal{N}} \tilde{g}_j^{(k)} w_j \tilde{y}_j$$

which is different from $\hat{Y}_I^{(k)}$ since $g_j^{(k)}$ in $\hat{Y}_I^{(k)}$ is replaced by $\tilde{g}_j^{(k)}$. The shortcut random group variance estimator is

$$v_S = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\hat{Y}_S^{(k)} - \frac{1}{K} \sum_{j=1}^{K} \hat{Y}_S^{(j)} \right)^2 \tag{4}$$

Note that

$$\hat{Y}_S^{(k)} = \sum_{i \in \mathcal{R}} g_i^{(k)} w_i y_i + (1 + u_i)$$

For reimputation,

$$\hat{Y}_{RI}^{(k)} = \sum_{i \in \mathcal{R}} g_i^{(k)} w_i y_i + \left(1 + u_i^{(k)}\right)$$

where $u_i^{(k)} = \sum_{j \in \mathcal{N}} K^{-1} g_j^{(k)} w_j d_{ij}^{(k)} / w_i$, $d_{ij}^{(k)} = 1$ if $\tilde{y}_j^{(k)} = y_i$, and $d_{ij}^{(k)} = 0$ otherwise. Thus, $\hat{Y}_{RI}^{(k)}$ is computationally simpler than $\hat{Y}_{RI}^{(k)}$ because we do not need to compute $u_i^{(k)}$ for every k or the reimputed values $\tilde{y}_j^{(k)}$.

Unfortunately, v_S may be inconsistent and seriously biased. To see this, we consider the special case of simple random sampling. Under simple random sampling, $w_i = N/n$, where n and N are respectively the sample size and population size, and $u_i = d_i$, the number of times respondent y_i is used to impute nonrespondents. If y_i is equal to a constant c for all $i \in \mathcal{P}$, then we can estimate $Y = cN$ perfectly by $\hat{Y}_I = cN$ (as long as we have at least one observed value) and the asymptotic variance of $\hat{Y}_I$ is 0. In this case, both v_I in (2) and v_{RI} in (3) are 0, since $\hat{Y}_I^{(k)} = \hat{Y}_{RI}^{(k)} = cN$. On the other hand, a straightforward calculation shows that $\hat{Y}_S^{(k)} = (cN/n)\sum_{i\in\mathcal{R}} g_i^{(k)}(1 + d_i)$ and

$$v_S = \frac{c^2N^2}{n^2K(K-1)}\sum_{K=1}^{K}\left(\sum_{i\in\mathcal{R}} g_i^{(k)}(1+d_i) - n\right)^2$$

which is not 0 and can be arbitrarily large when c^2 is arbitrarily large. The problem is caused by the fact that, when $y_i = c$ for all i, $\hat{Y}_S^{(k)}$ is not a prefect estimator, i.e.,$\sum_{i\in\mathcal{R}} g_i^{(k)}(1 + d_i) \neq n$.

Thus, we propose an adjustment to force the shortcut replicate estimator to be prefect in the special case of $y_i = c$ for all i. Note that

$$\hat{Y}_S^{(k)} = \frac{N}{n}\sum_{l=1}^{L}\sum_{i\in\mathcal{R}_l} g_i^{(k)} y_i(1 + d_i)$$

where $\mathcal{R}_l$ is the set of respondents in imputation cell l, $l = 1, \ldots, L$. Our adjustment is to divide each term in the previous sum by a factor

$$a_{k,l} = \frac{1}{n_l}\sum_{i\in\mathcal{R}_l} g_i^{(k)}(1 + d_i)$$

where n_l is the sample size for imputation cell l, i.e., the adjusted shortcut replicate estimator is

$$\tilde{Y}_S^{(k)} = \frac{N}{n}\sum_{l=1}^{L}\sum_{i\in\mathcal{R}_l}\frac{g_i^{(k)} y_i(1 + d_i)}{a_{k,l}}$$

In the special case of $y_i = c$ for all i,

$$\tilde{Y}_S^{(k)} = \frac{N}{n}\sum_{l=1}^{L} cn_l = cN$$

Although this adjustment is derived under the special case of simple random sampling, our result in Section 3 shows that the same adjustment produces a consistent and asymptotically unbiased shortcut random group variance estimator in the case of unequal w_i's.

Thus, in general, we define

$$\tilde{Y}_S^{(k)} = \sum_{l=1}^{L}\sum_{i\in\mathcal{R}_l}\frac{g_i^{(k)} w_i y_i(1 + u_i)}{a_{k,l}}$$

and the adjusted shortcut random group variance estimator

$$v_{AS} = \frac{1}{K(K-1)} \sum_{k=1}^{K} \left(\tilde{Y}_S^{(k)} - \frac{1}{K} \sum_{j=1}^{K} \tilde{Y}_S^{(j)} \right)^2 \tag{5}$$

3. Asymptotic Unbiasedness and Consistency

We consider the asymptotic setting in which S is sampled from a sequence of finite populations, there is a fixed number of imputation cells, the response probability within each imputation cell is constant, and the number of sampled units in any imputation cell tends to ∞. The estimator $\hat{Y}_I$ in (1) is well-defined as long as there is at least one respondent in each imputation cell. The estimators $\hat{Y}_I^{(k)}$, $\hat{Y}_{RI}^{(k)}$, and $\hat{Y}_S^{(k)}$ are well-defined as long as there is at least one respondent in each imputation cell and group k. If the response probability in each imputation cell is positive, then asymptotically these conditions are satisfied and $\hat{Y}_I$, $\hat{Y}_I^{(k)}$, $\hat{Y}_{RI}^{(k)}$, and $\hat{Y}_S^{(k)}$ are well-defined. From now on, all the expectations and variances are calculated conditional on the event that there is at least one respondent in each imputation cell and group. Thus, there are asymptotic expectations and variances.

For simplicity, we now assume that there is a single imputation cell ($L = 1$) and $\mathcal{P}(i \in \mathcal{R}) = \pi$ for all $i \in \mathcal{P}$. Since $\hat{Y}_I = \sum_{l=1}^{L} \hat{Y}_I^{(l)}$, where $\hat{Y}_I^{(l)} = \sum_{i \in \mathcal{R}_l} w_i y_i (1 + u_i)$ and $\mathcal{R}_l$ is the set of respondents in imputation cell l, all the results obtained are valid for the case of multiple imputation cells with a fixed L when $\hat{Y}_I^{(l)}$'s are independent or asymptotically independent.

3.1. Asymptotic Unbiasedness under Simple Random Sampling without Replacement

Under simple random sampling without replacement, $w_i = N/n$ for all i and u_i in the formula of $\hat{Y}_I$ equals to d_i, the number of times unit i is used as a donor. Furthermore, $P(d_{ij} = 1 | i \in \mathcal{R}, j \in \mathcal{N}, r) = 1/r$ and $P\left(d_{ij}^{(k)} = 1 \middle| i \in \mathcal{R}, j \in \mathcal{N}, r_k\right) = g_i^{(k)}/(Kr_k)$, where r and r_k are random variables representing the numbers of respondents in the sample and in group k, respectively. Let δ_i be the indicator of whether unit i is a respondent, $\boldsymbol{a} = \left\{(N/n)\delta_i(1 + d_i),\ i \in \mathcal{P}\right\}'$, and $y = (y_i,\ i \in \mathcal{P})'$. Then $\hat{Y}_I = \boldsymbol{a}'y$ and

$$\mathrm{Var}(\hat{Y}_I) = \boldsymbol{y}' E(\boldsymbol{a}\boldsymbol{a}')y - y' E(\boldsymbol{a}) E(a')y$$

Because missing is completely at random and imputation is random hot deck, components of $\boldsymbol{a}$ have the same distribution and, hence, $E(\boldsymbol{a}\boldsymbol{a}')$ and $E(\boldsymbol{a})E(\boldsymbol{a}')$ are linear combinations of the $N \times N$ identity matrix and the $N \times N$ matrix with all ones. Also, $E(\boldsymbol{a}\boldsymbol{a}')$ and $E(\boldsymbol{a})E(\boldsymbol{a}')$ do not depend on the y_i's. Then

$$\mathrm{Var}(\hat{Y}_I) = \alpha Y^2 + \beta \sum_{i \in \mathcal{P}} y_i^2 \tag{6}$$

where α and β are two constants not depending on the y_i's. We now determine values of α and β using two particular sets of y_i's. When $y_i = 1$ for all $i \in \mathcal{P}$, $\mathrm{Var}(\hat{Y}_I) = 0$ and result (6) becomes

$$\alpha N^2 + \beta N = 0 \tag{7}$$

If all y_i's are 0 except that $y_1 = 1$, then $\hat{Y}_I = Nn^{-1}\delta_1(1 + d_1)$, $Y = 1$, $\sum_{i\in\mathcal{P}} y_i^2 = 1$, and result (6) becomes

$$\alpha + \beta = \mathrm{Var}\,(Nn^{-1}\delta_1(1 + d_1))$$

Using result (21) in the Appendix, we obtain that

$$\alpha + \beta = \frac{N}{n}\left(1 - \pi + \frac{1}{\pi}\right) + O(1) + O(N/n^2) \tag{8}$$

where π is the response probability. From (6), (7), and (8), we obtain

$$\mathrm{Var}\,(\hat{Y}_I) = \frac{N^2}{n}\left(1 - \pi + \frac{1}{\pi}\right)S_N^2 + O(N) + O(N^2/n^2) \tag{9}$$

where

$$S_N^2 = \frac{1}{N}\sum_{i\in\mathcal{P}}(y_i - Y/N)^2$$

For $l = 1, \ldots, K$, let $a_l = \left\{(NK/n)\delta_{il}(1 + u_i^{(l)}),\ i \in \mathcal{P}\right\}'$ and let $\delta_{il} = 1$ if y_i is a respondent in group l and $\delta_{il} = 0$ otherwise. For the random group variance estimator with reimputation, by the exchangeability of group assignment

$$E(v_{RI}) = \frac{E\left(\hat{Y}_{RI}^{(k)}\right)^2 - E\left(\hat{Y}_{RI}^{(k)}\hat{Y}_{RI}^{(h)}\right)}{K} = \frac{y'E(\boldsymbol{a}_k\boldsymbol{a}_k')\boldsymbol{y} - \boldsymbol{y}'E(\boldsymbol{a}_k\boldsymbol{a}_h')y}{K} = \gamma Y^2 + \eta\sum_{i\in\mathcal{P}} y_i^2$$

where $k \neq h$, $k, h \in \{1, \ldots, K\}$, and γ and η are two constants not depending on y_i's. The last equality follows from the fact that $E(\boldsymbol{a}_k\boldsymbol{a}'_k)$ and $E(\boldsymbol{a}_k\boldsymbol{a}'_h)$ $(k \neq h)$ are linear combinations of the $N \times N$ identity matrix and the $N \times N$ matrix with all ones. Again, by considering $y_i = 1$ for all $i \in \mathcal{P}$, we obtain

$$\gamma N^2 + \eta N = 0 \tag{10}$$

by considering $y_i = 0$ for all i except $y_1 = 1$, we obtain that

$$\gamma + \eta = \frac{N}{n}\left(1 - \pi + \frac{1}{\pi}\right) + O\left(\frac{NK}{n^2}\right) \tag{11}$$

From (10) and (11), we obtain

$$E(v_{RI}) = \frac{N^2}{n}\left(1 - \pi + \frac{1}{\pi}\right)S_N^2 + O\left(\frac{N^2K}{n^2}\right) \tag{12}$$

Using the same technique, we obtain the following results for the two shortcut random group variance estimators

$$E(v_{AS}) = \frac{N^2}{n}\left(1 - \pi + \frac{1}{\pi}\right)S_N^2 + O\left(\frac{N^2K}{n^2}\right) \tag{13}$$

$$E(v_S) = \frac{N^2}{n}\left(1 - \pi + \frac{1}{\pi}\right)S_N^2 + \left(-\pi + \frac{1}{\pi}\right)\frac{Y^2}{n} + O\left(\frac{N^2}{n^2}\right) \tag{14}$$

Combining (9), (12), (13), and (14), we have the following result.

Theorem 1. *Suppose that S_N^2 and $\bar{Y} = Y/N$ are bounded. Then, under simple random sampling without replacement, v_{RI} and v_{AS} are asymptotically unbiased in the sense that*

$$\frac{n}{N^2}[E(v_{RI}) - \mathrm{Var}(\hat{Y}_I)] \to 0 \quad \textit{and} \quad \frac{n}{N^2}[E(v_{AS}) - \mathrm{Var}(\hat{Y}_I)] \to 0$$

when $n \to \infty$, $n/K \to \infty$, and $n/N \to 0$. The unadjusted shortcut random group variance estimator v_S is asymptotically biased in the sense that

$$\frac{n}{N^2}[E(v_S) - \mathrm{Var}(\hat{Y}_I)] = \left(-\pi + \frac{1}{\pi}\right)\bar{Y}^2 + O(1/n) + O(1/N)$$

where $\bar{Y} = Y/N$.

3.2. Asymptotic Unbiasedness under Unequal Probability Sampling with Replacement

For unequal probability sampling, we consider sampling with replacement. Under unequal probability sampling without replacement, the derivation of a valid variance estimator may be too difficult or even impossible and, hence, at the stage of variance estimation, we treat S as a sample obtained with replacement. This may overestimate the variance when the original sampling design is without replacement and sampling fractions are not negligible. But it is often used in practice because of its simplicity when no valid variance estimator for without replacement sampling is available.

Let $w_i = 1/(np_i)$, $S_N^2 = \sum_{i\in\mathcal{P}} p_i\{y_i/(Np_i) - \bar{Y}\}^2$, $P(d_{ij} = 1|\mathcal{R}, i \in \mathcal{R}, j \in \mathcal{N}) = w_i/\sum_{i\in\mathcal{R}} w_i$ and $P\left(d_{ij}^{(k)} = 1|\mathcal{R}, i \in \mathcal{R}, j \in \mathcal{N}\right) = g_i^{(k)}w_i/\sum_{i\in\mathcal{R}}\left(g_i^{(k)}w_i\right)$, $i, j \in \mathcal{P}$ where p_i is the probability that y_i is sampled in each draw. We assume that

C1. $M_1N/n < w_i < M_2N/n$ for all $i \in \mathcal{P}$, where M_1 and M_2 are constants;
C2. $|\bar{Y}| = |Y/N| < M$ and $S_N^2 < M$ for a constant $M > 0$.

Condition C1 means that there is no unit that has extremely large or small inclusion probability. Under Conditions C1 and C2, it is shown in the Appendix that

$$\frac{n}{N^2}\mathrm{Var}(\hat{Y}_I) = \frac{S_N^2}{\pi} + (1-\pi)\left(\bar{Y}^2\left[\psi + \frac{1}{\pi}\{(1-2\pi)\beta + 1\}\right]\right.$$
$$\left. + \bar{Y}\left\{\frac{2\varphi}{\pi} + 2\alpha(\beta - 1)\right\} + \frac{1}{N}\sum_{i\in\mathcal{P}} y_i^2\right) + O\left(\frac{1}{n}\right) \tag{15}$$

where $\alpha = \sum_{i\in\mathcal{P}} y_i p_i$, $\beta = N^{-2}\sum_{i\in\mathcal{P}} 1/p_i$, $\varphi = N^{-2}\sum_{i\in\mathcal{P}} y_i/p_i$, and $\psi = N\sum_{i\in\mathcal{P}} p_i^2$. Under C1, α, β, φ, and ψ are bounded. Hence, C1–C2 and result (15) show that $\mathrm{Var}(\hat{Y}_I)$ is $O(N^2/n)$.

Without loss of generality we assume that n/K is an integer, which is the "sample size" of each group (replicate). With reimputation, it is clear that, for a fixed group, $\hat{Y}_{RI}^{(k)}$ behaves like $\hat{Y}_I$ with a sample size n/K. In particular, $K^{-1}\mathrm{Var}\left(\hat{Y}_{RI}^{(k)}\right) = \mathrm{Var}(\hat{Y}_{RI}) + o(N^2/n)$.

It follows from exchangeability that, for $k \neq h$,

$$E(v_{RI}) = \frac{1}{K}\left[E\left(\hat{Y}_{RI}^{(k)}\right)^2 - E\left(\hat{Y}_{RI}^{(k)}\hat{Y}_{RI}^{(h)}\right)\right]$$

Let E_G be the expectation taken under a fixed assignment of groups. Then, for $k \neq h$,

$$E\left(\hat{Y}_{RI}^{(k)}\hat{Y}_{RI}^{(h)}\right) = E_G\left(\hat{Y}_{RI}^{(k)}\hat{Y}_{RI}^{(h)}\right) = E_G\left(\hat{Y}_{RI}^{(k)}\right)E_G\left(\hat{Y}_{RI}^{(h)}\right) = E\left(\hat{Y}_{RI}^{(k)}\right)E\left(\hat{Y}_{RI}^{(h)}\right),$$

where the first and the last equalities are because of the exchangeability and the second equality is because data in different groups are independent and the imputation is also independently carried out in different groups. Hence, $\hat{Y}_{RI}^{(k)}$ and $\hat{Y}_{RI}^{(h)}$ are uncorrelated and

$$E(v_{RI}) = \frac{1}{K}\mathrm{Var}\left(\hat{Y}_{RI}^{(k)}\right) = \mathrm{Var}\,(\hat{Y}_I) + o(N^2/n),$$

showing that v_{RI} is asymptotically unbiased.

For the adjusted shortcut estimators, $\tilde{Y}_S^{(k)}$ and $\tilde{Y}_S^{(h)}$ are not uncorrelated but the correlation converges to 0, i.e.,

$$\frac{n}{N^2K}\mathrm{Cov}\left(\tilde{Y}_S^{(k)}, \tilde{Y}_S^{(h)}\right) = O(K^{-1}) \to 0, \quad k \neq h\,, \tag{16}$$

when $K \to \infty$ (see the Appendix). Then, for the adjusted shortcut random group variance estimator v_{AS} in (5),

$$\begin{aligned}\frac{n}{N^2}E(v_{AS}) &= \frac{n}{N^2K}\left[E\left(\tilde{Y}_S^{(k)}\right)^2 - E\left(\tilde{Y}_S^{(k)}\tilde{Y}_S^{(h)}\right)\right] \\ &= \frac{n}{N^2K}\left[E\left(\tilde{Y}_S^{(k)}\right)^2 - E\left(\tilde{Y}_S^{(k)}\right)E\left(\tilde{Y}_S^{(h)}\right)\right] + O(K^{-1}) \\ &= \frac{n}{N^2K}\mathrm{Var}\left(\tilde{Y}_S^{(k)}\right) + O(K^{-1}).\end{aligned} \tag{17}$$

The form of $\mathrm{Var}\left(\tilde{Y}_S^{(k)}\right)$ is not simple because of the shortcut and adjustment. In the Appendix, we derive a formula for $\mathrm{Var}\left(\tilde{Y}_S^{(k)}\right)$, which leads to the following result.

Theorem 2. *Under C1–C2, both v_{RI} and v_{AS} are asymptotically unbiased, i.e.,*

$$\frac{n}{N^2}[E(v_{RI}) - \mathrm{Var}\,(\hat{Y}_I)] \to 0 \quad \text{and} \quad \frac{n}{N^2}[E(v_{AS}) - \mathrm{Var}\,(\hat{Y}_I)] \to 0$$

when $n/K \to \infty$ and $K \to \infty$.

3.3. Consistency

To establish the consistency of v_{RI} and v_{AS}, we focus on v_{AS} because the discussion for v_{RI} is similar. By exchangeability, $E\left(\tilde{Y}_S^{(k)}\right)$ does not depend on $k\,(k = 1, \ldots, K)$. Letting

$\theta = E\left(\tilde{Y}_S^{(k)}\right)$ and $\bar{\theta} = K^{-1}\sum_{k=1}^{K}\tilde{Y}_S^{(k)}$, we obtain that

$$\frac{n}{N^2}v_{AS} = \frac{n}{N^2K(K-1)}\sum_{k=1}^{K}\left(\tilde{Y}_S^{(k)} - \theta\right)^2 - \frac{n}{N^2(K-1)}(\theta - \bar{\theta})^2. \tag{18}$$

The second term on the right-hand side of (18) converges to 0 in probability (see the proof of Theorem 3 in the Appendix). The first term on the right-hand side of (18) involves a sum of variables $\left(\tilde{Y}_S^{(k)} - \theta\right)^2$, $k = 1, \ldots, K$. If these variables are independent, then the consistency of v_{AS} follows from Khintchine's law of large numbers and the fact that $nN^{-2}K^{-1}E\left(\tilde{Y}_S^{(k)} - \theta\right)^2 = nN^{-2}\mathrm{Var}(\hat{Y}_I) + O(n^{-1}K) + O(n^{-1})$ (see the proof of Theorem 2 in the Appendix). Although $\left(\tilde{Y}_S^{(k)} - \theta\right)^2$, $k = 1, \ldots, K$, are not independent, the correlation among them converges to 0 and we can use a modified Khintchine's law of large numbers (Lemma 1 stated in the Appendix) to establish the consistency of v_{RI} and v_{AS}. The proof of the following result is given in the Appendix.

Theorem 3. *When $n/K \to \infty$ and $K \to \infty$,*

$$\frac{n}{N^2}[v_{RI} - \mathrm{Var}(\hat{Y}_I)] \to_p 0 \quad \textit{and} \quad \frac{n}{N^2}[v_{AS} - \mathrm{Var}(\hat{Y}_I)] \to_p 0, \tag{19}$$

where $\to_p$ is convergence in probability.

The asymptotic results require both n/K and K (the number of groups) to be large. In applications, constructing groups with a large K may not be easy when n is not very large. We can use the following idea proposed by Rao and Shao (1996). Suppose that we independently construct R sets of groups with the rth set containing K groups that result in the variance estimator $v_{AS}^{(r)}$ according to formula (5). Then, we use $v_{AS} = R^{-1}\sum_{r=1}^{R}v_{AS}^{(r)}$ as the adjusted shortcut variance estimator. This estimator is approximately as good as the v_{AS} with KR groups. v_{RI} with R sets of groups can be similarly obtained.

4. Empirical Results for the National Immunization Survey

The 2007 National Immunization Survey is used here to illustrate the application of the random group variance estimators and to examine their finite sample performance. This survey was conducted jointly by the National Center for Immunization and Respiratory Diseases and the National Center for Health Statistics (see http://www.cdc.gov/nis/data_files.htm). The survey design is stratified unequal probability sampling with 64 geological strata, each of which is used as an imputation cell. Duration of breast feeding (in days) is chosen to be the variable y of interest, because it has appreciable nonresponse. The population size N is 6,025,053 and the total sample size is 24,807. Ranges of the population sizes, sample sizes, sampling fractions, nonresponse rates, estimated stratum totals, and estimated stratum standard deviations over 64 strata are listed in Table 1. Nonrespondents were imputed by random hot deck imputation within each imputation cell. Based on the imputed data set, four random group variance estimators, v_I, v_{RI}, v_S, and v_{AS}, were computed for the estimated population total. Ten random groups were constructed within each imputation cell and grouping was independently repeated ten times (see the discussion at the end of Section 3). The results are listed in Table 2. From Table 2, v_I is much smaller than other variance estimates (e.g., 32% smaller than v_{RI}),

Table 1. Ranges of some quantities and estimates over 64 strata for the duration of breast feeding from the 2007 National Immunization Survey

Quantity	Range over 64 strata
Population size	9,766 ~ 506,300
Sample size	244 ~ 519
Sampling fraction	0.062% ~ 4.32%
Nonresponse rate	13.48% ~ 46.86%
Estimated total	$2.197 \times 10^6 \sim 1.467 \times 10^8$
Estimated standard deviation	142.05 ~ 199.96

which indicates its underestimation. On the other hand, the shortcut estimate v_S is about 40% larger than v_{RI}, which suggests the overestimation of v_S as indicated by our analysis in Section 2. The adjusted shortcut estimate v_{AS} is close to v_{RI}, both of which are shown to be asymptotically unbiased in Section 3.

To examine the finite sample performance of the four random group variance estimators, we conducted a simulation study based on a population generated using the observed data. Our simulation procedure is as follows.

1. Population. Within each stratum $h = 1, \ldots, 64$, population y-values are generated by first taking a probability proportional to survey weight sample of size equaling the original population size of stratum h with replacement from the respondents in the hth stratum, and then adding a random noise to each generated population value. The random noises have mean 0 and standard deviation equal to 10^{-3} times the observed standard deviation within each stratum and are independent within and across strata.
2. Within the hth population stratum generated in Step 1, draw a sample with replacement of the original sample size. The inclusion probability is proportional to the inverse of the original survey weight. Independent samples are obtained for $h = 1, \ldots, 64$.
3. Within each stratum, generate nonrespondents (missing completely at random) with the observed nonresponse rate. Respondents in different strata are obtained independently.
4. Perform random hot deck imputation for nonrespondents within each stratum and independently across strata.
5. Compute the estimated total $\hat{Y}_I$ and random group variance estimates v_I, v_S, v_{AS}, and v_{RI} with $K = 10$ and $R = 10$.
6. Repeat steps 2–5 for 1,000 times. Compute V = the sample variance of the 1,000 $\hat{Y}_I$'s. Compute the averages of v_I, v_S, v_{AS}, and v_{RI} over 1,000 simulations, and their standard errors based on 1,000 simulations.

Table 2. Random group variance estimates for the estimated total of the duration of breast feeding from the 2007 National Immunization Survey

$\hat{Y}_I$	v_I	v_S	v_{AS}	v_{RI}
1.36×10^9	2.25×10^{14}	4.65×10^{14}	3.14×10^{14}	3.32×10^{14}

Table 3. Simulation results of the random group variance estimators for the estimated total of the duration of breast feeding from the 2007 National Immunization Survey

	v_I	v_S	v_{AS}	v_{RI}	V
Mean	2.0819	3.9796	2.7378	2.7603	2.8233
Standard error	0.0110	0.0219	0.0152	0.0142	

All numbers have been divided by 10^{14}
V: simulation sample variance of $\hat{Y}_I$'s

The results are tabulated in Table 3. In theory, the simulation variance V is close to the true variance when the simulation size is large. From Table 3, it is clear that v_{AS}, v_{RI}, and V are very close to each other; the naive estimate v_I is about 26% smaller than V; and v_S is 41% larger than V. These results are consistent with our theory in Section 3 and support our conclusions based on the results in Table 2.

Appendix

Since y_i, $i \in \mathcal{P}$, are sampled with replacement in Sections 3.2–3.3, to simplify the proofs, throughout, we view $\hat{Y}_I$, v_I, v_S, v_{AS} and v_{RI} as functions of $Y_1, \ldots, Y_n$ and $W_1, \ldots, W_n$, where $Y_l\,(l = 1, \ldots, n)$ denotes the random outcome from the lth sampling, which has discrete uniform distribution on $\mathcal{P}$, and W_l is the associating weight of Y_l. The pairs (Y, W_l), $l = 1, \ldots, n$, are independent of each other. Correspondingly, let $\mathcal{R} = \{l\text{: } Y_l \text{ is a respondent}\}$, and let $\mathcal{N} = \{m\text{: } Y_m \text{ is a nonrespondent}\}$, but S is unchanged. Now, we introduce some additional notation. Define $U_l^{(k)} = (1 + u_l)g_l^{(k)}$, $U_l = (1 + u_l)$ and $Z_l = Y_l W_l$, where $u_l = \sum_{m \in \mathcal{N}} W_m d_{lm}/W_l$ and $d_{lm} = 1$ if the imputed value $\tilde{Y}_m = Y_l$, 0 otherwise. Let $G = \{g_l^{(k)}, l = 1, \ldots, n, k = 1, \ldots, K\}$ denote the group assignment and $D = \{d_{lm}, l \in \mathcal{R}, m \in \mathcal{N}\}$ denote the imputation. Let $r = r_1 + \cdots + r_K$, where $r_k\,(k = 1, \ldots, K)$ denotes the number of respondents in the kth group. For any pair of event or variable B and C, let $E_{B|C}$ and $\mathrm{Var}_{B|C}$ be the conditional expectation and variance taken with respect to B given C.

Proof of (8) Since $E[\delta_1(1 + d_1)|\delta_1 = 0] = 0$, $E[\delta_1^2(1 + d_1)^2|\delta_1 = 0] = 0$, and $P(\delta_1 = 1) = nN^{-1}\pi$,

$$\mathrm{Var}(n^{-1}N\delta_1(1+d_1)) = n^{-2}N^2E[\delta_1^2(1+d_1)^2] - n^{-2}N^2\{E[\delta_1(1+d_1)]\}^2$$
$$= n^{-1}N^1\pi E[\delta_1^2(1+d_1)^2|\delta_1=1] - \pi^2\{E[\delta_1(1+d_1)|\delta_1=1]\}^2$$

Recall that d_l is the number of times y_1 is used as a donor to impute missing values. Conditional on $\delta_1 = 1$ and the number of respondents $r > 0$, d_1 has the binomial distribution with size $n - r$ and probability r^{-1}. Then, Var $(n^{-1}N\delta_1(1 + d_1))$ equals

$$n^{-1}N\pi E[1 + r^{-1}(n-r)\{2 + r^{-1}(n-1)\}|\delta_1 = 1] - \pi^2[E(r^{-1}n|\delta_1 = 1)]^2 \tag{20}$$

Since we consider asymptotic expectations with $r > 0$, as $n \to \infty$, $E(r^{-1}|\delta_1 = 1) = (n\pi)^{-1} + O(n^{-2})$ and $E(r^{-2}|\delta_1 = 1) = (n\pi)^{-2} + O(n^{-3})$. Substituting these results into (20), we obtain that

$$\operatorname{Var}(n^{-1}N\delta_1(1+d_1)) = Nn^{-1}(1-\pi+\pi^{-1}) + O(1) + O(Nn^{-2}) \tag{21}$$

Proof of (15) We show (15) in three steps. First, let

$$A_n = nN^{-2}[E\{\hat{Y}_I - E(\hat{Y}_I)\}^2 - E(\hat{Y}_I - Y)^2] = -nN^{-2}\{E(\hat{Y}_I) - Y\}^2 \tag{22}$$

To simplify (22), we have that $E(\hat{Y}_I) = E\big(\sum_{l\in\mathcal{R}} Z_l U_l\big)$, which equals

$$E\left[\sum_{l\in\mathcal{R}}\big\{E_{S|\mathcal{R}}(Z_l) + Z_l E_{D|S,\mathcal{R}}(u_l)\big\}\right] = E\left\{\frac{rY}{n} + \frac{N(n-r)}{n}E_{S|\mathcal{R}}\left(\frac{\sum_{l\in\mathcal{R}} Z_l}{\sum_{l\in\mathcal{R}} W_l}\right)\right\}$$

$$= E(rn^{-1}Y) + n^{-1}NE\left[(n-r)\left\{\frac{E_{S|\mathcal{R}}\left(\sum_{l\in\mathcal{R}} Z_l\right)}{E_{S|\mathcal{R}}\left(\sum_{l\in\mathcal{R}} W_l\right)} + O_p(nr^{-2})\right\}\right]$$

$$= E(rn^{-1}Y) + n^{-1}NE[(n-r)\{\bar{Y} + O_p(nr^{-2})\}] = Y + O(n^{-1}N)$$

where the third equality is due to the Taylor expansion argument and Conditions C1–C2. Hence, (22) equals $O(n^{-1})$.

Second, let $B_n = nN^{-2}\{E(\hat{Y}_I - Y)^2 - E(\hat{Y}_I - \tilde{Y})^2\}$, which equals

$$nN^{-2}E\big[2E\{\hat{Y}_I(\tilde{Y} - Y)\} + Y^2 - E(\tilde{Y}^2)\big] = B_{n1} + B_{n2}$$

where $\tilde{Y} = n^{-1}Y\sum_{l\in\mathcal{R}}(1+u_l)$, $B_{n1} = 2nN^{-2}E\{\hat{Y}_I(\tilde{Y} - Y)\}$ and $B_{n2} = nN^{-2}\times\{Y^2 - E(\tilde{Y}^2)\}$. By Conditions C1–C2, we have

$$B_{n1} = 2(1-\pi)\bar{Y}\big\{-\pi^{-1}(\bar{Y}+\varphi) + \alpha\beta + \pi^{-1}(2-\pi)\bar{Y}\beta\big\} + O(n^{-1}) \tag{23}$$

$$B_{n2} = -3\pi^{-1}(1-\pi)(\beta-1)\bar{Y}^2 + O(n^{-1}) \tag{24}$$

Third, consider $C_n = n/N^2E(\hat{Y}_1 - \tilde{Y})^2$. We have

$$C_n = \frac{n}{N^2}\left[E\left\{\sum_{l\in\mathcal{R}}(Z_l - n^{-1}Y)^2U_l^2\right\} + E\left\{\sum_{l_1\neq l_2\in\mathcal{R}}(Z_{l_1} - n^{-1}Y)(Z_{l_2} - n^{-1}Y)U_{l_1}U_{l_2}\right\}\right]$$

$$= \pi^{-1}S_N^2 + (1-\pi)\left(N^{-1}\sum_{i=1}^{N} y_i^2 - 2\bar{Y}\alpha + \bar{Y}^2\psi\right) + O(n^{-1}) \tag{25}$$

Since, $nN^{-2}\operatorname{Var}(\hat{Y}_I) = A_n + B_{n1} + B_{n2} + C_n$, by (23)–(25) and the fact $A_n = O(n^{-1})$, (15) is proved.

Proof of (16) For $k = 1, \ldots, K$, let

$$f_k(t) = \{ta_k + (1-t)E(a_k)\}^{-1}\left\{t\hat{Y}_s^{(k)} + (1-t)E\left(\hat{Y}_S^{(k)}\right)\right\}$$

where $0 < t < 1$ with $f_k(1) = \tilde{Y}_S^{(k)}$. By the Taylor expansion argument,

$$f_k(1) = f_k(0) + f_k'(0) + 2^{-1}f_k''(\tilde{t}) \tag{26}$$

where $0 < \tilde{t} < 1$. Then, by the exchangeability of group assignment, $\mathrm{Cov}\{f_k'(0), f_h''(\tilde{t})\} = \mathrm{Cov}\{f_h'(0), f_k''(\tilde{t})\}$ for $k \neq h$, and this, together with the fact that $f_k(0)$ and $f_h(0)$ are constant, gives:

$$\begin{aligned}\mathrm{Cov}\left(\tilde{Y}_S^{(k)}, \tilde{Y}_S^{(h)}\right) &= \mathrm{Cov}\{f_k(0) + f_k'(0) + 2^{-1}f_k''(\tilde{t}), f_h(0) + f_h'(0) + 2^{-1}f_h''(\tilde{t})\} \\ &= \mathrm{Cov}\{f_k'(0), f_h'(0)\} + \mathrm{Cov}\{f_k'(0), f_h''(\tilde{t})\} + 4^{-1}\mathrm{Cov}\{f_k''(\tilde{t}), f_h''(\tilde{t})\}.\end{aligned} \tag{27}$$

To simplify (27), we work on $\mathrm{Cov}\{f_k'(0), f_h'(0)\}$ first. Since $E(a_k) = 1$, we have

$$\begin{aligned}\mathrm{Cov}\{f_k'(0), f_h'(0)\} = \mathrm{Cov}\left(\hat{Y}_S^{(k)}, \hat{Y}_S^{(h)}\right) + \mathrm{Cov}(a_k, a_h) \\ \times\left\{E\left(\hat{Y}_S^{(k)}\right)\right\}^2 - 2\mathrm{Cov}\left(\hat{Y}_S^{(k)}, a_h\right)E\left(\hat{Y}_S^{(k)}\right).\end{aligned} \tag{28}$$

Now, we calculate the three terms of the above separately. For the first term,

$$E\left(\hat{Y}_S^{(k)}\hat{Y}_S^{(h)}\right) = E\left\{E_{G,D|\mathcal{R},S}\left(\sum_{l_1 \neq l_2 \in \mathcal{R}} Z_{l_1}Z_{l_2}U_{l_1}^{(k)}U_{l_2}^{(h)}\right)\right\} \tag{29}$$

where the inside expectation of the right-hand side of (29) can be calculated by taking another conditional expectation:

$$E_{G,D|\mathcal{R},S,r_k,r_h}\left(U_{l_1}^{(k)}U_{l_2}^{(h)}\right) = \frac{r_k r_h K^2}{r(r-1)}\left\{1 + \frac{2\sum_{m\in\mathcal{N}} W_m}{\sum_{l\in\mathcal{R}} W_l} + \frac{\sum_{m_1 \neq m_2 \in \mathcal{N}} W_{m1}W_{m2}}{\left(\sum_{l\in\mathcal{R}} W_l\right)^2}\right\}$$

Plugging this back into (29), taking expectation with respect to S gives:

$$K^2Y^2E[r_k r_h\{r(r-1)\}^{-1}\{1 - (3n-2r)n^{-2} + O_p(nr^{-2})\}]$$

Taking expectation of the above with respect to r_k, r_h given r and then taking expectation with respect to r yields:

$$E\left(\hat{Y}_S^{(k)}\hat{Y}_S^{(h)}\right) = Y^2 + O(n^{-1}N^2) \tag{30}$$

By the exchangeability of group assignment, $E\left(\hat{Y}_S^{(k)}\right)E\left(\hat{Y}_S^{(h)}\right) = \left\{E\left(\hat{Y}_S^{(k)}\right)\right\}^2$, which equals $\{Y + O(n^{-1}N)\}^2$. Then, this together with (30) and Condition C1 gives:

$$\mathrm{Cov}\left(\hat{Y}_S^{(k)}, \hat{Y}_S^{(h)}\right) = Y^2[1 + O(n^{-1}) - \{1 + O(n^{-1})\}^2] = O(n^{-1}N^2) \tag{31}$$

The calculation of the first term of (28) is now completed. Next, for the second term of (28), observe that:

$$\sum_{k\neq h}\mathrm{Cov}(a_k,a_h)+\sum_{k=1}^{K}\mathrm{Cov}(a_k,a_k)=\mathrm{Cov}\left(\sum_{k=1}^{K}a_k,\sum_{k=1}^{K}a_k\right)=\mathrm{Var}\left(\sum_{k=1}^{K}a_k\right)=\mathrm{Var}(K)=0$$

By the exchangeability of group assignment, we have that

$$\mathrm{Cov}(a_k, a_h) = -\{K(K-1)\}^{-1}\sum_{k=1}^{K}\mathrm{Var}\,(a_k) = -(K-1)^{-1}\mathrm{Var}\,(a_1),$$

where $\mathrm{Var}\,(a_1) = \mathrm{Var}\{E_{G|\mathcal{R},D,r_1}(a_1)\} + E\{\mathrm{Var}_{G|\mathcal{R},D,r_1}(a_1)\}$ equals

$$\begin{aligned}
&\mathrm{Var}\,(Kr_1r^{-1}) + E\left[n^{-2}\sum_{l\in\mathcal{R}}(1+d_l)^2\{K^2r_1r^{-1} - (Kr_1r^{-1})^2\}\right.\\
&\qquad\left. + \; n^{-2}\sum_{l_1\neq l_2\in\mathcal{R}}(1+d_{l_1})(1+d_{l_2})\left\{K^2\frac{r_1(r_1-1)}{r(r-1)} - (Kr_1r^{-1})^2\right\}\right]\\
&= E\{(K-1)r^{-1}\} + K^2E\left[\frac{r_1(r-r_1)}{r(r-1)}\left\{n^{-2}\sum_{l\in\mathcal{R}}(1+d_l)^2 - r^{-1}\right\}\right] = O(n^{-1}K)
\end{aligned}$$

Then, the second term in (28),

$$\mathrm{Cov}(a_k, a_h)\left\{E\left(\hat{Y}_S^{(k)}\right)\right\}^2 = O(n^{-1}N^2) \tag{32}$$

For the third term in (28), observe

$$\sum_{k\neq h}\mathrm{Cov}\left(\hat{Y}_S^{(k)},a_h\right)+\sum_{k=1}^{K}\mathrm{Cov}\left(\hat{Y}_S^{(k)},a_k\right)=\mathrm{Cov}\left(\sum_{k=1}^{K}\hat{Y}_S^{(k)},\sum_{k=1}^{K}a_k\right)=\mathrm{Cov}\left(\sum_{k=1}^{K}\hat{Y}_S^{(k)},K\right)=0$$

By the exchangeability of group assignment, $\mathrm{Cov}\left(\hat{Y}_S,a_h\right)=(1-K)^{-1}\mathrm{Cov}\left(\hat{Y}_S^{(1)},a_1\right)$ for $k \neq h$, where

$$\begin{aligned}
\mathrm{Cov}\left(\hat{Y}_S^{(1)}, a_1\right) &= \mathrm{Cov}\left\{E_{G|\mathcal{R},S,D,r_1}\left(\hat{Y}_S^{(1)}\right), E_{G|\mathcal{R},S,D,r_1}(a_1)\right\} + E\left\{\mathrm{Cov}_{G|\mathcal{R},S,D,r_1}\left(\hat{Y}_S^{(1)}, a_1\right)\right\} \\
&= \mathrm{Cov}\left(Kr_1 r^{-1}, Kr_1 r^{-1}\sum_{l\in\mathcal{R}} Z_l U_l\right) + E\left\{\sum_{l\in\mathcal{R}} Z_l \mathrm{Cov}_{G|\mathcal{R},S,D,r_1}\left(U_l^{(1)}, a_1\right)\right\} \\
&= E\left\{\sum_{l_1,l_2\in\mathcal{R}} n^{-1} Z_{l_1} U_{l_1}(1+d_{l_2})\mathrm{Cov}_{G|\mathcal{R},S,D,r_1}\left(g_{l_1}^{(1)}, g_{l_2}^{(1)}\right)\right\} + Y(K-1)E\{r^{-1} + O(nr^{-3})\} \\
&= E\left[K^2 \frac{r_1(r-r_1)}{r(r-1)}\sum_{l\in R} Z_l U_l\{n^{-1}(1+d_l) - r^{-1}\}\right] + Y(K-1)E\{r^{-1} + O(nr^{-3})\} \\
&= O(n^{-1}NK) + O(n^{-1}N).
\end{aligned}$$

Hence, the third term of (28),

$$\mathrm{Cov}\left(\hat{Y}_S^{(k)}, a_h\right) E\left(\hat{Y}_S^{(k)}\right) = O(n^{-1}N^2). \tag{33}$$

With the fact that $E\left(\hat{Y}_S^{(k)}\right) = O(N)$ combining (31), (32) and (33) gives, the first term in (27),

$$\mathrm{Cov}\{f_k'(0), f_h'(0)\} = O(n^{-1}N^2).$$

Following similar argument, it can be shown that the remaining terms of (27) are of the same order as the first term. This implies,

$$nN^{-2}K^{-1}\mathrm{Cov}\left(\tilde{Y}_S^{(k)}, \tilde{Y}_S^{(h)}\right) = O(K^{-1}).$$

Therefore, (16) is proved.

Proof of Theorem 2 Observe that $E(v_{RI}) = K^{-1}\left\{E\left(\hat{Y}_{RI}^{(1)}\right)^2 - E\left(\hat{Y}_{RI}^{(1)}\hat{Y}_{RI}^{(2)}\right)\right\}$. Since, for $k \neq h$, $\hat{Y}_{RI}^{(k)}$ and $\hat{Y}_{RI}^{(h)}$ are uncorrelated, $nN^{-2}E(v_{R1}) = nN^{-2}K^{-1}\mathrm{Var}\left(\hat{Y}_{RI}^{(k)}\right)$, which equals

$$\begin{aligned}
&nN^{-2}K^{-1}\left[\mathrm{Var}\left\{E_{S,\mathcal{R},D|G}\left(\hat{Y}_{RI}^{(k)}\right)\right\} + E\left\{\mathrm{Var}_{S,\mathcal{R},D|G}\left(\hat{Y}_{RI}^{(k)}\right)\right\}\right] \\
&\quad = nN^{-2}K^{-1}\mathrm{Var}_{S,\mathcal{R},D|G}\left\{\sum_{l\in R_k} Y_l W_l^*\left(1+u_l^{(k)}\right)\right\},
\end{aligned}$$

where $W_l^* = KW_l$ and the equality follows from the fact that $E_{S,\mathcal{R},D|G}\left(\hat{Y}_{RI}^{(k)}\right)$ and $\mathrm{Var}_{S,\mathcal{R},D|G}\left(\hat{Y}_{RI}^{(k)}\right)$ do not depend on G. Conditional on group assignment G, we view $\hat{Y}_{RI}^{(k)} = \sum_{l\in R_k} Y_l W_l^*\left(1+u_l^{(k)}\right)$ as a replicate of $\hat{Y}_I = \sum_{l\in R} Y_l W_l(1+u_l)$ calculated based on a random sample of size nK^{-1}. Since (15) holds for all n, we have:

$$nN^{-2}E(v_{RI}) = nN^{-2}K^{-1}\mathrm{Var}\left(\hat{Y}_{RI}^{(k)}\right) = nN^{-2}\mathrm{Var}\left(\hat{Y}_I\right) + O(n^{-1}) + O(n^{-1}K),$$

which means that v_{RI} is asymptotically unbiased when $n^{-1}K \to 0$.

Next, we show the asymptotic unbiasedness *of* v_{AS}. By (17), $nN^{-2}E(v_{As})$ is asymptotically equivalent to $nN^{-2}K^{-1}\mathrm{Var}\left(\hat{Y}_S^{(k)}\right)$. Note that

$$nN^{-2}K^{-1}\left[E\left\{\tilde{Y}_S^{(k)} - E\left(\tilde{Y}_S^{(k)}\right)\right\}^2 - E\left(\tilde{Y}_S^{(k)} - Y\right)^2\right] = O(n^{-1}K). \tag{34}$$

Hence, to show the asymptotic unbiasedness of v_{AS}, it is sufficient to show $E\left(\tilde{Y}_S^{(k)} - Y\right)^2$ is asymptotically unbiased for $\mathrm{Var}(\hat{Y}_I)$. We have

$$\begin{aligned}
&nN^{-2}\left\{K^{-1}E\left(\tilde{Y}_S^{(k)} - Y\right)^2 - E(\hat{Y}_I - Y)^2\right\} \\
&= nN^{-2}K^{-1}E\left\{\sum_{l\in\mathcal{R}} a_k^{-2}\left(Z_l U_l^{(k)}\right)^2 - K\sum_{l\in\mathcal{R}}(Z_l U_l)^2 + a_k^{-2}\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}U_{l_1}^{(k)}Z_{l_2}U_{l_2}^{(k)}\right. \\
&\quad \left. - K\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}U_{l_1}Z_{l_2}U_{l_2} - 2Y\left(a_k^{-1}\sum_{l\in\mathcal{R}} Z_l U_l^{(k)} - K\sum_{l\in\mathcal{R}} Z_l U_l\right) + Y^2(1-K)\right\}
\end{aligned} \tag{35}$$

To simplify the above, we use the Taylor expansion argument similar to (26) to expand the three ratios with a_k or a_k^2 as the denominator. For the first ratio $a_k^{-2}\left(Z_l U_l^{(k)}\right)^2$, after expanding it to the first order term and taking the expectation, the residual part is already of order $O(n^{-2}N^2K^2)$, which multiplies with the factor $nN^{-2}K^{-1}$ converging to 0 when $n^{-1}K \to 0$. For $a_k^{-2}\sum_{l_1,l_2\in\mathrm{R}} Z_{l_1}U_{l_1}^{(k)}Z_{l_2}U_{l_2}^{(k)}$ and $a_k^{-1}\sum_{l\in\mathrm{R}} Z_l U_l^{(k)}$, in order to obtain a residual part with order $O(n^{-2}N^2K)$, we expand them to high order Taylor series. Details are the following. By Condition C1 and the fact $E_{G,D|\mathcal{R},S}(a_k) = 1$,

$$\begin{aligned}
&E_{G,D|\mathcal{R},S}\left(a_k^{-2}\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}U_{l_1}^{(k)}Z_{l_2}U_{l_2}^{(k)}\right) = E_{G,D|\mathcal{R},S}\left(\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}Z_{l_2}U_{l_1}^{(k)}U_{l_2}^{(k)}\right) \\
&+ 3E_{G,D|\mathcal{R},S}\left(\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}Z_{l_2}U_{l_1}^{(k)}U_{l_2}^{(k)}\right)\mathrm{Var}_{G,D|\mathcal{R},S}(a_k) \\
&- 2E_{G,D|\mathcal{R},S}\left(\sum_{l_1,l_2\in\mathcal{R}} Z_{l_1}Z_{l_2}\right)\mathrm{Cov}_{G,D|\mathcal{R},S}\left(U_{l_1}^{(k)}U_{l_1}^{(k)}, a_k\right) + O_p(r^{-2}K^2N^2)
\end{aligned} \tag{36}$$

and $E_{G,D|\mathcal{R},S}\left(a_k^{-1}\sum_{l\in\mathcal{R}} Z_l U_l^{(k)}\right)$ equals

$$\begin{aligned}
&E_{G,D|\mathcal{R},S}\left(\sum_{l\in\mathcal{R}} Z_l U_l^{(k)}\right) + E_{G,D|\mathcal{R},S}\left(\sum_{l\in\mathcal{R}} Z_l U_l^{(k)}\right)\mathrm{Var}_{G,D|\mathcal{R},S}(a_k) \\
&- E_{G,D|\mathcal{R},S}\left(\sum_{l\in R} Z_l\right)\mathrm{Cov}_{G,D|\mathcal{R},S}\left(U_l^{(k)}, a_k\right) + O_p(r^{-2}K^2N)
\end{aligned} \tag{37}$$

where $\mathrm{Cov}_{G,D|\mathcal{R},S}\left(U_l^{(k)}, a_k\right)$ equals:

$$\frac{K}{n}\left\{\frac{\sum_{m\in\mathcal{N}} W_m}{\sum_{l\in\mathcal{R}} W_l}+\frac{(n-r-1)W_l\sum_{m\in\mathcal{N}} W_m}{\left(\sum_{l\in\mathcal{R}} W_l\right)^2}+\frac{(n-r)W_l}{\sum_{l\in\mathcal{R}} W_l}\right\}+O_p(r^{-2}K^2), \tag{38}$$

$$\mathrm{Var}_{G,D|\mathcal{R},S}(a_k)=\frac{K}{n}\left\{\frac{2(n-r)}{n-1}+\frac{(n-r)^2\sum_{l\in\mathrm{R}} W_l^2}{(n-1)\left(\sum_{l\in\mathrm{R}} W_l\right)^2}\right\}+O_p(r^{-2}K^2), \tag{39}$$

and $\mathrm{Cov}_{G,D|\mathcal{R},S}\left(U_{l_1}^{(k)}U_{l_2}^{(k)}, a_k\right)$ equals:

$$\begin{aligned}&\frac{K}{n}\left[2\left\{\frac{\sum_{m\in\mathcal{N}} W_m+\sum_{m_1\neq m_2\in\mathcal{N}} W_{m_1}W_{m_2}}{\left(\sum_{l\in\mathcal{R}} W_l\right)^2}\right\}\right.\\&\left.+(W_{l_1}+W_{l_2})\left\{\frac{n-r}{\sum_{l\in\mathrm{R}} W_l}+\frac{2(n-r-1)\sum_{m\in\mathcal{N}} W_m}{\left(\sum_{l\in\mathcal{R}} W_l\right)^2}+\frac{(n-r-2)\sum_{m_1\neq m_2\in\mathcal{N}} W_{m_1}W_{m_2}}{\left(\sum_{l\in\mathcal{R}} W_l\right)^3}\right\}\right]\\&+O_p(r^{-2}K^2).\end{aligned} \tag{40}$$

By plugging (37) and (36) into (35) and using (38)–(40) and the Taylor Expansion argument for simplification, we obtain, for $k=1,\ldots,K$,

$$nN^{-2}K^{-1}E\left(\tilde{Y}_S^{(k)}-Y\right)^2=nN^{-2}\mathrm{Var}\hat{Y}_I+O(n^{-1}K)+O(n^{-1}).$$

This together with (16) gives the asymptotic unbiasedness of v_{AS}:

$$\begin{aligned}nN^{-2}E(v_{AS})&=nN^{-2}K^{-1}\left\{E\left(\tilde{Y}_S^{(1)}-Y\right)^2+\mathrm{Cov}\left(\tilde{Y}_S^{(1)},\tilde{Y}_S^{(2)}\right)\right\}+O(n^{-1}K)\\&=nN^{-2}\mathrm{Var}\hat{Y}_I+O(n^{-1}K)+O(K^{-1})+O(n^{-1}).\end{aligned} \tag{41}$$

Theorem 2 is now proved.

Lemma 1. *Let X_k^K, $k=1,\ldots,K$, be identically distributed positive random variables with a common mean $\mu<\infty$ and a constant $\mathrm{Cov}\left(X_k^K, X_h^K\right)$ for $k\neq h$ and any fixed*

K. Assume that $\mathrm{Cov}\left(X_k^K, X_h^K\right) \to 0$ *as* $K \to \infty$. *Then, for any* $\varepsilon > 0$,

$$\lim_{K\to\infty} P\left(\left|\frac{1}{K}\sum_{k=1}^{K} X_k^K - \mu\right| > \varepsilon\right) = 0 \tag{42}$$

Lemma 1 can be proved by slightly modifying the proof of Khintchine's law of large numbers (see, e.g., Chung 1974, pp. 109–110) and using the fact that $\mathrm{Cov}\left(X_k^K, X_h^K\right) \to 0$.

Proof of Theorem 3. The second term on the right-hand side of (18) converges to 0 in probability, because, by Chebyshev's inequality, for any $\varepsilon > 0$ and $k \neq h$,

$$\begin{aligned} P\left(\sqrt{\frac{n(\theta-\bar{\theta})^2}{N^2(K-1)}} > \varepsilon\right) &\leq \frac{nE(\theta-\bar{\theta})^2}{\varepsilon^2 N^2(K-1)} \\ &= \frac{n\mathrm{Var}\left(\tilde{Y}_S^{(k)}\right)}{\varepsilon^2 N^2(K-1)K} + \frac{n\mathrm{Cov}\left(\tilde{Y}_S^{(k)}, \tilde{Y}_S^{(h)}\right)}{\varepsilon^2 N^2 K} \\ &= \frac{nVar\left(\tilde{Y}_S^{(k)}\right)}{\varepsilon^2 N^2(K-1)K} + O(K^{-1}) \\ &\to 0 \end{aligned}$$

when $K \to \infty$, where the second last equality follows from (16) and the last equality follows from $\mathrm{Var}\left(\tilde{Y}_S^{(k)}\right) = O(N^2K/n)$. It remains to show the first term on the right-hand side of (18) is consistent. For $k = 1, \ldots, K$, let $X_k^K = nN^{-2}(K-1)^{-1}\left(\tilde{Y}_S^{(k)} - \theta\right)^2$. Following similar argument for the proof of (16), it can be shown that $\mathrm{Cov}\left(X_k^K, X_h^K\right) \to 0$ when $nK^{-1} \to \infty$ and $K \to \infty$. By Theorem 2, (15) and C1, $E\left(X_k^K\right)$ is bounded for all k. Then, by Lemma 1, (42) holds. This, together with the proved fact $\lim_{K\to\infty} E\left(X_k^K\right) - nN^{-2}\mathrm{Var}\left(\hat{Y}_I\right) = 0$ in Theorem 2, implies v_{AS} is asymptotically consistent. Thus, Theorem 3 is proved.

5. References

Chung, K.L. (1974). A Course in Probability Theory, (Second Edition). New York: Academic Press.

Haziza, D., Thompson, J.K., and Yung, W. (2010). The Effect of Nonresponse Adjustments on Variance Estimation. Survey Methodology, 36, 35–43.

Kalton, G. and Kasprzyk, D. (1986). The Treatment of Missing Data. Survey Methodology, 12, 1–16.

Moore, R.A. (2006). Random Group Variance Adjustments when Hot Deck Imputation is used to Compensate for Nonresponse. Technical Report, U.S. Census Bureau.

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

Rao, J.N.K. and Shao, J. (1996). On Balanced Half-Sample Variance Estimation in Stratified Sampling. Journal of the American Statistical Association, 91, 343–348.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.
Shao, J. (2001). Replication Methods for Variance Estimation in Complex Surveys with Imputed Data. In Survey Nonresponse, R. Groves, D. Dillman, J. Eltinge, and R. Little (eds). New York: Wiley and Sons, 303–314.
Shao, J. and Sitter, R.R. (1996). Bootstrap for Imputed Survey Data. Journal of the American Statistical Association, 91, 1278–1288.
Shao, J., Chen, Y., and Chen, Y. (1998). Balanced Repeated Replications for Stratified Multistage Survey Data under Imputation. Journal of the American Statistical Association, 93, 819–831.
Thompson, K.J. and Yung, W. (2006). To Replicate (a Weight Adjustment Procedure) or not to Replicate? An Analysis of the Variance Estimation Effects of a Shortcut Procedure Using the Stratified Jackknife. Proceedings of the American Statistical Association, Section on Survey Research Methods, 3772–3779.
Wolter, K.M. (2007). Introduction to Variance Estimation, (Second Edition). New York: Springer-Verlag.

Received October 2009
Revised November 2010