Sankhyā: The Indian Journal of Statistics
Special Issue on Sample Surveys
1999, Volume 61, Series B, Pt. 1, pp. 187-201

APPROXIMATE BALANCED HALF SAMPLE AND RELATED REPLICATION METHODS FOR IMPUTED SURVEY DATA*

By JUN SHAO
and
YINZHONG CHEN
University of Wisconsin, Madison

SUMMARY. The balanced half sample (BHS) method is a popular method for variance estimation under stratified multistage sampling. The standard BHS works when two units (or two first-stage clusters) are sampled from each stratum and when there is no imputed data. We consider the situation where the first-stage sample sizes are larger than two in some strata and where the data set contains nonrespondents imputed using some popular methods such as ratio and random hot deck imputation. The method we propose is a combination of an approximate BHS method (such as the grouped BHS method and its extensions) and the adjustment for imputation considered in Shao, Chen and Chen (1998). Consistency of the proposed BHS variance estimators is established under some regularity conditions. Some examples are presented for illustration.

1. **Introduction**

The balanced half sample (BHS) method is one of the most popular variance estimation methods for nonlinear survey estimators from stratified multistage sampling designs. Survey agencies such as the U.S. Census Bureau, the U.S. Bureau of Labor Statistics, and Westat have computer softwares for computing BHS variance estimators. Compared with the linearization method for variance estimation, BHS requires more computations but has the advantages of (1) requiring no theoretical derivations of a variance formula for each problem, which

AMS (1991) *subject classification.* Primary 62D05; secondary 62G05, 62G99.
Key words and phrases. Grouped half-samples, pseudo-strata; random repeated replication; ratio imputation; hot deck imputation; variance estimation.

* This work was finished when the first author was an ASA/NSF/BLS/Census Senior Research Fellow at the U.S. Census Bureau and the U.S. Bureau of Labor Statistics. The research was also partially supported by National Science Foundation Grants DMS-9504425 and DMS-9803112 and by National Security Agency Grant MDA904-99-1-0032.

cannot be avoided when applying the linearization method; (2) programming ease in complex situations; (3) using a unified recipe for various problems, for example, variance estimation for functions of estimated totals and for sample quantiles. Compared with another population resampling method, the jackknife, BHS has a wider application scope, since the jackknife is known to have problems in estimating variances of nonsmooth estimators such as sample quantiles.

The BHS method was first proposed by McCarthy (1969) for the case of stratified sampling with two sampled units per stratum. This method is based on the creation of a set of "balanced" half samples (each half sample contains one unit from each stratum), which can be done easily using Hadamard matrices. Although the idea of BHS can be extended to cases where some strata contain more than two units, it is much more difficult to construct balanced sub-samples in general. Thus, various approximate BHS or related replication methods are proposed in the literature. We consider the following three methods: (1) The grouped BHS method or its extension, which works by grouping the units in each stratum into two pseudo-units and then applying BHS to the pseudo-units; (2) The pseudo-strata BHS method, which works by creating some pseudo-strata, each containing exactly two units, and then applying BHS to the pseudo-strata; (3) The random repeated replication (RRR) method: which works by randomly forming some half samples or sub-samples. More details about these methods are given in Section 2. The grouped BHS is currently used in the Current Employment Survey (the U.S. Bureau of Labor Statistics) and the RRR method is adopted in the Transpotation Annual Survey (the U.S. Census Bureau).

Most surveys have nonrespondents. For unit nonresponse, weighting adjustment is usually applied to produce a new data set containing respondents. the BHS, approximate BHS, or other related replication method can then be applied to the new data set. For item nonresponse, however, imputation is commonly applied to compensate for nonresponse. When the rate of imputation is not negligible, applying BHS designed for the case of no missing data by treating the imputed values as observed data usually produces negatively biased and inconsistent variance estimators (see, for example, Rao and Shao, 1992). An adjusted BHS is proposed in Shao, Chen and Chen (1998) and the resulting adjusted BHS variance estimator is shown to be asymptotically unbiased and consistent.

The purpose of this article is to study applications of approximate BHS and RRR in the presence of imputed data. In Section 2, we introduce the BHS, approximate BHS and RRR methods in more detail. Special attention is paid to applications of these methods in cases where some stratum sizes are odd numbers. For imputed data sets, we apply the idea of adjustment in Shao, Chen and Chen (1998) to produce adjusted approximate BHS or replication variance estimators. In Section 3, these variance estimators are shown to be consistent under some regularity conditions. Some examples are presented in Section 4 for illustration.

2. Methodology

2.1. *Sampling design.* Throughout this article we consider the following commonly used stratified multistage sampling design. The finite population under consideration has been stratified into H strata with N_h first-stage units (ultimate units for single stage sampling or clusters for multistage sampling) in the hth stratum. From stratum h, $n_h \geq 2$ first-stage units are selected without replacement according to some probability sampling plan, independently across the strata. We assume that the first stage sampling fraction n/N is negligible, where $n = \sum_h n_h$ and $N = \sum_h N_h$, although some within-stratum sampling fractions n_h/N_h may be non-negligible. For multistage sampling, a second-stage sample, a third-stage sample, and so on, may be selected from each sampled cluster, using some sampling methods independently across the clusters. We denote the ultimate units in the population by (h,i,k), $h = 1, ..., H$, $i = 1, ..., N_h$, $k = 1, ..., N_{hi}$, where N_{hi} is the number of ultimate units in stratum h and cluster i, and the sample by $s = \{(h,i,k) \text{ in the sample}\}$. From the specification of the sampling design, a survey weight w_{hik} attached to the sample ultimate unit (h,i,k) is obtained. Using the survey weights, an unbiased estimator of the population total for an item y is of the form

$$\hat{Y} = \sum_{(h,i,k)\in s} w_{hik} y_{hik}, \qquad \ldots (1)$$

where y_{hik} is the value for item y associated with unit (h,i,k).

The most commonly used survey estimator is a function of a vector of estimated totals of the form (1); that is, $\hat{\theta} = g(\hat{\boldsymbol{Y}})$, where $\hat{\boldsymbol{Y}}$ is defined by (1) with y_{hik} replaced by $\boldsymbol{y}_{hik}$, the vector of values of several items associated with unit (h,i,k), and g is a known function. Examples of such estimators include ratio or regression estimators of a population total, and the ratio of two estimated totals. Another important class of survey estimators is the class of sample quantiles and their related functions. For an item y, the empirical distribution function is defined as

$$\hat{F}(t) = \sum_{(h,i,k)\in s} w_{hik} I(y_{hik} \leq t) \Bigg/ \sum_{(h,i,k)\in s} w_{hik},$$

where $I(\cdot)$ is the usual indicator function. For $p \in (0,1)$, the pth sample quantile is $\hat{F}^{-1}(p)$, where $\hat{F}^{-1}$ is the usual inverse of a distribution function.

To study asymptotic properties of $\hat{\theta} = g(\hat{\boldsymbol{Y}})$ or $\hat{\theta} = \hat{F}^{-1}(p)$, we assume that the finite population under study is a member of a sequence of finite populations indexed by λ, but λ is suppressed for simplicity of notation. All limiting processes are understood to be as $\lambda \to \infty$. For $\hat{\theta} = g(\hat{\boldsymbol{Y}})$, $\hat{\theta}$ is asymptotically normal with mean $\theta = g(\boldsymbol{Y})$, $\boldsymbol{Y} = E\hat{\boldsymbol{Y}}$, if the following regularity conditions are satisfied:

C1: As $\lambda \to \infty$, $n \to \infty$, $n/N \to 0$, and $\max_{h,i} \sum_k w_{hik}/M = O(n^{-1})$, where M is the total number of ultimate units in the population.

C2: $0 < \lim\inf[nV(\hat{\boldsymbol{Y}}/M)]$.

C3: $M^{-1}\sum_h \sum_i n_h^{-1} E\|\boldsymbol{z}_{hi} - E(\boldsymbol{z}_{hi})\|^{2+\delta} = O(n^{-(1+\delta)})$ for some $\delta > 0$, where $\|\ \|$ is the usual vector norm and $\boldsymbol{z}_{hi} = \sum_k n_h w_{hik}\boldsymbol{y}_{hik}$.

For $\hat{\theta} = \hat{F}^{-1}(p)$, it is asymptotically normal with mean $\theta = F^{-1}(p)$, where F is the finite population distribution function for item y and $\theta = F^{-1}(p)$, if C1 and the following condition are satisfied:

C4: There is a sequence of densities $\{f_\lambda\}$ such that $0 < \inf_\lambda f_\lambda(\theta) \leq \sup_\lambda f_\lambda(\theta) < \infty$ and

$$\lim_{\lambda\to\infty}\left[\frac{F\left(\theta + O(n^{-1/2})\right) - F(\theta)}{O(n^{-1/2})} - f_\lambda(\theta)\right] = 0.$$

2.2 *BHS, approximate BHS, and other replication methods.* BHS in the case of $n_h = 2$ for all h can be described as follows. A set of R half samples is formed by deleting one first-stage unit from the sample in each stratum. Let $(\epsilon_{rh})_{R\times H}$ be an $R \times H$ matrix with $\epsilon_{rh} = +1$ or -1 according to whether the first or the second first-stage unit of the hth stratum is in the r-th half sample. A set of R half samples is said to be balanced if $\sum_{r=1}^R \epsilon_{rh}\epsilon_{rh'} = 0$ for all $h \neq h'$. A minimal set of R balanced half samples can be constructed from an $R \times R$ Hadamard matrix by choosing any H columns excluding the column of all +1's, where $H + 1 \leq R \leq H + 4$. Once we have a set of R balanced half samples, the BHS variance estimator for a survey estimator $\hat{\theta}$ (either $g(\hat{\boldsymbol{Y}})$ or a sample quantile) is

$$v_{\text{BHS}} = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)} - \hat{\theta}\right)^2, \qquad \ldots(2)$$

where $\hat{\theta}^{(r)}$ is the same as $\hat{\theta}$ but computed based on the r-th half sample. The estimator $\hat{\theta}^{(r)}$ can be obtained using the same formula for $\hat{\theta}$ with w_{hik} changed to $w_{hik}^{(r)}$ which equals $2w_{hik}$ or 0 according as whether the (h,i)th first-stage unit is selected or not selected in the r-th half sample. When $\hat{\theta} = \hat{Y}$ in (1), v_{BHS} is exactly the same as the standard text book variance estimator for $\hat{Y}$ given in, for example, Cochran (1977, equation (11.78), p.319), because of the balancedness of the half samples (see, e.g., Wolter, 1985). In general, consistency of v_{BHS} can be established under conditions C1-C4 (Krewski and Rao 1981, Shao and Wu 1992, Shao and Rao 1994).

When some $n_h > 2$, we consider the following three methods:

(1) The grouped BHS method (Kish and Frankel 1970, Wolter 1985, pp. 131-132). In this method, the sample first-stage units in each stratum is first divided at random into two groups containing $[n_h/2]$ and $n_h - [n_h/2]$ first-stage units, where $[x]$ is the integer part of x. The BHS method is then applied to

the groups and the grouped BHS variance estimator, denoted by v_{GBHS}, can be obtained using formula (2). There are two problems in this method. The first problem is that v_{GBHS} is biased if there are odd n_h's and the bias is not negligible if odd n_h's are small. This problem can be solved by modifying the estimators $\hat{\theta}^{(r)}$ in (2) as follows. For each r, $\hat{\theta}^{(r)}$ is computed using the same formula for $\hat{\theta}$ with w_{hik} changed to

$$w_{hik}^{(r)} = \left(1 + \sqrt{\frac{n_h - m_{rh}}{m_{rh}}}\right) w_{hik} \qquad \text{or} \qquad \left(1 - \sqrt{\frac{m_{rh}}{n_h - m_{rh}}}\right) w_{hik} \qquad \ldots (3)$$

according as whether the (h, i)th first-stage unit is in or not in the r-th half sample, where m_{rh} = the number of first-stage units in the r-th half sample ($m_{rh} = [n_h/2]$ or $n_h - [n_h/2]$). Note that in the standard grouped BHS method, weights $w_{hik}^{(r)} = 2w_{hik}$ or 0 is used. Clearly, $w_{hik}^{(r)}$ in (3) is equal to $2w_{hik}$ or 0 for even n_h's.

The second problem with the grouped BHS method is that v_{GBHS} is inconsistent in cases where H is small and $\min_h n_h \to \infty$ (Valliant 1987, Rao and Shao 1996). Rao and Shao (1996) proposed the following repeatedly grouped BHS method to overcome this difficulty: We first repeat independently the random grouping T times and then obtain the repeatedly grouped BHS variance estimator v_{RGBHS} as the average of the T grouped BHS variance estimators. Rao and Shao (1996) established the consistency of v_{RGBHS}, assuming that H is fixed, $\min_h n_h \to \infty$ and $T \to \infty$.

(2) The pseudo-strata BHS method (Wolter 1985, pp. 132-133). Consider first the case where all n_h are even. The hth stratum is first divided at random into $n_h/2$ pseudo-strata, each of size 2, $h = 1, ..., H$. The BHS method with formula (2) is then applied to the expanded set of $\sum_h n_h/2 = n/2$ strata, using a Hadamard matrix of size R, $n/2 + 1 \leq R \leq n/2 + 4$. The resulting variance estimator, v_{PSBHS}, is shown to be consistent and have better finite sample performance than v_{GBHS} or v_{RGBHS} in a simulation study (Rao and Shao 1996).

When some n_h's are odd, we have the following two suggestions. The first one is to divide the hth stratum (with an odd n_h) at random into $(n_h - 1)/2$ pseudo-strata, one pseudo-stratum of size 3 and the rest of them of size 2. We then have a data set with pseudo-stratum sizes either 2 or 3. Balanced subsamples can be obtained using the method of orthogonal arrays described in Wu (1991). Our second suggestion is to combine the grouped BHS and the pseudo-strata BHS methods; i.e., we first randomly construct $(n_h - 1)/2$ pseudo-strata, one pseudo-stratum of size 3 and the rest of them of size 2, and then randomly form two groups in the pseudo-stratum of size 3. Formula (3) should be used if there are many pseudo–strata of size 3.

(3) The random repeated replication (RRR) method (Shao and Wu 1992, Shao and Tu 1995, Section 6.2.3). Let G be an integer $\leq n_h$ for all h (if some $n_h = 2$, then $G = 2$). The RRR method can be described in the following steps:

Step 1: In stratum h, randomly group the n_h sampled first-stage units into $G+1$ groups (or G groups when n_h/G is an integer), where each of the first G groups contains $[n_h/G]$ first-stage units and the last group contains $n_h - G[n_h/G]$ first-stage units. Do this independently for all h.

Step 2: For $g = 1, ..., G-1$, let $\hat{\theta}^{(g)}$ be the same as $\hat{\theta}$ with w_{hik} changed to

$$w_{hik}^{(g)} = \left(1 + \sqrt{\tfrac{n_h - [n_h/G]}{[n_h/G]}}\right) w_{hik}$$

for units in stratum h and in group g (formed in step 1) or

$$w_{hik}^{(g)} = \left(1 - \sqrt{\tfrac{[n_h/G]}{n_h - [n_h/G]}}\right) w_{hik}$$

for units in stratum h and not in group g (including the $(G+1)$th group when n_h/G is not an integer).

Step 3: Repeat steps 1-2 independently T times to obtain $\hat{\theta}^{(g)}$, $g = 1, ..., R$, where $R = T(G-1)$. The RRR variance estimator is defined as

$$v_{\mathrm{RRR}} = \frac{1}{R} \sum_{g=1}^{R} \left(\hat{\theta}^{(g)} - \hat{\theta}\right)^2 .$$

Note that the total number of sub-samples created in Steps 1-3 is $R = T(G-1)$. When G is small (e.g., when some n_h are small), we need a large T. When we choose $G = 2$, the sub-samples are half samples when all n_h are even and nearly half samples in general. Hence, the RRR method can be viewed as a half sample or repeated replication method, but the half samples or sub-samples are obtained randomly and are not exactly balanced.

For all three methods described here, the resulting variance estimator is not the same as the standard text book variance estimator in the special case of $\hat{\theta} = \hat{Y}$. However, the expectation (with respect to grouping or creating pseudo-strata) of the approximate BHS or RRR variance estimator is exactly the same as the standard text book variance estimator when $\hat{\theta} = \hat{Y}$.

2.3 *Imputation.* We consider marginal imputation or item imputation; that is, nonrespondents for an item are imputed using the respondents from the same item (and some auxiliary data observed from all sampled units). Imputation is done independently across items. Marginal imputation is simple but does not preserve the relationship between two items so that it provides biased estimators of population parameters such as the correlation coefficients and cell proportions in two-way tables. In this article we consider cases where $\hat{\theta} = g(\hat{\boldsymbol{Y}})$ or $\hat{\theta} = \hat{F}^{-1}(p)$ for a given item. In such cases marginal imputation provides asymptotically unbiased and consistent estimators, under a condition such as C5 or C5′ stated later.

Imputation is often done by first forming several imputation classes (according to the values of an auxiliary variable observed for all sampled units) and then performing imputation using the data within an imputation class, independently across the imputation classes. There are two popular ways to form imputation classes. Under the customary design-based approach, we assume

C5: The response probability is constant within each imputation class. Imputation classes are not necessarily the same for different items and units in the same imputation class may have different response probabilities for different items. Under the model-based approach (Särndal, Swennson and Wretman 1992), we assume

C5$'$: For an item y being imputed, the response probability does not depend on y-value (but may depend on a vector of covariates $\boldsymbol{x}$ used in imputation) and within each imputation class, there is a model that relates y and $\boldsymbol{x}$:

$$y_{hik} = \beta' \boldsymbol{x}_{hik} + \sigma s(\boldsymbol{x}_{hik}) e_{hik}, \qquad \ldots (4)$$

where β and σ are unknown parameters (which may be different in different imputation classes), β' is the transpose of β, $s(\cdot)$ is a known function, and the e_{hik} are independent and identically distributed random variables with mean 0 and variance 1.

When $\boldsymbol{x}_{hik} \equiv 1$ (no covariates), model (4) means that the y_{hik} are homogeneous within an imputation class. Within an imputation class, we assume that imputation is done by using one of the following methods.

1. *Ratio imputation.* When there are auxiliary data $\{x_{hik}\}$ (available for all sampled units), the ratio imputation method imputes a missing item by $\hat{\rho}_\nu x_{hik}$, where $\hat{\rho}_\nu = \sum_{(h,i,k)\in A_\nu} w_{hik} y_{hik} / \sum_{(h,i,k)\in A_\nu} w_{hik} x_{hik}$, and A_ν is the index set of all respondents for item y in the νth imputation class.

2 *Regression imputation.* With auxiliary data $\{\boldsymbol{x}_{hik}\}$, the regression imputation method imputes a missing item by $\boldsymbol{x}'_{hik}\hat{\beta}_\nu$, where

$$\hat{\beta}_\nu = \left(\sum_{(h,i,k)\in A_\nu} w_{hik} s^{-2}(\boldsymbol{x}_{hik}) \boldsymbol{x}_{hik} \boldsymbol{x}'_{hik} \right)^{-1} \sum_{(h,i,k)\in A_\nu} w_{hik} s^{-2}(\boldsymbol{x}_{hik}) \boldsymbol{x}_{hik} y_{hik}$$

is the weighted least squares estimator of β under model (4) within the νth imputation class.

3 *Random hot deck.* The simple random hot deck method imputes nonrespondents by a simple random sample with replacement from respondents for the same item (within an imputation class). Rao and Shao (1992) introduced the following weighted random hot deck method: within an imputation class, a missing value is imputed by a value randomly selected with replacement from the respondents for the same item with probability proportional to the survey weights w_{hik} attached to the respondents.

These imputation methods are commonly used in practice, for example, in the U.S. Bureau of Labor Statistics (West 1984), the U.S. Bureau of the Census (King and Kornbau 1994), and Statistics Canada (Lee, Rancourt and Särndal 1994).

Once the data set is imputed, survey estimator $\hat{\theta}$ is computed by using standard formulas and treating imputed values as observed values.

Ratio or regression imputation is deterministic imputation, since missing values are imputed by nonrandom values given the data set. An advantage of using ratio or regression imputation is that it incorporates auxiliary data. However, ratio or regression imputation does not provide consistent sample quantiles. Random hot deck imputation methods are not as efficient as deterministic imputation methods for estimating totals. The simple random hot deck method provides $\hat{\theta}$ that is consistent from the model-based viewpoint under assumption C5$'$ (i.e., the imputation classes are internally homogeneous), whereas the weighted random hot deck method provides $\hat{\theta}$ that is consistent from the model-based viewpoint under assumption C5$'$ as well as from the design-based viewpoint under assumption C5.

For ratio or regression imputation, we assume that C2 and C3 are also satisfied when $\boldsymbol{y}$ is replaced by $\boldsymbol{x}$. For variance estimation with imputed data, we also need the following moment condition:

C6: The population $(2+\delta)$th moments for $\boldsymbol{y}$ (and $\boldsymbol{x}$ if it is used in imputation), $\lambda = 1, 2, ...$, are bounded, where $\delta > 0$ is a constant.

2.4 *Adjusted approximate BHS and RRR for imputed data.* For imputed data, applying the standard BHS, approximate BHS, or RRR by treating the imputed values as observed data usually does not produce correct variance estimators. We consider the following method of adjustment proposed in Shao, Chen and Chen (1998). Consider first the case of $\hat{\theta} = g(\hat{\boldsymbol{Y}})$. Suppose that R half samples are obtained using one of the three methods described in Section 2.2. For every r, we adjust every imputed value $\tilde{y}_{hik}$ in the r-th half sample to

$$\tilde{y}_{hik}^{(r)} = \tilde{y}_{hik} + E_{A_\nu}^{(r)}(\tilde{y}_{hik}) - E_{A_\nu}(\tilde{y}_{hik}), \qquad \ldots(5)$$

where E_{A_ν} is the expectation with respect to the original imputation procedure within the νth imputation class, and $E_{A_\nu}^{(r)}$ is the same as E_{A_ν} except that the imputation is performed using data in the r-th half sample. The adjusted approximate BHS or RRR variance estimator is then obtained by applying (2) with $\hat{\theta}^{(r)}$ replaced by $\hat{\theta}_{adj}^{(r)}$ which is the same as $\hat{\theta}^{(r)}$ but based on the adjusted data set.

For any deterministic imputation method, such as the mean, ratio, and regression imputation, $E_{A_\nu}(\tilde{y}_{hik}) = \tilde{y}_{hik}$ and $E_{A_\nu}^{(r)}(\tilde{y}_{hik})$ is the imputed value for a missing y_{hik} based on the data in the r-th half sample only. Therefore, adjustment (5) is the same as re-imputing missing values in the r-th half sample using the data in the r-th half sample. For example, a nonrespondent y_{hik} in the νth imputation class is imputed by $\tilde{y}_{hik} = \hat{\rho}_\nu x_{hik}$ in ratio imputation (Section 2.3); the adjusted value is then $\tilde{y}^{(r)}_{hik} = \hat{\rho}^{(r)}_\nu x_{hik}$, where $\hat{\rho}^{(r)}_\nu$ is computed the same as $\hat{\rho}_\nu$ except that w_{hik} is replaced by $w^{(r)}_{hik}$.

For the weighted random hot deck method,

$$E_{A_\nu}(\tilde{y}_{hik}) = \sum_{(h,i,k)\in A_\nu} w_{hik} y_{hik} \Bigg/ \sum_{(h,i,k)\in A_\nu} w_{hik} \qquad \ldots (6)$$

and

$$E^{(r)}_{A_\nu}(\tilde{y}_{hik}) = \sum_{(h,i,k)\in A_\nu} w^{(r)}_{hik} y_{hik} \Bigg/ \sum_{(h,i,k)\in A_\nu} w^{(r)}_{hik}. \qquad \ldots (7)$$

The adjustment for the simple random hot deck method can be obtained using (6)-(7) and setting w_{hik} to 1. For sample quantiles, we focus on the hot deck imputation method, since ratio or regression imputation does not provide valid distribution and quantile estimators. The adjusted approximate BHS or RRR variance estimator can be obtained as follows. First, apply adjustment (5) to indicator functions, instead of y values; that is, for an imputed value $\tilde{y}_{hik}$, $I(\tilde{y}_{hik} \le t)$ is adjusted to

$$I^{(r)}_{adj}(\tilde{y}_{hik} \le t) = I(\tilde{y}_{hik} \le t) + \bar{F}^{(r)}_\nu(t) - \bar{F}_\nu(t),$$

where

$$\bar{F}_\nu(t) = \sum_{(h,i,k)\in A_\nu} w_{hik} I(y_{hik} \le t) \Bigg/ \sum_{(h,i,k)\in A_\nu} w_{hik}$$

and

$$\bar{F}^{(r)}_\nu(t) = \sum_{(h,i,k)\in A_\nu} w^{(r)}_{hik} I(y_{hik} \le t) \Bigg/ \sum_{(h,i,k)\in A_\nu} w^{(r)}_{hik}.$$

For a fixed r,

$$\hat{F}^{(r)}_{adj}(t) = \sum_{(h,i,k)\in s} w^{(r)}_{hik} I^{(r)}_{adj}(\tilde{y}_{hik} \le t) \Bigg/ \sum_{(h,i,k)\in s} w^{(r)}_{hik},$$

where $\tilde{y}_{hik} = y_{hik}$ if y_{hik} is a respondent. However, $\hat{F}^{(r)}_{adj}$ is not a distribution function due to the adjustment and, hence, $(\hat{F}^{(r)}_{adj})^{-1}(p)$ is not defined. Note that

$$\hat{F}^{(r)}_{adj}(t) = \hat{F}^{(r)}(t) + \sum_\nu c^{(r)}_\nu \left[\bar{F}^{(r)}_\nu(t) - \bar{F}_\nu(t)\right],$$

where

$$\hat{F}^{(r)}(t) = \sum_{(h,i,k)\in s} w^{(r)}_{hik} I(\tilde{y}_{hik} \le t) \Bigg/ \sum_{(h,i,k)\in s} w^{(r)}_{hik},$$

$$c_\nu^{(r)} = \sum_{(h,i,k)\in A_\nu^c} w_{hik}^{(r)} \Big/ \sum_{(h,i,k)\in s} w_{hik}^{(r)},$$

and A_ν^c is the index set of all nonrespondents for item y in the νth imputation class. Hence, we define

$$\hat{\theta}_{adj}^{(r)} = (\hat{F}^{(r)})^{-1}(p) + \sum_\nu c_\nu^{(r)} \left[(\bar{F}_\nu^{(r)})^{-1}(p) - \bar{F}_\nu^{-1}(p) \right].$$

The adjusted approximate BHS or RRR variance estimator is then obtained using (2) with $\hat{\theta}^{(r)}$ replaced by $\hat{\theta}_{adj}^{(r)}$.

3. Asymptotic Results

The following results establish the consistency of the approximate BHS and RRR variance estimators described in the previous section.

Consider first the case of functions of estimated totals. We assume that imputation is done using one of the method described in Section 2.3.

THEOREM 1. *Let $\hat{\theta} = g(\hat{\boldsymbol{Y}})$, where g is continuously differentiable in a neighbourhood of $E(\hat{\boldsymbol{Y}})$. Assume C1-C3, C5 (or C5′) and C6.*

(i) *Suppose that $H \to \infty$ and n_h's are bounded. Then*

$$\frac{v}{V(\hat{\theta})} \to 1 \qquad \textit{in probability,} \qquad \ldots(8)$$

where $V(\hat{\theta})$ denotes the asymptotic variance of $\hat{\theta}$ and v is one of the grouped BHS variance estimator v_{GBHS}, the pseudo-strata BHS variance estimator v_{PSBHS}, and the RRR variance estimator v_{RRR}, with the adjustment described in Section 2.4 for imputation (for v_{RRR}, $R \to \infty$ is assumed).

(ii) Suppose that H is fixed and $\min_h n_h \to \infty$. Then result (8) holds for $v = v_{\text{PSBHS}}$ or v_{RRR}. Result (8) does not hold for v_{GBHS}, but it holds for the repeatedly grouped BHS variance estimator v_{RGBHS} (with the adjustment for imputation) with $T \to \infty$.

PROOF. We prove the univariate case with $\hat{\theta} = \sum_{(h,i,k)\in s} w_{hik}\tilde{y}_{hik}$, where $\tilde{y}_{hik}$ is y_{hik} if y_{hik} is observed and is an imputed value otherwise. The multivariate case where $\theta = g(\boldsymbol{Y})$ can be handled using Taylor's expansion. For notation simplicity, we assume that there is only one imputation class. Also, we consider the weighted random hot deck imputation. The case of deterministic imputation is simpler and omitted.

The r-th half sample (replicate) estimate after the adjustment is

$$\hat{\theta}_{adj}^{(r)} = \sum_{(h,i,k)\in s} a_{hik} w_{hik}^{(r)} y_{hik} + \sum_{(h,i,k)\in s} (1 - a_{hij}) w_{hij}^{(r)} \left[\tilde{y}_{hij} + \bar{y}^{(r)} - \bar{y} \right],$$

where $\bar{y}$ and $\bar{y}^{(r)}$ are given by (6) and (7), respectively (no need for the index ν since we consider one imputation class). Then

$$\hat{\theta}^{(r)}_{adj} - \hat{\theta} = \left[\bar{y}^{(r)}\hat{M}^{(r)} - \bar{y}\hat{M}\right] + \sum_{(h,i,k)\in s} d^{(r)}_{hik} t_{hik},$$

where $\hat{M} = \sum_s w_{hik}$, $\hat{M}^{(r)} = \sum_s w^{(r)}_{hik}$, $d^{(r)}_{hik} = w^{(r)}_{hik}/w_{hik} - 1$, and $t_{hik} = (1 - a_{hik})w_{hik}(\tilde{y}_{hik} - \bar{y})$. Hence,

$$\frac{1}{R}\sum_{r=1}^{R}\left(\hat{\theta}^{(r)}_{adj} - \hat{\theta}\right)^2 = A + B + C$$

with

$$A = \frac{1}{R}\sum_{r=1}^{R}\left[\bar{y}^{(r)}\hat{M}^{(r)} - \bar{y}\hat{M}\right]^2,$$

$$B = \frac{1}{R}\sum_{r=1}^{R}\left(\sum_{(h,i,k)\in s} d^{(r)}_{hik} t_{hik}\right)^2,$$

and

$$C = \frac{2}{R}\sum_{r=1}^{R}\left[\bar{y}^{(r)}\hat{M}^{(r)} - \bar{y}\hat{M}\right]\sum_{(h,i,k)\in s} d^{(r)}_{hik} t_{hik}.$$

Note that $V(\hat{\theta}) = V[\tilde{E}(\hat{\theta})] + E[\tilde{V}(\hat{\theta})]$, where $\tilde{E}$ and $\tilde{V}$ are the mean and variance under weighted hot deck imputation. Also, $\tilde{E}(\hat{\theta}) = \bar{y}\hat{M}$. Hence, from the results in Rao and Shao (1996),

$$n\{A - V[\tilde{E}(\hat{\theta})]\} \to 0 \quad \text{in probability.}$$

Let E_* be the expectation with respect to random grouping, or creating pseudo-strata. Then

$$E_*(B) = \sum_{h=1}^{H}\sum_{i=1}^{n_h} t^2_{hi} - \sum_{h=1}^{H}\sum_{i\neq j}\frac{1}{n_h - 1} t_{hi}t_{hj},$$

where $t_{hi} = \sum_k t_{hik}$. Since $\tilde{E}(t_{hi}) = 0$, $\tilde{E}(t^2_{hi}) = \tilde{V}(t_{hi})$, and $\tilde{E}(t_{hi}t_{hj}) = 0$ for $i \neq j$, $\tilde{E}E_*(B) = \tilde{V}(\hat{\theta})$. Using condition C6 and the law of large numbers,

$$n\{B - E[\tilde{V}(\hat{\theta})]\} \to 0 \quad \text{in probability.}$$

Similarly we can show that

$$nC \to 0 \quad \text{in probability}$$

and hence the result.

Next, consider the case of $\hat{\theta} = \hat{F}^{-1}(p)$, assuming that the nonrespondents are imputed by weight random hot deck (or simple hot deck if C5$'$ is assumed).

Theorem 2. *Assume C1 and C5 (or C5$'$). The conclusions in Theorem 1 still hold for $\hat{\theta} = \hat{F}^{-1}(p)$.*

Proof. Using the same arguments in Shao (1994) we can establish the following Bahadur-type representations:

$$\hat{F}^{-1}(p) - F^{-1}(p) = \frac{F(\theta) - \hat{F}(\theta)}{f(\theta)} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

$$(\hat{F}^{(r)})^{-1}(p) - F^{-1}(p) = \frac{F(\theta) - \hat{F}^{(r)}(\theta)}{f(\theta)} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

$$\bar{F}^{-1}(p) - F^{-1}(p) = \frac{F(\theta) - \bar{F}(\theta)}{f(\theta)} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

and

$$(\bar{F}^{(r)})^{-1}(p) - F^{-1}(p) = \frac{F(\theta) - \bar{F}^{(r)}(\theta)}{f(\theta)} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Then

$$\hat{\theta}^{(r)}_{adj} - \hat{\theta} = \frac{\hat{F}(\theta) - \hat{F}^{(r)}(\theta)}{f(\theta)} + \sum_{\nu} c^{(r)}_{\nu} \frac{\bar{F}(\theta) - \bar{F}^{(r)}(\theta)}{f(\theta)} + o_p\left(\frac{1}{\sqrt{n}}\right).$$

Using the same arguments in the proofs of Lemmas 3.1-3.3 in Shao and Rao (1994) we can show that

$$v = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{\hat{F}(\theta) - \hat{F}^{(r)}(\theta)}{f(\theta)} + \sum_{\nu} c^{(r)}_{\nu} \frac{\bar{F}(\theta) - \bar{F}^{(r)}(\theta)}{f(\theta)}\right)^2 + o_p\left(\frac{1}{n}\right).$$

The result is then a consequence of Theorem 1 by considering $\hat{\theta} = \hat{F}(\theta) = \sum_s w_{hik}\tau_{hik}$ with $\tau_{hik} = I_{\tilde{y}_{hik}}(\theta)$.

4. Examples

In this section we illustrate the application of the approximate BHS or RRR methods using a data set from the Transportation Annual Survey (TAS) conducted by the U.S. Census Bureau.

The TAS is a survey of firms with one or more establishments that are primarily engaged in providing commercial motor freight transportation or public warehousing services in U.S. A one-stage stratified simple random sample is selected without replacement from employers contained on the Census Bureau's

Standard Statistical Establishment List. The strata, which are also the imputation classes in this example, are constructed according to business type and size. Two survey forms, warehousing and trucking, are usually used. We consider only those establishments receiving the warehousing form. Table 1 lists sample sizes (n_h) and survey weights ($w_{hik} = w_h$) for this data set.

There are various variables in this survey. We use the current year annual revenue (y) as the variable of interest. There are nonrespondents for y. The response size in each stratum is given in Table 1.

Table 1. Sample sizes and weights

Stratum		Sample Sizes and Weights		
industry code	size	sample size	response size	weight
4221	1	14	6	12.43
4221	2	11	7	8.91
4221	3	10	4	6.10
4221	4	11	6	5.73
4221	5	16	12	2.70
4221	6	18	13	2.16
4222	1	8	2	32.91
4222	2	13	10	9.85
4222	3	11	9	10.82
4222	4	12	10	6.08
4222	5	13	10	3.60
4225	1	14	9	87.91
4225	2	11	8	67.39
4225	3	13	10	44.48
4225	4	14	13	25.28
4225	5	16	13	15.57
4225	6	18	12	9.80
4225	7	15	11	6.23
4225	8	15	14	4.68
4225	9	40	33	2.13
4226	1	7	5	32.14
4226	2	14	6	15.56
4226	3	11	8	11.73
4226	4	14	12	7.00
4226	5	13	9	6.18
4226	6	11	7	4.70
4226	7	17	12	3.31
4226	8	20	15	1.80
4226	9	23	17	1.74

First, we consider ratio imputation introduced in Section 2.3. The x variable in ratio imputation is the previous year annual revenue. Table 2 lists the estimated population total and its variance estimates computed using the grouped BHS method and the RRR method introduced in Section 2.2. In each case, “adjusted for imputation” means that the variance estimate is computed using adjustment (5) for imputed values, whereas “unadjusted” means that the variance estimate is computed treating imputed values as observed data (which leads to an underestimation of the variance). For the grouped BRR, a Hadamard matrix of size 32 is used to construct grouped half samples. For the RRR, $G = 2$ and $T = 32$ are used.

Next, we consider weighted random hot deck imputation (values of x are not used). Results similar to those in Table 2 are given in Table 3.

Table 2. Results for ratio imputation

Estimated total = 4.095×10^9		
	Variance estimates	
	Grouped BHS	RRR
Adjusted for imputation	5.802×10^{16}	5.844×10^{16}
Unadjusted	5.294×10^{16}	5.317×10^{16}
Ratio (unadjusted over adjusted)	0.912	0.910

Table 3. Results for weighted hot deck imputation

Estimated total = 4.216×10^9		
	Variance estimates	
	Grouped BHS	RRR
Adjusted for imputation	3.036×10^{17}	3.108×10^{17}
Unadjusted	2.408×10^{17}	2.773×10^{17}
Ratio (unadjusted over adjusted)	0.793	0.892

Table 4. Results for weighted hot deck imputation
(based on the modified data set with even stratum sample sizes)

Estimated total = 4.318×10^9			
	Variance estimates		
	Grouped BHS	Pseudo-strata BHS	RRR
Adjusted for imputation	3.494×10^{17}	3.187×10^{17}	3.349×10^{17}
Unadjusted	2.775×10^{17}	2.705×10^{17}	2.926×10^{17}
Ratio (unadjusted over adjusted)	0.794	0.849	0.874

To illustrate the pseudo-strata BHS method introduced in Section 2.2, we delete some units in the data set to create a modified data set of even stratum sample sizes (sample weights are adjusted accordingly). This is just for convenience (i.e., it is difficult to apply pseudo-strata BHS to data with odd stratum sample sizes) and for a fair comparison among three methods. Also, we only compute the results for weighted random hot deck imputation (Table 4).

References

Cochran, W. G. (1977). *Sampling Techniques*, third edition. Wiley, New York.

King, C. and Kornbau, M. (1994). *Inventory of Economic area Statistical Practices.* ESMD Report Series 9401, Bureau of the Census, Washington D.C.

Kish, L. and Frankel, M.R. (1970). Balanced repeated replication for standard errors, *J. Amer. Statist. Assoc.*, **65**, 1071–1094.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: properties of the linearization, jackknife and balanced repeated replication methods, *Ann. Statist.*, **9**, 1010–1019.

Lee, H., Rancourt, E. and Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values. *Journal of Official Statistics*, **10**, 231-243.

McCarthy, P.J. (1969). Pseudo-replication: half samples, *Rev. Internat. Statist. Inst.*, **37**, 239–264.

Rao, J.N.K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation. *Biometrika* **79**, 811-822.

−−−− (1996). On balanced half-sample variance estimation in stratified sampling. *J. Amer. Statist. Assoc.* **91**, 343-348.

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

Shao, J. (1994). L-statistics in complex survey problems. *Ann. Statist.* **22**, 946-967.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replications for stratified multistage survey data under imputation. *J. Amer. Statist. Assoc.* **93**, 819-831.

Shao, J. and Rao, J.N.K. (1994). Standard errors for low income proportions estimated from stratified multi-stage samples. *Sankhya B*, Special Volume 55, 393-414.

Shao, J. and Tu, D. (1995). *Jackknife and Bootstrap.* Springer, New York.

Shao, J. and Wu, C.F.J. (1992). Asymptotic properties of the balanced repeated replication method for sample quantiles. *Ann. Statist.* **20**, 1571-1593.

Valliant, R. (1987). Some prediction properties of balanced half-sample variance estimators in single-stage sampling. *J. Roy. Statist. Soc. B*, **49**, 68-81.

West, S.A. (1984). A comparison of estimators for the variance of regression-type estimators in a finite population. *Proceedings of the Section of Survey Research Methods*, American Statistical Association, 170-175.

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

Wu, C.F.J. (1991). Balanced repeated replications based on mixed orthogonal arrays, *Biometrika*, **78**, 181–188.

Jun Shao and Yinzhong Chen
Department of Statistics
University of Wisconsin
Madison, WI 53706
e-mail: shao@stat.wise.edu