Survey Methodology, December 2001
Vol. 27, No. 2, pp. 189-196
Statistics Canada, Catalogue No. 12-001

A Repeated Half-Sample Bootstrap and Balanced Repeated Replications for Randomly Imputed Data

Hiroshi Saigo, Jun Shao and Randy R. Sitter [1]

Abstract

In this paper, we discuss the application of the bootstrap with a re-imputation step to capture the imputation variance (Shao and Sitter 1996) in stratified multistage sampling. We propose a modified bootstrap that does not require rescaling so that Shao and Sitter's procedure can be applied to the case where random imputation is applied and the first-stage stratum sample sizes are very small. This provides a unified method that works irrespective of the imputation method (random or nonrandom), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). In addition, we discuss the proper Monte Carlo approximation to the bootstrap variance, when using re-imputation together with resampling methods. In this setting, more care is needed than is typical. Similar results are obtained for the method of balanced repeated replications, which is often used in surveys and can be viewed as an analytic approximation to the bootstrap. Finally, some simulation results are presented to study finite sample properties and various variance estimators for imputed data..

Key Words: Hotdeck; Percentile method; Monte Carlo; Imputation; Bootstrap sample size.

1. Introduction

Item nonresponse is a common occurrence in surveys and is usually handled by imputing missing item values. The various imputation methods used in practice can be classified into two types: deterministic imputation, such as mean, ratio and regression imputation, typically using the respondents and some auxiliary data observed on all sampled elements; and random imputation. In both cases the imputation is often applied within imputation classes formed on the basis of auxiliary variables. This article focuses on random imputation.

Typically, random imputation is done in such a way that applying the usual estimation formulas to the imputed data set produces asymptotically unbiased and consistent survey estimators (*e.g.*, means, totals, quantiles). More details about random imputation are provided in section 2. It is common practice to also treat the imputed values as true values when estimating variances of survey estimators. This leads to serious underestimation of variances if the proportion of missing data is appreciable, and to poor confidence intervals.

There have been some proposals in the literature to circumvent this difficulty. For random imputation, Rubin (1978) and Rubin and Schenker (1986) proposed the multiple imputation method to account for the inflation in the variance, which can be justified from a Bayesian perspective (Rubin 1987). Adjusted jackknife methods for variance estimation have been proposed for both random and deterministic imputations (Rao and Shao 1992; Rao 1993; Rao and Sitter 1995; Sitter 1997), under stratified multistage sampling. However, it is well known that the jackknife cannot be applied to non-smooth estimators, *e.g.*, a sample quantile or an estimated low income proportion (Mantel and Singh 1991).

There are two methods available for handling randomly imputed data for both smooth and non-smooth estimators: the adjusted balanced repeated replication (BRR) methods proposed by Shao, Chen and Chen (1998); and the bootstrap method proposed by Shao and Sitter (1996) (see also Efron 1994) with a re-imputation step to capture the imputation variance. The bootstrap method is more computer intensive but is easy to motivate and understand, and provides a unified method that works irrespective of the imputation method (random or nonrandom), the type of $\hat{\theta}$ (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation).

In this article we continue the work by Shao and Sitter (1996). First, we show in section 3 how Shao and Sitter's bootstrap procedure can be modified to handle very small stratum sizes (*e.g.*, two psu's per stratum). Second, we discuss in section 4 the proper Monte Carlo approximation to the bootstrap estimators, a problem for which more care is needed when random re-imputation is applied than is typical. This has no detrimental effect on bootstrap confidence intervals based on the percentile method, but if done incorrectly, will cause the bootstrap-t to perform poorly. Third, we consider a BRR variance estimation method with a re-imputation step, which can be viewed as an analytic and symmetric approximation to the bootstrap method. Finally, we present some simulation results to study properties of various bootstrap and BRR variance estimators.

1. Hiroshi Saigo, School of Political Science and Economics, Waseda University, 1-6-1 Nishiwaseda Shinjuku, Tokyo, 169-8050 Japan; Jun Shao, Professor, Department of Statistics, University of Wisconsin, Madison, WI 53706, U.S.A.; Randy R. Sitter, Professor, Department of Statistics and Acturial Science, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6.

2. Stratified Multistage Sampling and Random Imputation

Though the methods discussed in this article can be more generally applied, we restrict attention to the commonly used stratified multistage sampling design. Suppose that the population contains H strata and in stratum h, n_h clusters are selected with probabilities p_{hi}, $i = 1, ..., n_h$. Samples are taken independently across strata. In the case of complete response on item y, let

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi} \Big/ (n_h\ p_{hi})$$

by a linear unbiased estimator of the stratum total Y_h, where $\hat{Y}_{hi}$ is a linear unbiased estimator of the cluster total Y_{hi} for a selected cluster based on sampling at the second and subsequent stages. A linear unbiased estimator of the total, $Y = \sum Y_h$, is given by $\hat{Y} = \sum \hat{Y}_h$, which may be written as

$$\hat{Y} = \sum_{(hik)\varepsilon s} w_{hik}\ y_{hik}, \qquad (1)$$

where s is the complete sample of elements, and w_{hik} and y_{hik} respectively denote the sampling weight and the item value attached to the $(hik)^{\text{th}}$ sampled element.

Often a survey estimator, $\hat{\theta}$, can be expressed as a function of a vector of estimated totals as in (1). If one is interested in the population distribution function, it can be estimated by $\hat{F}_n(t) = \sum_s w_{hik}\, I(y_{hik} \le t)/\hat{U}$, where $I(\cdot)$ is the usual indicator function and $\hat{U} = \sum_s w_{hik}$. Some non-smooth estimators that are of interest are the p^{th} sample quantile, $\hat{F}^{-1}(p)$, where $\hat{F}^{-1}$ is the quantile function of $\hat{F}$, and the sample low income proportion $\hat{F}[1/2\ \hat{F}^{-1}(1/2)]$.

Suppose that the value y_{hik} is observed for $(hik)\,\varepsilon\, s_r \subset s$, termed a respondent, while for others, $(hik)\,\varepsilon\, s_m$, it is missing, termed a nonrespondent, with $s = s_r \cup s_m$. When there are missing data, it is common practice to use $\{y_{hik} : (hik)\,\varepsilon\, s_r\}$ to obtain imputed values $\tilde{y}_{hik}$ for $(hik)\,\varepsilon\, s_m$ and then treat these imputed values as if they were true observations and estimate Y with

$$\hat{Y}_I = \sum_{s_r} w_{hik}\ y_{hik} + \sum_{s_m} w_{hik}\ \tilde{y}_{hik}. \qquad (2)$$

In practice, the accuracy of the imputation is improved by first forming several imputation classes using control variables observed on the entire sample, and then imputing within imputation class. For simplicity we consider a single imputation class.

Random imputation entails imputing the missing data by a random sample from the respondents, or, in the presence of auxiliary data, by using a random sample or residuals. If the imputation is suitably done, the estimator $\hat{Y}_I$ in (2) is asymptotically unbiased and consistent, although it is not as efficient as $\hat{Y}$ in (1). Throughout this article, we assume that, either

> within each imputation cell, the response probability for a given variable is a constant, the response statuses for different units are independent, and imputation is carried out within each imputation cell and independently across the imputation cells,

or

> within each imputation cell, the response probability of a given variable does not depend on the variable itself (but may depend on the covariates used for imputation), imputation is carried out independently across the imputation cells, and within an imputation cell, imputation is performed according to a model that relates the variable being imputed to the covariates used for imputation.

We also assume the same asymptotic setting as that in Shao *et al.* (1998). Thus, consistency (or asymptotic unbiasedness) refers to convergence of estimators (or expectations of estimators) under the assumption in Shao *et al.* (1998), as the first-stage sample size $n = \sum n_h$ increases to infinity.

There are many methods random imputation. We consider only two in this article: the weighted hotdeck considered in Rao and Shao (1992), which we refer to simply as random imputation, and the adjusted weighted hotdeck proposed in Chen, Rao and Sitter (2000), which we refer to as adjusted random imputation. Our results can be easily extended to random imputation with residuals in the presence of auxiliary data (*e.g.*, random regression imputation). Generalizations to other types of random imputation may be possible, but will not be considered here.

Random imputation randomly selects donors, $\tilde{y}_{hik}$ from $\{y_{hik} : (hik)\,\varepsilon\, s_r\}$ with replacement with probabilities $w_{hik}/\hat{T}$, where $\hat{T} = \sum_{s_r} w_{hik}$. In this case $E_I(\hat{Y}_I) = (\hat{S}/\hat{T})\,\hat{U} = \hat{Y}_r$, a ratio estimator which is asymptotically unbiased and consistent for Y, where $\hat{S} = \sum_{s_r} w_{hik}\ y_{hik}$. Here E_I denotes expectation under the random imputation. The variance of $\hat{Y}_I$ is larger than the variance of $\hat{Y}_r$ because of the random imputation. However, the distribution of item values in the imputed data set is preserved.

Adjusted random imputation simply uses $\tilde{\eta}_{hik} = \tilde{y}_{hik} + (\hat{S}/\hat{T} - \tilde{S}/\tilde{T})$ as the imputed values instead of $\tilde{y}_{hik}$, where $\tilde{S} = \sum_{s_m} w_{hik}\ \tilde{y}_{hik}$, $\tilde{T} = \sum_{s_m} w_{hik}$ and $\tilde{y}_{hik}$ are the imputed values from random imputation. Chen *et al.* (2000) show that this method completely eliminates the variability due to the random imputation for estimating the population total. That is $\tilde{Y}_I = \sum_{s_r} w_{hik}\ y_{hik} + \sum_{s_m} w_{hik}\ \eta_{hik} = \hat{Y}_r$. The method also retains the distribution of item values in the imputed data set. However, the resulting imputed values need not be actual realizations.

An imputed estimator of the distribution function under random imputation is given by

$$\hat{F}_I(t) = \left[\sum_{s_r} w_{hik}\, I(y_{hik} \le t) + \sum_{s_m} w_{hik}\, I(\tilde{y}_{hik} \le t) \right] \Big/ \hat{U}. \qquad (3)$$

An imputed estimator of the distribution function under adjusted random imputation, denoted $\tilde{F}_I(t)$, is simply obtained by replacing $\tilde{y}_{hik}$ in (3) by $\tilde{\eta}_{hik}$. For estimating the

distribution function, adjusted random imputation does not eliminate the imputation variance as it does for estimating the total. However, Chen *et al.* (2000) show that it does significantly reduce the imputation variance when compared to random imputation. Both $\hat{F}_I(t)$ and $\tilde{F}_I(t)$ are asymptotically unbiased and consistent.

For studying variance estimation with resampling methods, we assume that n/N is negligible, where $n = \sum n_h$, $N = \sum N_h$ and N_h is the number of first-stage clusters in the population.

3. A Repeated Half-Sample Bootstrap

When there are imputed missing data, naive bootstrap variance estimators obtained by treating the imputed data set, $\boldsymbol{Y}_I$, as $\boldsymbol{Y} = \{y_{hik} : (hik) \in s\}$, the data set of no missing values, do not capture the inflation in variance due to imputation and/or missing data and lead to serious underestimation. As a result, they are inconsistent. This is so, because simply treating $\boldsymbol{Y}_I$ as $\boldsymbol{Y}$ ignores the imputation process. This was noted by Shao and Sitter (1996) and they proposed re-imputing the bootstrap data set in the same way as the original data set was imputed. The bootstrap procedure in Shao and Sitter (1996) can be described as follows.

1. Draw a simple random sample $\{\boldsymbol{y}^*_{hi} : i = 1, ..., n_h - 1\}$ with replacement from the sample $\{\tilde{\boldsymbol{y}}_{hi} : i = 1, ..., n_h\}$, $h = 1, ..., H$, independently across the strata, where $\tilde{\boldsymbol{y}}_{hi} = \{y_{hij} : (h, i, j) \in s_r\} \cup \{\tilde{y}_{hij} : (h, i, j) \in s_m\}$.
2. Let a^*_{hij} be the response indicator associated with y^*_{hij}, $s^*_m = \{(h, i, j) : a^*_{hij} = 0\}$ and $s^*_r = \{(h, i, j) : a^*_{hij} = 1\}$. Apply the same imputation procedure used in constructing the imputed data sey $\boldsymbol{Y}_I$ to the "nonrespondents" in s^*_m, using the "respondents" in s^*_r. Denote the bootstrap analogue of $\boldsymbol{Y}_I$ by $\boldsymbol{Y}^*_I$.
3. Obtain the bootstrap analogue $\hat{\theta}^*_I$ of $\hat{\theta}$, based on the imputed bootstrap data set $\boldsymbol{Y}^*_I$. For example, if $\hat{\theta} = \hat{Y}$ in (1) and $\hat{\theta}_I = \hat{Y}_I$ in (2), then

$$\hat{\theta}^*_I = \hat{Y}^*_I = \sum_{s^*_r} w^*_{hik}\, y^*_{hik} + \sum_{s^*_m} w^*_{hik}\, \tilde{y}^*_{hik}, \qquad (4)$$

where $\tilde{y}^*_{hik}$ is the imputed value using the bootstrap data and w^*_{hik} is $n_h/(n_h - 1)$ times the survey weight associated with y_{hik} (to reflect the fact that the bootstrap sample size is $n_h - 1$, not n_h). The bootstrap estimator of $\mathrm{Var}(\hat{\theta}_I)$ is

$$\upsilon_B(\hat{\theta}_I) = \mathrm{Var}^*(\hat{\theta}^*_I), \qquad (5)$$

where Var^* is the conditional variance with respect to $\boldsymbol{Y}^*_I$, given $\boldsymbol{Y}_I$.

Shao and Sitter (1996) show that the bootstrap estimator defined in (5) is consistent for both smooth and nonsmooth estimators $\hat{\theta}$. When a random imputation method is considered, an implicit condition in their development is that $n_h/(n_h - 1)$ goes to 1. This can be seen from the special case of $\hat{\theta} = \hat{Y}$. From (2),

$$\mathrm{Var}(\hat{Y}_I) = \mathrm{Var}[E_I(\hat{Y}_I)] + E[\mathrm{Var}_I(\hat{Y}_I)]$$
$$= \mathrm{Var}\left(\frac{\sum_{s_r} w_{hik} y_{hik} \sum_{s} w_{hik}}{\sum_{s_r} w_{hik}} \right) + E\left(\hat{\sigma}^2 \sum_{s_m} w^2_{hik} \right), \qquad (6)$$

where

$$\hat{\sigma}^2 = \sum_{s_r} w_{hik} (y_{hik} - \bar{y}_r)^2 \Big/ \sum_{s_r} w_{hik},$$
$$\bar{y}_r = \sum_{s_r} w_{hik}\, y_{hik} \Big/ \sum_{s_r} w_{hik}.$$

Similarly, by (4),

$$\mathrm{Var}^*(\hat{Y}^*_I) = \mathrm{Var}^*\left(\frac{\sum_{s^*_r} w^*_{hik}\, y^*_{hik} \sum_{s^*} w^*_{hik}}{\sum_{s^*_r} w^*_{hik}} \right)$$
$$+ E^*\left(\hat{\sigma}^{*2} \sum_{s^*_m} w^{*2}_{hik} \right), \qquad (7)$$

where

$$\hat{\sigma}^{*2} = \sum_{s^*_r} w^*_{hik} (y^*_{hik} - \bar{y}^*_r)^2 \Big/ \sum_{s^*_r} w^*_{hik},$$
$$\bar{y}^*_r = \sum_{s^*_r} w^*_{hik}\, y^*_{hik} \Big/ \sum_{s^*_r} w^*_{hik}.$$

From the theory of the bootstrap, the first terms on the right hand side of (6) and (7) converge to the same quantity, as do $\hat{\sigma}^2$ and $\hat{\sigma}^{*2}$. Thus, Shao and Sitter's bootstrap is consistent if $\sum_{s^*_m} w^{*2}_{hik}$ and $\sum_{s_m} w^2_{hik}$ converge to the same quantity, which is true if $n_h/(n_h - 1)$ converges to 1 for all h, because

$$E^*\left(\sum_{s^*_m} w^{*2}_{hik} \right) = E^*\left[\sum_{s^*} (1 - a^*_{hik})\, w^{*2}_{hik} \right]$$
$$= \sum_{s} (1 - a_{hik})\, w^2_{hik}\, n_h/(n_h - 1).$$

The second term on the right hand side of (6) is the variance component corresponding to random imputation, which is typically a small portion of the overall variance. Thus, the overestimation due to $n_h/(n_h - 1)$ is serious only when the n_h's are very small. The case $n_h = 2$ is, however, an important special case.

We now propose a bootstrap method which has no difficulty in the case of very small n_h's while remaining valid more generally. Note that the use of bootstrap sample size $n_h = 1$ is to ensure that the first term on the right hand side of (7) has the same limit as the first term on the right

hand side of (6) (Rao and Wu 1988). When n_h is used as the bootstrap sample size in stratum h, Rao and Wu (1988) showed that in the case of no missing data, the bootstrap variance estimator underestimates. They proposed a rescaling to circumvent the problem, but rescaling does not produce correct bootstrap estimators in the presence of imputed data.

What is ideally required for our problem is a bootstrap method with the bootstrap sample size equal to the original sample size n_h which produces an asymptotically unbiased variance estimator (in the case of no missing data) without rescaling. We now show that this can be accomplished as follows. Suppose that there is no missing data and that all of the $n_h = 2m_h$'s are even. Take a simple random sample of size m_h without replacement independently from $\{\boldsymbol{y}_{hi}: i = 1, ..., n_h\}$ and repeat each obtained unit twice to get $\{\boldsymbol{y}^*_{hi}: i = 1, ..., n_h\}$. We call this method the repeated half-sample bootstrap. The resulting υ_B will then be approximately unbiased and consistent. In the linear case where $\hat{Y} = \sum_{(hik)} w_{hik}\, y_{hik} = \sum_h \sum_{i=1}^{n_h} y_{hi}/n_h = \sum_h \bar{y}_h$ and $y_{hi} = \sum_{k=1}^{n_{hi}} n_h\, w_{hik}\, y_{hik}$, the consistency of υ_B follows from

$$\begin{aligned}
\mathrm{Var}^*(\hat{Y}^*) &= \sum_h \mathrm{Var}^*(\bar{y}^*_h) = \sum_h \mathrm{Var}^*\left(\frac{1}{n_h}\sum_{i=1}^{n_h} y^*_{hi}\right) \\
&= \sum_h \mathrm{Var}^*\left(\frac{2m_h}{n_h}\,\frac{1}{m_h}\sum_{i=1}^{m_h} y^*_{hi}\right) \\
&= \sum_h \mathrm{Var}^*\left(\frac{1}{m_h}\sum_{i=1}^{m_h} y^*_{hi}\right) \\
&= \sum_h \frac{(1-1/2)}{m_h}\,\frac{1}{n_h-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2 \\
&= \sum_{hh} s_h^2/n_h,
\end{aligned}$$

the usual approximately unbiased and consistent estimator of variance, where $s_h^2 = (n_h-1)^{-1}\sum_{i=1}^{n_h}(y_{hi}-\bar{y}_h)^2$. The consistency of υ_B for a nonlinear $\hat{\theta}_I$ follows from the linear case and Taylor's expansion, when $\hat{\theta}_I$ is a function of weighted averages, or the arguments used in Shao and Rao (1994), Shao and Sitter (1996), and Shao *et al*. (1998) when $\hat{\theta}_I$ is non-smooth such as a median.

If $n_h = 2m_h + 1$ is odd, it is not possible to take an exact half-sample. In this case, the following two results lead us to an adaptation of the above idea:

i) If we choose a simple random resample of size $m_h = (n_h - 1)/2$ without replacement and repeat each unit twice, we end up with $n_h - 1$ units. If we obtain an additional unit by selecting one at random from the $n_h - 1$ units already resampled, $\mathrm{Var}^*(\hat{Y}^*) = \sum_h (n_h + 3)\, s_h^2/n_h^2$;

ii) If we choose a simple random resample of size $m_h + 1$ without replacement and repeat each unit twice, we end up with $n_h + 1$ units. If we discard one of these at random, $\mathrm{Var}^*(\hat{Y}^*) = \sum_h (n_h + 1)\, s_h^2/n_h^2$.

Thus, if we used method (i) with probability 1/4 and method (ii) with probability 3/4 at each bootstrap replication, we obtain the desired result. This repeated half-sample bootstrap method yields approximately unbiased variance estimates without rescaling and has a bootstrap sample size equal to the original sample size. Thus, if we use this bootstrap for Step 1 of the method of Shao and sitter (1996) as described above, the resulting bootstrap estimators are asymptotically unbiased and consistent for any n_h, under the regularity conditions stated in Shao and Sitter (1996) and Shao *et al*. (1998).

4. The Proper Monte Carlo for the Bootstrap

If υ_B in (5) has no explicit form, one may use the Monte Carlo approximation

$$\upsilon_B(\hat{\theta}_I) \approx \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{I(b)} - \bar{\theta}^*_I)^2, \tag{8}$$

where $\bar{\theta}^*_I = B^{-1}\sum_{b=1}^{B}\hat{\theta}^*_{I(b)}$, $\hat{\theta}^*_{I(b)} = \hat{\theta}(\boldsymbol{Y}^*_{I(b)})$, and $\boldsymbol{Y}^*_{I(b)}$, $b = 1, ..., B$, are independent re-imputed bootstrap data sets. It is common practice in many applications of the bootstrap to replace the average of the bootstrap estimators $\bar{\theta}^*_I$ in (8) by the original estimator $\hat{\theta}_I$ (see Rao and Wu 1985, page 232). The latter is simpler to use and is thus the most common. With no imputed data, this is usually correct. However, using the analogue with the re-imputed bootstrap is not correct. The reason is that $\hat{\theta}_I$ is the result of a single realization of the random imputation, while $\bar{\theta}^*_I \approx E^*(\hat{\theta}^*_I) \approx E_I(\hat{\theta}_I)$ since we are averaging over repeated re-imputations, and $\hat{\theta}_I$ and $E_I(\hat{\theta}_I)$ are not close for random imputation. When $\hat{\theta}_I = \hat{Y}_I$, for example, $E_I(\hat{Y}_I) = \hat{Y}_r$ given in section 2 and the difference $\hat{Y}_I - \hat{Y}_r$ is not a relatively negligible term when random imputation is used. Thus,

$$\begin{aligned}
\upsilon_{B2} &= \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{I(b)} - \hat{\theta}_I)^2 \\
&= \frac{1}{B}\sum_{b=1}^{B}(\hat{\theta}^*_{I(b)} - \bar{\theta}^*_I)^2 + (\bar{\theta}^*_I - \hat{\theta}_I)^2
\end{aligned}$$

and the first term goes to $\mathrm{Var}^*(\hat{\theta}^*_I)$ as $B \to \infty$ but the second term does not go to zero which implies that υ_{B2} badly overestimates the variance. This is not only true for the proposed repeated half-sample bootstrap but also for those considered in Shao and Sitter (1996).

One should also note that using the $\hat{\theta}^*_{I(b)}$, $b = 1, ..., B$ to obtain bootstrap confidence intervals via the percentile method avoids this concern since the histogram of these values will be correctly centered about $E^*(\hat{\theta}^*_I)$. However, one must take more care with boostrap$-t$ confidence

intervals. It is important that one define $t_b^* = (\hat{\theta}_{I(b)}^* - \bar{\theta}_{I(\cdot)}^*)/\sigma_b^*$ (not $t_b^* = (\hat{\theta}_I^{*b} - \hat{\theta}_I)/\sigma_b^*$) and use $\{\hat{\theta}_I - t_U^* \sigma_b^*, \hat{\theta}_I - t_L^* \sigma_b^*\}$, where $\sigma_b^{*2} = \upsilon_B(\boldsymbol{Y}_I^{*b})$, $t_L^* = \hat{\mathrm{CDF}}_t^{-1}(\alpha)$, $t_U^* \hat{\mathrm{CDF}}_t^{-1}(1-\alpha)$ and $\hat{\mathrm{CDF}}_t(x) = \#\{t_b^* \le x;\ b = 1, \ldots, B\}/B$.

5. A Repeated BRR

We first describe the most common application of the BRR, $n_h = 2$ clusters per stratum (McCarthy 1969) in the setting of no missing data. A set of B balanced half-samples or replicates is formed by deleting one first-stage cluster from the sample in each stratum, where this set is defined by a $B \times H$ matrix $(\delta_{bh})_{B\times H}$ with $\delta_{bh} = +1$ or -1 according to whether the first or the second first-stage cluster of stratum h is in the b^{th} half-sample and $\sum_{r=1}^{B} \delta_{bh}\delta_{bh'} = 0$ for all $h \ne h'$; that is, the columns of the matrix are orthogonal. A minimal set of B balanced half-samples can be constructed from a $B \times B$ Hadamard matrix by choosing any H columns excluding the column of all +1's, where $H + 1 \le B \le H + 4$. Let $\hat{\theta}_{(b)}$ be the survey estimator computed from the b^{th} half sample. The estimator $\hat{\theta}_{(b)}$ can be obtained using the same formula as for $\hat{\theta}$ with w_{hik} changed to $w_{hik(b)}$, which equals $2w_{hik}$ or 0 according to whether or not the $(hi)^{\text{th}}$ cluster is selected in the b^{th} half-sample or not. The BRR variance estimator for $\hat{\theta}$ is then given by

$$\upsilon_{\text{BRR}} = \frac{1}{B}\sum_{b=1}^{B} (\hat{\theta}_{(b)} - \bar{\theta}_{(\cdot)})^2, \qquad (9)$$

where $\bar{\theta}_{(\cdot)} = \sum_b \hat{\theta}_{(b)}/B$, and is often replaced by $\hat{\theta}$. The variance estimator υ_{BRR} has been shown to be consistent for smooth functions of estimated totals by Krewski and Rao (1981) and for nonsmooth estimators by Shao, and Wu (1992) and Shao and Rao (1994).

A naive BRR for problems with randomly imputed data would be obtained as in (9) with $\hat{\theta}_{(b)}$ and $\bar{\theta}_{(\cdot)}$ replaced by $\hat{\theta}_{I(b)}$ and $\bar{\theta}_{(\cdot)} = B^{-1}\sum_b \hat{\theta}_{I(b)}$, where $\hat{\theta}_{I(b)}$ is the estimator calculated from $\boldsymbol{Y}_I$ using the BRR weights. But this produces inconsistent variance estimators because it fails to take into account the effect of missing data and the random imputation.

To correctly apply the BRR in the presence of random imputation by using re-imputation, we must deal with the issue of n_h being small. Recall that for the bootstrap such small n_h's caused difficulty because the stratum resample size, $n_h - 1$, was smaller than the original stratum sample size, n_h. This is true for the BRR, as well. We propose an easy way to circumvent this difficulty. Rather than obtaining the b^{th} BRR replicate of the estimator, $\hat{\theta}_{(b)}$, from the same formula as for $\hat{\theta}$ but with weights $w_{hik(b)}$ equal $2w_{hik}$ or 0 according as to whether the $(hi)^{\text{th}}$ cluster is selected in the b^{th} half-sample or not, instead use the original weights but include the $(hi)^{\text{th}}$ cluster twice or not at all according as to whether the $(hi)^{\text{th}}$ cluster is selected in the b^{th} half-sample or not. If we view the BRR in this way: i) the resulting υ_{BRR} in (9) remains the same; and ii) the resample size is the same as the original sample size. This repeated BRR can be viewed as a type of balanced bootstrap, however one should note that the balanced bootstrap described in Nigam and Rao (1996) for the case of no missing data does not work in this case because, though it uses a resample size $n_h = 2$ in each stratum, it does so in such a way as to still require rescaling and thus will not work in the presence of random imputation.

The proposed repeated BRR has no difficulty in the presence of random imputation. The procedure becomes

1. Form the set of half-samples, 1 unit per stratum, using a Hadamard matrix as described above.
2. Obtain the b^{th} BRR replicate by repeating each unit in the obtained half-sample twice. Denote this $\{\boldsymbol{y}_{hi}^*: i = 1, \ldots, n_h = 2\}$.
3. Let a_{hij}^* be the response indicator associated with y_{hij}^*, $s_m^* = \{(h, i, j): a_{hij}^* = 0\}$, and $s_r^* = \{(h, i, j): a_{hij}^* = 1\}$. Apply the same imputation procedure used in constructing $\boldsymbol{Y}_I$ to the units in s_m^*, using the "respondents" in s_r^*. Denote the b^{th} BRR replicate of $\boldsymbol{Y}_I$ by $\boldsymbol{Y}_{I(b)}^*$.
4. Obtain the BRR analogue $\hat{\theta}_{I(b)}^*$ of $\hat{\theta}$, based on the imputed BRR data set $\boldsymbol{Y}_{I(b)}^*$.
5. Repeat 1–4 for each row of the $B \times H$ matrix to get $\hat{\theta}_{I(b)}^*$ for $b = 1, \ldots, B$ and apply the standard BRR formula (9) to obtain BRR variance estimators for $\hat{\theta}_I$, with $\bar{\theta}_{(\cdot)} = B^{-1}\sum_b \hat{\theta}_{I(b)}^*$ (For the same reason that is discussed in section 4, we should not replace $\bar{\theta}_{I(\cdot)}$ by $\hat{\theta}_I$).

We can extend this idea to cases with $n_h > 2$ by using the same strategy with half-samples obtained from balanced orthogonal multi-arrays (BOMA's) (Sitter 1993). For example, Table 1 gives a set of $B = 24$ balanced resamples for $H = 7$ strata with $n_h = 4$ psu's in each stratum. It is derived using the BOMA given in Table 1 of Sitter (1993) and repeating each resampled unit twice as in Step 2 above. Using a BOMA in Steps 1 and 2 of the procedure above also results in an approximately unbiased variance estimator, BOMA's are fairly easily constructed for even n_h using balanced incomplete block designs and Hadamard matrices, but are difficult to construct for odd n_h. They can also handle unequal n_h's for different strata, though construction becomes a more serious problem (see Sitter 1993).

6. A Simulation

To study the properties of the proposed resampling variance estimators, we consider a finite population of $H = 32$ strata with N_h clusters in stratum h and ten ultimate units in each cluster. The characteristic of interest y_{hik} are generated as follows:

Table 1
A Set of Balanced Resamples Constructed from a BOMA

	h						
b	1	2	3	4	5	6	7
1	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)
2	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)
3	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)
4	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)
5	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)
6	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)
7	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)
8	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)
9	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)
10	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)
11	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)
12	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)
13	(1,1,3,3)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)
14	(1,1,4,4)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)
15	(1,1,2,2)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)
16	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)
17	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)
18	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)
19	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)	(2,2,4,4)
20	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)	(2,2,3,3)
21	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)	(3,3,4,4)
22	(1,1,3,3)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(2,2,4,4)	(1,1,3,3)	(1,1,3,3)
23	(1,1,4,4)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(2,2,3,3)	(1,1,4,4)	(1,1,4,4)
24	(1,1,2,2)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(3,3,4,4)	(1,1,2,2)	(1,1,2,2)

$$y_{hik} = y_{hi} + \varepsilon_{hik},$$

where $y_{hi} \sim N(\mu_h, \sigma_h^2)$ independent of $\varepsilon_{hik} \sim N(0, [1 - \rho]\,\sigma_h^2/\rho)$ and the parameter values are those given in Table 2. For a particular value of the intracluster correlation, ρ, a single finite population was thus generated and then fixed and repeatedly sampled from. Each simulation consisted of selecting $n_h = 2$ clusters with replacement from stratum h for $h = 1, \ldots, H$ and enumerating the entire cluster. Each ultimate unit in the obtained cluster was independently declared a respondent or nonrespondent with probability p and $(1 - p)$ respectively, *i.e.*, uniform response. The nonrespondents were then imputed both using random imputation and adjusted random imputation and the population total and distribution function, for various values of $F(t)$, were estimated. Two values of ρ, 0.1 and 0.3, and two values of p, 0.6 and 0.8, were considered. Note that the first-stage sampling fraction is quite small (0.064), so that with-replacement and without replacement sampling are essentially equivalent.

To compare the performance of the different variance estimators we calculated the percent relative bias and relative instability for each, defined as

$$\%\mathrm{RB} = \frac{100}{S}\sum_{s=1}^{S} \upsilon_s(\hat{\theta}_I)/\mathrm{MSE}(\hat{\theta}_I)$$

and

$$\mathrm{RI} = \left\{\frac{1}{S}\sum_{s=1}^{S}[\upsilon_s(\hat{\theta}_I) - \mathrm{MSE}(\hat{\theta}_I)]^2\right\}^{1/2} \Big/ \mathrm{MSE}(\hat{\theta}_I),$$

respectively, where the number of simulation runs was $S = 5{,}000$ and the true $\mathrm{MSE}(\hat{\theta}_I)$ was obtained through an independent set of 50,000 simulation runs. The bootstrap variance estimators were each based on $B = 2{,}000$ bootstrap resamples. We obtain results for estimating the variance of $\hat{\theta}_I$ equal to the imputed total and the imputed distribution function using: (i) the repeated half-sample bootstrap with proper Monte Carlo approximation, υ_B, as in equation (8) and with improper Monte Carlo approximation replacing $\bar{\theta}^*_{I(\cdot)}$ with $\hat{\theta}_I$, denoted υ_{B2}; and (ii) the proper repeated BRR, υ_{BRR}, as in equation (9) and the improper repeated BRR replacing $\bar{\theta}_{I(\cdot)}$ with $\hat{\theta}_I$, denoted υ_{BRR2}.

Table 3 summarizes the results for percent relative bias using random imputation and adjusted random imputation. Note that adjusted random imputation is not presented for estimating the population total, Y, as adjusted random imputation removes the imputation variance from the estimator and thus simpler methods of variance estimation are available (Chen *et al.* 2000). It is clear from the high %RB for υ_{B2} and υ_{BRR2} that one must not replace $\bar{\theta}_{I(\cdot)}$ and $\bar{\theta}^*_{I(\cdot)}$ by $\hat{\theta}_I$ in the bootstrap or the BRR, respectively. It is also clear that both the repeated half-sample bootstrap and the repeated BRR variance estimators, υ_B and υ_{BRR} have negligible bias when properly applied.

Table 2
Parameters of the Finite Population

h	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	200	20.0	17	31	150	15.0
2	16	175	17.5	18	31	140	14.0
3	20	150	15.0	19	31	130	13.0
4	25	190	19.0	20	34	120	12.0
5	25	165	16.5	21	34	110	11.0
6	25	190	19.0	22	34	100	10.0
7	25	180	18.0	23	34	150	15.0
8	28	170	17.0	24	37	125	12.5
9	28	160	16.0	25	37	100	10.0
10	28	180	18.0	26	37	150	15.0
11	31	170	17.0	27	37	125	12.5
12	31	160	16.0	28	39	100	10.0
13	31	150	15.0	29	39	75	7.5
14	31	180	18.0	30	42	75	7.5
15	31	170	17.0	31	42	75	7.5
16	31	160	16.0	32	42	75	7.5

Given the results of Table 3, we consider relative instability, RI, only for υ_B and υ_{BRR}. We also restrict our presentation to $\rho = 0.3$ and $p = 0.6$ as the RI results were qualitatively the same in the other three cases. These results are given in Table 4. As one can see, though the differences are small, υ_B is slightly more stable than υ_{BRR}. This was generally the case for all values of ρ and p. We also included the adjusted jackknife of Rao and Shao (1992) and the adjusted BRR of Shao *et al.* (1998) in simulations for $\theta = Y$ and υ_B again was uniformly more stable. For example, with $\rho = 0.3$ and $p = 0.6$ as in Table 4, RI for the adjusted jackknife and the adjusted BRR were both 0.27. This may be because the reimputation approach has an advantage in estimating the component of the variance due to the imputation against the adjustment approach, provided the resample size is large enough to eliminate Monte Carlo error as is the case in our simulations. But, when the number of reimputations is moderate (like in the BRR with reimputation or the bootstrap with $B = 1{,}000$), this advantage is not entirely realized.

Table 3
%RB for υ_B, υ_{B2}, υ_{BRR} and υ_{BRR2}

	Random imputation				Adjusted random imputation			
Estimand	υ_{BRR}	υ_{BRR2}	υ_B	υ_{B2}	υ_{BRR}	υ_{BRR2}	υ_B	υ_{B2}
				$\rho = 0.1$ and $p = 0.6$				
Y	0.00	21.54	0.79	21.60				
$F(t) = 0.0625$	– 1.09	15.92	– 0.52	15.88	0.46	19.64	1.24	19.51
$F(t) = 0.2500$	– 0.13	19.44	0.62	19.55	0.85	14.86	1.80	15.08
$F(t) = 0.5000$	– 0.36	21.68	0.52	21.55	0.55	10.73	1.24	10.76
$F(t) = 0.7500$	– 0.84	19.89	0.13	20.09	– 0.36	10.98	0.54	11.31
$F(t) = 0.9375$	0.05	21.92	0.57	21.66	0.81	19.12	1.39	18.91
				$\rho = 0.1$ and $p = 0.8$				
Y	– 0.63	15.06	0.36	15.37				
$F(t) = 0.0625$	– 1.99	10.30	– 1.72	10.16	– 1.65	10.97	– 1.08	11.13
$F(t) = 0.2500$	– 1.27	13.65	– 0.88	13.30	– 0.95	8.89	– 0.52	8.81
$F(t) = 0.5000$	– 0.72	15.26	0.02	15.26	– 0.12	6.58	0.25	6.53
$F(t) = 0.7500$	– 0.37	14.50	0.57	14.76	0.36	7.56	1.05	7.81
$F(t) = 0.9375$	– 0.14	16.16	0.75	16.36	0.56	13.04	1.22	13.08
				$\rho = 0.3$ and $p = 0.6$				
Y	0.25	21.34	0.78	21.09				
$F(t) = 0.0625$	– 1.39	11.45	– 0.86	11.37	– 0.35	15.38	0.64	15.64
$F(t) = 0.2500$	– 0.41	19.89	0.14	19.73	1.23	13.79	1.71	13.62
$F(t) = 0.5000$	– 0.10	20.25	0.37	19.89	0.29	8.97	0.78	8.88
$F(t) = 0.7500$	– 1.40	16.70	– 0.49	16.89	– 0.75	9.24	0.07	9.49
$F(t) = 0.9375$	0.71	17.78	1.03	17.57	0.91	15.07	1.34	15.04
				$\rho = 0.3$ and $p = 0.8$				
Y	0.01	15.22	0.93	15.51				
$F(t) = 0.0625$	– 1.09	7.54	– 0.56	7.69	– 1.24	8.64	– 0.35	9.07
$F(t) = 0.2500$	– 0.44	15.22	– 0.08	14.99	– 0.23	8.18	0.29	8.23
$F(t) = 0.5000$	0.05	14.92	0.71	14.84	0.43	6.21	0.86	6.20
$F(t) = 0.7500$	0.13	12.54	0.86	12.70	0.81	6.85	1.26	6.99
$F(t) = 0.9375$	1.62	13.13	2.06	13.01	1.86	11.04	2.34	11.02

Table 4
RI for υ_B and υ_{BRR} with $\rho = 0.3$ and $p = 0.6$

Estimand	Random imputation		Adjusted random imputation	
	υ_{BRR}	υ_B	υ_{BRR}	υ_B
Y	0.27	0.23		
$F(t) = 0.0625$	0.60	0.59	0.57	0.56
$F(t) = 0.2500$	0.35	0.32	0.37	0.35
$F(t) = 0.5000$	0.27	0.23	0.28	0.26
$F(t) = 0.7500$	0.29	0.26	0.30	0.28
$F(t) = 0.9375$	0.48	0.46	0.48	0.46

7. Conclusion

We proposed repeated half-sample bootstrap and balanced repeated replication methods for variance estimation in the presence of random imputation that capture the imputation variance by reimputing for each replication using the same random imputation method as in the original sample. These repeated half-sample methods are valid in stratified multi-stage sampling, even when the number of psu's sampled in each stratum is very small, *e.g.*, 2. The key is that these methods use a stratum resample size that is equal to the original sample size without resorting to rescaling. These provide a unified method that works irrespective of the imputation method (random or non-random), the stratum size (small or large), the type of estimator (smooth or nonsmooth), or the type of problem (variance estimation or sampling distribution estimation). It is important to note that using reimputation to capture the imputation variance requires that one take greater care in the definition of the BRR and the Monte Carlo approximation to the bootstrap variance. In both cases it is important to use the mean of the replicates in the definition as opposed to replacing it with the estimator applied to the original sample.

Acknowledgements

Hiroshi Saigo was supported by grants from the Promotion and Mutual Aid Corporation for Private Universities of Japan and the Japan Economic Research Foundation. Jun Shao was supported by National Science Foundation Grant DMS-0102223, and National Security Agency Grant MDA904-99-1-0032. Randy R. Sitter was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. The authors would like to thank all referees for their helpful comments and suggestions.

References

Chen, J., Rao, J.N.K. and Sitter, R.R. (2000). Adjusted imputation for missing data in complex surveys. *Statistics Sinica*, 10, 1153-1169.

Efron, B. (1994). Missing data, imputation, and the bootstrap. *Journal of the American Statistical Association*, 89, 463-479.

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.

Mantel, H.J., and Singh, A.C. (1991). Standard errors of estimates of low proportions: A proposed methodology. Technical Report, Statistics Canada.

McCarthy, P.J. (1969). Pseudoreplication half samples. *Review of the International Statistical Institute*, 37, 239-264.

Nigam, A.K., and Rao, J.N.K. (1996). On balanced bootstrap, for stratified multistage samples. *Statistica Sinica*, 6, 199-214.

Rao, J.N.K. (1993). Linearization variance estimators under imputation for missing data. Technical Report, Laboratory for Research in Statistics and Probability, Carleton University.

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 Sitter, R.R. (1995). Variance estimation under two-phase sampling with application to imputation for missing data. *Biometrika*, 82, 453-460.

Rao, J.N.K., and Wu, C.F.J. (1985). Inference from stratified samples: Second order analysis of three methods for non-linear statistics. *Journal of the American Statistical Association*, 80, 620-630.

Rao, J.N.K., and Wu, C.F.J. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83, 231-241.

Rubin, D.B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Rubin, D.B., and Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association*, 81, 366-374.

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.

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

Shao, J., and Sitter, R.R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91, 1278-1288.

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

Sitter, R.R. (1993). Balanced repeated replications based on orthogonal multi-arrays. *Biometrika*, 80, 211-221.

Sitter, R.R. (1997). Variance estimation for the regression estimator in two-phase sampling. *Journal of the American Statistical Association*, 92, 780-787.