Statistica Sinica **10**(2000), 1153-1169

EFFICIENT RANDOM IMPUTATION FOR MISSING DATA IN COMPLEX SURVEYS

J. Chen, J. N. K. Rao and R. R. Sitter*

University of Waterloo, Carleton University and Simon Fraser University

Abstract: A simple adjusted random imputation method for handling item nonresponse in complex surveys is presented. This method eliminates the imputation variance of the estimator of a mean or total, and at the same time preserves the distribution of item values. Jackknife and bootstrap variance estimators that depend only on the reported values in the data file are also proposed. It is necessary to identify the respondent and imputed values in the data file as well as the imputation class. Simulation results on the performance of the proposed method in estimating a total and distribution function are also presented.

Key words and phrases: Adjusted imputation, bootstrap, hotdeck, jackknife, mean imputation, stratified multistage sampling.

1. Introduction

Item nonresponse occurs frequently in sample surveys with many items. It is usually handled by some form of imputation to fill in the missing values. An advantage of imputation is that the same sampling weight is used for all the items, unlike the weight-adjustment method which is typically used for unit nonresponse. Commonly used imputation methods include deterministic imputation, such as mean imputation within imputation classes, and stochastic imputation, such as random imputation within classes. Deterministic imputation eliminates imputation variance of the estimator of a mean or total, but the distribution of item values is not preserved. For example, mean imputation leads to a spike at the point $\bar{y}_r$, the mean of the respondent values. On the other hand, random imputation preserves the distribution, but leads to imputation variance which can be a significant component of the total variance if the item response rate is not high.

In this paper we present a simple adjusted random imputation method that eliminates the imputation variance of the estimator of a mean or total, and at the same time preserves the distribution of item values, and in fact estimates of the distribution function based on the imputed data are shown to remain consistent and asymptotically normal. Though the method does not entirely eliminate

*Names of authors are in alphabetical order to indicate equal contribution.

imputation variance from the estimated distribution function, it is shown to significantly reduce it. This method is valid for general stratified multistage designs. Jackknife and bootstrap variance estimators are also considered. We assume that the imputed and respondent values are identified in the data file as well as the imputation class. Uniform response within classes is also assumed; that is, within an imputation class equal response probabilities on the item y and independent response across sampled units, but response probabilities can vary across classes.

Section 2 studies the case of simple random sampling. Results for stratified multistage designs are given in Section 3. Section 4 reports the results of a simulation study on the performance of the proposed method in estimating a total, population variance and distribution function. For simplicity of notation, we consider only the case of a single imputation class, but the results readily extend to multiple imputation classes formed on the basis of auxiliary variables observed on all the sampled units.

2. Simple Random Sampling

To fix ideas, we first consider simple random sampling. Suppose in a simple random sample, s, of size n, r elements, s_r, respond and m elements, s_m, do not respond to an item y. Let y_i^*, $i \in s_m$ be the imputed values for the missing data, based on the donor set $\{y_i, i \in s_r\}$. The imputed estimator of the population mean $\bar{Y} = (y_1 + \cdots + y_N)/N$ is then given by

$$\bar{y}_I = \frac{1}{n}(\sum_{s_r} y_i + \sum_{s_m} y_i^*). \tag{2.1}$$

Note that the same weight, $1/n$, is used in (2.1) for all items y_i.

2.1. Mean imputation

Mean imputation uses $\bar{y}_r$ as the imputed value, i.e., $y_i^* = \bar{y}_r$ for all $i \in s_m$. In this case $\bar{y}_I$ has no imputation variance because y_i^* is deterministic given s_r. We have $\bar{y}_I = \bar{y}_r$ for mean imputation, and it is unbiased for $\bar{Y}$ under uniform response.

The sample variance under mean imputation reduces to

$$s_{yI}^2 = \frac{1}{n-1}[\sum_{s_r}(y_i - \bar{y}_I)^2 + \sum_{s_m}(y_i^* - \bar{y}_I)^2] = \left(\frac{r-1}{n-1}\right) s_{yr}^2,$$

where s_{yr}^2 is the sample variance of respondents. Under uniform response, given r, the sample of respondent values is a simple random sample of size r from the finite population. Therefore,

$$E(s_{yI}^2|r) = \frac{r-1}{n-1} S_y^2 \doteq \frac{r}{n} S_y^2, \tag{2.2}$$

where S_y^2 is the population variance. It follows from (2.2) that the sample variance under mean imputation seriously underestimates S_y^2 when the response rate, r/n, is not high, i.e., the variability of item values is understated due to a spike at the point $\bar{y}_r$.

2.2. Random imputation

Random imputation selects a simple random sample of size m with replacement from s_r and then uses the associated y-values as donors, that is, $y_i^* = y_j$ for some $j \in r$. We have $E_* y_i^* = \bar{y}_r$ and $E_* \bar{y}_I = \bar{y}_r$, where E_* is the expectation under imputation given s_r. It follows that $\bar{y}_I$ for random imputation is also unbiased under uniform response. But the variance of $\bar{y}_I$, given r, is now larger than the variance under mean imputation by the factor $1 + \hat{p}\hat{q}$, where $\hat{p} = r/n$ is the observed response rate and $\hat{q} = 1 - \hat{p}$. The relative contribution of imputation variance is equal to $\hat{p}\hat{q}$ with a maximum of 25%.

The sample variance, s_{yI}^2, is approximately unbiased for S_y^2, so that the variability of item values is preserved under random imputation.

Rao and Shao (1992) proposed a jackknife variance estimator, v_J, of $\bar{y}_I$ for mean or random imputation which is approximately design-unbiased. It is calculated in the usual way except that, whenever a respondent $j \in s_r$ is to be deleted, each of the imputed values, y_i^*, is adjusted by an average amount $E_*^{(j)} y_i^* - E_* y_i^* = \bar{y}_r(j) - \bar{y}_r$ to reflect the fact that the donor set is changed, where $\bar{y}_r(j) = (r\bar{y}_r - y_j)/(r-1)$ and $E_*^{(j)}$ is the imputation expectation when the donor set excludes respondent j. The imputed values, y_i^*, remain unchanged whenever a nonrespondent, $j \in s_m$, is to be deleted because the donor set is unchanged. Note that we need identification flags to locate observed and imputed values in the data file and compute $\bar{y}_I^a(j)$, the imputed estimator based on the respondent values and the modified imputed values, for $j \in s_r$ and $j \in s_m$.

The jackknife variance estimator of $\bar{y}_I$ is given by

$$v_J = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I^a(j) - \bar{y}_I]^2 \doteq \frac{1}{n}(s_{yI}^2 - s_{yr}^2) + \frac{s_{yr}^2}{n}(1 + \hat{p}\hat{q}). \tag{2.3}$$

For mean imputation, $y_i^* = \bar{y}_r$, we have $\bar{y}_I^a(j) = \bar{y}_r(j)$ for $j \in s_r$ and $\bar{y}_I^a(j) = \bar{y}_r$ for $j \in s_m$.

2.3. Adjusted random imputation

The proposed adjusted random imputation simply uses $\tilde{y}_i = \bar{y}_r + (y_i^* - \bar{y}_m^*)$ as imputed values in the data file instead of y_i^* for $i \in s_m$, where $\bar{y}_m^*$ is the mean of y_j^* for $j \in s_m$, obtained from random imputation. The imputed estimator is then given by

$$\bar{y}_I = \frac{1}{n}\Big(\sum_{s_r} y_i + \sum_{s_m} \tilde{y}_i\Big). \tag{2.4}$$

We can also express $\tilde{y}_i$ in terms of the residuals $\epsilon_i^* = y_i^* - \bar{y}_r$ as $\tilde{y}_i = \bar{y}_r + (\epsilon_i^* - \bar{\epsilon}_m^*) = \bar{y}_r + \tilde{\epsilon}_i$, where $\bar{\epsilon}_m^*$ is the mean of ϵ_j^* for $j \in s_m$. The imputed estimator, $\bar{y}_I$, reduces to $\bar{y}_r$, noting that $\sum_{s_m} \tilde{\epsilon}_i = 0$. Therefore, imputation variance is eliminated by using $\tilde{y}_i$ instead of y_i^* for $i \in s_m$.

The sample variance under adjusted random imputation, viz.,

$$s_{yI}^2 = \frac{1}{n-1}[\sum_{s_r}(y_i - \bar{y}_I)^2 + \sum_{s_m}(\tilde{y}_i - \bar{y}_I)^2],$$

is approximately unbiased for S_y^2, noting that $E_* s_{yI}^2 = s_{yr}^2$. Therefore, the variability of item values is preserved under adjusted imputation.

Turning to jackknife variance estimation, we first note that y_i^* is modified to $z_i^*(j) = y_i^* + \bar{y}_r(j) - \bar{y}_r$ if $j \in s_r$ is deleted, and it remains unchanged if $j \in s_m$ is deleted, i.e., $z_i^*(j) = y_i^*$. Therefore, if $j \in s_r$ is deleted $\tilde{y}_i$ should be changed to

$$\tilde{z}_i(j) = \bar{y}_r(j) + \{z_i^*(j) - \frac{1}{m}\sum_{i \in s_m} z_i^*(j)\} = \tilde{y}_i - \frac{1}{r-1}(y_j - \bar{y}_r), \quad j \in s_r. \tag{2.5}$$

Similarly, if $j \in s_m$ is deleted $\tilde{y}_i$ should be changed to

$$\tilde{z}_i(j) = \bar{y}_r + \{z_i^*(j) - \frac{1}{m-1}\sum_{i \neq j; i \in s_m} z_i^*(j)\} = \tilde{y}_i + \frac{1}{m-1}(\tilde{y}_j - \bar{y}_r), \quad j \in s_m. \tag{2.6}$$

It is important to note that the adjusted values $\tilde{z}_i(j)$ given by (2.5) and (2.6) depend only on the reported values in the data file, viz., y_i, $i \in s_r$ and $\tilde{y}_i$, $i \in s_m$.

Denote the imputed estimator based on the respondent values y_i and the modified imputed values $\tilde{z}_i(j)$ as $\bar{y}_I^a(j)$ for $j \in s_r$ and $j \in s_m$. The jackknife variance estimator, v_J, is again given by (2.3) using these $\bar{y}_I^a(j)$-values and $\bar{y}_I$ given by (2.4). It is readily seen that $\bar{y}_I^a(j) = \bar{y}_r(j)$ for $j \in s_r$ and $\bar{y}_I^a(j) = \bar{y}_r$ for $j \in s_m$, so that v_J is identical to the jackknife variance estimator under mean imputation. This equivalence implies that v_J under adjusted random imputation is approximately design-unbiased.

2.4. Distribution function

The population distribution function is given by $F_N(t) = \sum_j I(y_j \leq t)/N$, where $I(\cdot)$ is the usual indicator function. The imputed estimator, $\hat{F}_I(t)$, is simply obtained from (2.1) by changing y_i for $i \in s_r$ to $I(y_i \leq t)$ and y_i^* for $i \in s_m$ to $I(y_i^* \leq t)$. Similarly, for the adjusted imputation we change y_i for $i \in s_r$ to $I(y_i \leq t)$ and $\tilde{y}_i$ for $i \in s_m$ to $I(\tilde{y}_i \leq t)$. Appendix 1 part (a) proves that the resulting estimator, $\tilde{F}_I(t)$, is consistent while Appendix 1 part (b) proves its asymptotic normality. The imputation variance is not completely eliminated in estimating $F_N(t)$ by the proposed method, but simulation results in

Section 4 indicate significant reduction in mean squared error (MSE) relative to random imputation. Appendix 1 part (c) gives a few cases where this reduction in variance is certain, the most important of these being when the finite population is generated from a normal superpopulation.

It is quite difficult to understand how one might adjust the jackknife in this situation, since deleting a respondent will affect the $\tilde{y}_i$ values inside the indicators as well as the number of indicators in the sum. Instead we consider the bootstrap method of Shao and Sitter (1996). This method is quite general and could be used for estimating the variance of $\bar{y}_I$ in Section 2.3 if desired. However, the jackknife is more stable for means and totals. We describe and investigate the performance of the bootstrap via simulation in Section 4.

It may be noted that $\hat{F}_I(t)$ for mean imputation will be seriously biased due to a spike at the point $I(y_i^* = \bar{y}_r \leq t)$.

Sarndal (1992) suggested adjusted random imputation similar to ours under a model-assisted approach, but he did not consider its advantages in eliminating imputation variance while preserving the distribution of item values, nor did he study variance estimators that depend only on the reported values in the data file.

3. Stratified Multistage Sampling

Consider now the case of stratified multistage surveys. We restrict to designs in which the first-stage units or clusters are selected with replacement or are so treated for variance estimation, with independent subsamples taken within clusters which are selected more than once. Suppose n_h clusters are selected with probabilities p_{hi} with replacement or with inclusion probabilities $\pi_{hi} = n_h p_{hi}$ independently in each stratum. In the case of complete response on item y, let $\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}/(n_h p_{hi})$ be a linear unbiased estimator of the stratum total Y_h, where $\hat{Y}_{hi}$ is a linear unbiased estimator of the stratum 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

$$\hat{Y} = \sum_{(hik)\in s} w_{hik} y_{hik}, \tag{3.1}$$

where s is the total sample of elements, and w_{hik} and y_{hik} respectively denote the sampling weight and the item value attached to the (hik)th sampled element $(k = 1, \ldots, n_{hi}; i = 1, \ldots, n_h; h = 1, \ldots, L)$.

To construct a jackknife variance estimator of $\hat{Y}$, we need to recalculate the weights w_{hik} each time a sample cluster gj is deleted $(j = 1, \ldots, n_g; h = 1, \ldots, L)$. This is done in a straightforward manner as follows: $w_{hik(gj)} = w_{hik} b_{gj}$, where $b_{gj} = 0$ if $(hi) = (gj)$; $= n_g/(n_g - 1)$ if $h = g$ and $i \neq j$; $= 1$ if $h \neq g$. Replacing

w_{hik} by the jackknife weights $w_{hik(gj)}$ in (3.1) we get $\hat{Y}_{(gj)}$, and the jackknife variance estimator is given by

$$v_J = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}_{(gj)} - \hat{Y})^2. \tag{3.2}$$

Suppose now that a subsample s_m of elements do not respond on item y, and we impute the values y^*_{hik} for the missing data based on respondents (donors) s_r. The imputed estimator of Y is then given by

$$\hat{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} y^*_{hik}. \tag{3.3}$$

Note that the same weight w_{hik} is used in (3.3) for all items y attached to the (hik)th sample element.

For jackknife variance estimation under imputation, we need to adjust y^*_{hik} by an average amount $E_*^{(gj)} y^*_{hik} - E_* y^*_{hik}$, where E_* denotes expectation with respect to imputation given s_r, and $E_*^{(gj)}$ denotes expectation with respect to imputation when the donor set is modified by excluding the respondents from sample cluster gj. The adjusted imputed values reflect the fact that the donor set is changed when a sample cluster is deleted. Denote the imputed estimator as $\hat{Y}^a_{I(gj)}$ when w_{hik} in (3.3) is replaced by $w_{hik(gj)}$ and y^*_{hik} by the adjusted imputed value $y^*_{hik} + E_*^{(gj)} y^*_{hik} - E_* y^*_{hik} = z^*_{hik(gj)}$. The jackknife variance estimator is then given by (3.2) with $\hat{Y}_{(gj)}$ changed to $\hat{Y}^a_{I(gj)}$ and $\hat{Y}$ to $\hat{Y}_I$:

$$v_J = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}^a_{I(gj)} - \hat{Y}_I)^2. \tag{3.4}$$

3.1. Mean imputation

Mean imputation uses $\hat{S}/\hat{T} = \sum_{s_r} w_{hik} y_{hik} / \sum_{s_r} w_{hik}$ as the imputed value, i.e. $y^*_{hik} = \hat{S}/\hat{T}$ for all $(hik) \in s_m$. In this case $\hat{Y}_I$ reduces to

$$\hat{Y}_I = (\hat{S}/\hat{T})\hat{U}, \tag{3.5}$$

where $\hat{U} = \sum_s w_{hik}$. This ratio estimator has no imputation variance and is approximately unbiased under uniform response. Note $(\hat{S}/\hat{T})\hat{U}$ is the weight-adjusted estimator $\hat{Y}_w$.

For mean imputation, $z^*_{hik(gj)} = \hat{S}_{(gj)}/\hat{T}_{(gj)}$, where $\hat{S}_{(gj)}$ and $\hat{T}_{(gj)}$ are obtained from $\hat{S}$ and $\hat{T}$ using $w_{hik(gj)}$ instead of w_{hik}. In this case $\hat{Y}^a_{I(gj)}$ reduces to $[\hat{S}_{(gj)}/\hat{T}_{(gj)}]\hat{U}_{(gj)}$, where $\hat{U}_{(gj)}$ is given by $\hat{U}$ with w_{hik} changed to $w_{hik(gj)}$. It

follows from Rao and Shao (1992) that the jackknife variance estimator, v_J, is design-consistent under uniform response.

As in the case of simple random sampling, mean imputation does not preserve the distribution of item values due to a spike at the point $\hat{S}/\hat{T}$.

3.2. Random imputation

Random imputation selects the donors $(gjl) \in s_r$ with replacement with probabilities $w_{hik}/\hat{T}$ and uses $y^*_{hik} = y_{gjl}$ to get $\hat{Y}_I$. In this case $E_*\hat{Y}_I = (\hat{S}/\hat{T})\hat{U}$, the estimator under mean imputation. The variance of $\hat{Y}_I$ is larger than the variance under mean imputation because of the random imputation. However, it preserves the distribution of item values.

The adjusted imputed values for random imputation are given by $z^*_{hik(gj)} = y^*_{hik} + \hat{S}_{(gj)}/\hat{T}_{(gj)} - \hat{S}/\hat{T}$. The resulting jackknife variance estimator, v_J, is design-consistent under uniform response (Rao and Shao (1992)).

3.3. Adjusted random imputation

The proposed adjusted random imputation simply uses

$$\tilde{y}_{hik} = \hat{S}/\hat{T} + (y^*_{hik} - \sum_{s_m} w_{hik} y^*_{hik} / \sum_{s_m} w_{hik}) \tag{3.6}$$

as imputed values in the data file instead of y^*_{hik} for $(hik) \in s_m$, where y^*_{hik} are the imputed values under random imputation. We may also express $\tilde{y}_{hik}$ in terms of the residuals $\epsilon^*_{hik} = y^*_{hik} - \hat{S}/\hat{T}$ as $\tilde{y}_{hik} = \hat{S}/\hat{T} + (\epsilon^*_{hik} - \bar{\epsilon}^*_m) = \hat{S}/\hat{T} + \tilde{\epsilon}_{hik}$, where $\bar{\epsilon}^*_m = \sum_{s_m} w_{hik}\epsilon^*_{hik} / \sum_{s_m} w_{hik}$. The imputed estimator is given by

$$\tilde{Y}_I = \sum_{s_r} w_{hik} y_{hik} + \sum_{s_m} w_{hik} \tilde{y}_{hik} \tag{3.7}$$

which reduces to (3.5), the estimator under mean imputation, noting that $\sum_{s_m} w_{hik}\tilde{\epsilon}_{hik} = 0$. Therefore, the method preserves the distribution of item values, but imputation variance is eliminated by using $\tilde{y}_{hik}$ instead of y^*_{hik} for $(hik) \in s_m$. Note that the same weight w_{hik} is used for all item values y.

Turning to jackknife variance estimation, we first note that y^*_{hik} is modified to $z^*_{hik(gj)} = y^*_{hik} + \hat{S}_{(gj)}/\hat{T}_{(gj)} - \hat{S}/\hat{T}$ if the (gj)th sample cluster is deleted. Therefore $\tilde{y}_{hik}$ should be changed, using the jackknife weights $w_{hik(gj)}$, to

$$\begin{aligned} \tilde{z}_{hik(gj)} &= \hat{S}_{(gj)}/\hat{T}_{(gj)} + (z^*_{hik(gj)} - \sum_{s_m} w_{hik(gj)} z^*_{hik(gj)} / \sum_{s_m} w_{hik(gj)}) \\ &= \hat{S}_{(gj)}/\hat{T}_{(gj)} + \tilde{y}_{hik} - \sum_{s_m} w_{hik(gj)} \tilde{y}_{hik} / \sum_{s_m} w_{hik(gj)}. \end{aligned} \tag{3.8}$$

The form (3.8) follows from (3.6). Note that (3.8) depends only on the item values reported in the data file, i.e., y_{hik}, $(hik) \in s_r$ and $\tilde{y}_{hik}$, $(hik) \in s_m$. It now follows from (3.8) that $\tilde{Y}^a_{I(gj)} = \sum_{s_r} w_{hik(gj)} y_{hik} + \sum_{s_m} w_{hik(gj)} \tilde{z}_{hik(gj)} = [\hat{S}_{(gj)}/\hat{T}_{(gj)}]\hat{U}_{(gj)}$, which is identical to $\hat{Y}^a_{I(gj)}$ for mean imputation. The jackknife variance estimator, v_J, is given by (3.2) with $\hat{Y}_{(gj)}$ changed to $\tilde{Y}^a_{I(gj)}$ and $\hat{Y}$ to $\tilde{Y}_I$. Since v_J is algebraically equivalent to v_J for mean imputation, it now follows that the jackknife variance estimator for adjusted random imputation is also design-consistent under uniform response.

3.4. Distribution function

The population distribution function is estimated by $\hat{F}_N(t) = \sum_s w_{hik} I(y_{hik} \leq t)/\sum_s w_{hik}$, in the case of complete response. The numerator of the imputed estimator $\hat{F}_I(t)$ is simply obtained from (3.3) by changing y_{hik} for $(hik) \in s_r$ to $I(y_{hik} \leq t)$ and y^*_{hik} for $(hik) \in s_m$ to $I(y^*_{hik} \leq t)$ for mean and random imputation; the denominator remains unchanged. Similarly, for the adjusted random imputation we change y_{hik} for $(hik) \in s_r$ to $I(y_{hik} \leq t)$ and $\tilde{y}_{hik}$ for $(hik) \in s_m$ to $I(\tilde{y}_{hik} \leq t)$ to get $\tilde{F}_I(t)$. The imputed estimator can be seriously biased for mean imputation due to a spike at the point $I(y^*_{hik} = \hat{S}/\hat{T} \leq t)$. On the other hand, it is approximately unbiased for random imputation under uniform response. Appendix 2 proves consistency and asymptotic normality of $\tilde{F}_I(t)$ for adjusted random imputation under some regularity conditions. Simulation results in Section 4 verify that the relative bias is generally small for finite sample size. The imputation variance is not completely eliminated in estimating $F(t)$ by the proposed method, but simulation results in Section 4 indicate significant reduction in MSE relative to random imputation.

The jackknife variance estimator of $\hat{F}_I(t)$ for random imputation is obtained from (3.2) by changing $\hat{Y}_{(gj)}$ and $\hat{Y}$ to $\hat{F}^a_{I(gj)}(t)$ and $\hat{F}_I(t)$, where $\hat{F}^a_{I(gj)}(t)$ is calculated by using indicator variables $I(y_{hik} \leq t)$ for $(hik) \in s_r$ and $I(y^*_{hik} \leq t)$ for $(hik) \in s_m$ in the formula for $\hat{Y}^a_{I(gj)}$ and then dividing by $\sum_s w_{hik(gj)}$. This jackknife variance estimator for random imputation is design-consistent under uniform response, following Rao and Shao (1992). It is difficult to understand how one might obtain an adjusted jackknife for $\tilde{F}_I(t)$ for adjusted random imputation. Instead we consider the bootstrap method of Shao and Sitter (1996) which is applicable to stratified multi-stage sampling. Simulation results in Section 4 indicate that the bootstrap variance estimator of $\tilde{F}_I(t)$ performs well.

4. Simulation Study

In this section we compare the proposed adjusted hot deck imputation method to the usual hot deck imputation method through a limited simulation. We first generated a finite population similar to that given in Hansen

and Tepping (1985) in the National Assessment of Educational Progress Study. The population consisted of $L = 32$ strata, with N_h clusters in stratum h and $M_h = 10$ ultimate units in each cluster. To create the population of y_{hik}, we first generated y_{hi} i.i.d. with $E(y_{hi}) = \mu_h$ and $V(y_{hi}) = v_h^2$ for $h = 1, \ldots, L$ and $i = 1, \ldots, N_h$ and then independently generated $u_{hik} \overset{iid}{\sim} N(0, \sigma_{uh}^2)$, $k = 1, \ldots, 20$ with $\sigma_{uh}^2 = (1-\rho)v_h^2/\rho$ and $\rho > 0$. We considered both a normal distribution for the y_{hi} and a shifted gamma distribution. The results using the normal and shifted gamma were qualitatively the same and thus only the results for the normal are presented. The ultimate units are then defined by $y_{hik} = y_{hi} + u_{hik}$. Note that $V(y_{hik}) = v_h^2/\rho$. The parameter values for the finite population are given in Table 1.

Table 1. Parameters of the finite population.

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

To obtain a sample, we drew a simple random sample with replacement of size $n_h = 2$ clusters from stratum h. Whenever a cluster is selected, all of the ultimate units within the cluster were selected. Thus the total sample is of size $n = 32 \cdot 2 \cdot 10 = 640$. Independent uniform(0,1) random variables r_{hik} were generated for each (hik). If r_{hik} is less than or equal to the chosen response rate, then $(hik) \in s_r$, otherwise $(hik) \in s_m$.

We then considered two methods of imputation, random imputation as described in Section 3.2 and the proposed adjusted random imputation method as described in Section 3.3. In this setting $w_{hik} = N_h/n_h = N_h/2$.

4.1. Population total

For each ρ and response rate combination, the finite population was created and $A = 10,000$ independent stratified cluster samples were drawn with observations missing at random as described above. The missing values were imputed using random imputation and adjusted random imputation and estimated totals were calculated for each, yielding $\hat{Y}_I$ and $\tilde{Y}_I$ respectively. Note that mean imputation would yield the same estimated total as adjusted random imputation and is thus excluded. The simulated percentage relative bias and mean square error of each estimator $\hat{\theta}$ were calculated as

$$\mathrm{RB}(\hat{\theta}) = 100 * (\bar{\theta}_{(\cdot)} - \theta)/\theta \tag{4.1}$$

and

$$\mathrm{MSE}(\hat{\theta}) = \frac{1}{A}\sum_{a=1}^{A}\{\hat{\theta}^{(a)} - \theta\}^2, \tag{4.2}$$

where $\bar{\theta}_{(\cdot)} = \sum_a \hat{\theta}^{(a)}/A$ and $\hat{\theta}^{(a)}$ is the value of the particular estimate $\hat{\theta}$ of θ for the a-th simulation run. To compare random imputation to adjusted random imputation we also calculated relative efficiencies $\mathrm{RE}(\tilde{Y}_I) = \mathrm{MSE}(\tilde{Y}_I)/\mathrm{MSE}(\hat{Y}_I)$ and $\mathrm{RE}(v_J(\tilde{Y}_I)) = \mathrm{MSE}(v_J(\tilde{Y}_I))/\mathrm{MSE}(v_J(\hat{Y}_I))$. Table 2 gives the RB and RE of $\tilde{Y}_I$ for various values of response rate and ρ. One can see that the RB is negligible in all cases and that by eliminating the variation due to random imputation, the proposed adjusted random imputation method reduces the MSE by as much as 20% depending on the correlation and the response rate. The gains increase as the response rate decreases and as ρ decreases.

Table 2. *RB* and *RE* for $\tilde{Y}_I$.

Corr.	*RB* (in %)				*RE*			
	Response Rate				Response Rate			
ρ	.5	.6	.7	.8	.5	.6	.7	.8
.10	0.02	-0.03	0.03	0.02	0.79	0.79	0.82	0.85
.30	0.02	-0.02	0.02	0.02	0.81	0.80	0.82	0.85
.50	0.02	-0.02	0.02	0.02	0.81	0.81	0.82	0.85

In a similar fashion, we investigated the performance of the jackknife variance estimator of $\tilde{Y}_I$ in each case as presented in Table 3. The absolute percentage relative biases of the jackknife variances estimator of $\tilde{Y}_I$ were less than 2.7% in all cases and less than 1% in most cases. Table 3 also illustrates that the jackknife variance estimator of $\tilde{Y}_I$ leads to significantly smaller MSE, as is demonstrated by RE ranging between 0.63 and 0.75.

Table 3. RB and RE for $v_J(\hat{Y}_I)$).

Corr.	RB (in %)				RE			
	Response Rate				Response Rate			
ρ	.5	.6	.7	.8	.5	.6	.7	.8
.10	0.46	0.00	2.65	-1.00	0.63	0.65	0.69	0.73
.30	-0.68	0.99	2.00	-0.68	0.66	0.67	0.69	0.74
.50	-1.04	1.42	1.51	-0.74	0.67	0.68	0.69	0.75

4.2. Distribution function

The same set of simulations were used to investigate the performance of estimators of the population distribution function, $F(t)$, for various values of t corresponding to fixed percentiles. The weight adjustment estimator, $\hat{F}_W(t)$, mean imputation estimator, $\hat{F}_M(t)$, random imputation estimator, $\hat{F}_I(t)$, and adjusted random imputation estimator, $\tilde{F}_I(t)$, were all considered; note that $\hat{F}_W(t)$ is obtained from $\hat{Y}_W$ by changing y_{hik} to $I(y_{hik} \leq t)$. The simulated percentage relative biases and mean square errors were obtained using (4.1) and (4.2) with $\hat{\theta}$ replaced by $\hat{F}(t)$ and θ by $F(t)$.

Table 4 shows that the percentage relative biases of $\tilde{F}_I(t)$ were less than 1% in all cases. This was not the case for $\hat{F}_M(t)$ as the "spike" at $\hat{S}/\hat{T}$ induces large biases for some values of t, as high as 20-50%. These are not presented to save space. Table 4 also shows that $\tilde{F}_I(t)$ yields significant reductions in MSE relative to $\hat{F}_I(t)$ as demonstrated by RE as low as 0.85.

We also investigated the performance of the bootstrap variance estimator of Shao and Sitter (1996) in a few cases. This method selects bootstrap samples of units $i \in s$ using any bootstrap which is consistent for complete response. The respondents in each bootstrap sample are then used to re-impute the non-respondents using the same imputation method as in the original sample. In our case, we selected $n_h^* = n_h - 1$ clusters with replacement from the n_h clusters in each stratum and then used random imputation from the bootstrap donors. This procedure is repeated large number of times, B, and the variance of $\tilde{F}_I(t)$ is estimated by $v_B = B^{-1} \sum_{b=1}^{B} (\tilde{F}_I^{(b)} - \bar{\tilde{F}}_I^{(\cdot)})^2$, where $\tilde{F}_I^{(b)}$ is the estimated distribution function using the bth re-imputed bootstrap sample and $\bar{\tilde{F}}_I^{(\cdot)} = \sum_b \tilde{F}_I^{(b)}/B$. It is important with this re-imputed bootstrap method to use $\bar{\tilde{F}}_I^{(\cdot)}$ in v_B instead of, as is commonly done, $\tilde{F}_I(t)$ itself (see Saigo, Shao and Sitter (1999) for discussion on this point). We should also note that under random imputation, when n_h's are small (i.e. $n_h = 2$) this method has a positive bias since the size of the bootstrap sample which is used to re-impute is smaller (half the size when $n_h = 2$), than the original sample. Methods to adjust the bootstrap to correct this bias in a general setting are being considered.

Table 4. RB and RE for $\tilde{F}_I(t)$.

Corr.		RB (in %)				RE			
		Response Rate				Response Rate			
ρ	$F(t)$	.5	.6	.7	.8	.5	.6	.7	.8
.10	.0625	0.29	0.23	-0.14	-0.12	0.97	0.95	0.96	0.95
	.2500	-0.19	-0.05	-0.10	-0.13	0.90	0.90	0.90	0.91
	.5000	-0.10	0.04	-0.04	-0.03	0.85	0.85	0.86	0.89
	.7500	-0.03	0.03	-0.02	-0.03	0.86	0.87	0.89	0.91
	.9375	-0.04	0.06	0.01	0.02	0.92	0.91	0.93	0.94
.30	.0625	0.40	-0.08	-0.26	-0.03	0.99	0.97	0.97	0.97
	.2500	-0.54	-0.21	-0.29	-0.33	0.89	0.89	0.89	0.90
	.5000	-0.11	0.07	-0.06	0.04	0.86	0.86	0.87	0.89
	.7500	0.04	0.09	0.05	0.05	0.89	0.89	0.90	0.92
	.9375	0.02	0.05	0.00	-0.01	0.94	0.94	0.96	0.97
.50	.0625	-0.03	-0.49	-0.60	-0.30	1.00	0.98	0.98	0.98
	.2500	-0.29	-0.01	-0.13	-0.25	0.90	0.89	0.89	0.91
	.5000	-0.25	-0.12	-0.12	0.00	0.86	0.87	0.87	0.90
	.7500	0.01	0.08	0.02	0.02	0.91	0.91	0.92	0.93
	.9375	0.02	0.05	-0.02	-0.01	0.95	0.95	0.97	0.97

Investigation of the performance of $v_B(\tilde{F}_I)$ was done through a separate simulation. The true MSE of $\tilde{F}_I(t)$ was obtained from a simulation of 50,000 runs. Then an independent simulation of $A = 5,000$ runs using $B = 2,000$ bootstrap samples was performed.

Table 5 illustrates that for $n_h = 2$, moderate range of $F(t)$ and all ρ, response rate combinations the bootstrap variance estimator has small relative bias. However for extreme values of $F(t)$ when ρ and the response rate are both small, the bootstrap variance estimator of $\tilde{F}_I(t)$ yields larger relative biases. To investigate this further, we increased the sample size to $n_h = 4$ in each stratum. We see from Table 5 that the perfromance is greatly improved.

5. Concluding Remarks

A simple adjusted random imputation method for the case of item nonresponse in complex surveys was proposed. It removes imputation variance of estimated means and totals while preserving the distribution of items. In addition it reduces the imputation variance in estimated distribution functions.

There have been other attempts to reduce imputation variance in the literature. The review paper by Brick and Kalton (1996, Sec. 2.1.4) gives a good discussion of methods for reducing imputation variance. In the context of simple random sampling, Kalton and Kish (1984) suggested that donors may be selected by stratified sampling within imputation class or by systematic sampling from a

list of respondents ordered by their y-values. They noted that by stratifying on y the procedure can be very effective in reducing imputation varaince. However, it is not clear how one extends these methods to the case of stratified multistage sampling with unequal weights.

Table 5. The RB (in %) of v_B for $\tilde{F}_I(t)$.

Resp rate	$F(t)$	$n_h = 2$ Corr ρ .1	.3	.5	$n_h = 4$ Corr ρ .1	.3	.5
.60	.0625	17.2	10.5	8.6	6.5	5.6	5.3
	.2500	11.7	10.9	8.9	3.2	5.9	5.7
	.5000	10.0	8.8	8.5	3.7	2.3	2.8
	.7500	10.9	8.2	7.4	5.0	3.6	1.8
	.9375	16.8	12.9	10.8	6.2	3.3	2.9
.80	.0625	10.2	6.6	4.4	4.3	4.7	4.2
	.2500	6.8	6.4	5.0	2.3	1.8	0.0
	.5000	4.4	4.0	3.2	3.6	1.4	1.0
	.7500	6.7	5.4	4.3	2.1	1.4	0.8
	.9375	11.7	8.2	5.4	3.3	1.7	1.2

Another approach is to use fractional imputation which involves dividing respondent's values into parts and imputing separately to each part (Fay (1996); Kalton and Kish (1984)). This is similar to multiple imputation (Rubin (1987)) which also reduces imputation variance. However, both these methods are operationally less convenient than single imputation.

Acknowledgement

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada. The authors thank Hiroshi Saigo for his assistance with the bootstrap simulations.

Appendix 1. Asymptotic Properties of $\tilde{F}_I(t)$ for SRSWOR

In Appendix 1, we study asymptotic properties of the estimator of the distribution function using the proposed adjusted hot deck imputation method under simple random sampling without replacement. We assume that there is a sequence of sampling designs and a sequence of finite populations, indexed by ν. The sample size n_ν and the population size N_ν approach infinity as $\nu \to \infty$. We also assume uniform response and that the size, m_ν, of the non-respondent set s_m satisfies $m_\nu/n_\nu \to \alpha < 1$. All limiting processes are understood to be as $\nu \to \infty$, but the index ν is suppressed to simplify notation.

(a) Consistency of $\tilde{F}_I(t)$

Condition 1. As $N \to \infty$: S_y^2 goes to a positive constant limit; there exists a cdf $F(t)$ with continuous density function $f(t)$ such that $|F_N(t) - F(t)| = o(1)$; and for any $a_N = o(1)$, $\sup_{|\delta| \le a_N} |[F_N(t+\delta) - F_N(t)] - [F(t+\delta) - F(t)]| = o(N^{-1/2})$.

Theorem 1. *Under Condition* 1, $\tilde{F}_I(t)$ *is consistent.*

Proof. Rewrite $\tilde{y}_i = y_i^* - a_n$ for $i \in s_m$, where $a_n = (\bar{y}_m^* - \bar{y}_r)$ will be refered to as the adjustment factor. If we then let $F_r(t)$ and $F_m^*(t)$ be the empirical cdf based on y_i, $i \in s_r$ and y_i^*, $i \in s_m$ respectively, $\tilde{F}_I(t) = (r/n)F_r(t) + (m/n)F_m^*(t + a_n)$.

(i) $m/n \to 0$. The consistency of $\tilde{F}_I(t)$ is a consequence of the consistency of $F_r(t)$.

(ii) $m/n \to \alpha$, where $0 < \alpha < 1$. It is simple to verify that $|F_m^*(t) - F_r(t)|$, $|F_r(t) - F_N(t)|$ and $|F_N(t) - F(t)|$ all converge to 0, which in turn implies that $|F_m^*(t) - F(t)| \to 0$. Since $F(t)$ is a continuous cdf, we further conclude that $\sup_t |F_m^*(t) - F(t)| \to 0$, and as a result,

$$|F_m^*(t + a_n) - F(t + a_n)| \to 0. \tag{A.1}$$

By the finiteness of the limit of S_y^2, $a_n = o_p(1)$. Consequently, by the continuity of $F(t)$, $F(t + a_n) - F(t) = o_p(1)$, which together with (A.1) implies the result.

(b) Asymptotic normality of $\tilde{F}_I(t)$

Let $H_m(a) = |[F_m^*(t+a) - F_m^*(t)] - [F_r(t+a) - F_r(t)]|$, and $H_r(a) = |[F_r(t + a) - F_r(t)] - [F_N(t + a) - F_N(t)]|$.

Lemma 1. *Under Condition* 1, (i) $H_m(a_n) = o_p(m^{-1/2})$ *and* (ii) $H_r(a_n) = o_p(m^{-1/2})$, *where* a_n *is the adjustment factor.*

The proofs of both (i) and (ii) are similar to the proof of Lemma 1 in Chen and Shao (1999). Thus, they are omitted.

Theorem 2. *Under Condition* 1, $\tilde{F}_I(t)$ *is* $AN(F_N(t), \tilde{\sigma}_n^2)$ *with*

$$\tilde{\sigma}_n^2 = (1 - \frac{r}{N})\{(r^{-1} + n^{-2}m)F_N(t)[1 - F_N(t)] + n^{-2}m[f^2(t)\sigma_N^2 + 2C_N(t)\ f(t)]\}, \tag{A.2}$$

where $C_N(t) = N^{-1} \sum_{i=1}^N (y_i - \bar{Y}_N) I(y_i \le t)$.

Proof. First, note that $F_m^*(t + a_n) - F_m^*(t) = F_r(t + a_n) - F_r(t) + o_p(m^{-1/2}) = F_N(t+a_n) - F_N(t) + o_p(m^{-1/2}) = f(t)a_n + o_p(m^{-1/2})$ by Lemma 1 and Condition 1. It then follows that $\tilde{F}_I(t) = \frac{r}{n}F_r(t) + \frac{m}{n}F_m^*(t) + \frac{m}{n}[F_m^*(t+a_n) - F_m^*(t)] = \frac{r}{n}F_r(t) + \frac{m}{n}F_m^*(t) + \frac{m}{n}f(t)[a_n + o_p(m^{-1/2})]$. The remainder of the proof is straightforward.

(c) Situations where $\tilde{F}_I(t)$ will outperform $\hat{F}_I(t)$

Note that the first term in (A.2) is the variance of $\hat{F}_I(t)$. Therefore,the asymptotic variance of $\tilde{F}_I(t)$ will be smaller than that of $\hat{F}_I(t)$ when $f(t)\sigma_N^2 + 2C_N(t) < 0$. To shed some light on when this might be true, let us consider two simple situations:

(i) If the finite population itself was an iid sample from a standard normal distribution (super-population), then $f(t) + 2cov(Y, I(Y \leq t)) \doteq -f(t)$, and thus $\tilde{F}_I(t)$ would have smaller asymptotic variance than $\hat{F}_I(t)$. This generalizes to any normal population.

(ii) If the finite population were generated as an iid sample from a gamma distribution instead with density $f(t) \propto t^{d-1}e^{-t}$, then $f(t)\sigma_N^2 + 2cov(Y, I(Y \leq t)) = (d - 2t)f(t)$ and there would be a gain in precision when $t > d/2$.

Appendix 2. Asymptotic Properties of $\tilde{F}_I(t)$ for Stratified Multistage Sampling

Let us first develop some necessary notation. Let $\tilde{S} = N^{-1}\hat{S}$, $\tilde{U} = N^{-1}\hat{U}$, $\tilde{T} = N^{-1}\hat{T}$, $\tilde{V} = N^{-1}\sum_{(hik)\in s_m} w_{hik} I(y_{hik} \leq t)$, $a_{hik} = 1$ when $(hik) \in s_r$ and 0 otherwise, and $n = \sum_h n_h$.

Let us also denote $F_r(t) = \hat{S}^{-1}\sum_{s_r} w_{hik} I(y_{hik} \leq t)$ and $F_m(t) = \hat{T}_m^{-1}\sum_{s_m} w_{hik} I(y^*_{hik} \leq t)$, where $\hat{T}_m^{-1} = \sum_{s_m} w_{hik}$. The adjustment factor then becomes $a_n = \hat{T}_m^{-1}\sum_{s_m} w_{hik} y^*_{hik} - \hat{S}^{-1}\sum_{s_r} w_{hik} y_{hik}$ and the adjusted imputation estimate of the distribution function can be rewritten as $\tilde{F}_I(t) = W_r F_r(t) + W_m F_m(t + a_n)$, where $W_r = \hat{U}^{-1}\hat{S}_r$ and $W_m = \hat{U}^{-1}\hat{T}_m$.

We assume the first stage units are sampled with replacement and we need the following additional conditions similar to those introduced in Rao and Shao (1992).

Condition 2. $n^{1+\delta}\sum\sum E|\tilde{r}^{(l)}_{hi} - E\tilde{r}^{(l)}_{hi}|^{2+\delta} = O(1)$ for $l = 1, 2, 3, 4$ as $n \to \infty$, where $\tilde{r}^{(l)}_{hi} = N^{-1}\sum_k w_{hik}\tilde{y}^{(l)}_{hik}$, $y^{(1)}_{hik} = a_{hik}y_{hik}$, $y^{(2)}_{hik} = a_{hik}$, $y^{(3)}_{hik} = a_{hik}I(y_{hik} \leq t)$ and $y^{(4)}_{hik} = 1$.

Condition 3. $n\times$ (covariance matrix of $\tilde{S}$, $\tilde{U}$, $\tilde{T}$, $\tilde{V}$) converges to a positive definite matrix as $n \to \infty$.

Condition 4. $\max_{h,i}\sum_k \tilde{w}_{hik} = O(n^{-1})$, where $\tilde{w}_{hik} = w_{hik}/N$.

Condition 5. $\sum\sum\sum \tilde{w}_{hik}|y_{hik} - \bar{Y}|^{2+\delta} = O_p(1)$.

Theorem 3. *Under Conditions* 1-5, *we have* $\tilde{F}_I(t) = W_r F_r(t) + W_m F_m(t) + W_m f(t)a_n + o_p(n^{-1/2})$. *Therefore,* $\tilde{F}_I(t)$ *is consistent and* $AN(F_N(t), \tilde{\sigma}^2)$ *with* $\tilde{\sigma}^2 = Var(W_r F_r(t) + W_m F_m(t) + W_m f(t)a_n)$.

Proof. The first part is a direct consequence of Lemma 2, below. As all components in the expansion of $\tilde{F}_I(t)$ are sums of independent random variables, the asymptotic normality is then straightforward by using Slutsky's theorem and Rao and Shao (1992).

Lemma 2. *Under Conditions* 1-5, $F_m(t+a_n) - F_m(t) = f(t)a_n + o_p(m^{-1/2})$.

Proof. Note that $W_m F_m(t) - W_m F_m(t') = \hat{T}_m^{-1} \sum_{h,i} \sum_j w_{hij} I(t' < y^*_{hij} \le t)$ and $\sum_j w_{hij} I(t' < y^*_{hij} \le t)$ are independent random variables for different h or i with upper bound $\Delta = \max_{h,i} \sum_j w_{hij}$. Also, $\sum_{h,i} Var\{w_{hij} I(t' < y^*_{hij} \le t)\} \le \sum_{h,i} E[w^2_{hij} I(t' < y^*_{hij} \le t)] \le N[F_N(t) - F_N(t')]\Delta$. Using Bernstein's inequality, we have $P(\hat{T}_m | W_m F_m(t) - W_m F_m(t') - E[W_m F_m(t) - W_m F_m(t')]| \ge nz) \le 2\exp(-\frac{n^2 z^2}{2N\Delta[F_N(t)-F_N(t')]+2nz\Delta/3})$. Recall that $\Delta = O(Nn^{-1})$ by Condition 4. Thus, by choosing $z' = (nz)/N$ and ignoring constant factors, the right hand side becomes

$$\exp\left(-\frac{nz'^2}{[F_N(t) - F_N(t')] + z'}\right),$$

and when $|t' - t| \le n^{-1/2}$, choosing $z' = n^{-3/4} \log n$ implies an upper bound of order n^{-1}. By using the same technique as in Serfling (1980, p.97), we conclude that for any $C > 0$,

$$\sup_{t':|t'-t|\le Cn^{-1/2}} \hat{T}_m |W_m F_m(t) - W_m F_m(t') - E[W_m F_m(t) - W_m F_m(t')]|$$
$$= O_p(Nn^{-3/4} \log n).$$

Combined with the fact that $\hat{T}_m = O(N)$, $a_n = O_p(n^{-1/2})$, $E[F_m(t)] = F_N(t)$ and W_m converges to a constant, this implies $F_m(t) - F_m(t+a_n) = F_N(t) - F_N(t+a_n) + o_p(m^{-1/2}) = f(t)a_n + o_p(m^{-1/2})$. The last equality is obtained by using Condition 1 and the differentiability of $F(t)$ which is the limit of $F_N(t)$. This completes the proof.

References

Brick, J. M. and Kalton, G. (1996). Handling missing data in survey research. *Statist. Methods in Medical Research* **5**, 215-238.

Chen, Y. and Shao, J. (1999). Inference with survey data imputed by hot deck when imputed values are nonidentifiable. *Statist. Sinica* **9**, 361-384.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* SIAM, Philadelphia, USA.

Fay, R. E. (1996). Alternative paradigms for the analysis of imputed survey data. *J. Amer. Statist. Assoc.* **91**, 490-498.

Kalton, G. and Kish, L. (1984). Some efficient random imputation methods. *Comm. Statist., Part A – Theory and Methods* **13**, 1919-1939.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York.

Saigo, H., Shao, J. and Sitter, R. R. (1999). A repeated half-sample bootstrap and balanced repeated replications for randomly imputed data. Unpublished Manuscript.

Sarndal, C. E. (1992). Methods for estimating the precision of survey estimators when imputation has been used. *Survey Methodology* **18**, 241-252.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed data. *J. Amer. Statist. Assoc.* **91**, 1278-1288.

Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, N2L 3G1 Canada.

School of Mathematics and Statistics, Carleton University, Ottawa, ON, K1S 5B6 Canada.

Department of Mathematics and Statistics, Simon Fraser University, Burnaby, BC, V5A 1S6 Canada.

(Received January 1998; accepted March 2000)