Survey Methodology, June 1994
Vol. 20, No. 1, pp. 45–52
Statistics Canada

Jackknife Variance Estimation of Imputed Survey Data

JOHN G. KOVAR and EDWARD J. CHEN[1]

ABSTRACT

Imputation is a common technique employed by survey-taking organizations in order to address the problem of item nonresponse. While in most of the cases the resulting completed data sets provide good estimates of means and totals, the corresponding variances are often grossly underestimated. A number of methods to remedy this problem exists, but most of them depend on the sampling design and the imputation method. Recently, Rao (1992), and Rao and Shao (1992) have proposed a unified jackknife approach to variance estimation of imputed data sets. The present paper explores this technique empirically, using a real population of businesses, under a simple random sampling design and a uniform nonresponse mechanism. Extensions to stratified multistage sample designs are considered, and the performance of the proposed variance estimator under non-uniform response mechanisms is briefly investigated.

KEY WORDS: Item nonresponse; Hot deck imputation; Nearest neighbour imputation; Nonrandom nonresponse; Complex survey design.

1. INTRODUCTION

All sample surveys suffer from varying degrees of nonresponse. While total or unit nonresponse is often redressed by appropriate survey weight adjustment, most survey taking organizations resort to imputation in the case of item nonresponse. In this way, plausible values are inserted in place of missing or inconsistent entries, thus simplifying estimation of means and totals at all levels of aggregation. As early as the 1950's however, Hansen, Hurwitz and Madow (1953) recognized that treating the imputed values as observed values can lead to underestimation of variances of these estimators if standard formulae are used; underestimation which becomes more appreciable as the proportion of imputed items increases.

A number of remedies to overcome this problem have been advanced. In particular, Rubin (1987) proposed multiple imputation to estimate the variance due to imputation by replicating the process a number of times and estimating the between replicate variation. More recently, Särndal (1990) outlined a number of model assisted estimators of variance, while Rao and Shao (1992) proposed a technique that adjusts the imputed values to correct the usual or naive jackknife variance estimator. The Särndal, and Rao and Shao methods, are appealing in that only the imputed file (with the imputed fields flagged) is required for variance estimation. No auxiliary files are needed. Särndal's model assisted approach yields unbiased variance estimators, provided the model holds (Lee, Rancourt and Särndal 1991). The Rao and Shao adjusted jackknife method is design consistent as well as model unbiased (Rao 1992). But while the model assisted approach requires different variance estimators for each imputation method, the adjusted jackknife method provides a unified approach that requires the implementation of only one estimator, the jackknife estimator, provided the imputed values are adjusted appropriately during the variance estimation stage.

In this paper we describe a simulation study that evaluates the adjusted jackknife variance estimator of Rao and Shao (1992). In Section 2 we motivate the present empirical study by demonstrating the characteristics of the naive variance estimator under four imputation methods in the case of simple random sampling. In Section 3 we briefly outline the Rao and Shao adjustment procedure and present the empirical results. Extensions to more complex designs and experiments with nonrandom nonresponse mechanisms are elaborated in Section 4. Finally, in Section 5 we offer some concluding remarks and recommendations, including areas for future study.

2. BACKGROUND

Following the notation of Rao (1992), we suppose that in a sample s, of size n, m units respond to item y, while $n - m$ units do not. Denote by y_i^* the imputed value for unit i, $i \in s\text{-}s_r$, where s_r is the set of responding units. The usual estimator of the mean $\bar{Y}$ under simple random sampling, based on the imputed file is given by

$$\bar{y}_I = \frac{1}{n}\left(\sum_{i \in s_r} y_i + \sum_{i \in s\text{-}s_r} y_i^* \right). \qquad (1)$$

[1] John G. Kovar, Business Survey Methods Division; Edward J. Chen, Social Survey Methods Division, Statistics Canada, R.H. Coats Building, Tunney's Pasture, Ottawa, Ontario, Canada K1A 0T6.

2.1 Imputation Methods

In the present simulation study we consider four simple methods of imputation, namely the mean of respondents, ratio, nearest neighbour and hot deck imputation methods. The reader is referred to Kalton and Kasprzyk (1986) for a thorough review of the topic of imputation. The simplest and most intuitive method of imputation, when the interest lies in estimating the mean of the item y, is to impute all missing items with the mean of the observed responding units. The imputed value y_i^*, for unit i, under the mean imputation method, is thus given by

$$y_i^* = \bar{y}_m = \sum_{j \in s_r} y_j / m. \tag{2}$$

In this case, the estimator of the population mean $\bar{Y}$ in (1) reduces to the estimator $\bar{y}_I = \bar{y}_m$. Due to the fact that this method has the undesirable property of distorting the distributions, it is used in practice usually only as a last resort. It is included here for illustrative purposes.

Secondly, we consider a ratio imputation method based on the assumption that a correlated auxiliary variable x, is available, and that the ratio $\bar{y}/\bar{x}$ is the same in the s_r and $s - s_r$ sets, as would be the case if the nonresponse occurred at random, for example. Under the ratio imputation method, we impute the predicted value in place of the missing y_i as follows:

$$y_i^* = \frac{\bar{y}_m}{\bar{x}_m} x_i, \tag{3}$$

where $\bar{x}_m$ is the mean of the x values of the respondent set s_r. The estimator of the population mean $\bar{Y}$ in (1) reduces to the double sampling estimator $\bar{y}_I = (\bar{y}_m/\bar{x}_m)\bar{x}$, by considering the respondents as the second phase sample.

The third imputation technique we consider is the nearest neighbour (NN) method. Under this method, the missing value is filled in by an observed value of another unit from the set s_r, whose distance to the nonresponding unit is minimum. In practice the distance functions used are usually the ζ_1, ζ_2, or ζ_∞ Minkowski's norms based on the auxiliary x-variables, assumed observed for all units in s. Thus

$$y_i^* = y_j, j \in s_r, \text{ such that } \| x_i - x_j \| \text{ is minimized,} \tag{4}$$

where $\| \cdot \|$ is one of the above mentioned norms.

The above three methods are often labelled deterministic, since, given the sample of respondents, the imputed values are determined uniquely. The fourth imputation method considered in this study, the hot deck method (HD), is non-deterministic, since the imputed values are chosen at random from the respondent set. While in practice imputation classes are often created and some sort of sequential procedure is usually implemented, we consider here the pure hot deck, whereby the donor unit (j) is chosen at random, with replacement, from the entire set s_r, that is,

$$y_i^* = y_j, \; j \in s_r. \tag{5}$$

2.2 Variance Due to Imputation

Treating the imputed values as observed values, leads to the incorrect variance estimator

$$v_{naive} = (1 - f)s_I^2/n, \tag{6}$$

where s_I^2 is the sample variance of the complete sample of responding and imputed values, and $(1 - f)$ is the finite population correction factor ($f = n/N$). It can be easily shown that the true variance of the estimator $\bar{y}_I$ in (1), $V(\bar{y}_I)$, can be written as (Särndal 1990)

$$V(\bar{y}_I) = V_{sam} + V_{imp} + V_{mix}, \tag{7}$$

where V_{sam} is the sampling variance component, V_{imp} is the variance introduced by the imputation method in question and V_{mix} is a covariance term between V_{sam} and V_{imp} which in most cases is negligible or zero. An estimator of V_{sam} could be obtained by adding to v_{naive} a term to correct for the fact that the standard formula understates the sampling variance component when there are imputed values in the data set. To estimate $V(\bar{y}_I)$, however, an additional component of variance due to the imputation mechanism, V_{imp}, must be estimated. This may be done explicitly, as in Rubin's (1987) multiple imputation, or by modifying common variance formulae as in Särndal (1990) and Rao and Shao (1992). Note that the interest lies in estimating the variance of the estimator at hand, that is, $V(\bar{y}_I)$, not the variance of an estimator that would have been obtained had there been no nonresponse.

2.3 Variance Underestimation

To illustrate the seriousness of the underestimation of $V(\bar{y}_I)$ by v_{naive}, and the dependence of the degree of underestimation on the imputation method, we first describe the simulation study used for this purpose. We consider a data set of 5,620 units with two variables: An auxiliary variable x, the Gross Business Income, available for all units, that can be used as a measure of size, and a related purchase variable y. The correlation between x and y in this particular data set is of the order of 0.92. Simple random samples of size 200 were selected without replacement. A fixed proportion of units were identified at random as nonrespondents, having their y-values deleted and imputed according to one of the four methods described above. Various rates of nonresponse were generated, though, for the most part, we confine our reporting to results based on 5 and 30% nonresponse rates.

To evaluate the performance of the proposed variance estimators, we calculate the percent relative bias of the variance estimator $v.$, given by

$$\text{Rel.Bias}(v.) = \sum_{k=1}^{K} \frac{(v_k - V(\bar{y}_I))/K}{V(\bar{y}_I)} \times 100, \quad (8)$$

where $V(\bar{y}_I)$ is obtained through simulation, and v_k is the k-th realization of the K simulated variance estimates in question. Similarly, the percent relative stability of the variance estimators is given by

$$\text{Rel.Stab.}(v.) = \sum_{k=1}^{K} \frac{\sqrt{(v_k - V(\bar{y}_I))^2/K}}{V(\bar{y}_I)} \times 100. \quad (9)$$

All simulations were performed on an IBM PC, using Microsoft's Fortran 77, Version 5.0. In the case of simple random sampling, results are based on averages of 100,000 replications (K = 100,000). With this number of replicates, the reported relative bias values were observed not to vary by more than one percentage point. The results are summarized below in Table 1 for the case of 5 and 30% nonresponse rates.

Table 1
Underestimation of Variance of $\bar{y}_I$ by the Naive Estimator Under Four Imputation Methods, and 5 and 30% Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
5%	$V(\bar{y}_I)$	9.9	10.3	9.5	9.5
	v_{naive}	8.9	9.4	9.2	9.3
	Rel.Bias(v_{naive})	−10.7%	−9.4%	−2.5%	−2.2%
30%	$V(\bar{y}_I)$	13.5	16.5	10.1	10.3
	v_{naive}	6.5	9.4	8.5	9.0
	Rel.Bias (v_{naive})	−51.4%	−43.4%	−15.3%	−12.8%

First, we note in Table 1, that the naive estimator underestimates the true variance of $\bar{y}_I$ by 10.7% in the case of mean imputation at a 5% level of nonresponse. About half of this underestimation is due to the fact that v_{naive} underestimates V_{sam} and the other half is due to the fact that v_{naive} ignores the component V_{imp}. Särndal (1990) obtains very similar results with respect to the partitioning of the underestimation in the case of mean imputation. Secondly, in the first row of Table 1, the true variance of $\bar{y}_I$ is larger in the case of the hot deck imputation as compared to the mean imputation, due to the procedure's inherent variability (*i.e.*, the V_{imp} component is larger). By contrast, $V(\bar{y}_I)$ is slightly lower in the case of the ratio and nearest neighbour imputation methods, since V_{imp} decreases as the imputation procedure is better able to predict the true unobserved values (Särndal 1990), as is the case in the present study due to the relatively high correlation between the x and y variables. Thirdly, as can be seen in Table 1, $V(\bar{y}_I)$ increases while v_{naive} decreases as the nonresponse rate becomes more elevated. As such, the underestimation of $V(\bar{y}_I)$, when the imputed values are treated as observed values, becomes more serious as the proportion of missing items increases. The problem is more pronounced in the case of the mean and hot deck imputation methods, which do not use auxiliary information. Note that underestimation of variance in the order of 50%, as was observed in this case, can lead to confidence intervals that are about 30% too short and to declaration of significance when none exists. Also of note is the similar behaviour of the ratio and nearest neighbour methods which will be exploited later.

3. JACKKNIFE VARIANCE ESTIMATOR

Let $\bar{y}_I(j)$ be the imputed estimator of $\bar{Y}$ obtained when the j-th unit is deleted from the sample. Then, in the case of simple random sampling, a naive jackknife variance estimator of $\bar{y}_I$ is given by

$$\bar{v}_J = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I(j) - \bar{y}_I]^2, \quad (10)$$

which can be shown to reduce to v_{naive} (Rao 1992).

3.1 Imputed Value Adjustment

In order to produce the "correct" (Rao 1990) jackknife variance estimator, Rao (1992) proposed to adjust the imputed values as described below. Intuitively, the adjustment is necessary whenever a responding unit is deleted from a jackknife replicate, since in the case of most imputation methods, all the imputed values depend directly or indirectly on the observed value that was deleted. This is clear in the case of mean imputation and ratio imputation, where all respondents contribute directly to the mean $\bar{y}_m$, but is less evident in nearest neighbour and hot deck imputation methods where the deleted unit contributes to the imputation process only in the sense that it is not available to be selected as a donor. Thus, whenever a responding unit is deleted, *all* imputed values in the sample must be adjusted before the "delete-one" imputed estimator of the mean is computed. The adjustment must clearly be a function of the imputation method used. In the case of the mean and the hot deck imputation methods, it can be shown that the following adjustment is appropriate (Rao 1992; Rao and Shao 1992). Let $z_i^*(j)$ be the adjusted value of the i-th imputed unit y_i^*, when the j-th unit has been deleted. Then $z_i^*(j)$ is given by

$$z_i^*(j) = \begin{cases} y_i^* + [\bar{y}_m(j) - \bar{y}_m] & \text{if } j \in s_r \\ y_i^* & \text{if } j \in s\text{-}s_r. \end{cases} \quad (11)$$

In other words, no adjustment is necessary if the deleted unit (j), has itself been imputed; that is, unit j is a non-respondent. In the case of the mean imputation, for example, when $j \in s_r$, the adjusted value reduces to $\bar{y}_m(j)$, the mean of the remaining $m - 1$ respondents, as desired.

The jackknife variance estimator is evaluated by first computing the adjusted imputed estimator $\bar{y}_I^a(j)$, as

$$\bar{y}_I^a(j) = \sum_{\substack{i \in s \\ i \neq j}} z_i^*(j)/(n-1), \quad (12)$$

and then letting

$$v_J(\bar{y}_I) = \frac{n-1}{n} \sum_{j=1}^{n} [\bar{y}_I^a(j) - \bar{y}_I]^2. \quad (13)$$

It can be shown that the adjusted jackknife variance estimator reduces to the correct variance estimator in the case of the mean imputation (Rao 1990), and provides a consistent estimator in the case of the hot deck imputation (Rao and Shao 1992).

In the case of the ratio imputation, the adjusted values are given by

$$z_i^*(j) = \begin{cases} y_i^* + \left[\dfrac{\bar{y}_m(j)}{\bar{x}_m(j)} x_i - \dfrac{\bar{y}_m}{\bar{x}_m} x_i\right] & \text{if } j \in s_r \\ y_i^* & \text{if } j \in s\text{-}s_r, \end{cases} \quad (14)$$

where $\bar{x}_m(j)$ is the mean of the $m - 1$ sample values of x of the responding units when unit j is deleted. The jackknife variance estimator $v_J(\bar{y}_I)$ is then computed as in (13) above, yielding the correct variance estimator. Furthermore, Rao (1992) shows that not only is the adjusted jackknife variance estimator design consistent (p-consistent) under uniform nonresponse irrespective of the model, but is also design-model unbiased (pm-unbiased) under the model (15) and any nonresponse mechanism that does not depend on the y-values.

$$E_m(y_i) = \beta x_i, \quad V_m(y_i) = \sigma^2 x_i, \quad \text{cov}_m(y_i, y_j) = 0 \quad i \neq j \in s. \quad (15)$$

Since the naive variance estimator under the nearest neighbour imputation was observed to behave much like the naive variance estimator under the ratio imputation, the adjustment for the ratio imputation given in (14) was used in the case of the nearest neighbour imputation. As well, an alternate adjustment was considered, whereby unit i was re-imputed using the nearest neighbour method, whenever the deleted unit (j) was used to impute unit i. That is, adjustment takes place only if the deleted unit is a respondent (as above), but only those nonrespondents in the j-th jackknife replicate that were actually imputed using unit j are re-imputed by one of the $m - 1$ remaining donors. (This corresponds to imputing the second nearest neighbour for these units.) We note that no theoretical justification exists for either of these adjustments. Since the latter adjustment performed worse than the ratio adjustment in our examples, and since its eventual implementation in production would be cumbersome, we omitted it from further consideration, even though it was always observed to be conservative.

We would like to stress here that for all imputation methods the adjustments are only performed for the purpose of variance estimation and can be made temporarily while the variance estimation program executes. No permanent adjustments are required on the imputed file used for the estimation of means and totals, though the imputed fields must be flagged appropriately.

3.2 Empirical Results

The jackknife variance estimator with adjustments corresponding to the four imputation methods described above, was computed in addition to v_{naive} in the simulation study outlined in Section 2. Nonresponse rates of 5 and 30% were considered and the relative biases were calculated. They are summarized in Table 2 below.

Table 2

Relative Biases of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse Rates

Non-response Rate	Variance Estimator	Imputation Method			
		Mean	HD	Ratio	NN
		in percent			
5%	v_{naive}	−10.7	−9.4	−2.5	−2.2
	v_J	2.7	3.6	3.4	3.7
30%	v_{naive}	−51.4	−43.4	−15.3	−12.8
	v_J	3.3	1.9	3.0	5.3

Since the adjusted jackknife variance estimator is design consistent (p-consistent) (Rao 1992), it performs well in the case of the mean, hot deck and ratio imputation under uniform response mechanism, as expected. (Equally good performance was observed with other data sets which do not follow the model (15) as well, but more work is needed on this front.) Of note is the relatively good performance under the nearest neighbour imputation. The proposed estimator tends to be somewhat conservative, due, in small part, to the fact that it does not incorporate the finite population correction.

4. EXTENSIONS

While the adjusted jackknife variance estimator has been shown to perform well in the case of simple random sampling under uniform nonresponse mechanism in one imputation class, we consider here extensions to more complex design, to more than one imputation class, and to nonrandom response mechanisms.

4.1 Complex Designs

In this section we describe a simulation study that evaluates the Rao and Shao (1992) adjusted jackknife variance estimator in comparison to the naive variance estimator, in the case of stratified multistage sampling and hot deck imputation. In particular, data from the Canadian Survey of Consumer Finances (SCF) that follows the design of the Canadian Labour Force Survey will be used. The variable of interest, y, is the total household income. The SCF follows a complex stratified multistage design with the primary sampling units (psu's) in the strata used in this study selected with probability proportional to the number of dwellings. Generally speaking, the psu's are collections of dwellings, corresponding to city blocks in urban areas and to groups of Census Enumeration Areas (EA's) in rural regions. We used as a population a sample of 3,870 households in 30 strata and sampled two psu's in each stratum. As in the case of the simple random sampling study, 5 and 30% uniform nonresponse rates were generated at the household level. The missing values were then imputed using the hot deck imputation method described in Rao and Shao (1992). Briefly, the imputation method consists of selecting the donors from the respondent set with replacement, with probability proportional to the survey weight of the donors.

We first consider the case of a single imputation class. Let y_{hik} be the observed value for the k-th unit in the i-th psu and the h-th stratum ($k = 1, \ldots, n_{hi}, i = 1, \ldots, n_h, h = 1, \ldots, L, n = \sum\sum n_{hi}$), and let y^*_{hik} be the corresponding imputed value whenever the (hik) unit is a nonrespondent, that is, whenever $(hik) \in s\text{-}s_r$. The imputed estimator of Y is then given by

$$\hat{Y}_I = \sum_{(hik)\in s_r} w_{hik} y_{hik} + \sum_{(hik)\in s\text{-}s_r} w_{hik} y^*_{hik}, \qquad (16)$$

where w_{hik} is the survey weight corresponding to unit (hik). Under the above hot deck imputation scheme, $\hat{Y}_I$ is asymptotically unbiased (Rao and Shao 1992).

The expectation of $\hat{Y}_I$ under the hot deck imputation procedure can be written as (Rao and Shao 1992):

$$E_*(\hat{Y}_I) = \left[\sum_{(hik)\in s_r} w_{hik} y_{hik} \Big/ \sum_{(hik)\in s_r} w_{hik}\right] \times \sum_{(hik)\in s} w_{hik}$$

$$= [\hat{S}/\hat{T}] \times \hat{U}, \qquad (17)$$

thus defining the terms $\hat{S}$, $\hat{T}$ and $\hat{U}$. The jackknife "delete-one" values are then given by

$$\hat{S}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} y_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik} y_{gik},$$

$$\hat{T}(gj) = \sum_{\substack{(hik)\in s_r \\ h\neq g}} w_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(gik)\in s_r \\ i\neq j}} w_{gik}, \qquad (18)$$

whenever the j-th psu in the g-th stratum is deleted. The adjustment of the imputed values is performed whenever the (gj)-th psu is deleted, $(hi) \neq (gj)$, and $(hik) \in s\text{-}s_r$, by letting

$$z^{(gj)}_{hik} = y^*_{hik} + \left[\frac{\hat{S}(gj)}{\hat{T}(gj)} - \frac{\hat{S}}{\hat{T}}\right]. \qquad (19)$$

Then, analogous to (12) and (13), the jackknife variance estimator is evaluated by first computing the adjusted imputed estimator $\hat{Y}^a_I$ when the (gj)-th psu is deleted as

$$\hat{Y}^a_I(gj) = \hat{S}(gj) + \sum_{(hik)\in s\text{-}s_r} w_{hik} z^{(gj)}_{hik} + \frac{n_g}{n_g - 1} \sum_{\substack{(hik)\in s\text{-}s_r \\ i\neq j}} w_{gik} z^{(gj)}_{gik}, \qquad (20)$$

and then setting

$$v_J(\hat{Y}_I) = \sum_{g=1}^{L} \frac{n_g - 1}{n_g} \sum_{j=1}^{n_g} (\hat{Y}^a_I(gj) - \hat{Y}_I)^2. \qquad (21)$$

It can be shown that v_J as defined in (21), is a consistent estimator of the variance of $\hat{Y}_I$ (Rao and Shao 1992).

We generated 10,000 samples of 60 psu's selected with probability proportional to size, and subjected the selected households to 5 and 30% uniform nonresponse. We then computed the naive variance estimator, and the adjusted jackknife variance estimator, v_J, in (21). The relative bias (8) and the relative stability (9) were computed for both of the variance estimators, and are summarized in Table 3 below.

Table 3

Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under 5 and 30% Nonresponse, in the Case of Stratified Multistage Sampling

Variance Estimator	Nonresponse Rate	
	5%	30%
	in percent	
v_{naive}	−10.3 (88)	−43.7 (84)
v_J	−0.9 (97)	1.2 (124)

As can be seen in Table 3, the naive variance estimator underestimates the true variance of Y at rates comparable to the simple random sampling case (Table 2), with the underestimation becoming more serious as the nonresponse rate increases. The adjusted jackknife variance estimator, on the other hand, performs well at both levels of nonresponse, at a relatively modest cost of a slight decrease in the stability of the variance estimator, as compared to v_{naive}.

4.2 Imputation Classes

Under the same sample design as in Section 4.1, we also considered the case of more than one imputation class as is the case in practice. The household size, known for all households in the sample, was used to form two imputation classes, namely one member households and more than one member households. This was done under the assumption that the propensity to respond is different between these two classes, while uniform response probability was assumed within the imputation classes. Two nonresponse schemes were evaluated. The first assumes a 5% uniform nonresponse in the single member household class and 10% uniform nonresponse in the multiple member household class, while the second scheme assumes rates of 25 and 30% in each of the classes respectively. The hot deck imputation, the imputed value adjustments, and the adjusted total calculations in (20), $\hat{Y}^q_{Iv}(gj)$, were performed independently within each imputation class denoted by v. The terms $\hat{Y}^q_{Iv}(gj)$ were then summed over the two imputation classes, yielding $\hat{Y}^q_I(gj)$, which was used in (21) to provide the estimate v_J. The results are summarized in Table 4.

Table 4

Relative Bias and Relative Stability (in Parentheses) of the Naive Variance Estimator and the Adjusted Jackknife Variance Estimator Under Two Nonresponse Schemes, in the Case of Stratified Multistage Sampling and Two Imputation Classes

Variance Estimator	Nonresponse Rate	
	5% and 10%	25% and 30%
	in percent	
v_{naive}	−16.7 (87)	−40.2 (84)
v_J	−1.0 (103)	1.1 (127)

As can be seen in Table 4, the adjusted jackknife variance estimator v_J, performs well under both nonresponse schemes. The results, along with those in Table 3, demonstrate the consistency and the reasonably good stability of the adjusted jackknife variance estimator, even in cases of elevated nonresponse rates.

4.3 Nonrandom Nonresponse

As demonstrated above, the adjusted jackknife variance estimator performs well when the nonresponse is random within imputation classes. To study its robustness against the uniform response mechanism assumption, we use the data set described in Section 2, and generated nonresponse as outlined in Lee, Rancourt and Särndal (1991). In particular, the probability of nonresponse is assumed to be related to the x-variable in two distinct ways:

$$P_L = 1 - \exp(-c_L x), \tag{22}$$

$$P_S = \exp(-c_S x), \tag{23}$$

where the constants c_L and c_S are chosen such that an expected 30% nonresponse rate is achieved. In the model P_L given in (22) the nonresponse is positively correlated with the x-variable, implying that large (L) units are more likely not to respond. The opposite is true in the model P_S given in (23), under which smaller (S) units are more likely not to respond. Imputation methods which ignore the x-variable (mean and hot deck) are expected to yield estimators of $\bar{Y}$ that underestimate the true mean under nonresponse model (22) and over estimate the true mean under the model (23). However, imputation methods that incorporate the auxiliary variable into the procedure (ratio and nearest neighbour), can be expected to produce better estimates of the mean. This has been confirmed by simulation as shown in Table 5 below. As before, 100,000 replicates were used.

Table 5

Estimates of the Mean $\hat{Y}$ as Percent of the True Mean when the Nonresponse is not Random, and the Nonresponse Rate is an Expected 30%

Nonresponse Model	Imputation Method			
	Mean	HD	Ratio	NN
	in percent			
P_L	60.4	60.4	94.7	93.5
P_S	132.7	132.7	102.0	101.4

Clearly, variance estimation is of no interest when the point estimators themselves are highly biased as is the case for the mean and hot deck methods. However, in the case of the ratio and nearest neighbour methods, under which the point estimators perform better, we investigated the performance of the adjusted jackknife variance estimator, as well as an estimator proposed by Särndal (1990), which can be written as (Rao 1992):

$$v_s(\bar{y}_I) = \left(\frac{\bar{x}}{\bar{x}_m}\right)^2 \frac{1}{m(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m} x_i\right)^2$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right) \frac{2m}{n^2(m-1)} \sum_{i \in s_r} \left(y_i - \frac{\bar{y}_m}{\bar{x}_m}\right) x_i \qquad (24)$$

$$+ \left(\frac{\bar{y}_m}{\bar{x}_m}\right)^2 \frac{1}{n(n-1)} \sum_{i \in s} (x_i - \bar{x})^2,$$

provided that the finite population correction factor is ignored, and that $(n-1)/n \cong 1$ and $(m-1)/m \cong 1$. The results are summarized in Table 6.

Table 6

Relative Bias of the Naive Variance Estimator, the Adjusted Jackknife Variance Estimator and Särndal's Variance Estimator Under 30% Nonrandom Nonresponse

Nonresponse Model	Variance Estimator	Imputation Method	
		Ratio	NN
		in percent	
P_L	v_{naive}	−22.7	−54.6
	v_J	3.9	−37.5
	v_S	−2.6	−36.8
P_S	v_{naive}	−4.0	−0.7
	v_J	3.7	7.2
	v_S	2.8	4.5

In the case of the ratio imputation, the naive variance estimator performs quite differently under the two nonresponse models (−22.7 versus −4.0%). This is due to the fact that while the reduction in effective sample size tends to decrease the variance in both cases, under the P_L model disproportionately more large units are missing which tends to accentuate this effect, whereas under the P_S model, where disproportionately more small units are missing, this effect tends to be partly compensated for. Secondly, the adjusted jackknife variance estimator performs well in the case of ratio imputation, but relatively poorly in the case of nearest neighbour imputation. This is due to the fact that the present data set follows the usual linear model (15) fairly well and the adjusted jackknife variance estimator has been shown to be model unbiased (Rao 1992) in the case of the ratio imputation. On the other hand, the ratio adjustment does not work well in the case of nearest neighbour imputation when the nonresponse is not uniform. The alternate adjustment for the nearest neighbour imputation described in Section 3, performs equally poorly in absolute terms (not shown here), though the estimates are always conservative. Thirdly, the performance of Särndal's estimator, v_S, is roughly equivalent to that of the adjusted jackknife estimator under either the ratio or the nearest neighbour imputation methods, and nonrandom nonresponse that depends only on x.

In cases where the response mechanism is not random, and when the propensity to respond is related to the variable subject to nonresponse (y), the point estimators are themselves severely biased under all four imputation methods. As such, variance estimation is of little interest, as the real interest lies in estimating the mean squared error. That is, more attention needs to be concentrated on improving the point estimates and their bias. Some preliminary results on this front have been put forth by Rancourt, Lee and Särndal (1992).

5. CONCLUDING REMARKS

It is well known that the usual variance estimator understates the variance of the estimate of $\bar{Y}$ in the presence of imputed values if these values are treated as having been observed. In this study we again demonstrated the high degree of underestimation of the naive variance estimator in the presence of imputed data. Several imputation methods were considered in order to illuminate the dependence of the degree of underestimation on the method of imputation. We evaluated a unified jackknife variance estimator proposed by Rao and Shao (1992), an estimator that incorporates the variance due to imputation component. The study demonstrated some desirable properties of the proposed estimator in the case of both simple random sampling as well as complex survey designs. Our findings can be summarized as follows.

(1) The extent of variance underestimation is highly dependent on both the imputation method's ability to predict the true values, and its ability to preserve the natural variation in the data.

(2) The proposed adjusted jackknife variance estimator offers a unified approach to variance estimation of imputed data, that is easy to implement under a number of imputation methods and under designs of varying complexity.

(3) Operationally, no modifications to the original imputed file are necessary and the estimation of means and totals is thus unaffected by the need to estimate variances.

(4) The proposed method is easily extended to more complex designs, more than one imputation class and, with care, to the case of nonrandom nonresponse that depends only on available auxiliary variables.

(5) The adjusted jackknife variance estimator performs well whenever the nonresponse is uniform or the usual linear model holds, demonstrating the fact that the estimator is both design consistent as well as design-model unbiased.

(6) In the case of the P_L model, under which units with large y-values are more likely to not respond, all three variance estimators perform extremely poorly.

(7) In the case of y-dependent nonresponse, better imputation techniques are needed and the bias of the point estimators needs to be studied further. Here the issue is primarily that of estimating the mean square error rather than the variance.

Given the relatively high degree of imputation in today's surveys, at least within some imputation classes, it is clear that the effect of imputation on variance estimation cannot be ignored. An overestimation of precision can lead to confidence intervals that are too short and to spurious declaration of significance. If implementation of the above suggested methods is deemed too onerous in any particular circumstance, at the very least studies should be conducted to evaluate the impact of imputation in some representative cases. An *ad hoc* variance inflation factor could then be implemented. With the emergence of generalized estimation software, however, there seems to remain little reason for not implementing variance estimators which correctly account for the effect of imputation.

There clearly remain many unsolved, and perhaps unsolvable problems. To begin with, much more theoretical work is needed with respect to nearest neighbour imputation. The jackknife adjustments considered for this imputation method fail to perform as well as those applied to the other methods. Perhaps smoother alternatives to the nearest neighbour method need to be developed. Secondly, the robustness of the proposed estimator must be investigated. It is clear that satisfactory performance can be obtained if the model (15) holds, and when nonresponse is random. Limited failure of either one of these conditions did not seem to detract from the good performance of the jackknife estimator in our limited experience, but further research along these lines is warranted. Departures from both of the conditions simultaneously are yet to be investigated. Cases of nonrandom nonresponse when the propensity of nonresponse is related to the y-variable are even less well understood, though the emphasis in this case must be placed on the estimation of the mean square error rather than the variance. Thirdly, comparisons to multiple imputation results should be considered. It must be recognized, however, that proper imputation methods (Rubin 1987) must first be established. We note that none of the imputation methods studied within are proper with respect to multiple imputation.

Extensions to other imputation methods and other parameters of interest should be undertaken. This study was limited to four simple imputation methods. In practice, much more complicated methods are used, often in conjunction with each other. The impact of more than one imputation method on the estimation of variance has been studied by Rancourt, Lee and Särndal (1993); more work is needed. With respect to other, more complicated methods of imputation, the effect of adding theoretical residuals to imputed data can, for example, be considered. However, this technique only addresses the underestimation of V_{sam} by v_{naive} and ignores the effect of V_{imp}. Finally, other parameters, such as the median for example, and the effect of imputation on their variance are yet to be evaluated. Multivariate extensions can likewise be considered: estimation of correlations, ratios and regression parameters in the presence of imputation would likely be of interest.

ACKNOWLEDGEMENTS

The authors are grateful to Prof. J.N.K. Rao for his continuous encouragement and support, and the Associate Editor for his constructive comments.

REFERENCES

HANSEN, M., HURWITZ, W., and MADOW, W. (1953). *Sample Survey Methods and Theory*. (Volume 2), New York: J. Wiley, 139-141.

KALTON, G., and KASPRZYK, D. (1986). The treatment of missing survey data. *Survey Methodology*, 12, 1-16.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1991). Experiments with variance estimation from survey data with imputed values. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 690-695.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Unpublished report, Statistics Canada.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Unpublished report, Statistics Canada.

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

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1992). Bias corrections for survey estimates from data with imputed values for nonignorable nonresponse. *Proceedings 1992 Annual Research Conference*, Bureau of the Census, 523-539.

RANCOURT, E., LEE, H., and SÄRNDAL, C.-E. (1993). Variance estimation under more than one imputation method. *Proceedings of the International Conference on Establishment Surveys, American Statistical Association*, 374-379.

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

SÄRNDAL, C.-E. (1990). Methods for estimating the precision of survey estimates when imputation has been used. Special invited lecture. *Proceedings: Symposium 90, Measurement and Improvement of Data Quality*, Statistics Canada, 337-350.