Survey Methodology, December 1992
Vol. 18, No. 2, pp. 241–252
Statistics Canada

Methods for Estimating the Precision of Survey Estimates when Imputation Has Been Used

CARL-ERIK SÄRNDAL[1]

ABSTRACT

In almost all large surveys, some form of imputation is used. This paper develops a method for variance estimation when single (as opposed to multiple) imputation is used to create a completed data set. Imputation will never reproduce the true values (except in truly exceptional cases). The total error of the survey estimate is viewed in this paper as the sum of sampling error and imputation error. Consequently, an overall variance is derived as the sum of a sampling variance and an imputation variance. The principal theme is the estimation of these two components, using the data after imputation, that is, the actually observed values and the imputed values. The approach is model assisted in the sense that the model implied by the imputation method and the randomization distribution used for sample selection will together determine the appearance of the variance estimators. The theoretical findings are confirmed by a Monte Carlo simulation.

KEY WORDS: Single value imputation; Variance estimation; Imputation model; Model assisted inference.

1. DIFFERENT TYPES OF IMPUTATION

This paper reports work carried out in connection with the development of Statistics Canada's Generalized Estimation System (GES). Variance estimates are to be routinely calculated in the different estimation modules that define the GES. There was a need to develop suitable methods for variance estimation when the data set contains imputed values, which is the case in practically all surveys.

Two principal approaches to estimation with missing data are weighting and imputation. In the recent literature, the weights used to compensate for nonresponse are usually viewed as the inverse of the response probabilities associated with an assumed response mechanism. Since the response probabilities are ordinarily unknown, they need to be estimated from the available data. Imputation, on the other hand, has the advantage that it yields a complete data matrix. Such a matrix simplifies data handling, but it does not imply that "standard estimation methods" can be used directly. The imputed values are sample-based, thus they have their own statistical properties, such as a mean and a variance.

In our age, imputation is an extensively used tool. It is interesting to note what Pritzker, Ogus and Hansen (1965) say about imputation policy at the US Bureau of the Census: "Basically our philosophy in connection with the problem of ... imputation is that we should get information by direct measurement on a very high proportion of the aggregates to be tabulated, with sufficient control on quality that almost any reasonable rule for ... imputation will yield substantially the same results ... With respect to imputation in censuses and sample surveys we have adopted a standard that says we have a low level of imputation, of the order of 1 or 2 percent, as a goal."

[1] Carl-Erik Särndal, Département de mathématiques et de statistique, Université de Montréal, C.P. 6128, succursale A, Montréal (Québec) H3C 3J7.

Ideally, we should still strive for the goal of only one to two percent imputation. But in our time most surveys carried out by large survey organizations show a rate of imputation that is much higher. Clearly, if 30% of the values are imputed, the effects of imputation can not be ignored. Imputation can create systematic error (bias) in the point estimate; this is perhaps the most serious concern. But even if an imputation method can be found such that there is no appreciable systematic error, one must not ignore the often considerable effect that imputation has on the precision (the variance) of the point estimate. There is a need for simple yet valid variance estimation methods for survey data containing imputations, so that the coefficients of variation of the survey estimates can be properly reported.

A variety of imputation methods have been proposed. These can be classified in different ways. One way to classify is by the number of imputations carried out. In **single imputation** methods, a single value is imputed for a missing value. A complete data matrix is obtained, in which the imputed values are flagged. Estimates are calculated with the aid of the completed set. In **multiple imputation**, two or more values are imputed for each missing value. Several completed data sets are thus obtained. Estimates are calculated with the aid of the completed data sets.

Imputation methods also differ with respect to the modeling underlying the imputation. Some imputation methods use an **explicit** model, as when the imputed value is obtained by a regression fit, a ratio or mean imputation. In other methods, the model is only **implicit**, as in hot deck imputation and nearest neighbour donor imputation. The distinctions just made are important for this paper.

Statistics Canada currently uses imputation methods such as nearest neighbour donor, current ratio, current mean, previous value, previous mean, auxiliary trend. All of these are single imputation methods. The imputed values originate in the Generalized Edit and Imputation System (GEIS), from where they enter into the Generalized Estimation System (GES), where the point estimates and the variance estimates are calculated in a number of different estimation modules. This paper deals in particular with current ratio imputation, which represents a case of explicit modeling.

2. SOME THOUGHTS ON MULTIPLE IMPUTATION

Multiple imputation was suggested by D.B. Rubin around 1977. His ideas are explained in a number of papers, of which Herzog and Rubin (1983) and Rubin (1986) are expository, and in a book, Rubin (1987). Multiple imputation has advantages as well as disadvantages; the same is true for single imputation.

Rubin (1986) sees as a disadvantage of single imputation that "... the one imputed value cannot in itself represent uncertainty about which value to impute: If one value were really adequate, then that value was never missing. Hence, analyses that treat imputed values just like observed values generally systematically underestimate uncertainty, even assuming the precise reason for nonresponse are known."

Multiple imputation is attractive because it communicates the idea that imputation has variability. It is precisely this variability – the variability within and between the several completed data sets – that is exploited in the variance estimation methods proposed under multiple imputation. These methods make powerful use of basic statistical concepts. (On the other hand, one can argue that sample selection also has variability, but most surveys cannot afford more than a single sample, and estimation must be carried out with this unique sample.)

Simple examples show that treating imputed values just like observed values can lead to severe underestimation of the true uncertainty; survey samplers have long been aware of this. And it is a fact that users sometimes treat imputed values just like observed values, with wrong statement of precision as a result. With modern computers, it is easy to impute by some rule or another, but not so easy to obtain valid variance estimates.

The citation above seems to conclude that because a single imputed value does not display variation, we cannot obtain reasonable variance estimates; we are necessarily led to underestimation. I do not share this opinion. The methods that I discuss show that valid variance estimation is indeed possible with single imputation.

A method for variance estimation in the presence of imputed values should have the following properties: (a) a sound theoretical backing; (b) robustness to the assumptions underlying the imputation; (c) it must be practical, easy to carry out, and readily accepted by users.

While multiple imputation has the ingredients (a) and (b), it is clear that, in some applications at least, it does not have the property (c). In the development of the GES we must depend on procedures that are easy to administer and easy to accept by the user. The user of a data set (someone who is not primarily a statistician) can easily understand that the statistician imputes once, with the objective to fill in the best possible value for one that is missing. While it is true that for some purposes, such as secondary analyses, it might be interesting to have several completed data matrices, the costs of storage of multiple data sets will often rule out this option.

Multiple imputation may well be useful in other contexts and for other reasons than those that are essential to the development of the GES. The multiple imputation method has indicated one way of handling the problem of understatement of the variance, at least for some situations. The method has recently come under criticism by Fay (1991) and is not the only answer. Let us see what can be done with single imputation methods. The method described below is based on Särndal (1990).

3. IMPUTATION VARIANCE AND SAMPLING VARIANCE

An imputation rule corresponds to an (explicit or implicit) model for the relationship among variables of interest to the survey. That is, when the analyst has fixed an imputation rule, he or she has in fact chosen a model. The principle for the developments that follow is that if this rule is considered good enough for the point estimates (no systematic error), the rule is also good enough for the corresponding estimates of variance. In other words, the model maker should take responsibility for control of the bias as well as for the appropriateness of the variance estimate.

Let $U = \{1, \ldots, k, \ldots, N\}$ be a finite population; let y denote one of the study variables in the survey. The objective is to estimate the population total of y, $t = \sum_U y_k$. (If C is any set of population units, where $C \subseteq U$, $\sum_C$ is used as shorthand for $\sum_{k \in C}$, for example, $t = \sum_U y_k$ means $\sum_{k \in U} y_k$.) A probability samples s is selected with a given sampling design. The inclusion probabilities are known, and ordinary design-based variance estimates would be obtained if all units $k \in s$ are observed. However, there are missing data. Let r be the subset s for which the values y_k are actually observed. For the complement, $s - r$, imputations are calculated. The **data after imputation** consist of the values denoted $y_{\bullet k}$, $k \in s$, such that

$$y_{\bullet k} = \begin{cases} y_k & \text{if } \; k \in r \\ y_{\text{imp},k} & \text{if } \; k \in s-r \,, \end{cases}$$

where y_k is an actually observed value, and $y_{\text{imp},k}$ denotes the imputed value for the unit k. The case $r = s$ implies no imputation; all data are actual observations.

Let us write the estimator of t that would be used in the case of 100% response (that is, $r = s$) as $\hat{t} = \sum_{k\epsilon s} w_k y_k = \sum_s w_k y_k$, where w_k is the weight given to the observation y_k. For example, in simple random sampling without replacement (SRSWOR) of n units from N, $w_k = N/n$ for all $k \epsilon s$ when the expanded sample mean is used to estimate t, and $w_k = (\bar{z}_U/\bar{z}_s)(N/n) = (\sum_U z_k)/(\sum_s z_k)$ for all $k \epsilon s$ when the ratio estimator is used with z as an auxiliary variable.

When the data contain imputations, the estimator of t is $\hat{t}_\blacklozenge = \sum_s w_k y_{\blacklozenge k}$. That is, we assume that the weights w_k are identical to those used when all data are actual observations. This principle is used in the estimation modules of the GES. It embodies an assumption that imputation by the chosen rule causes little or no systematic error in the estimates.

The variance of an estimated total is increased by imputation, because imputation does not (except in truly exceptional circumstances) reproduce the true value y_k. Concrete evidence of this is the fact that if the imputation rule is applied to the actually observed sample units, there will always be error. If the rule is not without error for the responding units, it is not without error for the nonresponding units either. In Section 4 we express the variance of $\hat{t}_\blacklozenge$ as a sum of two components, a sampling variance, and a variance due to imputation,

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}.$$

The imputation variance V_{imp} is zero if all data are actually observed values, or if the imputation procedure is capable of exactly reproducing the true value y_k for every unit requiring imputation. (Neither case is likely in practice.) The procedure given in Section 4 uses the data after imputation, $y_{\blacklozenge k}$, $k \epsilon s$, to obtain estimates of each of the two components, leading to

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}.$$

The component $\hat{V}_{\text{sam}}$ is calculated in two steps:

(1) Compute the standard design-based variance estimate using the data after imputation. (For example, if SRSWOR is used, and $r = s$, the standard unbiased variance estimate of $N\bar{y}_s$ is $N^2(1/n - 1/N)\sum_s(y_k - \bar{y}_s)^2/(n-1)$. This formula, calculated on the data after imputation, yields $N^2(1/n - 1/N)\sum_s(y_{\blacklozenge k} - \bar{y}_{\blacklozenge s})^2/(n - 1)$, where $\bar{y}_{\blacklozenge s}$ is the mean of the n values $y_{\blacklozenge k}$.)

(2) Add a term to correct for the fact that many imputation rules give data with "less than natural" variability, which would lead to understatement of the sampling variance unless corrective action is taken. Finally, the component $\hat{V}_{\text{imp}}$ is readily computed from the data after imputation. The user will easily accept the argument that the variance obtained by the standard formula is not sufficient in itself; something must be added because the imputation rule is less than perfect.

The method has the good property that if no imputation is required, that is, $r = s$, then $\hat{V}_{\text{imp}} = 0$ and $\hat{V}_{\text{sam}}$ equals the "standard variance estimator" that one would have used with 100% actually observed values.

4. THEORETICAL DEVELOPMENTS

The total error of $\hat{t}_\blacklozenge$ is decomposed as

$$\hat{t}_\blacklozenge - t = (\hat{t} - t) + (\hat{t}_\blacklozenge - \hat{t}) = \text{sampling error} + \text{imputation error}.$$

The imputation error is the difference between the unknown estimate that would have been calculated if the data had consisted entirely of actual observations and the estimate that can be calculated on the data after imputation. The imputation error is

$$\hat{t}_{\bullet} - \hat{t} = - \sum_{s-r} w_k e_k,$$

where

$$e_k = y_k - y_{\text{imp},k}$$

is an **imputation residual** which can not be observed for a unit $k \in s-r$. The magnitude of e_k depends on how well the imputation model fits. The residuals are small if the imputation method gives nearly perfect substitute values. To pursue the argument, different directions may be taken. Here, we use a **model assisted** approach in which three different probability distributions are considered. The corresponding expectation symbols are written as E_ξ, E_s, and E_r. Here, ξ indicates "with respect to the imputation model"; s indicates "with respect to the sampling design", and r indicates "with respect to the response mechanism, given s". The model is implied by the imputation rule, so it is known; the sampling design is the given probability sampling distribution, so it is also known; the response mechanism is an ordinarily unknown distribution governing the response, given the sample s.

The estimator $\hat{t}_{\bullet}$ is overall unbiased in the sense that $E_\xi E_s E_r(\hat{t}_{\bullet} - t) = 0$ if two conditions hold:

(a) the order of the expectation operators can be changed so that $E_\xi E_s E_r(\cdot)$ can be evaluated as $E_s E_r\{E_\xi(\cdot \mid s, r)\}$, and

(b) the imputation residual $e_k = y_k - y_{\text{imp},k}$ has zero model expectation for every $k \in r$, that is, $E_\xi(e_k) = 0$, which implies that $E_\xi(\hat{t}_{\bullet} - \hat{t}) = 0$.

Condition (a) is satisfied if the response mechanism is one that may depend on s and on auxiliary data, but not on the y-values, y_k, $k \in s$. That is, the probability $q(r)$ of realizing the response set r is of the form $q(r) = q(r \mid s, \{x_k : k \in s\})$, where $\{x_k : k \in s\}$ denote the auxiliary data. The response mechanism can then be said to be ignorable.

We now examine the overall variance given by

$$V_{\text{tot}} = E_\xi E_s E_r\{(\hat{t}_{\bullet} - t)^2\},$$

which may also be called the anticipated variance under the imputation model ξ. We obtain

$$\begin{aligned} V_{\text{tot}} &= E_{\xi sr}(\hat{t}_{\bullet}) = E_\xi E_s E_r\{(\hat{t}_{\bullet} - t)^2\} \\ &= E_\xi E_s E_r\{(\hat{t} - t) + (\hat{t}_{\bullet} - \hat{t})\}^2 \\ &= E_\xi V_p + E_s E_r V_{\xi c}, \qquad (4.1) \end{aligned}$$

where $V_p = E_s\{(\hat{t} - t)\}^2$ is the design-based variance of $\hat{t}$, supposing $\hat{t}$ is design unbiased for the total t. (For an estimator with some slight design bias, V_p is the design-based mean square error of $\hat{t}$.) Note that $(\hat{t} - t)$ depends on s only, and not on r. Moreover,

$$V_{\xi c} = E_\xi\{(\hat{t}_{\bullet} - \hat{t})^2 \mid s, r\}$$

is the model variance of the imputation error, conditionally on s and r. The subscript c stands for "conditional". The derivation of (4.1) assumes that condition (a) holds so that the expectation E_ξ can be moved inside E_sE_r, and that the mixed term

$$2E_\xi E_s[(\hat{t} - t)\{E_r(\hat{t}_\bullet - \hat{t}) \mid s\}] \tag{4.2}$$

vanishes or is sufficiently close to zero that we can ignore it. This would be the case if the expected imputation error is zero or negligible under the response mechanism, conditionally on the realized probability sample s. Even if (4.2) is not exactly zero for the mechanism that determines the response, we can in many cases approximate (4.2) by zero and still use the method below to obtain a variance estimate that is much better than pretending naively that imputed data are as good as actually observed data. For ratio imputation and SRSWOR, which is an application considered in Section 5, the term (4.2) is exactly zero.

If we denote $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_sE_rV_{\xi c}$ in (4.1), then

$$V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$$

or

$$\text{overall variance} = \text{sampling variance} + \text{imputation variance}.$$

The objective is to estimate the overall variance, so that a valid confidence interval for the unknown t can be calculated. Our approach is to obtain separate estimates, $\hat{V}_{\text{sam}}$ and $\hat{V}_{\text{imp}}$, of the two components $V_{\text{sam}} = E_\xi V_p$ and $V_{\text{imp}} = E_sE_rV_{\xi c}$. The data available for this estimation are $y_{\bullet k}$, $k \epsilon s$. The argument for obtaining $\hat{V}_{\text{sam}}$ and $\hat{V}_{\text{imp}}$ is as follows:

(i) Estimation of the sampling variance component. Let $\hat{V}_p$ be the standard (design-unbiased or nearly design-unbiased) estimator of the design variance V_s. Denote by $\hat{V}_{\bullet p}$ the quantity obtained by calculating $\hat{V}_p$ from the data after imputation, $y_{\bullet k}$, $k \epsilon s$. For many imputation rules, $\hat{V}_{\bullet p}$ underestimates V_{sam}. The underestimation is compensated in the following way. Evaluate the conditional expectation

$$E_\xi(\hat{V}_p - \hat{V}_{\bullet p} \mid s, r) = V_{\text{dif}}.$$

Then for given s and r, find a model unbiased estimator, denoted $\hat{V}_{\text{dif}}$, of V_{dif}. This will usually require the estimation of certain parameters of the model ξ. Consequently,

$$E_\xi(\hat{V}_{\text{dif}} \mid s, r) = E_\xi(\hat{V}_p - \hat{V}_{\bullet p} \mid s, r).$$

Then

$$\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}}$$

is overall unbiased for the component $V_{\text{sam}} = E_\xi V_p$, as the following derivation shows:

$$\begin{aligned} E_\xi E_s E_r(\hat{V}_{\text{sam}}) &= E_sE_r\{E_\xi(\hat{V}_{\bullet p}) + E_\xi(\hat{V}_{\text{dif}})\} \\ &= E_sE_r\{E_\xi(\hat{V}_p)\} = E_\xi E_s(\hat{V}_p) \\ &= E_\xi V_p = V_{\text{sam}}. \end{aligned}$$

(ii) Estimation of the imputation variance component. Simply find an estimator, $\hat{V}_{\xi c}$, that is model unbiased for $V_{\xi c}$. That is, $E_\xi(\hat{V}_{\xi c}) = V_{\xi c}$. Again, this may require the estimation of unknown parameters of the model ξ. Then $\hat{V}_{\xi c}$ is overall unbiased for the imputation variance component V_{imp}, since

$$E_s E_r E_\xi(\hat{V}_{\xi c}) = E_s E_r V_{\xi c} = V_{\text{imp}}.$$

Finally, an overall unbiased estimator of V_{tot} is given by

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}},$$

where $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}}$ and $\hat{V}_{\text{imp}} = \hat{V}_{\xi c}$. Note that the role of $\hat{V}_{\text{dif}}$ is to correct for the fact that the data after imputation may display "less than natural" variation. This often happens when $y_{\text{imp},k}$ equals the predicted value from a fitted regression, that is, "the value on the line". The variation around the line is not reflected in the predicted value.

To be overall unbiased, the estimator $\hat{V}_{\text{tot}}$ constructed above requires that condition (a) holds, that (4.2) is zero, and that the imputation model is correct, so that $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$ are model unbiased for V_{dif} and $V_{\xi c}$, respectively. Mild departures from the assumed imputation model may not have serious consequences, but if the imputation model is grossly misspecified it is clear that $\hat{V}_{\text{tot}}$ may be considerably biased because of the model bias of $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$. Monte Carlo simulations reported in Lee, Rancourt and Särndal (1992) show that the variance estimator $\hat{V}_{\text{tot}}$ is fairly robust to imputation model breakdown. To add the terms $\hat{V}_{\text{dif}}$ and $\hat{V}_{\xi c}$ is in any case a vast improvement on simply using the naive uncorrected variance estimator $\hat{V}_{\bullet p}$.

Note that if the imputation model holds, an unbiased variance estimate is obtained with the method even if the response probabilities differ among units, as long as they depend on the x_k-values only. That is, we can allow a systematic response pattern such that large x_k-value units are less likely to respond than small x_k-value units. If the response probabilities depend explicitly the y_k-values, then the situation is different; the response mechanism is nonignorable and condition (a) does not hold. There will now be bias in $\hat{V}_{\text{tot}}$ due to nonignorability; the simulations in Lee, Rancourt and Särndal (1992) throw some light on the magnitude of this bias.

Example. The sample s is drawn with SRSWOR; n units from N. Let m denote the size of the response set r. Suppose the respondent mean is imputed for units requiring imputation. The corresponding imputation model ξ states that $y_k = \beta + \epsilon_k$, where the ϵ_k are uncorrelated errors terms with $E_\xi(\epsilon_k) = 0$, $V_\xi(\epsilon_k) = \sigma^2$. That is, $y_{\bullet k} = y_k$ if $k \epsilon r$ and $y_{\bullet k} = \hat{\beta} = \bar{y}_r$ if $k \epsilon s - r$, and we obtain the estimator $\hat{t}_\bullet = (N/n) \sum_s y_{\bullet k} = N\bar{y}_r$. Here the standard design-based variance estimator for 100% response is $\hat{V}_p = N^2(1/n - 1/N) \sum_s (y_k - \bar{y}_s)^2/(n-1)$; when this formula is computed on data after imputation we get $\hat{V}_{\bullet p} = N^2(1/n - 1/N) \{(m-1)/(n-1)\} S_{yr}^2$, where $S_{yr}^2 = \sum_r (y_k - \bar{y}_r)^2/(m-1)$. Other derivations give $\hat{V}_{\text{dif}} = N^2(1/n - 1/N)\{(n-m)/(n-1)\} S_{yr}^2$ and $\hat{V}_{\text{imp}} = N^2(1/m - 1/n) S_{yr}^2$. Thus, $\hat{V}_{\text{sam}} = \hat{V}_{\bullet p} + \hat{V}_{\text{dif}} = N^2(1/n - 1/N) S_{yr}^2$, and $\hat{V}_{\text{tot}} = N^2(1/m - 1/N) S_{yr}^2$, which is easy to accept as a "good" variance estimator for this simple imputation rule. The following table shows the contribution of each of the three terms to the total variance estimator $\hat{V}_{\text{tot}}$, for different rates of imputation, assuming that N is large compared to m and n, and $(m-1)/m \approx (n-1)/n \approx 1$.

Imputation rate in %	% contribution to $\hat{V}_{\text{tot}}$		
100 (1 − m/n)	$\hat{V}_{\blacklozenge p}$	$\hat{V}_{\text{dif}}$	$\hat{V}_{\text{imp}}$
10	81	9	10
20	64	16	20
30	49	21	30

The table illustrates the dangers of acting as if imputations are real data: with 30% imputed values, the standard formula variance estimator $\hat{V}_{\blacklozenge p}$ in this example covers less than half of the correctly estimated total variance. Imputation by the respondent mean is useful as an example; the results are particularly simple. But usually in practice, respondent mean imputation is neither justified nor efficient. The underlying model is not sophisticated enough to avoid systematic error in the point estimates, and the residuals $e_k = y_k - \bar{y}_r$ can vary considerably.

5. APPLICATION TO IMPUTATION BY THE CURRENT RATIO METHOD

The method assumes that a positive auxiliary value x_k is known for every unit $k \epsilon s$. If $k \in s-r$, we impute $y_{\text{imp},k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$. The data after imputation are

$$y_{\blacklozenge k} = \begin{cases} y_k & \text{if} \quad k \in r \\ \hat{B}x_k & \text{if} \quad k \in s-r. \end{cases}$$

The model behind current ratio imputation is

$$y_k = \beta x_k + \epsilon_k, \tag{5.1}$$

where the ϵ_k are uncorrelated model errors such that

$$E_\xi(\epsilon_k) = 0, \quad V_\xi(\epsilon_k) = \sigma^2 x_k. \tag{5.2}$$

Suppose that the sample s is selected by SRSWOR. Let the respective sizes of s, r, and $s - r$ be n, m, and $n - m$. If no imputation was needed, the estimator of $t = \sum_U y_k$ would be $\hat{t} = N\bar{y}_s$.Using the data after imputation, we get

$$\hat{t}_\blacklozenge = (N/n) \sum_s y_{\blacklozenge k} = N\bar{x}_s\bar{y}_r/\bar{x}_r. \tag{5.3}$$

(Overbar and subscript s, r, or $s - r$ indicates "straight mean", for example, $\bar{y}_r = \sum_r y_k/m$, $\bar{x}_{s-r} = \sum_{s-r} x_k/(n - m)$, *etc.*) Using the results of the preceding section, we have $V_{\text{tot}} = V_{\text{sam}} + V_{\text{imp}}$ with $V_{\text{sam}} = E_\xi\{N^2(1/n - 1/N)S^2_{yU}\}$ and $V_{\text{imp}} = E_s E_r\{N^2(1/m - 1/n)C_1\sigma^2\}$, where $S^2_{yU} = \sum_U(y_k - \bar{y}_U)^2/(N - 1)$ and $C_1 = \bar{x}_s\bar{x}_{s-r}/\bar{x}_r$, a known constant. The mixed term (4.2) is exactly zero in this case. Our method of variance estimation gives $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$, where

$$\hat{V}_{\text{sam}} = N^2(1/n - 1/N)\{S^2_{y\blacklozenge s} + C_0\hat{\sigma}^2\}, \tag{5.4}$$

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)C_1\hat{\sigma}^2, \tag{5.5}$$

where $S^2_{y\blacklozenge s} = \sum_s(y_{\blacklozenge k} - \bar{y}_{\blacklozenge s})^2/(n - 1)$ is the variance calculated on data after imputation, and we have chosen to estimate σ^2 by the model unbiased formula

$$\sigma^2 = \frac{1}{\bar{x}_r\{1 - (1/m)(cv_{xr})^2\}} \frac{\sum_r(y_k - \hat{B}x_k)^2}{m - 1},$$

where $cv_{xr} = S_{xr}/\bar{x}_r$ is the coefficient of variation of x in the response set r. The constant C_0 is obtained as

$$C_0 = \frac{1}{\sigma^2}E_\xi(S^2_{ys} - S^2_{y\blacklozenge s}),$$

where

$$S^2_{ys} = \frac{1}{n - 1}\sum_s (y_k - \bar{y}_s)^2$$

is the (unknown) sample variance based on data with 100% actual observations. After evaluation,

$$C_0 = \frac{1}{n - 1}\left\{\sum_{s-r} x_k - \frac{\sum_{s-r}x_k^2}{\sum_r x_k} + \frac{1}{n}\frac{\sum_{s-r}x_k\sum_s x_k}{\sum_r x_k}\right\}.$$

If m is not too small, the approximations $\hat{\sigma}^2 \approx (\sum_r e_k^2)/(\sum_r x_k)$ with $e_k = y_k - \hat{B}x_k$ and $C_0 \approx (1 - m/n)\bar{x}_{s-r}$ are sufficiently good for most applications.

We can write the imputation variance component as

$$\hat{V}_{\text{imp}} = N^2(1/m - 1/n)A\bar{x}_s\hat{\sigma}^2,$$

where $A = \bar{x}_{s-r}/\bar{x}_r$. The constant A reflects the selection effect due to nonresponse. If large units are less inclined to respond than small units, then A may be considerably greater than unity, and, for a given a sample s and a given number m of respondents, the component $\hat{V}_{\text{imp}}$ tends to be large, relative to a case where, say, all units are equally likely to respond. This tendency makes good sense intuitively.

Two special cases are noted: (1) If all $x_k = 1$, the estimated total variance becomes simply

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}} = N^2(1/m - 1/N)S^2_{yr},$$

where S^2_{yr} is the variance of the m actual observations y_k. This agrees with the variance obtained under a two-phase sampling design with SRSWOR in each phase. (2) If no imputation is required, that is, if $s = r$, then $\hat{V}_{\text{imp}} = 0$, and

$$\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} = N^2(1/n - 1/N)S^2_{ys}.$$

That is, our method yields the well known variance estimator for SRSWOR.

A Monte Carlo study with 100,000 repeated response sets r was carried out to confirm the above results for current ratio imputation. A finite population of size $N = 100$ was generated according to the model consisting of (5.1) and (5.2). The typical response set r was obtained as follows: Draw a SRSWOR sample s of size $n = 30$; given s, generate r by a response mechanism in the form of independent Bernoulli trials, one for each $k \in s$, with probability θ_k for the outcome "response". Three different response mechanisms were used: Mechanism 1: θ_k increases with y_k in such a way that $\theta_k = 1 - \exp(-a_1 y_k)$; Mechanism 2: θ_k increases as y_k decreases in such a way that $\theta_k = \exp(-a_2 y_k)$; Mechanism 3: θ_k is constant at 0.7, that is, a uniform response mechanism. The constants a_1 and a_2 in the first two response mechanisms (which can be described as non-ignorable) were fixed to obtain an average response probability of 0.7. The sizes of the realized response sets r thus varied around a mean of 21 for all three mechanisms. For each r, the point estimate $\hat{t}_{\bullet}$ given by (5.3) was calculated as well as three different variance estimators, $\hat{V} = \hat{V}(\hat{t}_{\bullet})$. These were: (1) the **model assisted** variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ equal to the total of (5.4) and (5.5); (2) the **two-phase** sampling variance estimator $N^2(1/n - 1/N)S^2_{yr} + N^2(1/m - 1/n)\sum_r e_k^2/(m - 1)$, an estimator which follows from standard two-phase sampling theory with an assumption of SRSWOR subsampling of m respondents from the n units in the initial sample (Rao 1990); and (3) the **standard unadjusted** variance estimator $N^2(1/n - 1/N)S^2_{y\bullet s}$ obtained by acting as if imputations are as good as actual data. The results are shown in the following table.

Estimator $\hat{V}$	Relative bias of $\hat{V}$ in %		
	Mechanism 1	Mechanism 2	Mechanism 3
Model assisted	−0.20	−4.64	−3.99
Two-phase	9.95	−12.49	−1.11
Standard unadjusted	−25.73	−37.90	−33.21

The relative bias of an estimator $\hat{V}$ was calculated as $\{\text{mean}(\hat{V}) - \text{var}(\hat{t}_{\bullet})\}/\text{var}(\hat{t}_{\bullet})$, where mean $(\hat{V})$ is the mean of the 100,000 values of $\hat{V}$, and var$(\hat{t}_{\bullet})$ is the variance of the 100,000 values of $\hat{t}_{\bullet}$. The simulation shows that the model assisted variance estimator $\hat{V}_{\text{tot}} = \hat{V}_{\text{sam}} + \hat{V}_{\text{imp}}$ is nearly unbiased for all three response mechanisms. In a way, this is not surprising because the population was generated to agree with the ratio imputation model. Mechanisms 1 and 2 are of the nonignorable kind and do not verify condition (a) of Section 4 required for unbiasedness of $\hat{V}_{\text{tot}}$. Interestingly, though, in this example the bias of $\hat{V}_{\text{tot}}$ remains small despite this. The two-phase estimator works well for the uniform response mechanism 3, the case for which it was conceived; otherwise it is biased. Finally, to act as if imputed data are as good as actual data leads, as expected, to a dramatic understatement of the true variance for all three mechanisms. A more extensive Monte Carlo study of ratio estimation is reported in Lee, Rancourt and Särndal (1992). This paper gives an idea of the effect of imputation model misspecification, which is also discussed in Rao (1992).

6. IMPUTED VALUES THAT HAVE AN ADDED RESIDUAL

We can distinguish two types of imputed values: (1) the imputed value $y_{\text{imp},k}$ consists of a predicted value only, $y_{\text{pred},k}$, as when the value on a fitted regression line or surface is used. For example in the current ratio imputation method as used above, $y_{\text{imp},k} = y_{\text{pred},k} = \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$; (2) the imputed value $y_{\text{imp},k}$ consists of a predicted value and a residual, so that $y_{\mathrm{imp},k} = y_{\mathrm{pred},k} + e_k^*$. The residual term, whose purpose is to make imputed values more like actual observations, may be obtained by sampling the residuals $e_k = y_k - y_{\mathrm{pred},k}$ calculated for the responding units $k \epsilon r$. A scheme for this is given below. This type of imputation is sometimes recommended in the literature as a means of preserving the distributions of the imputed data; see, for example, the discussion in Little (1988). The imputation process then requires more effort to complete, and for the purposes of the GES (whose principal aim is valid estimation of the precision of survey estimates), it is not clear that the advantages gained are worth the extra effort.

Let us, however, indicate one scheme for imputation by "predicted value plus residual" in the case where the current ratio imputation model is taken as the point of departure: For $k \epsilon r$, calculate $e_k = y_k - \hat{B}x_k$ with $\hat{B} = (\sum_r y_k)/(\sum_r x_k)$, then $\tilde{e}_k = e_k/\sqrt{x_k}$. This gives a supply of m "standardized residuals" $\tilde{e}_k$. Then for a unit $k \in s-r$, calculate $e_k^0 = \sqrt{x_k}\tilde{e}_k$, where $\tilde{e}_k$ is drawn by SRSWR from the supply, and x_k belongs to the unit requiring imputation. Then large x-value units tend to obtain larger residuals e_k^o, which is consistent with the model. Then set $e_k^* = e_k^o - (\sum_{s-r} e_k^o)/(n - m)$. For $k \in s-r$, impute $y_{\mathrm{imp},k} = \hat{B}x_k + e_k^*$, $k \in s-r$; for $k \epsilon r$, we have actual observations, y_k. Since the e_k^* were made to sum to zero over $s - r$, the point estimator is given by $\hat{t}_\blacklozenge = (N/n)\sum_s y_{\blacklozenge k} = N\bar{x}_s\bar{y}_r/\bar{x}_r$ as in Section 5, but its variance is different. It can be shown that $E_\xi E_s E_r E_\# (S_{y\blacklozenge s}^2 - S_{ys}^2) \approx 0$, where $E_\#$ denotes average with respect to the random selection of a standardized residual. That is, the difference between the variance calculated on data after imputation, $S_{y\blacklozenge s}^2$, and the unknown variance of a sample consisting entirely of actual observations, S_{ys}^2, is approximately zero on the average. We can use $\hat{V}_{\mathrm{sam}} = N^2(1/n - 1/N)S_{y\blacklozenge s}^2$ as an approximately overall unbiased estimator of the sampling variance component. There is no need now to add a correction $\hat{V}_{\mathrm{dif}}$. However, an estimator of the imputation variance $V_{\mathrm{imp}} = N^2(1/m - 1/n)C_1\,\sigma^2$ must still be calculated and added to $\hat{V}_{\mathrm{sam}}$.

7. CONCLUDING REMARKS

The continued work on the variance estimation techniques outlined in this paper has the following objectives: (1) extensions to imputation procedures based on models that are implicit only, in particular the nearest neighbour donor method; (2) extensions to the case where there is a mixture of several imputation procedures in the same survey.

Deville and Särndal (1992) present results for an extension in which the Horwitz-Thompson estimator, $\hat{t} = \sum_s y_k/\pi_k$, serves as the prototype. The estimator using data after imputation is then

$$\hat{t}_\blacklozenge = \sum_r y_k/\pi_k + \Bigl(\sum_{s-r} x_k/\pi_k\Bigr)'\hat{B} = \sum_s y_k/\pi_k - \sum_{s-r} e_k/\pi_k,$$

where $e_k = y_k - x_k'B$ is the imputation residual for unit k obtained by multiple regression.

ACKNOWLEDGEMENTS

I am indebted to M. Hidiroglou, P. Lavallée, Y. Leblond, H. Lee, and G. Reinhardt of Statistics Canada for their collaboration in the work that led to this paper. The comments of two referees led to improvements in the original manuscript and are gratefully acknowledged.

REFERENCES

DEVILLE, J.C., and SÄRNDAL, C.-E. (1992). Variance estimation for survey data with regression imputation. Technical report.

HERZOG, T.N., and RUBIN, D.B. (1983). Using multiple imputations to handle nonresponse in surveys. In *Incomplete Data in Sample Surveys*. (Eds. W.G. Madow, I. Olkin and D.B. Rubin). New York: Academic Press, 209-245.

FAY, R.E. (1991). A design-based perspective on missing data variance. Proceedings, 1991 Annual Research Conference, U.S. Bureau of the Census, 429-440.

LEE, H., RANCOURT, E., and SÄRNDAL, C.-E. (1992). Experiments with variance estimation from survey data with imputed values. Report, Business Survey Methods Division, Statistics Canada, submitted for publication.

LITTLE, R.J.A. (1988). Missing-data adjustments in large surveys (with discussion). *Journal of Business and Economic Statistics*, 6, 287-301.

PRITZKER, L., OGUS, J., and HANSEN, M.H. (1965). Computer editing methods: some applications and results. *Bulletin of the International Statistical Institute*, 41, 442-466.

RAO, J.N.K. (1990). Variance estimation under imputation for missing data. Manuscript seen by courtesy of the author.

RAO, J.N.K. (1992). Jackknife variance estimation under imputation for missing survey data. Manuscript seen by courtesy of the author.

RUBIN, D.B. (1986). Basic ideas of multiple imputation for nonresponse. *Survey Methodology*, 12, 37-47.

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

SÄRNDAL, C.-E. (1990). Estimation of precision in the generalized estimation system when imputation is used. Report, Informatics and Methodology Field, Statistics Canada, March 31, 1990.