Survey Methodology, June 2001
Vol. 27, No. 1, pp. 75-83
Statistics Canada, Catalogue No. 12-001

Variance Estimation After Imputation

Jae-Kwang Kim [1]

Abstract

Imputation is commonly used to compensate for item nonresponse. Variance estimation after imputation has generated considerable discussion and several variance estimators have been proposed. We propose a variance estimator based on a pseudo data set used only for variance estimation. Standard complete data variance estimators applied to the pseudo data set lead to consistent estimators for linear estimators under various imputation methods, including without-replacement hot deck imputation and with-replacement hot deck imputation. The asymptotic equivalence of the proposed method and the adjusted jackknife method of Rao and Sitter (1995) is illustrated. The proposed method is directly applicable to variance estimation for two-phase sampling.

Key Words: Two phase sampling; Item nonresponse; Deterministic imputation; Random imputation.

1. Introduction

Imputation, inserting values for missing items, is commonly used for handling missing survey data. An advantage of imputation is its convenience. That is, we can apply standard complete data programs for computing point estimates to the imputed data set. Rubin (1996), Fay (1996), and Rao (1996) reviewed various issues on imputation.

All imputation methods use some type of model. After designating a model, we can use either deterministic imputation or random imputation based on the model. Under random imputation, missing values are imputed by the use of some form of probability sampling. We call this additional random mechanism the imputation mechanism. On the other hand, deterministic imputation does not introduce an additional random mechanism. When the set of respondents is viewed as a random sample from the original sample, the selection mechanism of the respondents is called the response mechanism. The response mechanism is often regarded as the second phase of sampling. See Särndal and Swensson (1987) for details.

With a suitable imputation model and method, the bias due to nonresponse can be greatly reduced relative to using only the observed data. However, it is well known that a variance estimator which uses the imputed data as if it were observed data is inconsistent.

Various methods have been proposed for variance estimation after imputation. Rubin and Schenker (1986) and Rubin (1987) advocate multiple imputation. Multiple imputation creates multiple data sets and calculates the complete data statistics for each imputed data set. The variance estimator is calculated by combining two terms, the within-dataset variance term and the between-dataset variance term. Multiple imputation applies standard variance estimators to each data set to compute within-dataset variance terms and applies the standard point estimators to compute a between-imputed-dataset variance term. This method requires the imputation method to be proper. That is, the imputation should satisfy conditions 1-3 in Rubin (1987, pages 118-119). These conditions are not always easy to achieve. (For example, see Fay 1992). Even the multiple imputation methods described in Schafer (1997) are not shown to be proper in the sense of Rubin. As noted by Rao (1996), some commonly used imputation methods, including hot deck imputation and regression imputation, are not proper.

Rao and Shao (1992) and Rao and Sitter (1995) proposed an adjusted jackknife variance estimator. The suggested procedure is applicable to a number of imputation methods and sample designs. The actual calculation using standard complete data software is not easy because special computations are performed to adjust the imputed values for each pseudo replicate. Also, Särndal (1992) proposed a variance estimation method that explicitly uses the model considered for imputation.

Essentially, Rubin's method generates several pseudo data sets for variance estimation and applies the standard variance estimators to each data set to compute the within-dataset variance terms, while Rao's method and Särndal's method apply a special variance estimator to the imputed data set. In this paper, a method to create a single pseudo data set for variance estimation is proposed. In section 2, the new method is introduced in a two-phase sampling set-up. In section 3, we illustrate extensions of the suggested method to the random imputation method. In section 4, we extend the suggested method to complex sampling designs. In section 5, comparisons are made with the adjusted jackknife variance estimator. In section 6, a limited simulation study is presented. Some concluding remarks are made in section 7. Outlines of some proofs are given in the appendix.

1. Jae-Kwang Kim, Westat, 1650 Research Boulevard, Rockville, Maryland, 20850, U.S.A.

2. A Variance Estimation Method

We outline a variance estimation procedure applicable for two-phase samples and for imputed samples. The procedure requires a separate data set for variance estimation in addition to the tabulation data set. To introduce the procedure and to illustrate the concepts, consider a two-phase sample. Let the second phase be a simple random sample of size r selected from the first phase, which is a simple random sample of size n selected from an infinite population. Let the regression estimator of the mean of a characteristic y be

$$\hat{\mu}_y = \bar{y}_2 + (\bar{x}_1 - \bar{x}_2)\hat{\beta}, \tag{1}$$

where

$$(\bar{y}_2, \bar{x}_2) = r^{-1}\sum_{i=1}^{r}(y_i, x_i),$$

$$\bar{x}_1 = n^{-1}\sum_{i=1}^{n} x_i,$$

$$\hat{\beta} = \left[\sum_{i=1}^{r}(x_i - \bar{x}_2)^2\right]^{-1}\sum_{i=1}^{r}(x_i - \bar{x}_2)(y_i - \bar{y}_2)$$

and the second phase units are indexed from one to r. It is well known (*e.g.*, Cochran 1977, equation 12.51) that the variance of the regression estimator is, approximately,

$$V\{\hat{\mu}_y\} = [n^{-1}\rho^2 + r^{-1}(1-\rho^2)]\sigma_y^2, \tag{2}$$

where ρ is the population correlation between y and x and σ_y^2, is the population variance of y. An estimator of the variance is, by classical regression theory,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + r^{-1}(r-2)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2 \tag{3}$$

where $\hat{y}_i = \bar{y}_2 + (x_i - \bar{x}_2)\hat{\beta}$ for $i = 1, 2, \ldots, n$, and $\bar{y}_I = n^{-1}\sum_{i=1}^{n}\hat{y}_i$. Observe that $\bar{y}_I$ is an alternative way of writing $\hat{\mu}_y$ in (1).

Let

$$c_r = [n(n-1)r^{-1}(r-2)^{-1}]^{1/2} \tag{4}$$

and

$$y_i^* = \begin{cases} \hat{y}_i, & i = r+1, r+2, \ldots, n \\ \hat{y}_i + c_r(y_i - \hat{y}_i), & i = 1, 2, \ldots, r. \end{cases} \tag{5}$$

Then,

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(y_i^* - \bar{y}_I)^2 \tag{6}$$

where $\bar{y}_I$ is the mean of the y_i^*, as well as the mean of the $\hat{y}_i$, because the sum of $y_i - \hat{y}_i$ is zero. Equation (6) is the operational form of the suggested estimator. The variance estimation data set contains the pseudo observation y_i^*.

To the extent that the model for imputation matches that of two-phase sampling, equation (6) is applicable to an imputed data set. For example, if we assume that missing data are missing at random and use regression to impute the missing value with $\hat{y}_i$, then equation (6) is immediately applicable. Of course, regression imputation or two-phase sampling can use a vector x.

3. Extensions to Random Imputation

A moderate extension of the method described in section 2 enables us to estimate the variance of a sample mean using random imputation. In fact, alternative approaches are possible.

As one approach, assume that the imputation model is the regression model

$$y_i = \mathbf{x}_i\beta + e_i \tag{7}$$

where the first element of every $\mathbf{x}_i$ is equal to 1 and the e_i are uncorrelated $(0, \sigma_e^2)$ random variables.

Assume the model is estimated and that the imputed values are

$$\ddot{y}_i = \hat{y}_i + \ddot{e}_i, \quad i = r+1, r+2, \ldots, n \tag{8}$$

where $\hat{y}_i = \mathbf{x}_i\hat{\beta}$ with $\hat{\beta} = (\sum_{i=1}^{r}\mathbf{x}_i'\mathbf{x}_i)^{-1}\sum_{i=1}^{r}\mathbf{x}_i'y_i$ and $\ddot{e}_i$ is chosen at random from the set $\hat{\mathbf{e}}_r = \{\hat{e}_i = y_i - \hat{y}_i; i = 1, 2, \ldots, r\}$. The estimator of the mean of y is

$$\hat{\mu}_y = n^{-1}\sum_{i=1}^{n}\ddot{y}_i \tag{9}$$

where $\ddot{y}_i = y_i$ if $i = 1, 2, \ldots, r$.

If the $\ddot{e}_i$ are chosen with replacement with equal probability from the set $\hat{\mathbf{e}}_r$, then the variance $\hat{\mu}_y$ is, approximately,

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}m)(1-R^2)]\sigma_y^2, \tag{10}$$

where $m = n - r$ and R^2 is the squared multiple correlation coefficient between y and $\mathbf{x}$. The increase in variance due to using random imputation with $\ddot{e}_i$, rather than using $\ddot{e}_i \equiv 0$, is $n^{-2}m(1-R^2)\sigma_y^2$.

Therefore, an estimator of the variance of the imputed sample mean is given by (6) where the c_r of (4) is

$$c_I = [n(n-1)(r^{-1} + n^{-2}m)(r-p)^{-1}]^{1/2}, \tag{11}$$

and p is the dimension of β. We have

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + (r^{-1} + n^{-2}m)(r-p)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2 \tag{12}$$

where $\bar{y}_I = \sum_{i=1}^{n} \hat{y}_i$. The estimator of the variance using c_I of equation (11) is an estimator of the unconditional variance, the average over all possible imputed sample. Derivations of (10) and (12) are given in Appendix A.

To consider an alternative variance estimation approach, we assume that a random selection procedure is used for imputation but place no restriction on the procedure, other than that the probabilities of selection are inversely proportional to the probability that the y-value responds. In addition, we record the number of times an $\hat{e}$ value is used as a donor in the imputation.

Let

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, ..., n \\ \hat{y}_i + c_r(y_i - \hat{y}_i) & i = 1, 2, ..., r \end{cases} \tag{13}$$

with

$$c_r = [n^{-1}(n-1)r(r-p)^{-1}]^{1/2}(1+d_i) \tag{14}$$

where d_i is the number of times $\hat{e}_i$ is used as a donor. The term $[n^{-1}(n-1)r(r-p)^{-1}]^{1/2}$ is used to adjust for the effect of estimating p regression parameters. Then, the variance estimator (6) can be written as

$$\hat{V}\{\hat{\mu}_y\} = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + n^{-2}r(r-p)^{-1}\sum_{i=1}^{r}(1+d_i)^2(y_i - \hat{y}_i)^2. \tag{15}$$

If the imputation method is simple random sampling with replacement, then, conditional on the sample and the respondents,

$$E_I\{(1+d_i)^2\} = \left(\frac{n}{r}\right)^2 + \frac{m}{r}\left(1 - \frac{1}{r}\right) \tag{16}$$

where the notation I is used here to denote the expectation with respect to the imputation mechanism generated by random imputation. The equality in (16) establishes the equivalence of (12) to (15) under with-replacement selection. It is shown in Appendix B that $\hat{V}\{\hat{\mu}_y\}$ in (15) is also a valid estimator when donors are selected without replacement. Since the proposed variance estimation method is the conditional variance given the realized imputed sample, it has wide applicability.

4. Complex Sampling Designs

4.1 Deterministic Imputation

The suggested method is applicable to complex sampling designs as well as to simple random sampling. Assume that the full sample estimator of the mean of y can be written as $\bar{y} = \sum_{i=1}^{n} w_i y_i$, where w_i is the sampling weight of unit i in the sample. Assume that $\sum_{i=1}^{n} w_i = 1$.

If the first r elements are observed and the remaining $n-r$ elements are missing, then the estimator of the mean of y under regression imputation is

$$\bar{y}_I = \sum_{i=1}^{r} w_i y_i + \sum_{i=r+1}^{n} w_i \hat{y}_i \tag{17}$$

where

$$\hat{y}_i = \mathbf{x}_i \hat{\beta},$$

$$\hat{\beta} = \left[\sum_{i=1}^{r} w_i^* \mathbf{x}_i' \mathbf{x}_i\right]^{-1} \sum_{i=1}^{r} w_i^* \mathbf{x}_i' y_i.$$

Here w_i^* is the sampling weight of unit i in the second-phase sample and is defined by

$$w_i^* = \left[\Pr(i \text{ is in the second phase sample} \mid i \text{ is in the first phase sample})\right]^{-1} w_i.$$

Also, $\sum_{i=1}^{r} w_i^* = 1$. If we assume that the second phase sample is a random sample of size r from the n first phase sample, then $w_i^* = nr^{-1}w_i$. Under certain conditions we can write the estimator in (17) as

$$\bar{y}_I = \sum_{i=1}^{n} w_i \hat{y}_i. \tag{18}$$

The representation (18) holds if $(w_i^*)^{-1}w_i$ is in the column space of the matrix $\mathbf{X} = (\mathbf{x}_1', ..., \mathbf{x}_r')'$ because then we have $\sum_{i=1}^{r} w_i(y_i - \hat{y}_i) = 0$ from $\sum_{i=1}^{r} w_i^* \mathbf{x}_i'(y_i - \hat{y}_i) = \mathbf{0}$.

We assume a sequence of samples and finite populations such as that described in Fuller (1998). Define $\bar{\mathbf{x}}_1 = \sum_{i=1}^{n} w_i \mathbf{x}_i$ and $(\bar{\mathbf{x}}_2, \bar{y}_2) = \sum_{i=1}^{r} w_i^*(\mathbf{x}_i, y_i)$. We also adopt the same assumptions as in Fuller (1998). That is

$$E(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2) = (\mu_x, \mu_x, \mu_y), \tag{19}$$

and

$$V\{(\hat{\beta} - \beta)', \bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \bar{y}_2\} = O(n^{-1}), \tag{20}$$

where $(\mu_{\mathbf{x}}, \mu_y) = N^{-1}\sum_{i=1}^{N}(\mathbf{x}_i, y_i)$ and $\beta = (\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i)^{-1} \sum_{i=1}^{N} \mathbf{x}_i' y_i$.

For $i = 1, 2, ..., N$, define

$$a_i = \begin{cases} 1 & \text{if unit } i \text{ responds when sampled} \\ 0 & \text{otherwise,} \end{cases}$$

and $\mathbf{a} = (a_1, a_2, ..., a_N)$. The extended definition of a_i is discussed by Fay (1991) and used in Shao and Steel (1999). Now, let

$$\bar{y}_{II} = \sum_{i=1}^{n} w_i \tilde{y}_i^* \tag{21}$$

where

$$\tilde{y}_i^* = \tilde{y}_i + a_i w_i^{-1} w_i^* (y_i - \tilde{y}_i) \tag{22}$$

with $\tilde{y}_i = \mathbf{x}_i\beta$. Then, we have $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. By (19) and (20), we have $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$ and $V(\bar{y}_I - \bar{Y}_N) = V(\bar{y}_{II} - \bar{Y}_N) + o(n^{-1})$. Now,

$$V(\bar{y}_{II} - \bar{Y}_N) = V[E(\bar{y}_{II} - \bar{Y}_N|\mathbf{a})] + E[V(\bar{y}_{II} - \bar{Y}_N)|\mathbf{a}]. \quad (23)$$

The first term on the right side of (23) is 0 because $E(\bar{y}_{II} - \bar{Y}_N|\mathbf{a}) = 0$ under model (7). To estimate the second term in (23), note that conditional on $\mathbf{a}$, $\bar{y}_{II}$ is a linear estimator. Hence, the standard variance estimation method applied to the pseudo data set $\tilde{\mathbf{Y}}^* \equiv \{\tilde{y}_i^*; i = 1, 2, ..., n\}$ will unbiasedly estimate the variance of $\bar{y}_{II} = \sum_{i=1}^n w_i\tilde{y}_i^*$. Since the set $\tilde{\mathbf{Y}}^*$ is not observable, we can use the set $\mathbf{Y}^* \equiv \{y_i^*; i = 1, 2, ..., n\}$, where

$$y_i^* = \hat{y}_i + a_i w_i^{-1} w_i^*(y_i - \hat{y}_i) \quad (24)$$

to get a consistent variance estimator.

To illustrate that the set $\mathbf{Y}^*$ can be used to approximate the variance estimator, assume that the full sample variance estimator of $\bar{y}$ can be written as

$$\hat{V} = \sum_{i=1}^L c_i(\bar{y}^{(i)} - \bar{y})^2$$

where L is the number of replications, c_i is the i^{th} replication factor, and $\bar{y}^{(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j$ is the i^{th} replicate of $\bar{y}$. The term $M_j^{(i)}$ is the replication multiplier applied to the weight of unit j at the i^{th} replication. For example, under simple random sampling, the jackknife multiplier is

$$M_j^{(i)} = \begin{cases} (n-1)^{-1}n & \text{if } i \neq j \\ 0 & \text{if } i = j. \end{cases}$$

Assume that the replicate variance estimator $\hat{V}$ is applied to the set $\mathbf{Y}^*$ to get

$$\hat{V}^* = \sum_{i=1}^L c_i(\bar{y}_I^{*(i)} - \bar{y}_I)^2$$

where $\bar{y}_I^{*(i)} = \sum_{j=1}^n w_j M_j^{(i)} y_j^*$ with y_j^* being defined in (24). Then, we have

$$\bar{y}_I^{*(i)} - \bar{y}_I = \bar{y}_{II}^{*(i)} - \bar{y}_{II} + (\bar{\mathbf{x}}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)} - \bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)(\hat{\beta} - \beta) \quad (25)$$

where

$$(\bar{\mathbf{x}}_1^{(i)}, \bar{\mathbf{x}}_2^{(i)}) = \sum_{j=1}^n w_j M_j^{(i)}(\mathbf{x}_j, a_j w_j^{-1} w_j^* \mathbf{x}_j).$$

It is shown in Appendix C that

$$\hat{V}^* = \sum_{i=1}^L c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2 + o_p(n^{-1}). \quad (26)$$

Therefore, the standard jackknife variance estimator applied to the pseudo data set $\mathbf{Y}^*$ can be used to approximate the standard jackknife variance estimator applied to the pseudo data set $\tilde{\mathbf{Y}}^*$.

4.2 Random Imputation

The arguments for variance estimation with random imputation are quite similar to those for deterministic imputation described in the previous subsection. First, define the imputation indicator function

$$d_{ij} = \begin{cases} 1 & \text{if unit } i \text{ is used as donor for unit } j \\ 0 & \text{otherwise.} \end{cases} \quad (27)$$

Then, the estimator of the mean of y using random imputation is

$$\bar{y}_I = \sum_{i=1}^n w_i y_i^* \quad (28)$$

where

$$\bar{y}_i^* = \hat{y}_i + a_i(1 + d_i)(y_i - \hat{y}_i) \quad (29)$$

and

$$d_i = \sum_{j=1}^n (1 - a_j) d_{ij} w_i^{-1} w_j. \quad (30)$$

If the original sample weights are the same, then d_i is the number of times that unit i is used as a donor. We assume that

$$E[a_i(1 + d_i)|F_1] = 1 \quad (31)$$

where $F_1 = \{(i, \mathbf{x}_i, y_i); i = 1, 2, ..., n\}$. The expectation in (31) is with respect to the joint distribution of the response mechanism and the imputation mechanism. Then, we have

$$E(\bar{y}_I|F_1) \doteq \bar{y}.$$

If we assume equal response probability, then, by (31), the probability of selection of donors should be proportional to the weights. This is the Rao and Shao (1992) setup for random imputation.

Now, let

$$\bar{y}_{II} = \sum_{i=1}^n w_i[\tilde{y}_i + a_i(1 + d_i)(y_i - \tilde{y}_i)] \quad (32)$$

where $\tilde{y}_i = \mathbf{x}_i\beta$. Then, we also have $\bar{y}_I = \bar{y}_{II} + (\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1)(\hat{\beta} - \beta)$ where $\bar{\mathbf{x}}_d = \sum_{i=1}^n w_i a_i(1 + d_i)\mathbf{x}_i$. By the assumption (31), we have $E(\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1|F_1) = 0$. Under mild conditions, $\bar{\mathbf{x}}_d - \bar{\mathbf{x}}_1 = O_p(n^{-1/2})$ and $\bar{y}_I = \bar{y}_{II} + O_p(n^{-1})$. Now,

$$V(\bar{y}_I - \bar{Y}_N) = V[E(\bar{y}_I - \bar{Y}_N|\mathbf{a},\mathbf{d})] + E[V(\bar{y}_I - \bar{Y}_N|\mathbf{a},\mathbf{d})]$$

where $\mathbf{d} = (d_1, d_2, ..., d_N)$. Conditional on $\mathbf{a}$ and $\mathbf{d}$, the estimator $\bar{y}_{II}$ is a linear estimator. Hence, the pseudo data

$$y_i^* = \hat{y}_i + a_i(1 + d_i)(y_i - \hat{y}_i) \quad (33)$$

can be used to estimate the variance of $\bar{y}_I$.

5. Comparisons with Adjusted Jackknife Method

Rao and Sitter (1995) proposed an adjusted jackknife variance estimator for the ratio imputation problem. Under the setup described in section 4, the ratio imputed estimator of μ_y is

$$\hat{\mu}_I = \sum_{i=1}^{n} w_i [a_i y_i + (1-a_i)\hat{y}_i]$$

with $\hat{y}_i = x_i \hat{R}$ and $\hat{R} = (\sum_{i=1}^{n} w_i a_i x_i)^{-1} \sum_{i=1}^{n} w_i a_i y_i$. The Rao and Sitter (1995) variance estimator is

$$V_a = \sum_{i=1}^{L} c_i (\hat{\mu}_I^{(i)} - \hat{\mu}_I)^2, \tag{34}$$

where the adjusted jackknife replicate at the i^{th} replication is

$$\hat{\mu}_I^{(i)} = \sum_{j=1}^{n} w_j M_j^{(i)} y_j^{*(i)} \tag{35}$$

where

$$y_j^{*(i)} = \begin{cases} x_j \hat{R}^{(i)} & \text{if } a_i = 1 \\ x_j \hat{R} & \text{if } a_i = 0 \end{cases} \tag{36}$$

with $\hat{R}^{(i)} = (\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j)^{-1} \sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j$. The adjusted values (36) in the Rao and Sitter (1995) method can also be regarded as pseudo data for variance estimation. Note that the calculation of the pseudo data (36) requires recalculation of $\hat{R}^{(i)}$ for each i with $a_i = 1$.

We modify the calculation of the pseudo values y_i^* in (5) to

$$y_i^* = \begin{cases} \hat{y}_i & \text{if } a_i = 0 \\ \hat{y}_i + c_r \left(\dfrac{\bar{x}_1}{\bar{x}_2} \right)(y_i - \hat{y}_i) & \text{if } a_i = 1, \end{cases} \tag{37}$$

where $\bar{x}_2 = \sum_{i=1}^{n} w_i r^{-1} n a_i x_i, \bar{x}_1 = n^{-1} \sum_{i=1}^{n} w_i x_i$ and $c_r \doteq r^{-1} n$. The term $(\bar{x}_1 / \bar{x}_2)$ is inserted to improve the conditional properties of V_J given the first phase sample. The resulting variance estimator is approximately equivalent to the adjusted jackknife variance estimator (34). To see this, note that the adjusted values (35) can be written in the form

$$\hat{\mu}_I^{(i)} = \left(\sum_{j=1}^{n} w_j M_j^{(i)} x_j \right) \frac{\sum_{j=1}^{n} w_j M_j^{(i)} a_j y_j}{\sum_{j=1}^{n} w_j M_j^{(i)} a_j x_j} =: \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}},$$

where $A =: B$ denotes that we define B to be A. Also, define $\hat{Z} = \sum_{j=1}^{n} w_j x_j, \hat{S} = \sum_{j=1}^{n} w_j a_j y_j$, and $\hat{T} = \sum_{j=1}^{n} w_j a_j x_j$. Then by the first order Taylor expansion,

$$\begin{aligned} \hat{Z}^{(i)} \frac{\hat{S}^{(i)}}{\hat{T}^{(i)}} &\doteq \hat{Z} \frac{\hat{S}}{\hat{T}} + (\hat{Z}^{(i)} - \hat{Z}) \frac{\hat{S}}{\hat{T}} \\ &\quad + (\hat{S}^{(i)} - \hat{S}) \frac{\hat{Z}}{\hat{T}} - (\hat{T}^{(i)} - \hat{T}) \frac{\hat{Z}\hat{S}}{\hat{T}^2} \\ &= \left[\hat{Z}^{(i)} \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} \left(\hat{S}^{(i)} - \hat{T}^{(i)} \frac{\hat{S}}{\hat{T}} \right) \right]. \end{aligned} \tag{38}$$

Note that the right side of (38) is exactly equal to

$$\sum_{j=1}^{n} w_j M^{(i)} \left[\frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_j \left(y_j - \frac{\hat{S}}{\hat{T}} \right) \right].$$

Thus, the pseudo data for variance estimation can be written as

$$y_i^* = \frac{\hat{S}}{\hat{T}} + \frac{\hat{Z}}{\hat{T}} a_i \left(y_i - \frac{\hat{S}}{\hat{T}} \right),$$

which reduces to (37). Hence, the proposed method is exactly a first order Taylor linearization of the Rao and Sitter method in the case of ratio imputation. Therefore, we can expect our proposed method to have the same asymptotic properties as the Rao and Sitter method up to the order of n^{-1}.

The variance estimation method using the pseudo data set calculated by (37) is easy to implement because we can directly use existing software, which is more difficult with the Rao and Shao (1992) or Rao and Sitter (1995) method. Furthermore, if we calculate the pseudo data by (13), then the data set works for without-replacement hot deck imputation as well as for with-replacement hot deck imputation.

6. A Simulation Study

The preceding theory was tested in a simulation study using an artificial, finite population, from which repeated samples were drawn. The population has $L = 32$ strata, N_h clusters in stratum h, and 20 ultimate units in each cluster. The values of the population parameters were chosen to correspond to real populations encountered in the U.S. National Assessment of Educational Progress Study (Hansen and Tepping 1985) and are listed in Table 1. The finite population units are

$$y_{hij} = y_{hi} + e_{hij},$$

where

$$y_{hi} \overset{\text{iid}}{\sim} N(\mu_h, \sigma_h^2),\ h = 1, 2, ..., L,\ i = 1, 2, ..., N_h,$$

and

$$e_{hij} \overset{\text{iid}}{\sim} N\left(0, \frac{1-\rho}{\rho}\sigma_h^2\right), \; j = 1, 2, ..., 20.$$

Shao, Chen and Chen (1998) also used the same population in their simulation study. The value of the intra-cluster correlation ρ considered in the simulation is $\rho = 0.3$. Simulations with other values of ρ produced similar results and are not listed here for brevity.

Table 1
Parameters of the Finite Population for Simulation

H	N_h	μ_h	σ_h	h	N_h	μ_h	σ_h
1	13	100.0	20.0	2	16	95.0	19.0
3	20	90.0	18.0	4	25	98.0	19.6
5	25	93.0	18.6	6	25	98.0	19.6
7	25	96.0	19.2	8	28	94.0	18.8
9	28	92.0	18.4	10	28	96.0	19.2
11	31	94.0	18.8	12	31	92.0	18.4
13	31	90.0	18.0	14	31	96.0	19.2
15	31	94.0	18.8	16	31	92.0	18.4
17	31	90.0	18.0	18	31	88.0	17.6
19	31	86.0	17.2	20	34	84.0	16.8
21	34	82.0	16.4	22	34	80.0	16.0
23	34	90.0	18.0	24	37	85.0	17.0
25	37	80.0	16.0	26	37	90.0	18.0
27	37	85.0	17.0	28	39	80.0	16.0
29	39	75.0	15.0	30	42	75.0	15.0
31	42	75.0	15.0	32	42	75.0	15.0

We consider a stratified cluster sampling design, where $n_h = 2$ clusters are selected with replacement from stratum h with equal probability and all of the ultimate units in the selected clusters are in the sample. The sampling fraction is 6.4%. For each sampled unit y_{hij}, a response indicator variable a_{hij} is generated from

$$a_{hij} \overset{\text{iid}}{\sim} \text{Bernoulli}\,(p),$$

and that a_{hij} is independent of y_{hij}. The value of p considered in the simulation are p = 0.9, 0.8, 0.7, 0.6, and 0.5.

A set of 5,000 samples were selected using the same sampling design. In each of the selected samples, three imputation methods are considered;

[M1] With-replacement weighted hot deck imputation considered by Rao and Shao (1992), where a missing value is imputed by a value randomly selected from the respondents with replacement with probability proportional to the survey weights.

[M2] Without-replacement weighted hot deck imputation, which is the same as [M1] expect that the selection was performed using a without-replacement sample. The without-replacement selection of donors is carried out systematically using the method described by Hansen, Hurwitz, and Madow (1953, page 343) from the respondents sorted by random order.

[M3] Overall mean imputation, where the weighted mean of the respondents in the sample is imputed.

Hence, all the imputation methods use a single imputation cell that collapses all the strata.

In each imputed data set we computed three variance estimators $\hat{V}_n$, naive variance estimator treating the imputed data as if it were observed data, $\hat{V}_a$, the adjusted jackknife variance estimator of Rao and Shao (1992) for [M1] and [M2] and of Rao and Sitter (1995) for [M3], and $\hat{V}^*$, the jackknife variance estimator based on the pseudo data. The pseudo data set is constructed by (29) for [M1] and [M2] and by (24) for [M3]. The complete sample variance estimator used a standard jackknife for stratified cluster sampling, in which a cluster is deleted for each replication. Note that the standard jackknife is a consistent estimator of the variance under the model with nonzero intracluster correlation. Thus, the standard jackknife method based on the pseudo data can be applicable to the data set considered. The point estimators of the population mean are unbiased under the three different imputation schemes and are not listed here.

Table 2 presents the relative bias of the three variance estimators, the standard error of the relative bias of the variance estimators, and the sample correlation coefficient between the Rao's adjusted jackknife variance estimator and the new variance estimator based on the 5,000 samples. The relative bias of $\hat{V}$ as an estimator of the variance of $\bar{y}_I$ is calculated by $[\text{Var}_B(\bar{y}_I)]^{-1}[E_B(\hat{V}) - \text{Var}_B(\bar{y}_I)]$, where the subscript B denotes the distribution generated by the Monte Carlo simulation. The correlation coefficients of the two variance estimators are computed to give a measure the relative linearity behavior of the two variance estimators.

Table 2
Relative Bias of the Variance Estimator, Standard Error of the Relative Bias, and Sample Correlation Coefficient Between the Rao's Variance Estimator and the New Variance Estimator Based on 5,000 Samples

Response Rate (p)	Imputation Method	Rel. Bias × 100 (S.E. × 100) Naive	Rao	New	Corr. Coeff. r
0.9	M1	-17.40 (2.02)	1.61 (2.03)	1.70 (2.04)	0.967
	M2	-17.50 (2.00)	1.41 (2.01)	0.81 (2.03)	0.974
	M3	-18.03 (2.03)	1.16 (2.05)	1.15 (2.04)	1.000
0.8	M1	-34.45 (2.01)	0.65 (2.03)	0.49 (2.05)	0.939
	M2	-32.89 (2.01)	2.49 (2.04)	0.19 (2.03)	0.947
	M3	-34.96 (2.01)	1.59 (2.03)	1.59 (2.03)	1.000
0.7	M1	-48.96 (2.01)	0.21 (1.99)	0.41 (2.04)	0.912
	M2	-44.76 (2.02)	5.31 (2.05)	0.76 (2.05)	0.920
	M3	-50.21(2.02)	1.53 (2.05)	1.52 (2.04)	1.000
0.6	M1	-59.80 (2.02)	1.58 (2.05)	1.27 (2.06)	0.892
	M2	-54.86 (2.03)	7.10 (2.07)	-0.75 (2.07)	0.899
	M3	-64.11 (2.00)	-0.35 (2.04)	-0.35 (2.01)	1.000
0.5	M1	-69.75 (1.99)	0.84 (2.03)	1.12 (2.03)	0.873
	M2	-59.90 (2.01)	15.07 (2.07)	2.27 (2.06)	0.872
	M3	-74.44 (1.97)	1.99 (2.00)	1.98 (2.00)	1.000

Table 2 supports our theory in the following ways.

1. As is well known, the naive variance estimator seriously underestimates the true variance. The adjusted jackknife variance estimator performs well for [M1] and [M3], but not for [M2]. The theory for the adjusted jackknife method assumes that hot deck imputations are done using the with-replacement selection which is not used in [M2]. As the response rate decreases in Table 2, the relative bias of the adjusted jackknife becomes larger.
2. The new method based on the pseudo data performs well even for the without-replacement imputation [M2]. As was discussed at the end of section 3, a single formula (29) can be used as the pseudo data for a large class of imputation methods.
3. As is observed in the correlation coefficients, the behaviors of the adjusted jackknife variance estimator and the proposed variance estimator are very similar for mean imputation [M3]. This is because the two variance estimators are asymptotically equivalent, as discussed in section 5.

7. Concluding Remarks

We have described methods of making pseudo data to be used for variance estimation. Generally speaking, the pseudo data can be described as

$$y_i^* = \begin{cases} \hat{y}_i & i = r+1, r+2, ..., n \\ \hat{y}_i + c_i g_i (y_i - \hat{y}_i) & i = 1, 2, ..., r, \end{cases} \tag{39}$$

where $\hat{y}_i$ is the predicted value of y_i under the model used for imputation. If $c_i g_i = 1$, then the variance estimator treats the imputed values as observations. A suitable choice of $c_i g_i > 1$ leads to a consistent variance estimator. If the imputation method is deterministic and the respondents are regarded as a random sample from the original sample, then $c_i \doteq r^{-1} n > 1$. For a two-phase sampling with a complex design, $c_i = w_i^{-1} w_i^*$, where w_i is the sampling weight of the unit i for the first-phase sample and w_i^* is the sampling weight of the unit i for the second-phase sample.

The g_i in (39) is the adjustement made to improve the conditional properties given the auxiliary variable x. For ratio imputation,

$$g_i = (\bar{x}_2)^{-1} \bar{x}_1$$

where $\bar{x}_2 = \sum_{i=1}^r w_i^* x_i$ and $\bar{x}_1 = \sum_{i=1}^n w_i x_i$. For regression imputation with scalar x,

$$g_i = 1 + (\bar{x}_1 - \bar{x}_2) \left\{ \sum_{k=1}^r w_k^* (x_k - \bar{x}_2)^2 \right\}^{-1} (x_i - \bar{x}_2).$$

In either case, we have

$$\sum_{i=1}^r w_i^* g_i x_i = \bar{x}_1.$$

While this paper was under review, Shao and Steel (1999) also provided similar methods in the case of deterministic imputation. Our method is more general in the sense that we also considered random imputation and introduced c_i term to improve finite sample properties.

Acknowledgements

The author thanks his thesis adviser Wayne A. Fuller for valuable discussions. The author also thanks Pamela Abbitt, F. Jay Breidt, Lou Rizzo, Richard Valliant, and the referees for helpful comments, which greatly improved the paper. Most of this work was done while the author was a graduate student at Iowa State University and was funded in part by cooperative agreement 68-3A75-43 between the USDA Natural Resources Conservation Service and Iowa State University and by Cooperative Agreement 43-3AEU-3-80088 between Iowa State University, the National Agricultural Statistics Service and the U.S. Bureau of Census.

Appendix

A. Proof of Equation (10) and (12)

The estimator $\hat{\mu}_y$ in (9) can be written as

$$\hat{\mu}_y = n^{-1} \sum_{i=1}^n \hat{y}_i + n^{-1} \sum_{i=1}^r (1 + d_i) \hat{e}_i \tag{A.1}$$

where d_i is the number of times that unit i is used as a donor. Under the equal probability and with-replacement imputation mechanism, we have

$$E_I(d_i) = r^{-1} m$$

and

$$\mathrm{Cov}_I(d_i, d_j) = \begin{cases} r^{-1} m (1 - r^{-1}) & \text{if } i = j \\ -r^{-2} m & \text{if } i \neq j \end{cases}$$

where the subscript I denotes the variation due to the imputation mechanism. It follows that $E_I(\hat{\mu}_y) = n^{-1} \sum_{i=1}^n \hat{y}_i$ and $V_I(\hat{\mu}_y) = n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2$. Hence,

$$V(\hat{\mu}_y) \doteq V\left(n^{-1} \sum_{i=1}^n \hat{y}_i \right) + E\left(n^{-2} r^{-1} m \sum_{i=1}^r \hat{e}_i^2 \right). \tag{A.2}$$

Now, by an similar argument similar to the one leading to (2), we have

$$\text{Var}\left(n^{-1}\sum_{i=1}^{n}\hat{y}_i \right) = [n^{-1}R^2 + r^{-1}(1-R^2)]\,\sigma_y^2. \qquad \text{(A.3)}$$

Since $\hat{y}_i - \bar{y}_I = (\mathbf{x}_i - \bar{\mathbf{x}}_i)\,\beta + o_p(1)$, we apply classical regression theory to get

$$E\left[(r-p)^{-1}\sum_{i=1}^{r}\hat{e}_i^2 \right] = (1-R^2)\,\sigma_y^2, \qquad \text{(A.4)}$$

and

$$E\left[(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 \right] = R^2\sigma_y^2. \qquad \text{(A.5)}$$

Therefore, (10) is proved and the estimator in (12) is consistent for the variance in (10).

B. Validity of (15) Under the Without-Replacement Imputation Mechanism

We assume that $m = kr + t$ where k and t are nonnegative integers and $t < r$. Let the estimator of the mean of y have the form (A.1). Let the imputation be performed such that t of the respondents are used $k+1$ times for imputation and $r-t$ units are used k times for imputation. The t of the respondents that are used $k+1$ times are chosen by simple random sampling without replacement. Then,

$$E_I(d_i) = k + r^{-1}t = r^{-1}m$$

and

$$\text{Cov}_I(d_i, d_j) = \begin{cases} r^{-1}t(1-r^{-1}t) & \text{if } i = j \\ -r^{-2}t & \text{if } i \neq j. \end{cases}$$

So, by similar arguments as in the proof of (A.2), we have

$$V(\hat{\mu}_y) \doteq V(\bar{y}_I) + E\left(n^{-2}r^{-1}t\sum_{i=1}^{r}\hat{e}_i^2 \right). \qquad \text{(B.1)}$$

Hence, using (A.3) and (A.4), we have

$$V\{\hat{\mu}_y\} = [n^{-1}R^2 + (r^{-1} + n^{-2}t)(1-R^2)]\,\sigma_y^2. \qquad \text{(B.2)}$$

Now, conditional on the realized sample and the respondents, we have

$$E_I\{(1+d_i)^2\} = \left(\frac{n}{r}\right)^2 + \frac{t}{r}\left(1 - \frac{t}{r}\right)$$

so that $\hat{V}\{\mu_y\}$ in (15) satisfies

$$E_I(\hat{V}\{\mu_y\}) \doteq n^{-1}(n-1)^{-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y}_I)^2 + [r^{-1} + n^{-2}t(1-r^{-1}t)](r-p)^{-1}\sum_{i=1}^{r}(y_i - \hat{y}_i)^2.$$

Therefore, using (A.4) and (A.5), we have the approximate unbiasedness of the $\hat{V}\{\mu_y\}$ under the without-replacement imputation mechanism.

C. Proof of Equation (26)

First, define $R_n^{(i)} = (\mathbf{x}_1^{(i)} - \bar{\mathbf{x}}_2^{(i)})(\hat{\beta} - \beta)$ and $R_n = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\hat{\beta} - \beta)$. From the equality (25),

$$\hat{V}^* = \sum_{i=1}^{L} c_i(\bar{y}_I^{*(i)} - \bar{y}_I)^2 = A_n + B_n + 2C_n$$

where $A_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})^2$, $B_n = \sum_{i=1}^{L} c_i(R_n^{(i)} - R_n)^2$, and $C_n = \sum_{i=1}^{L} c_i(\bar{y}_{II}^{*(i)} - \bar{y}_{II})(R_n^{(i)} - R_n)$. Hence, by the assumption (20), (26) follows because $A_n = O_p(n^{-1})$, $B_n = o_p(n^{-1})$, and $C_n = o_p(n^{-1})$. The last property comes from the Cauchy-Schwartz inequality, $C_n^2 \leq A_n B_n$.

References

Cochran, W.G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, Inc.

Fay, R.E. (1991). A design-based perspective on missing data variance. *Proceedings of the Bureau of the Census Annual Research conference*, 429-440.

Fay, R.E. (1992). When are inferences from multiple imputation valid? *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 227-232.

Fay, R.E. (1996). Alternative paradigms for the analysis of imputed survey data. *Journal of the American Statistical Association*, 91, 490-498.

Fuller, W.A. (1998). Replication variance estimation for two-phase samples. *Statistica Sinica*, 8, 1153-1164.

Hansen, M., Hurwitz, W.N. and Madows, W.G. (1953). *Sample Survey Methods and Theory*, Vol. I, New York: John Wiley & Sons, Inc.

Hansen, M., and Tepping, B.J. (1985). Estimation for Variance in NAEP. Unpublished memorandum, Westat, Washington, D.C.

Rao, J.N.K. (1996). On variance estimation with imputed survey data. *Journal of the American Statistical Association*, 91, 499-506.

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

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

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

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 473-489.

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

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

Särndal, C.-E., and Swensson, B. (1987). A general view of estimation for two phases of selection with applications to two-phase sampling and nonresponse. *International Statistical Review*, 55, 279-294.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Shao, J., Chen, Y. and Chen, Y. (1998). Balanced repeated replication for stratified multistage survey data under imputation. *Journal of the American Statistical Association*, 93, 819-831.

Shao, J., and Steel, P. (1999). Variance estimation for survey data with composite imputation and nonnegligible sampling fraction. *Journal of the American Statistical Association*, 94, 254-265.