JACKKNIFE VARIANCE ESTIMATION AFTER HOT DECK IMPUTATION

Jae-Kwang Kim and Wayne A. Fuller
Jae-Kwang Kim, Department of Statistics, Iowa State University, Ames, IA 50011

Key Words: Nonresponse, missing at random

1. Introduction

Item nonresponse occurs when a sampled unit cooperates in the survey but fails to respond to some of the items. To compensate for item nonresponse at the processing stage, various imputation procedures have been used in practice to fill in missing item values. Hot deck imputation is the imputation procedure in which the value assigned for a missing item is taken from respondents in the current sample.

Many of the hot deck imputation procedures start with a division of the sample into cells based on auxiliary variables known for both the respondents and nonrespondents. We will restrict our attention to the case where the imputed values are selected with a random mechanism from a set of donors in the same cell. This cell is called the *imputation cell.*

In spite of its convenience, treating the imputed values as if they are true values and making inference using standard formulas should be used with caution. The standard variance estimators, in particular, lead to underestimation because the additional variability due to missing values and imputation is not being taken into account. Rubin (1987) advocated multiple imputation to estimate the variance due to imputation. Multiple imputation is a model-based approach in that models are specified for the study variable, conditional on the realized sample and the respondents.

Rao and Shao (1992) proposed an adjusted jackknife variance estimator in the context of the model randomization approach, where the population values are treated as fixed and inferences are based on the sampling distribution generated by repetitions of the sample selection procedure and a model for response probabilities. In this paper, we propose a new replication variance estimator for hot deck imputation that can be used for a wide range of statistics and hot deck imputation methods.

2. Preliminaries

A population of N identifiable elements is denoted by $U = \{1, 2, ..., N\}$. A subset of the population is selected and called a sample. The selection of samples uses a set of probability rules called the *sampling mechanism.* Let A denote the set of indices for the elements in the sample.

Let Y_i denote the value for the i-th unit of some characteristic of interest. Let the population vector be

$$\mathbf{Y} = (Y_1, Y_2, \cdots, Y_N) .$$

Let the population quantity of interest be $\theta_N = \theta_N(\mathbf{Y})$ and let $\hat{\theta}$ be the estimator of θ_N based on the full sample.

An estimator of the variance of $\hat{\theta}$ is the replicate estimator

$$V_{JK} = \sum_{k=1}^{L} c_k \left(\hat{\theta}^{(k)} - \hat{\theta}\right)^2, \tag{1}$$

where $\hat{\theta}^{(k)}$ is the estimate of θ based on the observations included in the k-th replicate, L is the number of replicates, and c_k is a factor associated with replicate k determined by the replication method.

When the original estimator $\hat{\theta}$ is a linear estimator of the form

$$\hat{\theta} = \sum_{i \in A} w_i Y_i, \tag{2}$$

the k-th replicate of $\hat{\theta}$ can be written as

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} Y_i, \tag{3}$$

where $w_i^{(k)}$ denotes the replicate weight for the i-th unit of the k-th replication. Then,

$$\sum_{i \in A} w_i^{(k)} = \sum_{i \in A} w_i, \qquad k = 1, 2, \cdots, L. \tag{4}$$

Let us assume that the finite population U is made up of G imputation cells. Within each cell g, $g = 1, ..., G$, the elements are identically and independently distributed with mean μ_g and variance

σ_g^2 , i.e.

$$Y_i \overset{iid}{\sim} (\mu_g, \sigma_g^2), \qquad i \in U_g, \tag{5}$$

where U_g denotes the set of indices for the g^{th} imputation cell. We call the model (5) the *imputation cell model.*

The model-based approach in survey sampling makes inferences based on the conditional distribution of $\mathbf{Y}$ given the sample outcome A. This conditional distribution is determined by the sampling mechanism as well as by the distribution of the variable $\mathbf{Y}$. The dependence on the sampling mechanism can be avoided if the sampling mechanism is *ignorable*. Let the distribution of $\mathbf{Y}$ be denoted by $\mathcal{L}(\mathbf{Y})$ and call $\mathcal{L}(\mathbf{Y})$ the superpopulation model. Then, the sampling mechanism is ignorable under the superpopulation model if and only if

$$\mathcal{L}(\mathbf{Y} \mid A) = \mathcal{L}(\mathbf{Y}), \tag{6}$$

where $\mathcal{L}(\mathbf{Y} \mid A)$ is the conditional distribution of $\mathbf{Y}$ given the sample outcome A.

Under the existence of nonresponse, let A_R and A_M denote the set of indices of the respondents and nonrespondents, respectively. Define the response indicator function

$$R_i = \begin{cases} 1 & Y_i \text{ responds} \\ 0 & Y_i \text{ does not respond} \end{cases}, \qquad i \in A \tag{7}$$

and

$$\mathbf{R} = (R_i;\, i \in A).$$

The distribution of $\mathbf{R}$ is called the *response mechanism.* The response mechanism is usually unknown and is specified by a model. Conditional inference for $\mathbf{Y}$ given $\mathbf{R}$ requires the specification of the response mechanism.

Let $\mathcal{L}(\mathbf{Y} \mid A, A_R)$ be the conditional distribution of $\mathbf{Y}$ given the sample outcome A and the response outcome A_R. Then, the response mechanism is *ignorable* under the model if

$$\mathcal{L}(\mathbf{Y} \mid A, A_R) = \mathcal{L}(\mathbf{Y} \mid A). \tag{8}$$

If the sampling mechanism and the response mechanism are ignorable, then the imputation cell model still holds for the responding units as well as for nonrespondents. That is,

$$Y_i \mid (A, A_R) \overset{iid}{\sim} (\mu_g, \sigma_g^2), \qquad i \in U_g. \tag{9}$$

On the other hand, if one assumes the actual respondent observations satisfy (9), no other assumptions are necessary. Rubin (1976) and Scott and Smith (1977) discuss ignorability.

The hot deck imputation method that we consider is based on imputation cells. The hot deck imputation method assigns the value from a record with a response to the record with a missing value on that item in the same cell. The record with the response will be called the *donor* and the record with the missing value is the *recipient.* Often, the values for a vector of missing items are taken from the same donor.

Given the values of the respondents, the properties of the augmented sample are determined by the the choices of which donors go with which recipients. Define

$$d_{ij} = \begin{cases} 1 & \text{if } Y_i \text{ is used as donor for } Y_j \\ 0 & \text{otherwise} \end{cases}, \tag{10}$$

for $j \in A_M$ and

$$\mathbf{d} = (d_{ij};\, i \in A_R, j \in A_M). \tag{11}$$

Then the distribution of $\mathbf{d}$ is called the *imputation mechanism.*

The following class of imputation mechanisms is of particular interest.

(I.1) For any missing $j \in A_M$ and any responding $i \in A_R$,

$$\Pr(d_{ij} = 1 \mid \mathbf{Y}) = \Pr(d_{ij} = 1).$$

(I.2) If the unit $i \in A_r$ and the unit $j \in A_M$ belong to different imputation cells,

$$\Pr(d_{ij} = 1) = 0.$$

(I.3) If the unit $i \in A_R$ and the unit $j \in A_M$ belong to the same imputation cell,

$$0 < \Pr(d_{ij} = 1) < 1.$$

Assumptions (I.1) to (I.3) are sufficient conditions for the distributions of the observations after imputation to be the same as those of the observations before imputation. Hence, given the imputation cell model (9) and the three assumptions, we have

$$Y_i \mid (A, A_R, \mathbf{d}) \overset{iid}{\sim} (\mu_g, \sigma_g^2), \qquad i \in U_g. \tag{12}$$

3. Variance Estimation after Imputation

We consider a *pairwise imputation method,* where two distinct donors are selected for each missing

item. This is a special case of fractional imputation, proposed by Kalton and Kish (1984). We assume there are at least two donors in each imputation cell. The following assumption describes the pairwise imputation method.

(I.4) For each missing $j \in A_M$,

$$\sum_{i \in A_R} d_{ij} = 2,$$

where A_R is the set of respondents, A_M is the set of nonrespondents, and d_{ij} is the imputation indicator defined in (10).

When the original estimator $\hat{\theta}$ is a linear estimator of the form in (2), the linear estimator based on the augmented sample can be written as

$$\widehat{\theta}_I = \sum_{i \in A} a_i Y_i, \tag{13}$$

where

$$a_i = R_i \left(w_i + \sum_{j \in A_M} 0.5 d_{ij} w_j \right) \tag{14}$$

is the sum of the weights of the items that are imputed from unit i. If Y_i is missing, then $a_i = 0$. Notice that, under (I.2) and (I.4),

$$\sum_{i \in A \cap U_g} a_i = \sum_{i \in A \cap U_g} w_i, \qquad g = 1, 2, \cdots, G \tag{15}$$

because, by (14),

$$\begin{aligned} \sum_{i \in A \cap U_g} a_i &= \sum_{i \in A_R \cap U_g} w_i + \sum_{i \in A_R \cap U_g} \sum_{j \in A_M} 0.5 d_{ij} w_j \\ &= \sum_{i \in A_R \cap U_g} w_i + \sum_{j \in A_M \cap U_g} w_j \\ &= \sum_{i \in A \cap U_g} w_i \, . \end{aligned}$$

For simplicity of notation, let us define

$$A_g = A \cap U_g, \qquad g = 1, 2, \cdots, G.$$

In theorem 1, we establish properties of the estimator for the population total under the imputation cell model. The variance of the estimator is a function of the expectation of a_i^2. This expectation is a function of the procedure used to select donors. For example, with an equal probability design, the use of a procedure that produces nearly equal a_i will minimize variance.

Theorem 1 *Let the superpopulation model be (5). Assume the sampling mechanism and the response mechanism are ignorable, and that the imputation mechanism satisfies (I.1)- (I.4). Let $\hat{\theta}$ be a linear estimator of the form (2) constructed from the full sample that is design unbiased for the population quantity θ_N. Then, the linear estimator $\hat{\theta}_I$ based on the imputed sample satisfies*

$$E\left(\widehat{\theta}_I - \theta_N\right) = 0, \tag{16}$$

$$\begin{aligned} Var\left(\hat{\theta}_I\right) &= Var\left(\sum_{g=1}^{G} \sum_{i \in A_g} w_i \mu_g \right) \\ &\quad + E\left(\sum_{g=1}^{G} \sum_{i \in A_g} a_i^2 \sigma_g^2 \right), \end{aligned} \tag{17}$$

and, if $\theta_N = \sum_{i=1}^{N} Y_i$, then

$$\begin{aligned} Var\left(\hat{\theta}_I - \theta_N\right) &= Var\left(\sum_{g=1}^{G} \sum_{i \in A_g} w_i \mu_g \right) \\ &\quad + E\left\{ \sum_{g=1}^{G} \sum_{i \in A_g} \left(a_i^2 - a_i\right) \sigma_g^2 \right\} \end{aligned} \tag{18}$$

where the a_i are defined in (14), A is the set of sample indices defined in Section 2, A_R is the set of respondent indices defined in Section 2, G is the number of imputation cells, and A_g is the set of indices for the g^{th} imputation cell in the sample.

Proof. Let the linear estimator for the full sample be as in (2) and let $\hat{\theta}_I$ be the imputed estimator. For any measurable set B_1 and B_2 in the sigma-field $\sigma(Y)$ generated by the random variable Y, we have

$$\begin{aligned} &\Pr\left(Y_i \in B_1, Y_j \in B_2 \mid d_{ij} = 1\right) \\ &= \Pr\left(Y_i \in B_1, Y_j \in B_2\right) \times \frac{\Pr\left(d_{ij} = 1 \mid Y_i, Y_j\right)}{\Pr\left(d_{ij} = 1\right)}. \end{aligned}$$

So, by (I.1) and (I.3),

$$\Pr\left(Y_i \in B_1, Y_j \in B_2 \mid d_{ij} = 1\right) = \Pr\left(Y_i \in B_1, Y_j \in B_2\right). \tag{19}$$

Similarly,

$$\Pr\left(Y_i \in B_1, Y_j \in B_2 \mid d_{ij} = 0\right) = \Pr\left(Y_i \in B_1, Y_j \in B_2\right). \tag{20}$$

Hence, from (19) and (20),

$$\Pr\left(Y_i \in B_1, Y_j \in B_2 \mid d_{ij}\right) = \Pr\left(Y_i \in B_1, Y_j \in B_2\right). \tag{21}$$

To show the mean part (16), by (21),

$$E\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A_g} a_i E\left(Y_i \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A_g} a_i E\left(Y_i \mid A, A_R\right).$$

Under model (5), the ignorable sampling mechanism, and the ignorable response mechanism, we have

$$E\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A_g} a_i\mu_g = \sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g, \quad (22)$$

where the last equality comes from (15). Thus, by the design unbiasedness of $\hat{\theta}$,

$$E\left(\hat{\theta}_I\right) = E\left\{E\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right)\right\} = E\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right).$$

So, (16) is proved because $\sum_{g=1}^{G}\sum_{i\in A\cap U_g} w_i\mu_g$ is design unbiased for $E\left(\theta_N\right)$.

For the conditional variance of $\hat{\theta}_I$, by (13),

$$Var\left\{\hat{\theta}_I \mid A, A_R, \mathbf{d}\right\} = \sum_{i\in A}\sum_{j\in A} a_i a_j Cov\left(Y_i, Y_j \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A_g} a_i^2\sigma_g^2, \quad (23)$$

where the last equality comes from (21).

For the total variance of $\hat{\theta}_I$, note that

$$Var\left(\hat{\theta}_I\right) = Var\left\{E\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right)\right\} + E\left\{Var\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right)\right\}. \quad (24)$$

Inserting (16) and (23) into (24), the result (17) follows. Now,

$$Var\left(\hat{\theta}_I - \theta_N \mid A, A_R, \mathbf{d}\right) = Var\left(\hat{\theta}_I \mid A, A_R, \mathbf{d}\right) - 2Cov\left(\hat{\theta}_I, \theta_N \mid A, A_R, \mathbf{d}\right) + Var\left(\theta_N \mid A, A_R, \mathbf{d}\right). \quad (25)$$

Note that, for $\theta_N = \sum_{i=1}^{N} Y_i$,

$$Cov\left(\hat{\theta}_I, \theta_N \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A\cap U_g} a_i\sigma_g^2$$

and

$$Var\left(\theta_N \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in U_g}\sigma_g^2.$$

So, from (25),

$$Var\left(\hat{\theta}_I - \theta_N \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A\cap U_g}\left(a_i^2 - 2a_i\right)\sigma_g^2 + \sum_{g=1}^{G}\sum_{i\in U_g}\sigma_g^2$$

Note that, by (15),

$$E\left(\sum_{g=1}^{G}\sum_{i\in A_g} a_i\sigma_g^2\right) = E\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\sigma_g^2\right) = \sum_{g=1}^{G}\sum_{i\in U_g}\sigma_g^2. \quad (26)$$

So,

$$E\left\{\sum_{g=1}^{G}\sum_{i\in A\cap U_g}\left(a_i^2 - 2a_i\right)\sigma_g^2 + \sum_{g=1}^{G}\sum_{i\in U_g}\sigma_g^2\right\} = E\left\{\sum_{g=1}^{G}\sum_{i\in A\cap U_g}\left(a_i^2 - a_i\right)\sigma_g^2\right\}.$$

Therefore, using the decomposition (24) applied to $\hat{\theta}_I - \theta_N$, the result (18) follows. ■

If we treat the imputed values as if they are true values and apply the standard replication variance estimator V_{JK} in (1), then the naive variance estimator can be expressed as

$$V_{JK}^I = \sum_{k=1}^{L} c_k\left(\hat{\theta}_I^{(k)} - \hat{\theta}_I\right)^2 \quad (27)$$

where

$$\hat{\theta}_I^{(k)} = \sum_{i\in A} a_i^{(k)} Y_i \quad (28)$$

with

$$a_i^{(k)} = R_i\left(w_i^{(k)} + \sum_{j\in A_M} 0.5 d_{ij} w_j^{(k)}\right) \quad (29)$$

and $\hat{\theta}_I$ is defined in (13). The expectation of the naive variance estimator is given in the following theorem.

Theorem 2 *Let the assumptions of Theorem 1 hold. Assume the jackknife variance estimator V_{JK} for the full sample is design unbiased for the variance of $\hat{\theta}$. Then, the naive jackknife variance estimator V_{JK}^I applied to the augmented set satisfies*

$$E\left(V_{JK}^I\right) = Var\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) + \sum_{k=1}^{L} E\left\{\sum_{g=1}^{G}\sum_{i\in A_g} c_k\left(a_i^{(k)}-a_i\right)^2\sigma_g^2\right\}. \quad (30)$$

where a_i is defined in (14) and and $a_i^{(k)}$ is the k-th replication version of a_i defined in (29).

Proof. We write

$$\begin{aligned} & E\left\{V_{JK}^I\right\} \\ = \; & E\left\{E\left(V_{JK}^I \mid A, A_R, \mathbf{d}\right)\right\} \\ = \; & E\left\{E\left[\sum_{k=1}^{L} c_k\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I\right)^2 \mid A, A_R, \mathbf{d}\right]\right\}. \end{aligned} \quad (31)$$

Observe that

$$\begin{aligned} & E\left[\sum_{k=1}^{L} c_k\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I\right)^2 \mid A, A_R, \mathbf{d}\right] \\ = \; & \sum_{k=1}^{L} c_k\left[E\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I \mid A, A_R, \mathbf{d}\right)\right]^2 \\ & + \sum_{k=1}^{L} c_k Var\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I \mid A, A_R, \mathbf{d}\right). \end{aligned} \quad (32)$$

Under the ignorability of the sampling and the response mechanism and the assumption (I.1)-(I.4) of the imputation mechanism,

$$E\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{j\in A_g}\left(w_j^{(k)}-w_j\right)\mu_g$$

and

$$Var\left(\hat{\theta}_I^{(k)}-\hat{\theta}_I \mid A, A_R, \mathbf{d}\right) = \sum_{g=1}^{G}\sum_{i\in A_g}\left(a_i^{(k)}-a_i\right)^2\sigma_g^2.$$

Now, define

$$\ddot{\theta} = \sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g$$

the estimator of the same functional form as $\hat{\theta}$ with the means μ_g replacing the response variable Y_i, $i \in U_g$. The estimators $\hat{\theta}$ and $\ddot{\theta}$ are identical whenever $\sigma_g^2 \equiv 0$. The jackknife variance estimator applied to $\ddot{\theta}$ can be written as

$$\sum_{k=1}^{L} c_k\left(\ddot{\theta}^{(k)}-\ddot{\theta}\right)^2 = \sum_{k=1}^{L} c_k\left[\sum_{g=1}^{G}\sum_{j\in A_g}\left(w_j^{(k)}-w_j\right)\mu_g\right]^2$$

where $\ddot{\theta}^{(k)}$ is the k-th replicate of $\ddot{\theta}$. By the design unbiasedness of the jackknife variance estimator applied to $\ddot{\theta}$, we have

$$\begin{aligned} & E\left\{\sum_{k=1}^{L} c_k\left[\sum_{g=1}^{G}\sum_{j\in A_g}\left(w_j^{(k)}-w_j\right)\mu_g\right]^2\right\} \\ & = Var\left(\sum_{g=1}^{G}\sum_{i\in A_g} w_i\mu_g\right) \end{aligned} \quad (33)$$

where the expectation and the variance are calculated with respect to the sampling mechanism. Therefore, by (31), (32), and (33), we have (30). ■

By Theorem 1 and Theorem 2, we can calculate the bias of the naive variance estimator under the imputation cell model. The bias of V_{JK}^I is

$$Bias = E\left\{\sum_{g=1}^{G}\sum_{i\in A_g}\left[\sum_{k=1}^{L} c_k\left(a_i^{(k)}-a_i\right)^2 - a_i^2\right]\sigma_g^2\right\}. \quad (34)$$

To adjust for the bias, we must estimate σ_g^2. For each missing item $j \in A_M$, we have two distinct imputed values with each having weight $0.5w_j$. We treat each pair as a pseudo stratum and apply the two-per-stratum jackknife method to the pseudo strata in cell g to estimate σ_g^2.

To illustrate this, assume Y_s and Y_t are two distinct donors for the missing item Y_j. If the Y_s in pseudo stratum $j \in A_M$ is deleted, then the jackknife replicate for $\hat{\theta}_I$ can be written as

$$\hat{\theta}_I^{(-js)} = \sum_{i\in A} a_i Y_i + 0.5w_j Y_t - 0.5w_j Y_s \quad (35)$$

where $\hat{\theta}_I^{(-js)}$ is the jackknife replicate of $\hat{\theta}_I$ when the element s in the pseudo stratum $j \in A_M$ is deleted. So, for $j \in A_M \cap U_g$,

$$E\left\{\left(\hat{\theta}_I^{(-js)}-\hat{\theta}_I\right)^2 \mid A, A_R, \mathbf{d}\right\} = 0.5w_j^2\sigma_g^2. \quad (36)$$

The fact that the bias term (34) is also a linear function of σ_g^2s suggests a bias-corrected variance estimator.

The suggested variance estimator is of the form

$$V_{c1} = V_{JK}^I + \sum_{j \in A_M} \sum_{i \in A_R} d_{ij} q_{ji} \left(\hat{\theta}_I^{(-ji)} - \hat{\theta}_I \right)^2 \quad (37)$$

where d_{ij} is the imputation indicator function defined in (10) and the replication factors q_{ji} are to be determined. Note that

$$E \left\{ \sum_{j \in A_M} \sum_{i \in A_R} d_{ij} q_{ji} \left(\hat{\theta}_I^{(-ji)} - \hat{\theta}_I \right)^2 \mid A, A_R, \mathbf{d} \right\}$$
$$= \sum_{g=1}^{G} \sum_{j \in A_M \cap U_g} \sum_{i \in A_R} 0.5 d_{ij} q_{ji} w_j^2 \sigma_g^2.$$

If we want to estimate the variance of θ_I, then the determining equation for q_{ji} with $d_{ij} = 1$ is

$$\sum_{k=1}^{L} c_k \left(a_i^{(k)} - a_i \right)^2 + 0.5 \sum_{j \in A_M} d_{ij} q_{ji} w_j^2 = a_i^2.$$

A solution is

$$q_{ji} = \left(0.5 \sum_{j \in A_M} d_{ij} w_j^2 \right)^{-1} \left\{ a_i^2 - \sum_{k=1}^{L} c_k \left(a_i^{(k)} - a_i \right)^2 \right\}. \quad (38)$$

If we want to estimate the variance of $\theta_I - \theta_N$ with $\theta_N = \sum_{i=1}^{N} Y_i$, then a solution is

$$q_{ji} = \left(0.5 \sum_{j \in A_M} d_{ij} w_j^2 \right)^{-1} \times \left\{ a_i^2 - a_i - m \sum_{k=1}^{L} c_k \left(a_i^{(k)} - a_i \right)^2 \right\}.$$

Several simplifications of the suggested variance estimator in (37) are possible. Since the two squared jackknife deviates in a pseudo stratum have the same expectations, deleting only one element per each pseudo stratum reduces the number of replications. Then, the variance estimator can be written as

$$V_{c2} = V_{JK}^I + \sum_{j \in A_M} q_j \left(\hat{\theta}_I^{(j)} - \hat{\theta}_I \right)^2 \quad (39)$$

where $q_j = q_{js} + q_{jt}$ with q_{js} and q_{jt} are calculated from (38) when Y_s and Y_t are two distinct donors for the missing item Y_j. In this case, $\hat{\theta}_I^{(j)} = \hat{\theta}_I^{(-js)}$ defined in (35).

4. Comments

The model assumptions are those commonly made for the imputation cell model. Nontheless they are relatively strong assumptions. They are used in the proof of mean unbiasedness and in an even stronger way in the proof of variance unbiasedness.

Modest modification of an existing single imputation program is required to implement the fractional imputation procedure. In practice, we would employ a scheme such that donors are used approximately an equal number of times and such that a donor is never used twice for the same recipient. Given the data set, all estimation is conducted with the single data set. Under the model, all functions of the Y-variable, including the distribution function, are consistently estimated. Once the weights q_j are determined, variance calculation can be carried out with a program such as WesVarPC (1996). No additional programming is required.

The variance estimation procedure is relatively efficient because the degrees of freedom for the estimation of the imputation variance is equal to the number of recipients. Of course, one can reduce the number of replicates if desired.

REFERENCES

Kalton, G. and Kish, L. (1984). "Some efficient random imputation methods." *Communications in Statistics, Part A - Theory and Methods*, 13, 1919-1939.

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

Rubin, D. B. (1976). "Inference and missing data" *Biometrika*, 63, 581-590.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York : Wiley.

Scott, A. and Smith, T. M. F. (1973). "Survey design, symmetry and posterior distributions" *Journal of the Royal Statistical Society, Series B, Methodological*, 35, 57-60.

WesVarPC (1996). "A user's guide to WesVarPC", Version 2.0. Rockville, MD: Westat Inc.