FREQUENCY VALID MULTIPLE IMPUTATION FOR SURVEYS WITH A COMPLEX DESIGN

David A. Binder and Weimin Sun, Statistics Canada
Business Survey Methods Division, Statistics Canada, Ottawa, ON, Canada K1A 0T6

Key Words: Nonresponse; Missing data; Repeated imputation; Proper imputation; Stratified sampling; Multistage sampling; Variance estimation.

Abstract: General conditions required for valid design-based inference when using multiple imputation for dealing with missing values are considered. We focus on the means or totals and the estimation of their variances. We study multiple and proper imputation under a general setting, concentrating on the mathematical and statistical conditions required for valid design-based inference, assuming the nonresponse mechanism is an additional phase of sampling.

1. INTRODUCTION

In virtually all surveys, there is nonresponse to at least some of the information requested. A common method to compute the survey estimates when nonresponse is present is to impute for the missing values; that is, to derive valid values for the missing entries, and compute the estimates as if the data were complete in the first place. With a judicious choice of imputation method, this technique will often lead to acceptable point estimates, but if the fact that some data are imputed is ignored in the calculation of estimates of variance, these estimates can be biased downwards. This concern can be addressed in a variety of ways, some more complicated than others. One of the suggestions that is gaining increasing popularity is multiple imputation. The basics of this method are described in Rubin (1987).

Rubin (1987) defines *multiple imputation* as the creation (conceptually, at least) of a set of different data files, each with complete data derived from some imputation method. Typically, each data set is stochastically generated from the same stochastic model, although, in general, this is not necessary, as different models may be used for different data sets.

Proper imputation is defined as multiple imputation with certain frequency-based properties so that the sum of the between and within sums of squares can be used as an estimate of variance for valid frequency-based inferences, such as confidence intervals. For the frequency-based properties, we use the randomization distribution derived from the sample design (design-based) and the random mechanism that generates nonresponse, which is treated as an additional phase of sampling. We also include any randomness due to the imputations themselves being stochastic.

Some authors have expressed the concern that it may be difficult to find a multiple imputation scheme satisfying the requirements of proper imputation when the survey design is complex; for example, designs that are stratified and multistage. For a discussion of this point, see Fay (1996) and Rubin (1996) and discussions and rejoinders by Binder, Eltinge, Judkins, Rao, Fay and Rubin of these articles. The purpose of this paper is to study this problem in more depth.

The approach we take is as follows. We first consider methods of complete data analysis, where the estimators of the total are linear and the estimators of the variance are quadratic in the data. The reason why that we restrict ourselves to estimators of the total rather than more complex statistics, is that it is only when these estimates are frequency-valid that one would wish to pursue the more complex case. We then consider stochastic imputation schemes using means and ratios as the basis for the imputations. We discuss the properties of these schemes that are needed for the analysis of multiple imputation derived by generating different stochastically imputed data sets. In particular, we consider the properties that must be satisfied for the imputations to be proper.

2. COMPLETE DATA ANALYSIS

We first consider the situation where there is no nonresponse. In this section we give the expressions for the estimator, its mean square error (or variance) and the estimator of this mean square error.

In the case of simple random sampling without replacement (srswor), our usual estimator of the total is given by

$$\hat{Y} = (N/n)\mathbf{1}_n'\boldsymbol{Z}\boldsymbol{y}, \qquad (1)$$

where $\boldsymbol{Z}$ is an $n \times N$ matrix, such that z_{ij} is 1 if the i th unit selected in the sample is the j th unit of the population and is 0 otherwise; $\boldsymbol{y}$ is the $N \times 1$ vector of population values. In general, letting s be the set of units in the sample, we consider estimators of the form

$$\hat{Y} = \boldsymbol{d}(s)'\boldsymbol{y} = \boldsymbol{d}(s)'\boldsymbol{Z}'\boldsymbol{Z}\boldsymbol{y} \qquad (2)$$

where $\boldsymbol{d}(s)=(d_1(s),...,d_N(s))'$ is a random vector, $d_i(s)=0$ for $i \notin s$. For example, for Horvitz-Thompson (H-T) estimators, $\boldsymbol{d}(s)=\boldsymbol{D}_\pi^{-1}\boldsymbol{Z}'\mathbf{1}_n$ where $\boldsymbol{D}_\pi$ is a diagonal matrix of π_i's, the first order inclusion probabilities.

Under srswor, using the sample design as our randomization distribution, we have that

$$E[\hat{Y}]=\mathbf{1}_N'\boldsymbol{y}=Y, \tag{3}$$

since $E[\boldsymbol{Z}] = N^{-1}\mathbf{1}_n\mathbf{1}_N'$. In general, for estimators given by (2),

$$E[\hat{Y}]=E[\boldsymbol{d}(s)]'\boldsymbol{y}, \tag{4}$$

so that the condition for (2) to be asymptotic design unbiased (ADU) is that $E[\boldsymbol{d}(s)]\to\mathbf{1}_N$ as $n\to\infty$. In the case of Horvitz-Thompson estimators, $E[\boldsymbol{d}(s)]=\boldsymbol{D}_\pi^{-1}E[\boldsymbol{Z}'\mathbf{1}_n]=\boldsymbol{D}_\pi^{-1}\pi=\mathbf{1}_N$, where π is the vector of first order inclusion probabilities.

Under srswor, we have that

$$V[\hat{Y}]=\frac{N(N-n)}{n(N-1)}\boldsymbol{y}'\left(\boldsymbol{I}_N-N^{-1}\mathbf{1}_N\mathbf{1}_N'\right)\boldsymbol{y}=N(N-n)S^2/n \tag{5}$$

where $\boldsymbol{I}_N$ is the $N\times N$ identity matrix. In general, for estimators given by (2), we have

$$MSE[\hat{Y}]=\boldsymbol{y}'\Delta\boldsymbol{y}, \tag{6}$$

where $\Delta=E[(\boldsymbol{d}(s)-\mathbf{1}_N)(\boldsymbol{d}(s)-\mathbf{1}_N)']$. In the case of H-T estimators, since $E[\boldsymbol{Z}'\mathbf{1}_n\mathbf{1}_n'\boldsymbol{Z}]=\Pi^{(2)}$, where $\Pi^{(2)}$ is the $N\times N$ matrix of second order inclusion probabilities, we have $\Delta=\boldsymbol{D}_\pi^{-1}\Pi^{(2)}\boldsymbol{D}_\pi^{-1}-\mathbf{1}_N\mathbf{1}_N'$.

To estimate expression (5), the variance of the estimator under srswor, the usual estimator is

$$\begin{aligned} v[\hat{Y}] &= \frac{N(N-n)}{n(n-1)}\boldsymbol{y}'\boldsymbol{Z}'\left(\boldsymbol{I}_n-n^{-1}\mathbf{1}_n\mathbf{1}_n'\right)\boldsymbol{Z}\boldsymbol{y} \\ &= N(N-n)s^2/n, \end{aligned} \tag{7}$$

which is an unbiased estimate of (5). In the general case, we assume that the estimator of (6) is given by

$$mse[\hat{Y}]=\boldsymbol{y}'\boldsymbol{Z}'\hat{\Delta}(s)\boldsymbol{Z}\boldsymbol{y}, \tag{8}$$

where $\hat{\Delta}(s)$ is an $n\times n$ matrix. Normally, for asymptotic design consistency (ADC), $E[\boldsymbol{Z}'\hat{\Delta}(s)\boldsymbol{Z}]\to\Delta$ as $n\to\infty$.

3. STOCHASTIC IMPUTATION

We now consider the case where some data are missing. We denote by $\boldsymbol{Z}_1$ the $n_1\times n$ matrix, where z_{1ij} is 1 if the i th respondent is the j th unit of the original sample and is 0 otherwise. In general, we use the subscript 1 to denote that the expression is computed for respondents and subscript 2 for nonrespondents.

3.1 Simple Random Sampling - Mean Imputation with Random Noise

We first consider srswor. To impute the missing values, we use an imputation, denoted by the vector, $\boldsymbol{y}^*$, generated stochastically with common mean,

$$\bar{y}_1^* = \hat{Y}_1^*/N = n_1^{-1}\mathbf{1}_{n_1}'\boldsymbol{Z}_1\boldsymbol{Z}\boldsymbol{y}. \tag{9}$$

We consider the situation where the variance matrix for the imputations, $\boldsymbol{y}^*$, is

$$\Sigma_I=\sigma_I^2[(1-\rho_I)\boldsymbol{I}_{n_2}+\rho_I\mathbf{1}_{n_2}\mathbf{1}_{n_2}']; \tag{10}$$

that is, the imputation variance is σ_I^2 and the imputations have correlation ρ_I. In general, σ_I^2 and ρ_I may depend on the observed responses, $\boldsymbol{Z}_1\boldsymbol{Z}\boldsymbol{y}$. Imputations with mean and variance given by (9) and (10), respectively, are common in practice when there are no auxiliary data. For example, for deterministic mean imputation, we have $\sigma_I^2=0$. For hot deck imputation, where we select a respondent at random with replacement, $\sigma_I^2=(n_1-1)s_1^2/n_1$ and $\rho_I=0$. In the case of the Bayesian bootstrap - see, for example, Rubin (1987; p. 43) - we have $\sigma_I^2=(n_1-1)s_1^2/n_1$ and $\rho_I=(n_1+1)^{-1}$.

For the case where some values are missing, we have that the imputed estimate for the population total would be given by

$$\hat{Y}_I = (N/n)[\mathbf{1}_{n_1}'\boldsymbol{Z}_1\boldsymbol{Z}\boldsymbol{y}+\mathbf{1}_{n_2}'\boldsymbol{y}^*]. \tag{11}$$

To study the properties of (11), we need to make some assumptions about the response mechanism; that is, the mechanism that generates $\boldsymbol{Z}_1$. Throughout this section we perform the analysis conditional on n_1. For srswor, we assume that the response mechanism is exchangeable among sampled units. We first consider the conditional expectation and variance of $\hat{Y}_I$ over the stochastic imputation, given the pattern of nonresponse. This is given by

$$E[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] = N\bar{y}_1 = (N/n_1)\mathbf{1}_{n_1}'\boldsymbol{Z}_1\boldsymbol{Z}\boldsymbol{y}, \quad \text{and} \tag{12}$$

$$\begin{aligned} V[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] &= (N/n)^2\mathbf{1}_{n_2}'\Sigma_I\mathbf{1}_{n_2} \\ &= (N/n)^2 n_2\sigma_I^2[1+(n_2-1)\rho_I]. \end{aligned} \tag{13}$$

One of the conditions required for proper imputation is that

$$E[\hat{Y}_I | \boldsymbol{Z}] \approx \hat{Y}, \tag{14}$$

for large n and n_1, which can be seen to be satisfied here. A second condition for the imputation to be proper is

$$E_{\boldsymbol{Z}_1} V[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] \approx V_{\boldsymbol{Z}_1} E[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] \tag{15}$$

for large n. Exchangeability of the response mechanism implies that the second-order inclusion probabilities for response are like srswor of n_1 units out of n units. Therefore, condition (15) implies that

$$\frac{N^2 n_2}{n^2} E[\sigma_I^2\{1+(n_2-1)\rho_I\} | \boldsymbol{Z}] \approx \frac{N^2 n_2}{n n_1} s^2, \tag{16}$$

or, equivalently

$$E[\sigma_I^2\{1+(n_2-1)\rho_I\} | \boldsymbol{Z}] \approx n s^2 / n_1. \tag{17}$$

We now consider the complete data estimator of the variance, expression (7) above, applied to a singly imputed data set. We denote this by $v_I[\hat{Y}]$, given by

$$v_I[\hat{Y}] = \frac{N(N-n)}{n(n-1)} [\boldsymbol{y}' \boldsymbol{Z}' \boldsymbol{Z}_1', \; y^{*\prime}] \left(\boldsymbol{I}_n - n^{-1} \boldsymbol{1}_n \boldsymbol{1}_n'\right) \begin{bmatrix} \boldsymbol{Z}_1 \boldsymbol{Z} \boldsymbol{y} \\ y^* \end{bmatrix}. \tag{18}$$

A third condition for the imputation to be proper is that

$$E\left\{v_I[\hat{Y}] | \boldsymbol{Z}\right\} \approx v[\hat{Y}], \tag{19}$$

where $v[\hat{Y}]$ is given in expression (7). Expression (19) implies that

$$\begin{aligned} &\frac{N(N-n)n_2}{n^2(n-1)} E\left[\sigma_I^2\{n-1-(n_2-1)\rho_I\} | \boldsymbol{Z}\right] \\ &+ \frac{N(N-n)(n_1-1)}{n(n-1)} s^2 \approx \frac{N(N-n)}{n} s^2, \end{aligned} \tag{20}$$

or, equivalently

$$E\left[\sigma_I^2\{n-1-(n_2-1)\rho_I\} | \boldsymbol{Z}\right] \approx n s^2. \tag{21}$$

Combining (17) and (21), we have that for the imputation to be proper for large samples, it is necessary that

$$\sigma_I^2 \approx [(n_1+1)/n_1] s^2 \approx s^2 \tag{22}$$

and

$$\rho_I \approx 1/(n_1+1) \approx n_1^{-1}. \tag{23}$$

For example, the Bayesian bootstrap satisfies this property, but selecting a respondent at random with replacement or using deterministic mean imputation does not.

Conditions (14), (15) and (19) are necessary for imputations to be proper for the more complex cases as well. In the following sections, we consider their implications.

3.2 Simple Random Sampling - Ratio Imputation with Random Noise

More generally, for example when auxiliary information is available, other imputation methods are commonly used. To keep the derivations relatively simple, we consider ratio imputations given by

$$\boldsymbol{y}^* = \hat{R}_1 \boldsymbol{Z}_2 \boldsymbol{Z} \boldsymbol{x} + \epsilon^*, \tag{24}$$

where

$$\hat{R}_1 = \boldsymbol{1}_{n_1}' \boldsymbol{Z}_1 \boldsymbol{Z} \boldsymbol{y} / \boldsymbol{1}_{n_1}' \boldsymbol{Z}_1 \boldsymbol{Z} \boldsymbol{x} \tag{25}$$

and ϵ^* is a vector of random errors with mean $\boldsymbol{0}$ and variance matrix Σ_I, and $\boldsymbol{Z}_2$ is an $n_2 \times n$ matrix indicating the nonrespondents among the sampled units. For example, an extension of the Bayesian bootstrap for the ratio imputation case would be to let

$$\boldsymbol{y}^* = [\hat{R}_1 + \epsilon_{R_1}^*] \boldsymbol{Z}_2 \boldsymbol{Z} \boldsymbol{x} + \epsilon = \hat{R}_1 \boldsymbol{Z}_2 \boldsymbol{Z} \boldsymbol{x} + (\boldsymbol{Z}_2 \boldsymbol{Z} \boldsymbol{x} \epsilon_{R_1}^* + \epsilon), \tag{26}$$

where $\epsilon_{R_1}^*$ has mean 0 and variance $v[\hat{R}_1 | \boldsymbol{Z}]$, and ϵ has mean $\boldsymbol{0}$ and variance $\sigma^2 \boldsymbol{Z}_2 \boldsymbol{Z} \boldsymbol{D}_x \boldsymbol{Z}' \boldsymbol{Z}_2'$, where $\boldsymbol{D}_x$ is a diagonal matrix of the population x values.

The imputed estimate for the population total is given by

$$\hat{Y}_I = (N/n)[\boldsymbol{1}_{n_1}' \boldsymbol{Z}_1 \boldsymbol{Z} \boldsymbol{y} + \boldsymbol{1}_{n_2}' \boldsymbol{y}^*] = \hat{R}_1 \hat{X} + (N/n) \boldsymbol{1}_{n_2}' \epsilon^*. \tag{27}$$

Considering the conditional expectation and variance of $\hat{Y}_I$ given the pattern of nonresponse, we have

$$E[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] = \hat{R}_1 \hat{X}, \tag{28}$$

and

$$V[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] = (N/n)^2 \boldsymbol{1}_{n_2}' \Sigma_I \boldsymbol{1}_{n_2}. \tag{29}$$

We see that, since $\hat{R}_1 \to \hat{Y}/\hat{X}$, we have that condition (14) is satisfied, asymptotically.

Motivated by (26), we consider the particular case where $\Sigma_I = \boldsymbol{Z}_2\boldsymbol{Z}(\boldsymbol{D}_c + b\boldsymbol{x}\boldsymbol{x}')\boldsymbol{Z}'\boldsymbol{Z}_2'$, where b is a scalar and $\boldsymbol{D}_c$ is an $N \times N$ diagonal matrix with the diagonal being the elements of the vector $\boldsymbol{c}$. In this case,

$$V[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] = (N/n)^2 \mathbf{1}_{n_2}' \boldsymbol{Z}_2 \boldsymbol{Z}(\boldsymbol{D}_c + b\boldsymbol{x}\boldsymbol{x}')\boldsymbol{Z}'\boldsymbol{Z}_2' \mathbf{1}_{n_2}. \quad (30)$$

Next we consider condition (15) under the assumption that, given n_1, the response probabilities are exchangeable. Now,

$$E_{Z_1} V[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] \approx (N/n)^2 n_2 \left[\hat{C}/N + b n_2 (\hat{X}/N)^2\right] \quad (31)$$

and

$$V_{Z_1} E[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] = \hat{X}^2 V_{Z_1}[\hat{R}_1 | \boldsymbol{Z}]. \quad (32)$$

For large samples, taking a Taylor series expansion,

$$V_{Z_1} E[\hat{Y}_I | \boldsymbol{Z}, \boldsymbol{Z}_1] \approx (N^2 n_2 s_R^2)/(n_1 n), \quad (33)$$

where

$$s_R^2 = (\boldsymbol{y} - \hat{R}\boldsymbol{x})'\boldsymbol{Z}'\boldsymbol{Z}(\boldsymbol{y} - \hat{R}\boldsymbol{x})/(n-1) = s^2 - 2\hat{R}s_{xy} + \hat{R}^2 s_x^2 \quad (34)$$

and $\hat{R} = \hat{Y}/\hat{X}$. Therefore, condition (15) implies that, for large samples,

$$\hat{C}/N + b n_2 (\hat{X}/N)^2 \approx n s_R^2/n_1. \quad (35)$$

We now consider the complete data estimator of the variance, $v_I[\hat{Y}]$, given by expression (18) above, applied to a singly imputed data set. Condition (19) would imply that

$$n_1 s^2 + n_2 \hat{R}^2 s_x^2 + E[\mathrm{tr}(\Sigma_I) | \boldsymbol{Z}] - n^{-1} \mathbf{1}_{n_2}' E[\Sigma_I | \boldsymbol{Z}] \mathbf{1}_{n_2} \approx n s^2, \quad (36)$$

or, equivalently, for large samples,

$$\hat{C}/N + b(s_x^2 + (n_1/n)(\hat{X}/N)^2) \approx s^2 - R^2 s_x^2. \quad (37)$$

Combining (35) and (37) we have that for the imputation to be proper it is necessary that, for large samples,

$$\hat{C}/N \approx s^2 - R^2 s_x^2 \quad (38)$$

and

$$b \approx \frac{(n/n_1)s_R^2 - (s^2 - R^2 s_x^2)}{n_2 (\hat{X}/N)^2}. \quad (39)$$

We note that some simplification is possible when the following ratio model is valid:

$$y_i = \beta x_i + \epsilon_i, \quad (40)$$

where the distribution of ϵ_i given x_i has mean 0 and is uncorrelated with the other ϵ's. In this case, we have that (39) is equivalent to

$$b \approx N^2 s_R^2/n_1 \hat{X}^2. \quad (41)$$

Therefore, under model (40), we have that

$$V[\hat{R}_1 | \boldsymbol{x}] \approx N^2 s_R^2/n_1 \hat{X}^2 \approx b. \quad (42)$$

On the other hand, if the model is not valid, for example when

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (43)$$

then expression (42) is generally not satisfied. This seems to imply that for the ratio imputation to be proper, model (40) should be satisfied, even though the multiple imputation estimator is design-consistent.

3.3 Complex Sampling Schemes - Ratio Imputation with Random Noise

In Sections 3.1 and 3.2 we considered imputation under srswor. Here we extend this to more general sampling schemes. We generalize the nonresponse mechanism, by assuming that there are *imputation classes* $W_1, W_2, \ldots, W_K$ such that within class k there are n_{1k} respondents and n_{2k} nonrespondents, and given n_{1k}, the response mechanism is exchangeable within the imputation class. It is assumed that the response mechanism is independent between classes. This is a common assumption for nonresponse modelling. It is the usual assumption for justifying weighting class adjustment methods or response homogeneity group models; see, for example Särndal, Swennson and Wretman (1992; p. 578).

Without loss of generality, we relabel the population elements so that $\boldsymbol{y}' = (\boldsymbol{y}_1', \ldots, \boldsymbol{y}_K')$, where $\boldsymbol{y}_k$

corresponds to those elements of $\boldsymbol{y}$ in the kth class. We similarly partition $\boldsymbol{d}(s)'$ into $[\boldsymbol{d}_1(s)',...,\boldsymbol{d}_K(s)']$ and Δ into

$$\begin{bmatrix} \Delta_{11} & \cdots & \Delta_{1K} \\ \vdots & & \\ \Delta_{K1} & \cdots & \Delta_{KK} \end{bmatrix}.$$

Similar partitions of $\boldsymbol{Z}$, $\hat{\Delta}(s)$ and $\boldsymbol{Z}_1$ are also possible.

We consider now ratio imputations given by

$$\boldsymbol{y}_k^* = \hat{R}_{1k}\boldsymbol{Z}_{2kk}\boldsymbol{Z}_{kk}\boldsymbol{x}_k + \epsilon_k^*, \tag{44}$$

where $\hat{R}_{1k}$ depends on the observed data and ϵ_k^* is a vector of random errors with mean $\boldsymbol{0}$ and variance matrix $\boldsymbol{Z}_{2kk}\boldsymbol{Z}_{kk}(\boldsymbol{D}_{c_k} + b_k\boldsymbol{x}_k\boldsymbol{x}_k')\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'$, where $\boldsymbol{D}_{c_k}$ is an $N_k \times N_k$ diagonal matrix of a vector, $\boldsymbol{c}_k$. In order that condition (14) is satisfied, it is necessary that

$$\sum E[\hat{R}_{1k}|\boldsymbol{Z}](n_{2k}/n_k)\boldsymbol{d}_k(s)'\boldsymbol{x}_k = \sum (n_{2k}/n_k)\boldsymbol{d}_k(s)'\boldsymbol{y}_k. \tag{45}$$

We see that the weighted ratio using the original sampling weights for the respondents satisfies (45) asymptotically, but, in general, the unweighted ratio does not. Therefore, we take

$$\hat{R}_{1k} = \frac{\boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{1kk}'\boldsymbol{Z}_{1kk}\boldsymbol{Z}_{kk}\boldsymbol{y}_k}{\boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{1kk}'\boldsymbol{Z}_{1kk}\boldsymbol{Z}_{kk}\boldsymbol{x}_k} = \frac{\hat{Y}_{1k}}{\hat{X}_{1k}}. \tag{46}$$

When $\boldsymbol{x}_k = \boldsymbol{1}$, we have mean imputation within imputation class.

The imputed estimate for the population total is given by

$$\hat{Y}_I = \sum (\hat{Y}_{1k} + \hat{Y}_{2k}^*) = \sum [R_{1k}\hat{X}_k + \boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'\epsilon_k^*], \tag{47}$$

where $\hat{Y}_{2k}^* = \boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'\boldsymbol{y}_k^*$, $\boldsymbol{Z}_{2kk}$ is the $n_{2k} \times n_k$ matrix of 0's and 1's corresponding to nonresponding units in the sampled units from the kth class and

$$\hat{X}_k = \boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{kk}\boldsymbol{x}_k = \boldsymbol{d}_k(s)'\boldsymbol{x}_k. \tag{48}$$

The conditional expectation and variance of $\hat{Y}_I$, given the pattern of nonresponse, is given by

$$E[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] = \sum \hat{R}_{1k}\hat{X}_k, \tag{49}$$

and

$$V[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] = \sum b_k\hat{X}_{2k}^2 + \sum \boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'\boldsymbol{Z}_{2kk}\boldsymbol{Z}_{kk}\boldsymbol{D}_{c_k}\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'\boldsymbol{Z}_{2kk}\boldsymbol{Z}_{kk}\boldsymbol{d}_k(s), \tag{50}$$

where $\hat{X}_{2k} = \boldsymbol{d}_k(s)'\boldsymbol{Z}_{kk}'\boldsymbol{Z}_{2kk}'\boldsymbol{Z}_{2kk}\boldsymbol{Z}_{kk}\boldsymbol{x}$.

Next we consider condition (15). We have that, for large samples,

$$E_{Z_1}V[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] \approx \sum (n_{2k}/n_k)\boldsymbol{d}_k(s)'\boldsymbol{D}_{c_k}\boldsymbol{d}_k(s) + \sum b_k(n_{2k}\hat{X}_k/n_k)^2 \tag{51}$$

and

$$V_{Z_1}E[\hat{Y}_I|\boldsymbol{Z},\boldsymbol{Z}_1] \approx \sum n_k n_{2k} s_{R_k(d)}^2/n_{1k}. \tag{52}$$

where

$$s_{R_k(d)}^2 = (\boldsymbol{y}_k - \hat{R}_k\boldsymbol{x}_k)'\{\text{diag}[\boldsymbol{d}_k(s)]\}^2(\boldsymbol{y}_k - \hat{R}_k\boldsymbol{x}_k)/(n_k - 1). \tag{53}$$

Now, condition (15) implies that, for large samples,

$$\sum (n_{2k}/n_k)\boldsymbol{d}_k(s)'\boldsymbol{D}_{c_k}\boldsymbol{d}_k(s) + \sum b_k(n_{2k}\hat{X}_k/n_k)^2 \approx \sum n_k n_{2k} s_{R_k(d)}^2/n_{1k}, \tag{54}$$

which would be satisfied if we let

$$\boldsymbol{d}_k(s)'\boldsymbol{D}_{c_k}\boldsymbol{d}_k(s) + b_k(n_{2k}/n_k)\hat{X}_k^2 = n_k^2 s_{R_k(d)}^2/n_{1k}. \tag{55}$$

From (8), the complete data estimator of the variance when $\hat{Y}$ is unbiased is given by

$$v[\hat{Y}] = \sum_k \sum_\ell \boldsymbol{y}_k'\boldsymbol{Z}_{kk}'\hat{\Delta}_{k\ell}(s)\boldsymbol{Z}_{\ell\ell}\boldsymbol{y}_\ell, \tag{56}$$

since $\boldsymbol{Z}$ is block diagonal. Defining $\boldsymbol{Z}_{1kk}\boldsymbol{Z}_{kk}\boldsymbol{y}_k$, the observed data in the kth class, as $\boldsymbol{y}_{1k}$, and substituting imputation for nonresponding units, (56) becomes

$$v_I[\hat{Y}] = \sum_k \sum_\ell (\boldsymbol{y}_{1k}'\boldsymbol{Z}_{1kk}\hat{\Delta}_{k\ell}(s)\boldsymbol{Z}_{1\ell\ell}'\boldsymbol{y}_{1\ell} + 2\boldsymbol{y}_{1k}'\boldsymbol{Z}_{1kk}\hat{\Delta}_{k\ell}(s)\boldsymbol{Z}_{2\ell\ell}'\boldsymbol{y}_\ell^* + \boldsymbol{y}_k^{*\prime}\boldsymbol{Z}_{2kk}\hat{\Delta}_{k\ell}(s)\boldsymbol{Z}_{2\ell\ell}'\boldsymbol{y}_\ell^*), \tag{57}$$

and taking expectations, we have

$$E[v_1[\hat{Y}]\,|\,Z] \approx \sum_k \Big[(n_{1k}n_{2k}/n_k^2)(y_k - \hat{R}_k x_k)' Z_{kk}' \operatorname{diag}[\hat{\Delta}_{kk}(s)] Z_{kk}(y_k - \hat{R}_k x_k)\Big] + \sum_k \sum_\ell \Big\{[(n_{1k}/n_k)y_k + (n_{2k}/n_k)\hat{R}_k x_k]' Z_{kk}' \hat{\Delta}_{k\ell}(s) Z_{\ell\ell}[(n_{1\ell}/n_\ell)y_\ell + (n_{2\ell}/n_\ell)\hat{R}_\ell x_\ell]\Big\} + \sum_k \Big\{\operatorname{tr}\Big[(D_{c_k} + b_k x_k x_k')Z_{kk}'\Big\{(n_{2k}/n_k)^2\,\hat{\Delta}_{kk}(s) + (n_{1k}n_{2k}/n_k^2)\operatorname{diag}[\hat{\Delta}_{kk}(s)]\Big\}Z_{kk}\Big]\Big\}. \quad (58)$$

Condition (19) would imply that

$$\sum_k \sum_\ell y_k' Z_{kk}' \hat{\Delta}_{k\ell}(s) Z_{\ell\ell} y_\ell \approx \sum_k \Big[(n_{1k}n_{2k}/n_k^2)(y_k - \hat{R}_k x_k)' Z_{kk}' \operatorname{diag}[\hat{\Delta}_{kk}(s)] Z_{kk}(y_k - \hat{R}_k x_k)\Big] + \sum_k \sum_\ell \Big\{[(n_{1k}/n_k)y_k + (n_{2k}/n_k)\hat{R}_k x_k]' Z_{kk}' \hat{\Delta}_{k\ell}(s) Z_{\ell\ell}[(n_{1\ell}/n_\ell)y_\ell + (n_{2\ell}/n_\ell)\hat{R}_\ell x_\ell]\Big\} + \sum_k \Big\{\operatorname{tr}\Big[(D_{c_k} + b_k x_k x_k')Z_{kk}'\Big\{(n_{2k}/n_k)^2\,\hat{\Delta}_{kk}(s) + (n_{1k}n_{2k}/n_k^2)\operatorname{diag}[\hat{\Delta}_{kk}(s)]\Big\}Z_{kk}\Big]\Big\}. \quad (59)$$

Expression (59) is not very intuitive. To simplify this, we consider its model expectation under the ratio model,

$$\xi: y_k = \beta_k x_k + \epsilon_k, \quad (60)$$

where the ϵ_k's have mean $\mathbf{0}$ and covariance matrix D_{c_k}. After some simplification, we find that this implies that

$$b_k \approx \frac{n_k d_k(s)' D_{c_k} d_k(s)}{n_{1k}\hat{X}_k^2} \approx E_{Z_1} V_\xi[\hat{R}_{1k}], \quad (61)$$

where V_ξ represents the variance under model (60). From the above analysis, we see that under the condition that model (60) holds, we can get a reasonably good approximation for the imputation variance so that the multiple is proper.

4. DISCUSSION

We have considered the properties of randomly generated imputations under various sampling schemes. We have only considered the univariate cases, since, even though it is unrealistic for many applications, it does yield some enlightening results. For simple random sampling, we have found the conditions under which multiple imputation would be proper; that is, the variances derived from the multiple imputation estimator would be frequency-valid. However, when this result is extended to more general sampling schemes, using response homogeneity group model for the nonresponse mechanism, we find that the conditions required for proper imputation are generally complex. When we consider the case where the data actually satisfy the imputation model, we find that approximations are possible that greatly simplify the conditions for proper imputation.

Binder (1996) conjectured that confidence proper multiple imputation procedures for a complex design could not be achieved in general for complex surveys such as cluster or two-stage samples. Based on our findings here, it appears that the conditions required may be difficult to satisfy in practice *unless* the analyst's imputation model is valid. How robust this is to model failure deserves further study. The conditions for design-consistent point estimation are not as stringent as the conditions for design consistent variance estimation. As Kott (1995) observed, it is also apparent that the sampling weights must be used both for point estimation as well as for variance estimation in order to satisfy the conditions of being proper.

ACKNOWLEDGEMENTS

We are particularly grateful to J.N.K. Rao who pointed out an important omission in our definition of proper imputation in an earlier draft and to Don Rubin for his very useful suggestions that greatly improved the analysis. We also thank David Judkins for some of his discussions.

References

Binder, D.A. (1996). Comment to articles by Rao, Fay and Rubin. *Journal of the American Statistical Association, 91,* pp. 510-512.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Data. *Journal of the American Statistical Association, 91, No. 434.* pp. 490-498.

Kott, P.S. (1995) A Paradox of Multiple Imputation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association,* pp. 380-383.

Fay, R.E. (1996). Alternative Paradigms for the Analysis of Imputed Data. *Journal of the American Statistical Association, 91, No. 434.* pp. 490-498.

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. (with rejoinder to discussion) *Journal of the American Statistical Association, 91, No. 434.* pp. 473-489, 515-519.

Särndal, C.E., Swensson, B. and J. Wretman (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag New York Inc.