Proceedings of Statistics Canada Symposium 2001
Achieving Data Quality in a Statistical Agency: A Methodological Perspective

HOT DECK IMPUTATION FOR THE RESPONSE MODEL

Wayne A. Fuller[1] and Jae Kwang Kim[2]

ABSTRACT

Hot deck imputation, in which missing items are replaced with values from respondents, is often used in survey sampling. A model supporting such procedures is the model in which response probabilities are assumed equal within imputation cells. An efficient version of hot deck imputation is described and the variance of the efficient version derived under the cell response model. An approximation to the fully efficient procedure in which a small number of values are imputed for each nonrespondent is described. Variance estimation procedures are presented and illustrated in a Monte Carlo study.

KEY WORDS: Nonresponse; Fractional imputation; Response probability.

1. INTRODUCTION

Imputation is used in sample surveys as a method of handling item nonresponse. Kalton and Kaspryzk (1986) and Little and Rubin (1987) review various imputation procedures. In hot deck imputation, the imputed values are functions of the respondents in the current sample. Sande (1983) and Ford (1983) contain descriptions of hot deck imputation.

In one version of hot deck imputation, the imputed value is the value of a respondent in the same imputation cell, where the imputation cells form an exhaustive and mutually exclusive subdivision of the population. In random hot deck imputation, respondents are assigned values at random from respondents in the same imputation cell. The record providing the value is called the *donor* and the record with the missing value is called the *recipient*.

The variance of the imputed estimator is generally larger than the complete sample variance because nonresponse reduces sample size and because the imputed estimator may contain a component due to random imputation. Rao and Shao (1992) proposed an adjusted jackknife method for hot-deck imputation where with-replacement selection of donors is used for imputation. Rao and Sitter (1995) discussed the adjusted jackknife variance estimation method for ratio imputation. Rao (1996) and Sitter (1997) applied the adjusted jackknife method to regression imputation. Shao, Chen and Chen (1998) apply the idea of Rao and Shao (1992) to the BRR replication method. Shao and Steel (1999) propose variance estimation for survey data with composite imputation, where more than one imputation method is used, and the sampling fractions are included in the variance expressions. Yung and Rao (2000) applied the adjusted jackknife method to imputed estimators constructed with a poststratified sample. Rubin and Schenker (1986) suggested a multiple imputation procedure. Tollefson and Fuller (1992), and Särndal (1992) proposed imputation methods and corresponding variance estimators.

2. RESPONSE MODEL FOR IMPUTATION

Consider a population of N elements identified by a set of indices $U = \{1, 2, \ldots, N\}$. Associated with each unit i in the population there is a study variable y_i and a vector $\mathbf{x}_i$ of auxiliary information. The set of vectors, $(y_i, \mathbf{x}_i)$, $i = 1, 2, \ldots, N$, is denoted by F.

[1] Wayne A. Fuller, Iowa State University, Ames, IA USA, 50014
[2] Jae Kwang Kim, Westat, Rockville, MD, USA, 20850

Let A denote the indices of the elements in a sample selected by a set of probability rules called the *sampling mechanism*. The sample selection indicator function for element j is

$$I_j = \begin{cases} 1 & \text{if } j \in A \\ 0 & \text{if } j \notin A \end{cases}, \qquad j = 1, \ldots, N \tag{1}$$

and the associated vector of indicators is $\mathbf{B} = (I_1, \ldots, I_N)$. Let the population quantity of interest be $\theta_N = \theta(y_1, y_2, \ldots, y_N)$ and let $\hat{\theta}$ be a linear estimator of θ_N based on the full sample,

$$\hat{\theta} = \sum_{i \in A} w_i y_i . \tag{2}$$

If w_i is the inverse of the selection probability, then (2) is unbiased for the population total.

Let A_R and A_M denote the set of indices of the sample respondents and sample nonrespondents, respectively. Define the response indicator function

$$R_i = \begin{cases} 1 & \text{if } i \in A_R \\ 0 & \text{if } i \in A_M \end{cases} \tag{3}$$

and let $\mathbf{R} = \{(i, R_i);\ i \in A\}$. The distribution of $\mathbf{R}$ is called the *response mechanism*.

Assume that the finite population U is made up of G imputation cells. The set of elements in cell g is U_g. Let n_g be the number of sample elements in imputation cell g and let r_g, $r_g > 0$, be the number of respondents in imputation cell g. Assume the within-cell uniform response model in which the r_g responses in a cell are equivalent to a Poisson sample selected with equal probabilities from the n_g elements.

In the imputation procedure, let d_{ij} be the number of times that Y_i is used as donor for the missing Y_j and define $\mathbf{d} = \{d_{ij};\ i \in A_R, j \in A_M\}$. The distribution of $\mathbf{d}$ is called the *imputation mechanism*. Let w_{ij}^* be the factor applied to the original weight for element j when Y_i is used as a donor for element j. For element $j,\ j \in A_M$,

$$Y_{Ij} = \sum_{i \in A_R} d_{ij} w_{ij}^* y_i \tag{4}$$

is the weighted mean of the imputed values. The factor w_{ij}^* is called the *imputation fraction*. It is the fraction that donor i donates for the missing item Y_j. Note that $w_{ii}^* = 1$ for $i \in A_R$ and $w_{ii}^* = 0$ for $i \in A_M$. The d_{ij} are nonnegative and the sum of the imputation fractions of the donors for a missing item is restricted to be equal to one, that is,

$$\sum_{i \in A_R} d_{ij} w_{ij}^* = 1, \qquad \forall j \in A . \tag{5}$$

We call an estimator of the form (4) with some $w_{ij}^* < 1$ a *fractionally imputed* estimator.

A linear estimator using hot deck imputation can be written in the form

$$\hat{\theta}_I = \sum_{i \in A_R} \left(\sum_{i \in A} d_{ij} w_j w_{ij}^* \right) \tag{6}$$

$$=: \sum_{i \in A} \alpha_i Y_i , \tag{7}$$

where the notation $A =: B$ means that B is defined to be equal to A. The sum of $w_{ij}^* w_j$ over all recipients for which i is a donor (including acting as a donor for itself), denoted by α_i is the total weight of donor i. If a responding unit i is not used as a donor, except for itself, then $\alpha_i = w_i$.

3. FULLY EFFICIENT FRACTIONAL IMPUTATION

Assume all elements in an imputation cell have the same probability of responding and assume the responses are independent. Then the overall distribution of an imputed estimator under the response model can be obtained by using the probability structure of multiple phase sampling, where the response model is treated as the second phase sampling mechanism.

An imputed estimator $\hat{\theta}_I$ for the full sample estimator $\hat{\theta}$ is called co*nditionally unbiased* for $\hat{\theta}$ under the response model if it satisfies

$$E(\hat{\theta}_I - \hat{\theta} \mid \mathsf{F}, A) = 0, \tag{8}$$

where the expectation in (8) is taken with respect to the joint distribution defined by the response model and the imputation mechanism, given the realized sample A and the fixed finite population F. A necessary and sufficient condition for an imputed estimator of the form (7) to be conditionally unbiased under the assumed response model is

$$E(R_i - \alpha_i \mid \mathsf{F}, A) = w_i, \tag{9}$$

where α_i is the total weight of donor i defined in (7).

Let $\pi_{i2} = \Pr(i \in A_R \mid i \in A)$. If π_{i2} were known and $\pi_{i2} > 0$, then a hot deck imputation procedure satisfying

$$E(\alpha_i \mid A, A_R) = w_i \pi_{i2}^{-1} \tag{10}$$

will produce an unbiased estimator for the population total. In particular, hot deck imputation with $\alpha_i = w_i \pi_{i2}^{-1}$, for all $i \in A_R$, produces the most efficient imputed estimator in the sense of minimizing the imputation variance for estimators that are unbiased under the response model.

If π_{i2} is unknown and the response probabilities in a cell are uniform, then a reasonable estimator of the total is the weighted sum of ratio estimators

$$\hat{\theta}_{FE} = \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i \right) \frac{\sum_{i \in A_R \cap U_g} w_i y_i}{\sum_{i \in A \cap U_g} w_i}, \tag{11}$$

where w_i is proportional to the inverse of the selection probability. The estimator (11) is called fully efficient because it contains no variability due to random selection of donors. If the w_i are the same for all elements in a cell, the ratio

$$\left(\sum_{i \in A_R \cap U_g} w_i \right)^{-1} \sum_{i \in A_R \cap U_g} w_i y_i \tag{12}$$

is a simple mean and, hence, unbiased for the cell mean given that there is at least one respondent in the cell. If the w_i in a cell are not equal, then (12) is subject to ratio bias. It is possible for the number of elements in a cell, n_g, to be positive and the number of respondents, r_g, to be zero. If this occurs in practice, cells will be collapsed. To formally define an estimator with finite moments, we set the ratio in (12) equal to zero and set $(\sum_{i \in A_R \cap U_g} w_i)^{-1} \sum_{i \in A \cap U_g} w_i$ equal to one when $r_g = 0$.

The approximate variance of the estimator (11), given in (19) of Theorem 3.1, is the variance of the full sample estimator plus a term that depends on the cell response probabilities and the within-cell variances.

Theorem 3.1 *Assume a sequence of finite populations, indexed by* υ, *of size* N_υ, *where* $N_\upsilon > N_{\upsilon-1}$. *Let* y *be a characteristic of the population with bounded fourth moments. Let a sample of size* $n_\upsilon \ge n_{\upsilon-1}$ *be selected from the* υ*-th population with known probabilities of selection. Let the population be decomposed into* G_υ, $G_\upsilon \ge G_{\upsilon-1}$ *mutually exclusive and exhaustive cells. Let the population size of cell* g *be* $N_{g\upsilon}$, *let the sample size in cell* g *be* $n_{g\upsilon}$ *and let the number of respondents in cell* g *be* $r_{g\upsilon}$. *Assume*

$$K_{SL} G_{\upsilon}^{-1} \le N_{\upsilon}^{-1} N_{g\upsilon} \le K_{SU} G_{\upsilon}^{-1} \quad \textit{for all } \upsilon , \tag{13}$$

$$G_{\upsilon} < K_G n_{\upsilon}^{\lambda} \quad \textit{for all } \upsilon , \tag{14}$$

$$K_{wL} \le n_{\upsilon} w_i \le K_{wU} \quad \textit{for all } \upsilon , \tag{15}$$

$$K_{\pi} \le \pi_{g\upsilon} \quad \textit{for all } g \textit{ and all } \upsilon , \tag{16}$$

where K_{π}, K_{SL}, K_{SU}, K_G, K_{wL}, and K_{wU} are fixed positive constants, $0 \le \lambda < 0.5$, $\pi_{g\upsilon}$ is the common response probability in cell g of population υ, and the w_i are the weights of (2) for the estimator of the mean. Let the full sample estimator $\hat{\theta}_{\upsilon}$ be unbiased for the finite population mean.

Assume

$$V\{\hat{\theta}_{\upsilon} \mid \mathsf{F}_{\upsilon}\} < K_M V\{\hat{\theta}_{SRS,\upsilon} \mid \mathsf{F}\} \tag{17}$$

for a fixed K_M for any y with bounded fourth moments and $\hat{\theta}_{SRS,\upsilon}$ is the estimator of θ based on a simple random sample of size n_{υ}. Assume for every $i \ne j = 1, 2, \ldots, N_{\upsilon}$,

$$P(R_{i\upsilon} = 1,\ R_{j\upsilon} = 1) = P(R_{i\upsilon} = 1)\, P(R_{j\upsilon} = 1). \tag{18}$$

Then

$$\hat{\theta}_{FE\upsilon} - \hat{\theta}_{\upsilon} = \sum_{g_{\upsilon}=1}^{G_{\upsilon}} \sum_{i \in A_{g\upsilon}} w_{i\upsilon} (\pi_{g\upsilon}^{-1} R_{i\upsilon} - 1) e_{i\upsilon} + o_p(n_{\upsilon}^{-1/2}) .$$

Also

$$V(\tilde{\theta}_{FE\upsilon} \mid \mathsf{F}_{\upsilon}) = V(\hat{\theta}_{\upsilon} \mid \mathsf{F}_{\upsilon}) + E\left\{ \sum_{g_{\upsilon}=1}^{G_{\upsilon}} \pi_{g\upsilon}^{-1}(1-\pi_{g\upsilon}) \sum_{i \in A_{g\upsilon}} w_{i\upsilon}^2 e_{i\upsilon}^2 \mid \mathsf{F}_{\upsilon} \right\} , \tag{19}$$

where

$$\tilde{\theta}_{FE\upsilon} = \hat{\theta}_{\upsilon} + \sum_{g_{\upsilon}=1}^{G_{\upsilon}} \sum_{i \in A_{g\upsilon}} w_{i\upsilon} (\pi_{g\upsilon}^{-1} R_{i\upsilon} - 1) e_{i\upsilon} ,$$

F_{ν} is the υ-th population, $A_{g\upsilon}$ is the set of sample indices in the g-th cell for the υ-th sample, $e_{i\upsilon} = y_{i\upsilon} - \bar{Y}_{g\upsilon}$, and $\bar{Y}_{g\upsilon}$ is the population mean of the y-variable in cell g_{υ} of population F_{υ}.

The estimator (11) can be implemented by using fractional imputation in which every responding unit in an imputation cell is used as a donor for every nonrespondent in the cell, and the imputation weight is proportional to the sampling weight. Then, the estimator (11) can be written as the fractionally imputed estimator

$$\hat{\theta}_{FEFE} = \sum_{g=1}^{G} \sum_{j \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j w_{ij}^* y_i , \tag{20}$$

where $w_j w_{ij}^*$ is the weight of donor i for recipient j, w_{ij}^* is the imputation fraction of donor i for recipient j defined in (4), and

$$w_{ij}^* = \begin{cases} (\sum_{s \in A_R \cap U_g} w_s)^{-1} w_i R_i & \text{if } R_j = 0 \\ 1 & \text{if } R_j = 1 \text{ and } i = j . \end{cases} \tag{21}$$

The estimator (20) with w_{ij}^* of (21), algebraically equivalent to (11), is called the *fully efficient fractionally imputed* (FEFI) estimator.

We now consider variance estimation. Under complete response, let a replication variance estimator be

$$\hat{V}(\hat{\theta}) = \sum_{k=1}^{L} c_k (\hat{\theta}^{(k)} - \hat{\theta})^2 , \tag{22}$$

where $\hat{\theta}^{(k)}$ is the k-th estimate of θ_N 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 (2), the k-th replicate of $\hat{\theta}$ can be written

$$\hat{\theta}^{(k)} = \sum_{i \in A} w_i^{(k)} y_i , \tag{23}$$

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

A proposed replicate for the *FEFI* estimator $\hat{\theta}_{FEFI}$ is

$$\begin{aligned} \hat{\theta}_{FEFI}^{(k)} &= \sum_{g=1}^{G} \left(\sum_{i \in A \cap U_g} w_i^{(k)} \right) \frac{\sum_{i \in A_R \cap U_g} w_i^{(k)} y_i}{\sum_{i \in A_R \cap U_g} w_i^{(k)}} \\ &=: \sum_{g=1}^{G} \sum_{i \in A \cap U_g} \sum_{i \in A_R \cap U_g} w_j^{(k)} w_{ij}^{*(k)} y_i \end{aligned} \tag{24}$$

where $w_i^{(k)}$ is the complete sample replicate weight of unit i and $w_{ij}^{*(k)} = (\sum_{s \in A_R \cap U_g} w_s^{(k)})^{-1} w_i^{(k)}$. Using the replicates (24), the replicate variance estimator can be written as

$$\hat{V}_{FEFI} = \sum_{k=1}^{L} c_k (\hat{\theta}_{FEFI}^{(k)} - \hat{\theta}_{FEFI})^2 . \tag{25}$$

Note that the sum of the replication weights of the records for each recipient is the same as the replication weight for that unit in a complete sample.

The suggested procedure is related to the Rao and Shao (1992) variance estimator for single imputation. See also Yung and Rao (2000). Because (24) uses fractional replicates, the estimator (25) is appropriate for a vector of y-variables. Once computed, the replicate weights are appropriate for any smooth function of the vector y. The variance estimator is consistent.

THEOREM 3.2 *Let the assumptions of Theorem 3.1 hold. Let the replication variance estimator for the complete sample be of the form (22). Assume the replicates satisfy*

$$\left| \hat{\gamma}_{\upsilon}^{(k)} - \hat{\gamma}_{\upsilon} \right|^2 < \zeta_k^2 n_{\upsilon}^{-1} V\{\hat{\gamma}_{\upsilon}\} \tag{26}$$

for all k, where ζ_k are random variables with bounded fourth moments. Let $\hat{\theta}_{\upsilon}$ be the complete sample estimator of the mean. Assume the complete sample estimator of the variance of $\hat{\theta}_{\upsilon}$, denoted by $\hat{V}(\hat{\theta}_{\upsilon})$, satisfies

$$E\left\{ \left[\hat{V}(\hat{\theta}_{\upsilon}) - V(\hat{\theta}_{\upsilon} \mid \mathsf{F}_{\upsilon}) \right]^2 \mid \mathsf{F}_{\upsilon} \right\} = o(n_{\upsilon}^{-2}) \tag{27}$$

for any y with bounded fourth moments. Then for a y with bounded fourth moments, the variance estimator defined in (25) for a mean satisfies

$$\hat{V}_{FEFI} = V(\tilde{\theta}_{FE\upsilon} \mid \mathsf{F}_{\upsilon}) - N_{\upsilon}^{-2} \sum_{g_{\upsilon}=1}^{G_{\upsilon}} \sum_{i \in U_{g\upsilon}} \pi_{g\upsilon}^{-1} (1 - \pi_{g\upsilon}) e_{i\upsilon}^2 + o_p(n_{\upsilon}^{-1}) , \tag{28}$$

where $\tilde{\theta}_{FE\upsilon}$ is defined in Theorem 3.1, and the distribution is with respect to the sampling and response mechanisms.

If the finite population correction can be ignored, the estimator (25) is consistent for $V\{\hat{\theta}\}$. If the sample size is large relative to N, then an estimator of

$$N_{\upsilon}^{-2} \sum_{i=g_{\upsilon}}^{G_{\upsilon}} \sum_{i \in U_{g\upsilon}} \pi_{g\upsilon}^{-1} (1-\pi_{g\upsilon})\, e_{i\upsilon}^{2}$$

should be added to (25). An estimator is

$$N^{-1} \sum_{g=1}^{G} \left(\sum_{i \in A_g} w_i \right) \hat{\pi}_g^{-1} (1-\hat{\pi}_g)\,(r_g - 1)^{-1} \sum_{i \in A_R \cap U_g} (y_i - \bar{y}_{sg})^2 , \tag{29}$$

where $\hat{\pi}_g = n_g^{-1} r_g$ and $\bar{y}_{sg} = r_g^{-1} \sum_{i \in A_R \cap U_g} y_i$.

Estimator (29) can be constructed directly or by using replicates.

4. THE CELL MEAN MODEL

The imputation and variance estimation procedures outlined for the response model also produce consistent estimators for the cell mean model. Under the cell mean model, the elements within a cell of the finite population are a realization of independently and identically distributed random variables with mean μ_g and variance σ_g^2. Thus, for the cell mean model

$$Y_i \overset{i.i.}{\sim} (\mu_g, \sigma_g^2), \qquad i \in U_g , \tag{30}$$

where U_g denotes the set of indices for the g-th imputation cell and $\overset{i.i.}{\sim}$ is the abbreviation for *independently and identically distributed*. The imputation procedure based on the response model is not necessarily fully efficient for the population mean under the cell mean model, but the estimator of the mean and the estimator of the variance of the estimated mean are consistent.

The distribution of Y in the sample is determined by the sampling mechanism and by the distribution of the vector Y. If there is no dependence of the distribution of Y on the sampling mechanism, the sampling mechanism is said to be *ignorable*. The response mechanism is *ignorable* if the conditional distribution of Y given the sample outcome A and the set of respondents A_R is the same as the conditional distribution of Y given A.

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

$$Y_i \mid (A, A_R) \overset{i.i.}{\sim} (\mu_g, \sigma_g^2) \qquad i \in U_g . \tag{31}$$

The conclusions of Theorems 3.1 and 3.2 can be considered to be conditional results for a given finite population. If the cell mean model holds, we can relax the assumption of a common response probability.

Theorem 4.1 *Let the assumptions of Theorem 3.1 hold with the exception of (16). Assume there is some $K_\pi > 0$ such that*

$$K_\pi < \pi_{g\upsilon i} \tag{32}$$

for all g and all υ, where $\pi_{g\upsilon i}$ is the response probability of element i in population υ. Assume the cell mean model (31) holds. Then there exists a sequence of random variables $\tilde{\theta}_{FE\upsilon}$ with

$$p \lim_{\upsilon \to \infty} n_{\upsilon}^{1/2} (\hat{\theta}_{FE\upsilon} - \tilde{\theta}_{FE\upsilon}) = 0$$

and

$$V(\tilde{\theta}_{FE\upsilon}) = V(\hat{\theta}_{\upsilon}) + E\left\{ \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i\in A_{g\upsilon}} \tilde{\pi}_{g\upsilon}^{-2} \pi_{g\upsilon i}(1-\pi_{g\upsilon i}) w_{i\upsilon}^2 \sigma_{g\upsilon}^2 \right\}, \tag{33}$$

where

$$\tilde{\pi}_{g\upsilon} = \left(\sum_{i\in A_{g_\upsilon}} w_{i\upsilon} \right)^{-1} \sum_{i\in A_{g_\upsilon}} w_{i\upsilon}\pi_{g\upsilon i}.$$

Theorem 4.2 *Let the assumptions of Theorem 4.1 hold. Let the replication variance estimator for the complete sample be of the form (22). Let the replication variance estimator satisfy the assumptions of Theorem 3.2. Then the variance estimator for a mean defined by (25) satisfies*

$$\hat{V}_{FEFI} = V(\tilde{\theta}_{FE\upsilon} \mid \mathsf{F}_\upsilon) - N_\upsilon^2 \sum_{g_\upsilon=1}^{G_\upsilon} \sum_{i\in U_{g\upsilon}} \tilde{\pi}_{g\upsilon}^{-2} \pi_{g\upsilon i}(1-\pi_{g\upsilon i})\sigma_{g\upsilon}^2 + o_p(n_\upsilon^{-1}).$$

5. APPROXIMATIONS TO THE FULLY EFFICIENT PROCEDURE

In the previous sections, the estimator $\hat{\theta}_{FEFI}$ was constructed to produce zero imputation variance. The implementation of the fractional imputation procedure as described in (20), could require the use of a large number of donors for each recipient. Therefore, we outline a procedure with a fixed number of donors per recipient that is fully efficient for the grand total, but not necessarily fully efficient for subpopulations. The procedure assigns donors to produce small between-recipient variance of imputed values and modifies the weights of donors to attain full efficiency for the total.

Let w_i be the original weights of the respondents. Define

$$S_{gw} = \sum_{i\in A_{Rg}} w_i, \tag{34}$$

the sum of the weights of the respondents in cell g. We assume that M donors are to be assigned to each recipient. Donors are assigned such that the distribution of donors approximates the distribution of the respondents. One possible selection method is stratification, where donors are selected within a stratum with probability

$$P_{ijgh} = S_{gwh}^{-1} w_i. \tag{35}$$

and

$$S_{gwh} = \sum_{i\in A_{Rgh}} w_i$$

is the sum of the weights of donor elements assigned to stratum h in cell g. Given the set of donors for recipient j, the initial w_{ij}^* is

$$w_{ij0}^* = S_{gw}^{-1} S_{gwh},$$

where element i belongs to stratum gh.

A second possible selection procedure is to use systematic sampling with probability proportional to the weights to select donors for each recipient. With this procedure the initial w_{ij0}^* are M^{-1}.

After the donors are assigned, the initial weights, w_{ij}^* are adjusted with regression procedures so that the sum of the weights gives the fully efficient estimator of the mean of y, and such that the estimated cumulative distribution function based on the weights approximates the fully efficient estimator of the cumulative distribution function. Simulation studies have shown this procedure to work well with $M = 5$.

6. SUMMARY

The properties of fractional imputation, where several donors are used for each missing value and each donor is given a fraction of the weight of the nonrespondent, are described. It is shown that fractional imputation with a small number of donors per nonrespondent can give a fully efficient estimator of the mean. Fractional imputation permits the construction of a single set of variance estimation replicates that can be used for variance estimation for imputed variables, for variables observed on all respondents, and under model assumptions, for functions of the two types of variables. For example, the replicates give consistent estimates of the variances of domain means. Fractional imputation gives variance estimates with smaller bias and smaller variance than multiple imputation estimators with the same number of imputations.

ACKNOWLEDGMENTS

This research was partially supported by a subcontract between Westat and Iowa State University under Contract No. ED--99--CO—0109 between Westat and the Department of Education. We thank Jean Opsomer and Damiao Da Silva for useful comments.

REFERENCES

Fay, R. E. (1991), "A Design-Based Perspective on Missing Data Variance", *Proceedings of Bureau of the Census Annual Research Conference*, *American Statistical Association*, pp. 429−440.

Fay, R. E. (1992), "When are Inferences From Multiple Imputation Valid?" *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 227−232.

Ford, B. M. (1983), *An Overview of Hot-Deck Procedures*. *Incomplete Data in Sample Surveys*, Vol. 2. New York:Academic Press.

Kalton, G. and Kasprzyk, D. (1986), "The Treatment of Missing Survey Data", *Survey Methodology*, 12, pp. 1−16.

Little, R. J. A. and Rubin, D. B. (1987), "*Statistical Analysis with Missing Data*", New York:Wiley.

Rao, J. N. K. (1996), "On Variance Estimation with Imputed Survey Data", *Journal of the American Statistical Association*, 91, pp. 499−506.

Rao, J. N. K. and Shao, J. (1992), "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation", *Biometrika*, 79, pp. 811−822.

Rao, J. N. K. and Sitter, R. R. (1995), "Variance Estimation Under Two-Phase Sampling with Applications to Imputation for Missing Data", *Biometrika*, 82, pp. 453−460.

Rubin, D. B. (1976), "Inference and Missing Data", *Biometrika*, 63, pp. 581−590.

Rubin, D. B. (1978), "Multiple Imputations in Sample Surveys: A Phenomenological Bayesian Approach to Nonresponse", *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 20−28.

Rubin, D. B. and Schenker (1986), "Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse", *Journal of the American Statistical Association*, 81, pp. 366−374.

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

Sande, I. G. (1983), *Hot-Deck Imputation Procedures*, *Incomplete Data in Sample Surveys*, Vol. 3. New York:Academic Press.

Särndal, C.-E. (1992), "Methods for Estimating the Precision of Survey Estimates When Imputation Has Been Used", *Survey Methodology*, 18, pp. 241–252.

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, pp. 819–831.

Shao, J. and Steel, P. (1999), "Variance Estimation For Survey Data with Composite Imputation and Nonnegligible Sampling Fractions", *Journal of the American Statistical Association*, 94, pp. 254–265.

Sitter, R. R. (1997), "Variance Estimation for the Regression Estimator in Two-Phase Sampling", *Journal of the American Statistical Association*, 92, pp. 780–787.

Tollefson, M. and Fuller, W. A. (1992), "Variance Estimation for Sampling with Random Imputation", *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, pp. 140–145.

Yung, W. and Rao, J. N. K. (2000), "Jackknife Variance Estimation Under Imputation for Estimators Using Poststratification Information", *Journal of the American Statistical Association*, 95, pp. 903–915.