Resampling Methods for Imputing Missing Observations

by

M.S. Srivastava
Department of Statistics
University of Toronto

Technical Report No. 9706, March 3, (1997)

TECHNICAL REPORT SERIES

University of Toronto
Department of Statistics

Resampling Methods for Imputing Missing Observations

BY

M.S.SRIVASTAVA

Dept of Statistics, University of Toronto, Toronto, Ontario, Canada. M5S 3G3

ABSTRACT

Rubin has proposed multiple imputation to replace the missing observations when the non response is random. This requires the records of several completed data sets. Also, the multiply imputed estimator in any replicate has a larger variance than a singly imputed estimator. In this paper, a new multiply estimator is proposed overcoming the above shortcomings. For singly imputed data, Rao and Shao have provided an adjusted jackknife estimator for the variance. In this paper, a jackknife estimator is proposed requiring no special care for adjustment. A bootstrap estimator for the variance and a bootstrap distribution for the pivotal quantity to obtain confidence intervals for the population mean are given. Following the approach of Srivastava and Carter, it is shown that all the results of this paper require only the records of the respondents. Finally, it is shown that Fay's counter example does not apply to the methods of this paper.

Key Words and Phrases: Bootstrapping, Jackknifing, Missing observations, Simple Random Sampling, Single and Multiple Imputation.

1.INTRODUCTION

When the observations are missing at random, it is a common practice to replace the missing values by their imputed values and treat the data with imputed values as the complete data set. Inference such as estimation and confidence intervals are obtained for the population parameters or any function of these parameters from the completed data set. However, the usual estimating formulae, say, for the variance of the estimator do not provide correct answer since it fails to take into account the extra variability due to imputation and thus it provides an underestimate of the variance. Rao and Shao (1992) provided an adjusted jackknife estimator of the variance overcoming this shortcoming. Thus, the criticism against singly imputed data that it usually provides underestimates of the variance of the estimator has been somewhat muted.

Rubin (1978) proposed multiple imputation in which each missing value is replaced by two or more imputed value and argued that this will represent a distribution of likely values. Rubin and Schenker (1986) provided an improved asymptotic distribution of the mean estimator based on all imputations and showed that the coverage is considerably improved by multiple imputation as opposed to single imputation. However, many survey analysts have argued against multiple imputation since it is expensive and difficult to keep a record of several completed data sets. Fay (1993) has argued against it on technical grounds.

In this paper, we follow the approach of Srivastava and Carter (1986) to impute missing values, singly or multiply. The advantages of this method is that it neither requires the extra care as needed in Rao-Shao jackknife method in singly imputation nor does it require the survey analysts to keep a record of several imputed complete data sets as well as the record of the respondents as in Rubin (1978). What is required is simply the record of the respondents. We provide jackknife as well as bootstrap estimate of the variance of the imputed mean. We show that the variance of the Rubin-Schenker's type multiply imputed data mean obtained from any replicate is always larger than the variance of the corresponding singly imputed data mean. An alternative multiply imputed estimator is proposed overcoming the above shortcoming. We argue against using the grand mean of all the imputed data set as an estimator of the population mean since it gives considerably less weight to the imputed values. However, it provides an excellent coverage and may be used to obtain confidence intervals for the population parameters.

To fix our ideas, we consider only simple random sampling in this paper. It is assumed that the observations are missing at random. The organization of the paper is as follows. In section 2, we describe a single imputation method, obtain the imputed mean as an estimator of the population mean and obtain its variance. The jackknife and bootstrap estimator of the variance are given in Sections 3 and 4 respectively. In Section 5, we give the bootstrap distribution. Rubin and Schenker type estimator is considered in Section 6 while Section 7 gives a new multiply imputed estimator with the same variance as the singly imputed estimator. It is shown that only the record of the respondent's is needed to carry out alll the procedures described in this paper. In Section 8, Fay's (1993) counter example is revisited and it is shown that it does not apply to the methods of this paper. The paper concludes in Section 9.

2.Estimator Based on Single Imputation

In this section, we consider simple random samples of size n in which n_1 subjects responds and n_0 subjects do not respond to an item y. We shall denote the n_1 responses by $y_1, \ldots, y_{n1}$, and their sample mean and variance by

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i \ , \quad \text{and} \quad S_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 \ ,$$

respectively. The population mean and variance will be denoted by $\bar{Y}$ and S^2 respectively. In what follows, we shall igonre the finite population correction. To impute the n_0 missing values, we define the residuals

$$\xi_i = \left(\frac{n_1}{n_1 - 1}\right)^{\frac{1}{2}}(y_i - \bar{y}_1), \quad i = 1, \ldots, n_1 \tag{1}$$

from which a sample of size n_0 is drawn with replacement. We shall denote these values by $\xi_1^*, \ldots, \xi_{n_0}^*$. Since the probability is $1/n_1$ of drawing any of the $\xi_i's$,

$$E_*(\xi_i^*) = \frac{1}{n_1}\sum_{i=1}^{n_1}\xi_i = \frac{1}{n_1}\left(\frac{n_1}{n_1 - 1}\right)^{\frac{1}{2}}\sum_{i=1}^{n_1}(y_i - \bar{y}_1) = 0$$

and

$$E_*(\xi_i^{*2}) = \frac{1}{n_1}\left(\frac{n_1}{n_1 - 1}\right)\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2 = S_1^2$$

where E_* denotes the conditional expectation given $y_1, \ldots, y_{n1}$. Similarly, we shall denote by V_*

for the conditional variance. We define the n_0 imputed values by

$$y_l^* = \bar{y}_1 + \xi_l^* \quad , \quad l = 1, \ldots, n_0$$

The imputed mean based on the observed and imputed values is given by

$$\begin{aligned}\bar{y}_I^* &= \frac{\sum_{i=1}^{n_1} y_i + \sum_{i=1}^{n_0} y_l^*}{n} \\ &= \frac{\sum_{i=1}^{n_1} y_i + n_0\bar{y}_1 + \sum_{i=1}^{n_0} \xi_l^*}{n} \\ &= \frac{(n\bar{y}_1 + n_0\bar{\xi}_{n_0}^*)}{n} \qquad (2)\end{aligned}$$

where $\bar{\xi}_{n_0}^* = \frac{1}{n_0}\sum_{l=1}^{n_0} \xi_l^*$. It follows easily that

$$E(\bar{y}_I^*) = E[E_*(\bar{y}_1 + \frac{n_0}{n}\bar{\xi}_{n_0}^*)] = E(\bar{y}_1) = \bar{Y}$$

and

$$\begin{aligned}V(\bar{y}_I^*) &= V\Big(E_*(\bar{y}_I^*)\Big) + E[V_*(\bar{y}_I^*)] \\ &= \frac{S^2}{n_1} + \frac{n_0}{n^2}S^2 \\ &= \left(\frac{1}{n_1} + \frac{n_0}{n^2}\right)S^2. \qquad (3)\end{aligned}$$

Thus, a correct estimator of $V(\bar{y}_I^*)$ will *not* be S_I^{*2}/n but

$$\left(\frac{1}{n_1} + \frac{n_0}{n^2}\right)S_I^{*2} \quad ,$$

where

$$(n-1)S_I^{*2} = \sum_{i=1}^{n_1}(y_i - \bar{y}_I^*)^2 + \sum_{l=1}^{n_0}(y_l^* - \bar{y}_I^*)^2 \qquad (4)$$

The bias is of the order $O(n_0/n_1n^2)$ since

$$E_*(S_I^{*2}) = (1 - \frac{n_0}{n(n-1)})S_1^2 \; .$$

On the otherhand, the jackknife estimator of Rao and Shao (1992) has a bias of the order $O(n_0/n^3)$. The jackknife estimator not only reduces the bias but can easily be applied to obtain the variance of any nonlinear statistic. However, the jackknife estimator of Rao and Shao requires some adjustment and thus requires special care. In the next section, we provide a jackknife estimator based on the approach of this paper, which does not require any special care.

3. Jackknife Estimator of the Variance

Let A_1 denote the sample of respondents and A_0 non-respondetns. Define

$$\begin{aligned} \bar{y}^*_{I(j)} &= \frac{(n-1)\bar{y}_{1(j)} + n_0\bar{\xi}^*_{n_0}}{n-1} \ , \ j \in A_1 \\ &= \frac{(n-1)\bar{y}_1 + (n_0-1)\bar{\xi}^*_{n_0(j)}}{n-1} \ , \ j \in A_0 \ , \ j = 1, \ldots, n, \end{aligned} \tag{5}$$

where

$$\bar{y}_{1(j)} = \frac{1}{n_1 - 1}\sum_{i \neq j} y_i = \frac{n_1\bar{y}_1 - y_j}{n_1 - 1}$$

and

$$\bar{\xi}_{n_0(j)} = \frac{1}{n_0 - 1}\sum_{i \neq j} \xi_i = \frac{n_0\bar{\xi}^*_{n_0} - \xi^*_j}{n_0 - 1}$$

Then

$$\frac{1}{n}\sum_{j=1}^{n} \bar{y}^*_{I(j)} = \bar{y}^*_I \ .$$

The jackknife estimator is defined by

$$\begin{aligned} \frac{n}{n-1}v^*_J &= \sum_{j=1}^{n}\left(\bar{y}^*_{I(j)} - \bar{y}^*_I\right)^2 \\ &= \sum_{\substack{j=1 \\ j \in A_1}}^{n_1}\left(\bar{y}_{1(j)} + \frac{n_0}{n-1}\bar{\xi}^*_{n_0} - \bar{y}^*_I\right)^2 + \sum_{\substack{j=1 \\ j \in A_0}}^{n_0}\left(\bar{y}_1 + \frac{n_0-1}{n-1}\bar{\xi}^*_{n_0(j)} - \bar{y}^*_I\right)^2 \\ &= \frac{1}{(n-1)^2}\left[\sum_{j=1}^{n_1}(y_j - \bar{y}^*_I)^2 + \sum_{j=1}^{n_0}(y^*_j - \bar{y}^*_I)^2 + n_0 S^2_1 + \frac{n-1}{n_1-1}n_0 S^2_1\right] \end{aligned}$$

Thus,

$$v^*_J = \frac{S^{*2}_I}{n} + \frac{n_0}{n(n-1)}S^2_1 + \frac{n_0}{n(n_1-1)}S^2_1$$

$$
\begin{aligned}
&= \frac{1}{n}(S_I^{*2} - S_1^2) \;+\; \left(\frac{n_0}{n(n-1)} + \frac{n-1}{n(n_1-1)}\right) S_1^2 \\
&\simeq \frac{1}{n}(S_I^{*2} - S_1^2) \;+\; \left(\frac{1}{n_1} + \frac{n_0}{n^2}\right) S_1^2
\end{aligned}
\tag{6}
$$

and

$$
E_*(v_J^*) \simeq \left(1 + \frac{n_0 n_1}{n^2}\right)\frac{S_1^2}{n_1} - \frac{n_0}{n^3} S_1^2
$$

Hence, the bias is of the same order as for the Rao-Shao jackknife estimator but it does not require any special care.

4.Bootstrap Estimator of the Variance

We define the residuals as in (1). However, we draw a sample of size n with replacement from $\xi_1, \ldots, \xi_{n_1}$. We also replicate it m times. Denoting the first n_1 observations with its m replicates by ξ_{ij}^*, $j = 1, \ldots, n_1$, $i = 1, \ldots, m$. We impute the observed observations by

$$
y_{ij}^* \;=\; \bar{y}_1 + \xi_{ij}^* \;,\; j = 1, \ldots, n_1,\; i = 1, \ldots, m. \tag{7}
$$

Let

$$
\begin{aligned}
\bar{y}_{in_1}^* &= \frac{1}{n_1}\sum_{j=1}^{n_1} y_{ij}^* \\
&= \bar{y}_1 \;+\; \bar{\xi}_{in_1}^* \;,\; i = 1, \ldots, m
\end{aligned}
\tag{8}
$$

where $\bar{\xi}_{in_1}^* \;= {n_1}^{-1} \sum_{j=1}^{n_1} \; \xi_{ij}^*$. Next, denoting the remaining n_0 observations with its m replicates by ξ_{ij}^{**} , $j = 1, \ldots, n_0$, $i = 1, \ldots, m$, we define the imputed values of the missing observations as

$$
\begin{aligned}
y_{ij}^{**} &= \bar{y}_{in_1}^* + \xi_{ij}^{**} \\
&= \bar{y}_1 + \bar{\xi}_{in_1}^* + \xi_{ij}^{**} \;,\; j = 1, \ldots, n_0,\; i = 1, \ldots, m.
\end{aligned}
\tag{9}
$$

We now define the imputed estimator of the mean for the i^{th} replicate by

$$
\begin{aligned}
\bar{y}_{Ii}^{**} &= \frac{\sum_{j=1}^{n_1} y_{ij}^* + \sum_{l=1}^{n_0} y_{il}^{**}}{n} \\
&= \bar{y}_{in_1}^* + \frac{n_0}{n}\bar{\xi}_{in_0}^{**}
\end{aligned}
$$

$$= \bar{y}_1 + \bar{\xi}^*_{in_1} + \frac{n_0}{n}\bar{\xi}^{**}_{in_0} \ , \tag{10}$$

and the mean of all the replicates by

$$\begin{aligned}
\bar{y}^{**}_{I.} &= \frac{1}{m}\sum_{i=1}^{m}\bar{y}^{**}_{Ii} \\
&= \bar{y}^*_{.n_1} + \frac{n_0}{n}\bar{\xi}^{**}_{.n_0} \\
&= \bar{y}_1 + \bar{\xi}^*_{.n_1} + \frac{n_0}{n}\bar{\xi}^{**}_{.n_0}
\end{aligned} \tag{11}$$

where

$$\bar{\xi}^*_{.n_1} = \frac{1}{m}\sum_{i=1}^{m}\bar{\xi}^*_{in_0} \quad \text{, and} \quad \bar{\xi}^{**}_{.n_0} = \frac{1}{m}\sum_{i=1}^{m}\bar{\xi}^{**}_{in_0} \ .$$

Then, we define the bootstrap estimate of the variance by v^*_B given by

$$\begin{aligned}
(m-1)v^*_B &= \sum_{i=1}^{m}(\bar{y}^{**}_{Ii} - \bar{y}^{**}_{I})^2 \\
&= \sum_{i=1}^{m}[(\bar{\xi}^*_{in_1} - \bar{\xi}^*_{.n_1}) + \frac{n_0}{n}(\bar{\xi}^{**}_{in_0} - \bar{\xi}^{**}_{.n_0})]^2
\end{aligned} \tag{12}$$

Hence,

$$\begin{aligned}
(m-1)E_*(v^*_B) &= \frac{mS_1^2}{n_1} - \frac{mS_1^2}{mn_1} + \frac{n_0^2}{n^2}E_*\Big(\frac{mS_1^2}{n_0} - \frac{mS_1^2}{mn_0}\Big) \\
&= (m-1)\left[\frac{1}{n_1} + \frac{n_0}{n^2}\right]S_1^2
\end{aligned} \tag{13}$$

Thus, the boostrap estimate of the varaince is an unbiased estimator while the jackknife is not an unbiased estimator. It may be argued that most agency collecting the data do not like to maintain several set of completed data sets. But this is not what is required here in obtaining the bootstrap estimator. The only thing required is respondent's record.

5.Bootstrap Distribution

In this section, we wish to find the bootstrap distribution of the pivotal quantity

$$\frac{\bar{y}^*_I - \bar{Y}}{\sqrt{Var(\bar{y}^*_I)}} \ ,$$

where $\bar{y}_I^*$ is the mean of the completed data set with imputed values as obtained and defined in Section 2. To obtain the bootstrap distribution, we draw a sample of n observations with replacement from $\xi_1, \ldots, \xi_n$ and replicate it m times. We shall denote the observations so obtained by ξ_{ij}^* for the first n_1 observations and by ξ_{ij}^{**} for the remaining n_0 observations, $i = 1, \ldots, m$ and define

$$\bar{y}_{Ii}^{**} = \bar{y}_I^* + \bar{\xi}_{in_1}^* + \frac{n_0}{n}\bar{\xi}_{in_0}^{**} \ , \ i = 1, \ldots, m$$

Let

$$(n-2)S_i^{**2} = \left[\sum_{j=1}^{n_1}(\xi_{ij}^* - \bar{\xi}_{in_1}^*)^2 + \sum_{j=1}^{n_0}(\xi_{ij}^{**} - \bar{\xi}_{in_0}^{**})^2\right]\left(\frac{1}{n_1} + \frac{n_0}{n^2}\right)$$

Then as n → ∞, the limiting distributions of

$$\frac{\bar{y}_{Ii}^{**} - \bar{y}_I^*}{S_i^{**}} \quad \text{and} \quad \frac{\bar{y}_I^* - \bar{Y}}{\sqrt{v_J^*}}$$

are the same $N(0,1)$. Thus, a confidence interval for $\bar{Y}$ can be obtained by replicating the first pivotal quantity a large number of times (m large) and using its empirical distribution.

6.Rubin-Schenker Type Estimator

Rubin and Schenker (1986) proposed multiple imputation. Although, their imputation is slightly different, the properties of their estimator based on asymptotic Bayesian bootstrap method will be similar to the one described below. However, this method does not require the record of all the completed data sets. We impute the missing values as in (9), Section 4.

Define

$$\begin{aligned}
\bar{y}_{Ii}^{**}(s) &= \frac{\sum_{i=1}^{n_1} y_i + \sum_{i=l}^{n_0} y_{il}^{**}}{n} \\
&= \frac{n_1\bar{y}_1 + n_0\bar{y}_{in_0}^{**}}{n} \ , \ \bar{y}_{in_0}^{**} = n_0^{-1}\sum_{i=l}^{n_0} y_{il}^{**} \ , \ i = 1, \ldots, m \\
&= \bar{y}_1 + \frac{n_0}{n}(\bar{\xi}_{in_1}^* + \bar{\xi}_{in_0}^{**})
\end{aligned}$$

Thus

$$V[\bar{y}_{Ii}^{**}(s)] = \frac{S^2}{n_1} + \frac{n_0}{n_1 n}S^2 \ ,$$

$$\geq \quad V(\bar{y}_I^*) = \left(\frac{1}{n_1} + \frac{n_0}{n^2}\right) S^2 \ ,$$

where $\bar{y}_I^*$ is the mean based on single imputation as defined in (2). Thus, the variance of the imputed mean using multiple imputation in any replicate is larger than the variance of the mean based on single imputation. The mean of all the m replicates is defined by

$$\begin{aligned} \bar{y}_{I.}^{**}(s) &= \frac{1}{m}\sum_{i=1}^{m} \bar{y}_{Ii}^{**} \\ &= \bar{y}_1 + \frac{n_0}{n}(\bar{\xi}_{.n_1}^{*} + \bar{\xi}_{.n_0}^{**}). \end{aligned}$$

with

$$\begin{aligned} V(\bar{y}_{I.}^{**}(s) &= \frac{S^2}{n_1} + \frac{n_0}{mnn_1} S^2 \\ &\to \frac{S^2}{n_1} \text{ as } m \to \infty. \end{aligned}$$

The variance can be estimated by

$$T^{**} = \frac{1}{nm}\sum_{i=1}^{m} S_i^{**} + \frac{m+1}{m} B^{**} \quad ,$$

where

$$\begin{aligned} (n-1)S_i^{**^2} &= \sum_{j=1}^{n_1}[y_j - \bar{y}_{Ii}^{**}(s)]^2 + \sum_{j=1}^{n_0}[y_{ij}^{**} - \bar{y}_{Ii}^{**}(s)]^2 \\ &= \sum_{j=1}^{n_1}(y_j - \bar{y}_1)^2 + \sum_{j=1}^{n_0}(\bar{y}_{ij}^{**} - \bar{y}_{in_0}^{**})^2 + \frac{n_0 n_1}{n}(\bar{y}_1 - \bar{y}_{in_0}^{**})^2 \quad , \quad i = 1, \ldots, m \end{aligned}$$

and

$$\begin{aligned} (m-1)B^{**} &= \sum_{i=1}^{m}\left(\bar{y}_{Ii}^{**}(s) - \bar{y}_{I.}^{**}(s)\right)^2 \\ &= \left(\frac{n_0}{n}\right)^2 \sum_{i=1}^{m}(\bar{y}_{in_0}^{**} - \bar{y}_{.n_0}^{**})^2 \ , \ \bar{y}_{.n_0}^{**} = m^{-1}\sum_{i=1}^{m} \bar{y}_{in_0}^{**} \quad . \end{aligned}$$

It can be shown that

$$E(S_i^{**2}) = S^2$$

and

$$E(B^{**}) = \frac{n_0}{nn_1}S^2$$

Hence,

$$\begin{aligned} E(T^{**}) &= \frac{S^2}{n} + \frac{m+1}{m}\frac{n_0}{nn_1}S^2 \\ &= \frac{S^2}{n_1} + \frac{n_0}{mnn_1}S^2 . \end{aligned}$$

Thus, T^{**} is an unbiased estimator of $V(\bar{y}_{I.}^{**}(S)$. It can be shown that as $n \to \infty$ and $m \to \infty$.

$$\bar{y}_{I.}^{**}(s) - \bar{Y} \quad \to \quad N(0, T^{**}) .$$

Improvements given in Rubin and Schenker (1986) can also be used. However, in using $\bar{y}_{I.}^{**}(s)$ as an estimator of the population mean, we are giving very little weight to the imputed values. In effect, it amounts to using only the observed values. Thus, while $\bar{y}_{I.}^{**}(s)$ provides an excellent coverage for the population mean, it may not be used as an estimator. On the otherhand, the estimator $\bar{y}_{Ii}^{**}(s)$ gives proper weight to the imputed values. However, this estimator has a larger variance than the corresponding estimator obtained from a single imputation disucssed in Section 2. In the next section, we propose a multiply imputed estimator which does not have this shortcoming.

7.A New Multiply Imputed Estimator

In this section, we propose a new multiply imputed estimator which has the same varinace as a singly imputed estimator. In addition, it does not require double imputing i.e. imputing the observed values also. As noted earlier, none of the procedures in this paper require several completed data sets. We begin with the residuals, $\xi_1, \ldots, \xi_{n_1}$ defined in (1). We draw a sample of size n_0 with replacement from these residuals. We also replicate it m times. These observations will be denoted by ξ_{ij}^* , $j = 1, \ldots, n_0$, $i = 1, \ldots, m$. We impute the missing values by

$$y_{ij}^* = \bar{y}_1 + \xi_{ij}^* \ , \ j = 1, \ldots, n_0 \ , \ i = 1, \ldots, m \ .$$

We define the imputed mean for the i^{th} replicate by

$$\begin{aligned}\bar{y}^*_{Ii}(s) &= \frac{\sum_{l=1}^{n_1} y_l + \sum_{j=1}^{n_0} y^*_{ij}}{n} \\ &= \bar{y}_1 + \frac{n_0}{n}\bar{\xi}^*_{in_0} \ , \ i = 1, \ldots, m \ , \text{ where } \ \bar{\xi}^*_{in_0} = n_0^{-1}\sum_{j=1}^{n_0}\xi_{ij} \ .\end{aligned}$$

Then

$$V[\bar{y}^*_{Ii}(s)] = \frac{S^2}{n_1} + \frac{n_0}{n^2}S^2 \ ,$$

the same as $V(\bar{y}^*_I)$ given in Section 2. The mean of all the m imputation is defined by

$$\begin{aligned}\bar{y}^*_{I.}(s) &= \frac{1}{m}\sum_{i=1}^{m}\bar{y}^*_{Ii} \\ &= \bar{y}_1 + \frac{n_0}{n}\bar{\xi}^*_{.n_0} \ ,\end{aligned}$$

where $\bar{\xi}^*_{.n_0} = m^{-1}\sum_{i=1}^{m}\bar{\xi}^*_{in_0}$. Its variance is given by

$$V[\bar{y}^*_{I.}(s)] = \frac{S^2}{n_1} + \frac{n_0}{n^2 m} \quad .$$

To estimate this variance, we define

$$\begin{aligned}B^* &= \frac{1}{m-1}\sum_{i=1}^{m}\Big(\bar{y}^*_{Ii}(s) - \bar{y}^*_{I.}(s)\Big)^2 \\ &= \frac{1}{m-1}\left(\frac{n_0}{n}\right)^2\sum_{i=1}^{m}(\bar{\xi}^*_{in_0} - \bar{\xi}^*_{.n_0})^2\end{aligned}$$

Clearly,

$$E(B^*) = \frac{n_0}{n^2}S^2 \ .$$

Further, we define

$$\begin{aligned}(n-2)S_i^{*2} &= \sum_{l=1}^{n_1}(y_l - \bar{y}^*_{Ii}(s))^2 + \sum_{j=1}^{n_0}(y^*_{ij} - \bar{y}^*_{Ii}(s))^2 \\ &= \sum_{l=1}^{n_1}(y_l - \bar{y}_1)^2 + \sum_{j=1}^{n_0}(y^*_{ij} - \bar{y}^*_{in_0})^2 + \frac{n_0 n_1}{n}(\bar{y}_1 - \bar{y}^*_{in_0})^2\end{aligned}$$

$$= \sum_{l=1}^{n_1}(y_l - \bar{y}_1)^2 + \sum_{j=1}^{n_0}(\xi^*_{ij} - \bar{\xi}^*_{in_0})^2 + \frac{n_0 n_1}{n}\bar{\xi}^{*2}_{in_0}$$

Thus, we have

$$E(S_i^{*2}) = S^2 + \frac{n_1}{n(n-2)}S^2$$

Hence, we can use

$$T^* = \frac{1}{n_1 m}\sum_{l=1}^{m} S_i^{*2} + \frac{1}{m}B^*$$

as an estimator of $V(\bar{y}^*_{I.})$. Following Rubin and Schenker (1986), it can be shown that

$$\bar{y}^*_{I.}(S) - \bar{Y} \rightarrow N(0, T^*)$$

where

$$E(T^*) = \left(\frac{1}{n_1} + \frac{n_0}{mn^2}\right)S^2 + \frac{S^2}{m(n-2)} \rightarrow \frac{S^2}{n_1}\ , \text{ as } \ n, m \rightarrow \infty\ .$$

Thus, confidence intervals or any inference about the population mean $\bar{Y}$ can be obtained from the above asymptotic distribution. Improvements in distributions similar to Rubin and Schenker (1986), can also be obtained. Since $\bar{y}^*_{I.}(s)$ has a smaller variance than $\bar{y}^{**}_{I.}(s)$, it should provide a better coverage than using $\bar{y}^{**}_{I.}(s)$.

8.Fay's Counter Example Revisited.

Fay (1993) considers the simple case of estimating a bionomial proportion Θ. Suppose n_1 out of n sample cases have reported values, with missing data for the remaining cases and the proportion of responses, $r = n_1/n$, remains fixed as $n \rightarrow \infty$. Suppose further that the analyst attempts to make inferences about two subdomains which has been obtained by partitioning the original sample $n = n_b + n_b$, $n_1 = n_{1a} + n_{1b}$, $nr = n_a r_a + n_b r_b$ etc. Response rates r_a and r_b, the underlying proportion Θ, and the relative proportions $E(n_a)$ and $E(n_b)$ remain fixed as $n \rightarrow \infty$. The analysts forms separate estimates $\hat{\Theta}_{Ia}$ and $\hat{\Theta}_{Ib}$ for the two subdomains computed using only data from each respective subdomain. For example, $\hat{\Theta}_{Ia}$ would be computed only from the observed and imputed values in subdomain a. Thus, the estimates $\hat{\Theta}_{Ia}$ and $\hat{\Theta}_{Ib}$ do not exploit specific assumption that Θ was same in the two groups. Under this scenario, the imputed values must be obtained from each group seperately. Clearly, then there will be no covariance between $\hat{\Theta}_{Ia}$ and $\hat{\Theta}_{Ib}$ under any sampling plan of this paper. The expressions for the variances of these estimates are given in Sections 2,6 and 7 for various estimators and there is no contradiction. In fact, if one uses only W+B (instead of $W+\frac{m+1}{m}B$) as Fay does, the variance will be underestimated for Rubin-Schenker type of estimator.

We shall now consider the scenario in which Θ is assumed to be constant but the estimate of Θ are obtained from the two subdomains. Again, there will be no covariance between $\hat{\Theta}_{Ia}$ and $\hat{\Theta}_{Ib}$ for the estimators obtained by the methods of Section 2 and Section 7 but now there will be covariance between $\hat{\Theta}_{Ia}$ and $\hat{\Theta}_{Ib}$ by the method of Section 6, only if all the imputed values of the observed values are obtained at the same time using the mean of all the n_1 observations. This cannot be considered a valid method.

In conclusion, Fay's counter example does not apply to the methods of this paper.

9.Some Comments

In this paper, we have presented several imputed estimators of the popluation mean $\bar{Y}$. They are $\bar{y}_I^*$, $\bar{y}_{I.}^*$ and $\bar{y}_{I.}^{**}$. While the estimator $\bar{y}_I^*$ uses imputed values of the missing observations with variance $(1+1/n_1)S^2$, the estimators $\bar{y}_{I.}^*$ and $\bar{y}_{I.}^{**}$ use imputed values of the missing observations with variance $(1/m+1/n_1)S^2$ and $(1/m+2/n_1)$ respectively. Since they use average of m imputed values, the last two estimators, for m large, in effect are using $\bar{y}_1$ and $2\bar{y}_1$, respectively as the imputed values. The variance of the imputed values for the first estimator is closest to the variance of the observed values. I believe any imputed value should have this property. Thus, only $\bar{y}_I^*$ qualifies as a bonafide estimator of the population mean $\bar{Y}$ in the present situation in which the missing observations are being imputed. The bootstrap estimate of the variance of $\bar{y}_I^*$ and its bootstrap distribution appear attractive as they require only the record of the respondents. However, it remains to be seen which of the three methods of obtaining confidence intervals will provide a better coverage. The three methods are bootstrap and the two asymptotic distributions given in Section 6 and 7 respectively. A comparsion will be pursued in future communication.

References

Fay, R.E. (1993). Valid inferences from imputed survey data. *In Proceedings of the Section on Survey Research Methods, Washington, D.C. American Statistical Association.*

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. (1978). Multiple imputations in sample surveys - a phenomenological Bayesian approach to nonresponse. *In Proceedings of the section on Survey Research Methods, Washington, D.C. American Statistical Association.*

Rubin, D.B. and Schenker, N. (1986). Multiple imputations for interval estimation from single random samples with ignorable nonresponse. *J. Amer. Statist. Assoc* **81**, 366-374

Srivastava, M.S. and Carter, E.M. (1996). The maximun likelihood method for non-response in sample surveys. *Survey Methodology* **12**, 61-72.