The Koran Communications
in Statistics Vol. 11 No.2, 2004
pp. 399-411

A Naive Multiple Imputation Method for Ignorable Nonresponse[1]

Seung-Chun Lee[2]

Abstract

A common method of handling nonresponse in sample survey is to delete the cases, which may result in a substantial loss of cases. Thus in certain situation, it is of interest to create a complete set of sample values. In this case, a popular approach is to impute the missing values in the sample by the mean or the median of responders. The difficulty with this method which just replaces each missing value with a single imputed value is that inferences based on the completed dataset underestimate the precision of the inferential procedure. Various suggestions have been made to overcome the difficulty but they might not be appropriate for public-use files where the user has only limited information for about the reasons for nonresponse. In this note, a multiple imputation method is considered to create complete dataset which might be used for all possible inferential procedures without misleading or underestimating the precision.

Keywords : multiple imputation, ignorable nonresponse, Polya posterior

1. Introduction

Most sample surveys contain nonresponse, which become a problem when it comes time to analyze dataset if the user has only complete-data methods and has limited information about the reasons for nonresponse. Various suggestions have been made to handle the problem of nonresponse with or without much theoretical justification.

Recently imputation is becoming a standard approach for handling missing data problem. Note that Rubin (1987) argued that any sensible analysis must be based on an assumed model for the nonresponse. Since, in Bayesian point of view, the missing data can be treated as unknown quantities which are to be estimated together with parameters of interest on the assumed model, handling nonresponse is theoretically straightforward from the Bayesian viewpoint. A popular modeling of multivariate data relies on normality assumption and imputations based on normal models are reasonable in a range of scenarios, including cases

[1] This work was supported by Hanshin University Research Grant in 2004.

[2] Professor, Department of Statistics, Hanshin University, 411 Yangsan-dong, Osan, Korea, 447-791.
E-mail : seung@hanshin.ac.kr

where data are not normally distributed. See Rubin and Schenker (1986), and Schafer (1997). Thus theoretically the imputation for multivariate data is relatively easy to handle in some sense even though it might face various difficulties in practice.

On the other hand, in case of item nonresponse without auxiliary variables, only limited information can be incorporated to imputation, and because of the limitation, both Bayesian and frequentist methods of imputation often yield a single imputed value for nonresponse. However, it is well-known that inferences based on completed dataset with just a single imputed value underestimate the precision of the inferential procedure. See for example, Cohen (1996), Rao (1996), Ghosh and Meeden (1997), and Meeden (2000). To solve the problem, Rao and Shao (1992) considered a procedure that constructs a completed sample with a single imputed value but then used a jackknife-type estimate of variance to get a correct estimate of variance. See also Nusser, Carriquiry, Dodd, and Fuller (1996). They used unequal weights to solve the problem. However, those methods would not be appropriate in public-use files.

Another approach is multiple imputation. Note that an interesting approach for multiple imputation can be found in Meeden (2000). He provided a decision theoretic approach for multiple imputation which mainly focuses on the inference of the population mean. His imputation method have appealing features in that it yields an admissible solution and frequentist properties of a confidence interval for the mean based on the imputed values will be correct. Nonetheless one may criticize the discrepancy between the usual imputation methods. According to his method, each nonresponse is imputed by the mean of responders but the responders are adjusted so that the complete dataset would have desired properties. It is intuitively clear that one might want to keep the responders as they are and impute for the nonresponders in such a way that the distribution of complete dataset is as similar as that of responders.

In this note, we provide a multiple imputation method for the simplest situation where the characteristic of interest is univariate with and without an auxiliary variable. The created completed dataset would have desirable frequentist properties for the inferences of mean and median.

2. Theoretical Consideration for Multiple Imputation

Let y_i be the unknown value of some characteristic of interest. Suppose that a finite population U consists of N units and each element y_i in U is a real number. In the decision theoretic formulation, $y = (y_1, y_2, \cdots, y_N)$ is the unknown state of nature or parameter which belongs to N-dimensional Euclidean space. A subset s of $\{1, 2, \cdots, N\}$ is called a sample. Let s_r denote the subset of s that contains the labels of the units in the sample for which y was observed. Let n and n_r be the numbers of elements in s and s_r, respectively. For notational convenience, we assume that the first n units of the population are sampled and the responders are just the first n_r units of the sample. Thus $y_{n_r+1},\cdots,y_n$ are missing.

Let $\bar{y}(s_r)$ and $Var(y(s_r))$ denote the sample mean and the sample variance of the responders. In many cases the main interest may be the inference about the population mean. Since, assuming that the design is simple random sampling without replacement, the best frequentist estimates for the mean and variance of the population in this scenario are $\bar{y}(s_r)$ and $Var(y(s_r))$, respectively, one would estimate the population mean by $\bar{y}(s_r)$ and would construct the confidence interval for the population mean under the normality assumption.

Suppose that a created complete sample $a = (a_1, a_2, \cdots, a_n)$ is obtained by imputation. Let $\bar{a}$ and $Var(a)$ be the sample mean and the sample variance of a. From the frequentist viewpoint, it is reasonable that the created sample have the properties

$$\bar{a} = \bar{y}(s_r) \tag{2.1}$$

and

$$Var(a) = \frac{n(N-n_r)}{n_r(N-n)} Var(y(s_r)) \tag{2.2}$$

so that both inferences about the population mean based on the created complete sample and just on the responders would yield the same result.

In the decision theoretic formulation, a is an action. Given an appropriate loss function and the prior distribution of y, the best action in terms of the loss and the prior distribution can be obtained by finding the Bayes solution. If we assume the squared error loss with the two restrictions (2.1) and (2.2), the loss function can be defined as

$$L(a, y, s, s_r) = \frac{\sum_{i=1}^{n} (y_i - a_i)^2}{I\left(\bar{y}(s_r) = \bar{a}, \frac{n(N-n_r)}{n_r(N-n)} Var(y(s_r))\right)} \tag{2.3}$$

where I is the usual indicator function. Also Meeden and Vardeman (1991) provided a noninformative prior for this problem. Because of the limited information for this decision problem, seen and unseen are roughly exchangeable and we perfectly agree that Polya posterior is a reasonable posterior distribution of unseen based on seen. See for further details Meeden and Bryan (1996), and Ghosh and Meeden (1997).

Based on this belief and the loss function, the Bayes solution obtained by Meeden (2000) is given by

$$a_i = \begin{cases} \bar{y}(s_r) + \left(\frac{(n-1)n(N-n_r)}{(n_r-1)n_r(N-n)}\right)^{1/2} (y_i - \bar{y}(s_r)), & for\ i = 1, 2, \cdots, n_r \\ \bar{y}(s_r), & for\ i = n_r + 1, \cdots, n \end{cases} \tag{2.4}$$

The solution has a stepwise Bayes justification, and hence is admissible by the result of Lee and Meeden (1994).

However, the Bayes solution adjusts each responder in the sample. One might wish to keep the responders fixed and just adjust the imputed values for the nonresponders in such a way that one gets the correct mean and variance. That is, one might wish to impose the third constraint

$$a(s_r) = y(s_r) \tag{2.5}$$

where $a(s_r) = (a_1, \cdots, a_{n_r})$ and $y(s_r) = (y_1, \cdots, y_{n_r})$. Then the loss function is given by

$$L(a, y, s, s_r) = \frac{\sum_{i=1}^{n} (y_i - a_i)^2}{I\left(\bar{y}(s_r) = \bar{a}, \frac{n(N-n_r)}{n_r(N-n)} Var(y(s_r)), a(s_r) = y(s_r)\right)}. \tag{2.6}$$

Note that, under the Polya posterior, the expected value of unseen given seen is the sample mean of the responders. Thus minimizing the expected loss is equivalent to minimizing

$$\sum_{i=n_r+1}^{n} \left(\bar{y}(s_r) - a_i\right)^2 \tag{2.7}$$

subject to the constraints of (2.1), (2.2) and (2.5). Since the value of (2.7) is determined by (2.2), any solution satisfying (2.1), (2.2) and (2.5) is a Bayes with respect to the loss function (2.6).

One solution can be found in Cohen (1996). Although he did not consider the Bayesian approach, his solution with some obvious modifications indeed satisfies the constraints and hence is a Bayes for the problem. Even though Cohen provided a Bayes solution, he imputed the nonresponders by two different values. This might distort the actual distribution of sample. Also it is hard to recognize it as a multiple imputation method.

One might prefer to have the unique or the best solution here. However, it is not surprising that the Bayes solution is not unique because of the limited information. Many authors have suffered such difficulty. There might be two possible choices which can solve the problem. The first one is not to restrict the action space. Find the unique Bayes decision among all possible actions with minimal constraints of (2.1) and (2.2) as Meeden did. This strategy is successful but many statisticians would not agree. The second way is to impose more restrictions on the action space. To adopt the second strategy, more information on the decision problem may be required and hence may not be possible to obtain the unique or the best. However, we consider the restriction on the median of created dataset.

It is intuitively clear that the distribution of created sample should be as similar as that of responders so that the created sample does well for all possible inferential problem. For this purpose, we may create sample of size n whose mean as well as median agree with those of the responders. This may preserve the direction of skewness and we may hope that the created sample is good for other inferential problems. To incorporate this idea, we should set up the fourth constraint, namely

$$\tilde{y}(s_r) = \tilde{a} \tag{2.8}$$

where $\tilde{y}(s_r)$ and $\tilde{a}$ represent the median of $y(s_r)$ and a, respectively. Then the loss is defined to be

$$L(a,y,s,s_r) = \frac{\sum_{i=1}^{n} (y_i - a_i)^2}{I\left(\bar{y}(s_r) = \bar{a}, \frac{n(N-n_r)}{n_r(N-n)} Var(y(s_r)), a(s_r) = y(s_r), \tilde{a} = \tilde{y}(s_r)\right)} \tag{2.9}$$

Now the Bayes solution with respective to (2.9) should satisfy (2.1), (2.2), (2.5) and (2.8). Because the Bayes solution is not unique, various suggestions are possible. In subsequent sections we will provide a Bayes solution. Also we will demonstrate by simulation studies that the created sample works acceptably for the inference about the population median.

3. A Naive Multiple Imputation in Finite Population Sampling

In this section, we will find $a(s_r^c) = (a_{n_r+1}, \cdots, a_n)$ satisfying the four constraints. Because we are considering multiple imputation and the characteristic of interest y_i is assumed to be a real number, it might be preferable that $a_i \neq a_j$ for $i \neq j = n_r + 1, \cdots, n$. We will informally impose this restriction too.

Let $m = n - n_r$. The first constraint together with (2.5) implies that

$$\frac{1}{m} \sum_{i=n_r+1}^{n} a_i = \bar{y}(s_r) \tag{3.1}$$

Since the second constraint is given by

$$\frac{1}{n-1}\left(\sum_{i=1}^{n_r} (a_i - \bar{y}(s_r))^2 + \sum_{i=n_r+1}^{n} (a_i - \bar{y}(s_r))^2\right) = \frac{N-n_r}{N-n}\frac{n}{n_r}\frac{1}{n_r-1}\sum_{i=1}^{n_r} (y_i - \bar{y}(s_r))^2,$$

it can be written as:

$$\sum_{i=n_r+1}^{n} (a_i - \bar{y}(s_r))^2 = \left(\frac{N-n_r}{N-n}\frac{n}{n_r}\frac{n-1}{n_r-1} - 1\right)\sum_{i=1}^{n_r} (y_i - \bar{y}(s_r))^2. \tag{3.2}$$

Note that the right-hand side of (3.2) is fixed. We will denote the fixed value by Σ_1.

In what follows, we assume that $\tilde{y}(s_r)$ is uniquely defined. We are trying to create $a(s_r^c)$ satisfying (3.1) and (3.2), and the median of a, $\tilde{a}$ is equal to $\tilde{y}(s_r)$. To this end, two different situations shall be considered. At first we will consider the case that the number of missing observations is even. Then we may write $m = 2t$ for some integer t. For this case, the following action is considered:

$$a_{n_r+i} = \begin{cases} 2\bar{y}(s_r) - \tilde{y}(s_r) \pm \frac{i+1}{2}\lambda, & \text{if } i \text{ is odd} \\ \tilde{y}(s_r) \mp \frac{i}{2}\lambda, & \text{if } i \text{ is even} \end{cases} \tag{3.3}$$

where $D=\bar{y}(s_r)-\tilde{y}(s_r)$ and

$$\lambda=\frac{-t(t+1)D+\sqrt{t^2(t+1)^2D^2-\frac{t(t+1)(2t+1)}{3}(mD^2-\Sigma_1)}}{t(t+1)(2t+1)/3}.$$

The signs in (3.3) are determined by the inequality relationship between $\bar{y}(s_r)$ and $\tilde{y}(s_r)$. If $\tilde{y}(s_r)<\bar{y}(s_r)$, the odd cases take plus sign but the even cases take minus sign and vice versa.

Note that

$$\begin{aligned}\Sigma_1&=\left(\frac{N-n_r}{N-n}\frac{n}{n_r}\frac{n-1}{n_r-1}-1\right)\sum_{i=1}^{n_r}(y_i-\bar{y}(s_r))^2\\&>\frac{(n-n_r)(n+n_r-1)}{n_r(n_r-1)}\sum_{i=1}^{n_r}(y_i-\bar{y}(s_r))^2=\frac{2m(n_r+t-1/2)}{n_r(n_r-1)}\sum_{i=1}^{n_r}(y_i-\bar{y}(s_r))^2\\&>\frac{2m(n_r+t-1/2)}{n_r(n_r-1)}\frac{n_r}{2}D^2\\&>mD^2\end{aligned}\tag{3.4}$$

and λ is well-defined with property $\lambda>0$. Also it is easy to check that a_i's satisfy the constraint of (3.1) and

$$\sum_{i=n_r+1}^{n}(a_i-\bar{y}(s_r))^2=2tD^2+2t(t+1)D\lambda+\frac{t(t+1)(2t+1)}{3}\lambda^2.$$

Since λ is a solution to the equation

$$f(\Lambda)=2tD^2+2t(t+1)D\Lambda+\frac{t(t+1)(2t+1)}{3}\Lambda^2-\Sigma_1=0$$

the constraint of (3.2) is also satisfied by a_i's given in (3.3). Since $\lambda>0$, a_i's have the property that they are spreaded out from $\tilde{y}(s_r)$ and $2\bar{y}(s_r)-\tilde{y}(s_r)$ to either direction of tail sides. For example, if $\tilde{y}(s_r)<\bar{y}(s_r)$, all of the a_{n_r+i}'s with odd i are located on the left-hand side of $\tilde{y}(s_r)$ while with even i are on the right-hand side. Hence, the created sample has the unique median $\tilde{y}(s_r)$. Thus we have the following result.

Lemma 1. $a(s_r^c)$ given in (3.3) is well defined, and is a Bayes with respect to loss (2.9).

Suppose that the distribution of the responders is near symmetric. Then we might expect $\tilde{y}(s_r)\approx\bar{y}(s_r)$ and the imputation is done to preserve the symmetry. Suppose now the distribution is skewed to right so that $\tilde{y}(s_r)<\bar{y}(s_r)$. The imputation for the left-hand side of $\bar{y}\,(s_r)$ starts from $\tilde{y}\,(s_r)$ but starts from $2\bar{y}\,(s_r) - \tilde{y}\,(s_r)$ for right-hand side. Because the median is a robust alternative to the mean as a center of distribution, we might consider the median as the center for the skewed distribution. The imputation for right-side starts far from the center of distribution, and we might hope that the created sample preserve the skewness too.

Next, consider the case that $m = 2t + 1$ with $t \geq 1$. Define

$$a_{n_r+i} = \begin{cases} \tilde{y}\,(s_r), & \text{if } i = 1, \\ \bar{y}\,(s_r) + \dfrac{2t+1}{2t} D \pm \dfrac{i-1}{2}\lambda, & \text{if } i \text{ is odd}, i \neq 1 \\ \bar{y}\,(s_r) - \dfrac{2t-1}{2t} D \mp \dfrac{i}{2}\lambda, & \text{if } i \text{ is even} \end{cases} \tag{3.5}$$

where

$$\lambda = \frac{-t(t+1)D + \sqrt{t^2(t+1)^2 D^2 - \dfrac{t(t+1)(2t+1)}{3}\left[\left(m + \dfrac{1}{2t}\right)D^2 - \Sigma_1\right]}}{t(t+1)(2t+1)/3}.$$

(3.5) is well defined as before with properties (3.1) and (3.2), but theoretically we could not guarantee $\lambda > 0$. Note, however, mD^2 given in (3.4) is a loose lower limit of Σ_1. In practice we almost always have $\lambda > 0$ and $\tilde{a} = \tilde{y}\,(s_r)$.

We have provided a multiple imputation for a simple situation in the hope that the created sample is useful for various inferential problems such as the location estimation. We believe that it is useful for other inferential problem as well, because the created sample might preserve the skewness so that the distribution of created sample is not much different from that of responders.

4. Imputation with an Auxiliary Variable

Suppose that an auxiliary variable x_i, which is closely related to y_i, is observed for every unit in the sample so that one can employ the ratio type estimation for the population quantities, and we are interested in estimating the population mean. The ratio estimator of the population mean of y is $R_{s_r}\bar{x}$, where $R_{s_r} = \bar{y}\,(s_r)/\bar{x}\,(s_r)$, $\bar{x}\,(s_r)$ is the mean of auxiliary variable of the responders and $\bar{x}$ is the population mean of auxiliary variable x. It is well-known, that the variance of ratio estimator is estimated by

$$Var(R_{s_r}\bar{x}) = \frac{1-f_r}{n_r(n_r-1)} \sum_{i=1}^{n_r} (y_i - R_{s_r}x_i)^2$$

where $f_r = n_r/N$. See Cochran (1977). The ratio estimate of the mean based on the complete data a is just $R_a\bar{x}$ where $R_a = \bar{a}/\bar{x}\,(s)$ and $\bar{x}\,(s)$ is the mean of the auxiliary variable of the sample. As before, the frequentist confidence interval based on the responders and the created complete dataset will agree if and only if

$$R_{s_r} = R_a \tag{4.1}$$

and

$$Var(R_a\bar{x}) = Var(R_{s_r}\bar{x}). \tag{4.2}$$

where

$$Var(R_a\bar{x}) = \frac{1-f}{n(n-1)}\sum_{i=1}^{n} (a_i - R_a x_i)^2$$

and $f = n/N$. The constraint of (2.5) is also desirable, because we do not want to adjust the responders. It is easy to check that two constraints (4.1) and (4.2) together with (2.5) can be written as:

$$\bar{a}(s_r^c) = R_{s_r}\bar{x}(s_r^c) \tag{4.3}$$

and

$$\sum_{i=n_r+1}^{n} (a_i - R_{s_r}x_i)^2 = \left(\frac{1-f_r}{1-f}\frac{n}{n_r}\frac{n-1}{n_r-1} - 1\right)\sum_{i=1}^{n_r} (y_i - R_{s_r}x_i)^2. \tag{4.4}$$

Thus if we assume the squared error loss, we need to minimize the expected value of $\sum_{i=n_r+1}^{n} (y_i - a_i)^2$ subject to (4.3) and (4.4).

The Bayes solution can be obtained by defining an appropriate prior and the conditional expectation of unseen given seen. Various noninformative priors are possible. For example, Meeden (2000) argued that the ratios $r_i = y_i/x_i$ are roughly exchangeable provided that both y_i's and x_i's are all greater than zero and x_i's do not differ too much in size. Then the Polya posterior can be applied to the ratios $r_i = y_i/x_i$. The conditional expectation of $y_i/x_i, i \in s_r^c$ given $y(s_r)$ and $x(s)$ is just the sample mean of observed ratios $\overline{r_{s_r}} = n_r^{-1}\sum_{i=1}^{n_r}(y_i/x_i)$. This assertion was supported by Basu (1971) and Royall (1970). If we believe the ratios are roughly exchangeable, then the decision problem can be summarized to minimize $\sum_{i=n_r+1}^{n} (\overline{r_{s_r}}x_i - a_i)^2$ subject to (4.3) and (4.4). The solution of this problem was given by Meeden. Although he did not consider the constraint of (2.5), he provided the solution to this problem.

Another possibility, which we might prefer, is that $(y_i - rx_i)$'s are exchangeable. Formally we might assume that the population values (y_i, x_i) are random samples from a super population in which

$$y_i = rx_i + \epsilon_i \tag{4.5}$$

where ϵ_i's are independent of x_i's with $E(\epsilon_i) = 0$. Note that the method of moments produces the ratio estimator on the model and under certain conditions it can be the best linear unbiased estimator. Also it is intuitively clear that ϵ_i's might be roughly exchangeable without further assumptions. Thus the Polya posterior can be applied to ϵ_i's. That is, for each $i \in s_r^c$, we have

$$E(\epsilon_i | y(s_r), x(s), r) = \bar{\epsilon}(s_r) = \bar{y}(s_r) - r\bar{x}(s_r)$$

or equivalently

$$E(y_i | y(s_r), x(s), r) = \bar{y}(s_r) - r(\bar{x}(s_r) - x_i). \tag{4.6}$$

In Bayesian paradigm, it needs to define the prior distribution of r. Alternatively the empirical Bayesian method can be applied to here. Since ϵ_i's are assumed to be exchangeable, the least square estimator

$$R_{LS} = \sum_{i=1}^{n_r} x_i y_i / \sum_{i=1}^{n_r} x_i^2$$

is the best linear unbiased estimator. Thus we will replace r in (4.6) by R_{LS}. Then the decision problem is now to minimize

$$\sum_{i=n_r+1}^{n} (y_i^* - a_i)^2 \tag{4.7}$$

subject to (4.3) and (4.4) where $y_i^* = \bar{y}(s_r) - R_{LS}(\bar{x}(s_r) - x_i)$.

Theorem 1. Let Σ_2 be the value of the right-hand side of equation (4.4). Suppose $R_{LS} \neq R_{s_r}$ and the number of missing observations is greater than or equal to 2. Then for the problem of minimizing (4.7) with respect to a_i's subject to the constraints of (4.3) and (4.4), the solution is given by

$$a_i = R_{s_r} x_i + \delta(x_i - \bar{x}(s_r^c)),\ i = n_r + 1, \cdots, n$$

where δ is defined to be

$$\delta = \begin{cases} \eta & \text{if } \sum_{i=n_r+1}^{n} (R_{s_r} x_i - y_i^*)(x_i - \bar{x}(s_r^c)) < 0 \\ -\eta & \text{if } \sum_{i=n_r+1}^{n} (R_{s_r} x_i - y_i^*)(x_i - \bar{x}(s_r^c)) > 0 \end{cases}$$

with $\eta = \left(\Sigma_2 / \sum_{i=n_r+1}^{n} (x_i - \bar{x}(s_r^c))^2 \right)^{1/2}$.

Proof: It can be shown by a standard method using Lagrange multipliers. Let

$$L = \sum_{i=n_r+1}^{n} (a_i - y_i^*)^2 - 2\lambda_1 \left(\sum_{i=n_r+1}^{n} a_i - R_{s_r} \sum_{i=n_r+1}^{n} x_i \right) - \lambda_2 \left(\sum_{i=n_r+1}^{n} (a_i - \mu_i)^2 - \Sigma_2 \right)$$

where $\mu_i = R_{s_r} x_i$. Taking the partial derivative of L with respect to a_i and setting it equal to zero, we have

$$a_i = \frac{1}{1-\lambda_2}\left(y_i^* + \lambda_1 - \lambda_2\mu_i\right),\ i = n_r + 1,\cdots,n \tag{4.8}$$

and hence

$$\sum_{i=n_r+1}^{n} a_i = \frac{1}{1-\lambda_2}\left(\sum_{i=n_r+1}^{n} y_i^* + m\lambda_1 - m\lambda_2\bar{\mu}\right),$$

where $\bar{\mu} = \frac{1}{m}\sum_{i=n_r+1}^{n}\mu_i$. Note that, by (4.3), we must have $\bar{a}\ (s_r^c) = \bar{\mu}$ and the above equation can be written as:

$$\frac{\bar{a}\ (s_r^c)}{1-\lambda_2} = \frac{\overline{y}^* + \lambda_1}{1-\lambda_2}$$

where $\bar{y}^* = \frac{1}{m}\sum_{i=n_r+1}^{n} y_i^*$. This shows $\lambda_1 = \bar{a}\ (s_r^c) - \bar{y}^*$. Substituting λ_1 in (4.8) by $\bar{a}\ (s_r^c) - \bar{y}^*$ and subtract μ_i from both sides of equation (4.8), we get, for $i = n_r + 1,\cdots,n$,

$$\begin{aligned} a_i - \mu_i &= \frac{1}{1-\lambda_2}\left(\left(y_i^* - \bar{y}^*\right) - \left(\mu_i - \bar{\mu}\right)\right) \\ &= \frac{1}{1-\lambda_2}\left(R_{LS} - R_{s_r}\right)\left(x_i - \bar{x}\ (s_r^c)\right), \end{aligned} \tag{4.9}$$

and

$$\sum_{i=n_r+1}^{n}\left(a_i - \mu_i\right)^2 = \left(\frac{1}{1-\lambda_2}\right)^2\left(R_{LS} - R_{s_r}\right)^2\sum\left(x_i - \bar{x}\ (s_r^c)\right)^2.$$

Because of the constraint (4.4), it must be

$$\frac{1}{1-\lambda_2} = \pm\left[\frac{\Sigma_2}{\left(R_{LS} - R_{s_r}\right)^2\sum\left(x_i - \bar{x}\ (s_r^c)\right)^2}\right]^{1/2}.$$

Substituting above quantity into (4.9), we have the desired result.

Theorem 1 provides the method for creating complete dataset which have the correct frequentist properties under repeated sampling. It has the decision theoretic justification so that both Bayesian and frequentist would agree. The similar result can be found in Meeden (2000) under rough exchangeability assumption for the ratios y_i/x_i, which might be questionable in many situations. In fact, the assumption implies that the variance of y_i is proportional to x_i^2. More widely used model might be that the variance of y_i does not depend on x_i such as the usual normal model.

The created sample is selected with the goal of making inference about the population mean. Is it doing well for other inferential problems? In section 5, we will give an answer for the question.

5. Simulation Study and Conclusion

For the case of no auxiliary variable, we considered two artificial populations of size 500, sampled from a normal distribution with mean zero and variance 5, and a Gamma distribution with shape parameter 5 and scale parameter 1. After taking simple random sample of size 50 from each population, which are denoted by "All" in Table 1, some of sample values were removed randomly from "All," denotes it by "Res." The values of n_r considered here were 30, 40 and 45. "Nav" in the table represents the created complete sample by (3.3) or (3.5). A complete dataset, created by (2.4) was also obtained for comparison and denoted by "Bay." From each of the four dataset, the lower and the upper quartiles, and the median are calculated as estimates of the corresponding population quantities. After 1000 replications of this process, we obtained the mean absolute errors of the estimates. We also computed the MacKinnon 95% confidence interval for the population median and the average length of the intervals with coverage rate. Some simulation results are given in Table 1.

It can be observed that "Nav" would tend to give slightly smaller values of the lower quartile than those of "All" and "Rev" but give little bit larger values of the upper quartile for both symmetric and skewed populations. This kind of phenomenon is expected because the imputed values should be more dispersed than the responders because of the constraint (2.2). But the average length of confidence intervals for the median are comparable to those of "Res", while

Table 1 : Averages of the median, the lower and the upper quartiles for estimating the population quantities and their mean absolute errors, and the average length, divided by the standard deviation of population, of 95% MacKinnon confidence of intervals along with their coverage probabilities.

		Gamma(5,1)								N(0,5)							
No of Resp	Type	Ave. quantile			MAE			Length	Prob	Ave. quantile			MAE			Length	Prob
		q25	q50	q75	q25	q50	975			q25	q50	q75	q25	q50	q75		
	All	3.57	4.62	6.07	0.26	0.22	0.31	0.56	0.97	-1.50	-0.028	1.47	0.33	0.29	0.34	0.69	0.97
	Res	3.60	4.65	6.07	0.34	0.31	0.42	0.83	0.97	-1.46	-0.013	1.45	0.43	0.39	0.46	0.94	0.97
30	Nav	3.43	4.65	6.46	0.29	0.31	0.49	0.85	0.97	-1.65	-0.013	1.64	0.36	0.39	0.39	0.91	0.98
	Bay	3.92	4.98	5.20	0.49	0.44	0.89	0.03	0.06	-0.69	-0.000	0.66	0.96	0.33	0.93	0.002	0.003
	Pop	3.50	4.59	6.06						-1.61	-0.038	1.53					
	All	3.45	4.70	6.39	0.26	0.26	0.43	0.64	0.95	-1.44	-0.07	1.55	0.31	0.32	0.35	0.70	0.97
	Res	3.47	4.70	6.38	0.29	0.30	0.49	0.82	0.98	-1.44	-0.06	1.52	0.34	0.36	0.40	0.83	0.97
40	Nav	3.27	4.70	6.78	0.28	0.30	0.54	0.80	0.98	-1.51	-0.06	1.61	0.31	0.36	0.35	0.79	0.97
	Bay	3.44	5.08	6.04	0.30	0.41	0.63	0.38	0.76	-1.33	0.04	1.34	0.36	0.30	0.45	0.30	0.56
	Pop	3.38	4.72	6.48						-1.46	-0.10	1.56					
	All	3.56	4.89	6.35	0.24	0.25	0.37	0.64	0.96	-1.43	-0.11	1.53	0.25	0.27	0.35	0.69	0.96
	Res	3.57	4.89	6.35	0.27	0.28	0.40	0.73	0.97	-1.42	-0.11	1.53	0.27	0.30	0.38	0.79	0.97
45	Nav	3.46	4.89	6.60	0.25	0.28	0.45	0.73	0.97	-1.56	-0.11	1.70	0.28	0.30	0.36	0.79	0.97
	Bay	3.56	5.11	6.22	0.31	0.34	0.41	0.48	0.85	-1.41	0.03	1.45	0.29	0.28	0.43	0.51	0.83
	Pop	3.52	4.85	6.38						-1.46	-0.12	1.61					

Table 2 : Averages of the median, the lower and the upper quartiles for estimating the population quantities and their mean absolute errors, and the average correlation coefficents

		Popl.							Pop2.						
No of								Ave.							Ave.
Resp	Type	Ave. quantile			MAE			Corr.	Ave. quantile			MAE			Corr.
		q25	q50	q75	q25	q50	q75	corr	q25	q50	q75	q25	q50	q75	corr
30	All	25.03	29.14	34.18	0.77	0.86	1.20	0.95	23.48	28.64	35.66	1.14	1.13	1.66	0.71
	Res	25.11	29.17	34.00	1.02	1.13	1.54	0.94	23.54	28.68	35.52	1.48	1.53	2.19	0.70
	Rls	25.49	29.03	33.56	1.29	0.91	1.61	0.92	23.78	29.35	35.60	1.79	1.59	2.00	0.65
	Rme	25.63	29.20	33.64	1.36	0.91	1.66	0.92	23.83	29.37	35.54	1.80	1.62	1.99	0.65
40	All	25.00	29.12	34.16	0.75	0.85	1.20	0.95	23.36	28.65	35.76	1.11	1.19	1.70	0.71
	Res	25.05	29.11	34.09	0.86	0.97	1.38	0.94	23.35	28.59	35.64	1.27	1.31	1.90	0.70
	Rls	25.16	29.11	33.99	0.89	0.87	1.33	0.93	23.59	29.01	35.38	1.26	1.31	1.78	0.66
	Rme	25.25	29.12	33.91	0.90	0.88	1.33	0.93	23.61	29.00	35.37	1.26	1.32	1.76	0.66
45	All	25.00	29.13	34.23	0.81	0.88	1.20	0.95	23.53	28.74	35.87	1.15	1.18	1.80	0.71
	Res	25.03	29.16	34.20	0.86	0.95	1.28	0.95	23.55	28.73	35.86	1.19	1.25	1.90	0.71
	Rls	25.12	29.16	34.22	0.81	0.91	1.22	0.94	23.59	28.94	35.83	1.15	1.23	1.74	0.67
	Rme	25.11	29.15	34.14	0.81	0.91	1.24	0.94	23.60	28.94	35.83	1.14	1.24	1.74	0.67
	Pop	24.85	29.06	34.53				0.95	23.19	28.67	36.05				0.71

"Bay" does not give proper results. Thus we might conclude that "Nav" does not distort the distribution of responders too much.

Next, for the case of multiple imputation with an auxiliary variable x, we also considered two artificial populations of size 500. Each observation of auxiliary variable was obtained by setting to equal to 5 plus an observation from a gamma distribution with shape parameter 5 and scale parameter 1. Then two populations were created from normal distributions with mean $3x_i$ and variance 5 and $5x_i^2$, respectively. Note that ϵ's given in (4.5) are exchangeable in the first population, while the ratio y/x are exchangeable in the second population. As before, after taking random samples of size 50 from each population, some values of sample were randomly removed. Then the removed values were imputed by theorem 1, denoted by "Rls" and the result of Meeden (2000), denoted by "Rme." The sample quartiles were calculated from each of four dataset. After 1000 replications, we obtained the average of sample quartiles and their mean absolute error (MAE) with average sample correlation coefficient. The result is shown in Table 2.

In both populations the mean absolute errors of Rls and Rme are slight larger than those of Res when the number of responders is 30, but both imputation give better results when n_r is 40 or 45. Thus it seems that both imputation methods would be appropriate when the number of nonresponders is not large compared with the sample size. Also it can be observed that Rls gives slightly better results than Rme in the first population but other simulation results which are not appealed in here show that both methods are relatively competitive so that one would have a difficulty to choose the right method. In short, both methods provide reasonable imputation for most configurations.

References

[1] Basu, D. (1971). An Essay on the Logical Foundations of Survey, Part One, in foundations of Statistical Inference, eds. V. P. Godambe and D. A. Sproutt, Tortonto: Holt, Rinehardt and Winston.

[2] Cochran, W. G. (1977). *Sampling Techniques*, 3rd. ed. Wiley, New York.

[3] Cohen, M. P. (1996). A new approach to imputation, *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 293-298.

[4] Ghosh, M. and Meeden, G. (1997). *Bayesian Methods for Finite Population*, Chapman & Hall, London.

[5] Lee, S.-C., and Meeden, G. (1994). A minimal complete class theorem for decision problems where the parameter space contains only finite many points, *Metrika*, vol 41, 227-232.

[6] Meeden, G. (2000). A decision theoretic approach to imputation in finite population sampling, *Journal of American Statistical Association*, Vol. 95, 586-595.

[7] Meeden, G., and Bryan, M. (1996). An approach to the problem of nonresponse in sample survey using Polyar posterior, In Bayesian Analysis in Statistics and Econometrics Essays in Horner of Arnold Zeller, Wiley New York, pp 423-432.

[8] Meede, G., and Vardeman, S. (1991). A noninformative Bayesian approach to interval estimation in finite population sampling, *Journal of American Statistical Association*, Vol 86, 972-980.

[9] Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996). A semiparametric transformation approach to estimating usual intake distributions, *Journal of American Statistical Association*, Vol 91, 1440-1449.

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

[11] Rao, J. N. K. and Shao, J. (1992). Jackknife variance estimation with survey data under hot deck imputation, *Biometrika*, Vol 57, 377-387.

[12] Royall, R. M. (1970). On finite-population sampling theory under certain linear regression models, *Biometrika*, Vol. 57, 377-387.

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

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

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

[Received February 2004, Accepted July 2004]