STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, December 2018, pp 526–535.
Published online in International Academic Press (www.IAPress.org)

Improved Mean Methods of Imputation

Choukri Mohamed [1], Stephen A. Sedory [2], Sarjinder Singh [2,*]

[1] *Moody School, Corpus Christi, TX, USA*
[2] *Department of Mathematics, Texas A&M University Kingsville, Kingsville, USA*

Abstract Replacing missing values of a variable with the mean of the non-missing values is a simple and natural way to impute values fortunately in the case where data is missing completely at random. Following a short review of this method we consider three possible improvements, one called the shrinkage method, a second called the weighted interval method, and a third called the known variance method. Estimates of the population mean obtained from each of these methods are compared to the mean method both analytically and by means of numerical examples.

Keywords Estimation of mean, imputation, interval estimation, missing data, and shrinkage estimation.

AMS 2010 subject classifications 62D05

DOI: 10.19139/soic.v6i4.281

1. Introduction

The problem of missing data is a very common issue in real surveys as well as in experimental studies and may arise for any number of reasons. If the data is *missing completely at random* (MCAR) things can be very hard to handle in practice, particularly when no auxiliary information is available. Examples of data that is MCAR are laboratory sample that is dropped, or a questionnaire that is lost in a mail survey and so resulting observations became missing. Another recent example would be plant-studies that were affected by hurricane Mathew in Florida. MCAR is a more difficult situation than *missing at random* (MAR), or cases where there is deliberate non-response. More efforts are required to develop better imputation methods for MCAR where additional information could be little helpful or could lead to wrong prediction. In addition, in some situations, a cheap and fast method is an imputation technique which is being frequently used to substitute for missing values in order to improve inferential properties of an estimator. For more detail on the concept of MCAR, one could refer to [4, 11, 3].

A general problem with many frequently employed imputation methods is an introduction of bias and an increase in variance of the resultant estimators. A search for unbiased estimators under imputation should therefore be of interest. For details on the history of imputation methods, one could refer to [7], where one finds that several imputations methods such as hot deck, nearest neighbourhood, cold deck, warm deck, ratio method, regression method and power method of imputation have been proposed. These methods either make use of a "deck" from a past or the present survey, or make use of auxiliary information. A critical review in [7] shows that efforts have not been made to improve the mean method of imputation in the absence of auxiliary information. We also conclude that it is not an easy task to improve the mean method of imputation in the absence of any additional information at hand. In this paper we suggest three new imputing methods in the absence of auxiliary information. Consider a

*Correspondence to: Sarjinder Singh is with Department of Mathematics, Texas A&M University- Kingsville, Kingsville, TX 78363 (E-mail: kuss2008@tamuk.edu)

ISSN 2310-5070 (online) ISSN 2311-004X (print)

finite population, Ω, of N units as $\Omega = \{\omega_1, \omega_2, \ldots, \omega_i, \ldots, \omega_N\}$, Let y be the variable of interest in the population, and y_i the value of y for the unit i.

Let

$$\bar{Y} = \frac{1}{N}\sum_{i\in\Omega} y_i \tag{1}$$

be the true population mean of the study variable y. Assume a simple random without replacement sample (SRSWOR), s, of size n is drawn from the population Ω. Assume it was possible only to collect information on r units out of the sampled n units from the population. In particular, let the set of r responding units be denoted by $A \subseteq s$ and that of $(n-r)$ non-responding units be denoted by A^c. For every unit $i \in A$, the value of y_i is observed and the for that units $i \in A^c$, the value of y_i is missing. Thus the sample data values have the following structure:

$$y_{\bullet i} = \begin{cases} y_i & \text{if } i \in A \\ \text{missing} & \text{if } i \in A^c \end{cases} \tag{2}$$

Now the first choice is to forget or drop the missing $(n-r)$ data values in the set A^c from the sample s of n data values and consider an estimator of the population mean $\bar{Y}$ as:

$$\bar{y}_r = \frac{1}{r}\sum_{i\in A} y_i \tag{3}$$

which is the sample mean of the r values in the responding set A. Assuming the data is *missing completely at random* (MCAR), then applying the concept of two-phase sampling as given in [1], it is easy to verify that the sample mean $\bar{y}_r$ in (3) is an unbiased estimator of the population mean $\bar{Y}$ with conditional variance, for a given value of r, given by:

$$V(\bar{y}_r) = (\frac{1}{r} - \frac{1}{N})S_y^2 \tag{4}$$

where $S_y^2 = \frac{1}{N-1}\sum_{i\in\Omega}(y_i - \bar{Y})^2$ is the population mean squared error (or population variance) for the study variable.

Now consider imputing the missing data values by the mean method of imputation as follows:

$$\hat{y}_{\bullet i} = \begin{cases} y_i & \text{if } i \in A \\ \bar{y}_r & \text{if } i \in A^c \end{cases} \tag{5}$$

that is all the missing values are replaced by the sample mean $\bar{y}_r$ of the responding values. Now consider the point estimator of population mean, given by:

$$\bar{y}_{point} = \frac{1}{n}\sum_{i\in s}\hat{y}_{\bullet i} \tag{6}$$

On using (5) in (6) we have

$$\bar{y}_{point} = \frac{1}{n}\sum_{i\in s}\hat{y}_{\bullet i} = \frac{1}{n}[\sum_{i\in A}\hat{y}_{\bullet i} + \sum_{i\in A^c}\hat{y}_{\bullet i}] = \frac{1}{n}[\sum_{i\in A}y_i + \sum_{i\in A^c}\bar{y}_r] = \frac{1}{n}[r\bar{y}_r + (n-r)\bar{y}_r] = \bar{y}_r \tag{7}$$

From (7) and (3), the mean method of imputation leads to the same estimator($= \bar{y}_r$) of the population mean $\bar{Y}$ with the same variance as given in (4). Thus, although the mean method of imputation is helpful in completing the missing values in a sample, it does not provide any additional benefit in drawing inferences from the results.

In the following sections, we propose new imputing techniques which fill the missing data values with more accurate predicted values, and which lead to more efficient estimators of the population mean under various situations. In section 2, we introduce a new shrinkage estimation technique. In section 3, we introduce a new weighted interval method of imputation and in section 4 we introduce a new method of imputation when the population variance of the study variable is known.

2. SHRINKAGE METHOD OF IMPUTATION

Following [12, 13], we propose the following shrinkage method of imputation given by

$$\hat{y}_{\bullet i} = \begin{cases} y_i & \text{if } i \in A \\ \lambda \bar{y}_r & \text{if } i \in A^c \end{cases} \tag{8}$$

where λ is called the shrinkage parameter and which is to be determined based on some criterion, such as that the resultant estimator has minimum mean squared error. Under the proposed shrinkage method of imputation, the point estimator of the population mean $\bar{y}$ is given by

$$\bar{y}_{shrink} = (\frac{r}{n} + \lambda(1 - \frac{r}{n}))\bar{y}_r \tag{9}$$

The percentage relative bias in the proposed shrinkage estimator $\bar{y}_{shrink}$ is given by

$$RB(\bar{y}_{shrink}) = -(1 - \frac{r}{n})(1 - \lambda) \times 100\% \tag{10}$$

It may be worth pointing out that the value of percent relative bias is free from the value of the population mean. The mean squared error of the proposed shrinkage estimator $\bar{y}_{shrink}$ is given by

$$MSE(\bar{y}_{shrink}) = [\frac{r}{n} + \lambda(1 - \frac{r}{n})]^2(\frac{1}{r} - \frac{1}{N})S_y^2 + [(1 - \frac{r}{n})(1 - \lambda)\bar{Y}]^2 \tag{11}$$

The optimum value of λ which minimises the mean squared error of the proposed shrinkage estimator $\bar{y}_{shrink}$ is given by

$$\lambda = \frac{(1 - \frac{r}{n}) - \frac{r}{n}(\frac{1}{r} - \frac{1}{N})C_y^2}{(1 - \frac{r}{n})[1 + (\frac{1}{r} - \frac{1}{N})C_y^2]} \tag{12}$$

where $C_y = S_y/\bar{Y}$ denotes the value of the coefficient of variation of the study variable. The resultant minimum mean squared error of the shrinkage estimator $\bar{y}_{shrink}$ is given by

$$\min MSE(\bar{y}_{shrink}) = \frac{(\frac{1}{r} - \frac{1}{N})S_y^2}{1 + (\frac{1}{r} - \frac{1}{N})C_y^2} \tag{13}$$

The optimum percent relative bias in the proposed shrinkage estimator $\bar{y}_{shrink}$ is given by

$$RB(\bar{y}_{shrink})_o = -\frac{(\frac{1}{r} - \frac{1}{N})C_y^2}{\{1 + (\frac{1}{r} - \frac{1}{N})C_y^2\}} \times 100\% \tag{14}$$

The optimum percentage relative efficiency of the proposed shrinkage method of imputation over the mean method of imputation is given by

$$RE(\bar{y}_{shrink})_o = \{1 + (\frac{1}{r} - \frac{1}{N})C_y^2\} \times 100\% \tag{15}$$

Note that the percent relative efficiency is an increasing function of the value of coefficient of variation C_y and the difference between $\frac{1}{r}$ and $\frac{1}{N}$. The proposed shrinkage method of imputation obviously will perform better than the mean method of imputation in case the value of response r is low and the value of the coefficient of variation C_y is large. It seems that the proposed shrinkage method of imputation will be very useful when it is very expensive to obtain responses from the respondents and variation among the units in the population is large. Note that the percent relative efficiency $RE(\bar{y}_{shrink})_o$ value is a function of r, N and C_y, and the percent relative bias $RB(\bar{y}_{shrink})$ is a function of r, N, C_y^2 and S_y^2. We investigated the behaviour of $RE(\bar{y}_{shrink})_o$, $RB(\bar{y}_{shrink})_o$ and optimum value of λ for various choices of the parameter. In the study, we considered nine different populations with different values of the coefficient of variation C_y equal to $0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8$ and 0.9 each of size $N = 1000$ units. From each population we considered a sample of size $n = 15$ units, and assumed several different values of number of respondents r equal to $3, 5, 7$ and 9. It is noted that for $r = 3$, as the value of C_y varies between 0.1 and 0.5, the absolute value of the optimum relative bias $RB(\bar{y}_{shrink})_o$ remains less than 10% [1] and the percent relative efficiency $RE(\bar{y}_{shrink})_o$ varies between 100.33% and 108.31%; for $r = 5$ as the value of C_y varies between 0.1 and 0.7, the absolute value of the optimum relative bias $RB(\bar{y}_{shrink})_o$ remains less than 10% and the percent relative efficiency $RE(\bar{y}_{shrink})_o$ varies between 100.20% and 109.75%; for $r = 7$ as the value of C_y varies between 0.1 and 0.8, the absolute value of the optimum relative bias $RB(\bar{y}_{shrink})_o$ remains less than 10% and the percent relative efficiency $RE(\bar{y}_{shrink})_o$ varies between 100.14% and 109.08%; and for $r = 9$ as the value of C_y varies between 0.1 and 0.9, the absolute value of the optimum relative bias $RB(\bar{y}_{shrink})_o$ remains less than 10% and the percent relative efficiency $RE(\bar{y}_{shrink})_o$ varies between 100.11% and 108.92%. From the analysis, we conclude that if a good guess at or the true value of, the coefficient of variation C_y of the study variable is available, then it can be used to obtain better imputed values than the mean method of imputation. It may be worth mentioning that if the value of λ is unknown, it can be estimated by using a consistent estimator given by

$$\hat{\lambda} = \frac{(1 - \frac{r}{n}) - \frac{r}{n}(\frac{1}{r} - \frac{1}{N})\hat{C}_y^2}{(1 - \frac{r}{n})[1 + (\frac{1}{r} - \frac{1}{N})\hat{C}_y^2]} \tag{16}$$

where $\hat{C}_y = s_{y(r)}/\bar{y}_r$ is a consistent estimator of C_y with $s_{y(r)}^2 = (r-1)^{-1}\sum_{i \in A}(y_i - \bar{y}_r)^2$.

In the next section, we introduce a new interval method of imputation by making use of the sample standard deviation of the responding units in addition to the sample mean. The standard interval method of estimation of population mean available in almost all introductory statistics text books motivated the authors to think whether such a method can be constructed to impute a pair of values for each respondent instead of a single value.

3. A NEW WEIGHTED INTERVAL METHOD OF IMPUTATION

In this section, we suggest a new method of imputation by using a weighted interval method of estimation given by

$$\hat{\hat{y}}_{\bullet i} = \begin{cases} y_i & \text{if } i \in A \\ \alpha_1(\bar{y}_r - \sqrt{n}\frac{s_{y(r)}}{(n-r)}) + \alpha_2(\bar{y}_r + \sqrt{n}\frac{s_{y(r)}}{(n-r)}) & \text{if } i \in A^c \end{cases} \tag{17}$$

where α_1 and α_2 are real constants such that $\alpha_1 + \alpha_2 = 1$. In (17) two values are imputed for each non-respondent, one value to the left of the sample mean and another to the right of the sample mean. If one decides to choose $\alpha_1 = \alpha_2 = \frac{1}{2}$, then the imputation method in (17) reduces to the mean method of imputation. Thus the question is to decide about the possible best choice of the values of α_1 and α_2. One possibility is to determine the values of α_1 and α_2 such that the mean squared error of the resultant estimator is minimum. The point estimator (6) under the weighted interval method of imputation in (17) becomes

$$\begin{aligned} \bar{y}_{point} &= \frac{1}{n}\sum_{i \in s}\hat{y}_{\bullet i} \\ &= \frac{1}{n}[\sum_{i \in A}y_i + \sum_{i \in A^c}\{\alpha_1(\bar{y}_r - \sqrt{n}\frac{s_{y(r)}}{(n-r)}) + \alpha_2(\bar{y}_r + \sqrt{n}\frac{s_{y(r)}}{(n-r)})\}] \\ &= \bar{y}_r + (\alpha_2 - \alpha_1)\frac{s_{y(r)}}{\sqrt{n}} = \bar{y}_{w(int)} \quad (say) \end{aligned} \tag{18}$$

In order to study the asymptotic properties of the newly proposed estimator $\bar{y}_{w(int)}$ based on the weighted interval method of imputation, we find the bias, to the first order of approximation, is given by

$$B(\bar{y}_{w(int)}) = \frac{(\alpha_2 - \alpha_1)}{\sqrt{n}}S_y[1 - \frac{1}{8}(\frac{1}{r} - \frac{1}{N})(\frac{\mu_4}{S_y^4} - 1)] \tag{19}$$

where $\bar{Y} = N^{-1} \sum\limits_{i \in \Omega} y_i$,$S_y^2 = (N-1)^{-1} \sum\limits_{i \in \Omega} (y_i - \bar{Y})^2$, $\mu_3 = (N-1)^{-1} \sum\limits_{i \in \Omega} (y_i - \bar{Y})^3$ and $\mu_4 = (N-1)^{-1} \sum\limits_{i \in \Omega} (y_i - \bar{Y})^4$. The mean squared error of the estimator $\bar{y}_{w(int)}$ is given by

$$MSE(\bar{y}_{w(int)}) \approx (\frac{1}{r} - \frac{1}{N})S_y^2 + (\alpha_2 - \alpha_1)^2 \frac{S_y^2}{n} + (\alpha_2 - \alpha_1)\frac{1}{\sqrt{n}}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3}{S_y} \tag{20}$$

which will be minimum if:

$$(\alpha_2 - \alpha_1) = -\frac{\sqrt{n}}{2}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3}{S_y^3}$$

The minimum mean squared error, to the first order of approximation, of the proposed estimator $\bar{y}_{w(int)}$ is given by

$$MSE(\bar{y}_{w(int)}) = (\frac{1}{r} - \frac{1}{N})S_y^2[1 - \frac{1}{4}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3^2}{S_y^6}] \tag{21}$$

The values of α_1 and α_2 are respectively given by

$$\alpha_1 = \frac{1}{2} - \frac{\sqrt{n}}{4}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3}{S_y^3} \tag{22}$$

and

$$\alpha_2 = \frac{1}{2} + \frac{\sqrt{n}}{4}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3}{S_y^3} \tag{23}$$

Note that the value of $(\mu_3/S_y^3) = \mu_3/(S_y^2)^{3/2} = \beta_1$ in (22) and (23) is the value of the coefficient of skewness of the variable of interest. Thus the optimum values of α_1 and α_2 in (22) and (23) convey an important message. If the value of the coefficient of skewness is positive, then the proposed left side interval estimate $(\bar{y}_r - \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ should be given smaller weight and the right side interval estimate $(\bar{y}_r + \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ should be given greater weight so that the overall imputed value is given by

$$\bar{y}_{w(int)} = \bar{y}_r + (\alpha_2 - \alpha_1)\frac{s_{y(r)}}{\sqrt{n}} < \bar{y}_r \tag{24}$$

Thus for a data set skewed to the right, the imputed value of a non-respondent should be smaller than the sample mean value of the responding units. If the value of the coefficient of skewness is negative, then the proposed left side interval estimate $(\bar{y}_r - \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ should be given more weight and the right side interval estimate $(\bar{y}_r + \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ should be given less weight such that the overall imputed value is given by

$$\bar{y}_{w(int)} = \bar{y}_r + (\alpha_2 - \alpha_1)\frac{s_{y(r)}}{\sqrt{n}} > \bar{y}_r \tag{25}$$

Similarly for a data set skewed to the left, the imputed value of a non-respondent should be greater than the sample mean value of the responding units.

Further note that if the value of the coefficient of skewness is zero, then the proposed left side interval estimate $(\bar{y}_r - \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ should be given same weight as the right side interval estimate $(\bar{y}_r + \sqrt{n}\frac{s_{y(r)}}{(n-r)})$ so that the overall imputed value is given by

$$\bar{y}_{w(int)} = \bar{y}_r \tag{26}$$

Thus for a data set which is symmetric, (say Normally distributed) the imputed value of a non-respondent reduced to the sample mean value of the responding units.

For the optimum value of $\alpha_1 - \alpha_2 = \frac{\sqrt{n}}{2}(\frac{1}{r} - \frac{1}{N})\frac{\mu_3}{S_y^3}$, the percent relative bias in the proposed weight interval method of imputation is given by

$$RB(\bar{y}_{w(int)}) = -\frac{1}{2}C_y(\frac{1}{r} - \frac{1}{N})\beta_1[1 - \frac{1}{8}(\frac{1}{r} - \frac{1}{N})(\beta_2 - 1)] \times 100\% \quad (27)$$

where $\beta_1 = \mu_3/(S_y^2)^{3/2}$ is the coefficient of skewness, $\beta_2 = \mu_4/S_y^4$ is the coefficient of kurtosis and C_y is the coefficient of variation. Note that if $\beta_1 = 0$ then there is no relative bias and the proposed weighted method of imputation reduces to the usual mean method of imputation for the optimum values of α_1 and α_2. If $\beta_1 < 0$ then the distribution is skewed to the left and if $\beta_1 > 0$ then the distribution is skewed to the right. If $\beta_2 = 3$ then the distribution is *mesokurtic* which stands for normal distribution; if $\beta_2 > 3$ then the curve will be more peaked than the normal curve and it is called *leptokurtic* curve, while if $\beta_2 < 3$ then the curve is flatter than the normal curve and it is called *platykurtic* curve. The percent relative efficiency of the proposed weighted interval method of imputation with respect to the mean method of imputation is defined as:

$$RE(\bar{y}_{w(int)}) = \frac{1}{[1 - \frac{1}{4}(\frac{1}{r} - \frac{1}{N})\beta_1^2]} \times 100\% \quad (28)$$

From (28), one obvious observation is that the value of the coefficient of skewness should be a real number satisfying the condition:

$$-\frac{2}{\sqrt{(\frac{1}{r} - \frac{1}{N})}} \le \beta_1 \le \frac{2}{\sqrt{(\frac{1}{r} - \frac{1}{N})}} \quad (29)$$

Thus we would be interested in studing the behaviour of the percent relative bias and percent relative efficiency for values of β_1 satisfying the condition (29), for various value s of C_y, and for a few values of β_2 showing both *platykurtic* and/ or *leptokurtic* type curves. In order to look at the magnitude of gain in efficiency for different choice of parameters involved in the percent relative bias and percent relative efficiency expressions, we generated four populations each of size $N = 10000$ units by using the model:

$$y_i = 10 + y_i^* \quad (30)$$

where $y_i^* \sim Gamma(\alpha, \beta)$, that is, y_i^* follows gamma distribution with parameters α and β. In other words, in each of the generated population, the study variable y_i follows gamma distribution with mean shifted up by 10 units. The four populations are generated with four different choices of shape parameter α equal to 0.05, 0.15, 0.25 and 0.35 and only one choice of scale parameter $\beta = 1$. One can see that the purpose of adding 10 to the population values is to reduce the value of the coefficient of variation to a reasonable value around 10% by following [1]. A graphical presentation of four such populations is given in Figure 1.

For a population, with $\alpha = 0.05$ we found $\bar{Y} = 1.5489$, $S_y = 0.2208$, $Y_{min} = 1.5$, $Y_{med} = 1.500$, $Y_{max} = 6.1792$, $\beta_1 = 8.76$ and $\beta_2 = 106.18$; with $\alpha = 0.15$ we found $\bar{Y} = 1.6495$, $S_y = 0.3826$, $Y_{min} = 1.5$, $Y_{med} = 1.506$, $Y_{max} = 8.8694$, $\beta_1 = 4.94$ and $\beta_2 = 36.91$; with $\alpha = 0.25$ we found $\bar{Y} = 1.7581$, $S_y = 0.5205$, $Y_{min} = 1.5$, $Y_{med} = 1.5438$, $Y_{max} = 8.8640$, $\beta_1 = 3.98$ and $\beta_2 = 22.42$; and with $\alpha = 0.35$ we found $\bar{Y} = 1.8492$, $S_y = 0.5875$, $Y_{min} = 1.5$, $Y_{med} = 1.6039$, $Y_{max} = 8.6437$, $\beta_1 = 3.28$ and $\beta_2 = 15.92$. Note that the moment generating function of a gamma random variable is given by

$$M_{y^*}(t) = (1 - \beta t)^{-\alpha} \quad (31)$$

One can easily see that:

$$\bar{Y} = E(y_i) = 10 + \alpha\beta \quad (32)$$

$\mu_2 = S_y^2 = E(y_i - \bar{y})^2 = \alpha\beta^2$;$\mu_3 = E(y_i - \bar{y})^3 = 2\alpha\beta^3$; and$\mu_4 = E(y_i - \bar{y})^4 = 3\alpha(\alpha + 2)\beta^4$. The value of the coefficient of skewness is given by $\beta_1 = 2/\sqrt{n}$. The value of the coefficient of kurtosis is given by $\beta_2 =$

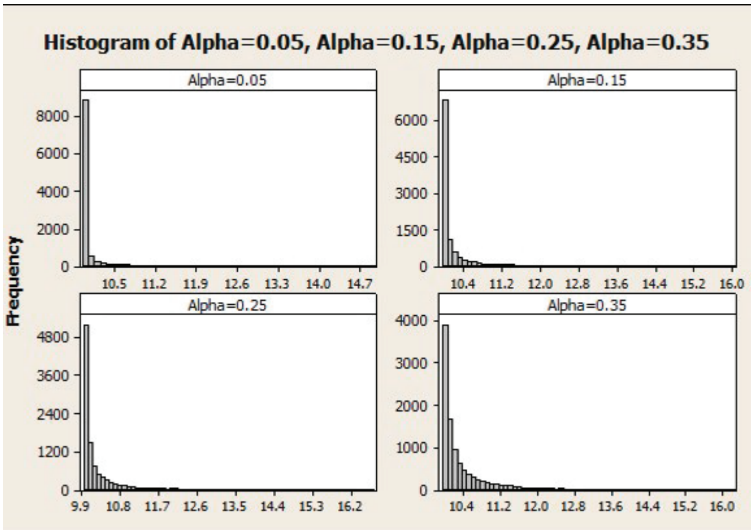

Figure 1. Four populations considered in the study

$3(\alpha+2)/\alpha$. Note that: $\beta_2-\beta_1^2-1>0$: The value of the coefficient of variation is given by $C_y=\sqrt{\alpha}\beta/(10+\alpha)$. Thus alternatively, with $\alpha=0.05$, we have $\bar{Y}=1.55$, $S_y=0.2236$, $C_y=0.1442$, $\beta_1=8.944$ and $\beta_2=123$; with $\alpha=0.15$, we have $\bar{Y}=1.65$, $S_y=0.3873$, $C_y=0.2347$, $\beta_1=5.164$ and $\beta_2=43$; with $\alpha=0.25$, we have $\bar{Y}=1.75$, $S_y=0.5000$, $C_y=0.2857$, $\beta_1=4.000$ and $\beta_2=27$; and with $\alpha=0.35$, we have $\bar{Y}=1.85$, $S_y=0.5916$, $C_y=0.3197$, $\beta_1=3.381$ and $\beta_2=20.14$;

Either one of these two sets of parameters can be used to study the percent relative efficiency and percent relative bias as both have identical results, and the difference is due to the finite population of size $N=10000$ taken from such super-populations. To be more precise, we used the alternatively produced parameters with $N=10000$ in order to find the percent relative efficiency and percent relative bias. We consider $n=500$, $r=50$, 100, 150, 200, 250, 300, 350, 400 and 450.

We found that, if $\alpha=0.05$ then the population is highly leptokurtic, having a value of $\beta_2=123$ and skewed to the right with a value of $\beta_1=8.944$ and a reasonable value of the coefficient of variation $C_y=14.43\%$, then the percent relative efficiency (RE) value increases from 104.4% to 166.1% as the value of r decreases from 450 to 50. The respective values of the percent relative bias (RB) remain negligible between -0.1369% and -1.2839% . If the value of α is increased to 0.15, then the population is still leptokurtic with a value of $\beta_2=43.0$ and skewed to the right with a value of $\beta_1=5.164$ and has a high value of the coefficient of variation $C_y=23.47\%$, so that the percent relative efficiency (RE) value increases from 101.4% to 115.3% as the value of r decreases from 450 to 50. The value of the percent relative bias (RB) still remains negligible between -0.1455% and -1.2061%. Now if the value of α is increased to 0.35, then the population is still leptokurtic with a value of $\beta_2=20.1$, skewed to the right with a value of $\beta_1=3.81$ and has a high value of the coefficient of variation $C_y=31.98\%$, then the percent relative efficiency (RE) value increases from 100.6% to 106.0% as the value of r decreases from 450 to 50. The value of the percent relative bias (RB) remains negligible between -0.1147% and -1.0757%. It seems that if the variation in a population is too large then it will be hard to impute missing values irrespective of the method one uses. Thus consistency of a population must be investigated before implementing any imputation method.

We conclude that the proposed weighted interval method of imputation can be useful in imputing the missing value if the value of the coefficient of skewness and coefficient of kurtosis in the population are known. The proposed weighted interval method of imputation can perform better than mean method of imputation if the distribution is skewed to right or left with a high value of β_2. In practice, the distribution of income is found to be skewed to the right in many populations of interest, and if it also has leptokurtic nature then the proposed weighted method of imputation can be useful in imputing missing values in such surveys.

Among others, [6] made use of the known population variance of the study variable in improving the estimator of population mean of the same study variable. In the next section, this motivated the authors to construct a new method of imputation in the presence of a known population variance of the same variable.

4. A NEW IMPUTATION METHOD WHEN POPULATION VARIANCE IS KNOWN

In this section, we introduce a naive variance dependent imputing method as follows:

$$\hat{\hat{y}}_{\bullet i} = \begin{cases} y_i & \text{if } i \in A \\ \bar{y}_r + \gamma\frac{n}{(n-r)}(S_y^2 - s_{y(r)}^2) & \text{if } i \in A^c \end{cases} \tag{33}$$

where γ is a constant to be determined such that the variance of the final estimator is minimum, and $S_y^2 = (N-1)^{-1}\sum_{i\in\Omega}(y_i - \bar{Y})^2$ is the known population variance of the study variable y.

If $\gamma = 0$ then the proposed variance dependent imputing method reduces to the usual mean method of imputation. Under the proposed variance dependent imputation method in (33), the point estimator (6) can be written as:

$$\bar{y}_{point} = \frac{1}{n}\sum_{i\in s}\hat{\hat{y}}_{\bullet i} = \bar{y}_r + \gamma(S_y^2 - s_{y(r)}^2) = \bar{y}_{n(var)} \tag{34}$$

The proposed variance dependent estimator $\bar{y}_{n(var)}$ is an unbiased estimator of the population mean. For the optimum value of γ, the minimum variance of the proposed naive variance dependent estimator $\bar{y}_{n(var)}$ is given by

$$Min.V(\bar{y}_{n(var)}) = (\frac{1}{r} - \frac{1}{N})S_y^2(1 - \frac{\beta_1^2}{(\beta_2 - 1)}) \tag{35}$$

The percent relative efficiency of the proposed naive variance dependent method of imputation estimator $\bar{y}_{n(var)}$ with respect to the mean method of imputation is defined as:

$$RE(\bar{y}_{n(var)}) = \frac{\beta_2 - 1}{\beta_2 - \beta_1^2 - 1} \times 100\% \tag{36}$$

It is interesting to note that the value of the percent relative efficiency is a function of only two parameters β_1 and β_2. We used the same four population for α equal to 0.05, 0.15, 0.25 and 0.35 as in the previous section. The computed optimum values of γ are found to be 0.328, 0.317, 0.308 and 0.199 respectively with the percent relative efficiency (RE) values of being $290.5\%, 273.9\%, 260.0\%$ and 248.1%. A further look at the behaviour of the percent relative efficiency as a function of value of α, finds that as the value of α increases from 1 to 49, the value of β_1 decreases from 2 to 0.286, the value of β_2 decreases from 9 to 3.122 (leptokurtic), the value of C_y decreases from 0.4 to 0.139, the optimum value of γ decreases from 0.25 to 0.019, and the percent relative efficiency decreases from 200.00% to 104.0%.

The following remark is devoted to answer very valuable question raised by one of the reviewers:

5. Remark

(1) Are the new methods competitive with the EM or MI algorithm?

(a) EM-Algorithm: As per our understanding, the EM-Algorithm is making an assumption of know distribution of data being imputed as is done in case of mathematical statistics. For example, refer to [5] and it seems it was introduced by [2]. In survey sampling methodology, we do not make any such assumption that the distribution of data is known or unknown. All the ratio, product and regression type estimators are free from such assumptions. However sometime they assume known values of a few constants being used at the estimation stage are functions of the parameters being estimated. Later they show that the replacement of those unknown constants with their consistent estimators is not altering the final mean square errors to the first order of approximation. For example, the difference estimator

$$\bar{y}_{dif} = \bar{y} + \beta(\bar{X} - \bar{x}) \tag{37}$$

depends on a constant β whose optimum value is given by $\beta = \frac{S_{xy}}{S_x^2}$, and it is estimated as $\hat{\beta} = \frac{s_{xy}}{s_x^2}$ and the difference estimator becomes a regression estimator given by

$$\bar{y}_{reg} = \bar{y} + \hat{\beta}(\bar{X} - \bar{x}) \tag{38}$$

It is shown is all textbooks, see [1], that $V_(\bar{y}_{dif}) = MSE(\bar{y}_{reg})$. One could also refer to [14].
In EM-Algorithm, it is not clear how one EM-Algorithm can be proved to be better than another EM-Algorithm except through simulation study based on couple of thousands of iterations. Those simulation results would change from data set to data set. In contrast, the ratio or regression type methods of imputation are comparable based on fixed parameter called Mean Square Error (MSE), which is easy to derive and compares based of theoretical justification. It results into one standard result, for example in the present paper it is shown that the Searls method of imputation always has MSE less than the mean method of imputation irrespective of the data set being used for imputation. Please see equation (15) on page 2. Such theoretical justification is not possible in case of EM Algorithms.

(b) Multiple Imputations: As per our understanding, the Multiple Imputations can be interpreted in two different meanings:

(i.) If we are imputing missing values for several variables, then the Searls imputing method proposed in this paper will always provide better results than the mean method of imputations for each variable being imputed. If missingness is considered as a very sensitive issue, then one should not impute several variables simultaneously. Imputation needs to be done very carefully for each variable separately, and if possible every single missing value should be imputed very carefully for every variable.

(ii.) If we are imputing missing values several times for a single variable, then no doubt an EM-Algorithm will provide a different value at each iterations of imputation based on the random start, and would lead to suspicious that which imputed value should be considered. In contrast, each one of the mean, ratio, regression, or the proposed Searls method of imputation will give us unique imputed value based on the method being used. There will be no confusion which imputed value to be considered or not.

(2) In what way can a measure of variability between imputations be produced from the new methods?

In mean, ratio and regression type methods of imputation, there is no question of variability between the imputed values. The imputed values are unique by a given method. However such a problem of variability between the imputed values by EM-Algorithm could make us suspicious which and why an imputed value should be considered?

(3) Is it possible to produce multiple imputations using any of the three methods?

It depends how you define Multiple Imputations, if you are imputing one variable, then the answer is yes, otherwise no. Here “ no ” is better than “ yes ”, because your imputed value is unique and removes any types of confusion.

(4) Can new method only be applied if the mechanism of missing values is completely random-MCAR? What should be done with the new methods in case of mechanism of missing values is MAR or MCAR?

So long as the mean method of imputation or the proposed Searls type method of imputation is concerned, it should be applied only to the situation of MCAR. However, if some auxiliary information is available, the proposed method can be extended on the lines of recent work of [8, 9, 10]. To our knowledge this idea of proposing Searls method of imputation, and other two methods in the paper are completely new.

Comments about the methods:

(1) From (37) and (38), and following [14], it is advisable to replace the parameters involved in the imputation methods by their consistent estimator. The resultant estimators will again be consistent and will have same asymptotic mean square errors. One could also refer to [8, 9, 10]. These findings of replacing unknown parameters by their consistent estimates are well known in the literature, so are not discussed in detail in this paper.

(2) If the knowledge of the population variance is lacking, then it can be estimated from the responding data set, and the resultant imputing method will be a consistent estimator and will have same mean squared error to the first order of approximation.

6. Conclusion

In conclusion, the results from the computations using the four populations that the proposed variance dependent method of imputation is more efficient than the mean method as well as more efficient than the proposed interval method. The reason may be due to the fact that the optimum value of γ makes use of both the value of coefficient of skewness β_1 and the value of the coefficient of kurtosis β_2, i.e., because it makes use of more information at the prediction stage, it is likely to be more efficient than its existing competitors. Overall, we conclude that the variance dependent method could be more efficiently used to impute missing values in comparison to other methods discussed in the present investigation.

Acknowledgement

The authors are thankful to the Field Editor: Paulo Rodrigues, Admin: David G. Yu and a learned referee for very constructive comments on the original version of this manuscript. The comments were so lovely that the authors decided to give special Remark in the revised version. The authors also would like to thank to Dr. Polly Allred, Department of Mathematics, Texas A&M University-Kingsville, TX for editing the manuscript.

REFERENCES

1. Cochran, W.G. (1963). *Sampling Techniques*. John Wiley and Sons: New York.
2. Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm (with Discussion). *Journal of the Royal Statistical Society*, B, 39(1), 1-38.
3. Hansen, M.H. and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *J . Amer. Statist. Assoc.*, 41,517–529.
4. Heitjan, D.F. and Basu S. (1996). Distinguishing Missing At Random and Missing Completely At Random. *The American Statistician*, 50, 207-213.
5. Johnson, Richard A., and Dean W. Wichern (1982). *Applied Multivariate Statistical Analysis*, pages 209-213: Prentice Hall Inc. Englewood Cliffs,.N.J.
6. Kataria, P. and Singh, S. (1989).On the estimation of mean when population variance is known. *J. Indian Soc. Agri. Statist.*, 41(2), 173-175.
7. Mohamed, C. (2015). *Improved Imputation Methods in Survey Sampling*. Unpublished MS thesis submitted to the Department of Mathematics, Texas A&M University-Kingsville, TX.
8. Mohamed, C., Sedory, S.A. and Singh, S. (2016). Comparison of different imputing methods for scrambled responses. *Handbook of Statistics: Data Gathering Analysis and Protection of Privacy Through Randomized Response Techniques: Qualitative and Quantitative Human Traits*, 34, 471-495.
9. Mohamed, C., Sedory, S.A. and Singh, S. (2017). Imputation using higher order moments of an auxiliary variable. *Communications in Statistics: Simulation and Computations*, 46(8), 6588-6617.
10. Mohamed, C., Sedory, S.A. and Singh, S. (2018). A fresh imputing survey methodology using sensible constraints on study and auxiliary variables: dubious random non-response. *Journal of Statistical Computations and Simulations*, 88:7, 1273-1294.
11. Rubin, D.B. (1976). *Inference and missing data. Biometrika*, 63(3), 581 -592
12. Searls, D.T. (1964). The utilization of a known coefficient of variation in the estimation procedure. *J. Amer. Statist. Assoc.*, 59, 1225–1226.
13. Searls, D.T. (1967). A note on the use of an approximately known co-efficient of variation. *American Statistician*, 21(2), 20–21.
14. Singh, S., Mangat, N.S., and Mahajan, P.K. (1995). General class of estimators. *J. Indian Soc. of Agricul. Statist.* 47, 129-133.