GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

Imputation Methods of Missing data for estimating the Population Mean using Simple Random Sampling

Ajeet Kumar Singh
Department of Statistics,
Banaras Hindu University
Varanasi,
India.

Priyanka Singh
Department of Statistics,
Banaras Hindu University
Varanasi,
India.
.

V.K. Singh
Department of Statistics,
Banaras Hindu University
Varanasi,
India.

Abstract

In this paper we propose three exponential type compromised imputation methods for estimating the population mean based upon an auxiliary variable in simple random sampling when some observations are missing. They are compared with other imputation methods such as mean method, ratio method and compromised method of imputation. Our proposed estimators are better than other above mentioned estimator. To support the discussed results the relative efficiencies of the estimator w.r.t.these estimators have been obtained using empirical data.

Keywords: Imputation methods, Bias and Mean square error, missing data, relative efficiency.

1. INTRODUCTION

In survey sampling situations, auxiliary information is generally used to improve the precision or accuracy of the estimator of unknown population parameter of interest under the assumption that all the observations in the sample are available. But in many survey sampling situations, this assumption is not true .This is the case of incomplete information which may arise due to some non- response in the given sample .Incomplete information is very common in the studies related to medical research, market research surveys, opinion polls socio economic investigations etc.

In statistical inference, sometimes the efficiency of parameters estimation may be reduced when some observations from the sample units are missing .Our aim is to try to impute the missing observations. To deal with missing values effectively Kalton *et al* (1981) and Sande (1979) suggested imputation methods that make an incomplete data set structurally complete and its analysis simple. Imputation may also be carried out with the aid of an auxiliary variate if it is available. For more about the missing data and the methods of imputation one can refer Rueda and Gonzalez (2008), Rao and Sitter (1995), Gonzalez *et al* (2008), Baraldi and Enders (2010), Bouza (2008). Based on auxiliary variable, recently Singh and Horn (2000) and Singh *et al* (2014) suggested some compromised method of imputation.

2. THE PROBLEM AND NOTATIONS:

Let a simple random sample S of size n without replacement be drawn from a finite population $U = (Y_1, Y_2, ..., Y_N)$ of size N and with study characteristic Y. Let $(\bar{Y}, \bar{X})$ be the population mean of the study variable Y and auxiliary variable X respectively. It is presumed

that the sample consists of r responding units (r < n) belonging to a set A and (n-r) non responding units belonging to a set A^C. Further let for every unit $i \in A$, the value y_i is observed and for the unit $i \in A^C$,the value y_i is missing for which suitable imputed value is to be derived .For this purpose, the ith value of the auxiliary variable is used as a source of imputation for missing data when $i \in A^C$.

In what follows, we shall use the following notations:

Z: Stands for either variable Y or variable X.

$\bar{z}_n$: Sample mean based on n observations for variable Z.

$\bar{z}_r$:Sample mean of the responding units based on r observations for the variable Z.

S_Z^2 :Population mean squares for the variable Z.

C_Z : Coefficient of variation (CV) for the variable Z; $C_Z = \frac{S_Z}{\bar{Z}}$.

ρ is the coefficient of correlation between the variable Y and X in the population .

$y_{.i}$: Imputed value for the ith value of $y_i (i = 1,2...n)$

$\theta_{n,N}, \theta_{r,N}, \theta_{r,n}$:Finite population corrections (fpc); $\left(\frac{1}{n} - \frac{1}{N}\right), \left(\frac{1}{r} - \frac{1}{N}\right), \left(\frac{1}{r} - \frac{1}{n}\right)$ respectively.

3. SOME IMPUTATION STRATEGIES

Before suggesting the proposed imputation strategy, we shall mention here some existing imputation strategies for readiness of the material which has a direct relevance with the present work. We shall denote by (D, T) denote a sampling strategy where D stands for simple random sampling without replacement sampling scheme and T for an estimator for population mean $\bar{Y}$. Followings are the some imputation methods and corresponding sampling strategies:

3.1 $(D, \bar{y}_r)$:*Mean method*

Here

$$y_{.i} = \begin{cases} y_i & \text{if } i \in A \\ \bar{y}_r & \text{if } i \in A^C \end{cases} \qquad (1)$$

The corresponding point estimator and its bias, B(.) and mean square error (MSE), M(.) are derived as

GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

$$\bar{y}_s = \frac{1}{n}\sum_{i\in s} y_{.i} = \bar{y}_r \tag{2}$$

$$B(\bar{y}_r) = 0 \tag{3}$$

$$M(\bar{y}_r) = \theta_{r,N}\bar{Y}^2 C_Y^2 \tag{4}$$

3.2 $(D, \bar{y}_{RAT})$: *Ratio method*

$$y_{.i} = \begin{cases} y_i & \text{if } i \in A \\ \hat{b}x_i & \text{if } i \in A^C \end{cases} \tag{5}$$

where $\hat{b} = \dfrac{\sum_{i\in A} y_i}{\sum_{i\in A} x_i}$

Then the point estimator, its bias and MSE are given by:

$$\bar{y}_{ratio} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{6}$$

$$B(\bar{y}_{ratio}) = \theta_{r,n}\bar{Y}\left[C_X^2 - \rho_{XY}C_X C_Y\right] \tag{7}$$

$$M(\bar{y}_{ratio}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,n}\bar{Y}^2\left[C_X^2 - 2\rho_{XY}C_X C_Y\right] \tag{8}$$

3.3 $(D, \bar{y}_{COMP})$: *Compromised method* (Singh and Horn, 2000)

$$y_{.i} = \begin{cases} \alpha \dfrac{n}{r} y_i + (1\text{-}\alpha)\hat{b}x_i & \text{if } i \in A \\ (1\text{-}\alpha)\,\hat{b}x_i & \text{if } i \in A^C \end{cases} \tag{9}$$

The point estimator is

$$\bar{y}_{COMP} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \quad ; \ \alpha \text{ being a suitable constant} \tag{10}$$

$$B(\bar{y}_{COMP}) = (1-\alpha)\theta_{r,n}\bar{Y}\left[C_X^2 - \rho_{XY}C_X C_Y\right] \tag{11}$$

$$M(\bar{y}_{COMP}) = \theta_{r,N}\bar{Y}^2 C_Y^2 + \theta_{r,n}\bar{Y}^2\left[(1-\alpha)^2 C_X^2 - 2(1-\alpha)\rho_{XY}C_X C_Y\right] \tag{12}$$

It can be seen that the estimator has minimum MSE for $\alpha = 1 - \rho \frac{C_Y}{C_X}$ for which

$$M(\bar{y}_{COMP})_{\min} = \bar{Y}^2\left[(\theta_{r,N} - \theta_{r,n}\rho^2)C_Y^2\right] \qquad (13)$$

4. PROPOSED IMPUTATION STRATEGY $(D, T_i, i = 1,2,3)$

Motivated with Singh *et al*. (2014) and Bahl and Tuteja (1991), we here proposed the following exponential-type estimators

$$T_1 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{X} - \bar{x}_n}{\bar{X} + \bar{x}_n}\right) \qquad (14)$$

$$T_2 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{x}_n - \bar{x}_r}{\bar{x}_n + \bar{x}_r}\right) \qquad (15)$$

$$T_3 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right) \qquad (16)$$

where k is a suitably chosen constant to be determined under certain conditions.

5. PROPERTIES OF PROPOSED IMPUTATION STRATEGY

In relation to bias, MSE, optimum value of the parameter k and corresponding minimum MSE, we have the following theorems:

5.1 *Theorem1:*

The bias and MSE of the proposed strategy (D, T_1) to the terms of order $O(n^{-1})$ are given by

$$Bias(T_1) = (1-k\ )\bar{Y}\left[\frac{3}{8}\theta_{n,N}C_X^2 - \frac{1}{2}\theta_{n,N}\rho C_Y C_X\right] \qquad (17)$$

$$M(T_1) = \bar{Y}^2\left[\theta_{r,N}C_Y^2 + (1-k\ )^2\theta_{n,N}\frac{C_X^2}{4} - (1-k\ )\theta_{n,N}\rho C_X C_Y\right] \qquad (18)$$

The minimum MSE of (T_1) occurs when $k = 1 - 2\rho\frac{C_Y}{C_X}$ for which MSE reduces to

$$M(T_1)_{\min.} = \bar{Y}^2\left[\theta_{r,N} - \theta_{n,N}\rho^2\right]C_Y^2 \qquad (19)$$

The proof of the theorem is given in the Appendix

5.2 Theorem2:

The bias and MSE of the proposed strategy (D,T_2) to the terms of order $O(n^{-1})$ are given by

$$Bias(T_2)=(1-k\ )\bar{Y}\theta_{r,n}\left[\frac{3}{8}C_X^2-\frac{1}{2}\rho C_Y C_X\right] \tag{20}$$

$$M(T_2)=\bar{Y}^2\left[\theta_{r,N}\ C_Y^2+(1-k\ )^2\theta_{r,n}\frac{C_X^2}{4}-(1-k\ )\theta_{r,n}\rho C_X C_Y\right] \tag{21}$$

The minimum MSE of (T_2)occurs when $k=1-2\rho\dfrac{C_Y}{C_X}$ for which MSE reduces to

$$M(T_2)_{\min.}=\bar{Y}^2[\theta_{r,N}-(\theta_{r,n}\rho^2]C_Y^2 \tag{22}$$

The proof of the theorem is given in the Appendix

5.3 *Theorem3*:

The bias and MSE of the proposed strategy (D,T_3) to the terms of order $O(n^{-1})$ are given by

$$Bias(T_3)=(1-k\ )\bar{Y}\theta_{r,N}\left[\frac{3}{8}C_X^2-\frac{1}{2}\rho C_Y C_X\right] \tag{23}$$

$$M(T_3)=\bar{Y}^2\left[\theta_{r,N}C_Y^2+(1-k\ )^2\theta_{r,N}\frac{C_X^2}{4}-(1-k\ )\theta_{r,N}\rho C_X C_Y\right] \tag{24}$$

The minimum MSE of (T_3)occurs when $k=1-2\rho\dfrac{C_Y}{C_X}$ for which MSE reduces to

$$M(T_3)_{\min.}=\bar{Y}^2\theta_{r,N}(1-\rho^2)C_Y^2 \tag{25}$$

The proof of the theorem is given in the Appendix

6. COMPARISON OF DIFFERENT STRATEGIES

(i)The estimator $\bar{y}_{ratio}$ based on the ratio method of imputation is more efficient than $\bar{y}_r$ if

$$\frac{C_X}{C_Y}<2\rho\ \textit{for}\ \mathrm{R}>0\ \text{ and }\ \frac{C_X}{C_Y}>2\rho\ \text{for}\ \mathrm{R}<0 \tag{26}$$

(ii) T_1 is more efficient than $\bar{y}_r$ if

$$k \; > 1-4\rho\frac{C_Y}{C_X} \quad \text{if k } <1$$
$$and \;\; k \; < 1-4\rho\frac{C_Y}{C_X} \quad \text{if k } >1 \qquad (27)$$

Further it can be seen that $M(T_1)_{\min}$ is always smaller than $V(\bar{y}_r)$. In similar, T_2 and T_3 is more efficient than $\bar{y}_r$

(iii) T_1 is more efficient than $\bar{y}_{ratio}$ if

$$\theta_{n,N}\rho^2\bar{Y}^2C_Y^2+\theta_{r,n}\bar{Y}^2(C_X^2-2\rho C_X C_Y)>0 \qquad (28)$$

(iv) T_2 is more efficient than $\bar{y}_{ratio}$ if

$$\theta_{r,n}\bar{Y}^2\left(\rho C_Y - C_X\right)^2>0 \qquad (29)$$

(v) T_3 is more efficient than $\bar{y}_{ratio}$ if

$$\theta_{n,N}\bar{Y}^2C_Y^2\rho^2+\theta_{r,n}\bar{Y}^2\left(\rho C_Y - C_X\right)^2>0 \qquad (30)$$

(vi) T_1 is more efficient than $\bar{y}_{comp}$ if

$$M(\bar{y}_{COMP})_{opt}-M(T_1)>0$$
$$\Rightarrow (\theta_{n,N}-\theta_{r,N})\rho^2\bar{Y}^2C_Y^2 > 0 \qquad (31)$$

(vii) T_2 is more efficient than $\bar{y}_{comp}$ if

$$M(\bar{y}_{COMP})_{opt}-M(T_2)=0$$
$$\Rightarrow M(\bar{y}_{COMP})_{opt}=M(T_2) \qquad (32)$$

(viii) T_3 is more efficient than $\bar{y}_{comp}$ if

$$M(\bar{y}_{COMP})_{opt}-M(T_3)>0$$
$$\Rightarrow \theta_{n,N}\bar{Y}^2\rho^2C_Y^2>0 \qquad (33)$$

(ix) It is now desirable to compare the three suggested strategies for their performances .We have

(x)D$_1$= $M(T_2)_{Min}$ - $M(T_1)_{Min}$ = $\left(\theta_{n,N}-\theta_{r,n}\right)\rho^2C_Y^2$ (34)

So, T_1 is better than T_2

$$\text{if} \qquad r > \frac{Nn}{2N-n} = \frac{n}{2-f} \;, \quad \text{where f} = \frac{\text{n}}{\text{N}} \quad \left(0 < f < 1\right)$$

Thus, theoretically

$$\left.\begin{array}{l} \text{when } f = 0 \Rightarrow r > \dfrac{n}{2} \\ \text{when } f = 1 \Rightarrow \quad r > n \end{array}\right\}$$

Therefore, strategy (D,T_1) would be preferable over (D,T_2) if the number of respondents in the sample is more than fifty percent, which generally occurs in most of the surveys.

(xi) D$_2$= $M(T_1)_{Min}$ - $M(T_3)_{Min}$ $=\theta_{r,n}\rho^2 C_Y^2$ (35)

which is always positive. Therefore strategy $(\mathrm{D},\mathrm{T}_3)$ is always better than strategy (D,T_1).

(xii)D$_3$= $M(T_2)_{Min}$ - $M(T_3)_{Min}$ $=\theta_{n,N}\rho^2 C_Y^2$ (36)

which is always positive. Therefore strategy (D,T_3).is always better than strategy (D,T_2).

Combining the results derived above, one can say that if the sample does not contain more than fifty percent non – respondents, then the following results holds.

$$M(T_3) \le M(T_1) \le M(T_2)$$

7. EMPIRICAL STUDY

We consider the data given in Shukla *et al* (2011). A generated artificial population of size N = 200 containing values of main variable Y and auxiliary variable X. Parameters of this are given below:

$\overline{Y}$ = 42.485; $\overline{X}$ = 18.515; S_Y^2 = 199.0598; S_X^2 = 48.5375;

ρ = 0.8652; C_X = 0.3763; C_Y = 0.3321;

Using random sample of size n = 20; f = 0.1 by SRSWOR.

The condition of bias and MSE of the existing and proposed estimator are computed based of 30,000 repeated samples drawn by SRSWOR from population N = 200. These computations, with respect to $\overline{y}_r$, are given in tables1 and 2 where efficiency measurement is considered as

$$\text{R.}\, e(\hat{y}) = \frac{M(\overline{y}_r)}{M(\hat{y})} * 100$$

with $M(\hat{y})$ the mean squared error of estimator $\hat{y}$.

The simulation procedure contains following steps :

Step 1: Draw a random sample of size 20 from the population of N = 200 by SRSWOR.

Step 2: Drop down 5 units randomly from each sample corresponding to Y.

Step 3: Compute and impute the dropped units of Y with the help of proposed methods and available methods.

Step 4: Repeat the above steps 30,000 times, which provides multiple sample based estimates $\hat{y}_1$, $\hat{y}_2$, $\hat{y}_3$,…….. $\hat{y}_{30000}$.

Step 5: Bias of $\hat{y}_1$ is obtained by

$$B(\hat{y}) = \frac{1}{30000} \sum_{i=1}^{30000} \left[(\hat{y}_i) - \overline{Y} \right]$$

Step 6: M.S.E. of $\hat{y}$ is computed by

$$M(\hat{y}) = \frac{1}{30000} \sum_{i=1}^{30000} \left[(\hat{y}_i) - \overline{Y} \right]^2$$

Table-1 Bias, MSE, Relative efficiency of strategies

Estimator	***Min M(.)***	***R.E***	***Bias (.)***
$\overline{y}_r$	12.1236	100	0.2683
$\overline{y}_{ratio}$	9.7669	124.12	0.3216
$\overline{y}_{COMP}$	9.5882	126.44	0.4119

Table-2 Bias, MSE, Relative efficiency of strategies

Estimator	*Min M(.)*	*R.E*	*Bias (.)*
T_1	5.2811	229.56	-0.0070
T_2	9.5862	126.46	0.0043
T_3	2.8866	411.99	-0.0053

8. CONCLUSIONS

The work presented a compromised imputation strategy and corresponding point estimator, utilizing the information on an auxiliary variable on the basis of ETE. On the basis of population, a comparative study for the efficiency of the proposed strategy with some existing strategies showed that it is always preferable over other estimators

9. REFERENCES

Bahl, S. Tuteja, R. K. (1991): Ratio and product type exponential estimator, Information and Optimization sciences, Vol, XII, I, 159-163.

Baraldi, A. N. and Enders, C. K.,: An introduction to modern missing data analyses.*J.Sch.Psychol.48, 5-37(2010).*

Bouza, C. N.: Estimation of population mean with missing observations using product type estimators. *Rev.Investing. Oper. 29(3) 207-233 (2008).*

Kalton, G., Kasprzyk, D. and Santos, R (1981) : Issues of non -response and imputation of income and program participation . Current topics in survey sampling. In D. Krevoski, R. Platek and J. N. K. Rao(Eds.), (pp.455-480).New York: Acad .Press.

Gonzalez, S., Rueda, M. and Arcos, A.: An improved estimator to analyse missing data.*Stat.Pap.49, 791-796(2008).*

Rueda, M. and Gonzalez, S.: A new ratio-type imputation with random disturbance .*Appl.Math.Lett.21, 978-982(2008)*

Rao, J. N. K. and Sitter. R. R. (1995): Variance estimation under two phase sampling with application to imputation of missing data .*Biometrica82, 453-460*.

Sande, I. G. (1979): A personal view of hot deck approach to automatic edit and imputation. Journal Imputation Procedures. *Survey Methodology*, 5, 238-246.

Singh, S. and Horn, S. (2000): Compromised imputation in Survey sampling. *Metrika*, 51, 267-276.

Singh, A. K, Singh, P. and Singh, V. K. (2014): Exponential-Type Compromised Imputation in Survey Sampling, *Journal of the Statistics Applications and Probability*,3(2),211-217.

Shukla, D., Singhai, R. and Thakur, N. S. (2011): A new imputation method for missing attribute values in data mining , *Journal of Applied Computer Science and Mathematics,10(5),14-19.*

Appendix

We have

$$T_1 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{X} - \bar{x}_n}{\bar{X} + \bar{x}_n}\right)$$

$$T_2 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{x}_n - \bar{x}_r}{\bar{x}_n + \bar{x}_r}\right)$$

$$T_3 = k\ \bar{y}_r + (1-k\ )\bar{y}_r \exp\left(\frac{\bar{X} - \bar{x}_r}{\bar{X} + \bar{x}_r}\right)$$

Now using the large sample approximations $\varepsilon = \frac{\bar{y}_r}{\bar{Y}} - 1,\ \delta = \frac{\bar{x}_r}{\bar{X}} - 1$ and $\eta = \frac{\bar{x}_n}{\bar{X}} - 1$.With the concept of two- phase sampling and following Rao and Sitter (1995) mechanism of MCAR, for given r and n, we have.

$$E(\varepsilon) = E(\delta) = E(\eta) = 0 \quad E(\varepsilon^2) = \theta_{r,N} C_Y^2; \quad E(\delta^2) = \theta_{r,N} C_X^2; \quad E(\eta^2) = \theta_{n,N} C_X^2;$$
$$E(\varepsilon\delta) = \theta_{r,N}\rho C_Y\ C_X;\ E(\varepsilon\eta) = \theta_{n,N}\rho C_Y\ C_X;\ E(\delta\eta) = \theta_{n,N} C_X^2;$$

The estimator (T_1), (T_2) and (T_3) in terms of ε, δ and η up to first order of approximation, could be expressed as:

$$T_1 = \bar{Y}\left[(1+\varepsilon) - (1-k\ )\left(\frac{\eta}{2} + \frac{\varepsilon\eta}{2} - \frac{3}{8}\eta^2\right)\right] \qquad (14)$$

$$T_2 = \bar{Y}\left[(1+\varepsilon) + (1-k\ )\left(\frac{\eta-\delta}{2} + \frac{\varepsilon\eta - \varepsilon\delta}{2} - \frac{\delta\eta}{4} - \frac{\eta^2}{8} + \frac{3}{8}\delta^2\right)\right] \qquad (15)$$

$$T_3 = \bar{Y}(1+\varepsilon) + (1-k\ )\bar{Y}\left(-\frac{\delta}{2} + \frac{3}{8}\delta^2 - \frac{\varepsilon\delta}{2}\right) \qquad (16)$$

The expression (14), (15) and (16), obtained assuming that $|\,\varepsilon\,| < 1, |\,\eta\,| < 1and|\,\delta\,| < 1$, are valid assumptions. Taking expectation of both the sides of (14), (15) and (16) and realising that $B(T_i) = E(T_i) - \bar{Y}, i = 1,2,3$ .we have the expressions (17), (20) and (23).

Similarly, squaring the expression (14), (15) and (16), neglecting the terms of $\varepsilon, \delta \; and \; \eta$ greater than two and realising that

$$M\left(T_i\right) = E[T_i^{\,2}] + \bar{Y}^2 - 2\bar{Y}E[T_i] \ , \qquad i = 1,2,3$$

The expressions (18), (21) and (24) could be obtained applying large sample approximation results as given above.