Journal of Reliability and Statistical Studies; ISSN (Print): 0974-8024, (Online): 2229-5666
Vol. 10, Issue 2 (2017): 33-41

A GENERALIZED CLASS OF ESTIMATORS FOR ESTIMATING POPULATION MEAN USING IMPUTATION TECHNIQUE

Prabhakar Mishra, Poonam Singh and *Rajesh Singh
Department of Statistics, Banaras Hindu University, Varanasi, India
E Mail: *rsinghstat@gmail.com
* Corresponding author

Received April 25, 2017
Modified October 12, 2017
Accepted October 30, 2017

Abstract

This article deals with estimation of population mean for missing data in simple random sampling. The properties of the proposed procedure are studied up-to first order of approximation and under the optimality conditions proposed estimator outperforms other existing estimators. A numerical illustration, based on the two real data sets, highlights the efficiency gain using our proposed estimator.

Key Words: Imputation, Missing Data, MSE, Bias, Efficiency, SRSWOR.

1. Introduction

It is known that use of auxiliary information always results in improved procedures, common examples are ratio, product and regression estimators. Many authors including Upadhyaya and Singh (1999), Abu-Dayyeh et al. (2003), Kadilar and Cingi (2005), Khoshnevisan et al. (2007), Singh et al. (2007), Singh et al. (2008), Singh and Kumar (2011), Singh et al. (2012) and Sanaullah et al. (2012) proposed improved procedurs.

When one or more values are missing for a case, it creates problem for the analysis of data. Missing data introduces bias and makes the handling and analysis of data more strenuous, which in turn causes reduction in efficiency. Imputation is a technique in which missing values are filled in by substitutes based on the information available. These substitutes can be constructed in several ways. After imputing missing values, we get a complete data set which can be analyzed using standard techniques. Kalton et al. (1981) caused one to think for imputation technique for missing survey responses. For variance estimation under imputation Lee et al. (1994) proposed some improved estimators. Singh and Horn (2000), Ahmed et al. (2006), Shukla et al. (2009), Shukla and Thakur (2011), Thakur et al. (2012), Shukla et al. (2013), Omari et al. (2013), Singh et al. (2014), Pandey et al. (2015) also studied the problem of missing values using imputation techniques.

Motivated by the work of above authors, in this paper we have proposed an exponential method of imputation for missing values and compared the efficiency with the other estimators considered in the paper.

2. Notations

Let $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i$. To estimate population mean a sample of size n is drawn. . Out of n units r units responds and (n-r) units did not responded. Also let A is the set of responding units and A^c is the set of non-responding units. General method of imputation can be expressed as:

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \hat{y}_i & \text{if} \quad i \in A^c \end{cases} \tag{2.1}$$

General point estimator of population mean is given by

$$\overline{y}_s = \frac{1}{n}\left[\sum_{i=1}^{r} y_i + \sum_{i=1}^{n-r} \hat{y}_i\right] \tag{2.2}$$

We will be using following approximations:

$$\overline{y}_r = \overline{Y}(1+e_0), \quad \overline{x}_r = \overline{X}(1+e_1), \quad \overline{x}_n = \overline{X}(1+e_2)$$

$$\overline{z}_r = \overline{Z}(1+e_3), \quad \overline{z}_n = \overline{Z}(1+e_4)$$

$$E(e_i) = 0 \qquad \forall \ i = 1,2,3,4,5$$

$$E(e_0^2) = \theta_{r,N} C_y^2 \ , \ E(e_1^2) = \theta_{r,N} C_x^2 \ , \ E(e_2^2) = \theta_{n,N} C_x^2$$

$$E(e_3^2) = \theta_{r,N} C_z^2 \ , \ E(e_4^2) = \theta_{n,N} C_z^2$$

Where $\theta_{r,N} = \frac{1}{r} - \frac{1}{N}$

$$\theta_{n,N} = \frac{1}{n} - \frac{1}{N}$$

where $S_y^2 = \frac{C_y^2}{\overline{Y}}$, $S_x^2 = \frac{C_x^2}{\overline{X}}$, $\rho = \frac{S_{xy}}{S_x S_y}$ and S_y^2, S_x^2, S_{xy} have their usual meaning.

3. Existing Imputation Methods

Some common methods of imputation are :

3.1. Imputation using simple mean

Here for missing data we have

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \bar{y}_r & \text{if} \quad i \in A^c \end{cases} \tag{3.1}$$

For unknown population mean $\bar{Y}$ the point estimate is

$$\bar{y}_{mean} = \frac{1}{n}\sum_{i \in s} y_{.i} = \bar{y}_r \tag{3.2}$$

where

$$\bar{y}_r = \frac{1}{r}\sum_{i \in A} y_{.i} \tag{3.3}$$

3.2. Imputation by ratio method

Here for missing data we have

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \hat{b}x_i & \text{if} \quad i \in A^c \end{cases} \tag{3.4}$$

The point estimator of $\bar{Y}$ is given as-

$$\bar{y}_{ratio} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{3.5}$$

where $\bar{x}_n = \frac{1}{n}\sum_{i \in A} x_i$, $\bar{x}_r = \frac{1}{r}\sum_{i \in A} x_i$

and

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

3.3. Singh and Horn (2000) method of imputation

Singh and Horn (2000) method of imputation is given as-

$$y_{.i} = \begin{cases} \alpha \frac{n}{r} y_i + (1-\alpha)\hat{b}x_i & \text{if} \quad i \in A \\ (1-\alpha)\hat{b}x_i & \text{if} \quad i \in A^c \end{cases} \tag{3.6}$$

where α is suitably chosen constant.

The point estimator for $\bar{Y}$ is given as-

$$\bar{y}_{comp} = \alpha \bar{y}_r + (1-\alpha)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{3.7}$$

4. Proposed Method of Imputation for Missing Data

Motivated by Shabbir and gupta (2015)we proposed following exponential estimator for missing data:-

$$y_{.i}=\begin{cases} k\frac{n}{r}y_i+(1-k)\bar{y}_r\exp\left(\frac{a(\bar{X}-\bar{x}_r)}{\bar{X}+(b-1)\bar{x}_r}\right)\exp\left(\frac{c(\bar{Z}-\bar{z}_r)}{\bar{Z}+(d-1)\bar{z}_r}\right) & \text{if } i\in A \\ (1-k)\bar{y}_r\exp\left(\frac{a(\bar{X}-\bar{x})}{\bar{X}+(b-1)\bar{x}_r}\right)\exp\left(\frac{c(\bar{Z}-\bar{z}_r)}{\bar{Z}+(d-1)\bar{z}_r}\right) & \text{if } i\in A^c \end{cases} \tag{4.1}$$

The point estimator of population mean under proposed method of imputation is

$$\bar{y}_{pr}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{a(\bar{X}-\bar{x}_r)}{\bar{X}+(b-1)\bar{x}_r}\right\}\exp\left\{\frac{c(\bar{Z}-\bar{z})}{\bar{Z}+(d-1)\bar{z}_r}\right\} \tag{4.2}$$

For different values of a, b, c and d we can generate following point estimators:

a	b	c	d	Estimators
1	2	0	-	$\bar{y}_{pr1}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{X}-\bar{x}_r)}{\bar{X}+\bar{x}_r}\right\}$
1	1	0	-	$\bar{y}_{pr2}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{X}-\bar{x}_r)}{\bar{X}}\right\}$
0	-	1	1	$\bar{y}_{pr3}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{Z}-\bar{z})}{\bar{Z}}\right\}$
1	1	1	1	$\bar{y}_{pr4}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{X}-\bar{x}_r)}{\bar{X}}\right\}\exp\left\{\frac{(\bar{Z}-\bar{z})}{\bar{Z}}\right\}$
1	1	1	2	$\bar{y}_{pr5}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{X}-\bar{x}_r)}{\bar{X}}\right\}\exp\left\{\frac{(\bar{Z}-\bar{z})}{\bar{Z}+\bar{z}_r}\right\}$
1	2	1	1	$\bar{y}_{pr6}=k\bar{y}_r+(1-k)\bar{y}_r\exp\left\{\frac{(\bar{X}-\bar{x}_r)}{\bar{X}+\bar{x}_r}\right\}\exp\left\{\frac{(\bar{Z}-\bar{z})}{\bar{Z}}\right\}$

Based on the above approximations equation (4.2) can be written as

$$\bar{y}_{pr} = k\bar{Y}(1+e_0) + (1-k)\bar{Y}(1+e_0)\exp\left\{\frac{a(-\bar{X}e_1)}{\bar{X}+(b-1)(\bar{X}+e_1\bar{X})}\right\} \exp\left\{\frac{c(-\bar{Z}e_3)}{\bar{Z}+(d-1)(\bar{Z}+e_3\bar{Z})}\right\} \tag{4.3}$$

Following results we obtain for our proposed procedure:

Theorem 4.1 : Bias of the suggested estimator of missing data up-to first order of approximation is given as:-

$$\text{Bias}(\bar{y}_{pr}) = \bar{Y} + \{(1-k)\left(\frac{h_1^2}{2}\theta_{r,N}C_x^2 + h_1h_2\theta_{r,N}C_x^2 + \left(\frac{h_3^2}{2}+h_3h_4\right)\theta_{r,N}C_z^2 - h_1\theta_{r,N}\rho_{yx}C_yC_x - h_3\theta_{r,N}\rho_{yz}C_yC_z + h_1h_3\theta_{r,N}\rho_{xz}C_xC_z\right)\} \tag{4.4}$$

Proof: Equation (4.3) can be written as:-

$$(\bar{y}_{pr} - \bar{Y}) = \bar{Y}\{ke_0 + (1-k)(e_0 + h_3h_4e_3^2 - h_3e_3 - h_3e_0e_3 + \frac{h_3^2e_3^2}{2} + h_1h_2e_1^2 - h_1e_1 - h_1e_0e_1 + h_1h_3e_1e_3 + \frac{h_1^2e_1^2}{2})\} \tag{4.5}$$

$$\text{Bias}(\bar{y}_{pr}) = \bar{Y} + \{(1-k)\left(\frac{h_1^2}{2}\theta_{r,N}C_x^2 + h_1h_2\theta_{r,N}C_x^2 + \left(\frac{h_3^2}{2}+h_3h_4\right)\theta_{r,N}C_z^2 - h_1\theta_{r,N}\rho_{yx}C_yC_x - h_3\theta_{r,N}\rho_{yz}C_yC_z + h_1h_3\theta_{r,N}\rho_{xz}C_xC_z\right)\}$$

Theorem 4.2: Minimum MSE of the proposed is given by

$$\min \text{MSE}(\bar{y}_{pr}) = \bar{Y}^2\left\{A - \frac{C^2}{B}\right\} \tag{4.6}$$

where

$$A = \theta_{r,N}C_y^2 + h_1^2\theta_{r,N} + h_3^2\theta_{r,N}C_z^2 - 2h_1\theta_{r,N}\rho_{yx}C_xC_y - 2h_3\theta_{r,n}\rho_{yz}C_yC_z + 2h_1h_3\theta_{r,N}\rho_{xz}C_xC_z \tag{4.7}$$

$$B = h_1^2\theta_{r,N}C_x^2 + h_3^2\theta_{r,N}C_z^2 + 2h_1h_3\theta_{r,N}\rho_{xz}C_xC_z \tag{4.8}$$

$$C = h_1^2\theta_{r,N}C_x^2 + h_3^2\theta_{r,N} - h_1\theta_{r,N}\rho_{yx}C_yC_x - h_3\theta_{r,N}\rho_{yz}C_yC_z + 2h_1h_3\theta_{r,N}\rho_{xz}C_xC_z \quad (4.9)$$

Proof: Squaring equation (4.5) on both the sides we get

$$\left(\bar{y}_{pr} - \bar{Y}\right)^2 = \bar{Y}^2\{(e_0^2 + h_1^2e_1^2 + h_3^2e_3^2 - 2h_1e_0e_1 - 2h_3e_0e_3 + 2h_1h_3e_1e_3) + k^2(h_1^2e_1^2 + h_3^2e_3^2 + 2h_1h_3e_1e_3) - 2k(h_1^2e_1^2 + h_3^2e_3^2 - h_1e_0e_1 - h_3e_0e_3 + 2h_1h_3e_1e_3)\} \quad (4.10)$$

$$MSE(\bar{y}_{pr}) = \bar{Y}^2\left\{A + k^2B - 2kC\right\} \quad (4.11)$$

Where values of A, B and C are as given in equation (4.7), (4.8) and (4.9) respectively.

Optimum value of K is obtained from equation (4.11) as

$$k_{opt} = \frac{C}{B} \quad (4.12)$$

Substituting optimum value of k in equation(4.11) we get min MSE of suggested estimator as

$$\min MSE(\bar{y}_{pr}) = \bar{Y}^2\left\{A - \frac{C^2}{B}\right\}.$$

5. Empirical Study

We have considered following two populations for numerical illustration.

	Population I (Shukla and Thakur (2008))	Population II (Singh et.Al. (1994))
$\bar{Y}$	42.485	1031.82
$\bar{X}$	18.515	2934.58
$\bar{Z}$	20.52	3651.49
C_Y	0.3287	1.59749
C_X	0.3755	2.00625
C_Z	0.3296	2.48654

ρ_{YX}	0.8734	0.93
ρ_{YZ}	0.8667	0.77
ρ_{XZ}	0.9943	0.84
N	200	200
N	30	20
R	22	15

Table 5.1: Data statistics for two populations

Using data for two different populations, percentage relative efficiency(PRE) of various estimators is listed in Table 5.2.

	Population I	Population II
Estimator	PRE	PRE
$\bar{y}_{mean}$	100.00	100.00
$\bar{y}_{ratio}$	126.0851	125.8006
$\bar{y}_{comp}$	129.6281	130.5069
$\bar{y}_{pr1}$	**421.6341**	**740.1925**
$\bar{y}_{pr2}$	**421.6341**	**740.1925**
$\bar{y}_{pr3}$	**401.879**	**245.6399**
$\bar{y}_{pr4}$	**415.8384**	**432.598**
$\bar{y}_{pr5}$	**418.5873**	**544.9725**
$\bar{y}_{pr6}$	**412.177**	**347.8837**

Table 5.2: PRE of the estimators with respect to usual mean estimator

From the Table 5.2, it can be seen that our proposed estimator $\bar{y}_{pr}$ performs better than other estimators as it achieves the highest efficiency among all the estimators.

6. Conclusion

In this article under imputation we have proposed a generalized class of estimators. Our generalized class includes various estimators that are obtained after taking different values of constants. Properties of the proposed class of estimators are studied under first order of approximation and found that our proposed estimator $\overline{y}_{pr}$ has highest efficiency among all other estimators and $\overline{y}_{pr1}$ and $\overline{y}_{pr2}$ turn out to be the best choices. Thus we conclude that our proposed method of imputation is preferable over other imputation methods considered in this article.

References

1. Abd-Elfattah, A.M., Sherpeny E.A., Mohamed S.M. and Abdou O.F. (2010). Improvement in estimating the population mean in simple random sampling using information on auxiliary attribute, Applied Mathematics and Computation, 215, p. 4198-4202.
2. Ahmed, M. S., Al-Titi, O., Al-Rawi, Z. and Abu-Dayyeh, W. (2006). Estimation of a population mean using different imputation methods, Statistics in Transition, 7(6), p. 1247, 1264.
3. Al-Omari, A. I., Bouza, C. N. and Herrera, C. (2013). Imputation methods of missing data for estimating the population mean using simple random sampling with known correlation coefficient, Quality & Quantity, 47(1), p. 353-365.
4. Kadilar, H. Cingi. (2005). A new estimator using two auxiliary variables, Applied Mathematics and Computation, 162, p. 901–908.
5. Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. In Proceedings of the Section on Survey Research Methods, American Statistical Association (Vol. 22, p. 31). American Statistical Association Cincinnati.
6. Khoshnevisan, M., Singh, R., Chauhan, P., Sawan, N. and Smarandache, F., (2007). A general family of estimators for estimating population mean using known value of some population parameter(s), Far East Journal of Theoretical Statistics, 22, p. 181–191.
7. Lee, H., and Särndal, C. E. (1994). Experiments with variance estimation from survey data with imputed values, Journal of Official Statistics, 10(3), p. 231-243.
8. Pandey, R., Thakur, N. S. and Yadav, K. (2015). Estimation of population mean using exponential ratio type imputation method under survey non-response, Journal of the Indian Statistical Association, 53(1), 89-107.
9. Sanaullah, A., Khan, H., Ali, H. A. and Singh, R. (2012). Improved exponential ratio-type estimators in survey sampling, Journal of Reliability and Statistical Studies, 5(2), p. 119-132.
10. Shabbir, J. and Gupta, S. (2015). A note on generalized exponential type estimator for population variance in survey sampling, Revista Colombiana de Estadística, 38(2), p. 385-397.

11. Shukla, D., Thakur, N. S., Pathak, S. and Rajput, D. S. (2009). Estimation of mean under imputation of missing data using factor-type estimator in two-phase sampling, Statistics in Transition, 10(3), p. 397-414.
12. Shukla, D., Thakur, N. S., Thakur, D. S. and Pathak, S. (2011). Linear combination based imputation method for missing data in sample, International Journal of Modern Engineering Research, Vol. 1(2), p-580-596.
13. 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), p. 211-217.
14. Singh, R., Chauhan, P., Sawan, N. and Smarandache, F. (2008. Ratio estimators in simple random sampling using information on auxiliary attribute, Pak. J. Stat. Oper. Res. 4(1), p. 47-53.
15. Singh, R. and Kumar, M. (2011). A note on transformations on auxiliary variable in survey sampling, MASA, 6:1, p. 17-19.
16. Singh, R., Malik, S., Chaudhary, M.K., Verma, H. and Adewara, A. A. (2012). A general family of ratio type- estimators in systematic sampling, Journal of Reliability and Statistical Studies, 5(1), p. 73-82).
17. Singh, S. and Horn, S. (2000). Compromised imputation in survey sampling, Metrika, 51(3), p. 267-276.
18. Thakur, N. S., Kalpana, Y. and Sharad, P. (2013). On mean estimation with imputation in two-phase sampling design, Research Journal of Mathematical and Statistical Sciences, 1(3), p. 1-9.
19. Thakur, N. S., Yadav, K. and Pathak, S. (2012). Some imputation methods in double sampling scheme to estimate the population mean, Research Journal of Mathematical and Statistical Sciences, 1(9), 1-10.
20. Upadhyaya, L. N. and Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population mean, Biometrical. Journal, 41(5), p. 627-636.