Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.18576/jsap/060304

A Class of Exponential Chain Type Estimator for Population Mean With Imputation of Missing Data Under Double Sampling Scheme

Abhishek Kumar[1,*], *Ajeet Kumar Singh*[2], *Priyanka Singh*[1] *and V. K. Singh*[1]

[1] Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi - 221005, India.
[2] Statistical Assistant, Directorate of Economics and Statistics, Civil Lines, New Delhi - 110054, India.

Received: 10 May 2017, Revised: 2 Jul. 2017, Accepted: 14 Jul. 2017
Published online: 1 Nov. 2017

Abstract: This paper aims to deal with the problem of non-response, by suggesting an exponential chain type class of imputation technique and corresponding point estimator in double sampling has been proposed for estimating finite population mean of the study variable when the information on another additional auxiliary variable is available along with the main auxiliary variable. The bias and mean square error of the proposed strategy have been obtained. Theoretical and empirical studies have been done to demonstrate the supremacy of the proposed strategy with respect to the strategies which utilize the information on one and two auxiliary characteristics.

Keywords: Auxiliary information, double sampling, exponential chain type estimator, imputation, non-response

1 Introduction

In survey sampling situations, it is common to utilize auxiliary information to improve the precision of an estimator of unknown population parameter of interest under the assumption that all the observations in the sample are available, but in real sense, sometimes this assumption is not fulfilled. This is the case of incomplete information which arises due to some non-response in the sample data. Incomplete information is a frequent issue in sample surveys and a common technique for managing it is imputation, where the missing values are filled in to create a complete data set that can be analyzed.

The use of ratio, product and linear regression strategies in survey sampling solely depends upon the knowledge of population mean $\bar{X}$ of the auxiliary variable X. However, in many circumstances, the population mean $\bar{X}$ is not known well in advance. In such a case, the two-phase (or double) sampling design is adopted to get an estimate of $\bar{X}$ with the aid of sample mean of a preliminary large sample on which only the variable X is measured. Sometimes, information on another additional auxiliary variable is available, which is relatively cheaper and less correlated to the main variable in comparison to the main auxiliary variable X. In such condition, this information may be used to get more efficient estimators of unknown $\bar{X}$, on the basis of ratio, product and regression-type estimators utilizing the second auxiliary character.

Chand (1975) [2] introduced a technique of chaining the information on auxiliary variables with the main variable. Further, his work was extended by Kiregyera (1980, 1984) [3,4], Mukerjee et.al. (1987) [5], Srivastava et.al. (1989) [10], Upadhyaya et. al. (1990) [12], Singh and Singh (1991) [8], Singh et.al. (1994) [9] and many others.

This paper aims at (i) to suggest new imputation technique on the basis of two auxiliary variables, (ii) to define family of point estimators for the population mean $\bar{Y}$, which are chain type estimators, (iii) to make a study of proposed imputation strategy in respect of bias and MSE and show its supremacy over some existing strategies and (iv) to apply the proposed strategy on some empirical populations for illustration purpose.

[*] Corresponding author e-mail: abhionline91@gmail.com

2 Problems and Notations

Let we have a finite population $U = \{U_1, U_2, ..., U_N\}$ of size N in which the study variable be Y whose mean $\overline{Y}$ is to be estimated. Let there be two auxiliary variables X and Z available in the population such that the variable X is highly correlated with Y and the variable Z is also correlated with Y but not with as much high correlation as between X and Y. We assume that the mean $\overline{X}$ is not known, hence it is to be estimated through a preliminary sample $s^{'}$ of size m. Further, let a sub-sample s of size n be selected from the preliminary sample $s^{'}$ $(n < m)$ which consists of r responding units belonging to a set R and $(n-r)$ non-responding units belonging to the set R^c such that $s = R \cup R^c$. Therefore, for every unit $i \in R$, the value y_i is observed and for the unit $i \in R^c$, the value y_i is missing for which suitable imputed value is to be derived.

We shall use the following notations:

$\overline{X}, \overline{Y}, \overline{Z}$: The population mean of the variates X, Y and Z respectively.
$\overline{x_m}, \overline{z_m}$: The sample mean of the variable X and Z respectively based on the sample $s^{'}$.
$\overline{y_r}, \overline{x_r}$: The sample mean of the variable Y and X respectively obtained for the set R.
ρ_{ab} : The coefficient of correlation between the variable a and b.
C_X, C_Y, C_Z : The coefficient of variation of X, Y and Z respectively.
$\theta_{a,b}$: The finite population correction (fpc) given by $\theta_{a,b} = (\frac{1}{a} - \frac{1}{b})$.

3 SOME EXISTING IMPUTATION METHODS

Some classical methods of imputation, which are available and frequently used, are as follows:

3.1 Mean Method of Imputation

Under mean method of imputation, the imputation scheme is

$$y_{.i} = \begin{cases} y_i & \text{if } i \in R \\ \overline{y_r} & \text{if } i \in R^c \end{cases} \tag{1}$$

the corresponding point estimator for the population mean $\overline{Y}$ is:

$$\overline{y_M} = \overline{y_r} = \frac{1}{r} \sum_{i \in R} y_i \tag{2}$$

The bias and mean square error of the estimator are derived as

$$B(\overline{y_M}) = 0 \tag{3}$$

$$M(\overline{y_M}) = V(\overline{y_M}) = \theta_{r,N} \overline{Y}^2 C_Y^2 \tag{4}$$

3.2 Ratio Method of Imputation

Under this method, the imputation scheme is given as

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

where

$$\hat{b} = \frac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i}; \tag{6}$$

therefore, the point estimator is

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{7}$$

For this estimator, the bias and mean square error are obtained as

$$B(\bar{y}_{RAT}) = \theta_{r,n}\bar{Y}[C_X^2 - \rho_{YX}C_YC_X] \tag{8}$$

$$M(\bar{y}_{RAT}) = \theta_{r,n}\bar{Y}^2C_Y^2 + \theta_{r,n}\bar{Y}^2[C_X^2 - 2\rho_{YX}C_YC_X]. \tag{9}$$

3.3 Compromised Method of Imputation

Singh and Horn (2000) [7] suggested this method of imputation, in which the data after imputation becomes

$$y_{.i} = \begin{cases} p\frac{n}{r} + (1-p)\hat{b}x_i & \text{if } i \in R \\ (1-p)\hat{b}x_i & \text{if } i \in R^c \end{cases} \tag{10}$$

where p is a suitably chosen constant, such that the variance of the resultant estimator is minimum. In this case, they used information from imputed values for the responding units in addition to non-responding units.

Thus, the point estimator of the population mean $\bar{Y}$ under the compromised method of imputation becomes

$$\bar{y}_{COMP} = p\bar{y}_r + (1-p)\bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{11}$$

The bias and mean square error of $\bar{y}_{COMP}$ are given as

$$B(\bar{y}_{COMP}) = (1-p)\theta_{r,n}\bar{Y}[C_X^2 - \rho_{YX}C_YC_X] \tag{12}$$

$$M(\bar{y}_{COMP}) = \theta_{r,n}\bar{Y}^2C_Y^2 + \theta_{r,n}\bar{Y}^2[(1-p)^2C_X^2 - 2(1-p)\rho_{YX}C_YC_X]. \tag{13}$$

For obtaining optimum value of p, i.e., p_{opt}, the MSE of $\bar{y}_{COMP}$ is minimized with respect to the constant p, then we get

$$p_{opt} = 1 - \rho_{YX}\frac{C_Y}{C_X} \tag{14}$$

and

$$M(\bar{y}_{COMP})_{min} = \bar{Y}^2[(\theta_{r,N} - \theta_{r,n}\rho_{YX}^2)C_Y^2]. \tag{15}$$

4 PROPOSED IMPUTATION STRATEGY

Bahl and Tuteja (1991) [1] defined the exponential - type estimator for population mean which was observed to be better in a wide range of $\rho_{YX}\frac{C_Y}{C_X}$ as compared to the usual mean estimator.
Movitaved by Bahl and Tuteja (1991), we here propose the following exponential type method of imputation, given as

$$y_{.i} = \begin{cases} y_i & \text{if } i \in R \\ \frac{1}{(n-r)}[n\bar{y}_rT - r\bar{y}_r] & \text{if } i \in R^c \end{cases} \tag{16}$$

where

$$T = \exp\left\{\alpha\left(\frac{\bar{x}_m\frac{\bar{Z}}{\bar{z}_m} - \bar{x}_r}{\bar{x}_m\frac{\bar{Z}}{\bar{z}_m} + \bar{x}_r}\right)\right\}; \tag{17}$$

α being a parameter, whose value can be obtained such that the variance of T is minimum.

Under the proposed method of imputation the point estimator for the population mean $\bar{Y}$ is obtained as

$$\bar{y}_T = \bar{y}_r T = \bar{y}_r \exp\left\{\alpha\left(\frac{\bar{x}_m \frac{\bar{Z}}{\bar{z}_m} - \bar{x}_r}{\bar{x}_m \frac{\bar{Z}}{\bar{z}_m} + \bar{x}_r}\right)\right\}. \qquad (18)$$

Obviously, $\bar{y}_T$ defines a family of chain type exponential estimator for $\bar{Y}$. We observe that for $\alpha = 0, \bar{y}_T = \bar{y}_r$, the point estimator under mean method of imputation.

For $\alpha = 1$ and -1 , $\bar{y}_T$ reduces to exponential chain ratio type estimator and exponential chain product type estimator respectively.

4.1 Bias and MSE of the Proposed Estimator

The bias B(.) and mean square error M(.) of the suggested estimator $\bar{y}_T$ up to the first order of approximations can be obtained under the following large sample approximations:

$\bar{y}_r = \bar{Y}(1+e_1)$; $\bar{x}_r = \bar{X}(1+e_2)$; $\bar{x}_m = \bar{X}(1+e_3)$ and $\bar{z}_m = \bar{Z}(1+e_4)$, such that $E(e_i) = 0$ for i = 1,2,3,4 and

$E(e_1^2) = \theta_{r,N}C_Y^2$; $E(e_2^2) = \theta_{r,N}C_X^2$; $E(e_3^2) = \theta_{m,N}C_X^2 = E(e_2e_3)$; $E(e_4^2) = \theta_{m,N}C_Z^2$; $E(e_1e_2) = \theta_{r,N}\rho_{YX}C_YC_X$;
$E(e_1e_3) = \theta_{m,N}\rho_{YX}C_YC_X$; $E(e_1e_4) = \theta_{m,N}\rho_{YZ}C_YC_Z$; $E(e_2e_4) = E(e_3e_4) = \theta_{m,N}\rho_{XZ}C_XC_Z$

Obviously, $B(\bar{y}_T) = E[\bar{y}_T - \bar{Y}] = E[\bar{y}_T] - \bar{Y}$.

Writing $\bar{y}_T$, given in (18), in terms of $e_i^{'}$s $(i = 1,2,3,4)$ and then using usual large sample approximations, the bias of the estimator $\bar{y}_T$, up to the order $O(n^{-1})$ is derived as:

$$B(\bar{y}_T) = \bar{Y}\left[\theta_{r,m}\left\{\left(\frac{\alpha}{4} + \frac{\alpha^2}{8}\right)C_X^2 - \frac{\alpha}{2}\rho_{YX}C_YC_X\right\} + \theta_{m,N}\left\{\left(\frac{\alpha}{4} + \frac{\alpha^2}{8}\right)C_Z^2 - \frac{\alpha}{2}\rho_{YZ}C_YC_Z\right\}\right]. \qquad (19)$$

Further, since, the expression of $M(\bar{y}_T)$, upto the first order approximation will be

$$M(\bar{y}_T) = E[\bar{y}_T - \bar{Y}]^2,$$

$$M(\bar{y}_T) = \bar{Y}^2\left[\theta_{r,N}C_Y^2 + \frac{\alpha^2}{4}\theta_{r,m}C_X^2 + \frac{\alpha^2}{4}\theta_{m,N}C_Z^2 - \alpha\theta_{r,m}\rho_{YX}C_YC_X - \alpha\theta_{m,N}\rho_{YZ}C_YC_Z\right]. \qquad (20)$$

Since the proposed imputation strategy depends on a constant α, it is therefore, desirable to obtain the optimum value of the constant, i.e., $\alpha_{opt}$and then using it in the expression of MSE so as to obtain minimum MSE. Hence differentiating the MSE equation with respect to the constant, we get the optimum value of the constant, which is

$$\alpha_{opt} = \frac{2\theta_{r,m}\rho_{YX}C_YC_X + 2\theta_{m,N}\rho_{YZ}C_YC_Z}{\theta_{r,m}C_X^2 + \theta_{m,N}C_Z^2}. \qquad (21)$$

On substituting the value of α_{opt} in the equation (20), one can get the minimum MSE of the estimator.

5 EMPIRICAL STUDY

It can be seen that theoretical comparison of the proposed strategy $\bar{y}_T$ with $\bar{y}_M$, $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$ yield no exclusive results, hence it is advisable to compare them on the basis of some practical examples. The various results obtained in previous sections are now examined with the help of two sets of empirical data:

5.1 Population P_1

This data has been taken from Sukhatme and Chand (1977) [11] which has been reproduced in Singh et. al. (1994) [9]. The particulars of the data are as under:

Y: Apple trees of bearing age in 1964
X: Bushels of apples harvested in 1964
Z: Bushels of apples harvested in 1959

For the data we have the following parametric values:
$\overline{Y} = 0.103182X10^4$; $\overline{X} = 0.293458X10^4$; $\overline{Z} = 0.365149X10^4$; $C_Y^2 = 2.55280$; $C_X^2 = 4.02504$; $C_Z^2 = 2.09379$; $\rho_{YX} = 0.93$; $\rho_{YZ} = 0.77$; $\rho_{XZ} = 0.84$.

The combination of r, n, m and N are respectively taken as (15, 20, 30, 200).

5.2 Population P_2

This data was artificially generated for three variables Y, X and Z by Shukla and Thakur (2008) [6]. Considering Y as study variable and X and Z respectively the main and additional auxiliary variables, we get the following population values:
$N = 200$; $\overline{Y} = 42.485$; $\overline{X} = 18.515$; $\overline{Z} = 20.52$; $C_Y = 0.3287$; $C_X = 0.3755$; $C_Z = 0.3296$; $\rho_{YX} = 0.8734$; $\rho_{YZ} = 0.8667$; $\rho_{XZ} = 0.9943$.

For the purpose, we select r = 22, n = 30, m = 80.
The values of MSE of the different strategies are shown in the following table. The table also depicts the percent relative efficiencies (PREs) of different strategies, with respect to $\overline{y}_M$.

In the following table, we have considered MSEs of $\overline{y}_{COMP}$ and $\overline{y}_T$ as obtained under their optimality condition.

Table 1: MSEs and PREs of Different Estimators

Estimators	**Population P$_1$**		**Population P$_2$**	
	MSE	**PRE**	**MSE**	**PRE**
$\overline{y}_M$	167600.40	100	7.89	100
$\overline{y}_{RAT}$	133227.10	125.80	6.26	126.09
$\overline{y}_{COMP}$	128422.70	130.51	6.09	129.63
$\overline{y}_T$	44114.26	379.92	1.90	415.09

From the table, it is evident that the proposed strategy dominates other imputation strategies in respect to its performance for both the populations. This suggests that even if the population mean of the auxiliary variable X is not known, the chaining of estimators technique is fruitful in order to improve the efficiency of the estimator.

6 CONCLUSIONS

The work presented in this paper, suggests an efficient imputation method based upon exponential type estimator and the concept of chaining the estimators when information on an additional auxiliary variable other than the main auxiliary variable is available. The method suggested has been seen to be precise enough in comparison of some existing imputation methods such as mean method, ratio method and compromised method.

References

[1] Bahl, S. and Tuteja, R. (1991): Ratio and product type exponential estimators, Journal of Information and Optimization Sciences, **12**, 159-164.

[2] Chand, L. (1975): Some Ratio-Type Estimators Based on Two or More Auxiliary Variables, Unpublished Ph.D. Thesis submitted to Iowa State University, Ames, Iowa, U.S.A.

[3] Kiregyera, B. (1980): A chain ratio-type estimator in finite population double sampling using two auxiliary variables, Metrika, **27**, 217-223.

[4] Kiregyera, B. (1984): Regression-type estimators using two auxiliary variables and the model of double sampling, Metrika, **31**, 215-226.

[5] Mukherjee, R., Rao, T. and Vijayan, K. (1987): Regression type estimators using multiple auxiliary information, Australian Journal of Statistics, **29**, 244-254.

[6] Shukla, D. and Thakur, N. S. (2008): Estimation of mean with imputation of missing data using factor-type estimator, Statistics in Transition, **9**, (1), 33-48.

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

[8] Singh, V. K. and Singh, G. N. (1991): Chain type regression estimators with two auxiliary variables under double samplimg scheme, Metron, **49**, 279-289.

[9] Singh, V. K., Singh, H. P., Singh, H. P. and Shukla, D. (1994): A general class of chain estimators for ratio and product of two means of a finite population, Communications in Statistics ? Theory and Methods, **23**, 1341-1355.

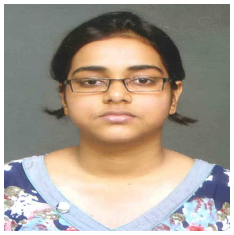
[10] Srivastava, R. S., Srivastava, S. and Khare, B. (1989): Chain ratio type estimator for ratio of two population means using auxiliary characters, Communications in Statistics ? Theory and Methods, **18**, 3917-3926.

[11] Sukhatme, B. V. and Chand, L. (1977): Multivariate ratio-type estimators, Proceedings of the American Statistical Association, Social Statistics Section, 927-931.

[12] Upadhyaya, L., Kushwaha, K. and Singh, H. (1990): A modified chain ratio-type estimator in two-phase sampling using multi auxiliary information, Metron, **48**, (1-4), 381-393.

Abhishek Kumar is research fellow in the Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India. He has obtained M.Sc. (Statistics) degree from Banaras Hindu University, Varanasi, India. His research interests are in the areas of sampling theory and methods of estimation.

Ajeet Kumar Singh is statistical assistant in Directorate of Economics and Statistics Vikas Bhawan -2 Civil lines. He received the Ph.d degree in Statistics. He has published more than ten research articles in reputed international journal.



Priyanka Singh is research fellow in the Department of Statistics, Institute of Science. Banaras Hindu University, Varanasi, India. She has completed M.Sc (Statistics) degree from BHU. Field of specializations is sampling theory. Published 8 research papers in refereed journals.

V. K. Singh presently working as Professor of Statistics, Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India since 2000. Joined the Department as Assistant Professor in 1972. Did M.Sc (Statistics) and Ph.D. (Statistics) from Banaras Hindu University in 1972 and 1979 respectively. Having 45 years teaching experience and 43 years research experience. Field of specializations are Sampling Theory, Stochastic Modelling, Mathematical Demography and Operations Research. Published 89 research papers in reputed international/national journals. Guided 15 Ph.D. scholars for their Ph.D. Degree. Visited United Kingdom, Australia and Sri Lanka for attending International Conferences and organizing Symposiums. Convened 2 national conferences. Life member of Indian Statistical Association, Member of International Association of Survey Statisticians (IASS), Associate Editor of Assam Statistical Review, India.