Journal of Statistics Applications & Probability
An International Journal

http://dx.doi.org/10.12785/jsap/030211

Exponential - Type Compromised Imputation in Survey Sampling

*Ajeet Kumar Singh**, *Priyanka Singh and V. K. Singh*

Department of Statistics, Banaras Hindu University, Varanasi-221005, India

Received: 2 Mar. 2014, Revised: 8 May 2014, Accepted: 10 May 2014
Published online: 1 Jul. 2014

Abstract: To deal with the problems of non - response, one parameter classes of imputation techniques have been suggested and their corresponding point estimators have been proposed. The proposed estimator is more efficient than several other estimators. A design based approach is used to compare the proposed strategy with existing strategies. Theoretical results have been verified through empirical studies handling real data set examples.

Keywords: Non - response, missing data, Imputation methods, Bias, Mean Square Error.

1 Introduction:

Incomplete data or non - response in the form of missingness, censoring or groupings are troubling issues for many data sets. Statisticians have recognized for some time that failure to account for the stochastic nature of incompleteness or non - response can spoil the nature of data . There are several factors that affect the non - response rate in any particular inquiry. Hansen and Hurwitz (1946) were the first to deal with the problem of incomplete samples in mail surveys. Mail surveys or telephone surveys are commonly used by bureaucratic or business organizations because of their low cost . In respect of non - response, Rubin (1976) defined two key concepts: Missing at random (MAR) and Observed at random (OAR).

1.1 Missing at random (MAR):

The data are MAR if the probability of observed missingness pattern given the observed and unobserved data does not depend on the values of the unobserved data. It will, therefore, include cases where the enumerator is not able to contact the respondents only by chance and had he been able to contact, the data would have been collected. For example when the information is kept on punched cards, the non - response due to the accidental loss of one or more cards is of the first category. Although this illusion is rather outdated in the world of modern computing but still there is a chance that some data files may get damaged due to virus attacks. This type of non - responce is called random non - response.

1.2 Observed at random (OAR):

The data are OAR, if for every possible value of the missing data, the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the observed data.The combination of MAR and OAR is called MCAR. In other words, the MCAR can be defined as $f(A|D)=f(A)$ for all D, where D is the data matrix and A is the missing data indicator matrix ($a_{ij}=1$ if d_{ij} is reported, $a_{ij}=0$ otherwise). Heitjan and Basu (1996) have also considered the problem of distinguishing between MAR and MCAR. Note that the concept of OAR is vestige

* Corresponding author e-mail: ajeetvns.singh@gmail.com

of Rubin (1976). Now a days people jump right from MAR to MCAR, which is a logical step and quite easy to follow . Among other methods to deal with the problem of non - response, one of the popular method is to impute the non - response units by suitably selected respondent units of the population.

Let $\bar{Y}$ be the mean of the finite population U of size N. A simple random sample without replacement (SRSWOR), s, of size n is drawn from U to estimate $\bar{Y}$. Let r be the number of responding units out of sampled n units. Let the set of responding units be denoted by A and that of non - responding units is denoted by A^C. For every unit $i \in$ A, the value y_i is observed. However for the units $i \in A^C$, the y_i values are missing and hence for them imputed values are derived . We assume that imputation is carried out with the aid of an auxiliary variable X such that X_i, the value of X for unit i, is known and positive for every $i \in s = A \cup A^C$. In other words, the data $x_i : i \in s$ are known.

2 Notations:

Let $U = U_1, U_2, \cdots U_N$ be the finite population of size of N and the character under study be denoted by Y. It is assumed that information on an auxiliary variate X (with the known population mean) is available at the beginning of the survey. A simple random sample (without replacement) s of n units is drawn from the population. Let the number of responding units out of sampled n units be denoted by r, the set of responding units by A, and the non responding unit by A^C. For every unit $i \in$ A the value y_i observed, but for the units $i \in A^C$, the y_i values are missing and for them imputed values are derived. The imputation is carried out with the aid of a quantitative auxiliary variate X, such that X_i, the value of X for unit i, is known for each $i \in$ s.

The following notations are used hereafter:
$\bar{X}, \bar{Y}$:The population mean of the variates X and Y respectively.
$\bar{x}_n$: The sample mean of X for the sample of size n.
$\bar{y}_r$: The mean of the variable Y for the set A.
ρ_{yx}: The correlation coefficient between the variates Y and X.
S_X^2, S_Y^2: The population mean squares of X and Y respectively.
C_X, C_Y: The coefficient of variation of X and Y respectively.

3 Some imputation methods:

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

3.1 Mean method of imputation:

Under this method, the study variate after imputations takes the form,

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \bar{y}_r & \text{if} \quad i \in A^c \end{cases} \tag{1}$$

Under this method of imputation, the point estimator of the population mean $\bar{Y}$ is given by

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

Where

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

3.2 Ratio method of imputation:

$$y_{.i} = \begin{cases} y_i & \text{if} \quad i \in A \\ \hat{b}x_i & \text{if} \quad i \in A^c \end{cases} \tag{3}$$

Under this method of imputation, the point estimator of the population mean $\bar{Y}$ is given by

$$\bar{y}_{RAT} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r} \tag{4}$$

Where

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

and

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

3.3 Compromised method of imputation:

Singh and Horn (2000) proposed the compromised imputation procedure, where the study variate after imputation takes the form

$$y_{.i} = \begin{cases} \alpha\frac{n}{r}y_i + (1-\alpha)\hat{b}x_i & \text{if} \quad i \in A \\ (1-\alpha)\hat{b}x_i & \text{if} \quad i \in A^c \end{cases} \tag{5}$$

Where α is a suitably chosen constant, such that the variance of the resultant estimator is minimum. In this case the information from the imputed values for the responding units is also used in addition to that from non - responding units.Thus the point estimator of the population mean under the above imputation method becomes,

$$\bar{y}_{COMP} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r\frac{\bar{x}_n}{\bar{x}_r} \tag{6}$$

On similar lines, Ahmed et al. (2006) proposed several new imputation techniques by introducing some unknown parameters and hence proposed the corresponding estimators for estimating the finite population mean $\bar{Y}$.

4 Proposed methods of imputation:

Motivated with Bahl and Tuteja (1991), we here propose the following exponential - type method of imputation

$$y_{.i} = \begin{cases} k\frac{n}{r}y_i + (1-k)\bar{y}_r \exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right) & if \quad i \in A \\ (1-k)\bar{y}_r \exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right) & if \quad i \in A^c \end{cases} \tag{7}$$

Which may be termed as exponential - type compromised imputation .

The point estimator of the population mean $\bar{Y}$ under the proposed method of imputation is

$$\bar{y}_{ET} = k\bar{y}_r + (1-k)\bar{y}_r exp\left(\frac{\bar{X}-\bar{x}_r}{\bar{X}+\bar{x}_r}\right) \tag{8}$$

4.1 Properties of the proposed estimator $\bar{y}_{ET}$:

The bias B(.) and mean square error M(.) of the estimator $\bar{y}_{ET}$ up to the first order of approximations is derived under the following transformations:

$$\bar{y}_r = \bar{Y}(1+e_1), \bar{x}_r = \bar{X}(1+e_2)$$

and

$$\bar{x}_n = \bar{X}(1+e_3)$$

such that

$$|e_i| < 1 \quad \forall \quad i = 1,2,3$$

Hence we have

$$E(e_i) = 0, \quad i = 1,2,3,; \quad E(e_1^2) = \frac{V(\bar{y}_r)}{Y^2}, \quad E(e_2^2) = \frac{V(\bar{x}_r)}{X^2}, \quad E(e_3^2) = \frac{V(\bar{x}_n)}{X^2}$$

Under the above transformations the estimator takes the following form :

$$\bar{y}_{ET} = k\bar{Y}(1+e_1) + (1-k)\bar{Y}(1+e_1)exp\left\{-\frac{e_2}{2}\left(1+\frac{e_2}{2}\right)^{-1}\right\} \tag{9}$$

Now we have the following theorems,

4.2 Theorem

The bias of the proposed estimator $\bar{y}_{ET}$ to the first order of approximations is given by

$$B(\bar{y}_{ET}) = (1-k)\left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}\left[\frac{3}{8}C_X^2 - \frac{1}{2}\rho_{YX}C_XC_Y\right] \tag{10}$$

Proof: we have

$$B(\bar{y}_{ET}) = E\left[\bar{y}_{ET} - \bar{Y}\right] = E\left[k\bar{Y}(1+e_1) + (1-k)\bar{Y}(1+e_1)exp\left\{-\frac{e_2}{2}\left(1+\frac{e_2}{2}\right)^{-1}\right\} - \bar{Y}\right] \tag{11}$$

Writing the expression of $\bar{y}_{ET}$ in terms of $e_i's$, expanding the right hand side of the above expression, taking expectations and collecting the terms up to the first order of approximations, we get the expression for bias of the estimator as given in (10)

4.3 Theorem

The mean square error of the proposed estimator up to the first order of approximations is given by

$$M(\bar{y}_{ET}) = \left(\frac{1}{r}-\frac{1}{N}\right)\bar{Y}^2\left[C_Y^2 + \frac{(1-k)^2}{4}C_X^2 - (1-k)\rho_{YX}C_XC_Y\right] \tag{12}$$

Proof:
By the definition of mean square error we have

$$M(\bar{y}_{ET}) = E\left[\bar{y}_{ET} - \bar{Y}\right]^2$$

Now using the expression given in equation (9) for $\bar{y}_{ET}$, expanding the terms and taking expectations and retaining the terms up to the first order of approximations we get expression for mean square error as given in equation (12)

4.4 Minimum mean square error of $\bar{y}_{ET}$

The mean square error of $\bar{y}_{ET}$ as given in (12) is a function of unknown constant k. Therefore , it is natural to search for an optimum value of k, such that the mean square error of the proposed estimators becomes minimum . Hence differentiating equation (12) with respect to k and equating to zero we get optimum value of k as

$$k = 1 - 2\rho_{YX}\frac{C_Y}{C_X} \tag{13}$$

4.5 Theorem

Putting the value of k as given in equation (13) in the equation (12) the minimum mean square error of $\bar{y}_{ET}$ is derived as

$$M(\bar{y}_{ET})_{min} = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2\left[C_Y^2(1-\rho_{XY}^2)\right] \tag{14}$$

In order to compare the proposed estimator $\bar{y}_{ET}$ with the imputed estimators $\bar{y}_r$, $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$, we give below the expression of bias and mean square error of these estimators . We have

$$B(\bar{y}_r) = 0; \quad V(\bar{y}_r) = \left(\frac{1}{r} - \frac{1}{N}\right)S_Y^2 \tag{15}$$

$$B(\bar{y}_{RAT}) = \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left[C_X^2 - \rho_{XY}C_YC_X\right] \tag{16}$$

$$M(\bar{y}_{RAT}) = \left(\frac{1}{n} - \frac{1}{N}\right){S_Y}^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left[S_Y^2 + R^2S_X^2 - 2RS_{XY}\right] \tag{17}$$

$$B(\bar{y}_{COMP}) = (1-\alpha)\left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}\left[C_X^2 - \rho_{XY}C_YC_X\right] \tag{18}$$

$$M(\bar{y}_{COMP}) = \left(\frac{1}{r} - \frac{1}{N}\right)\bar{Y}^2C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\bar{Y}^2\left[(1-\alpha)^2C_X^2 - 2(1-\alpha)\rho_{XY}C_XC_Y\right] \tag{19}$$

and

$$M(\bar{y}_{COMP})_{opt.} = M(\bar{y}_{RAT}) - \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - \rho_{YX}\frac{C_Y}{C_X}\right)^2\bar{Y}^2C_X^2 \tag{20}$$

Where

$$\alpha_{opt} = 1 - \rho_{YX}\frac{C_Y}{C_X}$$

5 Comparison of mean square errors:

On the basis of expressions of mean square errors of the proposed estimator $\bar{y}_{ET}$ with those of estimators $\bar{y}_r$, $\bar{y}_{RAT}$ and $\bar{y}_{COMP}$, we can observe the efficiency of the proposed estimator.

5.1

Comparing expressions (12) and (15), we observe that

$$M(\bar{y}_{ET})_{min} < V(\bar{y}_r)$$

when

$$k > 1 - 4\rho_{XY}\frac{C_Y}{C_X} \quad if \quad k < 1 \tag{21}$$

and

$$k < 1 - 4\rho_{XY}\frac{C_Y}{C_X} \quad if \quad k > 1 \tag{22}$$

Further it can be seen that $M(\bar{y}_{ET})_{min}$ is always smaller than $V(\bar{y}_r)$

5.2

Comparing expression (14) and (17), it is easy to see that

$$M(\bar{y}_{ET})_{min} < M(\bar{y}_{RAT})$$

if

$$\bar{Y}^2\left[\left(\frac{1}{r}-\frac{1}{n}\right)(C_X-\rho_{YX}C_Y)^2+\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{YX}^2C_Y^2\right] > 0 \tag{23}$$

Which is always true . Hence the estimator $\bar{y}_{ET}$ is always precised than the ratio method of imputation under optimality condition (13)

5.3

Finally comparison of the proposed imputation stratagy may be made with the compromised imputation strategy proposed by Singh and Horn (2000) . Using expression (14) and (20) we observe that,

$$M(\bar{y}_{COMP})_{opt} - M(\bar{y}_{ET})_{min} = \bar{Y}^2\left(\frac{1}{n}-\frac{1}{N}\right)\rho_{YX}^2C_Y^2 \tag{24}$$

Which is always true .Thus it can be concluded that it is always advisable to prefer exponential-type imputation strategy over compromised imputation strategy.

6 Empirical Study

For the empirical study of the proposed strategy with other existing imputation strategies we consider the following data.

Population 1. (Source: Mukhopadhyaya (2000)) .The population consists of N=20 jute mills. The data show the numbers of labourers X (in thousands) and quantity of raw materials required Y (in lakhs of bales).
Here we take n=7 and r=5. Further for the data, we have;

$\bar{X}$=441.95, $\bar{Y}$=41.5, S_Y^2=95.7368, S_X^2=10215.21, C_X=0.2286, C_Y=0.2358, ρ_{XY}=0.6521

Population 2. (Source: Giancarlo Diana and Pier Francesco Perri) . The data are taken from the survey of Household income and Wealth conducted by The Bank of Italy for the year (2002). The survey covers 8,011. Italian households composed of 22,148 individuals and 13,536 income-earners .In the analysis, we assume the 8,011 households as the target population on which the household net disposal income (Y) and the number of household income earners (X) are investigated . The following values are obtained for the considered variables.
N=8011,
we take $n = 400$, $r = 250$, $\bar{Y}$=28229.43, $\bar{X}$=1.69, S_Y=22216.56, S_X=0.78, ρ_{YX}=0.46

The following tables depicts the bias and mean square errors of different imputation strategies for the two populations.

In both the tables, MSEs of $\bar{y}_{ET}$ and $\bar{y}_{COMP}$ are minimum MSEs.

Table 1: Bias and Mean square errors (for Population. 1)

Estimators	Bias	MSE
$\bar{y}_r$	0	14.361
$\bar{y}_{RAT}$	0.0406	12.586
$\bar{y}_{COMP}$	0.1394	12.034
$\bar{y}_{ET}$	0.0169	8.2521

Table 2: Bias and Mean square errors (for Population.2)

Estimators	Bias	MSE
$\bar{y}_r$	0	1912690.2
$\bar{y}_{RAT}$	1.9449	1767867.7
$\bar{y}_{COMP}$	2.499	1756029.1
$\bar{y}_{ET}$	3.1236	1507964.7

7 Conclusions

The tables show that for both the populations, it is advisible to prefer the proposed estimator over other estimators under consideration. Further it is certainly better than the estimator proposed by Singh and Horn (2000).

References

[1] 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**, 1247-1264.
[2] Bahl, S and Tuteja, R.K. (1991) : Ratio and product - type exponential estimator, Information and Optimization Sciences, **XII**, 159-163.
[3] Giancarlo Diana and Pier Francesco Perri (2010): Improved Estimators of the population mean for missing data.
[4] Hansen, M.H and Hurwitz, W.N.(1946).: The Problem of non response in sample surveys. J. Amer. Statist.Assoc., **41**, 517 – 529.
[5] Heitjan, D.F and Basu, S.(1996): Distinguishing Missing at Random and Missing Completely at Random. Amer. Statist. **50**, 207-213.
[6] Kadilar, C and Cingi, H. (2008).: Estimators for the population mean in the case of missing data .Commun. Statist. Theor. Meth. **37**, 2226-2236.
[7] Mukhopadhyaya, P. (2000): Theory and methods of survey sampling. Prentice Hall of India Pvt. Ltd., New Delhi.
[8] Singh, S and Horn, S.(2000). Compromised imputation in survey sampling .
[9] Singh, S and Deo, B.(2003): Imputation by power transformation.Statist.papers.(555-579).
[10] Rubin, D. B. (1976a): Inference and missing data. Biometrica, **63**, 581-593.
[11] Rao, J.N.K and Sitter .R .R (1995): Variance estimation under two phase sampling with application to imputation for missing data. Biometrica **82**, 453 – 460.
[12] Shukla, D and Thakur, N.S (2008). Estimation of mean with imputation of missing data using factor type estimators. Statistics in Transition, **9**, 1, 33-48.
[13] Shukla, D., Thakur, N.S., Pathak Sharad 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**. 397-414.
[14] Singh, S.: Optimal method of imputation in survey sampling. Applied Mathematical Sciences, **3**, 2009, 1727 – 1737.