Journal of Statistical Theory and Practice (2019) 13:19
https://doi.org/10.1007/s42519-018-0016-5

ORIGINAL ARTICLE

# Estimation of Population Mean Using Imputation Methods for Missing Data Under Two-Phase Sampling Design

**G. N. Singh[1] · S. Suman[1]**

Published online: 5 November 2018


**Abstract**
This manuscript emphasizes the estimation procedure of population mean in two-phase sampling when non-response occurs during survey in both phases of sample data. To cope with the problem of missing data, some new imputation methods have been suggested for estimating the population mean which utilize the information on two auxiliary variables. The properties of the resultant estimators are studied which are followed by empirical and simulation studies accomplished on real as well as on artificial data sets which justify the suggested imputation methods. Results are significantly analyzed, and appropriate suggestions are made to the survey practitioners.

**Keywords** Non-response · Auxiliary variable · Imputation · Bias · Mean square error · Sampling design

**Mathematics Subject Classification** 62D05

## 1 Introduction

Missing data are the most frequent occurring feature in sample surveys, and recognizing its stochastic nature is of utmost importance in order to use appropriate methodology for handling the data sets. Failure in recognition of its nature may distort the inferences about population characteristics/parameters; therefore, the assiduous attempt is needed for handling of the data sets with missing values. A fundamental query appears in this regard that what assumptions to be considered while justifying the ignorability of the complete mechanism. Rubin [1] discussed

✉ S. Suman
surbhi.iitism@yahoo.com

G. N. Singh
gnsingh_ism@yahoo.com

1 Department of Applied Mathematics, Indian Institute of Technology (Indian School of Mines), Dhanbad 826004, India

this fundamental query for missing data by establishing ignorability conditions under the classical and Bayesian approach for statistical inference. Further, [2, 3] subsequently generalized the [1] model to include other forms of incompleteness. Initially, [1] addressed three key concepts related to missing pattern of the data sets: missing at random (MAR), observed at random (OAR) and parameter distribution (PD). He mentioned "The data are MAR if the probability of the observed missingness pattern, given the observed and unobserved data, does not depend on the values of the unobserved data. 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 observed data." Later, the combination of MAR and OAR is called missing completely at random (MCAR). Heitain and Basu [4] have differentiated MAR and MCAR mechanism with series of examples. Based on these works, the pattern of the missing mechanism of data sets is recognized and inference related to population parameter is made under some strategies according to their obtained pattern. These methods are termed as "imputation methods." Imputation is the procedure of replacing missing data with fabricated values. Abundant of works have been carried out based on imputation methods, such as [5–17].

The information related to the auxiliary variable may be used either at the planning stage or at design stage or survey stage or at estimation stage to get the improved precision of the estimates. When the information on auxiliary variable correlated with study variable is readily available, ratio, regression and their transformed and improved methods have been widely used to obtain efficient estimates, anticipating the information on the population mean of the auxiliary variable. In spite of that, the knowledge of the population mean of the auxiliary variable is not always available. In such circumstances, two-phase sampling or double scheme is a widely used sampling scheme to obtain the reliable estimates of unknown population mean of auxiliary variable in survey studies. The presence of missing data during survey sampling under two-phase sampling design enforces the researchers to implement the imputation methods for obtaining trustworthy conclusion regarding population parameters. Several researchers like [18–21] and others have suggested some imputation methods for compensating existence of the missing data with the assumption that the complete response may not be available on the study variable as well as on the auxiliary variable in second-phase sample. It is worth to be mentioned that very limited attention has been paid to deal with the situations, when the complete response is not available in the first-phase sample as well.

Following the aforementioned arguments and motivated with the work of [9], authors have proposed some effective imputation methods under missing completely at random (MCAR) response mechanism, which result in the point estimators of the population mean of study variable in two-phase sampling design. The properties of the proposed estimators have been discussed. Empirical and simulation studies are accomplished to authenticate the propositions of the suggested imputation methods and resultant estimators. Suitable recommendations have been made to the survey practitioners for real-life applications.

## 2 Sampling Design and Notations

Let $P = (P_1, P_2 \ldots P_N)$ be a finite population of size $N$ indexed by triplet characters $(y, x, z)$. It is assumed that $y$ is the study variable and ($x$ and $z$) are the (first and second) auxiliary variables, respectively, such that $y$ is positively correlated with $x$ and $z$, while in comparison with $x$, it is remotely correlated with $z$. When the population mean $\bar{X}$ of the first auxiliary variable is not known but information on the second auxiliary variable $z$ is available for all the units of the population, the following two-phase sampling scheme has been designed for making inference about the population parameters.

Let $s'$ be the first-phase sample of size $n'$ drawn using simple random sampling without replacement (SRSWOR) scheme from the population and surveyed for the auxiliary variable $x$ to estimate its population mean $\bar{X}$. The second-phase sample of size $n < n'$ is drawn to measure the study characteristic $y$ under the following design:

*Design I* The second-phase sample $s$ is drawn from the first-phase sample $s'$

*Design II* The second-phase sample $s$ is independently drawn from the entire population.

We have assumed that non-response occurs in the first- and second-phase samples where $r'$ and $r$ are the number of responding units in the first- and second-phase samples of sizes $n'$ and $n$, respectively. The corresponding sets of responding units are denoted by ($R_1$ and $R_2$) and the sets of non-responding units by ($R_1^c$ and $R_2^c$), respectively. We have also assumed that sample units in the second-phase sample $s$ have been drawn from the responding set $R_1$.

## 3 Proposed Methods of Imputation and Subsequent Estimators

In this section, using the compromised method of imputation in the first-phase sample, we have proposed some new compromised imputation methods under MCAR response mechanism in the second-phase sample for missing data on the study variable $y$. The proposed imputation methods and resultant estimators are given below:

### 3.1 Imputation for Missing Data in the First-Phase Sample

To compensate the missing values on auxiliary variable $x$ in the first-phase sample, we considered the ratio method of imputation; hence, after imputation, the sample data in $x$ take the following form:

$$x_{.i} = \begin{cases} \dfrac{\alpha n' x_i}{r'} + (1-\alpha)\hat{b}' z_i & \text{if } i \in R_1 \\ (1-\alpha)\hat{b}' z_i & \text{if } i \in R_1^c \end{cases} \tag{1}$$

where $\hat{b}' = \dfrac{\sum_{i=1}^{r'} x_i}{\sum_{i=1}^{r'} z_i}$ and $\alpha$ is an unknown constant. Under the imputation method described in Eq. (1), the point estimator of the population mean $\bar{X}$ in the first-phase sample is derived as

$$\bar{x}' = \frac{1}{n'}\left\{\sum_{i \in R_1} x_{.i} + \sum_{i \in R_1^c} x_{.i}\right\}$$

which produces the point estimator of the population mean $\bar{X}$ in the first-phase sample as

$$\bar{x}' = \alpha\bar{x}_{r'} + (1-\alpha)\bar{x}_{r'}\frac{\bar{z}_{n'}}{\bar{z}_{r'}} \tag{2}$$

where $\bar{x}_{r'} = \dfrac{\sum_{i \in R_1} x_i}{r'}$, $\bar{z}_{r'} = \dfrac{\sum_{i \in R_1} z_i}{r'}$ and $\bar{z}_{n'} = \dfrac{\sum_{i=1}^{n'} z_i}{n'}$.

### 3.2 Imputation for Missing Data in the Second-Phase Sample

To derive the reliable substitutes for missing values in the second-phase sample, we suggest two new compromised imputation methods which are presented below:

*First Imputation Method* Under this method of imputation, sample data take the following forms

$$y_{.i} = \begin{cases} \dfrac{\alpha_1 n y_i c}{r} + (1-\alpha_1)\hat{b}z_i c & \text{if } i \in R_2 \\ (1-\alpha_1)c\hat{b}z_i & \text{if } i \in R_2^c \end{cases} \tag{3}$$

where $c = \dfrac{1}{\bar{x}_n}\alpha\bar{x}_{r'} + (1-\alpha)\bar{x}_{r'}\dfrac{\bar{z}_{n'}}{\bar{z}_{r'}}$, $\hat{b} = \dfrac{\sum_{i=1}^{r} y_i}{\sum_{i=1}^{r} z_i}$ and $\alpha_1$ is suitably chosen constant such that the mean square error of resultant estimator is minimum.

Under the imputation method described in Eq. (3), the point estimator of the population mean $\bar{Y}$ takes the following form

$$\zeta_1 = \frac{\left\{\alpha_1\bar{y}_r + (1-\alpha_1)\bar{y}_r\frac{\bar{z}_n}{\bar{z}_r}\right\}}{\bar{x}_n}\left\{\alpha\bar{x}_{r'} + (1-\alpha)\bar{x}_{r'}\frac{\bar{z}_{n'}}{\bar{z}_{r'}}\right\}. \tag{4}$$

*Second Imputation Method* Under this method of imputation, sample data take the following forms

$$y_{.i} = \begin{cases} \frac{\alpha_2 n y_i}{r} + (1-\alpha_2)\hat{b}z_i & \text{if } i \in R_2 \\ (1-\alpha_2)\hat{b}z_i + \frac{1}{n-r}\hat{b}_{yx}(r)\left\{ \alpha\bar{x}_{r'} + (1-\alpha)\bar{x}_{r'}\frac{\bar{z}_{n'}}{\bar{z}_{r'}} - \bar{x}_n \right\} & \text{if } i \in R_2^c \end{cases} \tag{5}$$

where $\hat{b}_{yx}(r) = \frac{s_{yx}}{s_x^2}$ and $\alpha_2$ is suitably chosen constant such that the mean square error of resultant estimator is minimum.

Under the imputation method described in Eq. (5), the point estimator of the population mean $\bar{Y}$ takes the following form

$$\zeta_2 = \left\{ \alpha_2\bar{y}_r + (1-\alpha_2)\bar{y}_r\frac{\bar{z}_n}{\bar{z}_r} \right\} + \hat{b}_{yx}(r)\left\{ \alpha\bar{x}_{r'} + (1-\alpha)\bar{x}_{r'}\frac{\bar{z}_{n'}}{\bar{z}_{r'}} - \bar{x}_n \right\}. \tag{6}$$

## 4 Properties of Estimators $\zeta_1$ and $\zeta_2$

The properties of the proposed estimators $\zeta_1$ and $\zeta_2$ have been explored under two different types of two-phase sampling design opted for MCAR response mechanism. Large sample approximations have been used in order to obtain the expressions of biases and mean square errors of the proposed estimators using the following transformations:

$$\begin{aligned} &\bar{y}_r = \bar{Y}(1+e_0), \bar{x}_r = \bar{X}(1+e_1), \bar{x}_{r'} = \bar{X}(1+e_1'), \bar{x}_n = \bar{X}(1+e_2), \bar{x}_{n'} = \bar{X}(1+e_2'), \\ &\bar{z}_{r'} = \bar{Z}(1+e_3'), \bar{z}_n = \bar{Z}(1+e_4), \bar{z}_{n'} = \bar{Z}(1+e_4'), s_{yx}(r) = S_{YX}(1+e_5), s_x^2(r) = S_X^2(1+e_6), \\ &\quad \text{such that } E(e_i') = E(e_i) = 0, |e_i'| \le 1 \text{ and } |e_i| \le 1 \forall i, i' = 0,1,2,3,4,5,6. \end{aligned}$$

Under the above transformations, the estimators $\zeta_1$ and $\zeta_2$ take the following forms:

$$\zeta_1 = \bar{Y}\left\{ \alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3} \right\}\frac{(1+e_1')}{(1+e_2)}\frac{(1+e_4')}{(1+e_3')} \tag{7}$$

and

$$\begin{aligned} \zeta_2 = \bar{Y}(1+e_0)&\left\{ \alpha_2 + (1-\alpha_2)\frac{(1+e_4)}{(1+e_3)} \right\} \\ &+ \beta_{YX}\frac{(1+e_5)}{(1+e_6)}\bar{X}\left[ (1+e_1')\left\{ \alpha + (1-\alpha)\frac{(1+e_4')}{(1+e_3')} \right\} - (1+e_2) \right] \end{aligned} \tag{8}$$

where $\beta_{YX} = \frac{S_{YX}}{S_X^2}$.

### 4.1 Biases and Mean Square Errors of Estimators $\zeta_1$ and $\zeta_2$

Let $B(.)_d$ and $\mathrm{MSE}(.)_d$ be the bias and mean square error, respectively, of an estimator under a given two-phase sampling design $d(= I, II)$.

**Theorem 4.1** *The biases of the estimators* $\zeta_1$ *and* $\zeta_2$ *are given by*

$$B(\zeta_1)_I = \bar{Y}\left[\delta_2(C_X^2 - \rho_{YX}C_YC_X) + \{\delta_3(1-\alpha_1) + \delta_4(1-\alpha)\}(C_Z^2 - \rho_{YZ}C_YC_Z)\right] \tag{9}$$

$$\begin{aligned} B(\zeta_1)_{II} = \bar{Y}\Big[&f_1(C_X^2 - \rho_{YX}C_YC_X) + \delta_3(1-\alpha_1)(C_Z^2 - \rho_{YZ}C_YC_Z) \\ &+ \delta_4(1-\alpha)(C_Z^2 - \rho_{XZ}C_XC_Z)\Big] \end{aligned} \tag{10}$$

$$\begin{aligned} B(\zeta_2)_I = {}& \bar{Y}\delta_3(1-\alpha_2)(C_Z^2 - \rho_{YZ}C_YC_Z) + \beta_{YX}\bar{X}\left[\delta_4(1-\alpha)(C_Z^2 - \rho_{XZ}C_XC_Z)\right] \\ &+ \beta_{YX}\bar{X}\left[\frac{\delta_2}{\bar{X}}\left(\frac{\mu_{030}}{\mu_{020}} - \frac{\mu_{120}}{\mu_{110}}\right) + \frac{(1-\alpha)(\delta_4)}{\bar{Z}}\left(\frac{\mu_{021}}{\mu_{020}} - \frac{\mu_{111}}{\mu_{110}}\right)\right] \end{aligned} \tag{11}$$

$$\begin{aligned} B(\zeta_2)_{II} = {}& \bar{Y}\delta_3(1-\alpha_2)(C_Z^2 - \rho_{YZ}C_YC_Z) + \beta_{YX}\bar{X}\left[\delta_4(1-\alpha)(C_Z^2 - \rho_{XZ}C_XC_Z)\right] \\ &+ \beta_{YX}\bar{X}\frac{f_1}{\bar{X}}\left(\frac{\mu_{120}}{\mu_{110}} - \frac{\mu_{030}}{\mu_{020}}\right) \end{aligned} \tag{12}$$

where

$$\delta_1 = \left(\frac{1}{r} - \frac{1}{N}\right), \delta_2 = \left(\frac{1}{n} - \frac{1}{r'}\right), \delta_3 = \left(\frac{1}{r} - \frac{1}{n}\right), \delta_4 = \left(\frac{1}{r'} - \frac{1}{n'}\right), \delta_5 = \left(\frac{1}{n'} - \frac{1}{N}\right),$$

$\delta_6 = \left(\frac{1}{r'} - \frac{1}{N}\right)$ and $f_1 = \left(\frac{1}{n} - \frac{1}{N}\right)$.

***Proof*** The bias of the estimator $\zeta_1$ is derived as

$$\begin{aligned} B(\zeta_1)_d &= E(\zeta_1 - \bar{Y}) \\ &= E\left[\bar{Y}\left\{\alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{(1+e_1')}{(1+e_2)}\frac{(1+e_4')}{(1+e_3')} - \bar{Y}\right] \end{aligned} \tag{13}$$

Now, expanding the right-hand sides of Eq. (13) binomially, taking expectation under the sampling designs I and II, respectively, and retaining the terms up to the first order of approximations, we get the expression of the bias of the proposed estimator $\zeta_1$ under sampling designs I and II as obtained in Eqs. (9)–(10).

In similar fashion, we derive the expression of bias of the proposed estimator $\zeta_2$ under sampling designs I and II as obtained in Eq. (11)–(12). □

**Theorem 4.2** *The mean square errors of the estimators $\zeta_1$ and $\zeta_2$ are given by*

$$\begin{aligned}\mathrm{MSE}(\zeta_1)_I = \bar{Y}^2\Big[&\delta_1 C_Y^2 + \delta_2(C_X^2 - 2\rho_{YX}C_YC_X) + \delta_3\{(1-\alpha_1)^2C_Z^2 \\ &- 2(1-\alpha_1)\rho_{YZ}C_YC_Z\}\} + \delta_4\{(1-\alpha)^2C_Z^2 - 2(1-\alpha)\rho_{YZ}C_YC_Z\}\Big]\end{aligned} \tag{14}$$

$$\begin{aligned}\mathrm{MSE}(\zeta_1)_{II} = \bar{Y}^2\Big[&\delta_1 C_Y^2 + \delta_6 C_X^2 + f_1(C_X^2 - 2\rho_{YX}C_YC_X) + \delta_3\{(1-\alpha_1)^2C_Z^2 \\ &- 2(1-\alpha_1)\rho_{YZ}C_YC_Z\} + \delta_4\{(1-\alpha)^2C_Z^2 - 2(1-\alpha)\rho_{XZ}C_XC_Z\}\Big]\end{aligned} \tag{15}$$

$$\begin{aligned}\mathrm{MSE}(\zeta_2)_I = \bar{Y}^2\Big[&(\delta_1 - \delta_2\rho_{YX}^2)C_Y^2 + \delta_3\{(1-\alpha_2)^2C_Z^2 - 2(1-\alpha_2)\rho_{YZ}C_YC_Z\}\Big] \\ &+ \delta_4\Big[(1-\alpha)^2\beta_{YX}^2\bar{X}^2C_Z^2 - 2(1-\alpha)\bar{Y}\bar{X}\beta_{YX}\rho_{YZ}C_YC_Z)\Big]\end{aligned} \tag{16}$$

*and*

$$\begin{aligned}\mathrm{MSE}(\zeta_2)_{II} = \bar{Y}^2\big[&(\delta_1 - f_1\rho_{YX}^2)C_Y^2 + \delta_3\{(1-\alpha_2)^2C_Z^2 - 2(1-\alpha_2)\rho_{YZ}C_YC_Z\}\big] \\ &+ \beta_{YX}^2\bar{X}^2\big[\delta_4\{(1-\alpha)^2C_Z^2 - 2(1-\alpha)\rho_{XZ}C_XC_Z)\} + \delta_6 C_X^2\big].\end{aligned} \tag{17}$$

***Proof*** The mean square error of the estimator $\zeta_1$ is derived as

$$\begin{aligned}\mathrm{MSE}(\zeta_1)_d &= E(\zeta_1 - \bar{Y})^2 \\ &= E\left[\bar{Y}\left\{\alpha_1(1+e_0) + (1+\alpha_1)(1+e_0)\frac{1+e_4}{1+e_3}\right\}\frac{(1+e_1')}{(1+e_2)}\frac{(1+e_4')}{(1+e_3')} - \bar{Y}\right]^2\end{aligned} \tag{18}$$

Now, expanding the right-hand sides of Eq. (18) binomially, taking expectation under the sampling designs I and II, respectively, and retaining the terms up to the first order of approximations, we get the expressions of the mean square error of the proposed estimator $\zeta_1$ under sampling designs I and II as obtained in Eqs. (14)–(15).

In similar fashion, we derive the expression of mean square error of the proposed estimator $\zeta_2$ under sampling designs I and II as obtained in Eqs. (16)–(17). □

### 4.2 Minimum Biases and Mean Square Errors of the Estimators $\zeta_1$ and $\zeta_2$

Since the mean square errors of estimators $\zeta_1$ and $\zeta_2$ under two types of sampling designs mentioned in Eqs. (14)–(17) are the functions of unknown scalars $\alpha, \alpha_1$ and $\alpha_2$, the optimum choices of $\alpha, \alpha_1$ and $\alpha_2$ are obtained by minimizing the mean square errors given in Eqs. (14)–(17) with respect to $\alpha, \alpha_1$ and $\alpha_2$ as

$$\alpha_{1(\text{opt})_I} = \alpha_{1(\text{opt})_{II}} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \tag{19}$$

$$\alpha_{2(\text{opt})_I} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \quad \text{and} \quad \alpha_{2(\text{opt})_{II}} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \tag{20}$$

For estimator $\zeta_1$, we have

$$\alpha_{(\text{opt})_I} = 1 - \rho_{YZ}\frac{C_Y}{C_Z} \quad \text{and} \quad \alpha_{(\text{opt})_{II}} = 1 - \rho_{XZ}\frac{C_X}{C_Z} \tag{21}$$

For estimator $\zeta_2$, we have

$$\alpha_{(\text{opt})_I} = 1 - \frac{\rho_{YZ}C_X}{\rho_{YX}C_Z} \quad \text{and} \quad \alpha_{(\text{opt})_{II}} = 1 - \rho_{XZ}\frac{C_X}{C_Z} \tag{22}$$

The optimum biases of the proposed estimators $\zeta_1$ and $\zeta_2$ have been obtained by putting the optimum choices of $\alpha, \alpha_1$ and $\alpha_2$ from Eqs. (19)–(22) in Eqs. (9)–(12). The optimum biases of the proposed estimators $\zeta_1$ and $\zeta_2$ under two types of two-phase sampling designs are given as

$$B^*(\zeta_1)_I = \bar{Y}\left[\delta_2(C_X^2 - \rho_{YX}C_YC_X) + (\delta_3 + \delta_4)(\rho_{YZ}C_YC_Z - \rho_{YZ}^2C_Y^2)\right] \tag{23}$$

$$\begin{aligned} B^*(\zeta_1)_{II} = \bar{Y}\Big[&f_1(C_X^2 - \rho_{YX}C_YC_X) + \delta_3(\rho_{YZ}C_YC_Z - \rho_{YZ}^2C_Y^2) \\ &+ \delta_4(\rho_{XZ}C_XC_Z - \rho_{XZ}^2C_X^2)\Big] \end{aligned} \tag{24}$$

$$\begin{aligned} B^*(\zeta_2)_I = {}& \bar{Y}\delta_3(\rho_{YZ}C_YC_Z - \rho_{YZ}^2C_Y^2) + \beta_{YX}\bar{X}\left[\delta_4\frac{\rho_{YZ}}{\rho_{YX}}(C_XC_Z - \rho_{XZ}C_X^2)\right] + \beta_{YX}\bar{X} \\ & \left[\frac{\delta_2}{\bar{X}}\left(\frac{\mu_{030}}{\mu_{020}} - \frac{\mu_{120}}{\mu_{110}}\right) + \frac{(1-\alpha)(\delta_4)}{\bar{Z}}\left(\frac{\mu_{021}}{\mu_{020}} - \frac{\mu_{111}}{\mu_{110}}\right)\right] \end{aligned} \tag{25}$$

$$B^{*}(\zeta_2)_{II} = \bar{Y}\delta_3(\rho_{YZ}C_YC_Z - \rho_{YZ}^2C_Y^2) + \beta_{YX}\bar{X}\left[\delta_4(\rho_{XZ}C_XC_Z - \rho_{XZ}^2C_Z^2)\right]$$
$$+ \beta_{YX}f_1\left(\frac{\mu_{120}}{\mu_{110}} - \frac{\mu_{030}}{\mu_{020}}\right). \tag{26}$$

The minimum mean square errors of the proposed estimators $\zeta_1$ and $\zeta_2$ have been obtained by putting the optimum choices of $\alpha, \alpha_1 and \alpha_2$ from Eqs. (19)–(22) in Eqs. (14)–(17). The optimum mean square errors of the proposed estimators $\zeta_1$ and $\zeta_2$ under two types of two-phase sampling designs are denoted by $M(\zeta_1)_d$ and $M(\zeta_1)_d$, respectively, and given as

$$M(\zeta_1)_I = \bar{Y}^2\left[\{\delta_1 - (\delta_3 + \delta_4)\rho_{YZ}^2\}C_Y^2 + \delta_2(C_X^2 - 2\rho_{YX}C_YC_X)\right] \tag{27}$$

$$M(\zeta_1)_{II} = \bar{Y}^2\left[(\delta_1 - \delta_3\rho_{YZ}^2)C_Y^2 + (\delta_6 - \delta_4\rho_{XZ}^2)C_X^2 + f_1(C_X^2 - 2\rho_{YX}C_YC_X)\right] \tag{28}$$

$$M(\zeta_2)_I = \bar{Y}^2\left[(\delta_1 - \delta_2\rho_{YX}^2)C_Y^2 - (\delta_3 + \delta_4)\rho_{YZ}^2\right] \tag{29}$$

and

$$M(\zeta_2)_{II} = \bar{Y}^2\left[(\delta_1 - f_1\rho_{YX}^2) - \delta_3\rho_{yz}^2\right]C_Y^2 + \beta_{yx}^2\bar{X}^2C_X^2\left(\delta_6 - \delta_4\rho_{xz}^2\right). \tag{30}$$

## 5 Some Well-Known Methods of Imputation

In the single-phase sampling design when the sample of size $n$ is selected from the population under SRSWOR scheme and the non-response occurs in the sample data, some classical methods of imputation are presented in this section under the assumption that information on the auxiliary variable $x$ is available for each and every units of the population.

### 5.1 Mean Method of Imputation

The mean method of imputation gives the data as:

$$y_{.i} = \begin{cases} y_i & \text{if } i \in R \\ \bar{y}_r & \text{if } i \in R^c \end{cases} \tag{31}$$

Under the imputation method discussed in Eq. (31), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_m = \frac{1}{r}\sum_{i=1}^{r} y_{.i} = \bar{y}_r. \tag{32}$$

The variance of the estimator $\bar{y}_m$ is obtained as

$$v(\bar{y}_m) = \delta_1 \bar{Y}^2 C_Y^2. \tag{33}$$

### 5.2 Ratio Method of Imputation

The ratio method of imputation gives the data as:

$$y_{.i} = \begin{cases} y_i & \text{if } i \in R \\ \hat{b} x_i & \text{if } i \in R^c \end{cases} \tag{34}$$

where $\hat{b} = \dfrac{\sum_{i\in R}^{\cdot} y_i}{\sum_{i\in R}^{\cdot} x_i}$.

Under the imputation method discussed in Eq. (34), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_{\text{rat}} = \frac{1}{n}\sum_{i=1}^{n} y_{.i} = \bar{y}_r \frac{\bar{x}_n}{\bar{x}_r}. \tag{35}$$

The mean square error of the estimator $\bar{y}_{\text{rat}}$ up to the first order of approximations is obtained as

$$M(\bar{y}_{\text{rat}}) = \bar{Y}^2\left[\delta_1 C_Y^2 + \delta_3 (C_X^2 - 2\rho_{YX} C_Y C_X)\right]. \tag{36}$$

### 5.3 Regression Method of Imputation

The regression method of imputation gives the data as

$$y_{.i} = \begin{cases} y_i & \text{if } i \in R \\ \hat{a} + \hat{b}_{yx} x_i & \text{if } i \in R^c \end{cases} \tag{37}$$

where $\hat{b}_{yx} = \dfrac{s_{yx}(r)}{s_x^2(r)}$ and $\hat{a} = \left(\bar{y}_r - \hat{b}_{yx}\bar{x}_r\right)$. Under the imputation method discussed in Eq. (37), the corresponding point estimator of the population mean $\bar{Y}$ is derived as

$$\bar{y}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n} y_{.i} = \bar{y}_r + \hat{b}_{yx}\left(\bar{x}_n - \bar{x}_r\right). \tag{38}$$

The mean square of the estimator $\bar{y}_{\text{reg}}$ up to the first order of approximations is obtained as

$$M(\bar{y}_{\text{reg}}) = \bar{Y}^2 C_Y^2\left[\delta_1 - \delta_3 \rho_{yx}^2\right]. \tag{39}$$

## 6 Analytical Comparison

In this section, we compare the suggested estimators with existing classical estimators $\bar{y}_m$ , $\bar{y}_{\text{rat}}$ and $\bar{y}_{\text{reg}}$.

**Lemma 6.1**

(i) *The proposed estimator* $\zeta_1$ *under first-phase design is more efficient than* $\bar{y}_m$ *if*

$$M(\zeta_1)_I - v(\bar{y}_m) < 0 \Rightarrow \frac{1-2\rho_{YX}}{\rho^2_{YZ}} < \frac{\delta_3+\delta_4}{\delta_2}.$$

(ii) *The proposed estimator* $\zeta_1$ *under second--phase design is more efficient than* $\bar{y}_m$ *if*

$$M(\zeta_1)_{II} - v(\bar{y}_m) < 0 \Rightarrow 1-2\rho_{YX} < \frac{\delta_3\rho^2_{YZ}+\delta_4\rho^2_{XZ}-\delta_6}{f_1}.$$

(iii) *The proposed estimator* $\zeta_2$ *under first-phase design is more efficient than* $\bar{y}_m$ *if*

$$M(\zeta_2)_I - v(\bar{y}_m) < 0 \Rightarrow \delta_2\rho^2_{YZ} + (\delta_3+\delta_4)\rho^2_{YZ} > 0$$

*which is always true.*

(iv) *The proposed estimator* $\zeta_2$ *under second-phase design is more efficient than* $\bar{y}_m$ *if*

$$M(\zeta_2)_{II} - v(\bar{y}_m) < 0 \Rightarrow \bar{Y}^2(f_1\rho^2_{YX}+\delta_3\rho^2_{YZ}) > \bar{X}^2\beta^2_{YX}(\delta_6-\delta_4\rho^2_{XZ})$$

**Lemma 6.2**

(i) *The proposed estimator* $\zeta_1$ *under first-phase design is more efficient than* $\bar{y}_{\text{rat}}$ *if*

$$M(\zeta_1)_I - M(\bar{y}_{\text{rat}}) < 0 \Rightarrow \frac{1-2\rho_{YX}}{\rho^2_{YZ}} < \frac{\delta_3+\delta_4}{\delta_2-\delta_3}.$$

(ii) *The proposed estimator* $\zeta_1$ *under second-phase design is more efficient than* $\bar{y}_{\text{rat}}$ *if*

$$M(\zeta_1)_{II} - M(\bar{y}_{\text{rat}}) < 0 \Rightarrow 1-2\rho_{YX} < \frac{\delta_3\rho^2_{YZ}+\delta_4\rho^2_{XZ}-\delta_6}{f_1-f_3}.$$

(iii) *The proposed estimator $\zeta_2$ under first-phase design is more efficient than $\bar{y}_{\text{rat}}$ if*

$$M(\zeta_2)_I - M(\bar{y}_{\text{rat}}) < 0 \Rightarrow \delta_2\rho_{YX}^2 + (\delta_3 + \delta_4)\rho_{YZ}^2 + \delta_3(1 - 2\rho_{YX}) > 0$$

*which is always true if $\rho_{YX} > \frac{1}{2}$.*

(iv) *The proposed estimator $\zeta_2$ under second-phase design is more efficient than $\bar{y}_{\text{rat}}$ if*

$$M(\zeta_2)_{II} - M(\bar{y}_{\text{rat}}) < 0 \Rightarrow 1 - 2\rho_{YX} > \frac{\beta_{YX}^2\bar{X}^2(\delta_6 - \delta_4\rho_{XZ}^2) - (\delta_3\rho_{YZ}^2 + f_1\rho_{YX}^2)\bar{Y}^2}{\delta_3\overline{Y}^2}.$$

**Lemma 6.3**

(i) *The proposed estimator $\zeta_1$ under first-phase design is more efficient than $\bar{y}_{\text{reg}}$ if*

$$M(\zeta_1)_I - M(\bar{y}_{\text{reg}}) < 0 \Rightarrow \delta_3\rho_{YX}^2 + \delta_2(1 - 2\rho_{YX}) < (\delta_3 + \delta_4)\rho_{YZ}^2.$$

(ii) *The proposed estimator $\zeta_1$ under second-phase design is more efficient than $\bar{y}_{\text{reg}}$ if*

$$M(\zeta_1)_{II} - M(\bar{y}_{\text{reg}}) < 0 \Rightarrow \delta_3\rho_{YX}^2 + f_1(1 - 2\rho_{YX}) < (\delta_4\rho_{XZ}^2 + \delta_3\rho_{YZ}^2) - \delta_6.$$

(iii) *The proposed estimator $\zeta_2$ under first-phase design is more efficient than $\bar{y}_{\text{reg}}$ if*

$$M(\zeta_2)_I - M(\bar{y}_{\text{reg}}) < 0 \Rightarrow (\delta_3 - \delta_2)\rho_{YX}^2 < (\delta_3 + \delta_4)\rho_{YZ}^2$$

(iv) *The proposed estimator $\zeta_2$ under second-phase design is more efficient than $\bar{y}_{\text{reg}}$ if*

$$M(\zeta_2)_{II} - M(\bar{y}_{\text{reg}}) < 0 \Rightarrow \bar{Y}^2\left\{(\delta_3 - f_1)\rho_{YX}^2 - \delta_3\rho_{YZ}^2\right\} + \bar{X}^2\beta_{YX}^2(\delta_6 - \delta_4\rho_{XZ}^2) < 0.$$

***Remark 6.1*** It may be assumed that $C_Y \approx C_X \approx C_Z$ in the population.

## 7 Efficiency Comparison

In this section, empirical and simulation studies have been carried out to demonstrate the accomplishment of the proposed methods of imputation and resultant estimators over mean, ratio and regression methods of imputation.

### 7.1 Empirical Study

To show the practicability of the proposed methods of imputation in the real-life scenario, four natural populations from various survey studies have been chosen for empirical study. The optimum mean square errors of proposed estimators are taken under consideration in empirical study. The percent relative efficiencies of the proposed methods with respect to the classical methods of imputations (mean, ratio and regression) are obtained as

$$E_{11} = \frac{v(\bar{y}_m)}{M(\zeta_1)} \times 100, \quad E_{12} = \frac{M(\bar{y}_{\text{rat}})}{M(\zeta_1)} \times 100, \quad E_{13} = \frac{M(\bar{y}_{\text{reg}})}{M(\zeta_1)} \times 100;$$
$$E_{21} = \frac{v(\bar{y_m})}{M(\zeta_2)} \times 100, \quad E_{22} = \frac{M(\bar{y}_{\text{rat}})}{M(\zeta_2)} \times 100 \quad \text{and} \quad E_{23} = \frac{M(\bar{y}_{\text{reg}})}{M(\zeta_2)} \times 100.$$

The detailed information of populations is given below:

*Population I [Source* [22]] *(Page No. 58)*

$Y$: Head length of second son
$X$: Head length of first son
$Z$: Head breadth of first son
$N = 25, n' = 18, r' = 11, n = 9, r = 7.$

*Population II [Source:* [23] ] *(Page No. 399)*

$Y$: Area under wheat in 1964
$X$: Area under wheat in 1963
$Z$: : Cultivated area in 1961
$N = 34, n' = 22, r' = 14, n = 11, r = 8.$

*Population III [Source:* [24]] *(Page No. 182)*

$Y$: Number of ‘placebo’ children
$X$: Number of paralytic polio cases in the placebo group
$Z$: Number of paralytic polio cases in the ‘not inoculated’ group
$N = 33, n' = 22, r' = 18, n = 12, r = 8.$

*Population IV [Source:* [25] *(Page No. 349)*

$Y$: Volume
$X$: Diameter
$Z$: Height
$N = 31, n' = 22, r' = 16, n = 10, r = 7.$

**Table 1** Percent relative efficiencies of the proposed methods of imputation with respect to mean method of imputation

| Population | Design I | | Design II | |
|---|---|---|---|---|
| | $E_{11}$ | $E_{21}$ | $E_{11}$ | $E_{21}$ |
| I | 166.7263 | 120.0599 | 170.2578 | 151.791 |
| II | 344.1896 | 347.9319 | 344.3223 | 336.9121 |
| III | 153.1210 | 197.6898 | 204.4901 | 197.8982 |
| IV | 170.9074 | 103.5229 | 166.3121 | 159.7961 |

**Table 2** Percent relative efficiencies of the proposed methods of imputation with respect to ratio method of imputation

| Population | Design I | | Design II | |
|---|---|---|---|---|
| | $E_{12}$ | $E_{22}$ | $E_{12}$ | $E_{22}$ |
| I | 143.9904 | 103.6878 | 147.0403 | 131.0918 |
| II | 226.3443 | 228.8053 | 226.4315 | 221.5584 |
| III | 123.7454 | 95.73043 | 133.0007 | 130.4266 |
| IV | 127.7331 | 147.7498 | 152.8322 | 147.9056 |

**Table 3** Percent relative efficiencies of the proposed methods of imputation with respect to regression method of imputation

| Population | Design I | | Design II | |
|---|---|---|---|---|
| | $E_{13}$ | $E_{23}$ | $E_{13}$ | $E_{23}$ |
| I | 140.7310 | 101.3406 | 143.7118 | 128.1244 |
| II | 226.2768 | 228.7371 | 226.3640 | 221.4924 |
| III | 102.8689 | 79.58021 | 110.5628 | 108.4229 |
| IV | 108.9643 | 126.0398 | 130.3755 | 126.1727 |

The percent relative efficiencies are computed for the above-mentioned populations under both sampling designs I and II and shown in Tables 1, 2 and 3.

## 7.2 Simulation Study

A computer simulation is an endeavor to model a real-life or hypothetical scenarios on a computer so that it may be studied to see how the proposed system, strategies or methods works. The inference may be made about the behavior of the proposed system, strategies or methods by changing parameters in the simulation study. It is a tool to virtually investigate the behavior of the method or system under study. Inspired by this argument, we have run simulation study to investigate the behavior of the proposed imputation methods with respect to classical methods of imputation. The simulation studies have been performed on three artificial computer generated data sets to know the percent relative efficiencies and losses of proposed estimators due to the presence of non-response in the population. The description of artificial data sets is given as:

*Population V Source: [Artificially Generated Data Set]*

A population of size $N = 2000$ are generated from the multivariate normal distribution in R software. The study variable $y$ is positively correlated with auxiliary variables with fixed correlations $\rho_{YX} = 0.7$, $\rho_{YZ} = 0.6$ and $\rho_{XZ} = 0.5$. The parameters used for this population are $n' = 800, r' = 640, n = 256, r = 204$.

*Population VI Source: [Artificially Generated Data Set]*

The triplet $(y, x, z)$ is generated of size $N = 200$ . The study variable $y$ is highly correlated with auxiliary variables with fixed correlations $\rho_{YX} = 0.93$, $\rho_{YZ} = 0.87$ and $\rho_{XZ} = 0.95$. We have taken $n' = 80, r' = 64, n = 50, r = 40$.

*Population VII Source: [Artificially Generated Data Set]*

The triplet $(y, x, z)$ is generated of size $N = 1000$ such that $x \sim gamma(4, 2.5), e \sim N(0, 1), z = 1.5x^{0.5} + e, y = 8x + 7z + e$ where $\rho_{YX} > \rho_{Yz}$. We have taken $n' = 400, r' = 320, n = 128, r = 102$.

In this simulation studies, the following steps have been followed:

*Step I* Draw a random sample $s'$ of size $n'$ from population size $N$.
*Step II* Take out $(n' - r')$ sample units randomly from the first-phase sample each time. Impute dropped units using imputation method contemplated for the first-phase sample.
*Step III* Draw a random subsample of size $n$ from $s'$ for design I and independent random sample $n$ from $N$ for design II.
*Step IV* Take out $(n - r)$ sample units randomly from the second-phase sample each time. Impute dropped units using proposed method of imputation contemplated for the second-phase sample.
*Step V* Compute relevant statistics.
*Step VI* Repeat the above steps $\binom{N}{n=M}$ (say) times .

The simulated variance and mean square errors of the existing and proposed estimators are obtained as:

$$\mathrm{var}^*(\bar{Y}_M) = \frac{1}{M}\sum_{j=1}^{M}((\bar{y}_m)_j - \bar{Y})^2, M^*(\bar{y}_{\mathrm{rat}}) = \frac{1}{M}\sum_{j=1}^{M}(\bar{y}_{\mathrm{rat}})_j - \bar{Y})^2, M^*(\bar{y}_{\mathrm{reg}})$$
$$= \frac{1}{M}\sum_{j=1}^{M}((\bar{y}_{\mathrm{reg}})_j - \bar{Y})^2,$$

$$M^*(\zeta_1)_d = \frac{1}{M}\sum_{j=1}^{M}((\zeta_1)_{dj} - \bar{Y})^2 \quad \text{and} \quad M^*(\zeta_2)_d = \frac{1}{M}\sum_{j=1}^{M}((\zeta_1)_{dj} - \bar{Y})^2$$

The simulated percent-related efficiencies are given as

$$E'_{11} = \frac{\mathrm{var}^*(\bar{y_m})}{M^*(\zeta_1)_d} \times 100, \quad E'_{12} = \frac{M^*(\bar{y}_{\mathrm{rat}})}{M^*(\zeta_1)_d} \times 100, \quad E'_{13} = \frac{M^*(\bar{y}_{\mathrm{reg}})}{M^*(\zeta_1)_d} \times 100;$$
$$E'_{21} = \frac{\mathrm{var}(\bar{y_m})}{M^*(\zeta_2)_d} \times 100, \quad E'_{22} = \frac{M^*(\bar{y}_{\mathrm{rat}})}{M^*(\zeta_2)_d} \times 100 \quad \text{and} \quad E'_{23} = \frac{M^*(\bar{y}_{\mathrm{reg}})}{M^*(\zeta_2)_d} \times 100.$$

**Table 4** Bias of proposed, mean, ratio and regression estimators under imputation method

| Population | $B(\xi_1)_I$ | $B(\xi_2)_I$ | $B(\xi_1)_{II}$ | $B(\xi_2)_{II}$ | $B(\bar{y}_m)$ | $B(\bar{y}_{\text{rat}})$ | $B(\bar{y}_{\text{reg}})$ |
|---|---|---|---|---|---|---|---|
| I | − 0.0035 | − 0.0185 | 0.0224 | 0.0170 | − 0.0071 | − 0.0158 | − 0.0069 |
| II | − 0.7398 | − 0.3272 | − 0.8430 | − 0.5335 | 0.0920 | − 0.0525 | 0.0915 |
| III | 0.2768 | 0.2418 | 0.2661 | 0.2195 | 0.0665 | 0.2234 | − 0.0359 |
| IV | − 0.2358 | − 0.0651 | − 0.1998 | − 0.0664 | − 0.0034 | − 0.0913 | − 0.0613 |

**Table 5** Percent relative efficiencies of proposed method with respect to mean, ratio and regression methods of imputation under design I

| Population | $E'_{11}$ | $E'_{12}$ | $E'_{13}$ | $E'_{21}$ | $E'_{22}$ | $E'_{23}$ |
|---|---|---|---|---|---|---|
| *V* | 94.79711 | 111.2117 | 172.9588 | 242.4857 | 284.4733 | 442.419 |
| *VI* | 202.5033 | 167.5122 | 204.283 | 131.2021 | 108.5314 | 132.3552 |
| *VII* | 198.655 | 151.9309 | 202.3274 | 124.9503 | 95.56174 | 127.2602 |

**Table 6** Percent relative efficiencies of proposed method with respect to mean, ratio and regression methods of imputation under design II

| Population | $E'_{11}$ | $E'_{12}$ | $E'_{13}$ | $E'_{21}$ | $E'_{22}$ | $E'_{23}$ |
|---|---|---|---|---|---|---|
| *V* | – | – | – | 166.4769 | 295.3405 | 226.3282 |
| *VI* | 110.6395 | 91.0507 | 111.6266 | 132.9174 | 109.3844 | 134.1034 |
| *VII* | 165.2897 | 127.6943 | 167.7603 | 135.1092 | 104.3784 | 137.1287 |

The percent relative losses in efficiencies due to non-response of the estimators $\zeta_1$ and $\zeta_2$ are obtained with respect to the similar estimators when non-response has not observed in any phase. The estimators $T_1$ and $T_2$ are defined under the similar circumstances as the estimators $\zeta_1$ and $\zeta_2$, respectively, but under complete response. The simulated percent relative losses in efficiencies of the proposed estimators $\zeta_1$ and $\zeta_2$ with respect to $T_1$ and $T_2$, respectively, under their respective design are given as

$$l_1 = \frac{M'(\zeta_1)_d - \text{MSE}(T_1)_d}{M'(\zeta_1)_d} \times 100 \quad \text{and} \quad l_2 = \frac{M'(\zeta_2)_d - \text{MSE}(T_2)_d}{M'(\zeta_1)_d} \times 100$$

where

$$\text{MSE}(T_1)_d = \frac{1}{M} \sum_{j=1}^{M} ((T_1)_{dj} - \bar{Y})^2 \quad \text{and} \quad \text{MSE}(T_2)_d = \frac{1}{M} \sum_{j=1}^{M} ((T_1)_{dj} - \bar{Y})^2.$$

In this study, $M = 50{,}000$ has been taken for convenience in calculation. The values of $E'_{ij}(i = 1, 2, ), (j = 1, 2, 3)$ and $l_k(k = 1, 2)$ are calculated based on the above procedures and presented in Tables 5, 6, 7, 8, 9 and 10.

Following the above-mentioned simulation study, we have also calculated the biases of the resultant estimators $\zeta_1$, $\zeta_2$ and existing estimators $\bar{y}_m$, $\bar{y}_{\text{rat}}$ and $\bar{y}_{\text{reg}}$ for populations I-IV and shown in Table 4.

## 8 Interpretations of Empirical and Simulation Results

The following interpretation may be read out form Tables 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10:

(i) From Tables 1, 2 and 3, it is seen that the percent relative efficiencies of proposed estimators $\zeta_1$ and $\zeta_2$ with respect to the estimators $\bar{y}_m$, $\bar{y}_{\text{rat}}$ and $\bar{y}_{\text{reg}}$ are more than 100 in almost cases when percent relative efficiencies have been obtained using the large sample approximations. This reflects the dominance nature of the proposed method of imputations and resultant estimators over the classical method of imputations.

(ii) From Tables 5 and 6, it is observed that simulated percent relative efficiencies of proposed estimators $\zeta_1$ and $\zeta_2$ with respect to the estimators $\bar{y}_m$, $\bar{y}_{\text{rat}}$ and $\bar{y}_{\text{reg}}$ are more than 100 in most of the cases when simulation studies are performed on artificial data sets.

(iii) From Tables 7, 8, 9 and 10, it is indicated that the percent relative losses in efficiencies $l_1$ and $l_2$ of the estimators $\zeta_1$ and $\zeta_2$ under two types of two-phase sampling designs are not more than 30% for both artificial and real populations.

(iv) From Tables 7 and 8, the negative percent relative losses in efficiencies are observed for some cases under two-phase sample design I which indicates the gain in the precision of estimate.

(v) From Tables 8, 9 and 10, it is also seen that the percent relative losses in efficiencies $l_1$ and $l_2$ are decreasing as the values of $r$ increase for fixed values of $N, n', r'$ and $n$ under both types of two-phase sampling designs. This shows that the percent relative losses in efficiencies are decreasing as percentage of non-response in the second-phase sample decreases.

In Tables 7 and 8, the impact of percent relative losses in efficiencies of the proposed estimators is observed very closely taking into consideration of minor change in percentage of non-response in the second-phase sample and results are shown graphically in Figs. 1, 2, 3, 4, 5 and 6 to get more visible pattern under sampling designs I and II separately.

**Table 7** Percent relative loss in efficiencies of $T_1$ and $T_2$ for population V

| Non-response in % | $r$ | Design I | | Design II | |
|---|---|---|---|---|---|
| | | $l_1$ | $l_2$ | $l_1$ | $l_2$ |
| 20.3 | 204 | − 5.78817 | 18.37585 | 9.91844 | 22.05380 |
| 19.9 | 205 | − 6.31469 | 18.92934 | 9.78423 | 22.37468 |
| 19.5 | 206 | − 5.99639 | 17.22236 | 10.37980 | 21.97567 |
| 19.1 | 207 | − 6.44455 | 17.85349 | 9.99347 | 21.88525 |
| 18.8 | 208 | − 6.25436 | 16.28825 | 9.78489 | 21.35387 |
| 18.4 | 209 | − 6.44913 | 16.77161 | 9.70870 | 21.05562 |
| 18.0 | 210 | − 6.86914 | 16.30936 | 10.00350 | 20.77964 |
| 17.6 | 211 | − 7.01849 | 15.29330 | 9.81893 | 20.39861 |
| 17.2 | 212 | − 7.70971 | 15.23885 | 9.63560 | 19.20669 |
| 16.8 | 213 | − 6.02526 | 15.57119 | 9.45744 | 19.39193 |
| 16.4 | 214 | − 7.16053 | 14.64099 | 9.57376 | 18.67739 |
| 16.0 | 215 | − 7.52787 | 13.78702 | 9.46119 | 19.30768 |
| 15.6 | 216 | − 7.04825 | 14.22896 | 9.38657 | 18.36987 |
| 15.2 | 217 | − 7.09088 | 13.32298 | 9.12777 | 17.94182 |
| 14.8 | 218 | − 7.64929 | 13.68114 | 9.22407 | 17.84795 |
| 14.5 | 219 | − 7.83837 | 13.46266 | 9.19650 | 17.82550 |
| 14.1 | 220 | − 7.45660 | 12.21393 | 8.91465 | 16.92621 |
| 13.7 | 221 | − 7.39892 | 11.83445 | 9.06312 | 16.69474 |
| 13.3 | 222 | − 8.33907 | 11.31900 | 9.54730 | 16.56260 |
| 12.9 | 223 | − 8.10236 | 11.96917 | 9.03036 | 15.71908 |
| 12.5 | 224 | − 8.25863 | 11.31934 | 9.14293 | 16.08700 |
| 12.1 | 225 | − 7.81661 | 11.24739 | 9.04696 | 15.42324 |
| 11.7 | 226 | − 8.40448 | 10.23890 | 8.74597 | 14.73185 |
| 11.3 | 227 | − 7.98930 | 10.33325 | 8.79993 | 15.01819 |
| 10.9 | 228 | − 8.76567 | 8.98472 | 9.08217 | 14.12681 |
| 10.5 | 229 | − 8.45887 | 9.25380 | 8.48976 | 13.52911 |
| 10.2 | 230 | − 8.50701 | 9.03267 | 8.75179 | 13.04728 |
| 9.8 | 231 | − 9.09961 | 8.03963 | 9.19510 | 13.53535 |
| 9.4 | 232 | − 9.10836 | 7.93695 | 8.73893 | 13.19960 |
| 9.0 | 233 | − 8.44382 | 7.93043 | 8.75542 | 12.76529 |
| 8.6 | 234 | − 9.17175 | 7.76826 | 8.41679 | 12.15552 |
| 8.2 | 235 | − 9.96900 | 7.02693 | 8.81511 | 11.55354 |
| 7.8 | 236 | − 8.82106 | 7.28959 | 8.72726 | 11.69547 |
| 7.4 | 237 | − 8.60567 | 7.07465 | 8.59408 | 11.36768 |
| 7.0 | 238 | − 9.49687 | 5.97648 | 8.64117 | 11.28274 |
| 6.6 | 239 | − 9.05269 | 5.69001 | 8.10258 | 10.63448 |
| 6.3 | 240 | − 9.99699 | 4.80093 | 8.57878 | 10.74080 |
| 5.9 | 241 | − 9.53311 | 5.08852 | 8.59809 | 10.46995 |
| 5.5 | 242 | − 9.63400 | 4.76279 | 8.51161 | 9.71565 |
| 5.1 | 243 | − 9.42295 | 4.44962 | 8.53528 | 9.47632 |
| 4.7 | 244 | − 10.00734 | 3.71636 | 8.34055 | 9.10798 |
| 4.3 | 245 | − 10.48103 | 3.39492 | 8.15473 | 8.24409 |

**Table 7** (continued)

| Non-response in % | $r$ | Design I | | Design II | |
|---|---|---|---|---|---|
| | | $l_1$ | $l_2$ | $l_1$ | $l_2$ |
| 3.9 | 246 | – 9.88941 | 3.38665 | 8.16975 | 8.46270 |
| 3.5 | 247 | – 10.06214 | 3.29086 | 8.04340 | 7.55059 |
| 3.1 | 248 | – 10.14132 | 2.57034 | 8.20028 | 7.52925 |
| 2.7 | 249 | – 10.46531 | 2.45727 | 8.01292 | 7.55005 |
| 2.3 | 250 | – 10.15615 | 1.52099 | 7.69954 | 7.04355 |
| 2.0 | 251 | – 10.54069 | 1.41321 | 8.54029 | 7.05718 |
| 1.6 | 252 | – 11.12250 | 1.11616 | 8.16445 | 6.66917 |
| 1.2 | 253 | – 10.53713 | 0.20925 | 7.92645 | 5.82650 |
| 0.8 | 254 | – 10.27741 | 0.14196 | 7.95137 | 5.64613 |
| 0.4 | 255 | – 10.39877 | – 0.24243 | 7.94458 | 5.42982 |
| 0.0 | 256 | – 10.94064 | – 0.27996 | 7.60032 | 4.94298 |

**Table 8** Percent relative loss in efficiencies of $T_1$ and $T_2$ for population VI

| Non-response in % | $r$ | Design I | | Design II | |
|---|---|---|---|---|---|
| | | $l_1$ | $l_2$ | $l_1$ | $l_2$ |
| 20 | 40 | 0.38499 | 3.22463 | 20.39023 | 4.10889 |
| 18 | 41 | – 0.50571 | 2.65795 | 19.58656 | 3.69353 |
| 16 | 42 | – 0.98140 | 2.63887 | 19.85537 | 3.05953 |
| 14 | 43 | – 1.62884 | 1.97126 | 19.52627 | 2.68474 |
| 12 | 44 | – 2.40481 | 1.53769 | 19.07874 | 2.46456 |
| 10 | 45 | – 2.94636 | 1.15623 | 18.81600 | 1.88539 |
| 8 | 46 | – 3.93481 | 0.73010 | 18.34428 | 1.44364 |
| 6 | 47 | – 4.59481 | 0.31150 | 18.46878 | 1.18639 |
| 4 | 48 | – 4.98863 | – 0.07682 | 17.29139 | 0.69541 |
| 2 | 49 | – 5.69632 | – 0.49788 | 17.29067 | 0.33158 |
| 0 | 50 | – 6.27116 | – 0.81286 | 16.90846 | 0.00644 |

**Table 9** Percent relative loss in efficiencies of $T_1$ and $T_2$ for population II

| Non-response in % | $r$ | Design I | | Design II | |
|---|---|---|---|---|---|
| | | $l_1$ | $l_2$ | $l_1$ | $l_2$ |
| 33.3 | 8 | 36.95582 | 14.83099 | 41.03305 | 12.51196 |
| 25.0 | 9 | 30.64520 | 8.46692 | 34.14653 | 8.73106 |
| 16.7 | 10 | 23.86985 | 7.58180 | 27.67686 | 5.44361 |
| 08.3 | 11 | 17.51420 | 3.97377 | 21.32365 | 2.58095 |
| 00.0 | 12 | 15.93921 | 0.00102 | 15.31889 | 0.00006 |

**Table 10** Percent relative loss in efficiencies of $T_1$ and $T_2$ for population VII

| Non-response in % | $r$ | Design I | | Design II | |
|---|---|---|---|---|---|
| | | $l_1$ | $l_2$ | $l_1$ | $l_2$ |
| 0.195 | 103 | 26.81336 | 7.73479 | 29.08972 | 8.33188 |
| 0.188 | 104 | 26.24445 | 7.42685 | 28.39203 | 7.63929 |
| 0.180 | 105 | 26.23111 | 7.57448 | 28.12601 | 7.05548 |
| 0.172 | 106 | 25.03845 | 6.62386 | 26.57549 | 7.06049 |
| 0.164 | 107 | 24.07782 | 6.23184 | 26.92988 | 6.30257 |
| 0.156 | 108 | 22.02978 | 6.14109 | 25.23823 | 6.22763 |
| 0.148 | 109 | 22.86167 | 5.88638 | 25.19880 | 5.47770 |
| 0.141 | 110 | 21.85859 | 5.56049 | 24.44542 | 5.28713 |
| 0.133 | 111 | 20.30682 | 5.20738 | 24.28342 | 5.20218 |
| 0.125 | 112 | 20.31858 | 4.73168 | 23.29412 | 4.91795 |
| 0.117 | 113 | 19.25704 | 4.37953 | 22.84402 | 4.29005 |
| 0.109 | 114 | 18.74893 | 4.29919 | 22.22405 | 4.03677 |
| 0.102 | 115 | 17.83962 | 4.10569 | 21.72554 | 3.62004 |
| 0.094 | 116 | 17.08092 | 3.48148 | 21.07267 | 3.56363 |
| 0.086 | 117 | 16.17240 | 3.34654 | 20.96048 | 3.25696 |
| 0.078 | 118 | 15.21242 | 2.96510 | 20.23740 | 2.76381 |
| 0.070 | 119 | 14.06151 | 2.70694 | 19.59230 | 2.46887 |
| 0.063 | 120 | 13.79710 | 2.45448 | 19.47380 | 2.18053 |
| 0.055 | 121 | 13.94238 | 2.43216 | 18.36667 | 2.11684 |
| 0.047 | 122 | 11.94824 | 1.80634 | 17.72348 | 1.67985 |
| 0.039 | 123 | 10.74092 | 1.53694 | 17.81411 | 1.41050 |
| 0.031 | 124 | 10.49605 | 1.29132 | 17.09411 | 1.06919 |
| 0.023 | 125 | 10.02008 | 1.20352 | 16.58742 | 0.83095 |
| 0.016 | 126 | 9.02589 | 0.89909 | 15.76101 | 0.56016 |
| 0.008 | 127 | 7.42204 | 0.63633 | 15.51021 | 0.20878 |
| 0.000 | 128 | 7.52170 | 0.36594 | 14.21055 | 0.02435 |

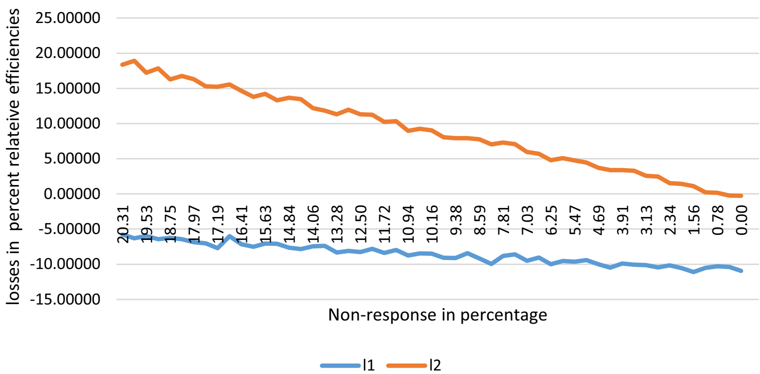


**Fig. 1** Losses in percent relative efficiencies under design I for population V

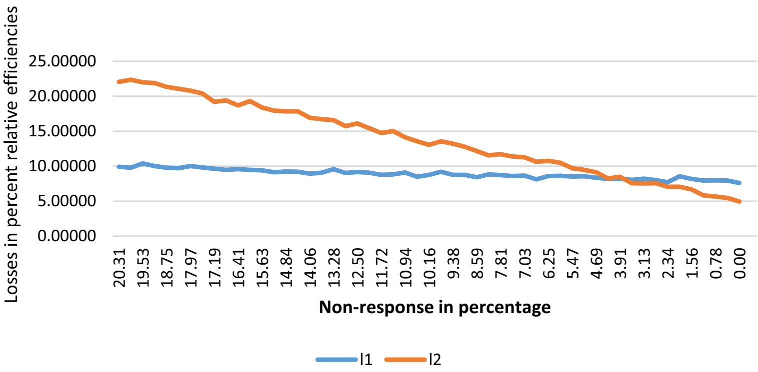


**Fig. 2** Losses in percent relative efficiencies under design II

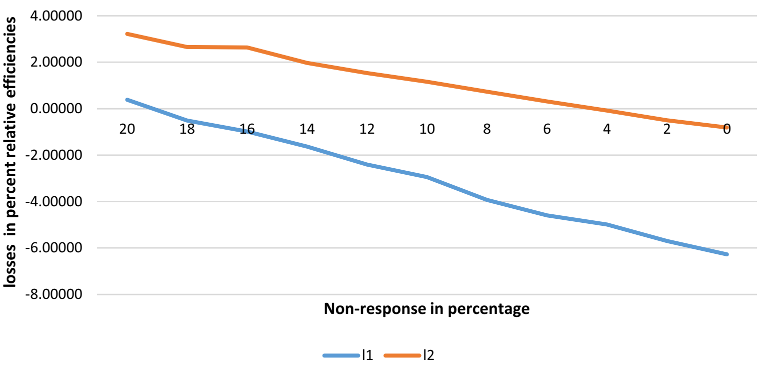


**Fig. 3** Losses in percent relative efficiencies under design I for population VI

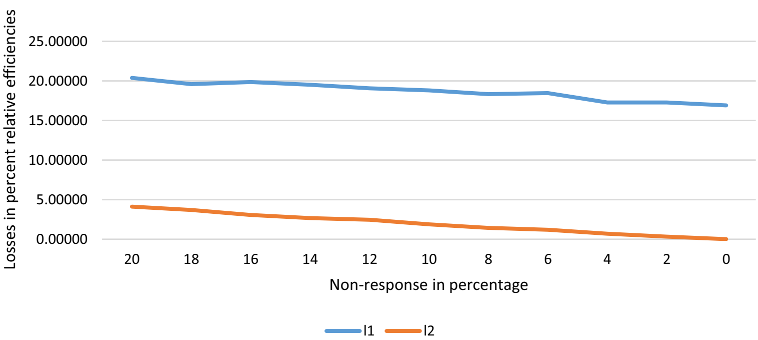


**Fig. 4** Losses in percent relative efficiencies under design II for population VI

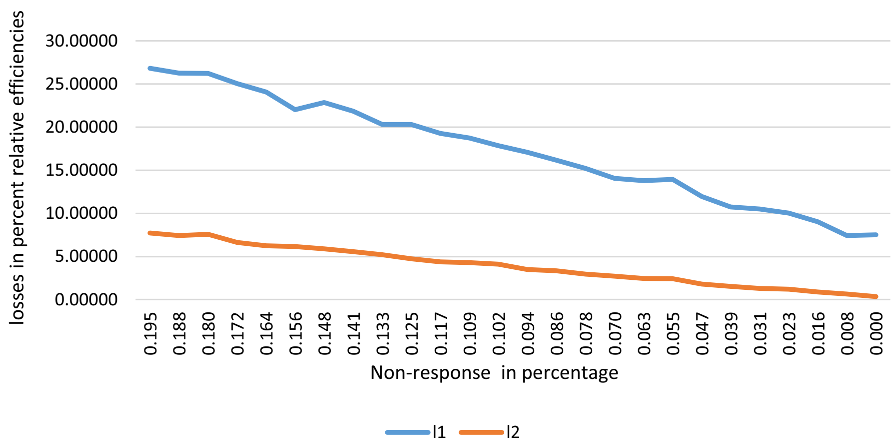


**Fig. 5** Losses in percent relative efficiencies under design I for Population VII

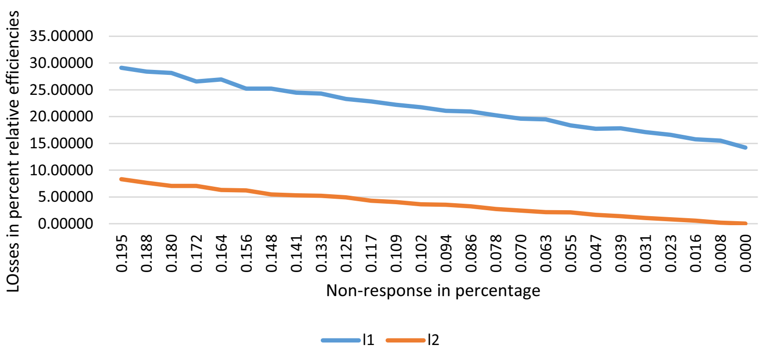


**Fig. 6** Losses in percent relative efficiencies under design II for Population VII

From Figs. 1, 2, 3, 4, 5 and 6, it is easily seen that the percent relative losses in efficiencies of proposed estimators are decreasing as the percentage of non-response decreases under both types of sampling designs.

## 9 Conclusions and Recommendations

When the proposed methods of imputation under study have implemented in real-life scenario, proposed methods are remunerating in terms of percent relative efficiencies. These strategies are also showing their superiority in terms of percent relative efficiencies over classical imputation methods namely mean, ratio and regression methods of imputation when simulation studies have been performed over artificial data sets. The percent relative losses in efficiency of proposed estimators are less than 30% whenever non-response occurs 20% or less of sample size. These results support that the proposed methods of imputations described in this study are appreciatively favorable in diminishing the pessimistic effect of non-response on inference to a greater extend as compared to the classical methods of imputation. Hence, looking on the persuaded behavior of the suggested imputation methods, survey practitioner may be encouraged for their practical applications, whenever non-response is inescapable in the survey data.

**Acknowledgements** Authors are thankful to the Indian Institute of Technology (Indian School of Mines), Dhanbad, for providing necessary support to carry out the present research work. Authors are also thankful to the reviewers for their valuable suggestions which improved the quality of the paper.

## References

1. Rubin DB (1976) Inference and missing data. Biometrica 63:581–593
2. Heitain FD, Rubin BD (1991) Ignorablity and coarse data random. Annu Stat 50(3):207–213
3. Heitain FD (1994) Ignoriablity in general complete-data models. Biometrika 81:701–708
4. Heitain FD, Basu S (1996) Distinguishing “missing at random” and “missing completely at random”. Am Stat 50(3):207–213
5. Sande IG (1979) A personal view of hot deck approach to automatic edit and imputation. J Imput Proced Surv Methodol 5:238–246
6. Kalton G, Kasprzyk D, Santos R (1981) Issues of non-response and imputation in the survey of income and program participation. In: Krewski D, Platek R, Rao JNK (eds) Current topics in survey sampling. Academic Press, New York, pp 455–480
7. Lee H, Rancourt E, Sarndal CE (1994) Experiments with variance estimation from survey data with imputed values. J Off Stat 10(3):231–243
8. Lee H, Rancourt E, Sarndal CE (1995) Variance estimation in the presence of imputed data for the generalized estimation system. In: Proceeding of the American Statistical Association (Survey Research Methods Section of the American Statistical Association (ASA)). pp 384–389
9. Singh S, Horn S (2000) Compromised imputation in survey sampling. Metrika 51:266–276
10. Singh S, Deo B (2003) Imputation by power transformation. Stat Pap 44:555–579
11. Ahmed MS, Al-Titi O, Al-Rawi Z, Abu-Dayyeh W (2006) Estimation of population mean using different imputation methods. Stat Transit 7(6):1247–1264
12. Kadilar C, Cingi H (2008) Estimators for the population mean in the case of missing data. Commun Stat Theory Methods 37:2226–2236
13. Singh S (2009) A new method of imputation in survey sampling. Statistics 43(5):499–511
14. Diana G, Perri PF (2010) Improved estimators of the population mean for missing data. Commun Stat Theory Methods 39:3245–3251
15. Singh GN, Karna JP (2010) Some imputation methods to minimize the effect of non response in two-occasion rotation patterns. Commun Stat Theory Methods 39(18):3264–3281
16. Gira Abdeltawab A (2015) Estimation of population mean with a new imputation methods. Appl Math Sci 9(34):1663–1672
17. Bhushan S, Pandey PP (2016) Optimality of ratio type estimation methods for population mean in presence of missing data. Commun Stat Theory Methods. https://doi.org/10.1080/03610926.2016.1167906
18. Thakur NS, Yadav K, Pathak S (2011) Estimation of mean in presence of missing data under two-phase sampling scheme. J Reliab Stat Stud 4(2):93–104
19. Thakur NS, Yand Pathak S (2012) Some imputation methods in double sampling scheme for estimation of population mean. Int J Mod Eng Res 2(1):200–207
20. Thakur NS, Yadav K, Pathak S (2013) On mean estimation with imputation in two-phase sampling. Res J Math Stat Sci 1(13):1–9
21. Pandey R, Yadav K (2016) Mean estimation under imputation based on two-phase sampling design using an auxiliary variable. Pak J Stat Oper Res XII(4):639–658

22. Anderson TW (1958) An introduction to multivariate statistical analysis. Wiley, New York
23. Murthy MN (1967) Sampling theory and methods. Statistical Publishing Society, Calcutta
24. Cochran WG (1977) Sampling techniques. Wiley, New-York
25. Wang SG, Chow SC (1994) Advanced linear models: theory and applications. Marcel Dekker, Inc., New York