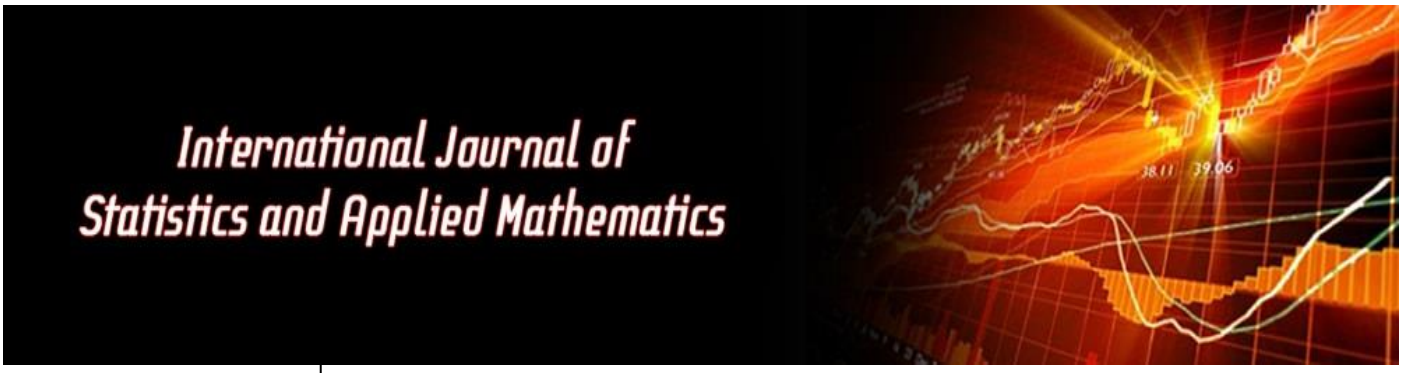
International Journal of Statistics and Applied Mathematics 2025; 10(6): 134-144



ISSN: 2456-1452
Maths 2025; 10(6): 134-144

www.mathsjournal.com
Received: 03-05-2025
Accepted: 04-06-2025

Bhargob Gohain
North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India

Deepjan Gohain
North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India

BK Singh
North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India

Corresponding Author:
Bhargob Gohain
North Eastern Regional Institute of Science and Technology, Nirjuli, Arunachal Pradesh, India

Exponential logarithmic approach for estimation of population parameter in two-phase sampling

Bhargob Gohain, Deepjan Gohain and BK Singh

DOI: https://www.doi.org/10.22271/maths.2025.v10.i6b.2067

Abstract
Imputation is a commonly used approach to substitute missing data in sample surveys to address the issue of lack of response. The objective of the current study is to develop various imputation techniques for mean estimation of study variable in situations where the population mean of the auxiliary variable is unknown. Three imputation strategies for the exponential-logarithmic type estimators along with their respective point estimators have been suggested. Also, some expressions for the bias and the mean square error of the proposed estimators have been derived. Furthermore, some numerical data sets have been used to test the percentage relative efficiency of the proposed estimators. A simulation study has also been done for the efficiency of the estimators. Higher percentage relative efficiency indicates better performance. The proposed exponential-logarithmic imputation techniques provide more efficient and accurate ways to handle non-response in surveys compared to traditional imputation methods.

Keywords: Mean Squared Error (MSE), Missing data, bias, two-phase sampling, ratio, product, large sample approximation, simple random sampling

1. Introduction
Survey sampling in statistical terminology describes the process of choosing a sample from an interest group that is the target population. A brief interview or discussion with people may be referred to as a survey. Depending on the desired approach and the goals to be met, there are numerous ways in which a survey can be conducted. In survey sampling, missing data or missing values occur when a chosen respondent refuses to participate in a survey for various reasons. Various techniques are available to handle the problems of missing data. One of the best techniques to deal with missing data is imputation. It is a familiar and special method of handling non-response in which various options are used to impute missing values for one or more study variables. The approach of the current study is to identify certain classes of estimators that are more effective than other pertinent estimators and are typically in practice in two-phase sampling.

An imputation technique that solves missing item values in an efficient manner, disallowing incompleteness and making its analysis simple. Recently, many authors have given notable attention to the study of imputation techniques. In case of observing missing values of the variable under study and its estimation there upon using two auxiliary variables in two phase sampling in a predictive approach with unknown population mean of the first auxiliary variable has been studied [1]. Using an auxiliary variable, an effective class of estimators for a finite population in stratified random sampling has been suggested [2]. A study has been attempted modestly to offer some effective imputation techniques of the difference and ratio types when there are missing values under rank set sampling [3]. A factor-type exponential ratio estimator as a technique for estimating population mean has been addressed, where the estimators make use of additional data for an auxiliary variable, including kurtosis and coefficient of variation [4]. To estimate the population mean of the study variable in stratified two-phase sampling while taking measurement error and non-response into account concurrently, three classes of more efficient and generalized combined regression-cum-ratio estimators have been introduced [5].

Using imputation of missing data under two-phase sampling, an efficient exponential chain ratio estimator for estimating the finite population mean of the study variable has been proposed [6].
To handle the missing data, a few new classes of exponential-logarithmic imputation approaches have been given [7]. A family of logarithmic cum exponential estimators using two-phase sampling for population mean with risk analysis has been suggested [8]. In two phase sampling, an efficient class of estimators for finite population mean estimation has been addressed when both measurement errors and non-response occur at the same time [9]. Efficient estimation combining exponential and ln functions under two-phase sampling has been suggested [10]. During survey sampling, there is a chance that the population mean of the auxiliary variable will be unknown to the practitioners. The solution to this problem of an unknown value is two-phase sampling, often known as double sampling.
This research aims to explore the properties of the Das and Singh (2023) estimators and propose double sampling variants of them [7]. The simple random sampling without replacement (SRSWOR) scheme has been taken into consideration throughout the work. To show how well the recommended estimators work in comparison to alternative estimators, an empirical study is conducted.

Let in a finite population $\Omega = \{\Omega_1, \Omega_2, \Omega_3, \ldots \Omega_N\}$ of size N, the value of the variables on the i^{th} unit $\Omega_i, \forall\, i = 1, 2, 3 \ldots N$, be $(y_i; x_i)$. Let $\bar{Y} = \sum_{i=1}^{N} \frac{y_i}{N}$ and $\bar{X} = \sum_{i=1}^{N} \frac{x_i}{N}$ be the population means of the study variable Y and the auxiliary variable X respectively. A simple random sample of size n is taken from the population Ω to estimate the population mean $\bar{Y}$. Let $\bar{y}_n = \sum_{i=1}^{n} \frac{y_i}{n}$ and $\bar{x}_n = \sum_{i=1}^{n} \frac{x_i}{n}$ be the sample means of study variable Y and the auxiliary variable X.

The two-phase sampling in which unknown population mean $\bar{X}$ is estimated by selecting a preliminary large sample S' of size n' from the population. The sample mean $\bar{x}' = \sum_{i \in s'} \frac{x_i}{n'}$ is considered as estimate of $\bar{X}$. The second sample S of size n is selected either from S' or Ω.

Case 1: As a sub sample from sample S'.

Case 2: Independent to sample S'.

Let sample S contains two distinct subsets of total of n units containing r responding units with sub-set R and non-responding $(n - r)$ units $(r < n)$ forming a sub-set R^c with respect to study variable only in $S = R \cup R^c$ for every $y_i \in R^c$, the i^{th} value x_i of the auxiliary variable is used as source of imputation. The sample mean of the variable X based on the sample S' of size n' i.e.

$\bar{x}' = \sum_{i=1}^{n'} \frac{x_i}{n'}$, where $n' \in S'$

$\bar{y}_r, \bar{x}_r$ are the sample mean of the variable Y and X based on the subset R (responding units) of S. $\bar{y}_n, \bar{x}_n$ are the sample mean of the study variable Y and auxiliary variable X based on $R \cup R^c$. The following symbols are used as a notation respectively.

ρ_{YX} : The coefficient of correlation between Y and X.

C_Y, C_X : The population coefficient of the variation of Y and X.

$$\zeta_1 = \left(\frac{1}{r} - \frac{1}{n'}\right);\ \zeta_2 = \left(\frac{1}{n} - \frac{1}{n'}\right);\ \zeta_3 = \left(\frac{1}{n'} - \frac{1}{N}\right);\ \zeta_4 = \left(\frac{1}{r} - \frac{1}{N - n'}\right);\ \zeta_5 = \left(\frac{1}{n} - \frac{1}{N - n'}\right)$$

2. Methodology

Several widely available and conventional techniques for imputation include the following classical methods:

2.1 Mean Methods

In this method each non-responding data is replaced with the mean of responded data, Here

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

The unit estimator of the population mean under this imputation method, $\bar{Y}$ is given by

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

Bias of $\bar{y}_r$ is $B(y_r) = 0$ and MSE $M(\bar{y}_r) = \bar{Y}^2\left(\frac{1}{r} - \frac{1}{N}\right)C_Y^2$.

2.2 Ratio Method

Lee and Sardal proposed method

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

Where $\hat{l} = \dfrac{\sum_{i \in R} y_i}{\sum_{i \in R} x_i}$

This imputation method has the following unit estimator

$$\bar{y}_{RAT} = \bar{y}_r\left(\frac{\bar{x}_n}{\bar{x}_r}\right)$$

Where $\bar{x}_n = n^{-1}\sum_{i \in R} x_i$, $\bar{x}_r = r^{-1}\sum_{i \in R} x_i$ and $\bar{y}_r = r^{-1}\sum_{i \in R} y_i$

Bias and the MSE of this unit estimator $\bar{y}_{RAT}$,

$$B(\bar{y}_{RAT}) = \bar{Y}\left(\frac{1}{r} - \frac{1}{n}\right)\left(C_X^2 - \rho_{YX}C_Y C_X\right)$$

$$M(\bar{y}_{RAT}) = \bar{Y}^2\left\{\left(\frac{1}{r} - \frac{1}{N}\right)C_Y^2 + \left(\frac{1}{r} - \frac{1}{n}\right)\left(1 - 2\rho_{YX}\frac{C_Y}{C_X}\right)C_X^2\right\}$$

2.3 Compromised Method

The compromised method of imputation is suggested by Singh and Horn

$$y_{.i} = \begin{cases} \alpha\dfrac{n}{r}y_i + (1-\alpha)\hat{l}x_i & if\ i \in R \\ (1-\alpha)\hat{l}x_i & if\ i \in R^c \end{cases}$$

Where α is a chosen constant.

The unit estimator of this imputation method is $\bar{y}_{COMP} = \alpha\bar{y}_r + (1-\alpha)\bar{y}_r\dfrac{\bar{x}_n}{\bar{x}_r}$

Bias and MSE of the unit estimator $\bar{y}_{COMP}$

$$B(\bar{y}_{COMP}) = \bar{Y}\left(\frac{1}{r} - \frac{1}{n}\right)(1-\alpha)(1-\phi_{YX})C_X^2$$

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

This estimator is minimum at $\alpha_{opt} = 1 - \rho_{YX}\dfrac{C_Y}{C_X}$

The optimum mean square error (MSE) of $\bar{y}_{COMP}$ is

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

2.4 Proposed Imputation Techniques and Corresponding unit Estimators

Three imputation techniques with their corresponding unit estimators are suggested as follows

(1)

$$y_{1.i} = \begin{cases} y_i & if\ i \in R \\ \dfrac{1}{n-r}\bar{y}_r\left[n\exp\left\{\alpha_1\log\left(1+\dfrac{\bar{x}'^{\frac{a}{h}}-\bar{x}_r^{\frac{a}{h}}}{\bar{x}'^{\frac{a}{h}}+\bar{x}_r^{\frac{a}{h}}}\right)\right\}\right] & if\ i \in R^c \end{cases}$$

The unit estimator of this proposed method of imputation is

$$\bar{y}_{BG1} = \bar{y}_r\left[\exp\left\{\alpha_1\log\left(1+\frac{\bar{x}'^{\frac{a}{h}}-\bar{x}_r^{\frac{a}{h}}}{\bar{x}'^{\frac{a}{h}}+\bar{x}_r^{\frac{a}{h}}}\right)\right\}\right] \quad (2.1)$$

(2)

$$y_{2.i} = \begin{cases} y_i & if\ i \in R \\ \dfrac{1}{n-r}\bar{y}_r\left[n\exp\left\{\alpha_2\log\left(1+\dfrac{\bar{x}'^{\frac{a}{h}}-\bar{x}_n^{\frac{a}{h}}}{\bar{x}'^{\frac{a}{h}}+\bar{x}_n^{\frac{a}{h}}}\right)\right\}\right]-r & if\ i \in R^c \end{cases}$$

The unit estimator of this proposed method of imputation is

$$\bar{y}_{BG2} = \bar{y}_r\left[\exp\left\{\alpha_2\log\left(1+\frac{\bar{x}'^{\frac{a}{h}}-\bar{x}_n^{\frac{a}{h}}}{\bar{x}'^{\frac{a}{h}}+\bar{x}_n^{\frac{a}{h}}}\right)\right\}\right] \quad (2.2)$$

(3)

$$y_{3.i} = \begin{cases} y_i & if\ i \in R \\ \dfrac{1}{n-r}\bar{y}_r\left[n\exp\left\{\alpha_3\log\left(1+\dfrac{\bar{x}_n^{\frac{a}{h}}-\bar{x}_r^{\frac{a}{h}}}{\bar{x}_n^{\frac{a}{h}}+\bar{x}_r^{\frac{a}{h}}}\right)\right\}\right]-r & if\ i \in R^c \end{cases}$$

The unit estimator of this proposed method of imputation is

$$\bar{y}_{BG3} = \bar{y}_r\left[\exp\left\{\alpha_3\log\left(1+\frac{\bar{x}_n^{\frac{a}{h}}-\bar{x}_r^{\frac{a}{h}}}{\bar{x}_n^{\frac{a}{h}}+\bar{x}_r^{\frac{a}{h}}}\right)\right\}\right] \quad (2.3)$$

Where $\alpha_i\ \forall\ i = 1, 2$ and 3 are constants to be determined such that MSEs of the proposed classes of estimators are $\bar{y}_{BG1}, \bar{y}_{BG2}, \bar{y}_{BG3}$ minimum and a and $h \neq 0$ are real numbers.

2.4.1 Properties of the proposed estimators

Consider the following approximations to find the bias and MSE of the proposed estimators

$\bar{y}_r = \bar{Y}(1+e_1)$; $\bar{x}_r = \bar{X}(1+e_2)$; $\bar{x}_n = \bar{X}(1+e_3)$; $\bar{x}' = \bar{X}(1+e_3')$

Where $e_i, e_3' \in (-1,1)\ \forall\, i = 1,2,3$ and $E(e_i) = 0, E(e_3') = 0\ \forall\, i = 1,2,3.$

$E(e_1^2) = \xi_1 C_Y^2$; $E(e_2^2) = \xi_1 C_X^2$; $E(e_3^2) = \xi_2 C_X^2$; $E(e_3'^2) = \xi_3 C_X^2$; $E(e_1 e_2) = \xi_1 \rho_{YX} C_Y C_X$; $E(e_1 e_3) = \xi_2 \rho_{YX} C_Y C_X$;

$E(e_1 e_3') = \xi_3 \rho_{YX} C_Y C_X$; $E(e_2 e_3) = \xi_2 C_X^2$; $E(e_2 e_3') = \xi_3 C_X^2$; $E(e_3 e_3') = \xi_3 C_X^2$.

With the above approximations, the unit estimators $\bar{y}_{BG1}$, $\bar{y}_{BG2}$, and $\bar{y}_{BG3}$ take the following forms in terms of e_i and e_i'

$$\bar{y}_{BG1} = \bar{Y}\left\{1 + e_1 + \frac{1}{2}\alpha_1 \frac{a}{h}(e_3' - e_2 + e_1 e_3' - e_1 e_2) + \frac{1}{4}(e_3'^2 - e_2^2)\left(\alpha_1 \frac{a}{h}\left(2\frac{a}{h} - 1\right)\right)\right\} \tag{2.4}$$

$$\bar{y}_{BG2} = \bar{Y}\left\{1 + e_1 + \frac{1}{2}\alpha_2 \frac{a}{h}(e_3' - e_3 + e_1 e_3' - e_1 e_3) + \frac{1}{4}(e_3'^2 - e_3^2)\left(\alpha_2 \frac{a}{h}\left(2\frac{a}{h} - 1\right)\right)\right\} \tag{2.5}$$

$$\bar{y}_{BG3} = \bar{Y}\left\{1 + e_1 + \frac{1}{2}\alpha_3 \frac{a}{h}(e_3 - e_2 + e_1 e_3 - e_1 e_3) + \frac{1}{4}(e_3^2 - e_2^2)\left(\alpha_3 \frac{a}{h}\left(2\frac{a}{h} - 1\right)\right)\right\} \tag{2.6}$$

2.4.2 Bias and Mean square error of $\bar{y}_{BG1}, \bar{y}_{BG2}, \bar{y}_{BG3}$

Constructing $(\bar{y}_{BG1} - \bar{Y})$ from the equation (2.4) and taking expectations on both sides, we will get

$$E(\bar{y}_{BG1} - \bar{Y}) = E\left[\bar{Y}\left\{e_1 + \frac{1}{2}K(e_3' - e_2 + e_1 e_3' - e_1 e_2) + \frac{1}{4}(e_3'^2 - e_2^2)K\left(2\frac{a}{h} - 1\right)\right\} - \frac{1}{4}(e_3' - e_2)^2\left(K\frac{a}{h}(1-\alpha)\right)\right]$$

Putting the corresponding values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall\, i, j = 1, 2$ and 3 under design I, the bias of $\bar{y}_{BG1}$ is

$$B(\bar{y}_{BG1})_I = \bar{Y}\left\{\frac{1}{2}K(\zeta_3 - \zeta_1)\rho_{YX} C_Y C_X + \frac{1}{4}(\zeta_3 - \zeta_2)\left(\frac{3aK}{h} - K(1+K)\right)C_X^2\right\}$$

Using equation (2.4) to construct $(\bar{y}_{BG1} - \bar{Y})^2$ and keeping the terms up to the second-degree of approximations, we will get

$$(\bar{y}_{BG1} - \bar{Y})^2 = \bar{Y}^2\left\{e_1^2 + \frac{1}{4}K^2(e_3' - e_2)^2 + K(e_1 e_3' - e_1 e_2)\right\} \tag{2.7}$$

Using equation (2.7) in taking expectation on both sides to put the values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall\, i, j = 1, 2$ and 3 under design I, the mean square error of $\bar{y}_{BG1}$ is

$$M(\bar{y}_{BG1})_I = \bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}K^2(\zeta_1 - \zeta_3)C_X^2 + K(\zeta_3 - \zeta_1)\rho_{YX} C_Y C_X\right\} \tag{2.8}$$

Where $K = \alpha_i \frac{a}{h}\ \forall\, i = 1,2,3$ respectively.

Constructing $(\bar{y}_{BG2} - \bar{Y})$ from the equation (2.5) and taking expectations on both sides, we will get

$$E(\bar{y}_{BG2} - \bar{Y}) = E\left[\bar{Y}\left\{e_1 + \frac{1}{2}K(e_3' - e_3 + e_1 e_3' - e_1 e_3) + \frac{1}{4}(e_3'^2 - e_3^2)K\left(2\frac{a}{h} - 1\right)\right\} - \frac{1}{4}(e_3' - e_3)^2\left(K\frac{a}{h}(1-\alpha_2)\right)\right]$$

Putting the corresponding values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall i, j = 1, 2$ and 3 under design I the bias of $\bar{y}_{BG2}$ is

$$B(\bar{y}_{BG2})_I = \bar{Y}\left\{\frac{1}{2}K(\zeta_3 - \zeta_2)\rho_{YX}C_Y C_X + \frac{1}{4}(\zeta_3 - \zeta_2)\left(\frac{3aK}{h} - K(1+K)\right)C_X^2\right\}$$

Using equation (2.5) to construct $(\bar{y}_{BG2} - \bar{Y})^2$ and keeping the terms up to the second-degree of approximations, we will obtain

$$(\bar{y}_{BG2} - \bar{Y})^2 = \bar{Y}^2\left\{e_1^2 + \frac{1}{4}K^2(e_3' - e_3)^2 + K(e_1 e_3' - e_1 e_3)\right\} \quad (2.9)$$

Using equation (2.9) in taking expectation on both sides to put the values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall i, j = 1, 2$ and 3 under design I, the mean square error of $\bar{y}_{BG2}$ is

$$M(\bar{y}_{BG2})_I = \bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}K^2(\zeta_2 - \zeta_3)C_X^2 + K(\zeta_3 - \zeta_2)\rho_{YX}C_Y C_X\right\} \quad (2.10)$$

Constructing $\left(\bar{y}_{BG3} - \bar{Y}\right)$ from the equation (2.6) and taking expectations on both sides, we will get

$$E(\bar{y}_{BG3} - \bar{Y}) = E\left[\bar{Y}\left\{e_1 + \frac{1}{2}K(e_3 - e_2 + e_1 e_3 - e_1 e_2) + \frac{1}{4}(e_3^2 - e_2^2)K\left(2\frac{a}{h} - 1\right)\right\} - \frac{1}{4}(e_3 - e_2)^2\left(K\frac{a}{h}(1 - \alpha_3)\right)\right]$$

Putting the corresponding values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall i, j = 1, 2$ and 3 under design I, the bias of $\bar{y}_{BG3}$ is

$$B(\bar{y}_{BG3})_I = \bar{Y}\left[\frac{1}{4}K(\zeta_2 - \zeta_1)\left\{2\rho_{YX}\frac{C_Y}{C_X} + \left(\frac{2a}{h} - 1\right) + \frac{a}{h}(1 - \alpha_3)\right\}\right]C_X^2$$

Using equation (2.6) to construct $(\bar{y}_{BG3} - \bar{Y})^2$ and keeping the terms up to the second-degree of approximations, we will obtain

$$(\bar{y}_{BG3} - \bar{Y})^2 = \bar{Y}^2\left\{e_1^2 + \frac{1}{4}K^2(e_3 - e_2)^2 + K(e_1 e_3 - e_1 e_2)\right\} \quad (2.11)$$

Using equation (2.11) in taking expectation on both sides to put the values of $E(e_i^2), E(e_3'^2), E(e_i e_j), E(e_i e_3'), \forall i, j = 1, 2$ and 3 under design I, the mean square error of $\bar{y}_{BG3}$ is

$$M(\bar{y}_{BG3})_I = \bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}K^2(\zeta_1 - \zeta_2)C_X^2 - K(\zeta_1 - \zeta_2)\rho_{YX}C_Y C_X\right\} \quad (2.12)$$

2.4.3 *Optimum Mean Square errors of* $\bar{y}_{BG1}, \bar{y}_{BG2}$ *and* $\bar{y}_{BG3}$

The MSE of estimators $\bar{y}_{BG1}, \bar{y}_{BG2}$ and $\bar{y}_{BG3}$ mentioned in equations (2.8), (2.10) and (2.12) under design I are the functions with unknown scalars $\alpha_i \ \forall i = 1, 2, 3$.

Differentiating equation (2.8) in respect of α_1 for equating it to zero

$$\frac{d}{d\alpha_1}\left[\bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}\left(\alpha_1\frac{a}{h}\right)^2(\zeta_1 - \zeta_3)C_X^2 + \alpha_1\frac{a}{h}(\zeta_3 - \zeta_1)\rho_{YX}C_Y C_X\right\}\right] = 0$$

$$\Rightarrow \alpha_1 = \frac{2h}{a}\rho_{YX}\frac{C_Y}{C_X} = \alpha_{1(opt)\,I}$$

Putting the optimum value of α_1 i.e. $\alpha_{1(opt)I}$ in equation (2.8), the optimum MSE of $\bar{y}_{BG1}$

$$M(\bar{y})_{BG1(opt)I} = \bar{Y}^2\left\{\zeta_1 - (\zeta_1 - \zeta_3)\rho_{YX}^2\right\}C_Y^2$$

Differentiating equation (2.11) in respect of α_2 for equating it to zero

$$\frac{d}{d\alpha_2}\left[\bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}\left(\alpha_2\frac{a}{h}\right)^2(\zeta_2 - \zeta_3)C_X^2 + \left(\alpha_2\frac{a}{h}\right)(\zeta_3 - \zeta_2)\rho_{YX}C_Y C_X\right\}\right] = 0$$

$$\Rightarrow \alpha_2 = \frac{2h}{a}\rho_{YX}\frac{C_Y}{C_X} = \alpha_{2(opt)I}$$

Substituting the optimum value of α_2 i.e. $\alpha_{2(opt)I}$ in equation (2.10), the optimum MSE of $\bar{y}_{BG2}$

$$M(\bar{y})_{BG2(opt)I} = \bar{Y}^2\left\{\zeta_1 - (\zeta_2 - \zeta_3)\rho_{YX}^2\right\}C_Y^2$$

Differentiating equation (2.12) in respect of α_3 for equating it to zero

$$\frac{d}{d\alpha_3}\left[\bar{Y}^2\left\{\zeta_1 C_Y^2 + \frac{1}{4}\left(\alpha_3\frac{a}{h}\right)^2(\zeta_1 - \zeta_2)C_X^2 - \left(\alpha_3\frac{a}{h}\right)(\zeta_1 - \zeta_2)\rho_{YX}C_Y C_X\right\}\right] = 0$$

$$\Rightarrow \alpha_3 = \frac{2h}{a}\rho_{YX}\frac{C_Y}{C_X} = \alpha_{3(opt)I}$$

Substituting the optimum value of α_3 i.e. $\alpha_{3(opt)I}$ in equation (2.11), the optimum MSE of $\bar{y}_{BG3}$

$$M(\bar{y})_{BG3(opt)I} = \bar{Y}^2\left\{\zeta_1 - (\zeta_1 - \zeta_2)\rho_{YX}^2\right\}C_Y^2$$

3. Results and Discussion

3.1 Efficiency Comparison

This section presents a theoretical and empirical demonstration of the performance of the suggested methods of imputation and its resulting estimators over the mean, ratio and compromised methods of imputation.

3.1.1 Theoretical Comparison

Here, we derive the conditions under which the suggested estimators perform better than the mentioned traditional estimators under both the designs. Theoretical comparisons under design I are as follows:

Comparison with mean method:

a) $\Delta_{1r} = M(\bar{y}_r) - M\left(\bar{y}_{BG1(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 + (\zeta_1 - \zeta_3)\rho_{YX}^2\right\}C_Y^2 \geq 0$

So, the proposed estimator $\bar{Y}_{BG1}$ is better than $\bar{y}_r$.

b) $\Delta_{2r} = M(\bar{y}_r) - M\left(\bar{y}_{BG2(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 + (\zeta_2 - \zeta_3)\rho_{YX}^2\right\}C_Y^2 \geq 0$

So, the proposed estimator $\bar{y}_{BG2}$ is better than $\bar{y}_r$.

c) $\Delta_{3r} = M(\bar{y}_r) - M\left(\bar{y}_{BG3(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 + (\zeta_1 - \zeta_2)\rho_{YX}^2\right\}C_Y^2 \geq 0$

So, the proposed estimator $\bar{y}_{BG3}$ is better than $\bar{y}_r$ under both the designs.

Comparison with ratio method

a) $\Delta_{1RAT} = M(\bar{y}_{RAT}) - M\left(\bar{y}_{BG1(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 C_Y^2 + (\zeta_1 - \zeta_2)(C_X - \rho_{YX}C_Y)^2 - (\zeta_2 - \zeta_3)\rho_{YX}^2\right\} > 0$

So, the proposed estimator $\bar{y}_{BG1}$ is better than $\bar{y}_{RAT}$.

b) $\Delta_{2RAT} = M(\bar{y}_{RAT}) - M\left(\bar{y}_{BG2(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 C_Y^2 + (\zeta_1 - \zeta_2)(C_X - \rho_{YX}C_Y)^2 - (2\zeta_2 - \zeta_1 - \zeta_3)\rho_{YX}^2 C_Y^2\right\} > 0$

When $2\zeta_2 \leq \left(\frac{1}{r} - \frac{1}{N}\right)$.

So, the proposed estimator $\bar{y}_{BG2}$ is better than $\bar{y}_{RAT}$.

c) $\Delta_{3RAT} = M(\bar{y}_{RAT}) - M\left(\bar{y}_{BG3(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 C_Y^2 + (\zeta_1 - \zeta_2)(C_X - \rho_{YX}C_Y)^2\right\} > 0$

So, the proposed estimator $\bar{y}_{BG3}$ is better than $\bar{y}_{RAT}$.

Comparison with compromised method

a) $\Delta_{1comp} = M(\bar{y}_{comp}) - M\left(\bar{y}_{BG1(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 C_Y^2 - (\zeta_1 - \zeta_2)\rho_{YX}^2 C_Y^2\right\} > 0$

So, the proposed estimator $\bar{y}_{BG1}$ is better than $\bar{y}_{comp}$ under both the designs.

b) $\Delta_{2comp} = M(\bar{y}_{comp}) - M\left(\bar{y}_{BG2(opt)I}\right) = \bar{Y}^2\left\{\zeta_3 C_Y^2 + (2\zeta_2 - \zeta_1 - \zeta_3)\rho_{YX}^2 C_Y^2\right\} > 0$

So, the proposed estimator $\bar{y}_{BG2}$ is better than $\bar{y}_{comp}$ under both the designs.

c) $\Delta_{3comp} = M(\bar{y}_{comp}) - M\left(\bar{y}_{BG3(opt)I}\right) = \bar{Y}^2\zeta_3 C_Y^2 > 0$

So, the proposed $\bar{y}_{BG3}$ is better than $\bar{y}_{comp}$.

3.2 Empirical study

To analyse the efficacy of the proposed estimators, we have selected a population data sets. The specifics of these data sets are delineated below.

Population A Source: District Census Handbook, Orissa (1981) The demographic information pertaining to 109 Villages/ Towns/Wards within the urban area under the jurisdiction of police station Baria, Tahasil-Champua, Orissa has been incorporated for examination. In this context, X denotes the mean number of non-workers within the village, while Y signifies the mean number of literate individuals residing in the village.

Here, we have considered the sample sizes $n' = 50$ fixed and subsample (n) between 10% to 45% and the response rate (r) between the 75% to 95% with different correlation coefficients.

N=109; n =16, 27, 31; r =(8, 12, 14), (13, 19, 24), (19, 22, 28)

$\bar{Y} = 145.3028, \bar{X} = 259.083, \rho_{YX} = 0.875, C_Y^2 = 0.4759, C_X^2 = 0.2947$

The percentage relative efficiencies (PREs) of the proposed estimators and exiting estimators with respect to the mean method of imputation $\bar{y}_r$ have been calculated and presented in Tables 1 to 4. The expressions of PREs are given as

$$PRE(t) = 100 \times \frac{MSE(\bar{y}_r)}{MSE(t)}$$

Where $t = \bar{y}_{RAT}, \bar{y}_{comp}, \bar{y}_{BG1}, \bar{y}_{BG2}, \bar{y}_{BG3}$

We have computed MSE and the Percentage Relative Efficiencies (PRE) of different estimators under this population data set.

Table 1: MSE and PRE of the existing estimator estimators under design I

n	r	$M(\bar{y}_r)$	$M(\bar{y}_{RAT})$	$M(\bar{y}_{comp})$	$M(y_{BG1})$	$M(y_{BG2})$	$M(y_{BG3})$
16	8	1,163.7738	687.8505	682.9789	376.2975	857.0923	488.08385
16	12	745.1225	586.4814	584.8575	278.1761	438.4410	389.96246
16	14	625.5078	557.5188	556.8228	250.1414	318.8264	361.92777
27	13	680.7146	376.9915	373.8825	263.0805	569.9125	178.98748
27	19	436.6426	317.8937	316.6782	205.8761	325.8405	121.78311
27	24	326.4712	291.2176	290.8568	180.0547	215.6691	95.96169
31	19	436.6426	281.5029	279.9149	205.8761	362.6038	85.01982
31	22	364.5304	264.0422	263.0136	188.9748	290.4916	68.11853
31	28	266.6639	240.3455	240.0761	166.0373	192.6251	45.18106

Table 2: MSE and PRE of the proposed estimators under design I

n	r	$PRE(\bar{y}_r)$	$PRE(\bar{y}_{RAT})$	$PRE(\bar{y}_{comp})$	$PRE(y_{BG1})$	$PRE(y_{BG2})$	$PRE(y_{BG3})$
16	8	100	169.1899	170.3967	309.2696	135.7816	238.4373
16	12	100	127.0496	127.4024	267.8600	169.9482	191.0754
16	14	100	112.1949	112.3352	250.0617	196.1907	172.8267
27	13	100	180.5650	182.0664	258.7477	119.4419	380.3141
27	19	100	137.3549	137.8821	212.0900	134.0050	358.5412
27	24	100	112.1056	112.2447	181.3178	151.3760	340.2099
31	19	100	155.1112	155.9912	212.0900	120.4186	513.5774
31	22	100	138.0576	138.5975	192.8989	125.4874	535.1413
31	28	100	110.9502	111.0747	160.6047	138.4367	590.2116

3.3 Simulation Analysis

To determine how well the suggested imputation techniques perform in comparison to mean and ratio approaches of imputation, we conduct simulation research in this part utilizing statistical computational software [11]. We used the 'genCorGen' function in the package 'simstudy' to generate data sets for this manuscript [12]. We created data sets from the normal distribution using the specified parameters and correlation coefficients for the study and auxiliary variables. To create a dataset of the variables (X, Y) and determine the necessary parameters and correlation coefficient, we take into account a population size of N = 8500 and we consider one simulation combination for $\beta = 2$.

For data from Normal distribution: Data= (X, Y).

$X:(321.0029, 0.9986); Y:(457.0018, 0.9925)$

The following steps have been used for the simulation of the required population:

Step 1: The 'genCorGen' function is used to create a data set with the normal distribution of variables X and Y of size N = 8500 using the statistical computer program [12] Software.

Step 2: The parameters have been calculated.

Step 3: A sample of sizes $n' = 2500$, n = 340, 765, 1020, 1445, 1700, and 1870 (i.e., sample sizes lie between 4% and 22%) was chosen from this artificial data set with response rates r = (265, 282, 296, 313), (597,635, 666, 704), (796, 847, 887, 939), (1127, 1199, 1257, 1329), (1326, 1411, 1479, 1564), and (1459,552, 1627, 1720), respectively (i.e., the response rate lies between 78% and 92% of the total).

Step 4: Sample statistics i.e. sample mean, sample variance, and the values of the proposed and existing estimators are calculated for this sample under imputation techniques.

Step 5: Steps (3) and (4) are repeated 50,000 times.
Step 6: The MSE of every estimator is calculated through the formula of the MSE given by $MSE(\hat{T}) = \frac{1}{50000}\sum_{i=1}^{50000}(\hat{T}_i - \bar{Y})^2$.

MSE of the existing and proposed estimators under design I

Table 3: MSE of the existing and proposed estimators under design I

n	r	$M(\bar{y}_r)$	$M(\bar{y}_{RAT})$	$M(\bar{y}_{comp})$	$M(y_{BG1})$	$M(y_{BG2})$	$M(y_{BG3})$
340	265	757.8474	585.78187	585.77842	60.32251	232.39148	527.24886
340	282	710.6914	585.64898	585.64648	60.19057	185.23549	527.11692
340	296	675.9242	585.55101	585.54920	60.09329	150.46825	527.01964
340	313	637.8881	585.44382	585.44277	59.98686	112.43216	526.91321
765	597	322.8358	246.79776	246.79623	59.10533	135.14489	188.26667
765	635	302.0571	246.73920	246.73809	59.04719	114.36619	188.20853
765	666	286.8622	246.69638	246.69558	59.00468	99.17133	188.16602
765	704	270.0618	246.64904	246.64857	58.95767	82.37093	188.11901
1,020	796	236.0300	179.00149	179.00035	58.86245	115.89212	120.47078
1,020	847	220.3496	178.95730	178.95647	58.81857	100.21173	120.42691
1,020	887	209.3130	178.92620	178.92559	58.78769	89.17512	120.39603
1,020	939	196.3711	178.88973	178.88938	58.75148	76.23325	120.35982
1,445	1,127	159.5454	119.18156	119.18075	58.64844	99.01311	60.65119
1,445	1,199	148.5003	119.15043	119.14984	58.61754	87.96794	60.62028
1,445	1,257	140.5229	119.12795	119.12752	58.59521	79.99063	60.59796
1,445	1,329	131.5888	119.10278	119.10253	58.57022	71.05644	60.57296
1,700	1,326	131.9416	97.64619	97.64550	58.57120	92.86734	39.11594
1,700	1,411	122.5242	97.61965	97.61915	58.54485	83.44994	39.08959
1,700	1,479	115.7697	97.60061	97.60025	58.52595	76.69538	39.07069
1,700	1,564	108.1525	97.57915	97.57894	58.50464	69.07815	39.04938
1,870	1,459	117.6910	86.55212	86.55150	58.53133	89.67078	28.02194
1,870	1,552	109.1772	86.52813	86.52768	58.50751	81.15707	27.99812
1,870	1,627	103.0203	86.51078	86.51045	58.49028	75.00013	27.98089
1,870	1,720	96.1314	86.49137	86.49117	58.47101	68.11123	27.96161

PRE of the existing and proposed estimators under design I

Table 4: PRE of the existing and proposed estimators under design I

n	r	y_r	y_{RAT}	y_{comp}	y_{BG1}	y_{BG2}	y_{BG3}
340	265	100	129.3737	129.3744	1,256.3260	326.1081	143.7362
340	282	100	121.3511	121.3516	1,180.7355	383.6691	134.8261
340	296	100	115.4339	115.4342	1,124.7914	449.2138	128.2541
340	313	100	108.9580	108.9582	1,063.3796	567.3537	121.0613
765	597	100	130.8099	130.8107	546.2042	238.8812	171.4779
765	635	100	122.4196	122.4201	511.5520	264.1140	160.4906
765	666	100	116.2815	116.2819	486.1686	289.2592	152.4517
765	704	100	109.4923	109.4926	458.0606	327.8606	143.5590
1,020	796	100	131.8592	131.8601	400.9857	203.6636	195.9230
1,020	847	100	123.1297	123.1303	374.6259	219.8841	182.9737
1,020	887	100	116.9829	116.9833	356.0491	234.7213	173.8538
1,020	939	100	109.7722	109.7724	334.2403	257.5925	163.1534
1,445	1,127	100	133.8675	133.8685	272.0369	161.1356	263.0541
1,445	1,199	100	124.6326	124.6332	253.3376	168.8118	244.9679
1,445	1,257	100	117.9597	117.9601	239.8198	175.6742	231.8938
1,445	1,329	100	110.4834	110.4836	224.6684	185.1891	217.2401
1,700	1,326	100	135.1222	135.1231	225.2671	142.0754	337.3091
1,700	1,411	100	125.5119	125.5125	209.2827	146.8236	313.4447
1,700	1,479	100	118.6157	118.6162	197.8091	150.9474	296.3083
1,700	1,564	100	110.8356	110.8359	184.8613	156.5653	276.9633
1,870	1,459	100	135.9770	135.9779	201.0734	131.2478	419.9958
1,870	1,552	100	126.1754	126.1761	186.6038	134.5259	389.9450
1,870	1,627	100	119.0838	119.0842	176.1323	137.3602	368.1810
1,870	1,720	100	111.1457	111.1459	164.4087	141.1388	343.7978

4. Conclusion

In this paper, we have proposed three imputation methods along with their point estimators to estimate the population mean of study variable. The bias, mean square error and minimum mean square error have been derived. Empirical and simulation studies have been carried out to examine the efficiency of the proposed estimators.

The current study allows for an inference of the following conclusions. From Tables 1 to 4, for all cases where empirical studies have been conducted under two-phase sampling design. We generated a data set utilizing the statistical software R, derived from a normal distribution, and conducted a simulation to analyse the efficiencies of the proposed estimators. Our examination of tables (3-4) indicates that the efficiency outcomes of the proposed estimators exhibit almost a similar trend to that presented in tables (1-2). The MSE of proposed estimators $\bar{y}_{BG1}$, $\bar{y}_{BG2}$ and $\bar{y}_{BG3}$ are less than the estimators $\bar{y}_r$, $\bar{y}_{RAT}$, and $\bar{y}_{comp}$, showing more PRE for the proposed estimators in maximum cases. This demonstrates how the suggested imputation methods and its resulting estimators are better than the traditional imputation methods. This method improves the estimator's robustness and accuracy, which makes it a useful tool in contemporary data science and statistical analysis. Survey practitioners may be encouraged for their practical applications after observing promising behaviour of the suggested imputation methods, if non-response is not ignorable in the survey data.

5. References

1. Grover LK, Sharma A. Predictive estimation of finite population mean in case of missing data under two-phase sampling. Journal of Statistical Theory and Applications, 2023, https://doi.org/10.1007/s44199-023-00064-6
2. Singh H, Gupta A, Tailor R. Efficient class of estimators for finite population mean using auxiliary attribute in stratified random sampling. Scientific Reports, 2023, p. 13. https://doi.org/10.1038/s41598-023-34603-z
3. Bhushan S, Kumar A, Zaman T, Al Mutairi A. Efficient difference and ratio-type imputation methods under ranked set sampling. Axioms. 2023;12(6):588. https://doi.org/10.3390/axioms12060558
4. Yadav VK, Prasad S. A simulation based optimization of factor-type exponential estimators in sample surveys with coefficients of variation and kurtosis. Franklin Open. 2023; https://doi.org/10.1016/j.fraope.2023.100050
5. Sanaullah A, Sabir S. Combined type generalized classes of efficient estimators of mean in joint presence of measurement error and non-response using stratified two-phase sampling. Scientia Iranica, 2022. https://doi.org/10.24200/sci.2022.54750.3906
6. Gohain D, Gogoi UG, Singh BK. Exponential chain ratio estimator of finite population mean with imputation of missing data under two-phase sampling. Advances and Applications in Statistics. 2024;91(9):1113-1135. https://doi.org/10.17654/0972361724059
7. Das J, Singh BK. Some efficient exponential-logarithmic imputation techniques in the presence of non-response using simulation. Advances and Applications in Statistics. 2023;88(1):123-137. https://doi.org/10.17654/0972361723042
8. Shukla D, Jain A. A class of logarithmic-cum-exponential estimators for population mean with risk analysis using double sampling. Reliability: Theory & Applications. 2024;19(3). https://gnedenko.net/Journal/2024/032024/RTA_3_2024-19.pdf

9. Sabir S, Sanaullah A. Efficient estimation of mean in two-phase sampling when measurement error and non-response are simultaneously present. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences. 2022;92:633-646. https://doi.org/10.1007/s40010-022-00776-x
10. Hassan Y, Ismail M, Murray W, Shahbaz MQ. Efficient estimation combining exponential and ln functions under two-phase sampling. AIMS Mathematics. 2020;5(6):7605-7623. https://doi.org/10.3934/math.2020486
11. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/
12. Goldfeld K, Jens WJ. Simstudy: Simulation of study data. R package version 0.1.15, 2019. Available from: https://CRAN.R-project.org/package=simstudy