International Journal of Statistics and Systems
ISSN 0973-2675 Volume 21, Number 1 (2026) pp. 77-92
Research India Publications
https://www.ripublication.com

Auxiliary Information-Based Estimation of the Finite Population Mean in the Presence of Non-Response

Manoj K. Chaudhary* and Saurabh
Department of Statistics
Institute of Science
Banaras Hindu University
Varanasi, India
**Corresponding author: E-mail: mk15@bhu.ac.in*

Abstract

In this paper, we have suggested some families of estimators for estimating the mean of finite population in simple random sampling using the information on auxiliary variable(s) under non-response. We have considered three different cases; (i) study of a family of estimators of population mean using the information on an auxiliary variable in the presence of non-response (ii) study of a family of estimators of population mean utilizing the information on an auxiliary variable under the situation in which population mean of the auxiliary variable is unknown and non-response is observed on study variable only (iii) study of a family of estimators of population mean using the information on one more additional auxiliary variable when population mean of the first auxiliary variable is not known and study variable suffers from non-response. The properties of the proposed families have been discussed in detail. An empirical study along with the simulation analysis has also been carried out to support the theoretical results.

Keywords: Simple random sampling, population mean, families of estimators, auxiliary information, non-response

1. Introduction

The non-response is a common phenomenon in most of the sample surveys. It is generally assumed that the information would be gathered from all the selected units in the sample, but practically, it is not achievable because of the non-response. Some of the units may not respond or these may not be contacted during the survey period. This generates a huge difficulty in estimating the characteristics of the population as the error due to non-response which diminishes the prescribed sample on one hand and the efficiency of the estimator on the other hand. The problem of non-response was first considered by Hansen and Hurwitz (1946) in mail surveys and they introduced a technique of sub-sampling of non-respondents to cope with the situation of non-response. Cochran (1963), Kish (1965) and Sarndal *et al.* (1992) have cited a number of paradigms of non-response from census and mail surveys and showed the need for developing the techniques to handle the situation. Srivastava (1993) has studied several ratio, product and regression-type estimators under fixed and super population model approaches in the presence of non-response. Rao (1986, 1987, 1990), Singh *et al.* (2010), Kumar and Bhougal (2011), Anderson and Särndal (2016) and others have contributed a lot in dealing with the problem of non-response under simple random sampling. Dealing with stratified population, Rao (1973) and Khare (1987) have suggested the methods for deciding the sampling design in each stratum in the presence of non-response. The phenomenon of non-response in stratified random sampling has further been studied by the authors including Chaudhary *et al.* (2009), Chaudhary *et al* (2011), Singh and Mishra (2020), Singh and Nigam (2022), Wani and Rizvi (2023), Chaudhary *et al.* (2024), Chaudhary and Ray (2024), Chaudhary *et al.* (2026) and others.

Auxiliary information collected from the population units can be successfully utilized to design a convenient and efficient sampling design and after sample selection it can further be used to improve the efficiency of estimators. Khare (1992) has discussed the problem of two-phase sampling regression estimator under non-response. Okafor and Lee (2000) have discussed the method of double sampling for ratio and regression estimation with sub-sampling of non-respondents. Khare and Sinha (2004) have proposed some estimators for estimating the population ratio using two-phase sampling scheme in the presence of non-response. Tabasum and Khan (2006) have developed double sampling ratio estimator for the population mean in presence of non-response. Chaudhary and Kumar (2016) have suggested some estimators for estimating the population mean using double sampling under non-response. Chaudhary *et al.* (2021) have proposed some improved estimators of population mean using two-phase sampling scheme in the presence of non-response.

In the present paper, we have proposed some families of estimators for estimating the population mean in simple random sampling utilizing the auxiliary information under non-response. In order to propose the families of estimators, we have considered three different cases; (i) the case where information on an auxiliary variable with known mean is used (ii) the case where information on an auxiliary variable with unknown mean is utilized (iii) the case where information on one more additional auxiliary variable with known mean is used. It is further assumed that the study variable suffers from non-response and auxiliary variables are free from non-response. We have presented the detailed discussion on the proposed families of estimators. An empirical illustration along with the simulation analysis has also been presented to strengthen the theoretical results.

2. Sampling Strategies and Proposed Families of Estimators

In the subsequent sub-sections, we have discussed all the three cases one by one. The sampling strategies and proposed families of estimators along with their properties have been conferred in each case.

2.1 Utilization of Information on an Auxiliary Variable with Known Mean

Suppose a population consists of N units $(U_1, U_2, ..., U_N)$. Let Y and X be the study and auxiliary variables with respective population means $\bar{Y}$ and $\bar{X}$. Let y_i and $x_i \left(i = 1,2,...N\right)$ be the observation on the i^{th} unit in the population for study and auxiliary variables respectively. Our objective is to estimate the population mean $\bar{Y}$ in such a situation in which study variable suffers from non-response and auxiliary variable is free from non-response. Let a sample of n units be selected from N units by the method of simple random sampling without replacement (SRSWOR) scheme. It is further observed that out of n units, n_1 units respond and n_2 units do not respond on the study variable Y. Using Hansen and Hurwitz (1946) technique of sub-sampling of non-respondents, we now select a sub-sample of h_2 units from the n_2 non-responding units by SRSWOR ($h_2 = \frac{n_2}{L}, L > 1$) and collect the information from all the h_2 units. Thus, the estimator of population mean $\bar{Y}$ without using auxiliary information is given as

$$\bar{y}^* = \frac{n_1 \bar{y}_{n1} + n_2 \bar{y}_{h2}}{n} \tag{1}$$

where $\bar{y}_{n1}$ and $\bar{y}_{h2}$ are respectively the means based on n_1 responding units and h_2 non-responding units.

The variance of the estimator $\bar{y}^*$ is given by

$$V\left(\bar{y}^*\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_Y^2 + \frac{\left(L-1\right)}{n} W_2 S_{Y2}^2 \tag{2}$$

where S_Y^2 and S_{Y2}^2 are respectively the mean squares of entire group and non-response group for the study variable. W_2 is the non-response rate in the population.

Since, complete information is obtained on all n units for the auxiliary variable. Thus, the estimator of $\bar{X}$ is simply given as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

The variance of the estimator $\bar{x}$ is given by

$$V\left(\bar{x}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) S_X^2 \tag{4}$$

where S_X^2 is the mean square of entire group for the auxiliary variable.

Motivated by Singh et al. (2017), we now propose a family of estimators for estimating the population mean $\overline{Y}$ using the information on an auxiliary variable under non-response as

$$T_1^* = \bar{y}^* \left[\frac{a\overline{X} + b\bar{x}}{c\bar{x} + d\overline{X}} \right] \tag{5}$$

where a, b, c and d are suitably chosen constants. The constants a, b, c and d may assume real values as well as parametric values.

In order to obtain the bias and mean square error (MSE) of the proposed family T_1^*, we use large sample approximation. Let us consider

$\bar{y}^* = \overline{Y}(1 + e_0)$ and $\bar{x} = \overline{X}(1 + e_1)$

such that $E(e_0) = E(e_1) = 0$,

$$E(e_0)^2 = \frac{1}{\overline{Y}^2}\left[\left(\frac{1}{n} - \frac{1}{N}\right)S_Y^2 + \frac{W_2(k-1)}{n}S_{Y2}^2\right],\ E(e_1)^2 = \frac{1}{\overline{X}^2}\left(\frac{1}{n} - \frac{1}{N}\right)S_X^2 \text{ and}$$

$$E(e_0 e_1) = \frac{1}{\overline{X}\overline{Y}}\left(\frac{1}{n} - \frac{1}{N}\right)\rho_{XY} S_X S_Y$$

where ρ_{XY} is the population correlation coefficient between study and auxiliary variables.

Expressing the equation (5) in terms of e_0 and e_1, we get

$$T_1^* = \overline{Y}(1 + e_0)\left[\frac{a+b}{c+d}\right]\left[\left\{1 + b(a+b)^{-1}e_1\right\}\left\{1 + c(c+d)^{-1}e_1\right\}^{-1}\right]$$

Assuming $(a+b) = (c+d)$ and putting $\theta_1 = c(c+d)^{-1}$ and $\theta_2 = b(a+b)^{-1}$, the above expression reduces to

$$T_1^* = \overline{Y}(1 + e_0)\left[\{1 + (\theta_2 e_1)\}\{1 + (\theta_1 e_1)\}^{-1}\right]$$

Now, we assume $|\theta_1 e_1| < 1$, so that $(1 + \theta_1 e_1)^{-1}$ is expandable. Thus, on expanding the above expression and neglecting the terms having powers of e_0 and e_1 greater than two, we get

$$\left(T_1^* - \bar{Y}\right) = \bar{Y}\left(e_0 - \delta e_1 - \delta e_0 e_1 + \theta_1 \delta e_1^2\right) \qquad (6)$$

where $\delta = (\theta_1 - \theta_2)$.

Taking expectation on both the sides of equation (6), we get the bias of T_1^* up to the first order of approximation as

$$Bais\left(T_1^*\right) = \frac{\delta}{\bar{X}}\left(\frac{1}{n} - \frac{1}{N}\right)\left[\theta_1 S_X^2 R - \rho_{XY} S_X S_Y\right] \qquad (7)$$

where $R = \dfrac{\bar{Y}}{\bar{X}}$.

Squaring both the sides of equation (6) and then taking expectation on neglecting the terms involving powers of e_0, e_1 higher than two, we get the MSE of T_1^* up to the first order of approximation as

$$MSE\left(T_1^*\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[S_Y^2 + \delta^2 R^2 S_X^2 - 2\delta\rho_{XY} R S_X S_Y\right] + \frac{(k-1)}{n} W_2 S_{Y2}^2 \qquad (8)$$

Remark 1: The proposed family can produce a number of well known existing estimators for the suitable choices of a, b, c and d under non-response [See Singh et al. (2017)]. The non-response versions of some new members of the proposed family can also be obtained by considering the suitable choices of a, b, c and d. The Table 1 shows some well known existing estimators and some new members of the proposed family.

Table 1: Some Existing and New Members of Proposed Family T_1^*

Choice of a, b, c and d				Estimators
a	b	c	d	
1	1	1	1	$\bar{y}^*$ (usual mean)
1	0	1	0	$\bar{y}_R^*$ (ratio-type)
0	1	0	1	$\bar{y}_P^*$ (product-type)
C_X^2	ρ_{XY}	C_X^2	ρ_{XY}	$\bar{y}_1^* = \bar{y}^*\left[\dfrac{C_X^2\bar{X} + \rho_{XY}\bar{x}}{C_X^2\bar{x} + \rho_{XY}\bar{X}}\right]$
C_X	M^2	C_X	M^2	$\bar{y}_2^* = \bar{y}^*\left[\dfrac{C_X\bar{X} + M^2\bar{x}}{C_X\bar{x} + M^2\bar{X}}\right]$
ρ_{XY}^2	f	ρ_{XY}^2	f	$\bar{y}_3^* = \bar{y}^*\left[\dfrac{\rho_{XY}^2\bar{X} + f\bar{x}}{\rho_{XY}^2\bar{x} + f\bar{X}}\right]$
C_Y	M^2	C_Y	M^2	$\bar{y}_4^* = \bar{y}^*\left[\dfrac{C_Y\bar{X} + M^2\bar{x}}{C_Y\bar{x} + M^2\bar{X}}\right]$

Here $M = \rho_{XY} \dfrac{C_Y}{C_X}$, $C_Y = S_Y / \overline{Y}$, $C_X = S_X / \overline{X}$ and $f = n/N$.

2.2 Utilization of Information on an Auxiliary Variable with Unknown Mean

Sometimes, the information about the population mean of the auxiliary variable is not available. In such a situation, one can use the double sampling (or two-phase sampling) scheme in order to estimate the population mean $\overline{Y}$. To deal with the problem, first a larger preliminary sample is selected from the population to estimate the unknown population mean $\overline{X}$ and then a smaller sub-sample is selected from the preliminary sample. Therefore, a sample of n' units is selected from N units by SRSWOR scheme at the first phase and then a smaller sub-sample of n units is selected from n' ($n < n'$) units by SRSWOR at the second phase. At the first phase, the estimator of $\overline{X}$ is given as

$$\bar{x}' = \frac{1}{n'} \sum_{i=1}^{n'} x_i \qquad (9)$$

The variance of the estimator $\bar{x}'$ is given by

$$V\left(\bar{x}'\right) = \left(\frac{1}{n'} - \frac{1}{N}\right) S_X^2 \qquad (10)$$

At the second phase, it is noted that out of n units, there are n_1 responding units and n_2 non-responding units for the study variable only. Applying Hansen and Hurwitz's (1946) technique of sub-sampling of non-respondents, we select a sub-sample of h_2 units from the n_2 non-responding units by SRSWOR ($h_2 = \dfrac{n_2}{L}, L > 1$) and gather the information on all the h_2 units. Thus, the Hansen and Hurwitz (1946) estimator of $\overline{Y}$ at the second phase is given by equation (1). The estimator of $\overline{X}$ at the second phase is represented by equation (3).

Now, we propose a family of estimators for estimating the population mean $\overline{Y}$ using the information on an auxiliary variable under the situation in which the population mean of auxiliary variable, $\overline{X}$ is unknown and study variable suffers from non-response as

$$T_2^{**} = \bar{y}^* \left[\frac{a\bar{x}' + b\bar{x}}{c\bar{x} + d\bar{x}'} \right] \qquad (11)$$

To obtain the bias and MSE of the proposed family T_2^{**}, we use large sample approximation. Let us assume

$\bar{y}^* = \bar{Y}(1+e_0)$, $\bar{x} = \bar{X}(1+e_1)$, $\bar{x} = \bar{x}'(1+e_2)$ and $\bar{x}' = \bar{X}(1+e_1')$

such that $E(e_0) = E(e_1) = E(e_2) = E(e_1') = 0$,

$E(e_2)^2 = \frac{1}{\bar{X}^2}\left(\frac{1}{n} - \frac{1}{n'}\right)S_X^2$, $E(e_1')^2 = \frac{1}{\bar{X}^2}\left(\frac{1}{n'} - \frac{1}{N}\right)S_X^2$,
$E(e_0 e_2) = \frac{1}{\bar{X}\bar{Y}}\left(\frac{1}{n} - \frac{1}{n'}\right)\rho_{XY} S_X S_Y$,

$$E(e_0 e_1') = \frac{1}{\bar{X}\bar{Y}}\left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{XY} S_X S_Y, E(e_1 e_1') = \frac{1}{\bar{X}^2}\left(\frac{1}{n'} - \frac{1}{N}\right)S_X^2 \text{ and}$$

$$E(e_1 e_2) = \frac{1}{\bar{X}^2}\left(\frac{1}{n} - \frac{1}{n'}\right)S_X^2$$

Expressing the equation (11) in terms of e_0 and e_2 on assuming $(a+b) = (c+d)$, we get

$$T_2^{**} = \bar{Y}(1+e_0)\left[\{1+(\theta_2 e_2)\}\{1+(\theta_1 e_2)\}^{-1}\right]$$

We assume $|\theta_1 e_2| < 1$, so that $(1+\theta_1 e_2)^{-1}$ is expandable. Expanding the right hand side (RHS) of the above expression and then neglecting the terms involving powers of e_0 and e_2 higher than two, we get

$$\left(T_2^{**} - \bar{Y}\right) = \bar{Y}\left(e_0 - \delta e_2 - \delta e_0 e_2 + \theta_1 \delta e_2^2\right) \quad (12)$$

Taking expectation on both the sides of equation (12), we obtain the bias of T_1^{**} up to the first order of approximation as

$$Bias\left(T_2^{**}\right) = \frac{\delta}{\bar{X}}\left(\frac{1}{n} - \frac{1}{n'}\right)\left[\theta_1 S_X^2 R - \rho_{XY} S_X S_Y\right] \quad (13)$$

Squaring both the sides of equation (12) and then captivating expectation on ignoring the terms having powers of e_0, e_2 greater than two, we get the MSE of T_2^{**} up to the first order of approximation as.

$$MSE\left(T_2^{**}\right) = \left(\frac{1}{n} - \frac{1}{n'}\right)\left[S_Y^2 + \delta^2 R^2 S_X^2 - 2\delta\rho_{XY} R S_X S_Y\right] + \left(\frac{1}{n'} - \frac{1}{N}\right)S_Y^2 + \frac{(k-1)}{n} W_2 S_{Y2}^2$$

(14)

Remark 2: It is to be remembered that several well known existing estimators and some new estimators under non-response may be considered as the members of the proposed family T_2^{**}. For the suitable choices of a, b, c and d, one can get some well known existing estimators $\bar{y}^*$, $\bar{y}_R^{**}$, $\bar{y}_P^{**}$ and some new estimators $\bar{y}_1^{**}$, $\bar{y}_2^{**}$, $\bar{y}_3^{**}$, $\bar{y}_4^{**}$ in similar way as we have done in Table 1. The empirical analysis of these estimators has been given in Table 3.

2.3 Utilization of Information on Two Auxiliary Variables

The efficiency of the estimators can be improved by using the information on one more additional auxiliary variable. Let Z be the additional auxiliary variable with population mean $\bar{Z}$, closely related to the first auxiliary variable X but compared to X remotely related to Y. Let z_i $(i = 1,2,...N)$ be the observation on the i^{th} unit in the population for the variable Z .Our aim is to estimate the population mean $\bar{Y}$ using the information on two auxiliary variables X and Z under the situation in which the knowledge about the population mean $\bar{X}$ is not available even as the population mean $\bar{Z}$ is known and study variable suffers from non-response. It is further assumed that the auxiliary variable X has high degree of correlation with the study variable rather than the auxiliary variable Z has with the study variable. If the auxiliary variable Z is highly positively correlated with the variable X, we now propose a family of estimators for estimating the population mean $\bar{Y}$ as

$$T_3^{***} = \bar{y}^* \left[\frac{a\bar{x}'_R + b\bar{x}}{c\bar{x} + d\bar{x}'_R} \right] \tag{15}$$

where $\bar{x}'_R = \bar{x}' \dfrac{\bar{Z}}{\bar{z}'}$, $\bar{z}' = \dfrac{1}{n'}\sum_{i=1}^{n'} z_i$ and $\bar{Z} = \dfrac{1}{N}\sum_{i=1}^{N} z_i$.

Now, we use large sample approximation to get the bias and MSE of the proposed family T_3^{***}. Let

$$\bar{y}^* = \bar{Y}(1+e_0), \bar{x} = \bar{X}(1+e_1),\ \bar{x} = \bar{x}'(1+e_2),\ \bar{x}' = \bar{X}(1+e_1') \text{and } \bar{z}' = \bar{Z}(1+e_2')$$

$$\text{such that } E(e_0) = E(e_1) = E(e_2) = E(e_1') = E(e_2') = 0,$$

$$E(e_2')^2 = \frac{1}{\bar{Z}^2}\left(\frac{1}{n'} - \frac{1}{N}\right) S_Z^2,\ E(e_0 e_2') = \frac{1}{\bar{Y}\bar{Z}}\left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{YZ} S_Y S_Z ,$$

$$E(e_1 e_2') = \frac{1}{\bar{X}\bar{Z}}\left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{XZ} S_X S_Z \text{ and } E(e_1' e_2') = \frac{1}{\bar{X}\bar{Z}}\left(\frac{1}{n'} - \frac{1}{N}\right)\rho_{XZ} S_X S_Z$$

where S_Z^2 is the mean square of entire group for the auxiliary variable Z. ρ_{YZ} is the population correlation coefficient between Y and Z. ρ_{XZ} is the population correlation coefficient between X and Z.

Expressing the equation (15) in terms of e_0, e_1, e_1', and e_2', we get

$$T_3^{***} = \bar{Y}(1+e_0)\left[1+\theta_2' e_1' + \theta_2 (e_2' + e_1 + e_1 e_2')\right]\left[1+\{\theta_1' e_1' + \theta_1 (e_2' + e_1 + e_1 e_2')\}\right]^{-1}$$

where $\theta_1' = d(c+d)^{-1}$, $\theta_2' = a(a+b)^{-1}$, $\delta_1 = \theta_1' - \theta_2'$ and $(a+b) = (c+d)$.

Now, we assume $\left|\{\theta_1' e_1' + \theta_1 (e_2' + e_1 + e_1 e_2')\}\right| < 1$, so that $\left[1+\{\theta_1' e_1' + \theta_1 (e_2' + e_1 + e_1 e_2')\}\right]^{-1}$ is expandable. Expanding the above expression and then neglecting the terms having powers of e_0, e_1, e_1', e_2' greater than two, we get

$$[T_3^{***} - \bar{Y}] = \bar{Y}[e_0 + \delta_1 e_1' + \delta e_1 + \delta e_2' + \delta_1 \theta_1' e_1'^2 + \delta\theta_1 e_1^2 + \delta\theta_1 e_2'^2 - e_0 e_1 \theta_1 - e_0 e_2' \theta_1 - e_0 e_1' \theta_1'$$

$$+ \delta e_1 e_2' (2\theta_1 - 1) + e_1 e_1' (\delta_1 \theta_1 + \delta\theta_1') + e_2' e_1' (\delta_1 \theta_1 + \delta\theta_1')] \qquad (16)$$

Taking expectation on both the sides of equation (16), we get the bias of T_3^{***} up to the first order of approximation as

$$Bias(T_3^{***}) =$$

$$\bar{Y}\left[\begin{array}{l}\dfrac{S_X^2}{\bar{X}^2}\left\{\left(\dfrac{1}{n'}-\dfrac{1}{N}\right)(\delta\theta_1' + \delta_1\theta_1 + \delta_1\theta_1') + \left(\dfrac{1}{n}-\dfrac{1}{N}\right)\delta\theta_1\right\} + \dfrac{S_Z^2}{\bar{Z}^2}\left(\dfrac{1}{n'}-\dfrac{1}{N}\right)\delta\theta_1 - \dfrac{\rho_{XY}S_X S_Y}{\bar{X}\bar{Y}}\left\{\left(\dfrac{1}{n}-\dfrac{1}{N}\right)\theta_1 + \left(\dfrac{1}{n'}-\dfrac{1}{N}\right)\theta_1'\right\} - \\ \dfrac{\rho_{YZ}S_Y S_Z \theta_1}{\bar{Y}\bar{Z}}\left(\dfrac{1}{n'}-\dfrac{1}{N}\right) + \left(\dfrac{1}{n'}-\dfrac{1}{N}\right)\dfrac{\rho_{XZ}S_X S_Z}{\bar{X}\bar{Z}}\left(\delta(2\theta_1 - 1) + \delta_1\theta_1 + \delta\theta_1'\right)\end{array}\right]. \qquad (17)$$

Squaring both the sides of equation (16) and then taking expectation on ignoring the terms having powers of e_0, e_1, e_1', e_2' higher than two, we obtain the MSE of T_3^{***} up to the first order of approximation as

$$MSE(T_3^{***}) = \left(\frac{1}{n}-\frac{1}{N}\right)\left[S_Y^2 + \delta^2 R^2 S_X^2 - 2\delta\rho R S_X S_Y\right] + \frac{(k-1)}{n} W_2 S_{Y2}^2$$

$$\left(\frac{1}{n'}-\frac{1}{N}\right)\left[R^2 \delta_1 S_X^2 (2\delta + \delta_1) + R_1^2 \delta^2 S_Z^2 - 2R\delta_1 \rho_{XY} S_X S_Y - 2R_1 \delta\rho_{YZ} S_Y S_Z + 2\delta R R_1 \rho_{XZ} S_X S_Z (\delta + \delta_1)\right]$$

(18)

where $R_1 = \dfrac{\bar{Y}}{\bar{Z}}$.

Remark 3: The proposed family T_3^{***} can generate several well known existing estimators and other new estimators under non-response for the suitable choices of a, b, c and d. As we have done in Table 2.1, one can discover some well known existing estimators $\bar{y}^*$, $\bar{y}_R^{***}$, $\bar{y}_P^{***}$ and some new estimators $\bar{y}_1^{**}$, $\bar{y}_2^{**}$, $\bar{y}_3^{**}$, $\bar{y}_4^{**}$ for the suitable choices of a, b, c and d.To understand the behavior of these estimators, an empirical illustration has been given in Table 4.

3. Empirical Study

To realize the applications of the results obtained in this paper and to observe the performance of the estimators of the proposed families, it is imperative to illustrate whatever discussed in the previous sections with some numerical data. For this purpose, we have considered two different features of data analysis: (i) analysis through real data and (ii) analysis through simulation viewpoint.

3.1 Real Data Set

We have considered the data illustrated by Cochran (1977). The details are given below:

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 =34, n' =15, n =10, $\bar{Y}$ = 4.92, $\bar{X}$ = 2.59, $\bar{Z}$ =2.91, C_Y =1.0123, C_X =1.2318, C_Z =1.0720, ρ_{XY} =0.7326, ρ_{YZ} =0.6430, ρ_{XZ} =0.6837, S_y^2 =24.8055396, S_x^2 =10.17840969, S_z^2= 9.73140503, $S_{Y2}^2 = \frac{4}{5} S_Y^2$.

Tables 2, 3 and 4 represent the variance/MSE of the proposed families T_1^*, T_2^{**} and T_3^{***} respectively for the different choices of a, b, c, d and k at non-response rate W_2 =0.2.

Table 2: Variance/MSE of T_1^* for Different Choices of a, b, c, d and k

Estimators	k			$a=c$	$b=d$
	2	3	4		
$\bar{y}^*$	2.1478(100)	2.5447(100)	2.9416(100)	1	1
$\bar{y}_R^*$	1.6187(132.69)	2.0155(126.25)	2.4124(121.93)	1	0
$\bar{y}_P^*$	7.8623(27.31)	8.2592(30.81)	8.6561(33.98)	0	1
$\bar{y}_1^*$	1.3744(156.27)	1.7713(143.66)	2.1681(135.67)	C_X^2	ρ_{XY}
$\bar{y}_2^*$	1.2164(176.56)	1.6133(157.73)	2.0102(146.33)	C_X	M^2
$\bar{y}_3^*$	1.4573(147.37)	1.8542(137.23)	2.2516(130.67)	ρ_{XY}^2	f
$\bar{y}_4^*$	1.2515(171.62)	1.6483(154.37)	2.0452(143.82)	C_Y	M^2

Table 3: Variance/MSE of T_2^{**} for Different Choices of a, b, c, d and k

Estimators	k			$a=c$	$b=d$
	2	3	4		
$\bar{y}^*$	2.1478(100)	2.5447(100)	2.9416(100)	1	1
$\bar{y}_R^{**}$	1.8979(113.16)	2.2948(110.88)	2.6917(109.28)	1	0
$\bar{y}_P^{**}$	4.8463(44.31)	5.2432(48.53)	5.6401(52.15)	0	1
$\bar{y}_1^{**}$	1.7826(120.48)	2.1795(116.75)	2.5764(114.17)	C_X^2	ρ_{XY}
$\bar{y}_2^{**}$	1.7080(125.75)	2.1043(120.89)	2.5018(117.58)	C_X	M^2
$\bar{y}_3^{**}$	1.8218(117.89)	2.2186(114.69)	2.6155(112.46)	ρ_{XY}^2	f
$\bar{y}_4^{**}$	1.7245(124.54)	2.1214(119.95)	2.5183(116.80)	C_Y	M^2

Table 4: Variance/MSE of T_3^{***} for Different Choices of a, b, c, d and k

Estimators	k			$a=c$	$b=d$
	2	3	4		
$\bar{y}^*$	2.1478(100)	2.5447(100)	2.9416(100)	1	1
$\bar{y}_R^{***}$	1.6758(128.16)	2.0726(122.77)	2.4695(119.11)	1	0
$\bar{y}_P^{***}$	7.1412(30.07)	7.5381(33.75)	7.9350(37.07)	0	1
$\bar{y}_1^{***}$	1.4697(146.13)	1.8666(136.32)	2.2635(129.95)	C_X^2	ρ_{XY}
$\bar{y}_2^{***}$	1.3299(161.50)	1.7268(147.36)	2.1237(138.51)	C_X	M^2
$\bar{y}_3^{***}$	1.5426(139.22)	1.9395(131.20)	2.3364(125.90)	ρ_{XY}^2	f
$\bar{y}_4^{***}$	1.3612(157.78)	1.7581(144.74)	2.1550(136.50)	C_Y	M^2

[Parentheses figures show the percentage relative efficiency (PRE) of the proposed families with respect to the usual mean estimator $\bar{y}^*$]

3.2 Simulation Viewpoint

Let X and X_1 be the random variables from a normal population *i. e.* $X \sim N(5,3)$ and $X_1 \sim N(5,3)$. The variable X has been considered as the auxiliary variable. Now we define the study variable Y and additional auxiliary variable Z using the transformations $Y = \rho_{XY} X + \sqrt{1-\rho_{XY}^2}\, X_1$ and $Z = \rho_{XZ} X + \sqrt{1-\rho_{XZ}^2}\, X_1$ where $\rho_{XY} = 0.70$, $\rho_{XZ} = 0.65$ and $\rho_{YZ} = 0.40$. Further, we have considered 5000 values for each of the variables Y, X and Z by generating the values of X and X_1 using R software. Now, a first phase sample of $n' = 2000$ units has been selected from the population of $N = 5000$ units using SRSWOR scheme and then a second phase sample of $n = 200(200)600$ units has been selected from the first phase sample using SRSWOR scheme. In the next step, 20% units (considered as non-responding units) of the second phase sample were randomly dropped down and then a sub-sample of 50% units was selected from the non-responding units. At this stage, we have computed the estimates of $\bar{Y}$ using the proposed families of estimators. The procedure of selecting the second phase sample of size n, dropping down the non-responding units, selecting the sub-sample of non-responding units and computing the estimates of $\bar{Y}$ was repeated 500 times. Finally, we have computed the approximate MSE (AMSE) of the proposed families T_1^*, T_2^{**} and T_3^{***} using the following formulae:

$$AMSE(T_1^*) = \frac{1}{500}\sum_{\alpha=1}^{500}\left(T_{1\alpha}^* - \bar{Y}\right)^2$$

$$AMSE(T_2^{**}) = \frac{1}{500}\sum_{\alpha=1}^{500}\left(T_{2\alpha}^{**} - \bar{Y}\right)^2$$

$$AMSE(T_3^{***}) = \frac{1}{500}\sum_{\alpha=1}^{500}\left(T_{3\alpha}^{***} - \bar{Y}\right)^2;\ \alpha = 1,2,...,500$$

Tables 5, 6 and 7 depict the a AMSE of the families T_1^*, T_2^{**} and T_3^{***} respectively for the different choices of a, b, c, d and k.

Table 5: AMSE of the Family T_1^*

Estimators	n			$a = c$	$b = d$
	200	400	600		
$\bar{y}^*$	0.0001740	0.0000389	0.0000164	1	1
$\bar{y}_R^*$	0.0001042	0.0000245	0.0000102	1	0
$\bar{y}_P^*$	0.0008570	0.0001980	0.0000805	0	1
$\bar{y}_1^*$	0.0000824	0.0000177	0.0000078	C_X^2	ρ_{XY}
$\bar{y}_2^*$	0.0000588	0.0000127	0.0000057	C_X	M^2
$\bar{y}_3^*$	0.0000823	0.0000383	0.0000245	ρ_{XY}^2	f
$\bar{y}_4^*$	0.0000655	0.0000140	0.0000063	C_Y	M^2

Table 6: AMSE of the Family T_2^{**}

Estimators	n			$a=c$	$b=d$
	200	400	600		
$\bar{y}^*$	0.0002030	0.0000491	0.0000266	1	1
$\bar{y}_R^{**}$	0.0001950	0.0000455	0.0000245	1	0
$\bar{y}_P^{**}$	0.0008980	0.0002082	0.0001115	0	1
$\bar{y}_1^{**}$	0.0001540	0.0000365	0.0000197	C_X^2	ρ_{XY}
$\bar{y}_2^{**}$	0.0001200	0.0000293	0.0000161	C_X	M^2
$\bar{y}_3^{**}$	0.0001860	0.0000449	0.0000242	ρ_{XY}^2	f
$\bar{y}_4^{**}$	0.0001380	0.0000337	0.0000186	C_Y	M^2

Table 7: AMSE of the Family T_3^{***}

Estimators	n			$a=c$	$b=d$
	200	400	600		
$\bar{y}^*$	0.0002620	0.0000465	0.0000311	1	1
$\bar{y}_R^{***}$	0.0002430	0.0000444	0.0000305	1	0
$\bar{y}_P^{***}$	0.0011180	0.0001984	0.0001314	0	1
$\bar{y}_1^{***}$	0.0002400	0.0000423	0.0000273	C_X^2	ρ_{XY}
$\bar{y}_2^{***}$	0.0001580	0.0000279	0.0000190	C_X	M^2
$\bar{y}_3^{***}$	0.0002450	0.0000432	0.0000289	ρ_{XY}^2	f
$\bar{y}_4^{***}$	0.0001810	0.0000322	0.0000217	C_Y	M^2

4. Concluding Remarks

We have suggested some families of estimators of population mean in simple random sampling utilizing the information on auxiliary variable(s) under non-response. The expressions for the biases and mean square errors of the proposed families have been derived up to the first order of approximation. Some well-known existing estimators and some new members of the proposed families have been pioneered out. The theoretical results have been illustrated through both real data set and simulation analysis. From Tables 2, 3, 4, 5, 6 and 7, it is revealed that the estimators $\bar{y}_2^*$, $\bar{y}_2^{**}$ and $\bar{y}_2^{***}$ would be the appropriate choices of the researchers as the considerable gains in efficiency have been achieved by them. From Tables 2, 3 and 4, it is also revealed that the variance/MSE of the estimators increases with increase in sub sampling rate k. The results are intuitively expected.

Note: There are no relevant financial or non-financial competing interests to report.

Acknowledgement: This work is supported by Incentive Grant, IoE, Banaras Hindu University.

References

[1] Andersson, P. G. and Särndal, C. E., 2016, Calibration for non response treatment: In one or two steps, *Statistical Journal of the IAOS*, *32*(3), 375-381.

[2] Chaudhary, M. K., Abdelwahab, M. M., Bhardwaj, N. and Hasaballah, M. M. , 2026, Estimation through calibration under stratified sampling with non-response and measurement error effects, *Mathematics, MDPI*, *14*(3), 439, 1-11.

[3] Chaudhary, M. K. and Kumar, A., 2016, Estimation of mean of finite population using double sampling scheme under non-response, *Journal of Statistics Applications and Probability*, *5*(2), 287-297.

[4] Chaudhary, M. K., Kumar, A. and Vishwakarma, G. K., 2021, Some improved estimators of population mean using two-phase sampling scheme in the presence of non-response, *Pakistan Journal of Statistics and Operation research*, *17*(4), 911-919.

[5] Chaudhary, M. K. and Ray, B. K., 2024, Treating the problem of non-response in stratified random sampling under calibration approach, *Communications in Statistics-Simulation and Computation, 54*(10), 4472-4480.

[6] Chaudhary, M. K., Ray, B. K., Vishwakarma, G. K. and Kadilar, C., 2024, A calibration-based approach on estimation of mean of a stratified population in the presence of non-response, *Communications in Statistics-Theory and Methods*, *53*(19), 7054-7068.

[7] Chaudhary, M. K., Singh, R., Shukla, R. K., Kumar, M. and Smarandache, F., 2009, A family of estimators for estimating population mean in stratified sampling under non-response, *Pakistan Journal of Statistics and Operation Research, 5*(1), 47-54.

[8] Chaudhary, M. K., Singh, V. K., Singh, R. and Smarandache, F., 2011, On some new allocation schemes in stratified random sampling under non-response, *Studies in Sampling Techniques and Time series*, Zip Publishing 1313 Chesapeake Avenue Columbus, Ohio 43212, USA, 40-54.

[9] Cochran, W. G., 1963, *Sampling Techniques*, Wiley Eastern Limited, New Delhi II Edition.

[10] Cochran, W. G., 1977, *Sampling Techniques*, Wiley Eastern Limited, New Delhi III Edition.

[11] Hansen, M. H. and Hurwitz, W. N., 1946, The problem of non-response in sample surveys, *Journal of the American Statistical Association*, *41*, 517-529.

[12] Kish, L., 1965, *Survey Sampling*, Wiley and Sons, New York.

[13] Khare, B. B., 1987, Allocation In stratified sampling in presence of non-response, *Metron*, *45*(I/II), 213-221.

[14] Khare, B. B., 1992, Non-response and the estimation of population mean in its presence in sample surveys, *Seminar Proceedings, 12th Orientation Course (Science Stream)*, Academic staff College, Banaras Hindu University, Varanasi, India.

[15] Khare, B. B. and Sinha, R. R., 2004, Estimation of finite population ratio using two-phase sampling scheme in the presence of non-response, *Aligarh J. Stat*., 24, 43-56.

[16] Kumar, S. and Bhougal, S., 2011, Estimation of the population mean in presence of non-response, *Communications for Statistical Applications and Methods*, *18*(4), 537-548.

[17] Okafor, F. C. and Lee, H., 2000, Double sampling for ratio and regression estimation with sub-sampling the non-respondents, *Survey Methodology*, 26, 183-188.

[18] Rao, J. N. K., 1973, On double sampling for stratification and analytical surveys, *Biometrika*, *60*(1), 125-133.

[19] Rao, P. S. R. S., 1986, Ratio estimators with sub-sampling the non-respondents, *Survey Methodology*, *12*(2), 217-230.

[20] Rao, P. S. R. S., 1987, Ratio and regression estimators with sub-sampling the non-respondents, Paper presented at a special contributed session of the International Association Meetings, September 2-16, Tokyo, Japan.

[21] Rao, P. S. R. S., 1990, Regression estimators with sub-sampling of non-respondents, *Data Quality Control : Theory and Pragmatics*, Liepins, E. G. and Uppuluri, V. R. R. (eds.), Marcel Dekker, New York, 191-208.

[22] Sarndal, C. E., Swensson, B. and Wretman, J., 1992, *Model Assisted Survey Sampling*, Springer-Verlag, New York, Inc.

[23] Singh, H. P. and Nigam, P., 2022, Estimation of population mean in stratified sampling under non-response, *Revista Investigación Operacional*, *43*(5), 574-590.

[24] Singh, H. P., Pal, S. K. and Solanki, R. S., 2017, A new class of estimators of finite population mean in sample surveys, *Communications in Statistics-Theory and Methods*, *46*(6), 2630-2637.

[25] Singh, R. and Mishra, M., 2020, Estimation in stratified random sampling in the presence of errors, *Revista Investigación Operacional, 41*(1), 125-137.

[26] Singh, R., Kumar, M., Chaudhary, M. K. and Smarandache, F., 2010, Estimation of mean in presence of non-response using exponential estimator, *Multispace & Multistructure Neutrosophic Transdisciplinarity*, North-European Scientific Publishers Hanko, Finland, Vol. IV, 758-768

[27] Srivastava, S., 1993, Some problems on the estimation of population mean using auxiliary character in presence of non-response in sample surveys, *Ph. D. Thesis, Banaras Hindu University, Varanasi.*

[28] Tabasum, R., and Khan, I. A., 2006, Double sampling ratio estimator for the population mean in presence of non-response, *Assam Statistical Review*, *20*, 73-83.

[29] Wani, Z. H. and Rizvi, S. E. H., 2023, Assessing the effect of non-response in stratified random sampling using enhanced ratio type estimators under double sampling strategy, *Pakistan Journal of Statistics and Operation research*, *19*(3), 459-473.