AN IMPROVED METHOD TO HANDLE NON-IGNORABLE TWO-PHASE MISSING DATA IN THE ESTIMATION OF MEAN: SIMULATION AND EMPIRICAL ANALYSIS UNDER OBSERVED HETEROGENEITY

R. R. Sinha[1] and Anjali Gupta
Dr B R Ambedkar National Institute of Technology, India.

ABSTRACT
Missing data in decision theory significantly impacts real-world problems, distorting results and potentially leading to biased or incorrect decisions. Imputation and deletion of significant survey responses may impair the reliability and validity of the results. In order to address non-ignorable missing data, this article suggests new, improved exponential type estimators for estimating the study variable's mean using an auxiliary variable that shows non-response during two phases. The study examines the strength of estimators using mathematical expressions for bias and MSE under stratified two-phase sampling. The theoretical constraints have been given to strengthen the performance of the proposed estimators. To mechanize the efficiency of the proposed estimators, a numerical analysis on the simulated data-sets (symmetric and asymmetric) and real data-sets has been carried out using statistical packages of R-software.

KEYWORDS: Bias, Mean, Mean square error, Sub-sampling, Simulation.

MSC: 62D05

RESUMEN
La ausencia de datos en la teoría de decisiones impacta de manera significativa en problemas del mundo real, distorsionando los resultados y pudiendo conducir a decisiones sesgadas o incorrectas. La imputación y eliminación de respuestas relevantes en encuestas puede afectar la fiabilidad y validez de los resultados. Para abordar los datos faltantes no ignorables, este artículo propone nuevos estimadores de tipo exponencial mejorados para calcular la media de la variable de estudio utilizando una variable auxiliar que presenta no respuesta en dos fases. El estudio examina la solidez de los estimadores mediante expresiones matemáticas de sesgo y error cuadrático medio bajo muestreo estratificado en dos fases. Se han establecido restricciones teóricas para reforzar el desempeño de los estimadores propuestos. Para operacionalizar la eficiencia de los estimadores, se realizó un análisis numérico sobre conjuntos de datos simulados (simétricos y asimétricos) y reales, empleando paquetes estadísticos del software R.

PALABRAS CLAVE: Sesgo, Media, Error cuadrático medio, Submuestreo, Simulación.

1. INTRODUCTION

1.1. Motivation and Literature Review

In recent days internet-based platforms such as- online surveys, web-based questionnaire, email surveys, web-scraping, crowdsourcing etc. are being used extensively to gather information regarding variables under study, due to widespread availability of internet and the advantages it offers in terms of cost effectiveness, time effectiveness, real- time collection, and accessibility. Instead of having several advantages, internet-based surveys may suffer from non-response. Non-response occurs due to attrition of the survey unit, or may be survey unit fails to respond the survey invitation. Various fields such as – academics, health care, market research, public policy, political polling, behavioural sciences, non-profit and social services etc. use online survey for data collection. For instance, in health sector to analyse the preferences, perspectives, experiences, and many other important facets of medical assets, online survey to the medical practitioners and interns may be beneficial. Loss of information due to non-response or missing data can potentially affect the precision, power and generalizability of the results as respondent units may differ from non-respondent units. This necessitates the study of the effect of non-response in the estimation of mean of the population under study and its remedies. And so, the primary focus of this article is to examine the impact of total non-response of units selected in the sample to estimate the population mean of study variable. To address non-response, Hansen and Hurwitz [8] developed an unbiased estimator for the population mean by using additional effort to collect data on a sub-sample of non-responding units.

Let us consider $\mathcal{S}_m$ is a sample of size m drawn from the population U_N (of N units with survey variables Y, X, Z) using simple random sampling without replacement ($SiRS_{(wor)}$). Hansen and Hurwitz [8] considered that the population is made of two mutually exclusive groups, respondent (of size $N_{(1)}$) and non-respondent (of size $N_{(2)}$) such that $N_{(1)} + N_{(2)} = N$. Further from m sampled units, $m_{(1)}$ and $m_{(2)}$ denote the size of the respondent units and non-respondent units respectively i.e. $m_{(1)} + m_{(2)} = m$. Again, r $\left(r = \frac{m_{(2)}}{k}\ ; k > 1\right)$ is the size of $SiRS_{(wor)}$

[1] raghawraman@gmail.com

sub-sample drawn from the $m_{(2)}$ non-respondents for personal interview in order to obtain information on the goal of interest. Hence, based on $m_{(1)} + r$ units, Hansen and Hurwitz [8] defined an unbiased estimator $(\bar{y}^*_{HH})$ for population mean of the study variable as:

$$\bar{y}^*_{HH} = \left(m_{(1)}/m\right)\bar{y}_{m_{(1)}} + \left(m_{(2)}/m\right)\bar{y}_r \qquad (1.1)$$

where, $\bar{y}_{m_{(1)}}$ and $\bar{y}_r$ are sample means based on responding units $(m_{(1)})$ and sub-sampled units (r) and variance of $\bar{y}^*_{HH}$ is given by,

$$var(\bar{y}^*_{HH}) = \lambda S^2_Y + \theta W_2 S^2_{Y(2)} \qquad (1.2)$$

where, $\lambda = \frac{1}{m} - \frac{1}{N}$, $\theta = \frac{k-1}{m}$, $W_2 = \frac{N_{(2)}}{N}$, S^2_Y and $S^2_{Y(2)}$ are population variance of the study variable based on respondent and non-respondent group. Subsequently, a large number of researchers [1], [3], [4], [5], [6], [7], [11], [12], [15], [16], [17], [18], [19], [20], [22], [23], [24], and [25] have carried out commendable work to address the non-response issue by introducing better estimators of the population parameters. But, when the population mean of auxiliary variable is unknown ahead of time, the majority of authors prefer to use two phase sampling to estimate it, taking into account that the auxiliary variable is free from non-response during the first phase and suffers from non-response during the second phase. Chaudhary and Kumar [4] took into account the non-response during both phases and suggested conventional ratio, product, and regression estimators to estimate the mean of study variable (Y) when population mean of an auxiliary variable (X) is unknown, as

$$T_R = \frac{\bar{y}^*_{HH}}{\bar{x}^*_{HH}}\bar{x}^{*\prime}_{HH}, \qquad (1.3)$$

$$T_P = \frac{\bar{y}^*_{HH}}{\bar{x}^{*\prime}_{HH}}\bar{x}^*_{HH}, \qquad (1.4)$$

$$T_{Reg} = \bar{y}^*_{HH} + b^*(\bar{x}^{*\prime}_{HH} - \bar{x}^*_{HH}); \qquad (1.5)$$

where $\bar{x}^{*\prime}_{HH}, \bar{x}^*_{HH}$ are the Hansen Hurwitz estimators based on first and second phase samples respectively.

1.2. Notations and Methodology

On the basis of observance, heterogeneity is classified in to two categories: Observed heterogeneity (presence of some factors that can be measure directly such as : age, gender, education, income, size of the firm, market share, supply and demand, blood pressure, cholesterol level and body mass index (BMI), etc.) and Unobserved heterogeneity (presence of the factors that cannot be included in the analysis directly such as: unobservable genetic factors, variations in immune response, metabolic process, mental process and cognitive factors, etc.). The presence of heterogeneity in the study population, as it can affect the variability of the data and so the representativeness of the selected sample, is necessary to take in to account. The purpose of the article that is being presented is to address the problem of non-response in a heterogeneous population that has been observed. Specifically, we have taken into consideration study and auxiliary variables: Y and X, both of which have unit non-response while another variable, Z, does not (see Rao [13]). Now we define two different situations for variables X and Z as

Situation-I When auxiliary mean $\bar{X}$ is unknown but $\bar{Z}$ is known.

Situation-II When auxiliary mean $\bar{X}$ and $\bar{Z}$ both are unknown.

To elucidate the methodology, let us consider population U_N can be divided into $L, (h = 1,2,3, \ldots, L)$ exhaustive and mutually heterogeneous strata, where h^{th} stratum consist of N_h units such that $\sum_{h=1}^{L} N_h = N$. Following Hansen and Hurwitz [8] it is assumed that h^{th} stratum is consist of two mutually exclusive groups, respondents (R) and non-respondents (NR) with $N_{h(1)}$ and $N_{h(2)}$ units respectively. For defined situations, here we have used two phase sampling scheme to obtain the unknown auxiliary mean. Let $\mathcal{S}'_{n_h}$ and $\mathcal{S}_{n_h}$are first phase and second phase simple random samples $(SiRS_{(wor)})$ of size n'_h and n_h respectively drawn from h^{th} stratum. Additionally, $(Y_i, X_i$ and $Z_i)$ indicates observations made on the i^{th} population unit of the h^{th} stratum for the survey variables (Y, X, Z) respectively. Detailed sampling methodology required for these situations can be understand from the figure 1

Here, $k' > 1$, $k > 1$ and $n_h < n'_h$

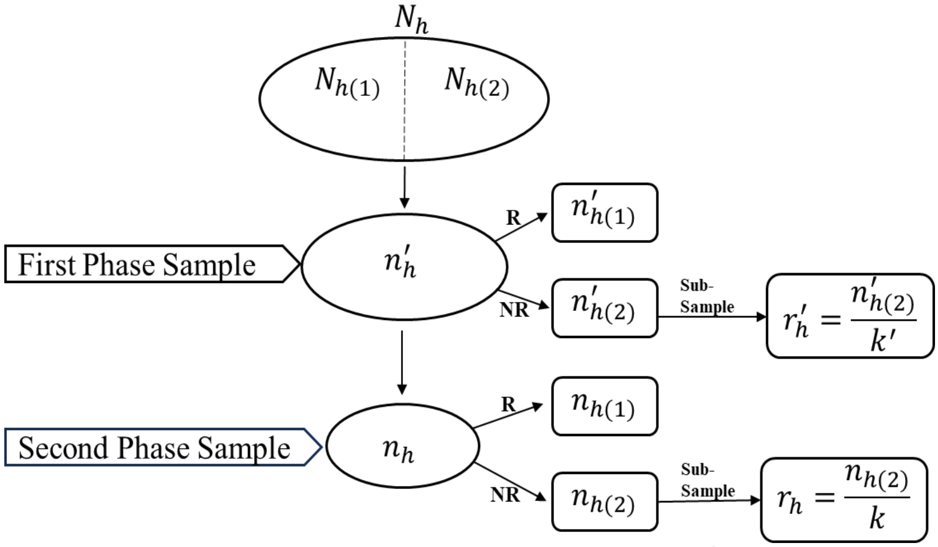

Figure 1: Execution of Two-Phase Sampling Technique for h$^{\text{th}}$ Stratum

To determine the variance and covariance of the variables under consideration, let us take into consideration the large sample approximations as:

$$e_0 = \frac{\bar{y}^*_{st}}{\bar{Y}} - 1, \qquad e_1 = \frac{\bar{x}^*_{st}}{\bar{X}} - 1, \qquad e_2 = \frac{\bar{z}_{st}}{\bar{Z}} - 1, \qquad e'_1 = \frac{\bar{x}^{*\prime}_{st}}{\bar{X}} - 1, \qquad e'_2 = \frac{\bar{z}'_{st}}{\bar{Z}} - 1.$$

Such that, $E(e_i) = 0; \forall i = 0,1,2, \quad E(e'_i) = 0, \forall i = 1,2$

$$\omega_0 = E(e_0^2) = \textstyle\sum_{h=1}^{L}\left(\lambda_h \tau_{Y_h} + \theta_h \tau_{Y_{h(2)}}\right), \omega_1 = E(e_1^2) = \sum_{h=1}^{L}\left(\lambda_h \tau_{X_h} + \theta_h \tau_{X_{h(2)}}\right), \omega_2 = E({e_2}^2) = \sum_{h=1}^{L}(\lambda_h \tau_{Z_h}),$$

$$\omega'_2 = E\left({e'_2}^2\right) = \textstyle\sum_{h=1}^{L}(\lambda'_h \tau_{Z_h}), \quad \omega'_1 = E\left({e'_1}^2\right) = \sum_{h=1}^{L}\left(\lambda'_h \tau_{X_h} + \theta'_h \tau_{X_{h(2)}}\right), \quad \omega_{02} = E(e_0 e_2) = \sum_{h=1}^{L}\left(\lambda_h \tau_{YZ_h}\right),$$

$$\omega'_{11} = E(e_1 e'_1) = \textstyle\sum_{h=1}^{L}\left(\lambda'_h \tau_{X_h} + \theta'_h \tau_{X_{h(2)}}\right), \quad \omega'_{02} = E(e_0 e'_2) = \sum_{h=1}^{L}\left(\lambda'_h \tau_{YZ_h}\right), \quad \omega_{12} = E(e_1 e_2) = \sum_{h=1}^{L}\left(\lambda_h \tau_{XZ_h}\right),$$

$$\omega'_{22} = E(e_2 e'_2) = \textstyle\sum_{h=1}^{L}(\lambda'_h \tau_{Z_h}), \quad \omega'_{12} = E(e'_1 e_2) = E(e_1 e'_2) = E(e'_1 e'_2) = \sum_{h=1}^{L}\left(\lambda'_h \tau_{XZ_h}\right),$$

$$\omega_{01} = E(e_0 e_1) = \textstyle\sum_{h=1}^{L}\left(\lambda_h \tau_{YX_h} + \theta_h \tau_{YX_{h(2)}}\right), \quad \omega'_{01} = E(e_0 e'_1) = \sum_{h=1}^{L}\left(\lambda'_h \tau_{YX_h} + \theta'_h \tau_{YX_{h(2)}}\right).$$

Here, we define a few terms for the variable V, which acknowledges the responding units Y, X, and Z. The non-responding units, which are only taken into account when looking at the variables Y and X, are denoted by the subscript (2).

$$\tau_{V_h} = P_h^2 S_{V_h}^2 / \bar{V}^2 \quad (for\ \ V = Y, X, Z)\ , \qquad \tau_{V_{h(2)}} = P_h^2 W_{h(2)} S_{V_{h(2)}}^2 / \bar{V}^2 \qquad (for\ V = Y, X),$$

$$\tau_{VV'_h} = P_h^2\ \rho_{V_h V'_h} S_{V_h} S_{V'_h} / \bar{V}\bar{V}' \quad (for\ V \neq V' = Y, X, Z), \qquad \tau_{YX_{h(2)}} = P_h^2 W_{h(2)}\ \rho_{Y_h X_{h(2)}} S_{Y_{h(2)}} S_{X_{h(2)}} / \bar{Y}\bar{X},$$

$$\lambda_h = \frac{1}{n_h} - \frac{1}{N_h}, \qquad \lambda'_h = \frac{1}{n'_h} - \frac{1}{N_h}, \qquad r_h = \frac{n_{h(2)}}{k}, \qquad r'_h = \frac{n'_{h(2)}}{k'}, \quad \theta_h = \frac{k-1}{n_h}, \theta'_h = \frac{k'-1}{n'_h}, \quad W_{h(2)} = \frac{N_{h(2)}}{N_h},$$

$$P_h = \frac{N_h}{N}, \qquad \varPi_h = \theta_h - \theta'_h, \quad \triangle_h = \lambda_h - \lambda'_h, \quad \bar{y}^*_{st} = \textstyle\sum_{h=1}^{L} P_h\, \bar{y}^*_h, \qquad \bar{x}^*_{st} = \sum_{h=1}^{L} P_h\, \bar{x}^*_h,$$

$$\bar{x}^{*\prime}_{st} = \textstyle\sum_{h=1}^{L} P_h\, \bar{x}^{*\prime}_h, \qquad \bar{z}_{st} = \sum_{h=1}^{L} P_h \bar{z}_h, \qquad \bar{z}'_{st} = \sum_{h=1}^{L} P_h \bar{z}'_h, \qquad \bar{z}'_h = \frac{1}{n'_h}\sum_{i=1}^{n'_h} z_i,$$

$$\bar{y}^*_h = \frac{n_{h(1)}\bar{y}^*_{h(1)} + n_{h(2)}\bar{y}^*_{h(2)}}{n_h}, \qquad \bar{x}^*_h = \frac{n_{h(1)}\bar{x}^*_{h(1)} + n_{h(2)}\bar{x}^*_{h(2)}}{n_h}, \quad \bar{x}^{*\prime}_h = \frac{n'_{h(1)}\bar{x}^{*\prime}_{h(1)} + n'_{h(2)}\bar{x}^{*\prime}_{h(2)}}{n'_h}, \quad \bar{z}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} z_i,$$

$$\bar{y}^*_{h(1)} = \frac{1}{n_{h(1)}}\sum_{i=1}^{n_{h(1)}} y_i, \qquad \bar{y}^*_{h(2)} = \frac{1}{r_h}\sum_{i=1}^{r_h} y_i, \qquad \bar{x}^*_{h(1)} = \frac{1}{n_{h(1)}}\sum_{i=1}^{n_{h(1)}} x_i, \qquad \bar{x}^*_{h(2)} = \frac{1}{r_h}\sum_{i=1}^{r_h} x_i$$

$$\bar{x}^{*\prime}_{h(1)} = \frac{1}{n'_{h(1)}}\sum_{i=1}^{n'_{h(1)}} x_i, \qquad \bar{x}^{*\prime}_{h(2)} = \frac{1}{r'_h}\sum_{i=1}^{r'_h} x_i, \qquad S_{V_h}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(V_i - \bar{V}_h)^2 \quad (for\ \ V = Y, X, Z),$$

$$S_{V_{h(2)}}^2 = \frac{1}{N_{h(2)}-1}\sum_{i=1}^{N_{h(2)}}(V_i - \bar{V}_{h(2)})^2\ (for\ V = Y, X)\ , \quad \rho_{V_h V'_h} = \frac{S_{V_h V'_h}}{S_{V_h} S_{V'_h}} \qquad (for\ V \neq V' = Y, X, Z),$$

$$\rho_{V_h V'_{h(2)}} = \frac{S_{V_h V'_{h(2)}}}{S_{V_{h(2)}} S_{V'_{h(2)}}} \quad (for\ V \neq V' = Y, X),\ S_{V_h V'_h} = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(V_{h_i} - \bar{V}_h)(V'_{h_i} - \bar{V}'_h) \ \ (for\ V \neq V' = Y, X, Z),$$

$$S_{V_h V'_{h(2)}} = \frac{1}{N_{h(2)}-1}\sum_{i=1}^{N_{h(2)}}(V_{h_i} - \bar{V}_{h(2)})(V'_{h_i} - \bar{V}'_{h(2)}) \quad (for\ V \neq V' = Y, X), \qquad \bar{V}_h = \frac{\sum_{i=1}^{N_h} V_i}{N_h} \quad (for\ V = Y, X, Z).$$

2. ADOPTED ESTIMATORS

Due to the heterogeneity of the population under investigation, stratified simple random sampling was employed. Within this framework, we adopted several well-known existing estimators, including those proposed by Chaudhary and Kumar [4], the conventional regression estimator (t_4), and an exponential estimator (t_5) motivated by the work of Kumar and Bhougal [10] in the context of stratified sampling. These estimators were considered to assess and validate the efficiency of the proposed improved exponential estimators, which are discussed in the following section.

The adopted estimators and their corresponding mean square errors (MSE) or minimum mean square errors $(M.MSE)$ are summarized in Table 1.

Estimators	$(\boldsymbol{MSE})$ / $(\boldsymbol{M.MSE})$
$t_1 = \frac{\bar{y}_{st}^{*}}{\bar{x}_{st}^{*}} \bar{x}_{st}^{*\prime}$ Chaudhary and Kumar [4]	$\sum_{h=1}^{L} P_h^2 \begin{pmatrix} \lambda_h' S_{Y_h}^2 + \triangle_h \left(S_{Y_h}^2 + R^2 S_{X_h}^2 - 2R\, \rho_{Y_h X_h} S_{Y_h} S_{X_h}\right) + \\ \theta_h W_{h(2)} S_{Y_h(2)}^2 + W_{h(2)} \left(R^2 S_{X_h(2)}^2 - 2R\, \rho_{Y_h X_h(2)} S_{Y_h(2)} S_{X_h(2)}\right) \varPi_h \end{pmatrix},$ where $R = \bar{Y} / \bar{X}$.
$t_2 = \frac{\bar{y}_{st}^{*}}{\bar{x}_{st}^{*\prime}} \bar{x}_{st}^{*}$ Chaudhary and Kumar [4]	$\sum_{h=1}^{L} P_h^2 \begin{pmatrix} \lambda_h' S_{Y_h}^2 + \triangle_h \left(S_{Y_h}^2 + R^2 S_{X_h}^2 + 2R\, \rho_{Y_h X_h} S_{Y_h} S_{X_h}\right) + \\ \theta_h W_{h(2)} S_{Y_h(2)}^2 + W_{h(2)} \left(R^2 S_{X_h(2)}^2 + 2R\, \rho_{Y_h X_h(2)} S_{Y_h(2)} S_{X_h(2)}\right) \varPi_h \end{pmatrix}.$
$t_3 = \bar{y}_{st}^{*} + b^{*}(\bar{x}_{st}^{*\prime} - \bar{x}_{st}^{*})$ where $b^{*} = \frac{s_{xy}^{*}}{s_x^{2*}}$ Chaudhary and Kumar [4]	$\sum_{h=1}^{L} P_h^2 \begin{pmatrix} \left(\lambda_h' + \triangle_h \left(1 - \rho_{Y_h X_h}^2\right)\right) S_{Y_h}^2 + \theta_h W_{h(2)} S_{Y_h(2)}^2 \\ + W_{h(2)} (\beta^2 S_{X_h(2)}^2 - 2\beta) \varPi_h \end{pmatrix}$, where, $\beta = \frac{s_{YX}}{s_X^2}$.
$t_4 = \bar{y}_{st}^{*} + \hat{\beta}^{*}(\bar{x}_{st}^{*\prime} - \bar{x}_{st}^{*})$; where $\hat{\beta}^{*}$ is estimate of regression coefficient.	$\left[\sum_{h=1}^{L} P_h^2 \left(\lambda_h S_{Y_h}^2 + \theta_h W_{h(2)} S_{Y_h(2)}^2\right)\right] - \frac{\left[\sum_{h=1}^{L} P_h^2 \left\{\triangle_h \rho_{Y_h X_h} S_{Y_h} S_{X_h} + \varPi_h W_{h(2)} \rho_{Y_h X_h(2)} S_{Y_h(2)} S_{X_h(2)}\right\}\right]^2}{\left[\sum_{h=1}^{L} P_h^2 \left\{\triangle_h S_{X_h}^2 + \varPi_h W_{h(2)} S_{X_h(2)}^2\right\}\right]}.$
$t_5 = \bar{y}_{st}^{*} exp \left[\alpha \left(\frac{\bar{x}_{st}^{*\prime} - \bar{x}_{st}^{*}}{\bar{x}_{st}^{*\prime} + \bar{x}_{st}^{*}}\right)\right]$; where α is an optimizing constant.	$\left[\sum_{h=1}^{L} P_h^2 \left(\lambda_h S_{Y_h}^2 + \theta_h W_{h(2)} S_{Y_h(2)}^2\right)\right] - \frac{\left[\sum_{h=1}^{L} P_h^2 \left\{\triangle_h \rho_{Y_h X_h} S_{Y_h} S_{X_h} + \varPi_h W_{h(2)} \rho_{Y_h X_h(2)} S_{Y_h(2)} S_{X_h(2)}\right\}\right]^2}{\left[\sum_{h=1}^{L} P_h^2 \left\{\triangle_h S_{X_h}^2 + \varPi_h W_{h(2)} S_{X_h(2)}^2\right\}\right]}.$

Table 1: Adopted Estimators and their (MSE) / $(M.MSE)$

3. PROPOSED ESTIMATORS: THEORETICAL PROPERTIES AND ADVANTAGES

Grasping motivation from Chaudhary and Kumar [4], this study introduces a set of novel improved exponential estimators aimed at estimating the population mean of the study variable. These estimators are designed to accommodate different conditions, particularly when the mean of the auxiliary variable is either known or unknown, as outlined in sub-section 1.2. The formulation is carried out under a stratified two-phase sampling framework and is presented as follows:

$$\mathcal{T}_{IE}^{(1)} = \bar{y}_{st}^{*} exp \left[\gamma_1 \left(\frac{\bar{x}_{st}^{*\prime} - \bar{x}_{st}^{*}}{\bar{x}_{st}^{*\prime} + \bar{x}_{st}^{*}}\right) + \gamma_2 \left(\frac{\bar{z}_{st} - \bar{Z}}{\bar{z}_{st} + \bar{Z}}\right)\right], \quad \text{(For Situation-I)} \tag{3.1}$$

and
$$\mathcal{T}_{IE}^{(2)} = \bar{y}_{st}^{*} exp \left[\mu_1 \left(\frac{\bar{x}_{st}^{*\prime} - \bar{x}_{st}^{*}}{\bar{x}_{st}^{*\prime} + \bar{x}_{st}^{*}}\right) + \mu_2 \left(\frac{\bar{z}_{st} - \bar{z}_{st}'}{\bar{z}_{st} + \bar{z}_{st}'}\right)\right], \text{(For Situation-II)} \tag{3.2}$$

where γ_1, γ_2, μ_1, μ_2 are optimizing constants to be used in determining the $M.MSE$ of the proposed estimators $\mathcal{T}_{IE}^{(1)}$ and $\mathcal{T}_{IE}^{(2)}$. We now state theorems to further clarify the traits of the proposed estimators based on approximation to the first degree [Appr $O(n^{-1})$]. Reddy [14] and Srivastava and Jhajj [21] suggested from a practical standpoint that the antecedent information from past data on the required parameters or their estimate can be used to obtain the values of unknown parameters. It is also possible to use sample observations, which won't have an impact on the performance of the estimator up to the first order of approximation.

The proposed estimators $\mathcal{T}_{IE}^{(1)}$ and $\mathcal{T}_{IE}^{(2)}$ exhibit important theoretical properties, which are rigorously supported by the following theorems.

Theorem 3.1- Bias and MSE of the proposed estimator $\mathcal{T}_{IE}^{(1)}$ are given by

$$Bias\left(\mathcal{T}_{IE}^{(1)}\right) = \bar{Y} \left[\gamma_1^2 \left\{\frac{(\omega_1 - \omega_{11}^{'})}{8}\right\} + \gamma_2^2 \frac{\omega_2}{8} + \gamma_1 \frac{2(\omega_{01}' - \omega_{01}) - (\omega_1' - \omega_1)}{4} + \gamma_2 \frac{(2\omega_{02} - \omega_2)}{4} + \gamma_1 \gamma_2 \frac{(\omega_{12}' - \omega_{12})}{8}\right] \tag{3.3}$$

$$MSE\left(\mathcal{T}_{IE}^{(1)}\right) = \bar{Y}^2 \left[\omega_0 + \gamma_1^2 \left(\omega_1 - \omega_{11}^{'}\right)/4 + \gamma_2^2\, \omega_2/4 + \gamma_1 \gamma_2 \left(\omega_{12}' - \omega_{12}\right)/2 + \gamma_1 (\omega_{01}' - \omega_{01}) + \gamma_2 \omega_{02}\right] \tag{3.4}$$

Proof: Under the above approximations given in sub-section1.2 the proposed estimator $\mathcal{T}_{IE}^{(1)}$ takes the following form as:

$$\mathcal{T}_{IE}^{(1)} = \bar{Y}\left[1 + e_0 + \gamma_1\left\{\frac{(e_1'-e_1)+(e_0e_1'-e_0e_1)}{2} - \frac{\left({e_1'}^2-{e_1}^2\right)}{4}\right\} + \gamma_2\left(\frac{e_2+e_0e_2}{2} - \frac{{e_2}^2}{4}\right) + \gamma_1^2\frac{\left(e_1'-e_1\right)^2}{8} + \gamma_2^2\frac{{e_2}^2}{8} + \gamma_1\gamma_2\frac{(e_1'e_2-e_1e_2)}{4}\right] \tag{3.5}$$

$$\mathcal{T}_{IE}^{(1)} - \bar{Y} = \bar{Y}\left[e_0 + \gamma_1\left\{\frac{(e_1'-e_1)+(e_0e_1'-e_0e_1)}{2} - \frac{\left({e_1'}^2-{e_1}^2\right)}{4}\right\} + \gamma_2\left(\frac{e_2+e_0e_2}{2} - \frac{{e_2}^2}{4}\right) + \gamma_1^2\frac{\left(e_1'-e_1\right)^2}{8} + \gamma_2^2\frac{{e_2}^2}{8} + \gamma_1\gamma_2\frac{(e_1'e_2-e_1e_2)}{4}\right] \tag{3.6}$$

On taking expectations on both side of the equation (3.6), and using the expected values we get the expression of $Bias$ as:

$$Bias\left(\mathcal{T}_{IE}^{(1)}\right) = \bar{Y}\left[\gamma_1^2\left\{\frac{(\omega_1-\omega_{11}')}{8}\right\} + \gamma_2^2\frac{\omega_2}{8} + \gamma_1\frac{2(\omega_{01}'-\omega_{01})-(\omega_1'-\omega_1)}{4} + \gamma_2\frac{(2\omega_{02}-\omega_2)}{4} + \gamma_1\gamma_2\frac{(\omega_{12}'-\omega_{12})}{8}\right]. \tag{3.7}$$

Now to study the characterization properties of proposed estimator, we get the expression for MSE of $\mathcal{T}_{IE}^{(1)}$ under situation-I by squaring and taking expectation of equation (3.6) as,

$$MSE\left(\mathcal{T}_{IE}^{(1)}\right) = \bar{Y}^2\left[\omega_0 + \gamma_1^2\left(\omega_1 - \omega_{11}'\right)/4 + \gamma_2^2\,\omega_2/4 + \gamma_1\gamma_2\left(\omega_{12}' - \omega_{12}\right)/2 + \gamma_1(\omega_{01}' - \omega_{01}) + \gamma_2\omega_{02}\right]. \tag{3.8}$$

Theorem 3.2- Minimum mean square error $(M.MSE)$ of the proposed estimator $\mathcal{T}_{IE}^{(1)}$ at the optimum value of $\gamma_1$and γ_2 is

$$M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) = \left[A_y - \frac{(B_{yx}^2A_z+B_{yz}^2A_x-2B_{yz}B_{xz}B_{yx})}{A_xA_z-B_{xz}^2}\right] \tag{3.9}$$

at, $\gamma_{1(opt)} = \frac{2[\omega_{02}(\omega_{12}'-\omega_{12})-\omega_2(\omega_{01}'-\omega_{01})]}{(\omega_1'+\omega_1-2\omega_{11}')\omega_2-(\omega_{12}'-\omega_{12})^2}$ and $\gamma_{2(opt)} = \frac{2[(\omega_{01}'-\omega_{01})(\omega_{12}'-\omega_{12})-\omega_{02}(\omega_1'+\omega_1-2\omega_{11}')]}{(\omega_1'+\omega_1-2\omega_{11}')\omega_2-(\omega_{12}'-\omega_{12})^2}$.

Proof: On minimizing the equation (3.8) with respect to $\gamma_1$and γ_2, we get the optimum value of the optimizing constants and then substituting them in the equation (3.8), the expression of $M.MSE$ of proposed estimator is obtained.

Corollary 3.2.1- $M.MSE$ of the proposed estimator $\mathcal{T}_{IE}^{(1)}$ under situation-I can also be obtained in terms of $M.MSE$ of the conventional regression estimator (t_4) as

$$M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) = M.MSE(t_4) - \frac{\left(B_{yz}A_x-B_{yx}B_{xz}\right)^2}{A_x\left(A_xA_z-B_{xz}^2\right)}, \tag{3.10}$$

here, $M.MSE(t_4) = A_y - \frac{B_{yx}^2}{A_x}$.

Proof: Now, on further simplifying the expression of minimum MSE of $\mathcal{T}_{IE}^{(1)}$ given in theorem 3.2 and analysing theoretically we can have the corollary 3.2.1 easily.

Theorem 3.3- Bias and mean square error (MSE) of the proposed estimator $\mathcal{T}_{IE}^{(2)}$ are given by

$$Bias\left(\mathcal{T}_{IE}^{(2)}\right) = \bar{Y}\left[\mu_1^2\frac{(\omega_1-\omega_{11}')}{8} + \mu_2^2\frac{(\omega_2-\omega_{22}')}{8} + \mu_1\left\{\frac{(\omega_{01}'-\omega_{01})}{2} - \frac{(\omega_1'-\omega_1)}{4}\right\} + \mu_2\left\{\frac{(\omega_{02}-\omega_{02}')}{2} - \frac{(\omega_2-\omega_2')}{4}\right\} + \mu_1\mu_2\frac{(\omega_{12}'-\omega_{12})}{8}\right] \tag{3.11}$$

$$MSE\left(\mathcal{T}_{IE}^{(2)}\right) = \bar{Y}^2\left[\omega_0 + \mu_1^2\left(\omega_1 - \omega_{11}'\right)/4 + \mu_2^2\left(\omega_2 - \omega_{22}'\right)/4 + \mu_1\mu_2\left(\omega_{12}' - \omega_{12}\right)/2 + \mu_1(\omega_{01}' - \omega_{01}) + \mu_2(\omega_{02} - \omega_{02}')\right] \tag{3.12}$$

Proof: Similarly, as proof of the previous theorem the proposed estimator $\mathcal{T}_{IE}^{(2)}$ takes the following form as:

$$\mathcal{T}_{IE}^{(2)} = \bar{Y}\left[1 + e_0 + \mu_1\left\{\frac{(e_1'-e_1+e_0e_1'-e_0e_1)}{2} - \frac{\left({e_1'}^2-{e_1}^2\right)}{4}\right\} + \mu_2\left\{\frac{(e_2-e_2'+e_0e_2-e_0e_2')}{2} - \frac{\left({e_2}^2-{e_2'}^2\right)}{4}\right\} + \mu_1^2\frac{\left(e_1'-e_1\right)^2}{8} + \mu_2^2\frac{\left({e_2}^2-{e_2'}^2\right)}{8} + \mu_1\mu_2\frac{(e_1'e_2-e_1e_2-e_1'e_2'+e_1e_2')}{4}\right]. \tag{3.13}$$

$$\mathcal{T}_{IE}^{(2)} - \bar{Y} = \bar{Y}\left[e_0 + \mu_1\left\{\frac{(e_1'-e_1+e_0e_1'-e_0e_1)}{2} - \frac{\left({e_1'}^2-{e_1}^2\right)}{4}\right\} + \mu_2\left\{\frac{(e_2-e_2'+e_0e_2-e_0e_2')}{2} - \frac{\left({e_2}^2-{e_2'}^2\right)}{4}\right\} + \mu_1^2\frac{\left(e_1'-e_1\right)^2}{8} + \mu_2^2\frac{\left({e_2}^2-{e_2'}^2\right)}{8} + \mu_1\mu_2\frac{(e_1'e_2-e_1e_2-e_1'e_2'+e_1e_2')}{4}\right]. \tag{3.14}$$

On taking expectations on both side of the equation (3.14), and using the expected values we get the expression of $Bias$ of $\mathcal{T}_{IE}^{(2)}$ as:

$$Bias\left(\mathcal{T}_{IE}^{(2)}\right) = \bar{Y}\left[\mu_1^2\frac{(\omega_1-\omega_{11}')}{8} + \mu_2^2\frac{(\omega_2-\omega_{22}')}{8} + \mu_1\left\{\frac{(\omega_{01}'-\omega_{01})}{2} - \frac{(\omega_1'-\omega_1)}{4}\right\} + \mu_2\left\{\frac{(\omega_{02}-\omega_{02}')}{2} - \frac{(\omega_2-\omega_2')}{4}\right\} + \mu_1\mu_2\frac{(\omega_{12}'-\omega_{12})}{4}\right]. \tag{3.15}$$

Now to study the characterization properties of proposed estimator, we get the expression for MSE of $\mathcal{T}_{IE}^{(2)}$ under situation-II by squaring and taking expectation of equation (3.14) as,

$$MSE\left(\mathcal{T}_{IE}^{(2)}\right) = \bar{Y}^2\left[\omega_0 + \mu_1^2\left(\omega_1 - \omega_{11}^{'}\right)/4 + \mu_2^2\left(\omega_2 - \omega_{22}'\right)/4 + \mu_1\mu_2\left(\omega_{12}' - \omega_{12}\right)/2 + \mu_1\left(\omega_{01}' - \omega_{01}\right) + \mu_2\left(\omega_{02} - \omega_{02}'\right)\right] \tag{3.16}$$

Theorem 3.4- Under situation-II the expression for minimum mean square error $(M.MSE)$ of the proposed estimator $\mathcal{T}_{IE}^{(2)}$ at the optimum value of $\mu_1$and μ_2 is

$$M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) = \left[A_y - \frac{\left(B_{yx}^2D_z + C_{yz}^2A_x - 2C_{yz}B_{xz}B_{yx}\right)}{A_xD_z - B_{xz}^2}\right]. \tag{3.17}$$

at, $\mu_{1(opt)} = \frac{2\left[\left(\omega_{02}-\omega_{02}'\right)\left(\omega_{12}'-\omega_{12}\right)-\left(\omega_2-\omega_{22}'\right)\left(\omega_{01}'-\omega_{01}\right)\right]}{\left(\omega_1-\omega_{11}^{'}\right)\left(\omega_2-\omega_{22}'\right)-\left(\omega_{12}'-\omega_{12}\right)^2}$ and $\mu_{2(opt)} = \frac{2\left[\left(\omega_{01}'-\omega_{01}\right)\left(\omega_{12}'-\omega_{12}\right)-\left(\omega_{02}-\omega_{02}'\right)\left(\omega_1-\omega_{11}^{'}\right)\right]}{\left(\omega_1-\omega_{11}^{'}\right)\left(\omega_2-\omega_{22}'\right)-\left(\omega_{12}'-\omega_{12}\right)^2}$.

Proof: On minimizing the equation (3.16) with respect to $\mu_1$and μ_2, we get the optimum value of the optimizing constants. Substituting these values in the equation (3.16) and after simplifying, we get the expression of minimum mean square error.

Corollary 3.4.1- The expression for $M.MSE$ of the proposed estimator $\left(\mathcal{T}_{IE}^{(2)}\right)$ under situation-II, can also be obtained in terms of $M.MSE$ of the conventional regression estimator (t_4) as

$$M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) = M.MSE(t_4) - \frac{\left(C_{yz}A_x - B_{yx}B_{xz}\right)^2}{A_x\left(A_xD_z - B_{xz}^2\right)}; \tag{3.18}$$

here, $M.MSE(t_4) = A_y - \frac{B_{yx}^2}{A_x}$.

Proof: Now, on further simplifying the expression of minimum MSE of $\mathcal{T}_{IE}^{(2)}$ given in theorem 3.4 and analysing theoretically we can have the corollary 3.4.1 easily.

here,

$A_y = \sum_{h=1}^{L} P_h^2\left(\lambda_h S_{Y_h}^2 + \theta_h W_{h(2)} S_{Y_{h(2)}}^2\right)$, $A_x = \sum_{h=1}^{L} P_h^2\{\triangle_h S_{X_h}^2 + \Pi_h W_{h(2)} S_{X_{h(2)}}^2\}$, $A_z = \sum_{h=1}^{L}\left(P_h^2\lambda_h S_{Z_h}^2\right)$

$D_z = \sum_{h=1}^{L}\left(\triangle_h P_h^2 S_{Z_h}^2\right)$, $B_{yx} = -\sum_{h=1}^{L} P_h^2\left(\triangle_h \rho_{Y_hX_h} S_{Y_h} S_{X_h} + \Pi_h W_{h(2)} \rho_{Y_hX_{h(2)}} S_{Y_{h(2)}} S_{X_{h(2)}}\right)$,

$B_{yz} = \sum_{h=1}^{L}\left(\lambda_h P_h^2 \rho_{Y_hZ_h} S_{Y_h} S_{Z_h}\right)$, $C_{yz} = \sum_{h=1}^{L}\left(\triangle_h P_h^2 \rho_{Y_hZ_h} S_{Y_h} S_{Z_h}\right)$, $B_{xz} = -\sum_{h=1}^{L}\left(\triangle_h P_h^2 \rho_{X_hZ_h} S_{X_h} S_{Z_h}\right)$

Thus, based on the established theorems, the proposed estimators $\mathcal{T}_{IE}^{(1)}$ and $\mathcal{T}_{IE}^{(2)}$ offer significant theoretical advantages. They enhance efficiency through exponential adjustments involving auxiliary variables, resulting in reduced bias and lower mean squared error (MSE) compared to traditional estimators. Their formulation accommodates both known and unknown auxiliary means, ensuring flexibility in diverse practical scenarios. The optimal values of the constants γ_1, γ_2, μ_1, and μ_2 are derived to minimize the MSE under first-order approximation. The unknown parameters required in the estimators are estimated from the sample, thereby maintaining robustness. Furthermore, the two-phase sampling design effectively integrates auxiliary information, improving accuracy while reducing data collection burden.

4. THEORETICAL COMPARISON OF EFFICIENCY

The efficacy of the proposed estimators has been compared in terms of mean square errors against all previously known competing estimators, with limitations arising from the use of various parameters and estimates. On putting the expressions of minimum MSE of the proposed estimators and respective adopted estimators in the inequality (i) to (xii) and on simplifying, we can easily get the following constraints and results for both the situations. The obtained results are shown in Table 2.

Situation-I	Situation-II
i. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < var\left(\bar{y}_{st}^*\right)$; if $\left(\frac{A_zB_{yx}^2 + A_xB_{yz}^2}{2B_{yx}B_{xz}B_{yz}} - 1\right) > 0.$	**vii.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < var\left(\bar{y}_{st}^*\right)$; if $\left(\frac{D_zB_{yx}^2 + A_xC_{yz}^2}{2B_{yx}B_{xz}C_{yz}} - 1\right) > 0.$
ii. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < MSE(t_1)$; if $2B_{yx}B_{xz} + A_x(RB_{xz} - B_{yz}) < A_z\frac{(RA_x - B_{yx})^2}{(RB_{xz} + B_{yz})}.$	**viii.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < MSE(t_1)$; if $2B_{yx}B_{xz} + A_x(RB_{xz} - C_{yz}) < D_z\frac{(RA_x - B_{yx})^2}{(RB_{xz} + C_{yz})}.$
ix. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < MSE(t_2)$; if $2B_{yx}B_{xz} - A_x(RB_{xz} + B_{yz}) < A_z\frac{(RA_x - B_{yx})^2}{(B_{yz} - RB_{xz})}.$	**ix.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < MSE(t_2)$; if $2B_{yx}B_{xz} - A_x(RB_{xz} + C_{yz}) < D_z\frac{(RA_x - B_{yx})^2}{(C_{yz} - RB_{xz})}.$

iv. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < MSE(t_3)$; if $2B_{yx}B_{xz} + A_x(\beta B_{xz} - B_{yz}) < A_z \frac{(\beta A_x + B_{yx})^2}{(\beta B_{xz} + B_{yz})}$; $\beta = \frac{S_{YX}}{S_X^2}$.	**x.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < MSE(t_3)$; if $2B_{yx}B_{xz} + A_x(\beta B_{xz} - C_{yz}) < D_z \frac{(\beta A_x + B_{yx})^2}{(\beta B_{xz} + C_{yz})}$; $\beta = \frac{S_{YX}}{S_X^2}$.
v. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < M.MSE(t_4)$; if $\left(B_{yz}A_x - B_{yx}B_{xz}\right)^2 > 0;$ $\textit{which is always true.}$	**xi.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < MSE(t_4)$; if $\left(C_{yz}A_x - B_{yx}B_{xz}\right)^2 > 0;$ $\textit{which is always true.}$
vi. $M.MSE\left(\mathcal{T}_{IE}^{(1)}\right) < M.MSE(t_5)$; if $\left(B_{yz}A_x - B_{yx}B_{xz}\right)^2 > 0;$ $\textit{which is always true.}$	**xii.** $M.MSE\left(\mathcal{T}_{IE}^{(2)}\right) < MSE(t_5)$; if $\left(C_{yz}A_x - B_{yx}B_{xz}\right)^2 > 0;$ $\textit{which is always true.}$

Table 2: Theoretical Constraints of the Proposed Estimators over Adopted Estimators

5. NUMERICAL ANALYSIS OF PERFORMANCE OF PROPOSED ESTIMATORS

This section aims to conduct two types of analyses for efficiency comparison: a simulation analysis using simulated data (symmetric and asymmetric) accounting for population type, and an empirical analysis using two different real-world data sets.

5.1. Efficiency Analysis on Simulated Data

We have used following statistical tools available in R software- mvrnorm (), unonr (), sample (), sampling (), moments ().

Algorithm for simulation study:

```
Input;
Generate Multivariate Categorical Data, $D = (Y_N, X_N, Z_N) \in M_{N\times 3}^{\mathbb{R}}$;

    $L$ :  Number of strata;
    $k$ :  Vector of sub-sampling factor;
    R:   Respondent Group;
    NR:  Non-respondent group;
    Rep: Replication needed;
Initialize;
    $\mathcal{S}'_{n_h} = \mathcal{S}_{n_h} = \emptyset$;
    Adopted Estimator (AE) Value = $\emptyset$ ;
    Proposed Estimator (PE) Value = $\emptyset$ ;
for j=1,…,length ($k$) do
      Stratify data $D$ into $L$ strata;
      Split each stratum into R and NR
    for i =1,…,Rep do
        Draw first phase sample of size $(n'_{h(1)}, n'_{h(2)}, r'_h)$ with $SiRS_{(wor)}$ from R and NR of each stratum.
        Estimate the unknown auxiliary mean from first phase sample.
        Draw second phase sample of size $(n_{h(1)}, n_{h(2)}, r_h)$ with $SiRS_{(wor)}$ from first phase sample.
        Estimate $\bar{Y}$ by AE and PE.
    end for
end for
Output;
Get mean square error of the AE and PE by the model;
        $MSE(T) = \frac{1}{Rep}\sum_{i=1}^{Rep}(T_i - \bar{Y})^2$; where $T$ = AE and PE;
```

For both situations I and II under two types of (symmetrical and asymmetrical) data sets, the mean square error and percentage relative efficiency (PRE) at different levels of k of the estimators are shown in Tables 3 to 4. The PRE of the estimators are calculated with respect to $\bar{y}_{st}^{*}$ using $PRE(\cdot) = \{var(\bar{y}_{st}^{*})/MSE(\cdot)\} \times 100$.

We have generated hypothetical data sets - symmetric and asymmetric with parameters (mentioned below with respect to variable $c(Y, X, Z)$ respectively), to perform the test the efficiency of the proposed estimators. The parameters are:

Symmetric Data-Set

Mean vector=$c(178, 37, 38)$,

Asymmetric Data Set

Mean vector=$c(1, 2, 3)$,

$$\text{Variance-covariance matrix} = \begin{bmatrix} 1.00 & 0.68 & 0.72 \\ 0.68 & 1.00 & 0.46 \\ 0.72 & 0.46 & 1.00 \end{bmatrix} \quad \text{Variance-covariance matrix} = \begin{bmatrix} 1.00 & 0.71 & 0.72 \\ 0.71 & 1.00 & 0.46 \\ 0.72 & 0.46 & 1.00 \end{bmatrix}$$

$$\rho = \begin{bmatrix} 1.00 & 0.69 & 0.69 \\ 0.69 & 1.00 & 0.47 \\ 0.69 & 0.47 & 1.00 \end{bmatrix} \qquad \rho = \begin{bmatrix} 1.00 & 0.69 & 0.70 \\ 0.69 & 1.00 & 0.45 \\ 0.70 & 0.45 & 1.00 \end{bmatrix}$$

k Estimators	Symmetric Data set			Asymmetric Data Set		
	2	3	4	2	3	4
$\bar{y}^*_{st}$	0.002012096 (100)	0.00229757 (100)	0.002852795 (100)	0.001870726 (100)	0.002019792 (100)	0.002957734 (100)
t_1	0.03490391 (5.764672)	0.04563346 (5.034836)	0.06382944 (4.469402)	0.001170022 (159.8881)	0.001350506 (149.5581)	0.001948345 (151.8075)
t_2	0.05727571 (3.513001)	0.07157728 (3.209916)	0.0974817 (2.926492)	0.003421357 (54.6779)	0.003995108 (50.55663)	0.005874892 (50.34533)
t_3	0.001313993 (153.1284)	0.001591433 (144.3712)	0.002006689 (142.1643)	0.001138077 (164.3761)	0.001439026 (140.3582)	0.002028566 (145.8042)
t_4	0.001315691 (152.9308)	0.001586917 (144.782)	0.002004802 (142.2981)	0.001135001 (164.8216)	0.001393617 (144.9316)	0.002023063 (146.2008)
t_5	0.001315691 (152.9307)	0.001586926 (144.7812)	0.002004784 (142.2993)	0.001136855 (164.5527)	0.001395599 (144.7258)	0.00202404 (146.1302)
$\mathcal{T}^{(1)}_{IE}$	0.0009052529 **(222.269)**	0.001122478 **(204.6873)**	0.001570297 **(181.6723)**	0.0007682207 **(243.5142)**	0.000981715 **(205.7411)**	0.00157702 **(187.5521)**

Table 3: MSE and $PRE(\cdot)$ of the Estimators on Simulated Data-Sets under Situation-I

k Estimators	Symmetric Data-Set			Asymmetric Data Set		
	2	3	4	2	3	4
$\bar{y}^*_{st}$	0.00204137 (100)	0.002632348 (100)	0.003241607 (100)	0.00195045 (100)	0.003018519 (100)	0.003040825 (100)
t_1	0.03630133 (5.623417)	0.05623772 (4.680752)	0.06354986 (5.100887)	0.00124344 (156.8582)	0.001868144 (161.5785)	0.001984735 (153.2107)
t_2	0.05859142 (3.484086)	0.08783845 (2.996806)	0.1021647 (3.172923)	0.00377029 (51.73206)	0.005675361 (53.18638)	0.005588416 (54.413)
t_3	0.00136931 (149.0803)	0.001769844 (148.7333)	0.002111472 (153.5235)	0.00126389 (154.321)	0.001824714 (165.4242)	0.002010917 (151.2158)
t_4	0.00136856 (149.1619)	0.001778227 (148.0322)	0.002139771 (151.4931)	0.00124991 (156.0469)	0.001863889 (161.9473)	0.002027081 (150.01)
t_5	0.00136855 (149.1624)	0.001778247 (148.0305)	0.002139767 (151.4935)	0.00125231 (155.7474)	0.001867794 (161.6088)	0.002025434 (150.132)
$\mathcal{T}^{(2)}_{IE}$	0.00102721 **(198.73)**	0.001478301 **(178.07)**	0.001841667 **(176.02)**	0.0009721707 **(200.6286)**	0.001532334 **(196.9883)**	0.001740313 **(174.7287)**

Table 4: MSE and $PRE(\cdot)$ of the Estimators on Simulated Data-Sets under Situation-II

5.2. Efficiency Analysis on Real Data

We have used Hypertension Arterial Mexico Data Set for the empirical study of numerical analysis available at https://www.kaggle.com/datasets/frederickfelix/hipertensin-arterial-mxico. The data set includes raw information (such as body mass index, height, gender, weight, different glucose results etc.) taken from the national health and nutrition survey (ENSANUT) https://ensanut.insp.mx/encuestas/ensanutcontinua2022/descargas.php.

In the present investigation, two distinct sets of variables are taken into consideration:

Combination-1

$\boldsymbol{Y}$: $valor\_colesterol\_total$
$\boldsymbol{X}$: $valor\_colesterol\_hdl$
$\boldsymbol{Z}$: $valor\_trigliceridos$

Combination-2

$\boldsymbol{Y}$: $valor\_hemoglobina\_glucosilada$
$\boldsymbol{X}$: $resultado\_glucosa$
$\boldsymbol{Z}$: $resultado\_glucosa\_promedio$

Using gender as the primary stratification criterion, we classified 20% of the units as non-respondent groups based on the specific circumstances surrounding their non-response. The required parameters are described below

Combination-1 $\bar{Y}$ =44.1389 $\bar{X}$ =36.03025 $\bar{Z}$ =137.2698

h	N_h	$N_{h(2)}$	n'_h	n_h	$n'_{h(2)}$	$n_{h(2)}$	$\bar{Y}_h$	$\bar{X}_h$	$\bar{Z}_h$
1	1687	337.4	1349	674	269	134	144.9887	35. 5602	143.508
2	2676	535.2	2140	1070	428	214	143.603	36.3267	133.337

	$\bar{Y}_{h(2)}$	$\bar{X}_{h(2)}$	$\bar{Z}_{h(2)}$	S_{Y_h}	S_{X_h}	S_{Z_h}	$S_{Y_{h(2)}}$	$S_{X_{h(2)}}$	$S_{Z_{h(2)}}$
1	143.7046	35.077	141.4154	29.4789	9.2631	91.4104	19.7273	4.6657	70.9033
2	142.9943	36.482	130.2079	27.403	7.2266	67.4521	21.3396	6.5817	61.3585

	$\rho_{Y_hX_h}$	$\rho_{Y_hZ_h}$	$\rho_{X_hZ_h}$	$\rho_{Y_hX_h(2)}$	$\rho_{Y_hZ_h(2)}$	$\rho_{X_hZ_h(2)}$
1	0.4158	0.5552	0.05756	0.45435	0.46237	-0.1266
2	0.5598	0.5177	0.00895	0.56319	0.3486	-0.0594

Combination-2 $\bar{Y}$ =5.452074 $\bar{X}$ =96.89466 $\bar{Z}$ =110.3149

h	N_h	$N_{h(2)}$	n'_h	n_h	$n'_{h(2)}$	$n_{h(2)}$	$\bar{Y}_h$	$\bar{X}_h$	$\bar{Z}_h$
1	1687	337.4	1349	674	269	134	5.3898	95.65027	108.3106
2	2676	535.2	2140	1070	428	214	5.4913	97.6792	111.5785

	$\bar{Y}_{h(2)}$	$\bar{X}_{h(2)}$	$\bar{Z}_{h(2)}$	S_{Y_h}	S_{X_h}	S_{Z_h}	$S_{Y_{h(2)}}$	$S_{X_{h(2)}}$	$S_{Z_{h(2)}}$
1	5.29331	92.9757	105.550	0.840756	28.1073	24.0496	0.3881	12.4809	11.0681
2	5.4307	95.0320	109.358	1.0999	53.7513	36.9487	0.9155	27.1248	26.0122

	$\rho_{Y_hX_h}$	$\rho_{Y_hZ_h}$	$\rho_{X_hZ_h}$	$\rho_{Y_hX_h(2)}$	$\rho_{Y_hZ_h(2)}$	$\rho_{X_hZ_h(2)}$
1	0.8577	0.9999	0.8582	0.6531	0.9997	0.6528
2	0.5776	0.8514	0.4867	0.8482	0.9927	0.8126

For both situations I and II under two types of combinations of variables, the mean square error and percentage relative efficiency (PRE) at different levels of k of the estimators are shown in Tables 5 and 6.

	Combination-1			**Combination-2**		
k **Estimators**	**2**	**3**	**4**	**2**	**3**	**4**
$\bar{y}^*_{st}$	0.3341742 (100)	0.3944839 (100)	0.4547936 (100)	0.0004302727 (100)	0.0005114625 (100)	0.0005926524 (100)
t_1	0.4480475 (88.0451)	0.4480475 (88.0451)	0.5018696 (90.61987)	0.0013886 (30.98608)	0.001510897 (33.85159)	0.005168526 (11.46656)
t_2	0.9809442 (34.06659)	1.118271 (35.27623)	1.270098 (35.80775)	0.003417028 (12.59201)	0.003793379 (13.48303)	0.00842079 (7.037966)
t_3	0.2682137 (124.5925)	0.3222577 (122.4126)	0.3693938 (123.1189)	0.0003001237 (143.3651)	0.0003602162 (141.9877)	0.0005687025 (104.2113)
t_4	0.2681772 (124.6095)	0.3222515 (122.4149)	0.3692491 (123.1671)	0.000299904 (143.4701)	0.0003593573 (142.327)	0.0004860617 (121.9295)
t_5	0.2681772 (124.6095)	0.3222515 (122.4149)	0.3692491 (123.1671)	0.000299904 (143.4701)	0.0003593573 (142.327)	0.0004860617 (121.9295)
$\mathcal{T}^{(1)}_{IE}$	0.1946878 **(171.6462)**	0.2487106 **(158.6116)**	0.2958468 **(153.726)**	0.0001426275 **(301.6759)**	0.0002116649 **(241.6379)**	0.0002835951 **(208.9783)**

Table 5: MSE and $PRE(\cdot)$ of the Estimators on Empirical Data-Sets under Situation-I

k Estimators	Combination-1			Combination-2		
	2	3	4	2	3	4
$\bar{y}_{st}^{*}$	0.3341742 (100)	0.3944839 (100)	0.4547936 (100)	0.0004302727 (100)	0.0005114625 (100)	0.0005926524 (100)
t_1	0.3808453 (87.7454)	0.4493367 (87.79249)	0.5140582 (88.47121)	0.001398142 (30.77462)	0.00151616 (33.73407)	0.001562741 (37.9239)
t_2	0.9770939 (34.20083)	1.101085 (35.82683)	1.274506 (35.68392)	0.00345917 (12.43861)	0.00375117 (13.63475)	0.003949056 (15.00745)
t_3	0.269741 (123.8871)	0.3247531 (121.4719)	0.3725691 (122.0696)	0.0002976758 (144.5441)	0.0003649452 (140.1478)	0.0004299789 (137.8329)
t_4	0.2697316 (123.8914)	0.3247528 (121.472)	0.3725558 (122.0739)	0.0002974204 (144.6682)	0.0003643492 (140.377)	0.0004281283 (138.4287)
t_5	0.2697316 123.8914 ()	0.3247528 (121.472)	0.3725558 (122.0739)	0.0002974204 (144.6682)	0.0003643492 (140.377)	0.0004281283 (138.4287)
$\mathcal{T}_{IE}^{(2)}$	0.09851726 **(339.2037)**	0.1675565 **(235.4334)**	0.2334231 **(194.8366)**	0.0001856999 **(231.7033)**	0.0002623509 **(194.9536)**	0.0003370512 **(175.8345)**

Table 6: MSE and $PRE(\cdot)$ of the Estimators on Empirical Data-Sets under Situation-II

6. CONCLUSION AND INTERPRETATION

Using a two-phase sampling scheme, several authors, including Singh and Kumar ([16], [18]), Khare and Kumar [9], Bhushan and Pandey [2], and many more, have produced promising research in parameter estimation under missing data due to non-response when auxiliary mean is unknown. Whereas, the majority of authors have taken into account that, although it appears implausible in real-world situations, the auxiliary variable does not suffer non-response during the first phase but does at the second. This article concerns the problem of having non-response at both the phases of survey sampling. In this regard, the main objective of this present paper is to produce exponential estimators that are more effective in estimating the mean of the variable under study using an enhanced methodology.

To justify the efficiency of the proposed estimators we have performed the test of efficiency using mean square error (MSE) and percentage relative efficiency (PRE) under the defined situations (I and II) of non-response for different values of sub-sampling factor (k). The MSE and PRE of the estimators are shown in table 3 to 6 for hypothetical data sets (symmetrical and asymmetrical) as well as empirical data-sets (Combination 1 and Combination 2). At last, after a thorough analysis of the data, we conclude that our suggested estimators outperform all adopted existing and conventional estimators in terms of PRE. Strong evidence is presented by the research findings for the preference of the proposed estimators in the context of non-response under observed heterogeneous population to obtain effective population mean estimate in the real-time problem domain.

Conflict of Interest:

There is nothing to reveal about our conflicts of interest. No pertinent financial interest exists. We thus confirm that this is the authors' original work and is not presently being considered for publication anywhere else.

Funding Declaration:

This study did not receive any financial support, grants, or other forms of funding from public, commercial, or not-for-profit sectors.

Acknowledgements:

The authors express their sincere gratitude to Prof. Carlos N. Bouza for his early and insightful evaluation of the manuscript. We also extend our appreciation to the Editor-in-Chief and the anonymous referees for their constructive and valuable suggestions, which have significantly contributed to enhancing the quality of this work.

RECEIVED: MARCH, 2024.
REVISED: AUGUST, 2025.

REFERENCES

[1] ALILAH, D.A., OUMA, C.O., OMBAKA, E.O. (2023). Efficiency of domain mean estimators in the presence of non-response using two-stage sampling with non-linear and linear cost function. **Annals of Data Science,** 10 (2), 291-316.

[2] BHUSHAN, S., PANDEY, A.P. (2019). An efficient estimation procedure for the population mean under non-response**. Statistica,** 79 (4), 363-378.

[3] CHAUDHARY, M.K., KUMAR, A., 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.

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

[5] CHAUDHARY, M.K., KUMAR, A. (2020). An improvement in estimation of population mean using two auxiliary variables and two-phase sampling scheme under non-response. **Journal of Reliability and Statistical Studies,** 13 (2-4), 349-362.

[6] CHAUDHARY, M.K., VISHWAKARMA, G.K. (2019). A general family of factor-type estimators of population mean in the presence of non-response and measurement errors**. International Journal of Mathematics and Statistics,** 20 (1), 83-93.

[7] COCHRAN, W.G. (1977). **Sampling Techniques**. John Wiley & Sons,

[8] HANSEN, M.H., HURWITZ, W.N. (1946). The problem of non-response in sample surveys. **Journal of the American Statistical Association,** 41 (236), 517-529.

[9] KHARE, B.B., KUMAR, S. (2011). Estimation of population mean using known coefficient of variation of the study character in the presence of non-response. **Communications in Statistics-Theory and Methods,** 40 (11), 2044-2058.

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

[11] KUMAR, S., CHOUDHARY, M. (2024). A general class of estimators in presence of non-response and measurement error under two-phase successive sampling. **Thailand Statistician**, 22 (1), 40-49.

[12] OKAFOR, F.C., LEE, H. (2000). Double sampling for ratio and regression estimation with sub-sampling the non-respondents. **Survey Methodology**, 26 (2), 183-188.

[13] RAO, P.S.R.S. (1990). **Regression estimators with sub-sampling of non-respondents, in-data quality control**. Theory and Pragmatics, (Eds.) Gunar E. Liepins and V.R.R. Uppuluri, Marcel Dekker, New York, 191–208.

[14] REDDY, V.N. (1978). A study on the use of prior knowledge on certain population parameters in estimation. **Sankhya C**, 40, 29-37.

[15] SINGH, H.P., KUMAR, S. (2008). A regression approach to the estimation of the finite population mean in the presence of non-response. **Australian & New Zealand Journal of Statistics**, 50 (4), 395-408.

[16] SINGH, H.P., KUMAR, S. (2010). Improved estimation of population mean under two-phase sampling with subsampling the non-respondents. **Journal of Statistical Planning and Inferenc***e*, 140 (9), 2536-2550.

[17] SINGH, H.P., KUMAR, S., KOZAK, M. (2010). Improved estimation of finite-population mean using sub-sampling to deal with non-response in two-phase sampling scheme. **Communications in Statistics-Theory and Methods,** 39 (5), 791-802.

[18] SINGH, H.P., KUMAR, S. (2010). Estimation of mean in presence of non-response using two phase sampling scheme. **Statistical Papers**, 51, 559-582.

[19] SINHA, R.R., KHANNA, B. (2023). Two-phase ratio estimation using ordinal and ratio auxiliary variables in non-response. **Proceedings of the National Academy of Sciences, India Section A: Physical Sciences**, 93 (4), 695-702.

[20] SINHA, R.R., KUMAR, V. (2021). Improved estimation of variance under complete and incomplete information. **Investigación Operacional**, 42 (1), 1-9.

[21] SRIVASTAVA, S.K., JHAJJ, H.S. (1983). A class of estimators of the population mean using multi-auxiliary information. **Calcutta Statistical Association Bulletin**, 32 (1-2), 47-56.

[22] TRIPATHI, T.P., KHARE, B.B. (1997). Estimation of mean vector in presence of non-response. **Communications in Statistics-Theory and Methods,** 26 (9), 2255-2269.

[23] UNAL, C., KADILAR, C. (2022). A new population mean estimator under non-response cases. **Journal of Taibah University for Science,** 16 (1), 111-119.

[24] UNAL, C., KADILAR, C. (2023). Improved population mean estimator with exponential function under non-response. **Applied Mathematics-A Journal of Chinese Universities**, 38 (4), 562-580.

[25] WANI, Z.H., RIZVI, S.E.H. (2024). Optimum estimation of population means in stratified random sampling using regression type estimator–non-response situation. **Investigación Operacional**, 45 (3), 381-395.