Mathematics

https://www.aimspress.com/journal/Math

AIMS Mathematics, 10(4): 8794–8817.
DOI:10.3934/math.2025403
Received: 12 March 2025
Revised: 08 April 2025
Accepted: 11 April 2025
Published: 17 April 2025

Research article

Estimation of the finite population mean using extreme values and ranks of the auxiliary variable in two-phase sampling

Hleil Alrweili[1] **and Fatimah A. Almulhim**[2,*]

[1] Department of Mathematics, College of Science, Northern Border University, Arar, Saudi Arabia; hleil.alrweili@nbu.edu.sa

[2] Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University, P. O. Box 84428, Riyadh 11671, Saudi Arabia

* **Correspondence:** Email: faalmulhim@pnu.edu.sa.

Abstract: This study aimed to improve the estimation of the mean of the dependent variable by incorporating the smallest and largest values and ranks of the independent variable. To achieve this, we introduce two new classes of estimators that offer enhanced accuracy compared with the existing approaches, as evaluated using the mean squared error (MSE) criterion. The key features of the proposed estimators are examined through a first-order approximation method, particularly focusing on the bias and mean squared error under two-phase sampling. In addition, their performance is assessed using simulated populations generated from six different distributions with varying parameter settings, along with three real datasets. Furthermore, the findings show that the new estimators achieve lower mean squared errors compared with existing methods.

Keywords: two-phase sampling; largest and smallest values; ranks; bias; mean squared error
Mathematics Subject Classification: 62D

1. Introduction

Supplementary information is a useful method in survey sampling for increasing estimators' precision. To enhance relative efficiency, methods such as ratio, regression, and product-type estimates utilize not only the data related to the primary study variables but also supplementary details from one or more auxiliary variables. Extensive research efforts have been dedicated to developing improved estimators for the overall parameters, including measures like the population mean, total, median, and various additional indicators. For more details about improved estimators and their properties, see [1–3] and references therein.

The sampling theory uses auxiliary variables as well as the study variable in order to get a better design, and it is also used to fine-tune the cost-effectiveness of the estimators by utilizing the relationship between the variables. In such situations, the two-phase sampling method is preferred, especially when the overall population mean is unavailable prior to conducting a survey. Survey methods that use two distinct stages to select samples from a population are called two-phase sampling methods. A preliminary phase can be conducted more cost-effectively or efficiently before the main phase when the main phase is more costly. The basic concept of two-phase sampling was initially introduced by [4]. Recently, it has attracted considerable interest because of its cost-effectiveness in assessing variables. A few investigations that have been conducted on two-phase sampling include [5–8].

The information obtained through a sample survey can produce unexpected results. When the sample contains extreme values, the mean estimator becomes susceptible to distortion, which may lead to biased results. Using the extreme values from known auxiliary variables, [9] first provided two estimators through the use of linear transformations. After this initial work, [10] suggested better estimators based on ratios, products, and regression to estimate population means more successfully by utilizing outlier values. In order to more accurately obtain the unknown population mean by considering extreme values, [11] introduced some estimators based on a two-phase strategy. A new set of transformations based on extreme values of auxiliary variables were provided by [12], which enhanced the estimation of the unknown population mean. Subsequently, [13] achieved significant improvements in increasing the precision of population mean estimations by introducing extreme values into a stratified random sampling technique. Using extreme value ideas as a foundation, [14,15] obtained a family of estimators intended to minimize the mean squared errors by using the population variance. In a recent work, [16–18] proposed various classes of highly efficient estimators by employing transformations on extreme values to accurately estimate the population variance. A ratio estimation of the population mean using auxiliary information under the optimal sampling design was proposed by [19]. Additional insights and related methodologies can be explored in [20–22].

1.1. Motivation and innovation

The motivation behind this study lies in the ongoing challenges of improving the accuracy and efficiency of population mean estimation, particularly in survey sampling. Traditional estimators often struggle with the presence of extreme values, such as the smallest and largest observations, which can distort the results. This issue is particularly critical in fields like economics, healthcare, and public health, where precise population estimates directly influence policy decisions and resource allocation.

Sometimes, there is a temptation to remove extreme observations from the data-sets. However, the accuracy of estimators often declines when the mean squared error (MSE) is calculated in scenarios involving extreme values. The innovation of this work lies in the development of these new estimators, which are designed to improve the precision of population mean estimation while minimizing the MSE compared with existing methods. In this article, we introduce different improved estimators that employ both the smallest and largest observations of independent variables and their ranks in a double-phase sampling design to improve the precision of estimation. The new estimators are widely applicable, with potential uses in areas like economics, healthcare, manufacturing, retail, and transportation. They are particularly effective in scenarios demanding precise predictions. The following are some implications of our suggested estimators.

1.2. Applications of the proposed estimators across domains

The new estimators are widely applicable, with potential uses in areas like economics, healthcare, manufacturing, retail, and transportation. They are particularly effective in scenarios demanding precise predictions.

1) **Industry:** In the industrial sector, our method supports predictive maintenance by monitoring equipment in manufacturing. For instance, in an automotive plant, it analyzes sensor data to forecast component failures, enabling timely repairs that reduce downtime and costs. This improves efficiency, lowers expenses, and ensures smoother production.
2) **Economics:** The estimators significantly improve economic forecasting by providing more accurate predictions of indicators like inflation, unemployment, and gross domestic product (GDP) growth. This makes it possible for the public and commercial sectors to allocate resources more effectively and make better policy decisions.
3) **Environmental science** The proposed estimators are important in environmental science, assisting in environmental forecasting and estimating pollution. They provide accurate predictions of ecological patterns, aiding in the projection of atmospheric shifts, pollution concentrations, and their effects on ecosystems.
4) **Transportation:** Through traffic pattern analysis and delay prediction, our approach improves route optimization for delivery services in the transportation industry. As a result, delivery times and fuel consumption can be reduced by logistics businesses planning more effective routes.
5) **Public health:** In public health, the estimators enhance epidemiological modeling and health outcome predictions. They enable more precise forecasting of disease spread, healthcare demands, and intervention planning. Organizations are better able to allocate resources and get ready for future healthcare demands due to the more accurate information.
6) **Retail:** Our approach more precisely predicts product demand in retail, which enhances inventory management. This enables retailers to keep inventory levels balanced, preventing both excess stock and shortages. These examples show how adaptable our methodology is and how it has the ability to significantly advance both industry and healthcare.
7) **Medicine:** Our approach can enhance early cancer detection by improving the accuracy of medical imaging analysis. When integrated with tools like magnetic resonance imaging (MRI) or computed tomography (CT) scans, it aids in identifying malignant tissues more effectively, enabling timely intervention. For example, it could improve tumor detection in breast cancer through better image segmentation and classification of abnormal tissue.

The structure of the paper is arranged as follows. The basic notations and concepts utilized in this study are explained in Section 2. Section 3 provides a summary of various existing estimators. The proposed classes of estimators are thoroughly discussed in Section 4. There is a rigorous mathematical comparison in Section 5 between existing and new proposed estimators. A description of a simulation conducted to produce some populations from various distributions is presented in Section 6, which is intended to evaluate the theoretical conclusions from Section 5. This section also presents numerical examples that illustrate the practical implications of the theoretical findings. Finally, Section 7 summarizes the overall conclusions and suggests some ideas for new research.

2. Framework for methodology and notation

A finite population can be represented as $\Omega = (\Omega_1, \Omega_2, \Omega_3, \ldots, \Omega_N)$, where N denotes the total number of units. In this context, let x_i represent the value of the independent variable X, r_i indicate the ranks associated with the independent variable R, and y_i denote the value of the response variable Y for the ith unit.

This research explores the impact of the variable X and presents two new estimators developed to obtain the population mean Y. The two-phase sampling scheme is described by the following definition.

(1) We collect a sample without replacement of size m_1 in the initial stage in order to accurately obtain the population mean, represented as $\bar{X}$. For accurate results, we must make sure that the sample size m_1 remains under N.

(2) A second sample (without replacement) size of m_2 observation (where $m_2 < m_1$) is selected to determine the variables y and x in the second phase.

The following are the formulas for the population means, which are calculated as

$$\bar{X} = \frac{\sum_{i=1}^{N} x_i}{N},$$

$$\bar{R} = \frac{\sum_{i=1}^{N} r_i}{N},$$

and

$$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N}.$$

Let the population variances for variables (X, R, Y) be defined as follows without replacement sampling:

$$S_x^2 = \frac{\sum_{i=1}^{N} \left(x_i - \bar{X}\right)^2}{N-1},$$

$$S_r^2 = \frac{\sum_{i=1}^{N} \left(r_i - \bar{R}\right)^2}{N-1},$$

$$S_y^2 = \frac{\sum_{i=1}^{N} \left(y_i - \bar{Y}\right)^2}{N-1},$$

while the population coefficients of variations for these variables are defined as:

$$C_x = \frac{S_x}{\bar{X}},$$

$$C_r = \frac{S_r}{\bar{R}},$$

and

$$C_y = \frac{S_y}{\bar{Y}},$$

respectively. Furthermore, the relationships among the variables (Y, X), (Y, R), and (X, R) are described through their population correlation coefficients as follows:

$$\rho_{yx} = \frac{S_{yx}}{S_y S_x},$$

$$\rho_{yr} = \frac{S_{yr}}{S_y S_r},$$

and

$$\rho_{xr} = \frac{S_{xr}}{S_x S_r},$$

where

$$S_{yx} = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(x_i - \bar{X}\right)}{N-1},$$

$$S_{yr} = \frac{\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)\left(r_i - \bar{R}\right)}{N-1},$$

and

$$S_{xr} = \frac{\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(r_i - \bar{R}\right)}{N-1},$$

are the population co-variances, respectively.

Now, we define the first-phase sample means and variances base on m_1 observations associated with the variables X and R, which are presented in the following manner:

$$\bar{x}' = \frac{\sum_{i=1}^{m_1} x_i}{m_1},$$

$$\bar{r}' = \frac{\sum_{i=1}^{m_1} r_i}{m_1},$$

$$s_x'^2 = \frac{\sum_{i=1}^{m_1} (x_i - \bar{x}')^2}{m_1 - 1},$$

and

$$s_r'^2 = \frac{\sum_{i=1}^{m_1} (r_i - \bar{r}')^2}{m_1 - 1}.$$

To calculate the sample means and variances for the second phase, a random sample without replacement of size m_2 observations is chosen based on the first phase $(m_2 < m_1)$, which are defined as:

$$\bar{x} = \frac{\sum_{i=1}^{m_2} x_i}{m_2},$$

$$\bar{r} = \frac{\sum_{i=1}^{m_2} r_i}{m_2},$$

$$\bar{y} = \frac{\sum_{i=1}^{m_2} y_i}{m_2},$$

$$s_x^2 = \frac{\sum_{i=1}^{m_2} (x_i - \bar{x})^2}{m_2 - 1},$$

$$x_r^2 = \frac{\sum_{i=1}^{m_2} (r_i - \bar{r})^2}{m_2 - 1},$$

and

$$s_y^2 = \frac{\sum_{i=1}^{m_2} (y_i - \bar{y})^2}{m_2 - 1}.$$

3. Existing estimators

The analysis of the mean squared errors and biases associated with the estimators used to determine the population mean is given in this section. These findings are then compared with those of the proposed classes of estimators to identify areas for improvement.

The conventional unbiased estimator is presented below:

$$\bar{y}^d = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{3.1}$$

The variance of the conventional unbiased estimator $\bar{y}^d$ is defined by:

$$V(\bar{y}^d) = \theta \bar{Y}^2 C_y^2, \tag{3.2}$$

where

$$\theta = \left(\frac{1}{m_2} - \frac{1}{N}\right)$$

is the sampling correction term for the second phase.

The following is an expression for a ratio-type estimator $\bar{y}_R^d$ according to [4]:

$$\bar{y}_R^d = \frac{\bar{y}}{\bar{x}} \bar{x}'. \tag{3.3}$$

The following formulas are used to express the bias and MSE of $\bar{y}_R^d$:

$$Bias\left(\bar{y}_R^d\right) \cong \theta'' \bar{Y}\left(C_x^2 - C_{yx}\right), \tag{3.4}$$

$$MSE\left(\bar{y}_R^d\right) \cong \bar{Y}^2\left(\theta C_y^2 + \theta'' C_x^2 - 2\theta'' C_{yx}\right), \tag{3.5}$$

where

$$\theta' = \left(\frac{1}{m_1} - \frac{1}{N}\right),$$

and

$$\theta'' = \left(\frac{1}{m_2} - \frac{1}{m}\right)$$

represent the sampling correction terms for the first and the intermediate stages.

The product estimator $\bar{y}_P^d$ is expressed as:

$$\bar{y}_P^d = \frac{\bar{y}}{\bar{x}'}\bar{x}. \tag{3.6}$$

The following formulas are used to express the bias and MSE of $\bar{y}_P^d$:

$$Bias\left(\bar{y}_P^d\right) \cong \bar{Y}\theta'' C_{yx}, \tag{3.7}$$

and

$$MSE\left(\bar{y}_P^d\right) \cong \bar{Y}^2\left(\theta C_y^2 + \theta'' C_x^2 + 2\theta'' C_{yx}\right). \tag{3.8}$$

The standard regression estimator is denoted as $\bar{y}_{lr}^d$, which is generally represented by:

$$\bar{y}_{lr}^d = \bar{y} + b_{yx}\left(\bar{x}' - \bar{x}\right), \tag{3.9}$$

where b_{yx} denotes the regression coefficient from the sample.

The following formulas are used to express the bias and MSE of $\bar{y}_{lr}^d$:

$$Bias\left(\bar{y}_{lr}^d\right) \cong -\theta''\beta\left(\frac{\mu_{12}}{S_{yx}} - \frac{\mu_{03}}{S_x^2}\right), \tag{3.10}$$

$$MSE\left(\bar{y}_{lr}^d\right) \cong \bar{Y}^2 C_y^2\left(\theta - \theta''\rho_{yx}^2\right), \tag{3.11}$$

where

$$\beta = \frac{S_{yx}}{S_x^2},$$

and

$$\mu_{rs} = \frac{\sum_{i=1}^N \left(y_i - \bar{Y}\right)^r \left(x_i - \bar{X}\right)^s}{N}.$$

According to [23], the definitions of the exponential ratio and product type estimators are given by

$$\bar{y}_{RMD}^d = \bar{y}\exp\left(\frac{\bar{x}' - \bar{x}}{\bar{x}' + \bar{x}}\right), \tag{3.12}$$

and

$$\bar{y}_{PMD}^d = \bar{y}\exp\left(\frac{\bar{x} - \bar{x}'}{\bar{x} + \bar{x}'}\right). \tag{3.13}$$

The expressions for MSE of $\bar{y}_{RMD}^d$ and $\bar{y}_{PMD}^d$ are, respectively, expressed as follows:

$$MSE\left(\bar{y}_{RMD}^d\right) \cong \bar{Y}^2\left[\theta C_y^2 + \theta'' C_x^2\left(\frac{1}{4} - C\right)\right], \tag{3.14}$$

and

$$MSE\left(\bar{y}_{PMD}^d\right) \cong \bar{Y}^2\left[\theta C_y^2 + \theta'' C_x^2\left(\frac{1}{4} + C\right)\right], \tag{3.15}$$

where

$$C = \left(\frac{\rho_{yx}}{C_x}\right)C_y.$$

The authors of [24] proposed a double sampling estimator is defined as

$$\bar{y}^d_{RP} \cong \bar{y}\left\{s\frac{\bar{x}'}{\bar{x}} + (1-s)\frac{\bar{x}}{\bar{x}'}\right\}, \tag{3.16}$$

where

$$s = \frac{1+C}{2}.$$

The following formulas are used to express the bias and MSE of $\bar{y}^d_{RP}$:

$$Bias(\bar{y}^d_{RP})_{\min} \cong \frac{\theta''}{2}\bar{Y}C_x^2\left[1 + C(1-2C)\right], \tag{3.17}$$

and

$$MSE(\bar{y}^d_{RP})_{\min} \cong \bar{Y}^2C_y^2\left[\theta(1-\rho_{yx}^2) + \theta'\rho_{yx}^2\right]. \tag{3.18}$$

4. Proposed generalized estimators

This section, inspired by the methodologies discussed in [25–27], presents some improved classes of estimators. These improved estimators employ both the smallest and largest observations of the independent variables and their ranks in a double-phase design to improve estimation precision. The proposed estimators are defined below:

$$\bar{y}^d_U = \left[k_1\bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_1} + k_2\left(\frac{\bar{x}'}{\bar{x}}\right)^{\alpha_2}\right]\exp\left[\frac{a_1\left(\bar{x}'-\bar{x}\right)}{a_1\left(\bar{x}'+\bar{x}\right)+2a_2}\right], \tag{4.1}$$

and

$$\bar{y}^d_e = \bar{y}\exp\left[k_3\left\{\frac{(\bar{x}'-\bar{x})}{(\bar{x}'-\bar{x})+2b_1}\right\}\right]\exp\left[k_4\left\{\frac{(\bar{r}'-\bar{r})}{(\bar{r}'+\bar{r})+2b_2}\right\}\right], \tag{4.2}$$

where the scalar quantities (α_1, α_2) can assume the values $(0, -1, 1)$. It is important to determine the unknown constant values of (k_1, k_2) by using the scalar quantities (α_1, α_2) so that the biases and mean squared errors can be minimized. Meanwhile, $b_1 = X_M - X_m$ represents the difference between extreme observations of the independent variable, and $b_2 = R_M - R_m$ represents the difference between the highest and lowest ranks of the independent variable. In addition, a_1 and a_2 are the various transformation parameter values. A detailed description of the subsets of the proposed estimator-I can be found in Table 1.

Table 1. Classification of estimators under the improved estimator-I.

Subsets of $\bar{y}^d_U$	α_1	α_2	a_1	a_2
$\bar{y}^d_{U_1} = \left[k_1\bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) + k_2\left(\frac{\bar{x}}{\bar{x}'}\right)\right]L$	1	-1	$-\beta_{2(x)}$	$x_M - x_m$
$\bar{y}^d_{U_2} = \left[k_1\bar{y}\left(\frac{\bar{x}}{\bar{x}'}\right) + k_2\left(\frac{\bar{x}'}{\bar{x}}\right)\right]L$	-1	1	$-c_x$	$x_M - x_m$
$\bar{y}^d_{U_3} = \left[k_1\bar{y}\left(\frac{\bar{x}}{\bar{x}'}\right) + k_2\left(\frac{\bar{x}}{\bar{x}'}\right)\right]L$	-1	-1	$x_M - x_m$	$-c_x$
$\bar{y}^d_{U_4} = \left[k_1\bar{y} + k_2\left(\frac{\bar{x}'}{\bar{x}}\right)\right]L$	0	1	$x_M - x_m$	$\beta_{2(x)}$
$\bar{y}^d_{U_5} = \left[k_1\bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) + k_2\left(\frac{\bar{x}'}{\bar{x}}\right)\right]L$	1	1	$x_M - x_m$	$-\beta_{2(x)}$
$\bar{y}^d_{U_6} = \left[k_1\bar{y} + k_2\left(\frac{\bar{x}}{\bar{x}'}\right)\right]L$	0	-1	$\beta_{2(x)}$	$x_M - x_m$
$\bar{y}^d_{U_7} = \left[k_1\bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) + k_2\right]L$	1	0	$x_M - x_m$	c_x
$\bar{y}^d_{U_8} = \left[k_1\bar{y}\left(\frac{\bar{x}}{\bar{x}'}\right) + k_2\right]L$	-1	0	c_x	$x_M - x_m$

where

$$L = \exp\left[\frac{a_1\left(\bar{x}' - \bar{x}\right)}{a_1\left(\bar{x}' + \bar{x}\right) + 2a_2}\right].$$

4.1. Properties of the improved estimator-I

To determine the mathematical properties of different estimators, the relative error terms are utilized:

$$e_0 = \left(\frac{\bar{y} - \bar{Y}}{\bar{Y}}\right), e_1 = \left(\frac{\bar{x} - \bar{X}}{\bar{X}}\right), e_2 = \left(\frac{\bar{x}' - \bar{X}}{\bar{X}}\right), e_3 = \left(\frac{\bar{r} - \bar{R}}{\bar{R}}\right), e_4 = \left(\frac{\bar{r}' - \bar{R}}{\bar{R}}\right),$$

such that $E\left(e_i\right) = 0$.
Moreover,

$$E\left(e_0^2\right) = \theta C_y^2, \quad E\left(e_1^2\right) = \theta C_x^2, \quad E\left(e_2^2\right) = \theta' C_x^2, \quad E\left(e_3^2\right) = \theta C_r^2, \quad E\left(e_4^2\right) = \theta' C_r^2, E\left(e_0 e_1\right) = \theta C_{yx},$$

$$E\left(e_0 e_2\right) = \theta' C_{yx}, \quad E\left(e_0 e_3\right) = \theta C_{yr}, \quad E\left(e_0 e_4\right) = \theta' C_{yr}, \quad E\left(e_1 e_2\right) = \theta' C_x^2, \quad E\left(e_1 e_3\right) = \theta C_{xr},$$

$$E\left(e_1 e_4\right) = \theta' C_{xr}, \quad E\left(e_2 e_3\right) = \theta' C_{xr}, \quad E\left(e_2 e_4\right) = \theta' C_{xr}, \quad E\left(e_3 e_4\right) = \theta' C_r^2.$$

In order to investigate the characteristics of the first improved estimator, we rewrite (4.1) using the error terms:

$$\begin{aligned}\bar{y}_U^d = &\left[k_1 \bar{Y}\left(1 + e_0\right)\left(1 + e_1\right)^{-\alpha_1}\left(1 + e_2\right)^{\alpha_1} + k_2\left(1 + e_1\right)^{-\alpha_2}\left(1 + e_2\right)^{\alpha_2}\right] \times \\ &\exp\left[\tfrac{g_1\left(e_2 - e_1\right)}{2}\left(1 + \tfrac{g_1}{2}\left(e_2 + e_1\right)\right)^{-1}\right],\end{aligned} \tag{4.3}$$

where

$$g_1 = \frac{a_1 \bar{X}}{a_1 \bar{X} + a_2}.$$

By performing a first-order Taylor series expansion and extending the right-hand sides of (4.3), excluding terms where $e_i > 2$, we get the following expression:

$$\begin{aligned}\bar{y}_U^d - \bar{Y} \cong -\bar{Y} + k_1 \bar{Y}&\Big[1 + e_0 - e_1\left(\alpha_1 + \tfrac{g_1}{2}\right) + e_2\left(\alpha_1 + \tfrac{g_1}{2}\right) + e_1^2\left(\tfrac{\alpha_1 g_1}{2} + \tfrac{3g_1^2}{8} + \tfrac{\alpha_1(\alpha_1+1)}{2}\right) \\ &+ e_2^2\left(\tfrac{\alpha_1 g_1}{2} - \tfrac{g_1^2}{8} + \tfrac{\alpha_1(\alpha_1-1)}{2}\right) - e_0 e_1\left(\alpha_1 + \tfrac{g_1}{2}\right) - e_0 e_2\left(\alpha_1 + \tfrac{g_1}{2}\right) - e_1 e_2\left(\alpha_1 + \tfrac{g_1}{2}\right)^2\Big] \\ &+ k_2\Big[1 - e_1\left(\alpha_2 + \tfrac{g_1}{2}\right) + e_2\left(\alpha_2 + \tfrac{g_1}{2}\right) + e_1^2\left(\tfrac{\alpha_2 g_1}{2} + \tfrac{3g_1^2}{8} + \tfrac{\alpha_2(\alpha_2+1)}{2}\right) \\ &+ e_2^2\left(\tfrac{\alpha_2 g_1}{2} - \tfrac{g_1^2}{8} + \tfrac{\alpha_2(\alpha_2-1)}{2}\right) - e_1 e_2\left(\alpha_2 + \tfrac{g_1}{2}\right)^2\Big].\end{aligned} \tag{4.4}$$

Using (4.4), the bias of $\bar{y}_U^d$ is given by

$$Bias\left(\bar{y}_U^d\right) \cong \left[-\bar{Y} + k_1 \bar{Y} D + k_2 G\right], \tag{4.5}$$

where

$$\begin{aligned}D = &\Big[1 + \theta\left\{C_x^2\left(\tfrac{4\alpha_1(\alpha_1+1+g_1)+3g_1^2}{8}\right) - C_{yx}\left(\tfrac{2\alpha_1+g_1}{2}\right)\right\} \\ &+ \theta'\left\{C_x^2\left\{\tfrac{-4\alpha_1(\alpha_1+g_1+1)-3g_1^2}{2}\right\} + C_{yx}\left(\tfrac{2\alpha_1+g_1}{2}\right)\right\}\Big],\end{aligned}$$

and

$$G = \left[1 + \theta C_x^2 \left(\frac{4\alpha_2\left(\alpha_2 + g_1 + 1\right) + 3g_1^2}{8}\right) + \theta' C_x^2 \left(\frac{-2\alpha_2\left(\alpha_2 + g_1 - 1\right) - 3g_1^2}{4}\right)\right].$$

By applying expectation after squaring both sides of (4.4), we derive the corresponding equation that expresses the MSE of $\bar{y}_U^d$.

$$MSE\left(\bar{y}_U^d\right) \cong \left[\bar{Y}^2 + \bar{Y}^2 k_1^2 A + k_2^2 B - 2\bar{Y}^2 k_1 D - 2\bar{Y} k_2 G + 2\bar{Y} k_1 k_2 F\right], \tag{4.6}$$

where

$$\begin{aligned} A = \Big[1 + \theta \Big\{C_y^2 + C_x^2 \Big\{\Big(\alpha_1 + \tfrac{g_1}{2}\Big)^2 + \Big(\alpha_1 g_1 + \tfrac{3g_1^2}{4} + \tfrac{\alpha_1(\alpha_1+1)}{2}\Big)\Big\} - 4C_{yx}\Big(\alpha_1 + \tfrac{g_1}{2}\Big)\Big\} \\ +\theta' \Big\{C_x^2 \Big\{\Big(\alpha_1 + \tfrac{g_1}{2}\Big)^2 + \Big(\alpha_1 g_1 - \tfrac{g_1^2}{4} + \alpha_1\left(\alpha_1 - 1\right)\Big) -4\Big(\alpha_1 + \tfrac{g_1}{2}\Big)\Big\} + 4C_{yx}\Big(\alpha_1 + \tfrac{g_1}{2}\Big)\Big\}\Big], \end{aligned}$$

$$\begin{aligned} B = \Big[1 + \theta C_x^2 \Big\{\Big(\alpha_2 + \tfrac{g_1}{2}\Big)^2 + \Big(\alpha_2 g_1 + \tfrac{3g_1^2}{4} + \alpha_2\left(\alpha_2 + 1\right)\Big)\Big\} + \theta' C_x^2 \Big\{\Big(\alpha_2 + \tfrac{g_1}{2}\Big)^2 \\ +\Big(\alpha_2 g_1 - \tfrac{g_1^2}{4} + \alpha_2\left(\alpha_2 - 1\right)\Big) - 4\Big(\alpha_2 + \tfrac{g_1}{2}\Big)^2\Big\}\Big], \end{aligned}$$

and

$$\begin{aligned} F = \Big[1 + \theta \Big\{C_x^2 \Big\{\Big(\tfrac{\alpha_1 g_1}{2} + \tfrac{3g_1^2}{8} + \tfrac{\alpha_1(\alpha_1+1)}{2}\Big) + \Big(\alpha_1 + \tfrac{g_1}{2}\Big)\Big(\alpha_2 + \tfrac{g_1}{2}\Big) + \Big(\tfrac{\alpha_2 g_1}{2} + \tfrac{3g_1^2}{8} + \tfrac{\alpha_2(\alpha_2+1)}{2}\Big) \\ - C_{yx}\left(\alpha_1 + \alpha_2 + g_1\right)\Big\} + \theta' \Big\{C_x^2 \Big\{\Big(\tfrac{\alpha_1 g_1}{2} - \tfrac{g_1^2}{8} + \tfrac{\alpha_1(\alpha_1-1)}{2}\Big) - \Big(\alpha_1 + \tfrac{g_1}{2}\Big)\Big(\alpha_2 + \tfrac{g_1}{2}\Big) \\ + \Big(\tfrac{\alpha_2 g_1}{2} - \tfrac{g_1^2}{8} + \tfrac{\alpha_2(\alpha_2-1)}{2}\Big) - \Big(\alpha_1 + \tfrac{g_1}{2}\Big)^2 - \Big(\alpha_2 + \tfrac{g_1}{2}\Big)^2\Big\} + C_{yx}(\alpha_1 + \alpha_2 + g_1)\,\}]\,. \end{aligned}$$

By minimizing Eq (4.6), the optimal values for k_1 and k_2 can be determined, as shown below

$$k_{1(opt)} = \frac{BD - FG}{AB - F^2},$$

and

$$k_{2(opt)} = \frac{\bar{Y}\left(AG - DF\right)}{AB - F^2}.$$

The minimum values for bias and MSE of $\bar{y}_U^d$ are determined by replacing the optimum k_1 and k_2 into (4.5) and (4.6). The resulting expressions are given below:

$$Bias\left(\bar{y}_U^d\right)_{min} \cong -\bar{Y}^2\left[1 - \frac{\left(AG^2 + BD^2 - 2DFG\right)}{AB - F^2}\right], \tag{4.7}$$

and

$$MSE\left(\bar{y}_U^d\right)_{min} \cong \bar{Y}^2\left[1 - \frac{\left(AG^2 + BD^2 - 2DFG\right)}{AB - F^2}\right]. \tag{4.8}$$

4.2. Properties of the improved estimator-II

To assess the behavior of the proposed estimator, we reformulate (4.2) using relative errors, which allows the computation of the Eqs (4.13) and (4.14), i.e.,

$$\begin{aligned}\bar{y}_e^d = \bar{Y}(1+e_0)\exp\left[k_3\left\{\tfrac{g_2(e_2-e_1)}{2}\left(1+\tfrac{g_2}{2}(e_1+e_2)\right)^{-1}\right\}\right]\exp\left[k_4\left\{\tfrac{g_3(e_4-e_3)}{2}\right.\right.\\ \left.\left.\left(1+\tfrac{g_3}{2}(e_3+e_4)\right)^{-1}\right\}\right],\end{aligned} \tag{4.9}$$

where

$$g_2 = \frac{\bar{X}}{\bar{X}+b_1},$$

and

$$g_3 = \frac{\bar{R}}{\bar{R}+b_2}.$$

By performing a first-order Taylor series expansion and extending the right-hand sides of (4.9), excluding terms where $e_i > 2$, we get the following expression:

$$\begin{aligned}\bar{y}_e^d - \bar{Y} \cong \bar{Y}\Big[e_0 - \tfrac{k_3g_2}{2}(e_1-e_2) - \tfrac{k_4g_3}{2}(e_3-e_4) + \left(\tfrac{k_3g_2^2}{4}+\tfrac{k_3^2g_2^2}{8}\right)e_1^2 - \left(\tfrac{k_3g_2^2}{4}-\tfrac{k_3^2g_2^2}{8}\right)e_2^2\\ +\left(\tfrac{k_4g_3^2}{4}-\tfrac{k_4^2g_3^2}{8}\right)e_3^2 - \left(\tfrac{k_4g_3^2}{4}-\tfrac{k_4^2g_3^2}{8}\right)e_4^2 - \tfrac{k_3g_2}{2}e_0e_1 + \tfrac{k_3g_2}{2}e_0e_2 - \tfrac{k_4g_3}{2}e_0e_3\\ +\tfrac{k_4g_3}{2}e_0e_4 - \tfrac{k_3^2g_2^2}{2}e_1e_2 + \tfrac{k_3k_4g_2g_3}{4}e_1e_3 - \tfrac{k_3k_4g_2g_3}{4}e_1e_4 - \tfrac{k_3k_4g_2g_3}{4}e_2e_3\\ +\tfrac{k_3k_4g_2g_3}{4}e_2e_4 - \tfrac{k_4^2g_3^2}{2}e_3e_4\Big].\end{aligned} \tag{4.10}$$

Using (4.10), the bias of $\bar{y}_e^d$ is given by:

$$\begin{aligned}Bias\left(\bar{y}_e^d\right) \cong \theta\bar{Y}\Big[\left(\tfrac{k_3^2g_2^2}{8}+\tfrac{k_3g_2^2}{4}\right)C_x^2 + \left(\tfrac{k_4^2g_3^2}{8}+\tfrac{k_4g_3^2}{4}\right)C_r^2 - \tfrac{k_3g_2}{2}C_{yx} - \tfrac{k_4g_3}{2}C_{yr}\\ +\tfrac{k_3k_4g_2g_3}{2}C_{xr}\Big] - \theta'\bar{Y}\Big[\left(\tfrac{k_3^2g_2^2}{8}+\tfrac{k_3g_2^2}{4}\right)C_x^2 + \left(\tfrac{k_4^2g_3^2}{8}+\tfrac{k_4g_3^2}{4}\right)C_r^2\\ -\tfrac{k_3g_2}{2}C_{yx} - \tfrac{k_4g_3}{2}C_{yr} + \tfrac{k_3k_4g_2g_3}{2}C_{xr}\Big].\end{aligned} \tag{4.11}$$

By applying expectation after squaring both sides of (4.10), we derive the corresponding equation that expresses the MSE of $\bar{y}_e^d$.

$$\begin{aligned}MSE\left(\bar{y}_e^d\right) \cong \theta\bar{Y}^2\Big[C_y^2 + \tfrac{k_3^2g_2^2}{4}C_x^2 + \tfrac{k_4^2g_3^2}{4}C_r^2 - k_3g_2C_{yx} - k_4g_3C_{yr} + \tfrac{k_3k_4g_2g_3}{2}C_{xr}\Big]\\ -\theta'\bar{Y}^2\left[\tfrac{k_3^2g_2^2}{4}C_x^2 + \tfrac{k_4^2g_3^2}{4}C_r^2 - k_3g_2C_{yx} - k_4g_3C_{yr} + \tfrac{k_3k_4g_2g_3}{2}\right],\end{aligned} \tag{4.12}$$

To express the bias and MSE for $\bar{y}_e^d$, we substitute the known constants k_3 and k_4 into the formulas in (4.11) and (4.12). After simplifying the expressions, we get the following results:

$$Bias\left(\bar{y}_e^d\right) \cong \theta''\bar{Y}\left[\frac{3}{8}\left(g_2^2C_x^2 + g_3^2C_r^2\right) - \frac{1}{2}\left(g_2C_{yx} + g_3C_{yr} + g_2g_3C_{xr}\right)\right], \tag{4.13}$$

and

$$MSE\left(\bar{y}_e^d\right) \cong \bar{Y}^2\left[\theta C_y^2 + \frac{\theta''}{4}\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right)\right]. \tag{4.14}$$

5. Mathematical comparison

In this section, we provide the efficiency conditions by using the mean squared error equation of the proposed estimators with the mean squared error equations of existing estimators, such as $\bar{y}^d$, $\bar{y}^d_R$, $\bar{y}^d_{lr}$, $\bar{y}^d_{RMD}$, $\bar{y}^d_{PMD}$, and $\bar{y}^d_{RP}$.

5.1. Improved estimator-I

Condition (i): We derived the following results from (3.2) and (4.8) to ensure that the suggested estimator is at least as efficient:

$$V(\bar{y}^d) > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$\theta C_y^2 + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

Condition (ii): The following expression obtained from (3.5) and (4.8), provides the required condition for the proposed estimator-I to be more efficient:

$$MSE(\bar{y}^d_R) > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$\left(\theta C_y^2 + \theta'' C_x^2 - 2\theta'' C_{yx}\right) + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

Condition (iii): The following condition obtained from (3.11) and (4.8), which provides that proposed estimator-I has a smaller mean squared error:

$$MSE(\bar{y}^d_{lr}) > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$C_y^2\left(\theta - \theta''\rho^2_{yx}\right) + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

Condition (iv): The required condition for the proposed estimator-I, determined from (3.14) and (4.8), shows that the proposed estimator-I is more efficient:

$$MSE(\bar{y}^d_{RMD}) > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$\left[\theta C_y^2 + \theta'' C_x^2\left(\frac{1}{4} - C\right)\right] + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

Condition (v): The resulting expression derived from (3.15) and (4.8) defines the necessary condition for the proposed estimator-I to show better efficiency:

$$MSE(\bar{y}^d_{PMD}) > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$\left[\theta C_y^2 + \theta'' C_x^2\left(\frac{1}{4} + C\right)\right] + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

Condition (vi): The following inequality, derived from (3.18) and (4.8), indicates the effectiveness of the proposed estimator-I:

$$MSE(\bar{y}^d_{RP})_{\min} > MSE\left(\bar{y}^d_U\right)_{min}, \quad \text{if}$$

$$C_y^2\left[\theta(1 - \rho^2_{yx}) + \theta'\rho^2_{yx}\right] + \left(\frac{AG^2 + BD^2 - 2DFG}{AB - F^2}\right) > 1.$$

5.2. Improved estimator-II

Condition (vii): From (3.2) and (4.14), we derive:

$$V(\bar{y}^d) > MSE\left(\bar{y}^d_e\right), \quad \text{if}$$

$$(\theta - \theta')\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right) < 0.$$

If θ is greater than $\acute{\theta}$ or, equivalently, $\theta > \theta'$, the following inequality is satisfied:

$$\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right) < 0. \tag{5.1}$$

In a similar way, when θ is less than $\acute{\theta}$, the following inequality holds true:

$$\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right) > 0. \tag{5.2}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d)$ when either (5.1) or (5.2) holds.

Condition (viii): From (3.5) and (4.14), we derive:

$$MSE(\bar{y}^d_R) > MSE\left(\bar{y}^d_e\right), \quad \text{if}$$

$$4\left(\theta - \theta'\right)\left(C_x^2 - 2C_{yx}\right) > (\theta - \theta')\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right).$$

If θ is greater than $\acute{\theta}$ or, equivalently, $\theta > \theta'$, the following inequality is satisfied:

$$4\left(C_x^2 - 2C_{yx}\right) > \left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right). \tag{5.3}$$

In a similar way, when θ is less than $\acute{\theta}$, the subsequent inequality holds true:

$$4\left(C_x^2 - 2C_{yx}\right) < \left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right). \tag{5.4}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d_R)$ when either (5.3) or (5.4) holds.

Condition (ix): From (3.11) and (4.14), we derive

$$MSE(\bar{y}^d_{lr}) > MSE\left(\bar{y}^d_e\right), \quad \text{if}$$

$$4\left(\theta - \theta'\right)\rho^{*2}_{yx} < (\theta - \theta')\left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right).$$

If θ is greater than $\acute{\theta}$ or, equivalently, $\theta > \theta'$, the following inequality is satisfied:

$$4\rho^{*2}_{yx} > \left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right). \tag{5.5}$$

In a similar way, when θ is less than $\acute{\theta}$, the following inequality holds true:

$$4\rho^{*2}_{yx} < \left(g_2^2C_x^2 + g_3^2C_r^2 - g_2C_{yx} - g_3C_{yr} + 2g_2g_3C_{xr}\right). \tag{5.6}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d_{lr})$ when either (5.5) or (5.6) holds.

Condition (x): From (3.14) and (4.14), we derive

$$MSE(\bar{y}^d_{RMD}) > MSE\left(\bar{y}^d_e\right)$$

$$4\left(\theta-\theta'\right)C_x^2\left(\frac{1}{4}-C\right) > \left(\theta-\theta'\right)\left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right).$$

In a similar way, when θ is less than $\acute{\theta}$, the following inequality holds true:

$$C_x^2\left(1-4C\right) > \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.7}$$

In a similar way, when θ is less than $\acute{\theta}$, the subsequent inequality holds true:

$$C_x^2\left(1-4C\right) < \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.8}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d_{RMD})$ when either (5.7) or (5.8) holds.

Condition (xi): From (3.15) and (4.14), we derive

$$MSE(\bar{y}^d_{PMD}) > MSE\left(\bar{y}^d_e\right), \quad \text{if}$$

$$4\left(\theta-\theta'\right)C_x^2\left(\frac{1}{4}+C\right) > \left(\theta-\theta'\right)\left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right).$$

If θ is greater than $\acute{\theta}$ or, equivalently, $\theta > \theta'$, the following inequality is satisfied:

$$C_x^2\left(1+4C\right) > \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.9}$$

In a similar way, when θ is less than $\acute{\theta}$, the following inequality holds true:

$$C_x^2\left(1+4C\right) < \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.10}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d_{PMD})$ when either (5.9) or (5.10) holds.

Condition (xii): From 3.18 and (4.14), we derive:

$$MSE(\bar{y}^d_{RP})_{\min} > MSE\left(\bar{y}^d_e\right), \quad \text{if}$$

$$4\left(\theta-\theta'\right)\rho_{yx}^2 < \left(\theta-\theta'\right)\left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right).$$

If θ is greater than $\acute{\theta}$ or, equivalently, $\theta > \theta'$, the following inequality is satisfied:

$$4\rho_{yx}^2 > \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.11}$$

In a similar way, when θ is less than $\acute{\theta}$, the following inequality holds true:

$$4\rho_{yx}^2 < \left(g_2^2C_x^2+g_3^2C_r^2-g_2C_{yx}-g_3C_{yr}+2g_2g_3C_{xr}\right). \tag{5.12}$$

The proposed estimator $\bar{y}^d_e$ shows better performance relative to $MSE(\bar{y}^d_{RP})$ when either (5.11) or (5.12) holds.

6. Results of the numerical comparison

In this section, we compare the mean squared errors of the improved classes of estimators with those of existing estimators in order to provide a comprehensive analysis. This analysis relies on both simulated data and distinct real datasets. By analyzing the MSE across these simulated and different datasets, we aim to provide a thorough assessment of the performance and reliability of the new proposed estimators.

6.1. Simulation study

The simulation method from [28] is in this section, where the auxiliary variable X is collected from six populations, each associated with a different probability distribution.

- Data-1: $X \sim Gamma(\eta_1 = 4, \eta_2 = 6)$,
- Data-2: $X \sim Gamma(\eta_1 = 8, \eta_2 = 10)$,
- Data-3: $X \sim Uniform(v_1 = 3, v_2 = 5)$,
- Data-4: $X \sim Uniform(v_1 = 0, v_2 = 1)$,
- Data-5: $X \sim Exponential(\mu = 5)$,
- Data-6: $X \sim Exponential(\mu = 10)$.

Afterwards, by utilizing the correlation coefficient's fixed value

$$r_{yx} = 0.79,$$

and the random error term

$$e \sim N(0, 1),$$

the dependent variable Y is computed by using the following equation:

$$Y = r_{yx} \times X + e.$$

On the basis of each distribution and correlation setting, we examined the MSEs of various suggested estimators and existing estimators to assess the robustness and efficiency.

Step 1: In order to generate 1000 observations, we start to obtain a population based on the different probability distributions mentioned above.

Step 2: In the first phase, simple random sampling without replacement (SRSWOR) is used to choose a sub-sample m_1 from a whole population of size N.

Step 3: From the first-phase sample, the SRSWOR technique is again applied to draw a second-phase sub-sample with size m_2.

Step 4: The total population, as well as the extreme observations of independent variables and their rankings, are calculated using the processes described above. We also determine the optimal values for the improved classes of estimators by considering unknown constants.

Step 5: Different sample sizes are computed for each population using the SRSWOR method.

Step 6: For each sample size under consideration, the MSE are calculated for all the estimators examined in this paper.

Step 7: After 45,000 repetitions of Steps 5 and 6; the MSE values for all the estimators are calculated on the basis of the formula defined as follows:

$$MSE(\bar{y}_k^d)_{\min} = \frac{\sum_{g=1}^{45000}\left(\bar{y}_{kg}^d - \bar{Y}\right)^2}{45000}, k = R,\ lr,\ RMD,\ PMD,\ RP,\ e,\ U_1,\ U_2, \ldots, U_8.$$

Step 8: Finally, Table 2 reports the MSE results for artificial populations.

Table 2. MSE computation for various estimators with artificial populations.

Estimator	$Gam(4,6)$	$Gam(8,10)$	$Uni(0,1)$	$Uni(3,5)$	$Exp(5)$	$Exp(10)$
$\bar{y}^d$	1.92e-2	2.28e-2	8.98e-3	9.90e-2	8.26e-2	9.95e-2
$\bar{y}_R^d$	1.73e-2	2.29e-2	4.70e-3	8.09e-2	7.66e-2	9.48e-2
$\bar{y}_{lr}^d$	1.68e-2	1.91e-2	3.10e-3	7.58e-2	7.01e-2	9.33e-2
$\bar{y}_{RMD}^d$	1.48e-2	1.72e-2	2.98e-3	7.24e-2	6.79e-2	9.07e-2
$\bar{y}_{PMD}^d$	2.29e-2	2.48e-2	9.25e-3	1.04e-1	9.25e-2	1.01e-1
$\bar{y}_{RP}^d$	7.62e-3	1.59e-2	2.79e-3	6.28e-2	6.09e-2	8.57e-2
$\bar{y}_e^d$	7.18e-3	1.40e-2	2.41e-3	5.60e-2	5.51e-2	7.54e-2
$\bar{y}_{U_1}^d$	3.33e-3	4.42e-3	8.60e-4	3.28e-3	2.86e-3	5.58e-3
$\bar{y}_{U_2}^d$	2.45e-3	4.56e-3	4.00e-4	3.05e-3	2.16e-3	1.54e-2
$\bar{y}_{U_3}^d$	3.62e-3	5.34e-3	1.73e-3	3.88e-3	4.53e-3	2.77e-2
$\bar{y}_{U_4}^d$	3.16e-3	6.63e-3	6.20e-4	3.13e-3	4.18e-3	3.00e-2
$\bar{y}_{U_5}^d$	2.17e-3	4.85e-3	1.58e-3	3.55e-3	2.46e-3	4.14e-3
$\bar{y}_{U_6}^d$	9.90e-4	1.63e-3	2.60e-4	1.20e-4	1.33e-3	2.05e-3
$\bar{y}_{U_7}^d$	6.90e-4	1.74e-3	2.20e-4	3.70e-4	8.70e-4	3.48e-3
$\bar{y}_{U_8}^d$	9.12e-6	6.32e-5	2.42e-6	6.10e-6	1.67e-5	6.30e-4

6.2. Numerical examples

We assessed the performance of different estimators by calculating the mean squared errors across three separate datasets. The aim of this analysis was to check the accuracy of the proposed estimators. Detailed descriptions of the datasets, along with their summary statistics, are provided below.

Dataset 1. (Source: [29], p. 226)

Y: Information on the number of employees in each department for the year 2012, which indicates the size of the workforce across different industries.

X: The number of factories each department officially authorized in 2012b by providing information on the productivity of the department and the existence of the industry.

R: The departments are arranged according to the count of factories in 2012, used to compare relative industrial activity.

Dataset 2. (Source: [29], p. 135)

Y: Denotes the overall count of students enrolled in educational departments in 2012;

X: Denotes the total number of schools funded by the government in 2012;

R: Denotes the ranks of the total number of schools funded by the government in 2012 according to the number of schools.

Dataset 3. (Source: [1], p. 24)

Y: The food costs that the families paid for, which are directly related to their jobs;

X: The total weekly income that the families made, which shows their financial resources for that time; R: the ranking of families according to their weekly income, which shows how much money they made compared with each other.

The above data sets are summarized in the following summary statistics given in Table 3, and mean squared comparison between new proposed estimators and existing estimators are displayed in Table 4.

Table 3. Summary statistics for different data-sets.

Parameters	Dataset 1	Dataset 2	Dataset 3	Parameters	Dataset 1	Dataset 2	Dataset 3
N	36	36	33	$\bar{Y}$	52432	148718	27.49
m_1	15	15	15	$\bar{X}$	335.78	1054.39	72.55
m_2	6	6	6	$\bar{R}$	18.51	18.50	17.00
X_M	2055	2370	95	S_y	178201	182315	10.13
X_m	24	39	58	S_x	451.136	402.61	10.58
R_M	36	36	33	S_r	10.53	10.54	9.64
R_m	1	1	1	C_y	3.40	1.23	0.37
C_x	1.34	0.38	0.15	C_r	0.57	0.56	0.57
ρ_{yx}	0.39	0.17	0.25	ρ_{yr}	0.36	0.19	0.20
ρ_{xr}	0.75	0.94	0.98	θ	0.14	0.14	0.14
θ'	0.10	0.10	0.10	θ''	0.04	0.04	0.04

Table 4. MSE computation for various estimators with real populations.

Estimator	Dataset 1	Dataset 2	Dataset 3
$\bar{y}^d$	4410841204	4616729172	13.97
$\bar{y}^d_R$	3934722123	4566840232	13.53
$\bar{y}^d_{lr}$	3934517862	4509295381	13.32
$\bar{y}^d_{RMD}$	4048702250	4511078294	13.35
$\bar{y}^d_{PMD}$	5021109648	4883614743	15.40
$\bar{y}^d_{RP}$	3180960973	4018903439	11.39
$\bar{y}^d_{Ue}$	3025382610	3918150768	11.29
$\bar{y}^d_{U_1}$	354493387	422589708	2.66
$\bar{y}^d_{U_2}$	2175126	7346603	0.05
$\bar{y}^d_{U_3}$	350163172	214844883	0.72
$\bar{y}^d_{U_4}$	212556223	352690922	2.04
$\bar{y}^d_{U_5}$	275131170	403622267	2.64
$\bar{y}^d_{U_6}$	377505122	216979294	0.74
$\bar{y}^d_{U_7}$	117242958	280960015	1.51
$\bar{y}^d_{U_8}$	1345452	6937497	0.04

6.3. Discussion

To identify the performance and quality of the newly improved estimators, we conducted simulations and analyzed three real-life datasets. The performance comparison was made using the MSE as the key criterion. The MSE values for both the new and existing estimators are listed in Table 2, while the results for the real datasets are shown in Table 4. From these evaluations, the following conclusions can be drawn.

- Findings from both the generated simulations and the actual datasets indicate that the newly introduced estimators consistently yield lower MSE values compared with the overall existing approaches that have already been discussed in the literature. This pattern is clearly reflected in Tables 2 and 4.
- Additionally, the MSE values for the proposed estimators are consistently lower than those for the existing ones, as illustrated by the declining trend of the graph lines in Figures 1 and 3, both for the simulation experiments and real-life data. This observation suggests that the proposed estimators outperform the other ones, with the MSE values for the new estimators showing an inverse pattern compared with those of the existing estimators.
- Figure 2 presents a grouped bar chart comparing the MSE of the proposed estimator across various artificial populations generated from gamma, uniform, and exponential distributions. Each group on the x-axis represents a distribution type, while the two bars within each group correspond to different parameter settings for that distribution (e.g., $\text{Gam}(4, 6)$ and $\text{Gam}(8, 10)$ for the gamma distribution). The y-axis shows the MSE values, with lower bars indicating better performance for that estimator. This visualization helps highlight how the estimator performs under different distributional shapes and scales, offering a clear comparison of its efficiency across varied population structures.
- Figure 4 shows a grouped bar chart of the MSE values for 15 estimators across three real datasets. The x-axis indexes the estimators from 1 to 15, while the y-axis uses a logarithmic scale to handle the large variation in MSE values particularly highlighting the contrast between the relatively large MSE values of Dataset 1 and Dataset 2 and the smaller values from Dataset 3. Each group of bars represents an estimator, with different colors indicating the datasets. This visualization allows for easy comparison of estimator performance across datasets, highlighting how some estimators perform consistently better, especially on Dataset 3. From this plot, it is evident that certain estimators (such as Estimators 14 and 15) consistently yield lower MSE values.

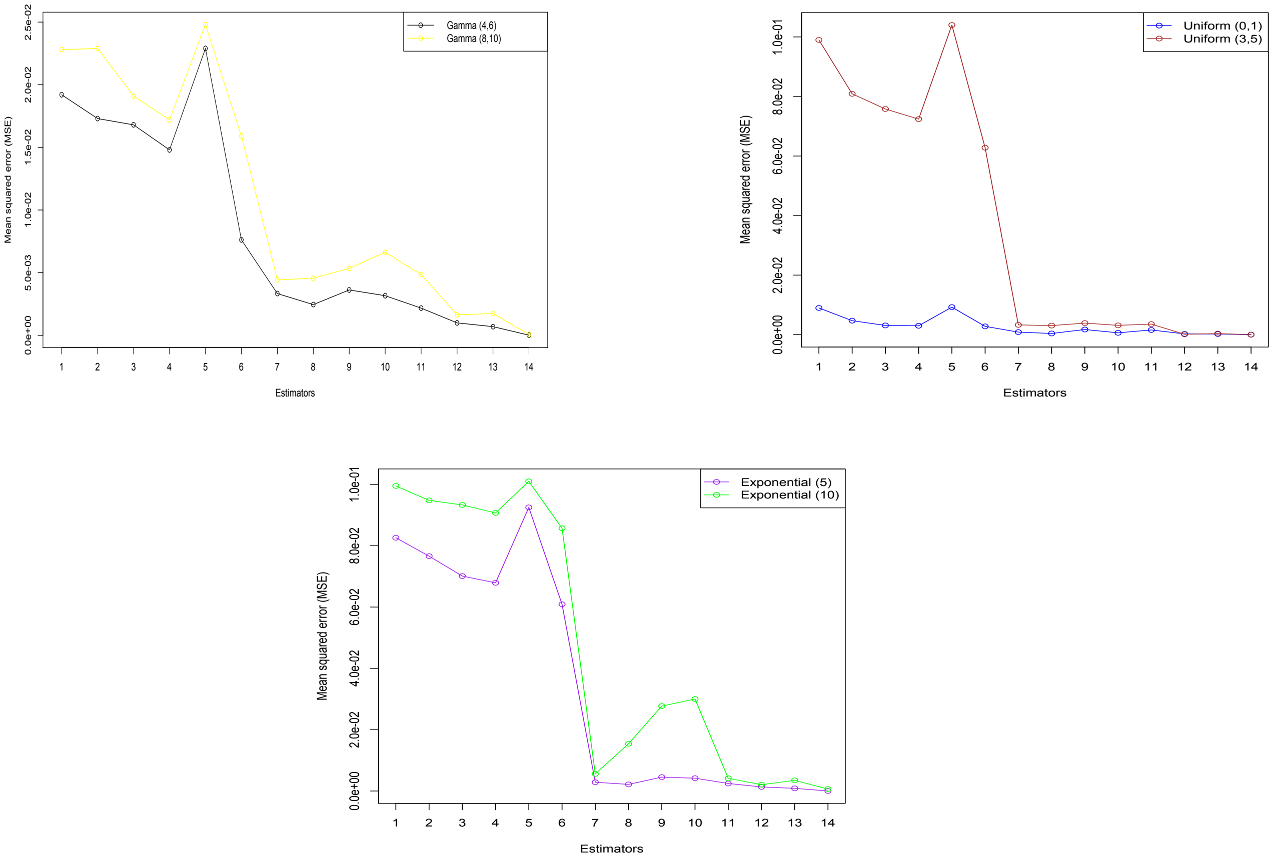

Figure 1. A visual representation of the mean squared errors (MSEs) results for all suggested and existing estimators using artificial data.

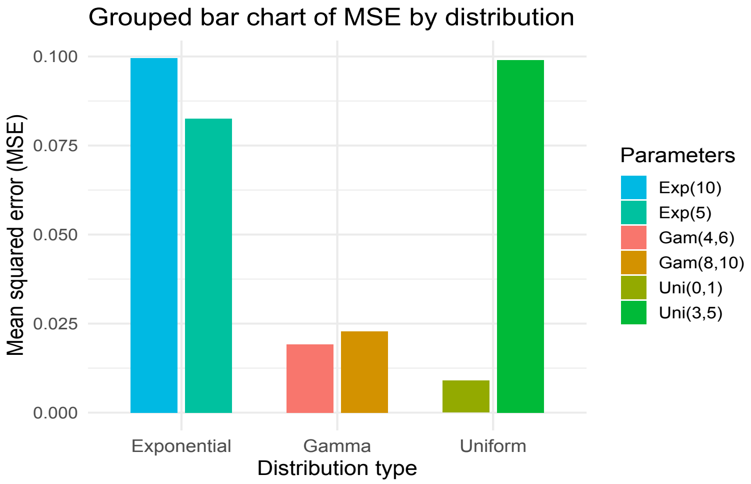

Figure 2. Visual representation of the mean squared errors (MSEs) for all suggested and existing estimators using artificial data.

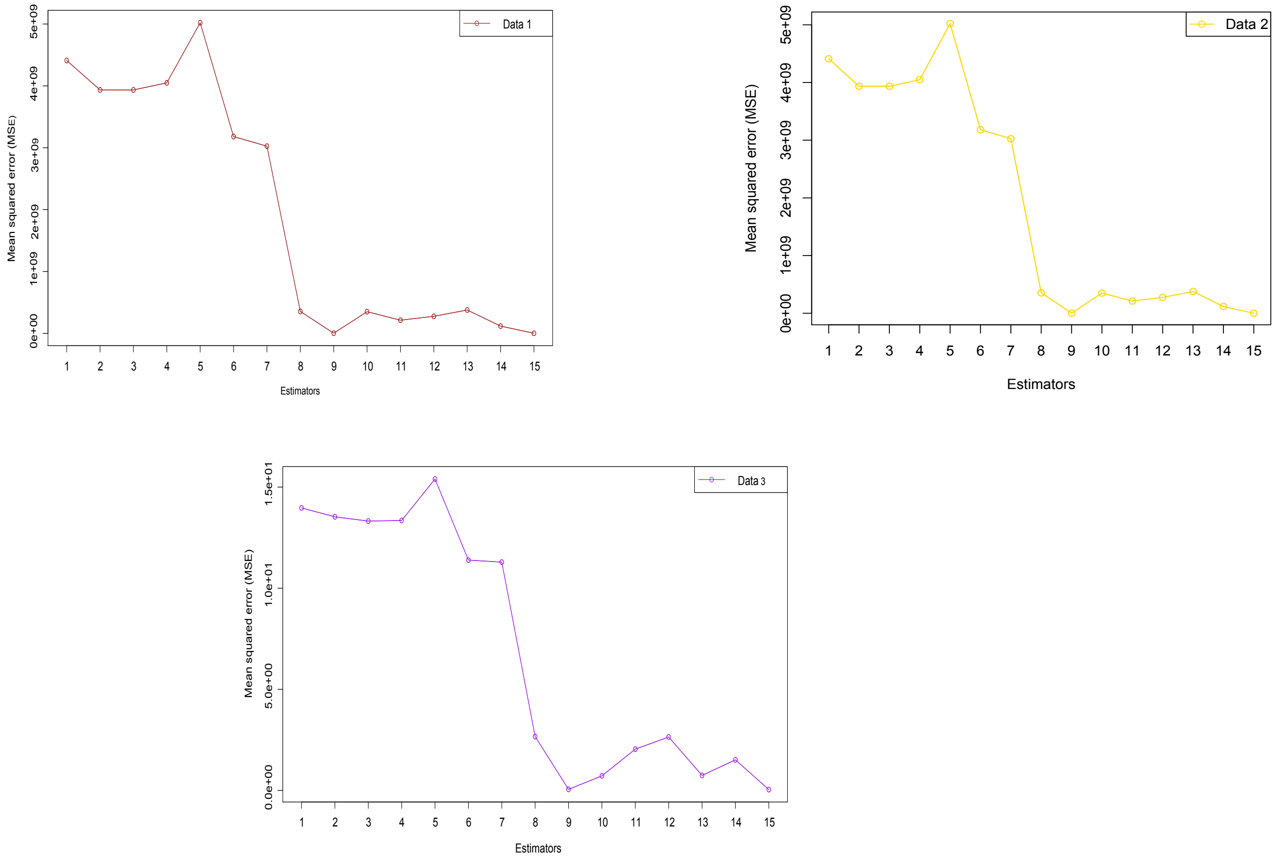

Figure 3. A visual representation of the mean squared errors (MSEs) results for all suggested and existing estimators using real data.

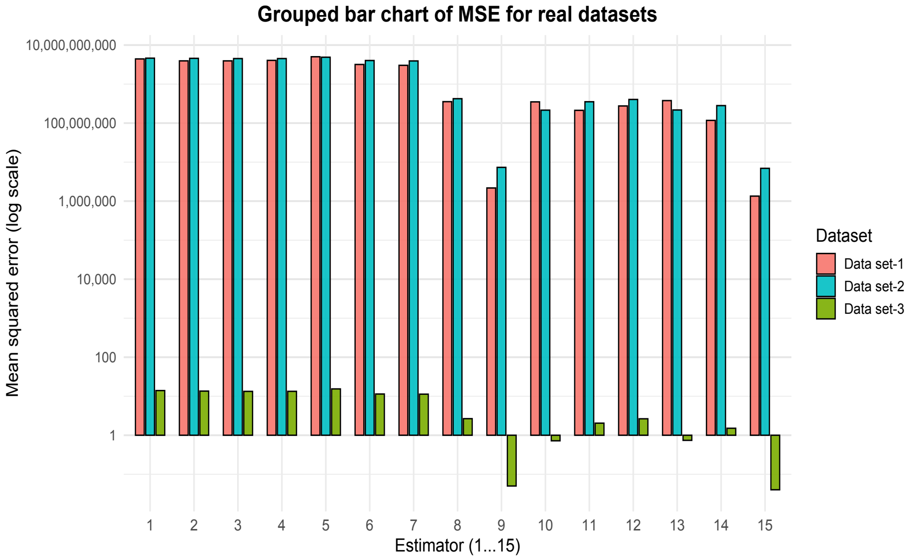

Figure 4. Visual representation of the mean squared errors (MSEs) for all suggested and existing estimators using real data.

7. Conclusions

This research introduced innovative sets of efficient newly enhanced estimators for determining the population mean, utilizing the minimum and largest values of the independent variable along with its known highest and lowest rank values. Section 5 outlines specific theoretical foundations to compare the proposed estimators with existing methods, demonstrating their effectiveness. To support these findings, simulation experiments were performed, along with an analysis of real-life datasets. The results show that the new estimators consistently perform better than the existing estimators, especially by having a lower mean squared error (MSE), as displayed in Table 2. These findings are also supported by the results in Table 4, which validates the theoretical conditions given in Section 5.

The simulation results and empirical studies clearly show that the suggested estimators $\bar{y}_i^d$ ($i = e, U_1, U_2, \ldots, U_8$) are more efficient than the other estimators being considered. We noticed that all the proposed estimators, $\bar{y}_{U_8}^d$ proves to be the most efficient option within the suggested group of estimators, making it strongly recommended.

This study introduced an enhanced method for population mean estimation by utilizing the smallest and largest values of the independent variables along with their ranks. The main advantage of this approach lies in its ability to significantly reduce the MSE, particularly in the presence of extreme values, offering improved precision over traditional methods. This method is valuable in fields such as economics, healthcare, public policy, and education, where accurate population estimates are essential. However, the method has some limitations. Its performance may vary, depending on the characteristics of the data, such as distribution and correlation, and the two-phase sampling design may add complexity and cost.

Moreover, our study examined the qualities and efficiency of the newly enhanced estimators by adopting a sampling two-phase technique. This approach allowed us to compare the performance of the enhanced estimators against traditional methods more effectively. Future research may explore the application of these methods in various sampling designs, including stratified and systematic approaches, and investigate optimal sample size allocation strategies to further improve efficiency. The use of additional auxiliary variables and assessment on large-scale datasets may provide deeper insights. Exploring the combination of these estimators with machine learning for adaptive sampling or real-time data processing can enhance estimation precision in dynamic environments. This topic offers an interesting direction for further investigation.

Author contributions

Hleil Alrweili: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing-original draft preparation, writing-review and editing, visualization, project administration; Fatimah A. Almulhim: Conceptualization, methodology, validation, formal analysis, investigation, data curation, writing-original draft preparation, writing-review and editing, visualization, supervision, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Use of Generative-AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This project was funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project. The authors, therefore, acknowledge with thanks Princess Nourah bint Abdulrahman University Researchers Supporting funds, Project number (PNURSP2025R515), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Conflict of interest

The authors declare no conflict of interest.

References

1. W. G. Cochran, *Sampling techniques*, 2 Eds., Hoboken: John Wiley and Sons, 1963.
2. C. E. Särndal, Sample survey theory vs. general statistical theory: estimation of the population mean, *Int. Stat. Rev.*, **40** (1972), 1–12. https://doi.org/10.2307/1402700
3. S. Y. Yang, D. B. Meng, H. F. Yang, C. Q. Luo, X. Y. Su, Enhanced soft Monte Carlo simulation coupled with support vector regression for structural reliability analysis, *P. I. Civil Eng.-Transp.*, **2024** (2024), 1–16. https://doi.org/10.1680/jtran.24.00128
4. B. V. Sukhatme, Some ratio-type estimators in two-phase sampling, *J. Am. Stat. Assoc.*, **57** (1962), 628–632. https://doi.org/10.2307/2282400
5. A. Y. Erinola, R. V. K. Singh, A. Audu, T. James, Modified class of estimator for finite population mean under two-phase sampling using regression estimation approach, *Asian Journal of Probability and Statistics*, **14** (2021), 52–64. https://doi.org/10.9734/ajpas/2021/v14i430338
6. S. Guha, H. Chandra, Improved estimation of finite population mean in two-phase sampling with subsampling of the nonrespondents, *Math. Popul. Stud.*, **28** (2021), 24–44. https://doi.org/0.1080/08898480.2019.1694325
7. A. Kamal, N. Amir, H. Dastagir, Some exponential type predictive estimators of finite population mean in two-phase sampling, *Journal of Statistics, Computing and Interdisciplinary Research*, **2** (2020), 51–57. https://doi.org/10.52700/scir.v2i1.10
8. T. Zaman, C. Kadilar, New class of exponential estimators for finite population mean in two-phase sampling, *Commun. Stat.-Theor. M.*, **50** (2019), 874–889. https://doi.org/10.1080/03610926.2019.1643480
9. S. Mohanty, J. Sahoo, A note on improving the ratio method of estimation through linear transformation using certain known population parameters, *Sankhya: The Indian Journal of Statistics*, **57** (1995), 93–102.

10. M. Khan, J. Shabbir, Some improved ratio, product, and regression estimators of finite population mean when using minimum and maximum values, *Sci. World J.*, **2013** (2013), 431868. https://doi.org/10.1155/2013/431868

11. M. Khan, Improvement in estimating the finite population mean under maximum and minimum values in double sampling scheme, *J. Stat. Appl. Pro. Lett.*, **2** (2015), 115–121.

12. G. S. Walia, H. Kaur, M. K. Sharma, Ratio type estimator of population mean through efficient linear transformation, *American Journal of Mathematics and Statistics*, **5** (2015), 144–149. https://doi.org/10.5923/j.ajms.20150503.06

13. U. Daraz, J. Shabbir, H. Khan, Estimation of finite population mean by using minimum and maximum values in stratified random sampling, *J. Mod. Appl. Stat. Meth.*, **17** (2018), eP2548. https://doi.org/10.22237/jmasm/1532007537

14. U. Daraz, M. Khan, Estimation of variance of the difference-cum-ratio-type exponential estimator in simple random sampling, *RMS Res. Math. Stat.*, **8** (2021), 1899402. https://doi.org/10.1080/27658449.2021.1899402

15. U. Daraz, M. A. Alomair, O. Albalawi, Variance estimation under some transformation for both symmetric and asymmetric data, *Symmetry*, **16** (2024), 957. https://doi.org/10.3390/sym16080957

16. A. S. Alghamdi, H. Alrweili, A comparative study of new ratio-type family of estimators under stratified two-phase sampling, *Mathematics*, **13** (2025), 327. https://doi.org/10.3390/math13030327

17. U. Daraz, J. B. Wu, D. Agustiana, W. Emam, Finite population variance estimation using Monte Carlo simulation and real life application, *Symmetry*, **17** (2025), 84. https://doi.org/10.3390/sym17010084

18. U. Daraz, D. Agustiana, J. B. Wu, W. Emam, Twofold auxiliary information under two-Phase sampling: An improved family of double-transformed variance estimators, *Axioms*, **14** (2025), 64. https://doi.org/10.3390/axioms14010064

19. C. X. Long, W. X. Chen, R. Yang, D. S. Yao, Ratio estimation of the population mean using auxiliary information under the optimal sampling design, *Probab. Eng. Inform. Sc.*, **36** (2022), 449–460. https://doi.org/10.1017/S0269964820000625

20. H. O. Cekim, H. Cingi, Some estimator types for population mean using linear transformation with the help of the minimum and maximum values of the auxiliary variable, *Hacet. J. Math. Stat.*, **46** (2017), 685–694. https://doi.org/10.15672/HJMS.201510114186

21. S. Chatterjee, A. S. Hadi, *Regression analysis by example*, 5 Eds., Hoboken: John Wiley & Sons, 2013.

22. A. S. Alghamdi, H. Alrweili, New class of estimators for finite population mean under stratified double phase sampling with simulation and real-life application, *Mathematics*, **13** (2025), 329. https://doi.org/10.3390/math13030329

23. H. P. Singh, G. K. Vishwakarma, Modified exponential ratio and product estimators for finite population mean in double sampling, *Aust. J. Stat.*, **36** (2007), 217–225.

24. H. P. Singh, M. R. Espeio, Double sampling ratio-product estimator of a finite population mean in sample surveys, *J. Appl. Stat.*, **34** (2007), 71–85. https://doi.org/10.1080/02664760600994562

25. U. Daraz, J. B. Wu, O. Albalawi, Double exponential ratio estimator of a finite population variance under extreme values in simple random sampling, *Mathematics*, **12** (2024), 1737. https://doi.org/10.3390/math12111737

26. U. Daraz, J. B. Wu, M. A. Alomair, L. A. Aldoghan, New classes of difference cum-ratio-type exponential estimators for a finite population variance in stratified random sampling, *Heliyon*, **10** (2024), e33402. https://doi.org/10.1016/j.heliyon.2024.e33402

27. M. A. Alomair, U. Daraz, Dual transformation of auxiliary variables by using outliers in stratified random sampling, *Mathematics*, **12** (2024), 2839. https://doi.org/10.3390/math12182839

28. U. Daraz, M. A. Alomair, O. Albalawi, A. S. Al Naim, New techniques for estimating finite population variance using ranks of auxiliary variable in two-stage sampling, *Mathematics*, **12** (2024), 2741. https://doi.org/10.3390/math12172741

29. *Punjab Development Statistics*, Bureau of Statistics, Government of the Punjab: Lahore, Pakistan, 2013. Available from: `https://bos.punjab.gov.pk`.

AIMS Press