

Article

A Novel Robust Transformation Approach to Finite Population Median Estimation Using Monte Carlo Simulation and Empirical Data

Huda M. Alshanbari 

Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdulrahman University,
P.O. Box 84428, Riyadh 11671, Saudi Arabia; hmalshanbari@pnu.edu.sa

Abstract

This study develops an improved family of estimators for estimating the finite population median within a two-phase sampling method. The proposed estimators, which use transformation techniques to reduce survey costs when full auxiliary information is unavailable, yield more accurate results than traditional methods. These transformations employ robust statistical measures such as Hodges–Lehmann location, Gini mean difference, and Bowley’s skewness, which strengthen resistance against outliers and heavy-tailed distributions. Through the use of these modern tools within the two-phase sampling framework, the proposed estimators achieve greater flexibility and robustness compared to conventional quantile-based approaches. A first-order approximation is employed to derive the bias and mean squared error expressions. The performance of the proposed estimators is examined through simulation experiments across multiple distributional scenarios and validated using real datasets against standard approaches. Findings based on percent relative efficiency confirm that the proposed estimators improve the accuracy and efficiency of median estimation in two-phase sampling, demonstrating superiority over conventional methods across various practical scenarios.

Keywords: two-phase sampling; transformation-based estimators; Monte Carlo simulation; bias; percent relative efficiency

MSC: 62D05



Academic Editor: Sat Gupta

Received: 28 August 2025

Revised: 18 September 2025

Accepted: 25 September 2025

Published: 29 September 2025

Citation: Alshanbari, H.M. A Novel Robust Transformation Approach to Finite Population Median Estimation Using Monte Carlo Simulation and Empirical Data. *Axioms* **2025**, *14*, 737. <https://doi.org/10.3390/axioms14100737>

Copyright: © 2025 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In survey methodology, the median is widely recognized as a reliable measure of central tendency, especially under skewed distributions or when outliers are present. The stratified two-phase sampling design enhances the efficiency of such estimation by first obtaining auxiliary information within each stratum and then selecting a sub-sample in the second phase for detailed measurement. This arrangement is useful when full auxiliary data are initially absent, since it reduces costs while preserving accuracy. As a result, the combined use of stratification and two-phase sampling provides a practical framework for precise and dependable median estimation in complex populations. More details about auxiliary information can be found at [1–4].

In many areas of applied statistics, the median has proven to be a more reliable summary measure than the mean, particularly when populations are skewed or contain a high frequency of extreme values. Its robustness has made it valuable in fields such as economic surveys, medical outcome analysis, and environmental monitoring, where averages may

be misleading. A key aspect of improving median estimation has been the use of auxiliary variables, which allow researchers to reduce cost while increasing precision. Early studies provided the theoretical groundwork [5–7], followed by the introduction of improved estimators such as ratio and regression forms [8,9]. As the discipline advanced, methods incorporating two-phase or double-sampling designs, as well as approaches utilizing multiple auxiliary variables, further enhanced efficiency [10–13]. More recent contributions emphasize the role of robust, transformation-based estimators under both simple and stratified designs [14–18], reflecting the ongoing importance of median estimation in survey methodology [19–25].

In many real-world applications, survey data are rarely symmetric and often include extreme observations. Agricultural yields fluctuate due to weather patterns, industrial outputs may be distorted by production defects, and educational test scores frequently deviate from normality. In such contexts, the median provides a more stable and reliable indicator of central tendency than the mean. However, traditional estimators such as ratio, regression, product, and exponential types were mostly developed under the assumption of normality and symmetry. As a result, they are highly sensitive to skewness and outliers, leading to inefficiency in complex populations. Motivated by the limitations of traditional approaches and supported by recent contributions by employing the methodologies outlined in [26–28], we develop a new class of estimators under a stratified two-phase sampling design. Stratification ensures that population heterogeneity is properly represented, while the two-phase design allows auxiliary information from the first phase to be utilized before more detailed data are collected in the second phase. By applying transformation techniques, the proposed estimators achieve improved robustness and precision, reducing sensitivity to irregularities in the data. The combined use of stratification, two-phase sampling, and transformations ensures higher efficiency and greater consistency, making the proposed estimators well suited for practical survey applications.

Rationale for the Proposed Class of Estimators

In addition to the conventional sample median, a number of robust summary measures have been proposed to capture distributional features such as location, scale, skewness, and tail behavior. These measures serve as useful transformation components that can improve the efficiency and reliability of median-based estimation, particularly in the presence of outliers or departures from normality. The motivation for introducing this new family of estimators with robust transformations can be summarized as follows:

- **Handling outliers and heavy tails:** Conventional median-type estimators often lose efficiency when the auxiliary variable is affected by extreme values. Robust transformations such as Hodges–Lehmann location and Gini mean difference, and resistant averages such as trimmed and winsorized means provide stable adjustments that reduce the influence of outliers.
- **Quartile-based measures:** Median ratio, Bowley’s skewness, interquartile range, geometric quartile mean, quartile deviation, and percentile ranges describe spread and asymmetry.
- **Robust dispersion:** The median absolute deviation and skewness-adjusted indices help reduce the influence of extreme values.
- **Variability and shape:** The coefficient of variation and Moors’ kurtosis provide insight into variability and tail behavior, especially in skewed or heavy-tailed data.
- **Improved efficiency:** By combining resistant measures of location, scale, and shape, the proposed class can achieve lower bias and mean squared error compared to traditional quantile-based methods.

- **Novelty in median estimation:** While robust measures such as trimmed means or Gini mean difference have been studied in the context of mean estimation, their use in designing median estimators under two-phase sampling has not been explored. This gives the proposed class originality and adds value to the literature.
- **Practical Relevance:** Many real-life datasets in economics, health, and social sciences exhibit skewness or contain irregular observations. The proposed transformations make the estimators more reliable for such applications.
- **Overall contribution:** Together, these measures and transformations enhance the efficiency, stability, and reliability of median-based estimation in the presence of outliers or non-normality.

2. Survey Design and Preliminaries

A comprehensive discussion of the methods used for the proposed median class of estimators and those previously introduced by various authors is presented in this section. Consider a population of N members, written as $T = \{T_1, T_2, \dots, T_N\}$. The research investigates a study variable Y along with an auxiliary variable X . From this population, a simple random sample without replacement (SRSWOR) of size n is selected, producing observed values y_i and x_i for $i = 1, \dots, n$. While Y and X are strongly related, the median Φ_x for X in the population remains unknown. The stages of the two-phase sampling method are specified as follows.

- The process begins with selecting a sample of m_1 elements. At this stage, information is collected only on the auxiliary factor x , which is then used to approximate the median value Φ_x of the population.
- A second-phase sub-sample of size m_2 ($m_2 < m_1$) is taken from the first-phase selection. At this stage, data are collected on both the main study characteristic y and the auxiliary factor x .

Population medians of the study and auxiliary variables are represented by Φ_y and Φ_x . From the first-phase data, the median is written as ϕ'_x , while the second-phase produces the estimators $\hat{\phi}_y$ and $\hat{\phi}_x$. The probability density functions for these population medians are denoted by $f_y(\Phi_y)$ and $f_x(\Phi_x)$. The relationship between Φ_y and Φ_x is quantified by the correlation coefficient ρ_{yx} , given by:

$$\rho(\Phi_y, \Phi_x) = 4P_{11}(y, x) - 1,$$

where P_{11} represents

$$P(y \leq \Phi_y \cap x \leq \Phi_x).$$

The following expressions for relative error and corresponding expected values serve as the basis for first-order approximations of biases and mean squared errors:

$$\phi_0 = \left(\frac{\hat{\phi}_y - \Phi_y}{\Phi_y} \right), \quad \phi_1 = \left(\frac{\hat{\phi}_x - \Phi_x}{\Phi_x} \right), \quad \phi_2 = \left(\frac{\hat{\phi}_x - \Phi_x}{\Phi_x} \right),$$

such that $E(\phi_i) = 0$ for $i = 0, 1, 2$.

Moreover,

$$E(\phi_0^2) = \lambda_1 \Phi_{C_y}^2,$$

$$E(\phi_1^2) = \lambda_1 \Phi_{C_x}^2,$$

$$E(\phi_2^2) = \lambda_2 \Phi_{C_x}^2,$$

$$E(\phi_0\phi_1) = \lambda_1\Phi_{C_{yx}} = \rho_{yx}\Phi_{C_y}\Phi_{C_x},$$

$$E(\phi_0\phi_2) = \lambda_2\Phi_{C_{yx}} = \rho_{yx}\Phi_{C_y}\Phi_{C_x},$$

$$E(\phi_1\phi_2) = \lambda_2\Phi_{C_x}^2,$$

where

$$\Phi_{C_y} = \frac{1}{\Phi_y f_y(\Phi_y)},$$

$$\Phi_{C_x} = \frac{1}{\Phi_x f_x(\Phi_x)},$$

are the mathematical representations of the population coefficients of variation for the medians of Y and X , and

$$\lambda_1 = \frac{1}{4} \left(\frac{N - m_2}{m_2 \times N} \right),$$

$$\lambda_2 = \frac{1}{4} \left(\frac{N - m_1}{m_1 \times N} \right),$$

$$\lambda_3 = \frac{1}{4} \left(\frac{m_1 - m_2}{m_1 \times m_2} \right).$$

3. Existing Approaches for Median Estimation

The median estimators outlined below are reviewed and will later serve as benchmarks for comparison prior to presenting the proposed class of estimators. These estimators, developed by various researchers, along with their bias, variance, and mean squared error expressions are presented as follows:

The usual sample median estimator and its variance formula proposed by [5] are presented below.

$$\hat{\phi}_{GR} = \hat{\phi}_y$$

and

$$Var(\hat{\phi}_y) = \lambda_1 \Phi_y^2 \Phi_{C_y}^2. \quad (1)$$

In the context of two-phase sampling, [10] introduced a ratio-based estimator designed to enhance estimation efficiency, defined as:

$$\hat{\phi}_{SA} = \left(\frac{\hat{\phi}_y}{\hat{\phi}_x} \right) \hat{\phi}_x.$$

The first-order approximations for the bias and mean squared error (MSE) of $\hat{M}_{SA}$ are expressed as:

$$Bias(\hat{\phi}_{SA}) \cong \lambda_3 \Phi_y (\Phi_{C_x}^2 - \Phi_{C_{yx}}) \quad (2)$$

and

$$MSE(\hat{\phi}_{SA}) \cong \Phi_y^2 \left[\lambda_1 \Phi_{C_y}^2 + \lambda_3 (\Phi_{C_x}^2 - 2\Phi_{C_{yx}}) \right]. \quad (3)$$

According to the formulation in [12], the difference-type estimator $\hat{\phi}_S$ is defined as:

$$\hat{\phi}_S = \hat{\phi}_y + d_1 (\hat{\phi}_x - \hat{\phi}_y).$$

Using the first-order approximation, the minimum mean squared error expression for $\hat{\phi}_S$ at the optimum value of d_1 is given by:

$$MSE(\hat{\phi}_S)_{min} \cong \Phi_y^2 \Phi_{C_y}^2 (\lambda_1 - \lambda_3 \rho_{yx}^2), \quad (4)$$

where

$$d_{1(opt)} = \frac{\rho_{yx}\Phi_y\Phi_{C_y}}{\Phi_x\Phi_{C_x}}.$$

Under two-phase sampling, the median forms of the exponential ratio and product estimators introduced by [29] are presented below:

$$\hat{\phi}_{Re} = \hat{\phi}_y \exp\left(\frac{\hat{\phi}_x - \hat{\phi}_x}{\hat{\phi}_x + \hat{\phi}_x}\right)$$

and

$$\hat{\phi}_{Pe} = \hat{\phi}_y \exp\left(\frac{\hat{\phi}_x - \hat{\phi}_x}{\hat{\phi}_x + \hat{\phi}_x}\right).$$

At the first-order level of approximation, the expressions for the biases and mean squared errors of $(\hat{\phi}_{Re}, \hat{\phi}_{Pe})$ are given below:

$$Bias(\hat{\phi}_{Re}) \cong \frac{\lambda_3}{8}\Phi_y(3\Phi_{C_x}^2 - 4\Phi_{C_{yx}}), \quad (5)$$

$$Bias(\hat{\phi}_{Pe}) \cong \frac{\lambda_3}{8}\Phi_y(3\Phi_{C_x}^2 + 4\Phi_{C_{yx}}), \quad (6)$$

$$MSE(\hat{\phi}_{Re}) \cong \frac{\Phi_y^2}{4} [4\lambda_1\Phi_{C_y}^2 + \lambda_3\Phi_{C_x}^2(1 - 4V)] \quad (7)$$

and

$$MSE(\hat{\phi}_{Pe}) \cong \frac{\Phi_y^2}{4} [4\lambda_1\Phi_{C_y}^2 + \lambda_3\Phi_{C_x}^2(1 + 4V)], \quad (8)$$

where

$$V = \frac{\rho_{yx}\Phi_{C_y}}{\Phi_{C_x}}.$$

The difference-based estimators for median estimation proposed by [9,13] are defined as follows:

$$\hat{\phi}_R = d_2\hat{\phi}_y + d_3(\hat{\phi}_x - \hat{\phi}_x),$$

$$\hat{\phi}_{G_1} = [d_4\hat{\phi}_y + d_5(\hat{\phi}_x - \hat{\phi}_x)] \left(\frac{\hat{\phi}_x}{\hat{\phi}_x} \right),$$

$$\hat{\phi}_{G_2} = [d_6\hat{\phi}_y + d_7(\hat{\phi}_x - \hat{\phi}_x)] \left(\frac{\hat{\phi}_x - \hat{\phi}_x}{\hat{\phi}_x + \hat{\phi}_x} \right).$$

The optimal values of d_i ($i = 2, 3, \dots, 7$) yield the following first-order approximation formulas for the minimum biases and mean squared errors of the estimators $(\hat{\phi}_R, \hat{\phi}_{G_1}, \hat{\phi}_{G_2})$:

$$Bias(\hat{\phi}_R)_{min} \cong -\Phi_y \left[\frac{\lambda_1\Phi_y^2\Phi_{C_x}^2 - \lambda_3\Phi_{C_{yx}}^2}{\Phi_x^2(1 + \lambda_1\Phi_{C_y}^2) - \lambda_3\Phi_{C_{yx}}^2} \right], \quad (9)$$

$$Bias(\hat{\phi}_{G_1})_{min} \cong \Phi_y(d_{4(opt)} - 1) + \lambda_3[d_{4(opt)}\Phi_y(\Phi_{C_x}^2 - \Phi_{C_{yx}}) + d_{5(opt)}\Phi_x\Phi_{C_x}^2], \quad (10)$$

$$Bias(\hat{\phi}_{G_2})_{min} \cong \Phi_y \left[-1 + \frac{\lambda_3}{4}\Phi_{C_x}^2 \left\{ -1 + d_{6(opt)} \left(\frac{\Phi_x + 8\rho_{yx}\Phi_y}{\Phi_x} \right) \right\} \right], \quad (11)$$

$$MSE(\hat{\phi}_R)_{min} \cong \Phi_y^2 \left[\frac{\lambda_1\Phi_{C_y}^2\Phi_{C_x}^2 - \lambda_3\Phi_{C_{yx}}^2}{\Phi_{C_x}^2(1 + \lambda_1\Phi_{C_y}^2) - \lambda_3\Phi_{C_{yx}}^2} \right], \quad (12)$$

$$MSE(\hat{\phi}_{G_1})_{\min} \cong Bias(\hat{\phi}_{G_1})_{\min}^2 + \lambda_1 \Phi_y^2 d_{4(opt)}^2 \Phi_{C_y}^2 + \lambda_3 \left[\Phi_{C_x}^2 (d_{4(opt)} \Phi_y + d_{5(opt)} \Phi_x)^2 - 2d_{4(opt)} \Phi_y \Phi_{C_{yx}} (d_{4(opt)} \Phi_y + d_{5(opt)} \Phi_x) \right] \quad (13)$$

and

$$MSE(\hat{\phi}_{G_2})_{\min} \cong Bias(\hat{\phi}_{G_2})_{\min} + \frac{\lambda_3}{4} d_{6(opt)}^2 \Phi_y^2 \Phi_{C_x}^2, \quad (14)$$

where

$$\begin{aligned} d_{2(opt)} &= \frac{\Phi_{C_x}^2}{\Phi_{C_x}^2 (1 + \lambda_1 \Phi_{C_y}^2) - \lambda_3 \Phi_{C_{yx}}^2}, \\ d_{3(opt)} &= \frac{\Phi_y \Phi_{C_{yx}}}{\Phi_x [\Phi_{C_x}^2 (1 + \lambda_1 \Phi_{C_y}^2) - \lambda_3 \Phi_{C_{yx}}^2]}, \\ d_{4(opt)} &= \left[\frac{\Phi_{C_x}^2}{\Phi_{C_x}^2 (1 + \lambda_1 \Phi_{C_y}^2 \{1 + \lambda_3 \Phi_{C_x}^2\}) + \lambda_3 \Phi_{C_{yx}}^2 (1 + \Phi_{C_x}^2)} \right], \\ d_{5(opt)} &= \Phi_y \left[\frac{\Phi_{C_x}^2 (\lambda_1 \Phi_{C_y}^2 - 1) + \Phi_{C_{yx}} (1 - \lambda_3 \Phi_{C_{yx}})}{\Phi_{C_x}^2 (1 + \lambda_1 \Phi_{C_y}^2 \{1 + \lambda_3 \Phi_{C_x}^2\}) + \lambda_3 \Phi_{C_{yx}}^2 (1 + \Phi_{C_x}^2)} \right], \\ d_{6(opt)} &= \frac{1}{8} \left[\frac{8 - \lambda_2 \Phi_{C_x}^2}{1 + \lambda_1 \Phi_{C_x}^2 (1 - \rho_{yx}^2)} \right], \end{aligned}$$

and

$$d_{7(opt)} = \frac{\Phi_y}{\Phi_x} \left[\frac{1}{2} + d_{6(opt)} \left(\frac{\rho_{yx} \Phi_y}{\Phi_x} \right) - 1 \right].$$

4. Proposed Class of Robust Median Estimators

Inspired by the modified forms of estimators proposed by employing the methodologies outlined in [26–28], we introduce a generalized double exponential class of median estimators for two-phase sampling. This approach utilizes various transformations of the auxiliary variable to estimate the finite population median. The new family of estimators is presented below:

$$\hat{\phi}_V = \hat{\phi}_y \exp \left[\frac{\tau_1 (\hat{\phi}_x - \hat{\phi}_x)}{\tau_1 (\hat{\phi}_x + \hat{\phi}_x) + 2\tau_2} \right] \exp \left[\frac{\tau_3 (\hat{\phi}_x - \hat{\phi}_x)}{\tau_3 (\hat{\phi}_x + \hat{\phi}_x) + 2\tau_4} \right], \quad (15)$$

where the symbols $\tau_1, \tau_2, \tau_3, \tau_4$ denote known population quantities related to the auxiliary variable X . In practice, they are computed from the first-phase sample where X is observed, but for theoretical derivations they may be computed from the full population. Starting from Equation (15), we introduce new estimators by selecting different combinations of these population quantities; the specific settings appear in Table 1.

The specific choices of τ_1 – τ_4 in Table 1 are motivated by their robustness and ability to capture essential features of the auxiliary variable X . For instance, measures such as the Hodges–Lehmann estimator, trimmed mean, and Winsorized mean provide central tendency estimates that are less sensitive to outliers, while quantities like the interquartile range, median absolute deviation (MAD), and skewness adjustments reflect the spread and asymmetry of the distribution. By combining these transformations in the proposed exponential ratio–product framework, the estimators utilize both location and variability information from X , enhancing efficiency and robustness in the two-phase sampling design.

Table 1. New suggested transformations for proposed estimators under two-phase sampling.

Estimator	τ_1	τ_2	τ_3	τ_4
$\hat{\phi}_{V_1}$	Hodges–Lehmann(X)	GMD(X)	1	Winsor_range _{10%} (X)
$\hat{\phi}_{V_2}$	Trim–Mean _{10%} (X)	Winsorized Mean _{10%} (X)	1	MAD(X)
$\hat{\phi}_{V_3}$	$Q_3 - Q_1$	Skew_adj(X)	1	CV(X)
$\hat{\phi}_{V_4}$	Median_ratio(Q_3/Q_1)	1	1	Bowley_skewness(X)
$\hat{\phi}_{V_5}$	$\log(1 + HL(X))$	$\log(1 + GMD(X))$	1	$\log(1 + IQR)$
$\hat{\phi}_{V_6}$	$\sqrt{Q_1} \cdot Q_3$	$Q_3 - Q_1$	1	$X_{90\%} - X_{10\%}$
$\hat{\phi}_{V_7}$	Hodges–Lehmann(X)	MAD(X)	1	Skew_adj(X)
$\hat{\phi}_{V_8}$	$1/(1 + CV(X))$	$1/(1 + Kurtosis(X))$	1	Q_3/Q_2

- **Hodges–Lehmann:**

$$HL(X) = \text{median} \left(\frac{X_i + X_j}{2} : 1 \leq i \leq j \leq m_1 \right),$$

- **Gini Mean Difference:**

$$GMD(X) = \frac{1}{m_1(m_1 - 1)} \sum_{1 \leq i < j \leq m_1} |X_i - X_j|,$$

- **Trimmed Mean (10%):** Let $k_1 = \lfloor 0.10 m_1 \rfloor$. Then

$$TM_{10\%}(X) = \frac{1}{m_1 - 2k} \sum_{i=k+1}^{m_1-k} X_{(i)}, \quad k = \lfloor 0.1m_1 \rfloor,$$

- **Winsorized Mean (10%):** Let $k_1 = \lfloor 0.10 m_1 \rfloor$. Then

$$WM_{10\%}(X) = \frac{1}{m_1} \left(k_1 X_{(k_1+1)} + \sum_{i=k_1+1}^{m_1-k_1} X_{(i)} + k_1 X_{(m_1-k_1)} \right),$$

- **Median ratio:**

$$MR(X) = \frac{Q_3}{Q_1},$$

- **Bowley Skewness:**

$$BSkew(X) = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1},$$

- **Inter quartile range:**

$$QR(X) = Q_3 - Q_1,$$

- **Geometric Quartile Mean:**

$$\sqrt{Q_1 \cdot Q_3},$$

- **Quartile Deviation:**

$$QD(X) = \frac{Q_3 - Q_1}{2},$$

- **10–90% Range:**

$$R_{10-90}(X) = X_{(0.90m_1)} - X_{(0.10m_1)},$$

- **Median Absolute Deviation (MAD):**

$$MAD(X) = \text{median}(|X_i - Q_2|), \quad i = 1, \dots, m_1,$$

- **Skewness Adjusted:**

$$Skew_{adj}(X) = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1},$$

- **Coefficient of Variation:**

$$CV(X) = \frac{S_X}{Q_2},$$

- **Moors' kurtosis:** Let Q_p be the sample quantile of order p . It is simple, bounded, and robust, making it a suitable choice when the target is the median. Using octiles $Q_{k/8}$:

$$M_K = \frac{(Q_{7/8} - Q_{5/8}) + (Q_{3/8} - Q_{1/8})}{Q_{6/8} - Q_{2/8}}.$$

Theorem 1 provides the theoretical foundation for the proposed median-based estimators under two-phase sampling. It establishes key properties, such as bias and mean squared error expressions, which are essential for understanding their efficiency and reliability. Including this theorem helps to explain why the proposed estimators perform better under different population distributions and correlation scenarios.

Theorem 1. A two-phase sampling scheme is considered in which $\hat{\phi}_V$ is defined as an exponential ratio-product type estimator for the finite population median Φ_y . The corresponding bias and mean squared error formulations are given below:

$$Bias(\hat{\phi}_V) \cong \Phi_y \left[\frac{1}{2}(r_1 - r_2)\lambda_3\Phi_{C_{yx}} - \frac{1}{8}(r_1^2 - r_2^2 - 2r_1r_2)\lambda_3\Phi_{C_x}^2 \right]$$

and

$$MSE(\hat{\phi}_V) \cong \Phi_y^2 \left[\lambda_1\Phi_{C_y}^2 + \frac{1}{4}\lambda_3\Phi_{C_x}^2(r_1^2 + r_2^2 - 2r_1r_2) + \lambda_3\Phi_{C_{yx}}(r_1 - r_2) \right].$$

Proof. The detailed derivation is given in Appendix A. $\square$

5. Theoretical Comparison with Existing Estimators

In this section, we demonstrate that the proposed estimator, denoted by $\hat{\phi}_V$, consistently outperforms each of the existing estimators presented in Section 2. The superiority of $\hat{\phi}_V$ is established through a set of explicit inequalities, which highlight its improved efficiency and reduced error in comparison. These inequalities not only provide a rigorous theoretical justification but also confirm the practical advantages of adopting $\hat{\phi}_V$ over its counterparts.

- By comparing the mean squared error of the newly proposed family of estimators (A4) with the variance of the sample median (1), we arrive at the following condition:

$$MSE(\hat{\phi}_y) > MSE(\hat{\phi}_V) \text{ if}$$

$$\frac{1}{4}\Phi_{C_x}^2(r_1^2 + r_2^2 - 2r_1r_2) + \Phi_{C_{yx}}(r_1 - r_2) > 0.$$

- By comparing Equation (A4) for the proposed family of estimators with the MSE in Equation (3), we obtain the condition stated below:

$$MSE(\hat{\phi}_{SA}) > MSE(\hat{\phi}_V) \text{ if}$$

$$\Phi_{C_x}^2(r_1^2 + r_2^2 - 2r_1r_2) + 4\Phi_{C_{yx}}(2 + r_1 - r_2) < 4\Phi_{C_x}^2.$$

- (iii) Examining the mean squared error of the proposed estimators in Equation (A4) against the MSE in Equation (4) leads to the following condition:

$$MSE(\hat{\phi}_S)_{min} > MSE(\hat{\phi}_V) \text{ if}$$

$$\Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + 4\Phi_{C_{yx}} (r_1 - r_2) > 4\Phi_{C_y}^2 \rho_{yx}^2.$$

- (iv) Evaluating the MSE of the proposed family of estimators (A4) in relation to the expression in Equation (7) provides the following condition, which is important for establishing their comparative efficiency:

$$MSE(\hat{\phi}_{Re}) > MSE(\hat{\phi}_V) \text{ if}$$

$$\Phi_{C_x}^2 (1 - 4V) > \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + 4\Phi_{C_{yx}} (r_1 - r_2).$$

- (v) A comparison of Equation (A4) for the proposed family of estimators with the MSE in Equation (8) yields the condition below:

$$MSE(\hat{\phi}_{Pe}) > MSE(\hat{\phi}_V) \text{ if}$$

$$\phi_{C_x}^2 (1 + 4V) > \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + 4\Phi_{C_{yx}} (r_1 - r_2).$$

- (vi) By comparing the MSE expression of the proposed family of estimators in Equation (A4) with the MSE of the sample median in Equation (12), the following condition is derived.

$$MSE(\hat{\phi}_R)_{min} > MSE(\hat{\phi}_V) \text{ if}$$

$$\frac{1}{4} \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + \Phi_{C_{yx}} (r_1 - r_2) > \frac{\Phi_{C_{yx}}^2}{\Phi_{C_x}^2 (1 + \lambda_1 \Phi_{C_y}^2) - \lambda_3 \Phi_{C_{yx}}^2}.$$

- (vii) By comparing the MSE of the newly developed family of estimators (A4) with the MSE in equation $\hat{\phi}_{G_1}$, the following condition is obtained, capturing their relative performance:

$$MSE(\hat{\phi}_{G_1})_{min} > MSE(\hat{\phi}_V) \text{ if}$$

$$\frac{B_{G_1}}{\lambda_3} + \Psi_{G_1} > \frac{1}{4} \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + \Phi_{C_{yx}} (r_1 - r_2),$$

where

$$\Psi_{G_1} = \Phi_{C_x}^2 (d_{4(opt)} \Phi_y + d_{5(opt)} \Phi_x)^2 - 2 d_{4(opt)} \Phi_y \Phi_{C_{yx}} (d_{4(opt)} \Phi_y + d_{5(opt)} \Phi_x)$$

and

$$B_{G_1} = \frac{Bias(\hat{\phi}_{G_1})_{min}^2}{\Phi_y^2} + \lambda_1 (d_{4(opt)}^2 - 1) \Phi_{C_y}^2.$$

- (viii) Evaluating the MSE of the proposed family of estimators (A4) relative to the MSE in Equation (14) leads to a specific condition, which highlights their comparative performance:

$$MSE(\hat{\phi}_{G_2})_{min} > MSE(\hat{\phi}_V) \text{ if}$$

$$\frac{B_{G_2}}{\lambda_3} + \frac{1}{4} d_{6(opt)}^2 \Phi_{C_x}^2 - \frac{\lambda_1}{\lambda_3} \Phi_{C_y}^2 > \frac{1}{4} \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1r_2) + \Phi_{C_{yx}} (r_1 - r_2),$$

where

$$B_{G_2} = \frac{\text{Bias}(\hat{\phi}_{G_2})_{\min}}{\Phi_y^2}.$$

6. Results and Discussion

In this section, the efficiency of the proposed family of estimators is evaluated in relation to existing methods. For this purpose, five distinct simulated populations are constructed based on appropriate positively skewed probability distributions. Furthermore, three practical datasets are employed to validate and support the empirical performance of the proposed estimators.

6.1. Monte Carlo Simulation Study

The most appropriate distribution for median estimation is determined by the nature of the dataset and the distribution's properties. Because the median resists the influence of skewness, extreme observations, and non-normal patterns, it becomes particularly advantageous in such contexts. In this study, the variable X is generated from one of the five distributions described below.

- **Population 1:** The random variable X follows a Cauchy distribution characterized by parameters $t_1 = 18$ and $t_2 = 11$. Since the Cauchy distribution does not have a defined mean or variance, the theoretical population correlation with Y is undefined. For our simulation, we used a correlation of $\rho_{yx} = -0.53$, which refers to the sample correlation calculated from the generated data set and reflects the observed inverse association between X and Y .
- **Population 2:** X follows a uniform distribution bounded between 16 and 22, and is statistically independent of Y in terms of correlation ($\rho_{yx} = 0$).
- **Population 3:** The variable X is modeled using an exponential distribution with a strong skew, where the rate parameter is $\alpha = 0.5$. Its correlation with Y is positive and equals $\rho_{yx} = 0.85$.
- **Population 4:** The random variable X is distributed as a gamma law with shape $a_1 = 14$ and scale $a_2 = 8$. Its dependence on Y is positive, with $\rho_{yx} = 0.67$.
- **Population 5:** The variable X is assumed to follow a log-normal distribution with mild skewness, defined by parameters $\mu_1 = 7$ and $\mu_2 = 2$. Its correlation with Y is 0.59.

6.2. Simulation Steps Under Two-Phase Sampling

The percent relative efficiencies (PREs) of both the proposed and existing estimators were examined for each distribution and correlation scenario defined above. All computations were performed in R software (latest v. 4.4.0) to assess their robustness and efficiency. The simulation procedure consists of the following steps:

1. **Population setup:** Generate the study population by selecting one of the five distributions for the auxiliary variable X (Cauchy, Uniform, Exponential, Gamma, or Log-normal) with their specified parameters, where $N = 1400$. For each unit, the corresponding study variable Y is obtained from

$$Y = \rho_{yx} \times X + e, \quad e \sim N(0, 1).$$

2. **First-phase sampling:** From the generated population of size N , draw a simple random sample of size m_1 . At this stage, only the values of the auxiliary variable X are recorded.

3. **Second-Phase Sampling:** From the first-phase sample, select a subsample of size m_2 ($m_2 < m_1$). For the units in this subsample, observe both the auxiliary variable X and the study variable Y .
4. **Estimator formation:** Using the information from phase-I and the paired (X, Y) observations from phase II, construct the proposed estimators based on the sample median.
5. **Repetition:** Repeat the above process a large number of times (for instance, 25,000 iterations) to assess the sampling behavior of the estimators.
6. **Performance assessment:** For each estimator, mean squared error (MSE), relative efficiency, and, when applicable, the coverage probability of confidence intervals.
7. **Comparison across populations:** Finally, compare the results across the five populations to evaluate how the estimators perform under heavy-tailed, skewed, and moderately correlated settings.

The following formulas are used to calculate the MSEs and PREs:

$$MSE(\hat{\phi}_t)_{\min} = \frac{\sum_{s=1}^{25000} (\hat{\phi}_{ts} - \Phi_y)^2}{25000}$$

and

$$PRE = \frac{MSE(\hat{\phi}_y)}{MSE(\hat{\phi}_t)_{\min}} \times 100,$$

where t ($t = y, SA, S, Re, Pe, R, G_1, G_2, V_1, V_2, \dots, V_8$).

6.3. Application to Survey Data

This part of the paper provides an empirical evaluation based on three real population datasets, with their details outlined below. The MSEs for the existing estimators and the proposed estimators of the population median $\hat{\Phi}_y$ of y are shown in Tables 2–4.

Population 1. We use data from the 2013 edition of Punjab Development Statistics [30] (p. 226), which records the number of registered factories in 2010 and their corresponding employment across different districts and divisions. This dataset serves as a practical benchmark for comparing the proposed estimators with existing approaches. The Pakistan Bureau of Statistics website provides a download link: <https://www.pbs.gov.pk/content/microdata> (accessed on 27 August 2025).

- Y : Denotes the average number of employees per district in 2010, reflecting workforce distribution across regions.
- X : Represents the total value of factory registrations in the same year, indicating the scale of industrial activity in each district.

Population 2. We use data from the 2014 edition of Punjab Development Statistics [31] (p. 135) to study a real finite population of gender-specific enrollment in government primary and middle schools, capturing both boys' and girls' enrollment across different districts. This dataset provides a comprehensive view of the distribution of students, allowing for an empirical evaluation of the proposed estimators under realistic population conditions. The Pakistan Bureau of Statistics website provides access to the microdata at: <https://www.pbs.gov.pk/content/microdata> (accessed on 27 August 2025).

- Y : Aggregate number of students registered in all schools during the 2012–2013 academic session.
- X : Total number of government-managed middle school institutions recorded for the 2012–2013 academic session.

Population 3. We consider a real finite population using data from the 2013 edition of Punjab Development Statistics [30] (p. 226). The dataset provides detailed information on the number of

registered factories and employment levels across various districts and divisions, including both employment distribution and factory registration counts for the year 2012. We use this dataset to apply and evaluate the performance of our proposed estimators, allowing us to examine their effectiveness under realistic population conditions. It is accessible for download on the Pakistan Bureau of Statistics website using the following URL: <https://www.pbs.gov.pk/content/microdata> (accessed on 27 August 2025).

- Y : Aggregate employment across all industrial sectors in each district.
- X : Fraction of registered factories in the corresponding district for the year 2012.

The datasets are summarized in Table 2, and the percent relative efficiencies of the proposed and existing estimators, derived from both simulated and real data, are displayed in Tables 3 and 4. These comparisons emphasize the practical effectiveness of the proposed methods, confirming their improved precision over conventional estimators.

Table 2. A statistical summary of the various datasets.

Symbols	Set 1 (Statistic)	Set 2 (Statistic)	Set 3 (Statistic)
N	36	36	36
Φ_x	168.5	1016.5	171.5
Φ_y	10,484.5	115,223.5	10,494
$f_x(\Phi_x)$	0.00015	0.00022	0.00019
$f_x(\Phi_y)$	0.00016	0.00024	0.00021
HL	199.50	1027	201.25
GMD	192.22	441.95	396.49
$TM_{10\%}$	235.50	1019.13	236.00
$WM_{10\%}$	270.67	1031.61	270.67
Q_1	89.5	729	87.5
Q_3	347	1242.25	352.5
MR	3.88	1.703	4.03
$BSkew$	0.386	−0.1225	0.366
QR	257.5	512.25	175.62
QD	128.75	256.125	265
R_{10-90}	751	844	762.5
MAD	92.50	289	99
$Skew_{adj}$	0.386	−0.1225	0.366
CV	2.60	0.396	2.64
M_K	2.10	1.125	2.18
ρ_{yx}	0.912	0.796	0.519
λ_1	0.0107	0.0242	0.0288
λ_2	0.0035	0.0123	0.0140
λ_3	0.00072	0.0119	0.0148

Table 3. A comparison of PRE across five populations.

Estimator	Pop1	Pop2	Pop3	Pop4	Pop5
$\hat{\phi}_{GR}$	100.00	100.00	100.00	100.00	100.00
$\hat{\phi}_{SA}$	196.15	169.71	95.99	190.70	151.17
$\hat{\phi}_S$	164.68	159.12	99.42	232.71	139.17
$\hat{\phi}_{Re}$	160.64	169.87	114.34	159.98	193.08
$\hat{\phi}_{Pe}$	60.91	90.10	72.55	60.59	47.22
$\hat{\phi}_R$	199.77	244.77	127.85	242.25	200.80
$\hat{\phi}_{G_1}$	224.02	199.57	154.70	252.88	230.97
$\hat{\phi}_{G_2}$	200.79	238.59	143.50	250.09	255.57
$\hat{\phi}_{V_1}$	274.81	300.02	191.82	281.45	261.83
$\hat{\phi}_{V_2}$	284.20	310.00	200.71	292.62	274.32
$\hat{\phi}_{V_3}$	284.42	300.00	188.92	299.87	299.23
$\hat{\phi}_{V_4}$	300.65	319.99	199.79	200.30	259.98

Table 3. Cont.

Estimator	Pop1	Pop2	Pop3	Pop4	Pop5
$\hat{\phi}_{V_5}$	296.34	280.01	179.73	278.66	269.67
$\hat{\phi}_{V_6}$	272.02	315.02	254.16	281.47	251.80
$\hat{\phi}_{V_7}$	301.87	325.00	239.22	260.07	270.99
$\hat{\phi}_{V_8}$	294.46	276.00	205.80	255.72	310.09

Table 4. PRE values among real data sets.

Estimator	Set 1	Set 2	Set 3
$\hat{\phi}_y$	100.00	100.00	100.00
$\hat{\phi}_{SA}$	134.54	123.16	143.53
$\hat{\phi}_S$	140.82	145.87	155.98
$\hat{\phi}_{Re}$	130.74	142.56	175.06
$\hat{\phi}_{Pe}$	95.62	65.92	86.49
$\hat{\phi}_R$	206.36	245.20	216.71
$\hat{\phi}_{G_1}$	239.47	252.38	274.63
$\hat{\phi}_{G_2}$	227.10	285.95	251.85
$\hat{\phi}_{V_1}$	251.88	310.43	290.91
$\hat{\phi}_{V_2}$	258.79	315.91	301.07
$\hat{\phi}_{V_3}$	292.68	298.08	299.37
$\hat{\phi}_{V_4}$	299.47	302.83	308.74
$\hat{\phi}_{V_5}$	286.98	299.55	284.46
$\hat{\phi}_{V_6}$	261.80	290.24	310.50
$\hat{\phi}_{V_7}$	257.94	331.82	334.29
$\hat{\phi}_{V_8}$	297.72	352.29	321.82

6.4. Analysis of Simulation and Empirical Results

Results show that the proposed transformation-based estimators outperform existing methods by attaining higher PRE values in both simulated and real-life datasets. Their ability to remain accurate under symmetric, skewed, and heavy-tailed distributions further demonstrates their robustness. The estimators consistently outperform traditional methods, particularly when data contain outliers or deviate from normality, making them highly adaptable to practical situations. They also remain stable across different sample sizes, ensuring reliable results even for small or finite populations. Tables 3 and 4 summarize these findings and emphasize the consistent superiority of the proposed methods.

- **Performance on simulated data:** Table 3 provides strong evidence that the proposed estimators consistently deliver higher percent relative efficiency (PRE) than conventional approaches. This advantage is observed across a broad range of distributions including all five different cases. The graphical representation in Figure 1 further illustrates this pattern, where the proposed estimators are repeatedly identified as the most efficient, appearing across all scenarios. Such findings confirm their robustness under diverse distributional conditions.
- **Application to real data sets:** The benefits of the proposed estimators are not confined to artificial populations; they extend convincingly to real-world data. Table 4 summarizes their performance across socio-economic and environmental populations, where they consistently achieve higher PRE values than their traditional counterparts. This trend is echoed in Figure 2, which clearly highlights the superiority of the proposed methods over exponential and difference-type estimators in all three populations considered. These results underscore the practical reliability of the proposed class of estimators when applied to empirical datasets.
- **Impact of correlation and sample size:** Another important aspect of efficiency is its stability across varying survey conditions. Figures 1 and 2 indicate that the proposed estimators maintain their effectiveness regardless of the correlation level between the

study and auxiliary variables. Complementary evidence from Tables 3 and 4 shows that efficiency does not deteriorate even when the second-phase sample size m_2 is much smaller than the first-phase size m_1 . This property is particularly valuable for practitioners conducting surveys with limited resources, as it ensures that reliable results can still be obtained under constrained sampling conditions.

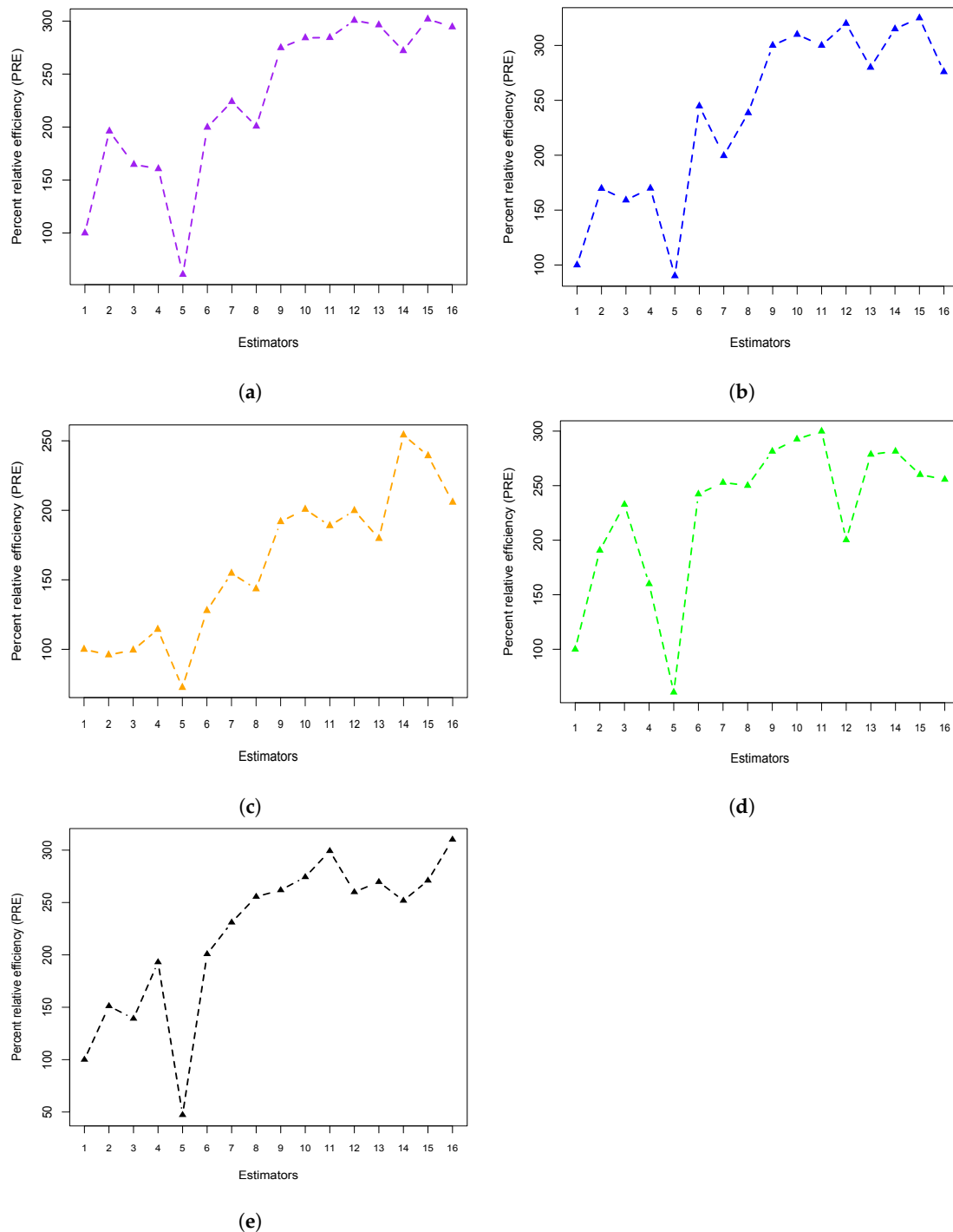


Figure 1. The graph depicts the PRE performance of the recommended estimators in comparison with traditional ones using simulated data, with estimators labeled sequentially from 1 through 16. For further details, please refer to Table 3. (a) $X \sim C(18, 11)$ with $\rho_{yx} = -0.53$. (b) $X \sim Uni(10, 15)$ with $\rho_{yx} = 0$. (c) $X \sim Exp(0.5)$ with $\rho_{yx} = 0.85$. (d) $X \sim Gam(14, 8)$ with $\rho_{yx} = 0.67$. (e) $X \sim LN(7, 2)$ with $\rho_{yx} = 0.59$.

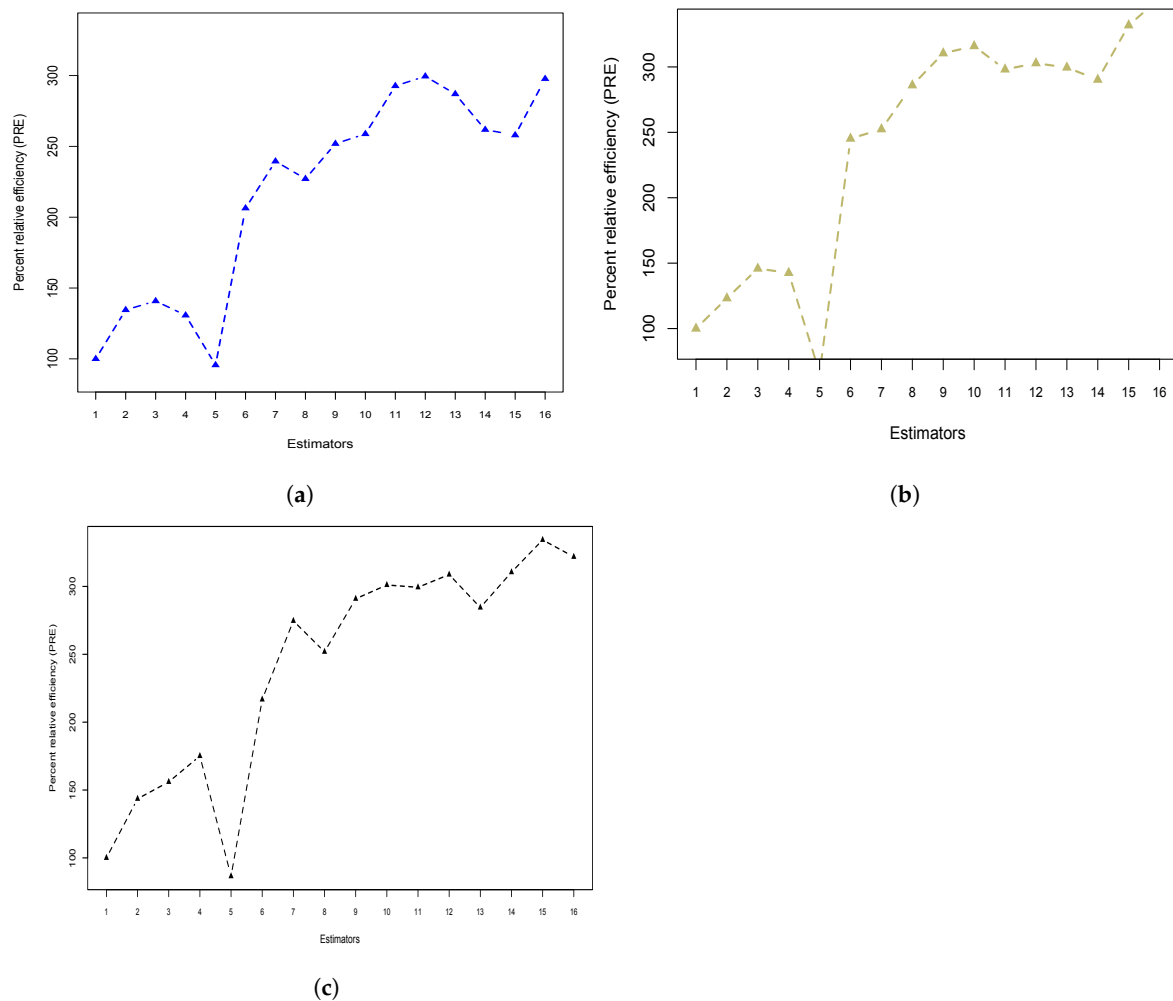


Figure 2. The figure shows the PRE performance of both proposed and traditional estimators on real-life data. For clarity, each estimator is assigned an index between 1 and 16. For further details, please refer to Table 4. (a) Source: [30]. (b) Source: [31]. (c) Source: [30].

7. Conclusions and Research Directions

This study presents a transformed class of median estimators developed for two-phase sampling when auxiliary information is expensive to obtain. Applying transformation strategies, the proposed estimators demonstrate clear improvements in efficiency and precision compared with conventional approaches.

Explicit expressions for bias and mean squared error were obtained, together with conditions under which these estimators consistently dominate classical counterparts. Theoretical properties were further confirmed through Monte Carlo experiments across different distributions and real survey applications, where the proposed methods consistently recorded higher PRE values.

Based on the PRE values reported in Tables 3 and 4, the $\hat{\phi}_{V_i}$ estimators consistently demonstrate superior performance across both artificial and real data sets. Specifically, for the artificial populations, the top three $\hat{\phi}_{V_i}$ estimators are $\hat{\phi}_{V_7}$, $\hat{\phi}_{V_3}$, and $\hat{\phi}_{V_6}$, as they exhibit the highest average PRE values across the five populations. For the real data sets, the leading estimators are $\hat{\phi}_{V_8}$, $\hat{\phi}_{V_7}$, and $\hat{\phi}_{V_4}$, which consistently achieve the largest PRE values across all three data sets. These results indicate that these $\hat{\phi}_{V_i}$ estimators provide the most efficient and reliable performance for the populations considered.

This study lays a strong foundation for robust median estimation under two-phase sampling, but several directions remain for future work. A natural extension is adapt-

ing these transformation-based estimators to stratified and multistage sampling, where auxiliary information is partly available at different levels. The proposed median-based estimators align with advances in robust and non-parametric statistics. By utilizing auxiliary information in a manner resistant to skewness and outliers, these estimators complement existing robust methods and offer practical value when standard parametric assumptions do not hold. Using multiple auxiliary variables together may also improve stability, especially in surveys where no single auxiliary measure explains population variability well.

Another direction lies in developing variance estimation procedures and confidence interval construction for the proposed estimators, whose applicability in practical survey settings can be expanded. Additionally, exploring their performance under nonresponse, measurement error, or incomplete auxiliary data can strengthen their robustness for real-life applications.

Funding: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2025R 299), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A

• A proof of Theorem 1, as detailed in Section 4

We revisit the following concepts, which are useful in proving the theorem:

$$\phi_0 = \left(\frac{\hat{\phi}_y - \Phi_y}{\Phi_y} \right),$$

$$\phi_1 = \left(\frac{\hat{\phi}_x - \Phi_x}{\Phi_x} \right),$$

and

$$\phi_2 = \left(\frac{\hat{\phi}_x - \Phi_x}{\Phi_x} \right),$$

such that $E(\phi_i) = 0$ for $i = 0, 1, 2$.

Moreover,

$$E(\phi_0^2) = \lambda_1 \Phi_{C_y}^2,$$

$$E(\phi_1^2) = \lambda_1 \Phi_{C_x}^2,$$

$$E(\phi_2^2) = \lambda_2 \Phi_{C_x}^2,$$

$$E(\phi_0 \phi_1) = \lambda_1 \Phi_{C_{yx}} = \rho_{yx} \Phi_{C_y} \Phi_{C_x},$$

$$E(\phi_0 \phi_2) = \lambda_2 \Phi_{C_{yx}} = \rho_{yx} \Phi_{C_y} \Phi_{C_x},$$

$$E(\phi_1 \phi_2) = \lambda_2 \Phi_{C_x}^2,$$

where

$$\Phi_{C_y} = \frac{1}{\Phi_y f_y(\Phi_y)},$$

$$\Phi_{C_x} = \frac{1}{\Phi_x f_x(\Phi_x)},$$

are the mathematical representations of the population coefficients of variation for the medians of Y and X , and

$$\begin{aligned}\lambda_1 &= \frac{1}{4} \left(\frac{N - m_2}{m_2 \times N} \right), \\ \lambda_2 &= \frac{1}{4} \left(\frac{N - m_1}{m_1 \times N} \right), \\ \lambda_3 &= \frac{1}{4} \left(\frac{m_1 - m_2}{m_1 \times m_2} \right).\end{aligned}$$

By expressing Equation (15) in terms of relative errors, we can derive analytical expressions for the bias and mean squared error of $\hat{\phi}_V$ as follows:

$$\hat{\phi}_V = \phi_y(1 + \phi_0) \exp \left[\frac{r_1(\phi_1 - \phi_2)}{2} \left(1 + \frac{r_1(\phi_1 + \phi_2)}{2} \right)^{-1} \right] \exp \left[\frac{-r_2(\phi_1 - \phi_2)}{2} \left(1 + \frac{r_2(\phi_1 + \phi_2)}{2} \right)^{-1} \right] \quad (\text{A1})$$

where r_1 , and r_2 are defined as:

$$r_1 = \frac{\tau_1 \Phi_x}{\tau_1 \Phi_x + \tau_2}$$

and

$$r_2 = \frac{\tau_3 \Phi_x}{\tau_3 \Phi_x + \tau_4}.$$

Applying a first-order Taylor series expansion to the right-hand side of Equation (A1) yields the desired approximation. Higher-order terms ($\phi_i > 2$) are excluded from consideration because their effect is insignificant, resulting in the following expression:

$$\begin{aligned}\hat{\phi}_V &= \phi_y(1 + \phi_0) \exp \left[\frac{r_1(\phi_1 - \phi_2)}{2} \left\{ 1 - \frac{r_1(\phi_1 + \phi_2)}{2} + \frac{r_1^2(\phi_1 + \phi_2)^2}{4} \right\} \right] \\ &\times \exp \left[-\frac{r_2(\phi_1 - \phi_2)}{2} \left\{ 1 - \frac{r_2(\phi_1 + \phi_2)}{2} + \frac{r_2^2(\phi_1 + \phi_2)^2}{4} \right\} \right]. \\ \hat{\phi}_V &= \phi_y(1 + \phi_0) \exp \left[\frac{r_1(\phi_1 - \phi_2)}{2} - \frac{r_1^2(\phi_1^2 - \phi_2^2)}{4} \right] \exp \left[\frac{r_2(\phi_2 - \phi_1)}{2} - \frac{r_2^2(\phi_1^2 - \phi_2^2)}{4} \right].\end{aligned}$$

After simplifying, we obtain:

$$\begin{aligned}\hat{\phi}_V - \Phi_y &\cong \Phi_y \left[\phi_0 + \frac{r_1}{2} (\phi_1 - \phi_2 + \phi_0 \phi_1 - \phi_0 \phi_2) + \frac{r_2}{2} (\phi_2 - \phi_1 - \phi_0 \phi_1 + \phi_0 \phi_2) \right. \\ &\left. - \frac{r_1^2}{8} (\phi_1^2 - 3\phi_2^2 + 2\phi_1 \phi_2) - \frac{r_2^2}{8} (\phi_1^2 - 3\phi_2^2 + 2\phi_1 \phi_2) - \frac{r_1 r_2}{4} (\phi_1^2 \phi_2^2 - 2\phi_1 \phi_2) \right].\end{aligned} \quad (\text{A2})$$

To compute the bias of $\hat{\phi}_V$, we take the expectation of Equation (A2) and replace each occurrence of $\phi_0, \phi_1, \phi_0^2, \phi_1^2, \phi_2^2, \phi_0 \phi_1, \phi_0 \phi_2, \phi_1 \phi_2$ with their respective expected values. This results in:

$$\begin{aligned}\text{Bias}(\hat{\phi}_V) &\cong \Phi_y \left[\frac{1}{2} (r_1 - r_2) (\lambda_1 \Phi_{C_{yx}} - \lambda_2 \Phi_{C_{yx}}) - \frac{1}{8} (r_1^2 - r_2^2) (\lambda_1 \Phi_{C_x}^2 - \lambda_2 \Phi_{C_x}^2) \right. \\ &\left. - \frac{r_1 r_2}{4} (\lambda_1 \Phi_{C_x}^2 - \lambda_2 \Phi_{C_x}^2) \right].\end{aligned}$$

After simplification, we obtain:

$$\text{Bias}(\hat{\phi}_V) \cong \Phi_y \left[\frac{1}{2} (r_1 - r_2) \lambda_3 \Phi_{C_{yx}} - \frac{1}{8} (r_1^2 - r_2^2 - 2r_1 r_2) \lambda_3 \Phi_{C_x}^2 \right], \quad (\text{A3})$$

where

$$\lambda_3 = \lambda_1 - \lambda_2.$$

The steps that follow outline the derivation of the mean squared error of $\hat{\phi}_V$ based on a first-order approximation. Squaring both sides of Equation (A2) and taking expecta-

tions, while neglecting terms involving ϕ 's with powers greater than two (ϕ 's > 2), and substituting the corresponding expected values, we obtain:

$$MSE(\hat{\phi}_V) \cong \Phi_y^2 \left[\lambda_1 \Phi_{C_y}^2 + \frac{1}{4} \lambda_3 \Phi_{C_x}^2 (r_1^2 + r_2^2 - 2r_1 r_2) + \lambda_3 \Phi_{C_{yx}} (r_1 - r_2) \right]. \quad (A4)$$

References

1. Daraz, U.; Alomair, M.A.; Albalawi, O.; Al Naim, A.S. New Techniques for Estimating Finite Population Variance Using Ranks of Auxiliary Variable in Two-Stage Sampling. *Mathematics* **2024**, *12*, 2741. [CrossRef]
2. Zaman, T.; Bulut, H. A simulation study: Robust ratio double sampling estimator of finite population mean in the presence of outliers. *Sci. Iran.* **2021**, *31*, 1330–1341. [CrossRef]
3. Daraz, U.; Shabbir, J.; Khan, H. Estimation of finite population mean by using minimum and maximum values in stratified random sampling. *J. Mod. Appl. Stat. Methods* **2018**, *17*, 1–15. [CrossRef]
4. Alomair, M.A.; Daraz, U. Dual transformation of auxiliary variables by using outliers in stratified random sampling. *Mathematics* **2024**, *12*, 2829. [CrossRef]
5. Gross, S. Median estimation in sample surveys. In Proceedings of the Section on Survey Research Methods, American Statistical Association Ithaca, Alexandria, VA, USA, 1980. Available online: http://www.asasrms.org/Proceedings/papers/1980_037.pdf (accessed on 27 August 2025).
6. Sedransk, J.; Meyer, J. Confidence intervals for the quantiles of a finite population: Simple random and stratified simple random sampling. *J. R. Stat. Soc. Ser. B* **1978**, *40*, 239–252. [CrossRef]
7. Philip, S.; Sedransk, J. Lower bounds for confidence coefficients for confidence intervals for finite population quantiles. *Commun. Stat. Theory Methods* **1983**, *12*, 1329–1344. [CrossRef]
8. Kuk, Y.C.A.; Mak, T.K. Median estimation in the presence of auxiliary information. *J. R. Stat. Soc. Ser. B* **1989**, *51*, 261–269. [CrossRef]
9. Rao, T.J. On certain methods of improving ratio and regression estimators. *Commun. Stat. Theory Methods* **1991**, *20*, 3325–3340. [CrossRef]
10. Singh, S.; Joarder, A.H.; Tracy, D.S. Median estimation using double sampling. *Aust. N. Z. J. Stat.* **2001**, *43*, 33–46. [CrossRef]
11. Khoshnevisan, M.; Singh, H.P.; Singh, S.; Smarandache, F. *A General Class of Estimators of Population Median Using Two Auxiliary Variables in Double Sampling*; Virginia Polytechnic Institute and State University: Blacksburg, VA, USA, 2002.
12. Singh, S. *Advanced Sampling Theory with Applications: How Michael Selected Amy*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2003; Volume 2.
13. Gupta, S.; Shabbir, J.; Ahmad, S. Estimation of median in two-phase sampling using two auxiliary variables. *Commun. Stat. Theory Methods* **2008**, *37*, 1815–1822. [CrossRef]
14. Aladag, S.; Cingi, H. Improvement in estimating the population median in simple random sampling and stratified random sampling using auxiliary information. *Commun. Stat. Theory Methods* **2015**, *44*, 1013–1032. [CrossRef]
15. Solanki, R.S.; Singh, H.P. Some classes of estimators for median estimation in survey sampling. *Commun. Stat. Theory Methods* **2015**, *44*, 1450–1465. [CrossRef]
16. Daraz, U.; Wu, J.; Albalawi, O. Double exponential ratio estimator of a finite population variance under extreme values in simple random sampling. *Mathematics* **2024**, *12*, 1737. [CrossRef]
17. Daraz, U.; Wu, J.; Alomair, M.A.; Aldoghan, L.A. New classes of difference cum-ratio-type exponential estimators for a finite population variance in stratified random sampling. *Heliyon* **2024**, *10*, e33402. [CrossRef]
18. Daraz, U.; Alomair, M.A.; Albalawi, O. Variance estimation under some transformation for both symmetric and asymmetric data. *Symmetry* **2024**, *16*, 957. [CrossRef]
19. Shabbir, J.; Gupta, S. A generalized class of difference type estimators for population median in survey sampling. *Hacet. J. Math. Stat.* **2017**, *46*, 1015–1028. [CrossRef]
20. Irfan, M.; Maria, J.; Shongwe, S.C.; Zohaib, M.; Bhatti, S.H. Estimation of population median under robust measures of an auxiliary variable. *Math. Probl. Eng.* **2021**, *2021*, 4839077. [CrossRef]
21. Shabbir, J.; Gupta, S.; Narjis, G. On improved class of difference type estimators for population median in survey sampling. *Commun. Stat. Theory Methods* **2022**, *51*, 3334–3354. [CrossRef]
22. Hussain, M.A.; Javed, M.; Zohaib, M.; Shongwe, S.C.; Awais, M.; Zaagan, A.A.; Irfan, M. Estimation of population median using bivariate auxiliary information in simple random sampling. *Heliyon* **2024**, *10*, e28891. [CrossRef]
23. Bhushan, S.; Kumar, A.; Lone, S.A.; Anwar, S.; Gunaim, N.M. An efficient class of estimators in stratified random sampling with an application to real data. *Axioms* **2023**, *12*, 576. [CrossRef]
24. Stigler, S.M. Linear functions of order statistics. *Ann. Math. Stat.* **1969**, *40*, 770–788. [CrossRef]

25. Daraz, U.; Almulhim, F.A.; Alomair, M.A.; Alomair, A.M. Population median estimation using auxiliary variables: A simulation study with real data across sample sizes and parameters. *Mathematics* **2025**, *13*, 1660. [\[CrossRef\]](#)
26. Daraz, U.; Khan, M. Estimation of variance of the difference-cum-ratio-type exponential estimator in simple random sampling. *Res. Math. Stat.* **2021**, *8*, 1–15. [\[CrossRef\]](#)
27. Daraz, U.; Wu, J.; Agustiana, D.; Emam, W. Finite population variance estimation using Monte Carlo simulation and real life application. *Symmetry* **2025**, *17*, 84. [\[CrossRef\]](#)
28. Daraz, U.; Agustiana, D.; Wu, J.; Emam, W. Twofold auxiliary information under two-phase sampling: An improved family of double-transformed variance estimators. *Axioms* **2025**, *14*, 64. [\[CrossRef\]](#)
29. Singh, H.P.; Vishwakarma, G.K. Modified exponential ratio and product estimators for finite population mean in double sampling. *Austrian J. Stat.* **2007**, *36*, 217–225. [\[CrossRef\]](#)
30. Bureau of Statistics. *Punjab Development Statistics Government of the Punjab, Lahore, Pakistan*; Bureau of Statistics: Islamabad, Pakistan, 2013.
31. Bureau of Statistics. *Punjab Development Statistics Government of the Punjab, Lahore, Pakistan*; Bureau of Statistics: Islamabad, Pakistan, 2014.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.