JIRSS (2026)
Vol. 25, No. 1, pp 1-16
DOI: 10.22034/jirss.2026.2064841.1120

A Dual Auxiliary Variable Approach to Finite Population Variance Estimation

Rabbia Mukhtar[1], Abid Hussain[1], Muhammad Asim Masood[1], Nasir Ali[1]

[1]Department of Statistics, PMAS-Arid Agriculture University, Rawalpindi, Pakistan.

Received: 2025-07-05, Accepted: 2026-05-18, Published online: 2026-05-19

Abstract. This study proposes a novel estimator for the finite population variance under simple random sampling. The estimator utilizes dual auxiliary information by incorporating the empirical cumulative distribution function (ECDF) of an auxiliary variable. The ECDF, which represents the stochastic process over the unit interval [0,1], is employed to enhance the estimation precision. The performance of the proposed estimator is evaluated through a comprehensive analysis. First, the bias and mean squared error (MSE) of the estimator are derived analytically. Second, a simulation study is conducted to investigate the estimator's behavior under various parametric settings. Finally, an empirical comparison is made with several well-established estimators using five real-world datasets. The results consistently demonstrate the superiority of the proposed estimator in terms of both bias and MSE, suggesting its practical utility.

Keywords. Auxiliary information, simple random sampling, mean squared error, empirical distribution function, simulation.

MSC: 62-XX, 62Dxx, 62D05.

1 Introduction

Variance estimation is crucial at all stages of survey sampling, from exploratory data analysis to advanced modeling, and is essential for making informed and reliable data-driven decisions. For instance, variance and the mean provide a summary of data dispersion; high variance may indicate the presence of outliers. It also helps evaluate the performance of machine learning models by examining the consistency of their predictions. In image processing, variance quantifies texture, distinguishing between different regions based on their variability. Overall,

Rabbia Mukhtar (rabbia.mukhtarbaig@gmail.com).
Corresponding Author: Abid Hussain (abid0100@gmail.com).
Muhammad Asim Masood (asimmasood2746@gmail.com).
Nasir Ali (nasir_stat@uaar.edu.pk).

variance is a fundamental statistical concept with wide-ranging applications in statistics, aiding in understanding data variability, assessing model performance, and making robust decisions. Variance estimation is also a significant challenge in the realm of large and complex data sets. Variance estimation serves as a cornerstone in statistical inference, playing a pivotal role in constructing confidence intervals, conducting hypothesis tests, and assessing the precision of parameter estimates. In survey sampling, accurate variance estimation is particularly critical as it directly impacts the reliability of conclusions drawn from sample data about finite populations. The importance of variance extends beyond traditional statistical applications—in machine learning, variance helps evaluate model stability; in quality control, it monitors process consistency; and in economic planning, it informs risk assessment and policy formulation.

Survey studies have shown that proper use of auxiliary variables is essential to improving estimates of population variance. It is additional data, often known as an auxiliary or benchmark variable, and is typically accessible through organizational databases, census data, or past experience, see for example, Singh et al. (2013). For more accuracy, the utility of auxiliary information in assisting the estimation operations while enumerating the character(s) of interest is well elaborated in the literature, see for example, Biemer and Peytchev (2013) and Chou et al. (2017). An overwhelming account of scientific inquiries from multidisciplinary research literature expounded upon the appropriate use of supportive information to control the variability of estimation procedure can be witnessed in Zhang and Chambers (2004); Kreuter et al. (2010) and Bai et al. (2021). The pioneer usage of additional information while estimating the parameter(s) of interest may be considered rooted in the argument provided by the prominent figure of eighteenth-century research circles – Pierre-Simon Laplace. Regarding the population estimation issue of France, Laplace advocated the use of supportive information by noting "The register of births, which are kept with care in order to assure the condition of the citizens, can serve to determine the population of great empire without resorting a census of its inhabitants. But for this it is necessary to know the ratio of population to annual birth." see, Lohr (1999). Then, Cochran (1940) proceeded on this front by offering the mathematical foundations to legitimize the influx of relevant auxiliary in the launched estimation procedure. By doing so, Cochran (1940) provided groundbreaking work materializing the well-cherished ratio estimator. The suggested estimator was capable of exploiting the underlying correlation structure of the study and the auxiliary variables. Unfortunately, the applicability of the devised mechanism remained restricted to positive correlation exhibition, only. Another advancement was made by Murthy (1964) with the derivation of a product estimator while enabling the analysts to exploit negative linear dependencies among study variables and associated auxiliaries. The final resolve came through Cochran (1977) with the proposition of a famous regression estimator capable of using both positive and negative linearity dictating the outlook of both variables of interest and the auxiliary.

The evolution of variance estimation using auxiliary information has followed a progressive trajectory. Following Cochran (1940) foundational work on ratio estimators, researchers have continuously sought to enhance estimation efficiency through innovative utilization of auxiliary variables. Isaki (1983) pioneered the direct application of auxiliary information for variance estimation, laying the groundwork for subsequent developments. The transition from simple ratio-type estimators to more sophisticated exponential-type estimators marked a significant advancement, offering improved flexibility in capturing nonlinear relationships (Bahl and Tuteja, 1991). More recently, the dual use of auxiliary information, as demonstrated by Haq et al. (2017) for mean estimation, has opened new avenues for efficiency gains by exploiting multiple aspects of auxiliary variables. Along with the above-documented developments, another frontier to enhance the estimation efficiency of existing schemes gained the attraction of investigative circles. Many researchers expounded on the idea of inflowing the supportive

information in the estimation process while using more delicate functional forms known for producing less variable estimators. Our goal was to focus on the competent incubation of exponential formations in existing estimating functionals and thus deriving novel pathways of incorporating auxiliary information. Resultantly, Bahl and Tuteja (1991) developed notable extensions to the classic ratio estimators and product estimators, namely, ratio-type exponentially and product-type exponentially estimators. The targets were obtained through the competent use of exponential functions. Further advancements were launched in Grover and Kaur (2014) suggesting a generalized formation based on exponent functionals. The flexibility and diversity of the proposed family were illustrated by showing numerous existing estimators as members of the devised scheme. The adaptability of the above-documented approaches can be assessed by more formal accounts provided in several studies, (see for example, Haq et al. 2017; Muneer et al. 2017, 2018; Irfan et al. 2020; Zaman et al. 2021; Pandey et al. 2021; Daraz and Khan 2021; Alomair and Daraz 2024; Daraz et al. 2024b,c,a, 2025a,b).

As in the near past, the argument of dual use of auxiliary information is generated and competently defended under the notion of maximal utility of information is generated. Haq et al. (2017) instigated the idea of using ranks of the auxiliary variable along with the usual approach of considering the mean to prompt the efficient estimation of the finite population mean of the study variable. Several existing estimators and families of estimators, such as the conventional mean, ratio, product, regression, exponential-ratio, and exponential-product estimates, were examined, (see for example, Rao 1991; Gupta and Shabbir 2008; Grover and Kaur 2014; Shabbir et al. 2014). A tedious comparative analysis established the dominance of Haq et al. (2017) dual use idea over the all above-documented contemporaries. Recently, Hussain et al. (2022) devised the duality by carefully combining the mean value and the empirical cumulative distribution function of auxiliary data while estimating the means of the studied variable's finite population. A detailed performance profiling of the newly suggested empirical cumulative distribution function based estimation scheme revealed the dominance of devised formation over the elegant scheme of Haq et al. (2017) and so over the other contemporaries. To the best of our knowledge, no other mechanism targeting the finite population mean estimation while considering dual use of auxiliary information is available, to produce more efficient estimates as compared to Hussain et al. (2022) proposition. Despite these advances, existing variance estimators primarily utilize only the mean or variance of auxiliary variables, leaving other distributional characteristics unexploited. The empirical cumulative distribution function (ECDF) offers a comprehensive representation of the auxiliary variable's entire distribution, yet its potential for variance estimation remains largely untapped. Hussain et al. (2022) demonstrated the effectiveness of ECDF-based dual auxiliary information for mean estimation, achieving superior performance compared to conventional approaches. However, no study to date has extended this promising framework to variance estimation—a gap this research aims to fill.

This research offers extensions to the Hussain et al. (2022) estimation strategy while targeting the finite population variance estimation. This article proposes a new estimator where auxiliary information use is profoundly argued on two fronts: by the exploitation of the empirical cumulative distribution function and through the exponential function. Moreover, the fairness in comparison is extended by considering the same data sets and same optimal conditions as were used and argued by the promising and previously dominating work. The various comparative subtleties along with a dominant performance of the newly devised variance estimating are elaborated on throughout this article.

The current article comprises five main parts. Section 2 lays the groundwork of methodologies that are extensively employed in this work, and details the contemporary, established methods that our results are intended to be compared with. Further, Section 3 provides the

mathematical foundations of the proposed scheme. Section 4 is dedicated to documenting the empirical performance of the competing techniques. In the last Section 5, we provide a comprehensive investigation of the proposed structure and a number of potential study locations.

2 Preliminaries

2.1 Symbolical representation

We assume a sequence of N elements of the desired study variable, say Y, can be written as:

$$P_Y(N) = Y_1(N), Y_2(N), \cdots, Y_N(N).$$

With the assumption of complete availability of the information, let's consider another vector, say, auxiliary variable (X), containing its attributes as:

$$P_X(N) = X_1(N), X_2(N), \cdots, X_N(N).$$

The mean of the study variable is denoted with $\bar{Y}$ whereas $\bar{X}$ represents the population mean of the auxiliary variable. The $\bar{F}$ is the mean of the empirical cumulative distribution function (ECDF) of the auxiliary variable. Also, S_Y^2, S_X^2, and S_F^2 indicate the population variances of the relevant variables along the co-efficient of variations are: $C_Y = \frac{S_Y}{\bar{Y}}$, $C_X = \frac{S_X}{\bar{X}}$, $C_F = \frac{S_F}{\bar{F}}$. Moreover, the following relative errors are used to compute the biases and mean square errors as $e_0 = \frac{s_y^2 - S_y^2}{S_y^2}$, $e_1 = \frac{s_x^2 - S_x^2}{S_x^2}$ and $e_2 = \frac{s_f^2 - S_F^2}{S_F^2}$, where, s_y^2, s_x^2 and s_f^2 are the corresponding sample variances. It is trivial to verify that $E(e_i) = 0 \; \forall \; i = 0, 1, 2$, whereas, $E(e_0^2) = \theta\Theta_{400}$, $E(e_1^2) = \theta\Theta_{040}$, $E(e_2^2) = \theta\Theta_{004}$. Similarly, $E(e_0e_1) = \theta\Theta_{220}$, $E(e_0e_2) = \theta\Theta_{202}$, $E(e_1e_2) = \theta\Theta_{022}$, with $\theta = (\frac{1}{n} - \frac{1}{N})$, whereas n is the sample size. Additionally,

$$\Theta_{pqr} = \Theta^*_{pqr} - 1,$$

$$\Theta^*_{pqr} = \frac{\nu_{pqr}}{\nu_{200}^{p/2}\nu_{020}^{q/2}\nu_{002}^{r/2}},$$

with

$$\nu_{pqr} = \frac{\sum (y_i - \bar{Y})^p (x_i - \bar{X})^q (f_i - \bar{F})^r}{N - 1}.$$

In the above expression, p, q, r are non-negative integers where, $\nu_{200}, \nu_{020}, \nu_{002}$ are the second-order moments from means, whereas the moment ratio is denoted with Θ^*_{pqr}.

Empirical cumulative distribution function (ECDF): For an auxiliary variable X with ordered observations $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$, the empirical cumulative distribution function at point t is defined as:

$$F_N(t) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \leq t)$$

where $I(\cdot)$ is the indicator function. For the i-th observation, the ECDF value is computed as:

$$F_i = F_N(X_i) = \frac{\text{Number of observations} \leq X_i}{N} = \frac{r_i}{N}$$

where r_i is the rank of X_i.

Population mean of the ECDF:

$$\bar{F} = \frac{1}{N}\sum_{i=1}^{N} F_i = \frac{1}{N}\sum_{i=1}^{N} \frac{r_i}{N} = \frac{N+1}{2N}$$

Population variance of the ECDF:

$$S_F^2 = \frac{1}{N-1}\sum_{i=1}^{N}(F_i - \bar{F})^2$$

Sample variance of the ECDF:

$$S_f^2 = \frac{1}{n-1}\sum_{i=1}^{n}(f_i - \bar{f})^2$$

Population variance of the study variable:

$$S_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

Sample variance of the study variable:

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

Population variance of the auxiliary variable:

$$S_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2$$

Sample variance of the auxiliary variable:

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

2.2 Prominent existing estimators

In this section, the most prominent existing estimators of population variance are discussed.

- The well-known estimator of variance is:

$$T_0 = \frac{\sum_{i=1}^{n}(y_i - \bar{Y})^2}{n-1}.$$

Its variance is given by

$$Var(T_0) = \theta S_y^4 \Theta_{400}. \tag{2.1}$$

- Isaki (1983) used auxiliary information for variance estimation first-time in literature as:

$$T_1 = s_y^2\left(\frac{S_x^2}{s_x^2}\right).$$

The expressions of bias and MSE of T_1 are given below:

$$Bias(T_1) = \theta S_y^2\Big(\Theta_{040} - \Theta_{220}\Big), \tag{2.2}$$

$$\text{MSE}(T_1) = \theta S_y^4\Big(\Theta_{400} + \Theta_{040} - 2\Theta_{220}\Big). \tag{2.3}$$

- Rao (1991) devised the following difference estimator as:

$$T_2 = \lambda_1 s_y^2 + \lambda_2(S_x^2 - s_x^2),$$

where,

$$\lambda_1 = \frac{\Theta_{040}}{\theta\Big(\Theta_{400}\Theta_{040} - \Theta_{220}^2\Big) + \Theta_{040}},$$

$$\lambda_2 = \frac{S_y^2\Theta_{220}}{S_x^2\Big(\theta(\Theta_{400}\Theta_{040} - \Theta_{220}^2) + \Theta_{040}\Big)}.$$

This estimator gives the following expressions of bias and MSE:

$$Bias(T_2) = S_y^2\left(\frac{\Theta_{400}}{\theta\Big(\Theta_{400}\Theta_{040} - \Theta_{220}^2\Big) + \Theta_{040} - 1}\right), \tag{2.4}$$

and

$$MSE(T_2) = \theta S_y^4\left(\frac{\Theta_{400}\Theta_{040} - \Theta_{220}^2}{\theta\Big(\Theta_{400}\Theta_{040} - \Theta_{220}^2\Big) + \Theta_{040}}\right). \tag{2.5}$$

- Singh et al. (2009) suggested an exponential based variance estimator as:

$$T_3 = s_y^2 \exp\left(\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right).$$

The expressions of bias and MSE of T_3 are as under:

$$Bias(T_3) = \theta s_y^2\left(\frac{3}{8}\Theta_{040} - \frac{1}{2}\Theta_{220}\right), \tag{2.6}$$

and

$$MSE(T_3) = \theta s_y^4\left(\Theta_{400} + \frac{1}{4}\Theta_{040} - \Theta_{220}\right). \tag{2.7}$$

- Recently, Ahmad et al. (2022) converted the well-known population mean estimator by Grover and Kaur (2011) for variance estimation. Their suggested estimator is in the following form:

$$T_4 = \lambda_3 s_y^2 + \lambda_4 (S_x^2 - s_x^2) \exp\left(\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right).$$

They derived the bias as:

$$Bias(T_4) = -S_y^2 - \lambda_3 S_y^2 \left(1 + \frac{3}{8}\Theta_{040} - \frac{1}{2}\Theta_{220}\right) + \frac{1}{2}\lambda_4 S_x^2 \theta \Theta_{040}, \tag{2.8}$$

where,

$$\lambda_3 = \frac{8\Theta_{040} - \theta\Theta_{040}^2}{8\left(\theta(\Theta_{400}\Theta_{040} - \Theta_{220}^2) + \Theta_{040}\right)},$$

$$\lambda_4 = \frac{S_y^2\left(\theta\Theta_{040}^2 - \theta\Theta_{040}\Theta_{220} + 8\Theta_{220} - 4\left(\Theta_{040} - \theta(\Theta_{400}\Theta_{040} - \Theta_{220}^2)\right)\right)}{8\left(\theta(\Theta_{400}\Theta_{040} - \Theta_{220}^2) + \Theta_{040}\right)}.$$

The MSE expression derived by them as:

$$MSE(T_4) = \frac{\theta S_y^4\left(64(\Theta_{400}\Theta_{040} - \Theta_{220}^2) - \theta\Theta_{040}^3 - 16\Theta_{040}(\Theta_{400}\Theta_{040} - \Theta_{220}^2)\right)}{64\left(\theta(\Theta_{400}\Theta_{040} - \Theta_{220}^2) + \Theta_{040}\right)}. \tag{2.9}$$

3 Proposed estimators

In the adjacent past, Hussain et al. (2022) suggested a dual formation of the supportive variable for estimating population parameters in the form of an empirical cumulative distribution function. To observe the complete outlook of the distributional structure of the random behavior, the empirical cdf is one of the best procedures. Hussain et al. (2022) showed that it would be beneficial to enhance the efficiency of the estimation procedures.

Let us consider a distribution function F that governs the dynamics of the supportive variable associated with order statistics as $X_1 < X_2 < \cdots < X_N$. For i^{th} object, it can be defined in the following expression:

$$F_i = \begin{cases} \frac{i}{N}; & X_i \le X < X_{i+1}, \quad \forall\, i = 1, 2, \cdots, N-1, \\ 1; & X_N \le X. \end{cases}$$

Our proposition is based on the motivation of Hussain et al. (2022) which is in the following form:

$$T_5 = \left(\lambda_{11} s_y^2 + \lambda_{12}(S_x^2 - s_x^2) + \lambda_{13}(S_f^2 - s_f^2)\right) \exp\left(\frac{S_x^2 - s_x^2}{S_x^2 + s_x^2}\right), \tag{3.1}$$

where λ_{11}, λ_{12}, and λ_{13} are the constants. For the calculations of bias and MSE, the above expression is further simplified in the notion of relative errors as:

$$T_5 = \left(\lambda_{11} s_y^2 + \lambda_{12}(S_x^2 - s_x^2) + \lambda_{13}(S_f^2 - s_f^2)\right) \exp\left(1 - \frac{1}{2}e_1 + \frac{3}{8}e_1^2 + \cdots\right).$$

To maintain the above expression up to second-order terms, we proceed as:

$$\begin{aligned}(T_5 - S_y^2) &= -S_y^2 + \lambda_{11}S_y^2 + \lambda_{11}S_y^2 e_0 - \frac{1}{2}\lambda_{11}S_y^2 e_1 - \lambda_{12}e_1 - \lambda_{13}e_2 + \frac{3}{8}\lambda_{11}S_y^2 e_1^2 \\ &\quad + \frac{1}{2}\lambda_{12}e_1^2 - \frac{1}{2}\lambda_{11}S_y^2 e_0 e_1 - \frac{1}{2}\lambda_{13}e_1 e_2.\end{aligned}$$

For the result of bias, we take the expectation of the above equation and get the following expression:

$$\begin{aligned}Bias(T_5) &= (\lambda_{11} - 1)S_y^2 + \frac{1}{8}\Big(4\theta\Theta_{040}\Big(S_x^2\Theta_{040}\lambda_{12} + S_r^2\Theta_{004}\lambda_{13}\Big(\frac{\Theta_{022}}{\sqrt{\Theta_{040}\Theta_{004}}}\Big)\Big) \\ &\quad + \theta\Theta_{040}\Big(3\Theta_{040} - 4\Theta_{400}\frac{\Theta_{220}}{\sqrt{\Theta_{400}\Theta_{040}}}\Big)\Big). \qquad (3.2)\end{aligned}$$

The optimum values of λ_{11}, λ_{12} and λ_{13} are given below:

$$\lambda_{11} = \frac{8 - \theta\Theta_{040}}{8\Big(1 + \theta\Theta_{400}(1 - A_2\Big)},$$

$$\lambda_{12} = \frac{S_y^2\left(\begin{array}{l}\theta\Theta_{040}^3\Big(\frac{\Theta_{022}}{\sqrt{\Theta_{040}\Theta_{004}}}\Big) + (-8\Theta_{400} + \theta\Theta_{040}\Theta_{400})\Big(\frac{\Theta_{220}}{\sqrt{\Theta_{400}\Theta_{040}}} - \frac{\Theta_{022}}{\sqrt{\Theta_{040}\Theta_{004}}}\Big) \\ \times \frac{\Theta_{202}}{\sqrt{\Theta_{400}\Theta_{004}}}\Big) + 4\Theta_{040}\Big(\frac{\Theta_{022}}{\Theta_{040}\Theta_{004}} - 1\Big)\Big(-1 + \theta\Theta_{400}(1 - A_2)\Big)\end{array}\right)}{8S_x^2\Theta_{040}(-1 + \frac{\Theta_{022}}{\Theta_{040}\Theta_{004}})\Big(1 + \theta\Theta_{400}(1 - A_2)\Big)},$$

and

$$\lambda_{13} = \frac{S_y^2(8 - \theta\Theta_{040})\Theta_{400}\Big(\frac{\Theta_{022}}{\sqrt{\Theta_{040}\Theta_{004}}}\frac{\Theta_{220}}{\sqrt{\Theta_{400}\Theta_{040}}} - \frac{\Theta_{202}}{\sqrt{\Theta_{400}\Theta_{004}}}\Big)}{8S_f^2\Theta_{004}(-1 + \frac{\Theta_{022}}{\sqrt{\Theta_{040}\Theta_{004}}})\Big(1 + \theta\Theta_{400}(1 - A_2)\Big)},$$

with

$$A_2 = \frac{\frac{\Theta_{220}^2}{\Theta_{400}\Theta_{040}} + \frac{\Theta_{202}^2}{\Theta_{400}\Theta_{004}} - \frac{2\Theta_{220}\Theta_{202}\Theta_{022}}{\Theta_{400}^2\Theta_{040}^2\Theta_{004}^2}}{1 - \frac{\Theta_{022}^2}{\Theta_{040}\Theta_{004}}}.$$

To follow with a similar track, the quantification of MSE of the proposed structure is provided in the following final form:

$$MSE(T_5) = \frac{\theta S_y^2\Big(64\Theta_{400}(1 - A_2) - \theta\Theta_{040}^2 - 16\theta\Theta_{400}\Theta_{040}(1 - A_2)\Big)}{64\Big(1 + \Theta_{400}(1 - A_2)\Big)}. \qquad (3.3)$$

4 Results and discussions

4.1 The simulation study

This section is dedicated to exploring the dynamic of the mean squared error (MSE) and percentage relative efficiency (PRE) associated with the proposed estimator by means of numerical

operation. Diverse parametric settings are involved by generating five finite populations each of the size of 5000 realizations. The populations are thought to be governed by multivariate normal distributions having common means of both study and auxiliary variables but varying variance-covariance matrices, such that;

$$\begin{pmatrix} Y \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \rho\sigma_y\sigma_x \\ \rho\sigma_y\sigma_x & \sigma_x^2 \end{pmatrix}\right).$$

We consider the same means as; $[Y, X] = [5, 5]$ for all the populations but the variance-covariance matrices project a varying degree of covariance structure, such as:

$$\text{Population 1} : \Sigma = \begin{bmatrix} 5 & 3 \\ 3 & 7 \end{bmatrix}$$

$$\text{Population 2} : \Sigma = \begin{bmatrix} 6 & 4.5 \\ 4.5 & 9 \end{bmatrix}$$

$$\text{Population 3} : \Sigma = \begin{bmatrix} 10 & 5.70 \\ 5.70 & 12 \end{bmatrix}$$

$$\text{Population 4} : \Sigma = \begin{bmatrix} 20 & 7.75 \\ 7.75 & 25 \end{bmatrix}$$

and

$$\text{Population 5} : \Sigma = \begin{bmatrix} 30 & 10.25 \\ 10.25 & 15 \end{bmatrix}.$$

A complete R-code is given in Appendix A. where for each generated population of size $N = 5,000$, we performed $R = 5,000$ independent iterations. In each iteration:

1. A simple random sample of size $n = 150$ was drawn without replacement from the population.
2. For each estimator T_j ($j = 0, 1, 2, 3, 4, 5$), the variance estimate was computed using the sample data.
3. The squared error for each estimator in the r-th iteration was calculated as:

   $$SE_j^{(r)} = (T_j^{(r)} - S_Y^2)^2$$

   where $T_j^{(r)}$ is the estimated variance from the r-th sample and S_Y^2 is the true population variance.
4. The empirical Mean Squared Error for each estimator was then computed as:

   $$MSE(T_j) = \frac{1}{R}\sum_{r=1}^{R}(T_j^{(r)} - S_Y^2)^2 = \frac{1}{5000}\sum_{r=1}^{5000} SE_j^{(r)}$$

5. The efficiency of the proposed and contemporary estimators is assessed using the Percentage Relative Efficiency (PRE) metric, defined as:

   $$\text{PRE} = \frac{\text{MSE}(T_0)}{\text{MSE}(T_j)} \times 100,$$

Based on the generated populations, the PREs of the proposed and contemporary estimators are computed relative to the variance of the study variable. The results, presented in Table 1, consistently demonstrate the superior performance of the proposed approach across all scenarios. This finding suggests that the proposed estimator is more precise and less susceptible to sampling variability compared to existing methods.

A practical application of our proposed estimator is now presented, supported by simulation evidence.

- The improved efficiency of the proposed estimator can lead to more accurate and reliable statistical inferences, such as hypothesis testing and confidence interval estimation.
- In practical applications, the higher efficiency of the proposed estimator may allow researchers to achieve the same level of precision with smaller sample sizes, thereby reducing costs and time.
- In policy-making and decision-making contexts, accurate and efficient statistical estimates are crucial. The proposed estimator can provide more reliable information, leading to better-informed decisions.
- It would be valuable to investigate the robustness of the proposed estimator to departures from the assumed model conditions, such as non-normality or heteroscedasticity.
- Exploring the applicability of the proposed estimator to more complex data structures, such as longitudinal or clustered data, could be a fruitful avenue for future research.
- Conducting comparative studies with a wider range of estimators, including those based on machine learning techniques, could provide further insights into the relative performance of the proposed approach.

Table 1: Percentage relative efficiency (PRE) of proposed and contemporary estimators.

Estimator	Population 1	Population 2	Population 3	Population 4	Population 5
T_0	100	100	100	100	100
T_1	66.69523	77.25621	69.99833	57.56807	67.46472
T_2	108.3465	115.3678	110.0802	102.6195	107.3446
T_3	100.5031	111.3678	103.5844	88.58833	99.4806
T_4	108.7184	115.7540	110.4612	102.9652	107.7140
T_5	108.7334	115.8988	110.4613	103.0353	107.7419

4.2 The empirical study

In this section, we provide a comprehensive detail of the data sets, which are used to evaluate the performance of the proposed with contemporary approaches. We chose five data sets of various types that are also well-known in the survey sample literature for general purposes. The used data sets' basic source and description are as follows:

- Data 1: This data set is borrowed from Singh (2003). The study variable (Y) remained the approximated sleep duration of individuals aged more than 50 years. The auxiliary variable (X) is considered the exact age of the corresponding person.

- Data 2: This data set is gathered from Gujarati (2009) reporting US egg production (in millions) in 1990 as study variable (Y), whereas, price per dozen (in cents) as the auxiliary variable (X).
- Data 3: The data is assembled from Aczel and Sounderpandian (2002), where study variable (Y) comprehends the magnitude of the US export to Singapore (in billions of Dollars). Further, the money supply figures (in billions of Dollars) are used as the supportive variable (X).
- Data 4: The data is selected from Singh (2003) reporting the estimated number of fish caught by marine recreational fishermen in 1995 as study variable (Y). Whereas, the auxiliary variable (X) is considered as the number of fish caught by marine recreational fishermen in 1994.
- Data 5: This data set is taken from Murthy (1967). The study variable (Y) is the output of the factory, whereas the auxiliary variable (X) indicates the total number of workers in the factory.

Determination of sample sizes: The sample sizes for each dataset were determined following standard survey sampling practices, where the sampling fraction (n/N) typically ranges between 10% and 30%. For each dataset, we selected a sample size that ensures:

1. A sampling fraction approximately between 10% and 20%,
2. The sample size is sufficiently large (≥ 5) to enable meaningful statistical inference, and
3. The sample sizes are comparable with those used in previous studies, such as Singh (2003) and Ahmad et al. (2022) to maintain consistency in comparative analysis.

Table 2 comprehends the detailed outlook of the selected data sets by providing relevant summary statistics, whereas empirical MSE was subsequently used to calculate the PRE values reported in Table 3. The complete R-code in Appendix A demonstrates the implementation of this procedure. The following are the key findings of our empirical study.

- Most encouragingly, it is observed that the empirical cumulative distribution function based newly proposed variance estimator outperforms the other competitors in all presented instances.
- The degree of correlation that is maintained between the research variable and the auxiliary variable continues to determine the efficiency improvement that results. The proportion of relative efficiency rises as the strength of the linear relationship between the two variables is increased.
- The minimal efficiency gain (48%) with respect to the proposed scheme is estimated to be associated with a correlation coefficient value of -0.2888328. Further, the maximum value of the percentage relative efficiency value of 1303.088 is found to be affiliated with the correlation coefficient value of 0.7751397.
- The closest competitor of the suggested estimator remained the proposition of the Ahmad et al. (2022) that is T_4. However, it is noteworthy that the new estimator surpasses the contemporary scheme in every considered scenario. The estimated performance hierarchy is noted such as; $T_5 > T_k$, where, $k = 0, 1, 2, 3, 4$.

Table 2: Summary statistics for five real-world datasets.

Parameter	Data Sets				
	Data 1	Data 2	Data 3	Data 4	Data 5
N	30	50	67	69	80
n	5	10	15	15	15
θ	0.1666667	0.08	0.05174129	0.05217391	0.05416667
$\bar{Y}$	384.2	1357.622	4.528358	4514.899	5182.637
$\bar{X}$	67.26667	78.29	6.908955	4954.435	285.125
$\bar{F}$	0.5255556	0.512	0.5226108	0.5072464	0.5067187
S_y^2	3582.579	2759726	0.6050927	37199578	3369642
S_x^2	85.23678	454.4344	1.343858	49829270	73132.09
S_f^2	0.08480332	0.08568163	0.08574828	0.08454106	0.08405041
C_y	0.1557903	1.223641	0.171779	1.350893	0.3541939
C_x	0.1372504	0.2722885	0.1677893	1.424781	0.9484593
C_f	0.5732115	0.5540996	0.5717075	0.5603176	0.5721408
ρ_{yx}	-0.8552412	-0.2888328	0.7751397	0.9601401	0.9149811
ρ_{yf}	-0.8377284	-0.2460575	0.7552868	0.7688589	0.9832342
ρ_{xf}	0.9879516	0.9460146	0.9919367	0.7543462	0.8906586
Θ_{400}	1.492915	4.623782	1.941397	6.544991	1.210342
Θ_{040}	1.092159	2.866077	0.5973322	8.69773	2.491817
Θ_{004}	0.673386	0.719782	0.7112738	0.7477145	0.7494113
Θ_{220}	0.9296233	-0.2469464	0.7347055	7.053057	1.294326
Θ_{202}	0.6902753	-0.276288	0.6680917	1.159139	0.8938889
Θ_{022}	0.8421904	0.971232	0.6438415	1.298094	0.9207891

Table 3: The percentage relative efficiency of the proposed and the contemporary estimators.

Estimator	Data 1	Data 2	Data 3	Data 4	Data 5
T_0	100	100	100	100	100
T_1	205.6847	57.9149	181.5547	575.8359	108.6964
T_2	237.6575	137.4526	197.1270	826.8919	231.5145
T_3	178.5076	82.75599	143.1683	392.7702	224.5658
T_4	250.1483	146.1565	198.7177	1018.524	242.0461
T_5	260.0297	148.4144	1303.088	1045.804	1029.126

5 Summary

This research contributes to the existing literature on estimating finite population variance by proposing a novel estimator that leverages the simultaneous use of an auxiliary variable and its empirical cumulative distribution function (ECDF) under simple random sampling. The proposed estimator's performance is evaluated through extensive numerical simulations across various parametric settings. Comparative analysis with well-established estimators

from the survey sampling literature, using five real-world datasets, consistently demonstrated the superiority of the proposed estimator. The enhanced efficiency of the new estimator, quantified by percentage relative efficiency, underscores its practical utility. While the proposed estimator offers significant advantages, it also presents certain limitations. The complexity of the estimator, involving multiple variables, hinders the derivation of exact bias and mean squared error expressions. Additionally, the computational demands of simulations for this estimator are higher compared to simpler estimators.

Future research may extend the proposed ECDF-based dual auxiliary framework to accommodate more sophisticated sampling designs, including stratified, cluster, and systematic sampling. The framework can also be adapted to estimate other finite population parameters, such as quantiles, correlation coefficients, and regression coefficients, thereby creating a comprehensive family of ECDF-based estimators.

Acknowledgment

We are deeply grateful to the Editor and the anonymous reviewers for their insightful comments and constructive suggestions, which have substantially strengthened this manuscript.

Disclosure statement

There are no conflicts of interest among authors about this work and its publication.

Data availability statement

The manuscript includes all of the used data sets.

Funding statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Ethical statement

The work has not been previously published or submitted in any other format, language, etc. The submission is entirely original.

References

Aczel AD, Sounderpandian J. Complete Business Statistics, vol. 5. McGraw Hill; 2002.

Ahmad S, Hussain S, Ullah K, et al. A simulation study: Improved ratio-in-regression type variance estimator based on dual use of auxiliary variable under simple random sampling. Plo One. 2022;17(11):e0276540.

Alomair MA, Daraz U. Dual transformation of auxiliary variables by using outliers in stratified random sampling. Mathematics. 2024;12(18):2839.

Bahl S, Tuteja RK. Ratio and product type exponential estimators. Journal of Information and Optimization Sciences. 1991;12(1):159–164.

Bai Y, Liang S, Yuan W. Estimating global gross primary production from sun-induced chlorophyll fluorescence data and auxiliary information using machine learning methods. Remote Sensing. 2021;13(5):963.

Biemer P, Peytchev A. Using geocoded census data for nonresponse bias correction: An assessment. Journal of Survey Statistics and Methodology. 2013;1(1):24–44.

Chou W, Imai K, Rosenfeld B. Sensitive survey questions with auxiliary information. Sociological Methods & Research. 2017;49(2):418–454.

Cochran WG. The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. The journal of agricultural science. 1940;30(2):262–275.

Cochran WG. Sampling Techniques. John Wiley & Sons; 1977.

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(1):64.

Daraz U, Alomair MA, Albalawi O, Al Naim AS. New techniques for estimating finite population variance using ranks of auxiliary variable in two-stage sampling. Mathematics. 2024;12(17):2741.

Daraz U, Khan M. Estimation of variance of the difference-cum-ratio-type exponential estimator in simple random sampling. RMS: Research in Mathematics & Statistics. 2021;8(1):1899402.

Daraz U, Wu J, Agustiana D, Emam W. Finite population variance estimation using Monte Carlo simulation and real life application. Symmetry. 2025;17(1):84.

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(11):1737.

Daraz U, Wu J, Alomair MA, Aldoghan LA. New classes of difference cum-ratio-type exponential estimators for a finite population variance in stratified random sampling. Heliyon. 2024;10(13):–.

Grover LK, Kaur P. Ratio type exponential estimators of population mean under linear transformation of auxiliary variable: theory and methods. South African Statistical Journal. 2011;45(2):205–230.

Grover LK, Kaur P. A generalized class of ratio type exponential estimators of population mean under linear transformation of auxiliary variable. Communications in Statistics-Theory and Methods. 2014;43(7):1552–1574.

Gujarati DN. Basic Econometrics. Tata McGraw-Hill Education; 2009.

Gupta S, Shabbir J. On improvement in estimating the population mean in simple random sampling. Journal of Applied Statistics. 2008;35(5):559–566.

Haq A, Khan M, Hussain Z. A new estimator of finite population mean based on the dual use of the auxiliary information. Communications in Statistics-Theory and Methods. 2017;46(9):4425–4436.

Hussain A, Ullah K, Cheema SA, Khan AA, Hussain Z. Empirical distribution function based dual use of auxiliary information for the improved estimation of finite population mean. Concurrency and Computation: Practice and Experience. 2022;34(27):e7346.

Irfan M, Javed M, Bhatti SH, Raza MA, Ahmad T. Almost unbiased optimum estimators for population mean using dual auxiliary information. Journal of King Saud University-Science. 2020;32(6):2835–2844.

Isaki CT. Variance estimation using auxiliary information. Journal of the American Statistical Association. 1983;78(381):117–123.

Kreuter F, Olson K, Wagner J, Yan T, Ezzati-Rice TM, et al. Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. Journal of the Royal Statistical Society Series A: Statistics in Society. 2010;173(2):389–407.

Lohr S. Sampling: Design and Analysis. Duxbury Press; 1999.

Muneer S, Khalil A, Shabbir J, Narjis G. A new improved ratio-product type exponential estimator of finite population variance using auxiliary information. Journal of Statistical Computation and Simulation. 2018;88(16):3179–3192.

Muneer S, Shabbir J, Khalil A. Estimation of finite population mean in simple random sampling and stratified random sampling using two auxiliary variables. Communications in Statistics-Theory and Methods. 2017;46(5):2181–2192.

Murthy MN. Product method of estimation. The Indian journal of statistics, series A. 1964;26(1):69–74.

Murthy MN. Sampling: Theory and Methods. Statistical Publication Society; 1967.

Pandey AK, Usman M, Singh GN. Optimality of ratio and regression type estimators using dual of auxiliary variable under non response. Alexandria Engineering Journal,. 2021;60(5):4461–4471.

Rao TJ. On certain methods of improving ratio and regression estimators. Communications in Statistics-Theory and Methods. 1991;20(10):3325–3340.

Shabbir J, Haq A, Gupta S. A new difference-cum-exponential type estimator of finite population mean in simple random sampling. Revista Colombiana de Estadística. 2014;37(1):199–211.

Singh HP, Kour SP, Kumar S, Chib R. A family of estimators for variance estimation. Research in Statistics. 2013;2(1):2350750.

Singh R, Chauhan P, Sawan N, Smarandache F. Improved exponential estimator for population variance using two auxiliary variables. Infinite Study. 2009;p. –.

Singh S. Advanced Sampling Theory With Applications: How Michael "Selected" Amy. Springer Science & Business Media; 2003.

Zaman T, Sagir M, Şahin M. A new exponential estimators for analysis of COVID-19 risk. Concurrency and Computation: Practice and Experience. 2021;34(10):e6806.

Zhang L, Chambers RL. Small area estimates for cross-classifications. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2004;66(2):479–496.

A Appendix

A simulation R-code for variance's estimators

```
library(mvtnorm)
z=rmvnorm(5000, mean = c(5,5), sigma=matrix(c(5,3,3,7), nrow=2), method = "chol")
x=z[,1]; y=z[,2]; N=length(x); n=5; h=(1/n)-(1/N); r=rank(x); f(x)
mean(x);mean(y);mean(r);mean(f(x));var(x);var(y);var(r);var(f(x))
cx=sd(x)/mean(x); cy=sd(y)/mean(y); cr=sd(r)/mean(r); cf=sd(f(x))/mean(f(x))
cor(y,x);cor(y,r);cor(y,f(x));cor(x,r);cor(x,f(x));cor(r,f(x))
a=sum((y-mean(y)) 4)/(N);b=sum((y-mean(y))\^2)/(N-1)
d=a/(b)$\wedge$2-1        ################### lambda 4000
a1=sum((x-mean(x))$\wedge$4)/(N);b1=sum((x-mean(x))$\wedge$2)/(N-1)
d1=a1/(b1)$\wedge$2-1     ################### lambda 0400
a2=sum((r-mean(r))$\wedge$4)/(N)
b2=sum((r-mean(r))$\wedge$2)/(N-1)
d2=a2/(b2)$\wedge$2-1     ################### lambda 0040
a3=sum((f(x)-mean(f(x)))$\wedge$4)/(N)
b3=sum((f(x)-mean(f(x)))$\wedge$2)/(N-1)
d3=a3/(b3)$\wedge$2-1     ################### lambda 0004
e1=sum((y-mean(y))$\wedge2$*(x-mean(x))$\wedge$2)/(N-1)
f1=e1/(var(y)*var(x))-1   ################### lambda 2200
e2=sum((y-mean(y))$\wedge$2*(r-mean(r))$\wedge$2)/(N-1)
f2=e2/(var(y)*var(r))-1   ################### lambda 2020
e3=sum((y-mean(y))$\wedge$2*(f(x)-mean(f(x)))$\wedge$2)/(N-1)
f3=e3/(var(y)*var(f(x)))-1################### lambda 2002
e4=sum((x-mean(x))$\wedge$2*(r-mean(r))$\wedge$2)/(N-1)
f4=e4/(var(x)*var(r))-1  ################### lambda 0220
e5=sum((x-mean(x))$\wedge$2*(f(x)-mean(f(x)))$\wedge$2)/(N-1)
f5=e5/(var(x)*var(f(x)))-1################### lambda 0202
e6=sum((r-mean(r))$\wedge$2*(f(x)-mean(f(x)))$\wedge$2)/(N-1)
f6=e6/(var(r)*var(f(x)))-1################### lambda 0022
##### MSE
T0=h*b$\wedge$2*d;T0
T1=h*b$\wedge$2*(d+d1-2*f1);T1
T2=h*b$\wedge$2*((d*d1-f1$\wedge$2)/(d1+h*(d*d1-f1$\wedge$2)));T2
T3=h*b$\wedge$2*(d+0.25*d1-f1);T3
q1=64*(d*d1-f1$\wedge$2)-h*d1$\wedge$3-16*h*d1*(d*d1-f1$\wedge$2);q1
q2=64*(h*(d*d1-f1$\wedge$2)+d1);q2
T4=h*(var(y))$\wedge$2*(q1/q2);T4
##### Proposed
y1=f1/sqrt(d*d1); y2=f3/sqrt(d*d3); y3=f5/sqrt(d1*d3)
t6=(y1$\wedge$2+y2$\wedge$2-2*y1*y2*y3)/(1-y3$\wedge$2)
q5=(64*d*(1-t6)-h*d1$\wedge$2-16*h*d1*d*(1-t6))
q6=64*(1+h*d*(1-t6))
T5=h*(var(y))$\wedge$2*(q5/q6)
##### PREs
(T0/T1)*100; (T0/T2)*100; (T0/T3)*100; (T0/T4)*100; (T0/T5)*100
```