An improved and robust class of variance estimator

M. Abid\textsuperscript{a,}* , R.A.Kh. Sherwani\textsuperscript{b} , M. Tahir\textsuperscript{c} , H.Z. Nazir\textsuperscript{d} , and M. Riaz\textsuperscript{d}

\textsuperscript{a} Department of Statistics, Government College University, Lahore, 54000, Pakistan.
\textsuperscript{b} College of Statistical and Actuarial Sciences, University of the Punjab Lahore, Pakistan.
\textsuperscript{c} Department of Statistics, University of Sargodha, Sargodha, Pakistan.
\textsuperscript{d} Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran, 31261, Saudi Arabia.

Received 22 February 2019; received in revised form 11 December 2019; accepted 21 September 2020

KEYWORDS
Auxiliary variable; Numerical methods; Mean square error; Monte Carlo; Outliers; Percentage relative efficiency; Simulation; Robust measures.

Abstract. The ratio, product, and regression estimators are commonly constructed based on conventional measures such as mean, median, quartiles, semi-interquartile range, semi-interquartile average, coefficient of skewness, and coefficient of kurtosis. In the case of the presence of outliers, these conventional measures lose their efficiency/performace ability and hence are less efficient as compared to those measures which performed efficiently in the presence of outliers. This study offers an improved class of estimators for estimating the population variance using robust dispersion measures such as probability-weighted moments, Gini, Downto and Bickel, and Lehmann measures of an auxiliary variable. Bias, mean square error and minimum mean square error of the suggested class of estimators have been derived. Application with two natural data sets is also provided to explain the proposal for practical considerations. In addition, a robustness study is also carried out to evaluate the performance of the proposed estimators in the presence of outliers by using environmental protection data. The results reveal that the proposed estimators perform better than their competitors and are robust, not only in simple conditions but also in the presence of outliers.

\copyright 2021 Sharif University of Technology. All rights reserved.

1. Introduction
Almost every sphere of life is prone to variation. For instance, a factory could produce five thousand piston rings of a specific size, and the exact size of a single piston ring could vary slightly, even though they are all intended to be ‘nominally’ of the same size. It is because of the unavoidable variability in the manufacturing process, many more situations (such as manufacturing industry, pharmaceutical laboratories, medical and biological sciences and agriculture sectors having populations of different nature) can come across in practice where variation is present (cf. [1]). Therefore, the estimation of population variability is very important and has a major impact on obtaining accurate results.

In survey sampling, the use of auxiliary information can be adopted in the form of a ratio, product, and regression methods of estimation to get more precise estimators of the population variance of the study variable. The ratio and product methods of estimation are useful when there exists a high positive and negative correlation between the study variable, say \( y \), and the auxiliary variable, say \( x \), respectively (cf. [2]). The ratio method of estimation loses its efficiency when the regression line of \( y \) on \( x \) does not pass through the origin and in this situation regression method of estimation is a suitable choice (cf. [2]).

---

* Corresponding author.
E-mail addresses: mabid@gcu.edu.pk (M. Abid); rehan.stat@pu.edu.pk (R.A.Kh. Sherwani); tahirquastat@yahoo.com (M. Tahir); hafiznazir@yahoo.com (H.Z. Nazir); riaz76quad@yahoo.com (M. Riaz)

Various types of estimators have been developed in the context of ratio, product, and regression estimators for estimating population variance using the known auxiliary information based on different conventional measures such as mean, median, quartiles, semi-interquartile range, semi-interquartile average, coefficient of skewness, coefficient of kurtosis, etc. In this regard, the works [3–25].

A major disadvantage related with the estimators proposed by Isaki [3], Upadhyaya and Singh [4], Kadilar and Cingi [5], Subramani and Kumarapandiyam [6–9], and Khan and Shabbir [10] is that these estimators lose their efficiency and performance ability in the presence of outliers (cf. [25–27]). For case with outliers, robust estimators are preferred and are of more practical use, because these estimators are not affected by the outliers. The development of a ratio estimator of variance that offers efficiency and robustness in the presence of outliers and measurement error in the data is a neglected area in survey sampling. So, the main concern of this work is to develop such kinds of estimators that perform well in the presence of outliers.

In this study, a general class of estimators using robust measures of dispersion such as Probability-Weighted Moments (PWM), Gini, Downton, and Bickel and Lehmann measures for estimating population variance of the study variable. A major advantage of assuming these measures is their ability to be stable in the presence of outliers. The theoretical properties of these robust measures have been thoroughly investigated by Muhammad and Riaz [28], Gini [29], Downton [30], and Bickel and Lehmann [31]; and these robust measures are considered to enhance the efficiency of the ratio estimator of variance in Simple Random Sampling Without Replacement (SRSWOR) scheme.

The rest of the article is planned as follows: Section 2 offers a comprehensive explanation of the existing variance estimators. The formation and performance assessment of the suggested estimators with existing estimators are given in Section 3. Section 4 includes a practical and robustness study of suggested estimators. In the end, concluding remarks are presented in Section 5.

### 2. Existing estimators

Isaki [3] proposed a usual ratio estimator to estimate the population variance ($\hat{S}_v^2$) of the study variable by using the known auxiliary information on the population variance ($S^2_x$) of the auxiliary variable and it is defined as:

$$\hat{S}_v^2 = s_y^2 \left( \frac{S^2_x}{\sigma^2_x} \right), \quad \hat{s}_v^2 \neq 0,$$

where $s_y^2$ and $s_x^2$ are the sample mean of the study and auxiliary variables.

The approximate bias and Mean Square Error (MSE) of the Isaki [3] is given as:

$$\text{Bias} \left( \hat{S}_v^2 \right) \approx \psi S^2_y \left( (\beta_2(x) - 1) - (\lambda_2 - 1) \right), \quad (2)$$

$$\text{MSE} \left( \hat{S}_v^2 \right) \approx \psi S^4_y \left( [\beta_2(y) - 1] + [\beta_2(x) - 1] - 2(\lambda_2 - 1) \right), \quad (3)$$

Upadhyaya and Singh [4] suggested an estimator to investigate population variance and showed that their suggested estimator is more efficient in comparison to the Isaki [3]. Upadhyaya and Singh [4] estimator is written as:

$$\hat{S}_{v_2}^2 = s_y^2 \left( \frac{s^2_X + \beta_2(x)}{s^2_x + \beta_2(x)} \right),$$

where $\beta_2(x)$ is the coefficient of kurtosis.

Based on the known auxiliary information on the coefficient of variation ($C_x$) and $\beta_2(x)$, Kadilar and Cingi [5] suggested some new estimators as follows:

$$\hat{S}_{v_4}^2 = s_y^2 \left( \frac{S^2_X + C_X}{s^2_x + C_x} \right), \quad \hat{S}_{v_5}^2 = s_y^2 \left( \frac{\beta_2(x) S^2_X + C_X}{\beta_2(x) s^2_x + C_x} \right),$$

$$\hat{S}_{v_5}^2 = s_y^2 \left( \frac{C_x s^2_X + \beta_2(x)}{C_x s^2_x + \beta_2(x)} \right).$$

Subramani and Kumarapandiyam [6] proposed an estimator by utilizing the information on the median ($M_d$) of an auxiliary variable and it is defined below:

$$\hat{S}_{v_6}^2 = s_y^2 \left( \frac{s^2_X + M_d}{s^2_x + M_d} \right).$$

Based on quartiles and their function of an auxiliary variable, Subramani and Kumarapandiyam [7] developed various ratio estimators which are shown below:

$$\hat{S}_{v_7}^2 = s_y^2 \left( \frac{S^2_X + Q_1}{s^2_x + Q_1} \right), \quad \hat{S}_{v_8}^2 = s_y^2 \left( \frac{S^2_X + Q_3}{s^2_x + Q_3} \right),$$

$$\hat{S}_{v_9}^2 = s_y^2 \left( \frac{S^2_X + Q_2}{s^2_x + Q_2} \right), \quad \hat{S}_{v_{10}}^2 = s_y^2 \left( \frac{s^2_X + Q_4}{s^2_x + Q_4} \right),$$

$$\hat{S}_{v_{11}}^2 = s_y^2 \left( \frac{S^2_X + Q_4}{s^2_x + Q_4} \right),$$

where $Q_1$, $Q_3$, $Q_r$, $Q_d$, and $Q_a$ are the first quartile, third quartile, inter-quartile range, semi-quartile range, and semi-quartile average, respectively.

To incorporate the information on the deciles of an auxiliary variable, Subramani and Kumarapandiyam [8] introduced the following estimators:

$$\hat{S}_{v_{12}}^2 = s_y^2 \left( \frac{s^2_X + D_1}{s^2_x + D_1} \right), \quad \hat{S}_{v_{13}}^2 = s_y^2 \left( \frac{s^2_X + D_2}{s^2_x + D_2} \right),$$
\[
\begin{align*}
S_{E14}^2 &= s_y^2 \left( \frac{S_X^2 + D_3}{s_x^2 + D_3} \right), & S_{E15}^2 &= s_y^2 \left( \frac{S_X^2 + D_4}{s_x^2 + D_4} \right), \\
S_{E16}^2 &= s_y^2 \left( \frac{S_X^2 + D_5}{s_x^2 + D_5} \right), & S_{E17}^2 &= s_y^2 \left( \frac{S_X^2 + D_6}{s_x^2 + D_6} \right), \\
S_{E18}^2 &= s_y^2 \left( \frac{S_X^2 + D_7}{s_x^2 + D_7} \right), & S_{E19}^2 &= s_y^2 \left( \frac{S_X^2 + D_8}{s_x^2 + D_8} \right), \\
S_{E20}^2 &= s_y^2 \left( \frac{S_X^2 + D_9}{s_x^2 + D_9} \right), & S_{E21}^2 &= s_y^2 \left( \frac{S_X^2 + D_{10}}{s_x^2 + D_{10}} \right).
\end{align*}
\]

Making use of the information of \( C_X \) and \( M_d \), Subramani and Kumarapandiy [9] introduced an estimator which is given below:

\[
S_{E22}^2 = s_y^2 \left( \frac{C_X S_X^2 + M_d}{C_X s_x^2 + M_d} \right).
\]

Khan and Shabbir [10] have suggested an estimator based on the coefficient of correlation (\( \rho \)) and \( Q_3 \) which is given below:

\[
S_{E23}^2 = s_y^2 \left( \frac{\rho S_X^2 + Q_0}{\rho s_x^2 + Q_0} \right).
\]

The bias and MSE of the existing estimators suggested by Upadhyaya and Singh [4], Kadilar and Cingi [3], Subramani and Kumarapandiy [6–9], and Khan and Shabbir [10] i.e., \( S_{Ei}^2, i = 2, 3, \ldots, 23 \) up to the first degree of approximation are listed below:

\[
\begin{align*}
\text{Bias} \left( S_{E1}^2 \right) &\approx \eta S_y^2 \left[ R_i^2 (\lambda_{04} - 1) - R_i (\lambda_{22} - 1) \right], \\
\text{MSE} \left( S_{E1}^2 \right) &\approx \eta S_y^4 \left[ (\lambda_{04} - 1) + R_i^2 (\lambda_{04} - 1) - 2R_i (\lambda_{22} - 1) \right].
\end{align*}
\]

where:

\[
\begin{align*}
R_2 &= \frac{S_X^2}{S_X^2 + \beta_2(X)}, & R_3 &= \frac{S_X^2}{S_X^2 + C_X}, \\
R_4 &= \frac{C_X S_X^2}{C_X S_X^2 + \beta_2(X)}, & R_5 &= \frac{\beta_2(X) S_X^2}{\beta_2(X) S_X^2 + C_X}, \\
R_6 &= \frac{S_X^2}{S_X^2 + M_d}, & R_7 &= \frac{S_X^2}{S_X^2 + Q_1}, \\
R_8 &= \frac{S_X^2}{S_X^2 + Q_3}, & R_9 &= \frac{S_X^2}{S_X^2 + Q_4}, \\
R_{10} &= \frac{S_X^2}{S_X^2 + Q_d}, & R_{11} &= \frac{S_X^2}{S_X^2 + Q_a}, \\
R_{12} &= \frac{S_X^2}{S_X^2 + D_1}, & R_{13} &= \frac{S_X^2}{S_X^2 + D_2}, \\
R_{14} &= \frac{S_X^2}{S_X^2 + D_3}, & R_{15} &= \frac{S_X^2}{S_X^2 + D_4}.
\end{align*}
\]

\[
\begin{align*}
R_{16} &= \frac{S_X^2}{S_X^2 + D_5}, & R_{17} &= \frac{S_X^2}{S_X^2 + D_6}, \\
R_{18} &= \frac{S_X^2}{S_X^2 + D_7}, & R_{19} &= \frac{S_X^2}{S_X^2 + D_8}, \\
R_{20} &= \frac{S_X^2}{S_X^2 + D_9}, & R_{21} &= \frac{S_X^2}{S_X^2 + D_{10}}, \\
R_{22} &= \frac{C_X S_X^2}{C_X S_X^2 + M_d}, & R_{23} &= \frac{\rho S_X^2}{\rho S_X^2 + Q_3}.
\end{align*}
\]

3. The proposed class of estimators

In this section, we propose a general class of ratio estimator of variance to evaluate \( S_y^2 \) by adopting the information on robust measures of dispersion such as PWM, Gini, Downton and Bickel, and Lehmann measures of an auxiliary variable. The PWM are investigated by Muhammad and Riaz [28] and they are defined as:

\[
PWM = \frac{\sqrt{\pi}}{\sqrt{2}} \sum_{i=1}^{N} (2i - N - 1) X(i),
\]

where \( X(i) \) is \( i \)th order statistic of the population. The PWM is more efficient than conventional measures against outliers and therefore, provides more effective estimators than conventional measures (cf. [26]). The next measure included in this study is the Gini (\( G \)) measure suggested by Gini [29] and it is stated as:

\[
G_X = \frac{4}{N - 1} \sum_{i=1}^{N} \left( \frac{2i - N - 1}{2N} \right) X(i).
\]

The \( G \) measure is also more robust than the conventional measures in the presence of outliers (cf. [32]). The Downton (\( D \)) measure is introduced by Downton [30] and is written as:

\[
D = \frac{2\sqrt{\pi}}{N(N-1)} \sum_{i=1}^{N} \left( i - \frac{N + 1}{2} \right) X(i).
\]

The main purpose to include \( D \) in this study is that it remains stable to outliers (cf. [26]). The last measure included in this study was suggested by Bickel and Lehmann [31] and hereafter named as \( B_n \). The \( B_n \) measure is obtained by replacing pairwise averages with pairwise distances and it is defined as:

\[
B_n = 1.0483 (\text{median} |x_i - x_j| : i < j).
\]

This robust measure also has an efficiency to perform better against outliers in the data (cf. [33]).

The proposed class of estimators to estimate \( S_y^2 \) is defined as:
\[
\hat{S}_p^2 = \lambda \left[ 2 \left( \frac{\psi S_X^2}{\psi S_X^2 + \delta} \right) \frac{S_Y^2}{S_X^2 + \delta} \right] - \Delta_j,
\]

where \( \lambda \) is the constant lies between 0 ≤ \( \lambda \) ≤ 1, \( \psi(\neq 0) \) and \( \delta \) are functions of the known parameters of auxiliary variable such as PWM, G, D, \( B_n \), \( C_x \), \( T \), etc.

To find the bias, the MSE, and the minimum MSE of the suggested class of estimators, \( \hat{S}_p^2 \), we used the following conventions:

\[
\varepsilon_0 = \frac{s_y^2 - s_p^2}{s_y^2}, \quad \varepsilon_1 = \frac{s_x^2 - s_p^2}{s_x^2},
\]

Further, we can write:

\[
s_y^2 = S_y^2(1 + \varepsilon_0) \quad \text{and} \quad s_x^2 = S_x^2(1 + \varepsilon_1),
\]

where \( \varepsilon_0 \) and \( \varepsilon_1 \) are the relative error of study and auxiliary variables, respectively.

Using the definition of \( \varepsilon_0 \) and \( \varepsilon_1 \), it can be shown that \( E(\varepsilon_0) = E(\varepsilon_1) = 0 \). Also \( E(\varepsilon_0^2) = \eta(\lambda_{40} - 1) \), \( E(\varepsilon_1^2) = \eta(\lambda_{41} - 1) \), and \( E(\varepsilon_0\varepsilon_1) = \eta(\lambda_{22} - 1) \), where:

\[
\lambda_{40} = \mu_{40}/\mu_{20}, \quad \lambda_{41} = \mu_{41}/\mu_{02},
\]

\[
\lambda_{22} = \mu_{22}/\mu_{02}\mu_{04},
\]

\[
\lambda_{rs} = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^r (X_i - \bar{X})^s,
\]

\[
S_Y^2 = (N - 1)^{-1} \sum_{i=1}^{N} (Y_i - \bar{Y})^2,
\]

\[
S_X^2 = (N - 1)^{-1} \sum_{i=1}^{N} (X_i - \bar{X})^2,
\]

\[
s_y^2 = (n - 1)^{-1} \sum_{i=1}^{n} (y_i - \bar{y})^2,
\]

and:

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

So, Eq. (6) can be written in terms of \( \varepsilon_0 \) and \( \varepsilon_1 \) as follows:

\[
\hat{S}_p^2 = \lambda \left[ S_y^2(1 + \varepsilon_0)(1 + \Delta j \varepsilon_1)^{-1} \right],
\]

\[
(1 + \Delta_j) \Delta_{41} = \frac{\rho S_X^2}{\rho S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{42} = \frac{\rho S_X^2}{\rho S_X^2 + PWM},
\]

\[
(1 + \Delta_j) \Delta_{43} = \frac{C_x S_X^2}{C_x S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{44} = \frac{S_X^2}{S_X^2 + G}.
\]

\[
\Delta_0 = \frac{\rho S_X^2}{\rho S_X^2 + PWM}, \quad \Delta_1 = \frac{S_X^2}{S_X^2 + D}, \quad \Delta_2 = \frac{\rho S_X^2}{\rho S_X^2 + D}, \quad \Delta_3 = \frac{C_x S_X^2}{C_x S_X^2 + PWM}, \quad \Delta_4 = \frac{S_X^2}{S_X^2 + G}.
\]

Expanding Eq. (7) up to the first degree of approximation provides:

\[
\hat{S}_p^2 - \hat{S}_0^2 = \lambda \left[ S_y^2 \left( 1 + \varepsilon_0 - \Delta_j \varepsilon_1 + \frac{1}{2} \Delta_j^2 (1 + \Delta_j) \varepsilon_1 \right) \right],
\]

\[
\Delta_j = \frac{S_X^2}{S_X^2 + D}, \quad \Delta_8 = \frac{\rho S_X^2}{\rho S_X^2 + D}, \quad \Delta_9 = \frac{C_x S_X^2}{C_x S_X^2 + D}, \quad \Delta_{10} = \frac{S_X^2}{S_X^2 + B_n}, \quad \Delta_{11} = \frac{\rho S_X^2}{\rho S_X^2 + B_n}, \quad \Delta_{12} = \frac{C_x S_X^2}{C_x S_X^2 + B_n}.
\]

So, Eq. (6) can be written in terms of \( \varepsilon_0 \) and \( \varepsilon_1 \) as follows:

\[
\hat{S}_p^2 = \lambda \left[ S_y^2(1 + \varepsilon_0)(1 + \Delta j \varepsilon_1)^{-1} \right],
\]

\[
(1 + \Delta_j) \Delta_{41} = \frac{\rho S_X^2}{\rho S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{42} = \frac{\rho S_X^2}{\rho S_X^2 + PWM},
\]

\[
(1 + \Delta_j) \Delta_{43} = \frac{C_x S_X^2}{C_x S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{44} = \frac{S_X^2}{S_X^2 + G}.
\]

The bias of an estimator \( \hat{S}_p^2 \) is defined as:

\[
\text{Bias}\left( \hat{S}_p^2 \right) = E \left( \hat{S}_p^2 - S_0^2 \right).
\]

Applying expectation on both sides of Eq. (8), we get:

\[
E \left( \hat{S}_p^2 - S_0^2 \right) = (\lambda - 1) S_y^2 + \lambda \left[ S_y^2 \left( E(\varepsilon_0) - \Delta_j E(\varepsilon_0 \varepsilon_1) \right) \right],
\]

\[
(1 + \Delta_j) \Delta_{41} = \frac{\rho S_X^2}{\rho S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{42} = \frac{\rho S_X^2}{\rho S_X^2 + PWM},
\]

\[
(1 + \Delta_j) \Delta_{43} = \frac{C_x S_X^2}{C_x S_X^2 + PWM}, \quad (1 + \Delta_j) \Delta_{44} = \frac{S_X^2}{S_X^2 + G}.
\]

The MSE of an estimator \( \hat{S}_p^2 \) is defined as:

\[
\text{MSE}\left( \hat{S}_p^2 \right) = E \left( \hat{S}_p^2 - S_0^2 \right)^2.
\]

Thus, squaring and then applying expectation on both sides of Eq. (8), we get:

\[
E \left( \hat{S}_p^2 - S_0^2 \right)^2 = [(\lambda - 1)^2 S_y^2 + \eta \lambda^2 S_y^2 \left( E(\varepsilon_0) - \Delta_j E(\varepsilon_0 \varepsilon_1) \right)] + \eta S_y^2 \left( \lambda^2 \Delta_j^2 (1 + 2 \Delta_j) - \lambda \Delta_j (1 + \Delta_j) \right) E(\varepsilon_0^2) \]

\[
+ \eta S_y^2 \left( \lambda^2 (1 + 2 \Delta_j) - \lambda (2 \Delta_j) E(\varepsilon_0 \varepsilon_1) \right) E(\varepsilon_0^2).
\]
After substituting the values of $E(x_0^2)$, $E(x_1^2)$ and $E(\xi_0 \xi_1)$ in the above equation, we obtain:

$$MSE\left(\hat{S}_y^2\right) \approx (\lambda - 1)^2 S_y^4 + \Gamma_1 (\lambda_{o4} - 1)$$

$$+ \Gamma_2 (\lambda_{o2} - 1) - \Gamma_3 (\lambda_{o2} - 1).$$

(10)

where

$$\Gamma_1 = \eta \lambda^2 S_y^4,$$

$$\Gamma_2 = \eta S_y^4 \left\{ \lambda^2 \Delta_j^3 (1 + 2 \Delta_j) - \lambda \Delta_j^3 (1 + \Delta_j) \right\},$$

$$\Gamma_3 = \eta S_y^4 \left\{ \lambda^2 (4 \Delta_j^2) - \lambda (2 \Delta_j^2) \right\}.$$

Differentiating Eq. (10) with respect to $\lambda$ and then setting it equal to zero, we get the equation shown in Box I, where:

$$A = 2 S_y^4 + \eta S_y^4 \left\{ \Delta_j^3 (1 + \Delta_j) (\lambda_{o4} - 1) - 2 \Delta_j^3 (\lambda_{o2} - 1) \right\},$$

$$B = 2 S_y^4 + \eta S_y^4 (\lambda_{o4} - 1) + \Delta_j^3 (1 + 2 \Delta_j) (\lambda_{o4} - 1) - 4 \Delta_j^2 (\lambda_{o2} - 1).$$

Putting $\lambda_{opt}$ in Eq. (10), we obtain the minimum MSE of $\hat{S}_y^2$:

$$MSE\left(\hat{S}_y^2\right)_{\min} = \left[ S_y^2 - \frac{A^2}{2B S_y^2} \right].$$

(11)

3.1. Some members of the proposed class of estimators

In this subsection, we will propose some new estimators which belong to the proposed class of estimators given in Eq. (6) by substituting different values of $\lambda$, $\psi$, and $\delta$.

**Remark 3.1.** If we put $(\lambda, \psi, \delta) = (1, 0, 1)$ in Eq. (6) then the proposed class reduces to the traditional simple random sampling estimator.

$$S_{E0}^2 = S_y^2.$$

**Remark 3.2.** If we substitute $(\lambda, \psi, \delta) = (1, 1, 0)$ in Eq. (6) then proposed class belongs to the Isaki [3] estimator as follows:

$$\hat{S}_{E1}^2 = S_y^2 \left( \frac{s_x^2}{S_y^2} \right), \quad s_x^2 \neq 0.$$

**Remark 3.3.** If we put different choices of $\psi(1, \rho, \tau_x)$ and $\delta (PWM, G, D, B_n)$ in Eq. (6) then some new members belong to the proposed class $\hat{S}_y^2$ as below:

$$\lambda_{opt} = \frac{2 S_y^4 + \eta S_y^4 \left\{ \Delta_j^3 (1 + \Delta_j) (\lambda_{o4} - 1) - 2 \Delta_j^3 (\lambda_{o2} - 1) \right\}}{2 S_y^4 + \eta S_y^4 (\lambda_{o4} - 1) + \Delta_j^3 (1 + 2 \Delta_j) (\lambda_{o4} - 1) - 4 \Delta_j^2 (\lambda_{o2} - 1)} = \frac{A}{B}.$$

Box I
these comparisons between the proposed and existing estimators. The performance comparison is carried out on the basis of MSE. The proposed family of estimators is compared with the estimators suggested by Upadhyaya and Singh [4], Kadilar and Cingi [5], Subramani and Kumarapandiyam [6-9], and Khan and Shabbir [10].

3.2. Efficiency comparisons

In this section, we find the conditions in which the proposed estimators perform more efficiently in comparison to the traditional and the existing ratio estimators of variance.

3.2.1. Comparison with traditional ratio estimator

The proposed estimators are more efficient than the traditional ratio estimator suggested by Isaki [3], if they have smaller values of MSE against Isaki [3] estimator. Mathematically, it is defined as:

$$MSE\left( \hat{S}_p^2 \right)_{\text{min}} < MSE\left( \hat{S}_a^2 \right)_{\text{Isaki}}.$$  

By Eqs. (3) and (11):

$$S_y^2 \left[ \frac{A^2}{2BS_y^2} \right] < \eta S_y^2 (\lambda_40 - 1) + (\lambda_{44} - 1)$$

$$- 2(\lambda_{22} - 1).$$

After some simplification, we get:

$$A^2 > 2BS_y^2 \left\{ 1 - \eta(\lambda_{40} + \lambda_{44} - 2\lambda_{22}) \right\}. \quad (12)$$

3.2.2. Comparison with existing estimators

The proposed estimators perform better if the values of MSE of the suggested estimators are lesser than the values of MSE of existing estimators proposed by Upadhyaya and Singh [4], Kadilar and Cingi [5], Subramani and Kumarapandiyam [6-9], and Khan and Shabbir [10] and in algebraic form, it is expressed as:

$$MSE\left( \hat{S}_p^2 \right)_{\text{min}} < MSE\left( \hat{S}_a^2 \right)_{\text{other}}, \quad i = 2, 3, \ldots, 23.$$  

By Eqs. (5) and (11):

$$S_y^2 \left[ \frac{A^2}{2BS_y^2} \right] < \eta S_y^2 (\lambda_40 - 1) + R_t^2 (\lambda_{44} - 1)$$

$$- 2R_t (\lambda_{22} - 1).$$

After solving the above equation, we obtain:

$$A^2 > 2BS_y^2 \left\{ 1 - \eta(\lambda_{40} - 1) + R_t^2 (\lambda_{44} - 1) \right\}$$

$$- 2R_t (\lambda_{22} - 1). \quad (13)$$

If the conditions given in Eqs. (12) and (13) are held then it indicates the supremacy of the proposed estimators against the existing estimators discussed in Section 2.

In the next section, we present the empirical and outliers study comparisons between the proposed and the existing estimators.

4. Empirical study

To assess the performance of the suggested estimators against their competing estimators, we use two real populations. The descriptions of two populations are given below:

**Population I** (Source: Murthy [34], page 228):

$$Y = \text{Output}, \quad X = \text{The number of workers},$$

$$N = 80, \quad n = 20, \quad \bar{Y} = 51.826,$$

$$\bar{X} = 2.851, \quad \rho = 0.915, \quad S_y = 18.357,$$

$$C_y = 0.354, \quad S_x = 2.704, \quad C_x = 0.948,$$

$$\lambda_{04} = 3.581, \quad \lambda_{40} = 2.267, \quad \lambda_{22} = 2.323,$$

$$M_d = 1.480, \quad Q_1 = 0.865, \quad Q_3 = 4.453,$$

$$Q_e = 3.588, \quad Q_d = 1.794, \quad Q_a = 2.659,$$

$$PWM = 2.448, \quad G = 2.797, \quad D = 2.479,$$

$$B_n = 2.101.$$

**Population II** (Source: Murthy [34], page 228):

$$Y = \text{Output}, \quad X = \text{The fixed capital},$$

$$N = 80, \quad n = 20, \quad \bar{Y} = 51.826,$$

$$\bar{X} = 11.265, \quad \rho = 0.941, \quad S_y = 18.357,$$

$$C_y = 0.354, \quad S_x = 8.456, \quad C_x = 0.751,$$

$$\lambda_{04} = 2.806, \quad \lambda_{40} = 2.267, \quad \lambda_{22} = 2.221,$$

$$M_d = 7.575, \quad Q_1 = 5.150, \quad Q_3 = 16.975,$$

$$Q_e = 11.825, \quad Q_d = 5.913, \quad Q_a = 11.063,$$

$$PWM = 7.913, \quad G = 9.040, \quad D = 8.013,$$

$$B_n = 6.714.$$

The validity and utility of an estimator are generally assessed by the metric called MSE (cf. [4,5,25]). The MSE measures the divergence of the estimator ($\hat{\theta}$) values from the true parameter ($\theta$) value and mathematically it is defined as:

$$MSE = E \left( \hat{\theta} - \theta \right)^2. \quad (14)$$

We also find the Percentage Relative Efficiencies (PREs) of proposed and existing estimators with respect to the traditional ratio estimator of variance.
The mathematical expressions of PREs are given below:

\[
PRE(E_i; E_1) = \frac{MSE\left(\hat{S}_{E_i}^2\right)}{MSE\left(\hat{S}_{E_1}^2\right)},
\]

(15)

\[
PRE(P_j; E_1) = \frac{MSE\left(\hat{S}_{P_j}^2\right)}{MSE\left(\hat{S}_{E_1}^2\right)},
\]

(16)

where, \( i = 2, 3, \ldots, 23 \), \( j = 1, 2, \ldots, 9 \). The values of MSE and PREs are given in Table 1.

To get more insight into the study, we also used a new measure called the Percentage Decrease (PD) in MSE which hereafter is named as \( MSE_{PD} \). The values of \( MSE_{PD} \) between existing and proposed estimators are given in Table 2. An estimator which has a larger \( MSE_{PD} \) is considered to be better as compared to other estimators. The \( MSE_{PD} \) can be computed as:

\[
MSE_{PD} = \frac{(MSE_E - MSE_{E_1})}{MSE_E} \times 100.
\]

From the analysis of Tables 1 and 2, the key observations about the study could be summarized as follows:

i. By comparing all existing estimators, it is found that the performance of the estimators \( \hat{S}_{E_9}^2 \) and \( \hat{S}_{E_{20}}^2 \) is relatively better for Populations I and II, respectively;

ii. By comparing all proposed estimators with the usual ratio estimator and existing estimators, it is observed that proposed estimators perform more efficiently due to smaller MSE values and higher PREs values;

iii. It is seen that proposed estimators have lesser values of MSE ranges from (2393, 2421) and (2195, 2102) against existing estimators for Populations I and II, respectively;

iv. It is also observed that the proposed estimators have higher values of PREs ranges from (209.81, 213.62) and (140.05, 150.97) against their competing estimators for Populations I and II, respectively;

v. The estimators \( \hat{S}_{P_8}^2 \) and \( \hat{S}_{P_9}^2 \) have smaller values of MSE (2393, 1950) and higher values of PREs (213.62, 150.97) by comparing all other proposed estimators for Populations I and II, respectively;

vi. The estimator \( \hat{S}_{P_8}^2 \) has a larger \( MSE_{PD} \) against other proposed estimators for Population I in comparison to existing estimators (cf. Table 2).

A graphical comparison between the proposed and the existing estimators is also made. Most efficient estimators from the existing estimators suggested by Isaki [3], Upadhyaya and Singh [4], Kadilar and Cingi [5], Subramani and Kumarapandiyam [6–9], and Khan and Shabbir [10] and from the proposed estimators are chosen and then compared. Figure 1(a) and (b), depict that the proposed estimators have smaller values of MSE in comparison to the existing
estimators which shows that proposed estimators are more efficient than the existing estimators.

### 4.1. Simulation study

A simulation study is also carried out to assess the performance of the existing and proposed estimators. The statistical programming language R is used to carry out the simulation study. The following procedure is adopted to compute the MSE of proposed and existing estimators:

1. Generating a random sample of size $n = 30$ and 40 from the bivariate normal distribution;
2. Computing the MSE of the proposed and existing estimators by using the expressions given in Eqs. (3), (5), and (11) of the random samples generated in Step (i);
3. Repeating Steps (i) and (ii) 20000 times to obtain MSEs;
4. Averaging these MSEs to obtain the value of MSE of proposed and existing estimators.

The results of the simulation study in terms of MSEs are reported in Table 3 and Figure 2(a) and (b). Following are the main findings based on these results:

- The proposed estimators are more efficient in comparison to the usual and existing ratio estimators of variance;
- As the value of $n$ increases, the value of MSE decreases, and vice versa;
- From all the proposed estimators, the estimator $S^2_{p_1}$ has the smaller MSE values for each choice of $n$.

### 4.2. Performance of the proposed estimators in case of outliers

As mentioned in the previous section, the estimators suggested in this study are robust and efficient for case with outliers in the data. So, the efficiency of the proposed estimators against outliers is evaluated in this section. For the said purpose, two natural populations taken from the Italian national institute for environment protection and research (cf. [35]) are considered. The comparison between proposed and
Figure 1. Mean Square Error (MSE) values of the proposed and existing estimators for (a) Population I and (b) Population II.

existing estimators for case with outliers is also done on the basis of MSE and PREs.

The characteristics of the two populations are listed below:

**Population III** (Source: ISPR [35]):

\[ Y = \text{Total amount of recycled waste collection in Italy in 2003}, \]

\[ X = \text{The number of inhabitants in 2003}, \]

\[ N = 103, \quad n = 40, \quad \bar{Y} = 626.212, \]

\[ X = 557.191, \quad \rho = 0.994, \quad S_y = 913.541, \]

\[ C_y = 1.457, \quad S_x = 818.112, \quad C_x = 1.468, \]

\[ \lambda_{04} = 37.322, \quad \lambda_{40} = 37.128, \quad \lambda_{22} = 37.206, \]

\[ M_d = 308.050, \quad Q_1 = 142.995, \quad Q_3 = 665.625, \]

\[ Q_r = 522.630 \quad Q_d = 261.315, \quad Q_a = 404.310, \]

\[ PWM = 556.424, \quad G = 633.885, \quad D = 561.879, \]

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( n = 30 )</th>
<th>( n = 40 )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Usual</strong></td>
<td><strong>MSE</strong></td>
<td><strong>MSE</strong></td>
</tr>
<tr>
<td>( \hat{S}_R )</td>
<td>2840.26</td>
<td>1704.16</td>
</tr>
<tr>
<td><strong>Existing:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( \hat{S}_{E1} )</td>
<td>1544.49</td>
<td>926.69</td>
</tr>
<tr>
<td>( \hat{S}_{E2} )</td>
<td>2237.78</td>
<td>1342.67</td>
</tr>
<tr>
<td>( \hat{S}_{E3} )</td>
<td>2639.79</td>
<td>1583.88</td>
</tr>
<tr>
<td>( \hat{S}_{E4} )</td>
<td>1522.53</td>
<td>913.52</td>
</tr>
<tr>
<td>( \hat{S}_{E5} )</td>
<td>2011.95</td>
<td>1207.17</td>
</tr>
<tr>
<td>( \hat{S}_{E6} )</td>
<td>2279.36</td>
<td>1367.61</td>
</tr>
<tr>
<td>( \hat{S}_{E7} )</td>
<td>1463.32</td>
<td>877.99</td>
</tr>
<tr>
<td>( \hat{S}_{E8} )</td>
<td>1543.69</td>
<td>926.21</td>
</tr>
<tr>
<td>( \hat{S}_{E9} )</td>
<td>1905.39</td>
<td>1143.23</td>
</tr>
<tr>
<td>( \hat{S}_{E10} )</td>
<td>1688.15</td>
<td>1012.89</td>
</tr>
<tr>
<td>( \hat{S}_{E11} )</td>
<td>2298.22</td>
<td>1438.93</td>
</tr>
<tr>
<td>( \hat{S}_{E12} )</td>
<td>2325.77</td>
<td>1395.46</td>
</tr>
<tr>
<td>( \hat{S}_{E13} )</td>
<td>2333.15</td>
<td>1339.89</td>
</tr>
<tr>
<td>( \hat{S}_{E14} )</td>
<td>2108.65</td>
<td>1265.19</td>
</tr>
<tr>
<td>( \hat{S}_{E15} )</td>
<td>2011.95</td>
<td>1207.17</td>
</tr>
<tr>
<td>( \hat{S}_{E16} )</td>
<td>1835.39</td>
<td>1101.24</td>
</tr>
<tr>
<td>( \hat{S}_{E17} )</td>
<td>1543.51</td>
<td>926.10</td>
</tr>
<tr>
<td>( \hat{S}_{E18} )</td>
<td>1430.51</td>
<td>858.30</td>
</tr>
<tr>
<td>( \hat{S}_{E19} )</td>
<td>1391.23</td>
<td>834.74</td>
</tr>
<tr>
<td>( \hat{S}_{E20} )</td>
<td>1468.12</td>
<td>880.87</td>
</tr>
<tr>
<td>( \hat{S}_{E21} )</td>
<td>1982.94</td>
<td>1189.77</td>
</tr>
<tr>
<td>( \hat{S}_{E22} )</td>
<td>1437.99</td>
<td>862.80</td>
</tr>
</tbody>
</table>

**Proposed:**

\( \hat{S}_{P1} \) 1368.40 827.85
\( \hat{S}_{P2} \) 1366.42 826.30
\( \hat{S}_{P3} \) 1365.74 825.84
\( \hat{S}_{P4} \) 1365.69 825.45
\( \hat{S}_{P5} \) 1369.00 827.05
\( \hat{S}_{P6} \) 1366.61 825.83
\( \hat{S}_{P7} \) 1366.24 826.25
\( \hat{S}_{P8} \) 1366.42 825.74
\( \hat{S}_{P9} \) 1365.44 825.30
\( \hat{S}_{P10} \) 1388.25 841.11
\( \hat{S}_{P11} \) 1373.94 831.66
\( \hat{S}_{P12} \) 1378.94 835.00
Figure 2. Mean Square Error (MSE) values of the proposed and existing estimators for simulation study: (a) \( n = 30 \) and (b) \( n = 40 \).

\[ B_n = 370.155. \]

Population IV (Source: ISPRA [35]):

- \( Y \) = Total amount of recycled waste collection in Italy in 2003,
- \( X \) = Total amount of recycled waste collection in Italy in 2002,

\[ N = 103, \quad n = 40, \quad \bar{Y} = 62.621, \]

\[ X = 556.554, \quad P = 0.730, \quad S_y = 91.354, \]

\[ C_y = 1.459, \quad S_x = 610.164, \quad C_x = 1.006, \]

\[ \lambda_{04} = 17.874, \quad \lambda_{40} = 37.128, \quad \lambda_{22} = 17.222, \]

\[ M_d = 373.820, \quad Q_1 = 259.383, \quad Q_{20} = 628.023, \]

\[ Q_r = 388.641, \quad Q_d = 183.320, \quad Q_a = 443.703. \]

Figure 3. Boxplot of (a) total amount of recycled waste collection in Italy in 2003, (b) the number of inhabitants in 2003, and (c) total amount of recycled waste collection in Italy in 2002.

\[ PWM = 423.086, \quad G = 481.985, \quad D = 427.234, \]

\[ B_n = 250.776. \]

In addition, boxplots of Populations III and IV are provided to display that these populations contain outliers. It is obvious from Figure 3(a)–(c), that there
Table 4. Values of Mean Square Error (MSE) and Percentage Relative Efficiency (PRE) in the presence of outliers.

<table>
<thead>
<tr>
<th>Estimators</th>
<th>Population III</th>
<th>Population IV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>PRE</td>
</tr>
<tr>
<td></td>
<td></td>
<td>100</td>
</tr>
<tr>
<td><strong>Existing:</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$S_E^1$</td>
<td>4099007390</td>
<td>100.03</td>
</tr>
<tr>
<td>$S_E^2$</td>
<td>4010410270</td>
<td>100.00</td>
</tr>
<tr>
<td>$S_E^3$</td>
<td>4010463050</td>
<td>100.00</td>
</tr>
<tr>
<td>$S_E^4$</td>
<td>4099530093</td>
<td>100.02</td>
</tr>
<tr>
<td>$S_E^5$</td>
<td>4089906140</td>
<td>100.26</td>
</tr>
<tr>
<td>$S_E^6$</td>
<td>4095358732</td>
<td>100.12</td>
</tr>
<tr>
<td>$S_E^7$</td>
<td>4079712800</td>
<td>100.51</td>
</tr>
<tr>
<td>$S_E^8$</td>
<td>4083523660</td>
<td>100.41</td>
</tr>
<tr>
<td>$S_E^9$</td>
<td>4091402020</td>
<td>100.22</td>
</tr>
<tr>
<td>$S_E^{10}$</td>
<td>4086941360</td>
<td>100.33</td>
</tr>
<tr>
<td>$S_E^{11}$</td>
<td>4086417550</td>
<td>100.04</td>
</tr>
<tr>
<td>$S_E^{12}$</td>
<td>4096178960</td>
<td>100.10</td>
</tr>
<tr>
<td>$S_E^{13}$</td>
<td>4093653130</td>
<td>100.15</td>
</tr>
<tr>
<td>$S_E^{14}$</td>
<td>4092604610</td>
<td>100.19</td>
</tr>
<tr>
<td>$S_E^{15}$</td>
<td>4089061140</td>
<td>100.26</td>
</tr>
<tr>
<td>$S_E^{16}$</td>
<td>4086507550</td>
<td>100.34</td>
</tr>
<tr>
<td>$S_E^{17}$</td>
<td>4081378400</td>
<td>100.46</td>
</tr>
<tr>
<td>$S_E^{18}$</td>
<td>4077961133</td>
<td>100.55</td>
</tr>
<tr>
<td>$S_E^{19}$</td>
<td>4065769700</td>
<td>100.85</td>
</tr>
<tr>
<td>$S_E^{20}$</td>
<td>1235356840</td>
<td>96.40</td>
</tr>
<tr>
<td>$S_E^{21}$</td>
<td>4093094760</td>
<td>100.18</td>
</tr>
<tr>
<td>$S_E^{22}$</td>
<td>4079603180</td>
<td>100.51</td>
</tr>
</tbody>
</table>

**Proposed:**

| $S_P^1$    | 4037111170     | 101.57        | 14171434       | 154.50        |
| $S_P^2$    | 4037153860     | 101.57        | 14176295       | 154.45        |
| $S_P^3$    | 4035575300     | 101.61        | 14170280       | 154.51        |
| $S_P^4$    | 4038109340     | 101.54        | 14173202       | 154.48        |
| $S_P^5$    | 4038167870     | 101.54        | 14178797       | 154.42        |
| $S_P^6$    | 4035923930     | 101.60        | 14171948       | 154.50        |
| $S_P^7$    | 4037175520     | 101.57        | 14171563       | 154.50        |
| $S_P^8$    | 4037218980     | 101.57        | 14176171       | 154.45        |
| $S_P^9$    | 4035969990     | 101.61        | 14170937       | 154.51        |
| $S_P^{10}$ | 4035263800     | 101.61        | 14160803       | 154.56        |
| $S_P^{11}$ | 4035339520     | 101.61        | 14168866       | 154.53        |
| $S_P^{12}$ | 4035115220     | 101.62        | 14165398       | 154.57        |

exist outliers in Populations III and IV, so we believe that the proposed class of estimators perform well against their competitor estimators.

The MSE and PREs values of the existing and the proposed estimators in the presence of outliers are reported in Table 4. It is noted that when outliers are present, the suggested estimators based on PWM, G, D, and $B_n$ show good resistance and perform efficiently against the existing estimators (cf. Table 4). On the other hand, the existing estimators perform badly in the presence of outliers in comparison to the proposed estimators (larger MSE values and smaller PREs values, (cf. Table 4)). For case with outliers, the estimator $S_{P12}$ performs well followed by the other proposed estimators for Populations III and IV (cf. Table 4). So, we can say that proposed estimators offer higher resistance against outliers as compared to the existing estimators considered in this study.

5. Concluding remarks

This study proposes a class of estimators for estimating population variance using robust measures of dispersion including probability-weighted moments, Gini, Downton, and $B_n$ measures. The performance of the proposed estimators is compared with the competing estimators suggested by Isaki [3]. Upadhyaya and Singh [4], Kadilar and Cingi [5], Subraman and Kumarapandian [6-9], and Khan and Shabbir [10]. The mean square error and the percentage relative efficiency are used as performance metrics. It is found that the suggested estimators offer higher efficiency than existing estimators due to smaller values of mean square error and higher percentage relative efficiency. When there are outliers in the data, the proposed estimators based on robust dispersion measures offer quite robust behavior relative to the existing estimators. In brief, the proposed estimators based on PWM, G, D and $B_n$ behave very well under empirical and outliers studies. The scope of the current work can also be extended to other sampling techniques such as stratified sampling, ranked set sampling, and systematic sampling.

References


Biographies

Muhammad Abid obtained his MSc and MPhil degrees in Statistics from Quaid-i-Azam University, Islamabad, Pakistan, in 2008 and 2010, respectively. He did his PhD in Statistics from the Institute of Statistics, Zhejiang University, Hangzhou, China, in 2017. He served as a Statistical Officer in the National Accounts Wing, Pakistan Bureau of Statistics (PBS) during 2010–2011. He is now serving as an Assistant Professor in the Department of Statistics, Government College University, Faisalabad, Pakistan, from 2017 to the present. He has published more than 60 research papers in research journals. His research interests include statistical process control, Bayesian statistics, non-parametric techniques, and survey sampling.

Rehan Ahmed Khan Sherwani is working as an Associate Professor in the College of Statistical and Actuarial Sciences, University of the Punjab, Lahore, Pakistan. His fields of interest are multilevel models, statistical process control, and computational statistics.

Muhammad Tahir obtained his PhD in Statistics from Quaid-i-Azam University, Islamabad, Pakistan, in 2017. Currently, he is an Assistant Professor of Statistics at Government College University, Faisalabad, Pakistan. He has published more than 25 research papers in national and international reputed journals. His research interests include Bayesian inference, reliability analysis, and mixture distributions.

Hafiz Zafar Nazir obtained his MSc and MPhil degrees in Statistics from the Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan, in 2006 and 2008, respectively. He received his PhD in Statistics from the Institute of Business and Industrial Statistics University of Amsterdam, The Netherlands, in 2014. He served as a Lecturer in the Department of Statistics from University of Sargodha, Pakistan, from 2009–2014. He is now serving as an Assistant Professor in the Department of Statistics, University of Sargodha, Pakistan, from 2015 until the present. His current research interests include statistical process control, non-parametric techniques, and robust methods.

Muhammad Riaz obtained his PhD in Statistics from the Institute for Business and Industrial Statistics, University of Amsterdam, The Netherlands, in 2008. He holds the position of Professor in the Department of Mathematics and Statistics, King Fahd University of Petroleum and Minerals, Dhahran, Saudi Arabia. His current research interests include statistical process control, non-parametric techniques, and experimental design.