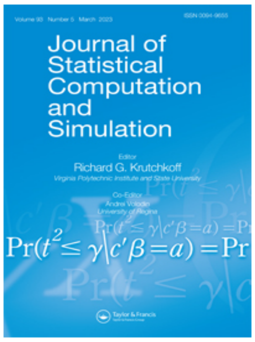

Journal of Statistical Computation and Simulation
Richard G. Krutchkoff
$\Pr(t^2 \leq \gamma | c'\beta = a) = \Pr$

Journal of Statistical Computation and Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Robust ratio estimators of population mean for skewed and contaminated population

Azaz Ahmed, Aamir Sanaullah, Evrim Oral & Muhammad Hanif

To cite this article: Azaz Ahmed, Aamir Sanaullah, Evrim Oral & Muhammad Hanif (2023) Robust ratio estimators of population mean for skewed and contaminated population, Journal of Statistical Computation and Simulation, 93:5, 800-817, DOI: 10.1080/00949655.2022.2120994

To link to this article: https://doi.org/10.1080/00949655.2022.2120994

Published online: 16 Sep 2022.

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
2023, VOL. 93, NO. 5, 800–817
https://doi.org/10.1080/00949655.2022.2120994

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

Robust ratio estimators of population mean for skewed and contaminated population

Azaz Ahmed[a], Aamir Sanaullah[b], Evrim Oral[c] and Muhammad Hanif[a]

[a]Department of Statistics, National College of Business Administration and Economics, Lahore, Pakistan; [b]Department of Statistics, COMSATS University Islamabad – Lahore Campus, Lahore, Pakistan; [c]Department of Statistics, LSU Health Sciences Center, New Orleans, LA, USA

ABSTRACT

The efficiency of the traditional ratio estimator decreases with the presence of outliers, inliers or when the underlying distribution is not normal. To improve the efficiency, we propose new robust ratio estimators utilizing the Lloyd's robust estimator and modified maximum likelihood estimator (MMLE), and provide their theoretical properties. We calculate the mean square error and relative efficiency of the proposed estimators and compare its performance with some other existing traditional estimators via a simulation study under various contamination or misspecification models. Further, we evaluate the exact value of the mean square error and to compare the performance of the proposed estimators using the numerical illustrations.

ARTICLE HISTORY
Received 21 January 2022
Accepted 31 August 2022

KEYWORDS
Auxiliary variable; robust ratio estimator; Weibull distribution; Lloyd's estimator; mean square error; outliers; inliers

1. Introduction

For practitioners, the sample mean $\bar{y}$ $(=\sum_{j=1}^{n} y_j/n)$ is an obvious choice to estimate the population mean. In order to enhance the precision of an estimate, the ratio estimator is suggested to be utilized by Cochran [1], for situations where the study variable is linearly correlated with the auxiliary variable, whose values are assumed to be known for each unit of the population. The ratio estimator is a good choice for estimating population mean $\bar{Y}$ if the survey population is assumed to be modelled as the realization of a super-population [2]. The study variable in the population is assumed to be the realization of the random variable $y_j, j = 1, 2, \ldots, N$, according to

$$y_j = \beta x_j + h(x_j)e_j, \tag{1}$$

where x_j is an auxiliary variable, $h(x_j)$ is a function of x_j and e_j is *i.i.d* random variable with mean zero and variance σ_e^2 that is assumed to follow a distribution $g(e_j)$, and $g(e_j)$ is the probability density function of e_j. We refer to the assumed model of Equation (1) as the working model; the accuracy of the estimates and statistical inference depend on the accuracy of the assumptions made on the working model. The Gaussian assumption on

CONTACT Azaz Ahmed azazpu17@gmail.com Department of Statistics, National College of Business Administration and Economics, Lahore, Pakistan

$g(e_j)$ provides the traditional ratio estimator,

$$\bar{y}_1 = ({}^{\bar{y}}\!/\!_{\bar{x}})\bar{X}, \tag{2}$$

with the mean square error (MSE)

$$\mathrm{MSE}(\bar{y}_1) = \mathrm{var}(\bar{y}) - 2R\mathrm{cov}(\bar{y},\bar{x}) + R^2\mathrm{var}(\bar{x}), \tag{3}$$

where $\bar{x} = \sum_{j=1}^{n} \bar{x}_j/n$, $\bar{X} = \sum_{j=1}^{N} x_j/N$, $\mathrm{var}(\bar{x}) = [(1-f)/n]S_x^2$, $\mathrm{var}(\bar{y}) = [(1-f)/n]S_y^2$, $\mathrm{cov}(\bar{y},\bar{x}) = \frac{1-f}{n}S_{xy}, f = n/N$ and $R = \bar{Y}/\bar{X}$.

The traditional ratio estimator in Equation (2) is more efficient than usual sample mean estimator if it fulfils the condition $\rho > V_x/2V_y$, where $V_x = S_x/\bar{X}$ and $V_y = S_y/\bar{Y}$; S_x and S_y are the standard deviations of x and y over the values of entire population, respectively, and ρ is the population correlation coefficient between x and y.

There are real-life situations where violations on a working model are expected to occur. Misspecification occurs when one assumes that the population under study follows a theoretical model, whereas in reality, a different model describes it better. In this situation, one should question if the estimate is still reliable. Contamination occurs when a few values of the sample data are outliers. These outliers might be so extreme that ignoring them or proceeding with them can lead to a seriously biased inference. In contrast of outliers, there are observations that belong to central tendency but are in error. These observations are known as inliers. It is difficult to identify the inlier. In under-developed countries, official statistics are fabricated like gross domestic product (GDP) and literacy rate. Due to fabricate information (inlier), results are not reliable. For this situation, one needs to modify the estimation procedure in order to make the estimator robust to outliers or inliers. This study explores a novel method for these cases so that the quality of statistical inference based on standard population models can be maintained where there is misspecification or contamination or inliers in the data.

In survey sampling, robust estimation of the mean in the presence of outliers and $g(e_j) \sim N(0,\sigma^2)$ has been discussed by a few authors to advocate the use of robust ratio-type estimators. For instance, [3,4] utilize M-estimation to robustify ratio-type estimators. Moreover, Refs. [5–9] provided robust estimators using the auxiliary information with robust regression methods.

The properties of ratio estimators were studied for both misspecification and contamination by Refs. [10–12] following the model (1) where $g(e_j)$ was assumed to be from a long-tailed symmetric (LTS) family. They integrated Tiku's modified maximum likelihood estimator (MMLE) into the ratio-type estimators. They showed that, when $g(e_j)$ follows a LTS distribution, or data have outliers, integrating MMLE provides more efficient estimates than the traditional ratio estimator; see also Tiku and Vellaisamy's [13].

Recently, Sanaullah et al. [14] highlighted various problems with MMLE's weight function and suggested to use the generalized least squares estimation (GLSE) as an alternative for the robustification process. They showed that when $g(e_j)$ follows a LTS distribution, integrating GLSE into the traditional ratio estimator yields more efficient ratio estimators with respect to the traditional, or MMLE integrated ratio estimator.

In many real-life applications, the assumption of normality is not realistic and the data can be fit by skewed distribution. Usually survey data about family expenditures, family size and household budget, reveal skewed distributions. In such cases, skewed distribution is a natural choice.

Thus, in this study, we assume that $g(e_j)$ is characterized by the skewed distribution in Equation (1). We integrate GLSE and MMLE into the traditional ratio estimator and study their properties and robustness under both misspecification and contamination models.

The rest of the paper is organized as follows: We provide MMLE and GLSE for the parameters of the Weibull distribution in Section 2. In Section 3, we construct the robust ratio estimators, derive their MSEs, and provide the conditions for which they perform better than the traditional ratio estimator. Furthermore, we construct the ratio estimator which is robust with respect to inliers and derive its properties. In Section 4, to provide numerical evidence, the performance of the proposed robust ratio estimators is evaluated in terms of the relative efficiencies by a simulation study. The practicality of the proposed estimators is justified by applying the proposed robust ratio estimators to a real-life data. In final section, the conclusion is presented.

2. MMLE and GLSE for location and scale parameters of the Weibull population

In this section, the parameter estimation of Skewed population using MMLE and GLSE is discussed. In order to exemplify the skewed population, the Weibull distribution is chosen because it is very flexible in terms of shapes. It covers the nearly symmetric to highly skewed shapes. Assume that the working model (1) follows the Weibull distribution,

$$f(y) = \frac{p}{\sigma_y^p}(y-\mu_y)^{p-1}exp\left[-\left(\frac{y-\mu_y}{\sigma_y}\right)^p\right],\ \mu_y < y < \infty\ ,\ p > 0 \tag{4}$$

where μ_y, σ_y and p are the location, scale and shape parameters, respectively. It should be noted that μ_y and σ_y are not the mean and standard deviation of random variable y but are related to the population mean θ and variance ϑ^2, respectively, when underlying distribution of the study variable y is Weibull distribution as,

$$\theta = \mu_y + \Gamma(1+1/p)\sigma_y \text{ and } \vartheta^2 = \sigma_y^2\left[\Gamma\left(1+\frac{2}{p}\right) - \left(\Gamma\left(1+\frac{1}{p}\right)\right)^2\right]. \tag{5}$$

The estimation of the location and scale parameters under known shape parameter is discussed in two cases, namely Case-I and Case-II. In Case-I, it is assumed that μ_y is unknown and σ_y is known. In Case-II, both μ_y and σ_y are unknown.

2.1. MMLE for Case-I

Let $y_{(1)} \le y_{(2)} \le \ldots \le y_{(n)}$ be the order statistics of the random sample $y_1, y_2, \ldots, y_n$ from (4). For deriving explicit estimator of μ_y, the MMLE is obtained by linearizing the likelihood equation around the first two terms of the Taylor series expansion. The solution of the equation is the MMLE following Case-I (see Ref. [15]):

$$\hat{\mu}_y = K - (\Psi/m)\sigma_y, \tag{6}$$

where $K = \sum_{j=1}^{n} \gamma_j y_{(j)}$; $\gamma_j = \delta_j/m, \delta_j = (p-1)\beta_{j0} + p\beta_j, m = \sum_{j=1}^{n} \delta_j, \psi_j = (p-1)\alpha_{j0} - p\alpha_j,$ $\Psi = \sum_{j=1}^{n} \psi_j, \beta_{j0} = t_{(j)}^{-2}, \beta_j = (p-1)t_{(j)}^{p-2}, \alpha_j = (2-p)t_{(j)}^{p-1}, \alpha_{j0} = 2t_{(j)}^{p-1}$ and $t_{(j)}$ takes the expected values of the order statistics of standardized variable $v = \frac{y-\mu_y}{\sigma_y}$ from (4) which may be determined by,

$$\int_{-\infty}^{t_{(j)}} f(v)dv = j/(n+1), f(v) = p(v)^{p-1} exp[-(v)^p], 0 < v < \infty. \tag{7}$$

It is to note that the MMLE for (4) is available for the Case-II, but we did not consider in this study because they are not useful in providing the robust estimator. This difficulty is not with the GLS estimators. Therefore, we consider it for both cases.

2.1.1. GLSE for Case-I

Following Lloyd [16]'s GLS estimation, we let $y_1, y_2, \ldots, y_n$ be a simple random sample (SRS) from (4). Let $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ be the order statistics from the sample above and $x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$ be the concomitants. Let $V_{(j)} = \frac{y_{(j)}-\mu_y}{\sigma_y}$ $(j = 1, 2, \ldots, n)$ be the standardized variate in (4). Let the means, variances and covariances of the order statistics $V_{(j)}$ be denoted by $t_{(j)}$, ω_{jj} and ω_{ji}, respectively. Further, let $\mathbf{y}' = (y_{(1)}, y_{(2)}, \ldots, y_{(n)})$, $\mathbf{t}' = (t_{(1)}, t_{(2)}, \ldots, t_{(n)})$, $\mathbf{1}' = (1, 1, \ldots, 1)$ and $\mathbf{\Omega} = \omega_{ji}$ for $j, i = 1, 2, \ldots, n$. The best linear unbiased estimator (BLUE) of μ_y for Case-I is given by the Chen and Chang [17] as,

$$\hat{\mu}_y^* = \frac{\mathbf{1}'\Omega\,\mathbf{y}}{\mathbf{1}'\Omega\mathbf{1}} - \frac{\mathbf{1}'\mathbf{\Omega}t'}{\mathbf{1}'\mathbf{\Omega}\mathbf{1}}\sigma_y, \; \hat{\mu}_y^* = \sum_{j=1}^{n} \phi_j y_{(j)} - \sum_{j=1}^{n} \phi_j t_{(j)}\,\sigma_y \tag{8}$$

where $\phi_j = \sum_{i=1}^{n} \nu_{ji} / \sum_{i=1}^{n}\sum_{j=1}^{n} \nu_{ji}$; ν_{ji} are the elements of the inverse matrix $\mathbf{\Omega}$ of order statistics of the Weibull distribution see, David and Nagaraja [18]. The elements of $\mathbf{\Omega}$ can be determined from the equation given as,

$$\omega_{ij} \cong p_i(1-p_j)/((n+2)f(u_i)f(u_j)), \tag{9}$$

where $p_i = i/(n+1), p_j = j/(n+1), F(u_i) = p_i$ and $f(v) = F'(v)$ is the density function of parent distribution, i.e. (4).

2.1.2. GLSE for Case-II

The GLSE of μ_y and σ_y for the Case-II are given by Lloyd [16] as,

$$\hat{\mu}_{y1}^* = \sum_{j=1}^{n} g_{1_j} y_{(j)} \text{ and } \hat{\sigma}_{y1}^* = \sum_{j=1}^{n} g_{2j} y_{(j)}, \tag{10}$$

where g_{1j} $(1 \leq j \leq n)$ are the elements of the vector

$$\mathbf{g_1} = \frac{1}{\Delta}(\mathbf{t}'(\mathbf{\Omega})^{-1}\mathbf{t}\mathbf{1}' - (\mathbf{1}'(\mathbf{\Omega})^{-1}\mathbf{t}')\mathbf{t}')(\mathbf{\Omega})^{-1}, \tag{11}$$

and g_{2j} $(1 \leq j \leq n)$ are the elements of the vector

$$\mathbf{g_2} = \frac{1}{\Delta}(-\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{t1}'+(\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{1})\mathbf{t}')(\mathbf{\Omega})^{-\mathbf{1}}, \tag{12}$$

where $\Delta = (\mathbf{t}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{t}\,(\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{1}) - (\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{t})^2)$.

The variances and covariance of (10) are given as:

$$\mathrm{var}(\hat{\mu}^*_{y1}) = \frac{\sigma_y^2}{\Delta}\mathbf{t}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{t},\ \mathrm{var}(\hat{\sigma}^*_{y1}) = \frac{\sigma_y^2}{\Delta}\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{1} \text{ and } \mathrm{cov}(\hat{\mu}^*_{y1}, \hat{\sigma}^*_{y1}) = -\frac{\sigma_y^2}{\Delta}\mathbf{1}'(\mathbf{\Omega})^{-\mathbf{1}}\mathbf{t}, \tag{13}$$

We can calculate $t_{(j)}$ and $\mathbf{\Omega}$ from Equations (7) and (9). Using $t_{(j)}$ and $\mathbf{\Omega}$, one may get the solutions for each of the Equations (6), (8) and (10)–(13).

3. Proposed robust ratio estimators

3.1. Proposed robust ratio estimators following case-I

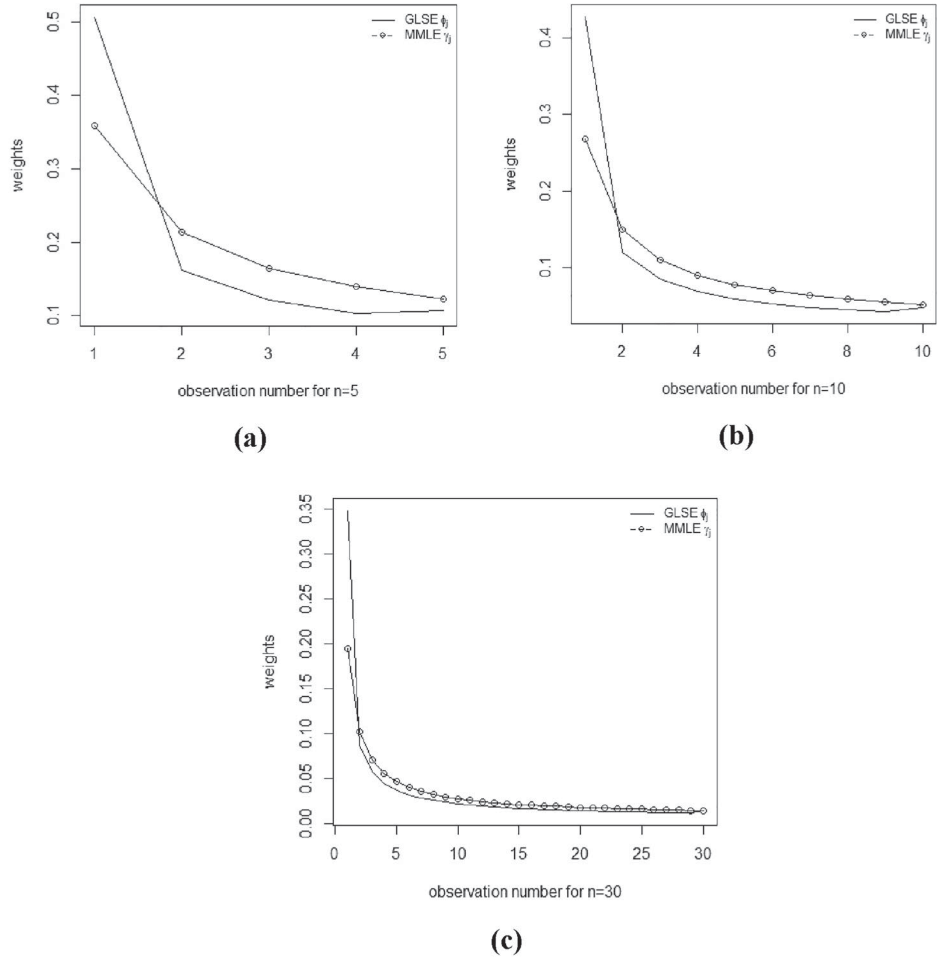
To evaluate the weight functions in (6) and (8) of MMLE and GLSE, we calculate the values of the coefficients γ_j and ϕ_j, respectively for $n = 5$, 10 and 30 using $t_{(j)}$ and $\mathbf{\Omega}$ for $p = 2$. These coefficients are presented in Figure 1. In Figure 1, it can be seen that the functions γ_j and ϕ_j allocate higher weights to the starting observations and lower weights to the upper extreme observations and accordingly, the upper extreme values get minimum weights, which trim down the effects of the outliers in a data set. In addition, it can be clearly noted that, the weight function ϕ_j as compared to the weight function γ_j assigns more appropriate weights not only to the starting observations but also to the upper extreme observations for small sample size say $n = 5$ and $n = 10$, see Figure 1(a,b). So the GLSE is considered to be a better alternate to the MMLE which imparts a more robust estimator of $\bar{Y}$ through the weight function ϕ_j in (8) for small sample. As the sample size increases say $n = 30$ the two weight functions γ_j and ϕ_j perform equally, see Figure 1(c). Consequently, both the coefficients γ_j and ϕ_j due to half umbrella ordering, provided one-sided robustness which is the upper extreme part of data.

Now in this Section, we advise to put forward these traits of MMLE and GLSE into a traditional ratio estimator and construct the new robust ratio estimators under the assumption that the working model (1) characterize the skewed distribution given in (4) with $E(e_j)$, $\mathrm{var}(e_j) = \sigma_e{}^2$ and $h(x_j) = 1$, where $j = 1, 2, \ldots, n$.

We propose the following robust ratio estimators as follows

$$\bar{y}_{pk} = \frac{\hat{\theta}_k}{\bar{x}}\bar{X},\ k = 1, 2, \tag{14}$$

where $\hat{\theta}_k$ for $k = 1$ and 2, are the MMLE and the GLSE of the mean θ for the population (4), respectively, as $\hat{\theta}_1 = \hat{\mu}_y + \Gamma(1 + 1/p)\sigma_y$ and $\hat{\theta}_2 = \hat{\mu}^*_y + \Gamma(1 + 1/p)\sigma_y$; where $\hat{\mu}_y$ and $\hat{\mu}^*_y$ are the MMLE and the GLSE estimators of location parameter μ_y following the Case-I which are given in (6) and (8), respectively.



(a)

(b)

(c)

Figure 1. Behaviour of the weights functions γ_j and ϕ_j in MMLE and GLSE to estimate the population mean.

In order to derive the approximate MSE of (14) under working model (1), we may consider (14) as,

$$\bar{y}_{pk} - \bar{Y} = \frac{\hat{\theta}_k}{\bar{x}}\bar{X} - \bar{Y} \Rightarrow \bar{X}(\hat{R}_k - R) \text{ where } \hat{R}_k = \frac{\hat{\theta}_k}{\bar{x}}, \text{and } R = \frac{\bar{Y}}{\bar{X}}. \tag{15}$$

Following the Taylor series approximation of $\hat{R}_k - R$ about $(\bar{X}, \bar{Y})$,

$$g(\bar{x}, \hat{\theta}_k) \cong g(\bar{X}, \bar{Y}) + (\bar{x} - \bar{X}) \left| \frac{\partial g(\bar{x}, \hat{\theta}_k)}{\partial \bar{x}} \right|_{\substack{\bar{x}=\bar{X} \\ \hat{\theta}_k=\bar{Y}}} + (\hat{\theta}_k - \bar{Y}) \left| \frac{\partial g(\bar{x}, \hat{\theta}_k)}{\partial \hat{\mu}_y^*} \right|_{\substack{\bar{x}=\bar{X} \\ \hat{\theta}_k=\bar{Y}}}, \tag{16}$$

where $g(\bar{x}, \hat{\theta}_k) = \hat{R}_k$ and $g(\bar{X}, \bar{Y}) = R$.

Now (15) can be expanded using (16) as,

$$\hat{R}_k - R \cong (\hat{\theta}_k - \bar{Y})\frac{1}{\bar{X}} - (\bar{x} - \bar{X})\frac{\bar{Y}}{\bar{X}^2},$$

or

$$\bar{X}(\hat{R}_k - R) \cong (\hat{\theta}_k - \bar{Y}) - (\bar{x} - \bar{X})R. \tag{17}$$

Taking square and then applying expectation to (17), we get,

$$\bar{X}^2 E(\hat{R}_k - R)^2 \cong E(\hat{\theta}_k - \bar{Y})^2 + E(\bar{x} - \bar{X})^2 R^2 - 2R(E(\hat{\theta}_k - \bar{Y})(\bar{x} - \bar{X})),$$

or the MSE$(\bar{y}_{pk})$ is obtained as,

$$\text{MSE}(\bar{y}_{pk}) \cong \text{var}(\hat{\theta}_k) + R^2\text{var}(\bar{x}) - 2R\text{cov}(\hat{\theta}_k, \bar{x}), \tag{18}$$

where $\text{var}(\hat{\theta}_k)$ and $\text{cov}(\hat{\theta}_k, \bar{x})$ are derived as, for $k = 1$, $\text{var}(\hat{\theta}_1) = \text{var}(\hat{\mu}_y) \Rightarrow (\boldsymbol{\gamma}'\,\boldsymbol{\Omega}\gamma)\,\sigma_y^2$ and for $k = 2$, $\text{var}(\hat{\theta}_2) = \text{var}(\hat{\mu}_y^*) \Rightarrow (\phi'\,\boldsymbol{\Omega}\,\phi)\,\sigma_y^2$, where $\boldsymbol{\gamma}$ and ϕ are the vectors consisting of the elements of the coefficients γ_j and ϕ_j, respectively. The $\text{cov}(\hat{\theta}_k, \bar{x})$ may be given as, for $k = 1$, $\text{cov}(\hat{\theta}_1, \bar{x}) = \text{cov}(\hat{\mu}_y + \Gamma(1 + 1/p)\sigma_y, \bar{x})$, or

$$\text{cov}(\hat{\theta}_1, \bar{x}) = \text{cov}(\sum_{j=1}^{n} \gamma_j y_{(j)} - \Psi/m + \Gamma(1 + 1/p)\sigma_y, \bar{x}),$$

as the entire sums will not be altered due to ordering, i.e.$\sum_{j=1}^{n} x_j = \sum_{j=1}^{n} x_{[j]}$, so

$$\text{cov}(\hat{\theta}_1, \bar{x}) = \text{cov}(\sum_{j=1}^{n} \gamma_j y_{(j)} - \Psi/m + \Gamma(1 + 1/p)\sigma_y, \sum_{j=1}^{n} x_{[j]}/n),$$

or alternatively,

$$\text{cov}(\hat{\theta}_1, \bar{x}) = \text{cov}(\sum_{j=1}^{n} \gamma_j[\mu_y + \sigma_y V_{(j)}] - \Psi/m + \Gamma(1 + 1/p)\sigma_y, \sum_{j=1}^{n} [\mu_x + \sigma_x U_{[j]}]/n), \tag{19}$$

where $V_{(j)} = \frac{y_{(j)} - \mu_y}{\sigma_y}$ and $U_{[j]} = \frac{x_{[j]} - \mu_x}{\sigma_y}$ and the covariance between $V_{(j)}$ and $U_{[j]}$ can be given by $\text{cov}(V_{(j)}\,,\, U_{[j]}) = \rho\sigma_x\sigma_y\boldsymbol{\Omega}$, see Refs. [19,20]. The expression of (19) after simplification may be given in matrix notation by,

$$\text{cov}(\hat{\theta}_1, \bar{x}) = \rho\sigma_x\sigma_y\,\boldsymbol{\gamma}'\,\boldsymbol{\Omega}\,\boldsymbol{\delta}, \tag{20}$$

where $\boldsymbol{\delta}$ is the $n \times 1$ vector with elements ${}^{1}\!/_{n}$ and $\boldsymbol{\gamma}' = (\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_n)$.

In similar manner, we may be obtained $\mathrm{cov}(\hat{\theta}_k, \bar{x})$ for $k = 2$ as,

$$\mathrm{cov}(\hat{\theta}_2, \bar{x}) = \mathrm{cov}(\hat{\mu}_y^* + \Gamma(1 + 1/p)\sigma_y, \bar{x}),$$

or

$$= \mathrm{cov}(\sum_{j=1}^{n} \phi_j y_{(j)} - \sum_{j=1}^{n} \phi_j t_{(j)}\, \sigma_y + \Gamma(1 + 1/p)\sigma_y, \bar{x}),$$

after simplification, the expression may also be given in matrix notation as

$$\mathrm{cov}(\hat{\theta}_2, \bar{x}) = \rho\sigma_x\sigma_y\, \phi' \,\boldsymbol{\Omega}\, \boldsymbol{\delta}, \tag{21}$$

where $\boldsymbol{\delta}$ is the $n \times 1$ vector with elements $^1\!/_n$ and $\phi' = (\phi_1, \phi_2, \ldots, \phi_n)$.

Finally, the expressions of $\mathrm{MSE}(\bar{y}_{p1})$ and $\mathrm{MSE}(\bar{y}_{p2})$are obtained respectively as,

$$\mathrm{MSE}(\bar{y}_{p1}) \cong (\boldsymbol{\gamma}' \,\boldsymbol{\Omega}\gamma)\, \sigma_y^2 + R^2 \mathrm{var}(\bar{x}) - 2R\rho\sigma_x\sigma_y\, \boldsymbol{\gamma}' \,\boldsymbol{\Omega}\, \boldsymbol{\delta} \tag{22}$$

and

$$\mathrm{MSE}(\bar{y}_{p2}) \cong (\phi' \,\boldsymbol{\Omega}\phi)\, \sigma_y^2 + R^2 \mathrm{var}(\bar{x}) - 2R\rho\sigma_x\sigma_y\, \phi' \,\boldsymbol{\Omega}\, \boldsymbol{\delta}. \tag{23}$$

3.1.1. Proposed robust ratio estimator following case-II and its properties

In this section, we proposed robust ratio estimator using GLSE following Case-II when the study variable follows (4).

Considering model (1) using (4) with $E(e_j)$, $\mathrm{var}(e_j) = \sigma_e{}^2$ and $h(x_j) = 1$, where $j = 1, 2, \ldots, n$. In order to estimate the population mean of y assuming the population mean of the auxiliary variable say $\bar{X}$ is known, we propose robust ratio estimator as,

$$\bar{y}_{p3} = \frac{\hat{\theta}_3}{\bar{x}} \bar{X} \tag{24}$$

where $\hat{\theta}_3 = \hat{\mu}_{y1}^* + \Gamma(1 + 1/p)\hat{\sigma}_{y1}^*$; where $\hat{\mu}_{y1}^*$ and $\hat{\sigma}_{y1}^*$ are the GLSE estimators respectively of μ_y and σ_y which are given in (10).

By substituting the (10), we have,

$\hat{\theta}_3 = \sum\limits_{j=1}^{n} g_{1j} y_{(j)} + \Gamma(1 + 1/p) \sum\limits_{j=1}^{n} g_{2j} y_{(j)}$, where g_{1j} and g_{2j} are given in (11) and (12).

Or alternatively,

$\hat{\theta}_3 \ = \sum\limits_{j=1}^{n} g_j^* y_{(j)}$, where $g_j^* = g_{1j} + \Gamma(1 + 1/p) g_{2j}$. (25)

Thus, the estimator $\hat{\theta}_3$ depends on the weights g_j^* which is the function of g_{1j} and g_{2j}. The values of g_j^* are calculated using the (11) and (12) following the (4) with $p = 2$ and $n = 10$, and given as,

0.1839, 0.0756, 0.0698, 0.0708, 0.0743, 0.0793, 0.0856, 0.0939, 0.1065, 0.1602.

It is to be noted that the values of g_j^* decreases until the centre of distribution and then increases. Thus, the central tendency of observations $y_{(j)}$ automatically receive small weights. As a result, $\hat{\theta}_3$ provides robustness to data anomalies in the central location of the sample which known as robustness in terms of inliers. This beneficial characteristic of $\hat{\theta}_3$ make the proposed estimator $\bar{y}_{p3}$ robust in term of inliers.

The MSE of the proposed estimator (24) can be derived in similar manner as in Section (3.1),

$$MSE(\bar{y}_p) \cong \operatorname{var}(\hat{\theta}_3) + R^2\operatorname{var}(\bar{x}) - 2R\operatorname{cov}(\hat{\theta}_3, \bar{x}), \tag{25}$$

where the exact value of $\operatorname{var}(\hat{\theta}_3)$ and $\operatorname{cov}(\hat{\theta}_3, \bar{x})$ may be obtained as,

$$\operatorname{var}(\hat{\theta}_3) = \operatorname{var}(\sum_{j=1}^{n} g_j^* y_{(j)}) \Rightarrow \mathbf{g}'\varOmega\,\mathbf{g}\,\sigma_y^2,$$

alternatively the $\operatorname{var}(\hat{\theta}_3)$ can be derived as,

$$\operatorname{var}(\hat{\theta}_3) = \operatorname{var}(\hat{\theta}_3) + [\Gamma(1+1/p)]^2\operatorname{var}(\hat{\sigma}_{y1}^*) + 2\Gamma(1+1/p)\operatorname{cov}(\hat{\mu}_{y1}^*, \hat{\sigma}_{y1}^*). \tag{26}$$

$$\operatorname{cov}(\hat{\theta}_3,\, \bar{x}) = \operatorname{cov}\left(\sum_{j=1}^{n} g_j^* y_{(j)},\, \sum_{j=1}^{n} x_j/n\right) \Rightarrow \rho\sigma_x\sigma_y\mathbf{g}'\,\mathbf{\Omega}\boldsymbol{\eta} \tag{27}$$

where $\mathbf{g}$ is the $n \times 1$ vector with elements g_j^* $(1 < j < n)$, $\mathbf{g}'$ is the transpose vector of $\mathbf{g}$ and $\boldsymbol{\eta}$ is the $n \times 1$ vector with element $1/n$.

Finally, the expression of $\operatorname{MSE}(\bar{y}_{p3})$ is obtained as,

$$\operatorname{MSE}(\bar{y}_{p3}) \cong \mathbf{g}'\varOmega\,\mathbf{g}\,\sigma_y^2 + R^2\operatorname{var}(\bar{x}) - 2R\rho\sigma_x\sigma_y\mathbf{g}'\,\mathbf{\Omega}\boldsymbol{\eta}. \tag{28}$$

3.2. Efficiency comparisons

In this section, we derive the conditions under which our proposed estimators perform better than the traditional estimators.

3.2.1. For proposed estimators following case-I with the traditional estimators

Suppose that underlying super-population is from (4). The MMLE and the GLS estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ in respectively, are calculated from the order statistics $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ for the random sample of size n. In order to establish that both estimators are more efficient than $\bar{y}$, we consider the inequality,

$$E(\bar{y} - \bar{Y})^2 > E(\hat{\theta}_k - \bar{Y})^2, \tag{29}$$

$$\text{where } E(\bar{y} - \bar{Y})^2 = \frac{\vartheta^2}{n}(1-f) \text{ and } E(\hat{\theta}_k - \bar{Y})^2 = \operatorname{var}(\hat{\theta}_k), \tag{30}$$

Substituting (29) in (30), we may have,

$$\frac{\vartheta^2}{n}(1-f) > \operatorname{var}(\hat{\theta}_k). \tag{31}$$

To show that $\hat{\theta}_k$ is more efficient than the $\bar{y}$, we simply have to show that (31). Since the exact expression of $\operatorname{var}(\hat{\theta}_k)$ for $k = 1$ and 2 given in (18) are calculated for shape parameter

Table 1. The (1) var($\bar{y}$), (2) var($\hat{\theta}_1$) and (3) var($\hat{\theta}_2$).

	$n = 5$			$n = 10$			$n = 20$		
p	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
1.5	0.0744	0.0448	0.0408	0.0368	0.0176	0.0157	0.0180	0.0068	0.0061
2	0.0424	0.0356	0.0341	0.0210	0.0162	0.0154	0.0103	0.0074	0.0070
3	0.0285	0.0272	0.0266	0.0141	0.0131	0.0128	0.0069	0.0063	0.0061

$p = 1.5$, 2 and 3 and sample size $n = 5$, 10 and 20. The results are calculated for $N = 500$ and given in Table 1. Table 1 shows that for skewed population (4), the $\hat{\theta}_k$ is more efficient than $\bar{y}$, furthermore, the GLSE estimator $\hat{\theta}_2$ is more efficient than the MMLE estimator $\hat{\theta}_1$.

In order to establish that $\bar{y}_{pk}$ is more efficient than $\bar{y}_1$ for a large sample, we may consider

$$MSE(\bar{y}_1) = E(\bar{y}_1^2) - \bar{Y}^2 \text{ and } MSE(\bar{y}_{pk}) = E(\bar{y}_{pk}^2) - \bar{Y}^2. \tag{32}$$

Considering the two random variables x and y, we may write $E(r(x)h(y)) = E(r(x)$ $E(h(y)|x)$, we may obtain the expected values as,

$$E(\bar{y}_1^2) = E((\bar{X}/\bar{x})^2\bar{y}^2) = \bar{X}E((1/\bar{x})^2E(\bar{y}^2|x)), \tag{33}$$

and

$$E(\bar{y}_{pk}^2) = E((\bar{X}/\bar{x})^2(\hat{\theta}_k)^2) = \bar{X}E((1/\bar{x})^2E((\hat{\theta}_k)^2|x)). \tag{34}$$

Since $E(\bar{y}^2) > E(\hat{\theta}_k)^2$ from (32), we have $E(\bar{y}^2|x) > E((\hat{\theta}_k)^2|x)$, which follows the fact that, on a probability space (Π, ξ, P),$P(E(y) \leq E(x)) = 1$ implies $P(E(y|\mathrm{H}) \leq E(x|\mathrm{H})) = 1$ for any $\mathrm{H} \subset \xi$ [21]. Consequently, $(1/\bar{x}^2)E(\bar{y}^2|x) > (1/\bar{x}^2)E((\hat{\theta}_k)^2|x)$ and $\bar{X}^2E((1/\bar{x}^2)E(\bar{y}^2|x))$ $> \bar{X}^2E((1/\bar{x}^2)E((\hat{\theta}_k)^2|x))$. Therefore, it is clear from (32) to (34) that $\mathrm{MSE}(\bar{y}_1) > \mathrm{MSE}(\bar{y}_{pk})$ for large sample sizes.

The suggested estimator $\bar{y}_{pk}$ performs better than $\bar{y}_1$ in a small sample if

$$\mathrm{MSE}(\bar{y}_{pk}) < \mathrm{MSE}(\bar{y}_1)$$

or

$$\mathrm{cov}(\bar{y}, \bar{x}) < C_{1k}, \text{ where } C_{1k} = \frac{1}{2R}[\mathrm{var}(\bar{y}) - \mathrm{var}(\hat{\theta}_k)] + \mathrm{cov}(\hat{\theta}_k, \bar{x}). \tag{35}$$

The proposed estimator $\bar{y}_{p2}$ performs better than the proposed estimator $\bar{y}_{p1}$ in a small sample if

$$\mathrm{MSE}(\bar{y}_{p2}) < \mathrm{MSE}(\bar{y}_{p1}),$$

after simplification, we have

$$\mathrm{cov}(\hat{\theta}_1,\, \bar{x}) < C_3, \text{ where } C_3 = \frac{1}{2R}[\mathrm{var}(\hat{\mu}_y) - \mathrm{var}(\hat{\mu}_y^*)] + \mathrm{cov}(\hat{\theta}_2, \bar{x}). \tag{36}$$

3.2.2. For proposed estimators following case-II with the traditional estimator

The conditions under which the proposed estimator is more efficient than the traditional ratio estimator can be derived as follow:

$$\mathrm{MSE}(\bar{y}_{p3}) < \mathrm{MSE}(\bar{y}_1),$$

after simplification, we have

$$\mathrm{cov}(\bar{y},\bar{x}) < \frac{1}{2R}[E(\bar{y}-\bar{Y})^2 - E(\hat{\theta}_3-\bar{Y})^2] + (1-f)\mathrm{cov}(\hat{\theta}_3,\bar{x}), \tag{37}$$

therefore, if $\mathrm{cov}(\bar{y},\ \bar{x}) < C_4$, where the bound C_4 is given as

$$C_4 = \frac{1}{2R}[E(\bar{y}-\bar{Y})^2 - E(\hat{\theta}_3-\bar{Y})^2] + (1-f)\mathrm{cov}(\hat{\theta}_3,\bar{x}).$$

Thus, (24) has smaller MSE with respect to the traditional ratio estimator if (37) is satisfied.

4. Numerical illustration

4.1. Simulation study for proposed estimators following case-I

To evaluate the performance of proposed estimator, we conduct a simulation study in this section.

Consider the model (1) in which e_j and x_j are generated independently, and compute y_j. Let e_j be a random observation of error of the skewed population (4) with mean $E(e_j)$and variance σ_e^2, $1 < j < N$. Let Π_N represents the related population consisting of $(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$. To assure that the correlation coefficient ρ is sufficiently high, the values of the parameter β in the model (1) are chosen such that the correlation coefficient is 0.65. The value of β which satisfies this condition is determined by $\beta^2 = \sigma_e^2\rho^2/(1-\rho^2)\mathrm{var}(x)$. To calculate the *MSE* of the estimators in (2) and (14), one has to calculate the $\bar{y}_1$,$\bar{y}_{p1}$ and $\bar{y}_{p2}$ from all possible samples say ${}_NC_n$ of size n each from Π_N. For illustration we take $N = 500$ and n =5, 10 and 30. Since ${}_{500}C_n$ is extremely large, we conduct a Monte Carlo study as follows.

We consider x_j from $U(0,1)$ and generate a population of Π_{500} pairs of (x_j,y_j) from a super-population. From the generated population Π_{500}, we choose $M = 50000$ possible SRS of size n which then gives 50,000 values of each estimator, i.e.$\bar{y}_1$, $\bar{y}_{p1}$ and$\bar{y}_{p2}$. In calculating the $\bar{y}_{p1}$ and $\bar{y}_{p2}$ using $n = 5, 10$ and 30, we calculate the coefficients γ_j and ϕ_j, respectively, using (7) and use them in (14). To compare the efficiencies, we calculate the values of the MSEs of each estimator following the expressions $\mathrm{MSE}(\bar{y}_1) = \sum_{i=1}^{M}(\bar{y}_1-\bar{Y})^2/M$, $\mathrm{MSE}(\bar{y}_{p1}) = \sum_{i=1}^{M}(\bar{y}_{p1}-\bar{Y})^2/M$ and $\mathrm{MSE}(\bar{y}_{p2}) = \sum_{i=1}^{M}(\bar{y}_{p2}-\bar{Y})^2/M$, respectively, under the following populations,

i) All of Nobservations are from Weibull distribution with $p = 2$, $\mu_y = 0$, $\sigma_y = 1$ and no outlier is present.

ii) The $N - N_o$observations from Weibull distribution with $p = 2$, $\mu_y = 0$, $\sigma_y = 1$ and N_o(we do not know which) from Lognormal distribution with $\mu_y = 0$ and $\sigma_y = d$, where $N_o$is calculated from the formula $\left[\left|\frac{N+5}{10}\right|\right]$.

iii) A proportion $1 - \varphi$ of observations are from Weibull distribution with $p = 2$, $\mu_y = 0$, $\sigma_y = 1$ and a proportion φ of the observations are from Lognormal distribution with $\mu_y = 0$ and $\sigma_y = d$.

iv) The first $N(1 - \varphi)$ observations are from Weibull distribution with $p = 2$, $\mu_y = 0$, $\sigma_y = 1$ and last $N(\varphi)$ observations are from Lognormal distribution with $\mu_y = 0$ and $\sigma_y = d$.

v) (a) All of Nobservations are from Weibull distribution with $p = 1.8$, $\mu_y = 0$, $\sigma_y = 1$

vi) (b) All of Nobservations are from Lognormal distribution with $\mu_y = 0$ and $\sigma_y = d$.

Realize that population (i) is the true super-population model, the (ii)–(iv) are contaminated populations and (v) is misspecified population. Where $\varphi(= 0.05, 0.10)$denotes the percentage of contamination and *d (=1.6, 1.7, 1.8)* refers to the extremity of the contamination.

The simulated values of the MSEs and their corresponding relative efficiencies $E_{1,h}$ are given in

Tables 2 and 3, respectively, where $E_{1,h} = {}^{\text{MSE}(\bar{y}_1)}/_{\text{MSE}(\bar{y}_h)}$ for $h = 1, p1$ and $p2$.

From Tables 2 and 3, we have noticed that the values of $\text{MSE}(\bar{y}_{p1})$ and $\text{MSE}(\bar{y}_{p2})$ are always less than the value of $\text{MSE}(\bar{y}_1)$, consequently the values of E_{1,p_1} and E_{1,p_2} are always greater than the value of $E_{1,1}$ under all populations considered in this study. So from this simulation study, we have established that proposed estimators are more efficient than the traditional estimators.

Table 2. The (1) $\text{MSE}(\bar{y}_1)$, (2) $\text{MSE}(\bar{y}_{p1})$ and (3) $\text{MSE}(\bar{y}_{p2})$

	$n = 5$			$n = 10$			$n = 30$		
d	(1)	(2)	(3)	(1)	(2)	(3)	(1)	(2)	(3)
(i)									
	0.2439	0.2139	0.2185	0.1133	0.0962	0.1022	0.0354	0.0299	0.0338
(ii)									
1.6	15.2703	6.06260	4.9105	7.7247	2.4326	2.4313	2.9761	1.1038	1.4150
1.7	24.8454	9.7340	7.7619	15.4807	4.6141	4.3683	4.4429	1.4862	1.8291
1.8	64.9346	24.8788	19.3535	30.4348	8.7855	8.0848	10.0848	2.6501	3.0447
(iii) Using $\varphi = 0.05$									
1.6	12.0657	4.7442	3.8286	4.6086	1.5783	1.5445	1.2426	0.5499	0.7850
1.7	18.5150	7.2078	5.7287	6.6973	2.0468	2.0337	3.1827	0.9483	1.2022
1.8	32.1257	12.3391	9.6721	15.5631	4.4715	4.1694	4.3307	1.1749	1.4562
(iii) Using $\varphi = 0.10$									
1.6	16.9632	6.7099	5.4159	8.4196	2.6346	2.6119	2.3988	0.9899	1.3030
1.7	27.2027	10.6184	8.4298	16.2874	4.8278	4.5677	5.0690	1.5681	1.9145
1.8	56.4135	21.6637	16.9007	27.9933	8.0889	7.4479	8.2935	2.3319	2.7124
(iv) Using $\varphi = 0.05$									
1.6	9.0761	3.6211	2.9850	6.8117	2.0523	2.0180	1.3764	0.5795	0.8114
1.7	12.3881	4.8971	3.9643	7.5782	2.2850	2.2389	2.4050	0.7947	1.0512
1.8	28.4946	10.9623	8.6023	23.1603	6.5464	5.9960	5.9334	1.4836	1.7620
(iv) Using $\varphi = 0.10$									
3	13.2386	5.2975	4.8351	7.9812	2.5155	2.5069	3.0617	1.1459	1.4477
4	25.5335	9.9815	7.9343	12.1663	3.7315	3.6152	6.1419	1.7307	2.0828
5	57.8472	22.2168	17.3702	30.4022	8.6753	7.9612	9.0961	2.4726	2.8618
(v)									
A	0.2624	0.2512	0.2502	0.1117	0.1051	0.1060	0.0330	0.0304	0.0313
B	1.6961	0.9497	0.7962	0.6510	0.2854	0.2280	0.1896	0.0595	0.0502

Table 3. Relative efficiencies of the estimators with respect to $\bar{y}_1$.

	$n = 5$			$n = 10$			$n = 30$		
d	$E_{1,1}$	$E_{1,p1}$	$E_{1,p2}$	$E_{1,1}$	$E_{1,p1}$	$E_{1,p2}$	$E_{1,1}$	$E_{1,p1}$	$E_{1,p2}$
(i)									
	1.0000	1.1403	1.1162	1.0000	1.1778	1.1086	1.0000	1.1839	1.0473
(ii)									
1.6	1.0000	2.5188	3.1097	1.0000	3.1755	3.1772	1.0000	2.6962	2.1033
1.7	1.0000	2.5524	3.2009	1.0000	3.3551	3.5439	1.0000	2.9894	2.4290
1.8	1.0000	2.6100	3.3552	1.0000	3.4642	3.7644	1.0000	3.8054	3.3122
(iii) Using $\varphi = 0.05$									
1.6	1.0000	2.5433	3.1515	1.0000	2.9200	2.9839	1.0000	2.2597	1.5829
1.7	1.0000	2.5687	3.2320	1.0000	3.2721	3.2932	1.0000	3.3562	2.6474
1.8	1.0000	2.6036	3.3215	1.0000	3.4805	3.7327	1.0000	3.6860	2.9740
(iii) Using $\varphi = 0.10$									
1.6	1.0000	2.5281	3.1321	1.0000	3.1958	3.2236	1.0000	2.4233	1.8410
1.7	1.0000	2.5618	3.2270	1.0000	3.3737	3.5658	1.0000	3.2326	2.6477
1.8	1.0000	2.6041	3.3379	1.0000	3.4607	3.7585	1.0000	3.5565	3.0576
(iv) Using $\varphi = 0.05$									
1.6	1.0000	2.5064	3.0406	1.0000	3.3191	3.3755	1.0000	2.3752	1.6963
1.7	1.0000	2.5297	3.1249	1.0000	3.3165	3.3848	1.0000	3.0263	2.2879
1.8	1.0000	2.5993	3.3124	1.0000	3.5379	3.8626	1.0000	3.9993	3.3674
(iv) Using $\varphi = 0.10$									
3	1.0000	2.4990	2.7380	1.0000	3.1728	3.1837	1.0000	2.6719	2.1149
4	1.0000	2.5581	3.2181	1.0000	3.2604	3.3653	1.0000	3.5488	2.9489
5	1.0000	2.6038	3.3303	1.0000	3.5045	3.8188	1.0000	3.6788	3.1785
(v)									
a	1.0000	1.0446	1.0488	1.0000	1.0628	1.0538	1.0000	1.0855	1.0543
b	1.0000	1.7859	2.1302	1.0000	2.2810	2.8553	1.0000	3.1866	3.7769

Furthermore, as the per cent of contamination increases from $\varphi = 0.05$ to $\varphi = 0.10$ the values of MSEs of the proposed estimators are increasing but relatively less as compared to traditional ratio estimator. Moreover, the value of MSE($\bar{y}_{p2}$) is smaller than the value of MSE($\bar{y}_{p1}$) for $n = 5$ and 10. While for $n = 30$ the proposed robust estimator $\bar{y}_{p1}$ is slightly more efficient than the proposed robust ratio estimator $\bar{y}_{p2}$ (see Table 3). This situation remains same for all population considered in this study. So, the performance of the traditional ratio estimator is seriously affected if underlying population is skewed and contains anomalies. It is therefore, the proposed robust ratio estimators are robust to plausible deviations from the assumed distribution.

4.1.1. Real-life application of proposed estimators for case-I

An example is given to illustrate how the proposed robust-ratio estimators are put to work in real situation. The data by Cochran ([2], p. 34) show the weekly family income (x) and the weekly expenditures on food (y) on 33 low-income families. Since the data intended as a population to illustrate the calculation.

In order to confirm, the data follow the normal distribution, some results are provided in Figure 2 and Table 4. From Figure 2(a), the box plot shows that the distribution of data seems to be right-skewed with an outlier. Furthermore, normal Q–Q plot in Figure 2(b) shows that data are not normally distributed. The value of coefficient of skewness $\sqrt{b_1}$ =1.43 and kurtosis b_2 =5.73 also revealed that the underlying distribution of study variable is positively skewed and leptokurtic. Moreover, we have tested the symmetry versus the alternate hypothesis that the distribution is positively skewed. The results of different tests such as Mira test [22], the Cabilio–Masaro test [23] and

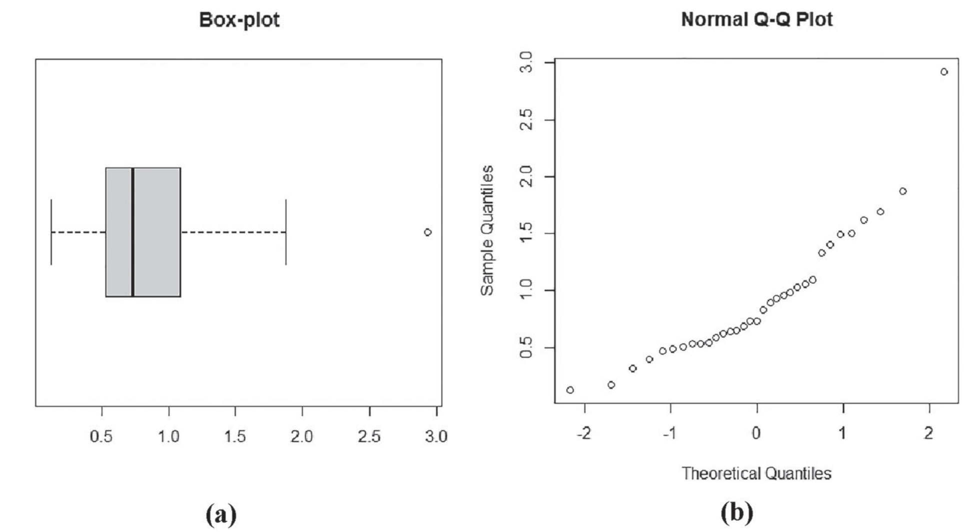


Figure 2. The box plot (a) and the Q–Q norm (b) for real-life example.

Table 4. Computational results of real-life application for case-I.

$N = 33; n = 5$	$\mathrm{cov}(\bar{y}, \bar{x}) = 4.5853$	$\mathrm{cov}(\hat{\theta}_1, \bar{x}) = 0.0012$	$\mathrm{MSE}(\bar{y}_1) = 16.6674$
$R = 0.3789$	$C_{11} = 22.9763$	$\mathrm{cov}(\hat{\theta}_2, \bar{x}) = 0.0009$	$\mathrm{MSE}(\bar{y}_{p1}) = 2.7564$
$\mathrm{var}(\bar{y}) = 17.4165$	$C_{12} = 22.9765$	$E_{1,p1} = 6.046$	$\mathrm{MSE}(\bar{y}_{p2}) = 2.7552$
$\mathrm{var}(\bar{x}) = 18.9858$	$C_3 = 0.0013$	$E_{1,p2} = 6.050$	$\mathrm{MSE}(\bar{y}_1) = 16.6674$

Miao–Gel–Gastwirth (MGG) test [24] with their p-values are Mira test $= 2.72$ with p-value $= .003$, Cabilio–Masaro test $= 2.46$ with p-value $= .006$ and MGG test $= 2.63$ with p-value $= .004$. Therefore, it is reasonable to conclude that distribution of the study variable is right-skewed.

However, to create Q–Q plot, ones need to calculate the population quantiles that require the values of the location, scale and shape parameters of the three-parameter Weibull distribution. The initial value of location and scale parameters are obtained using Equation (10) as $\mu_y = 11.196$ and $\sigma_y = 17.71$. Substituting these values of location and scale parameter the population quantiles, $t_{(i)}$, are determined from (7) for $1 \leq i \leq n$, where n is the sample size. The Q–Q plot is created by plotting the population quantiles against the ordered v values and presented in Figure 3.

From Figure 3, The Q–Q plot for the shape parameter $p = 1.5$ value provide the most closely approximates a straight line. Therefore, the Weibull distribution for $p = 1.5$ is a plausible model for this data. Due to the intrinsic robustness of the MMLEs and GLSEs, close shape parameter values will yield essentially the same estimates and MSEs. Furthermore, we have tested that the sample belonging from a Weibull pdf with known parameters $p = 1.5$ using the Kolmogorov–Smirnov test with test statistic $= 0.2223$ and p-value $= .077$. Therefore, it is reasonable to conclude that distribution of the study variable is the Weibull distribution with $p = 1.5$.

To evaluate the performance of the proposed robust ratio estimators with its competitor (2), we will report the results for the MSEs given in (3), (22) and (23), and the relative

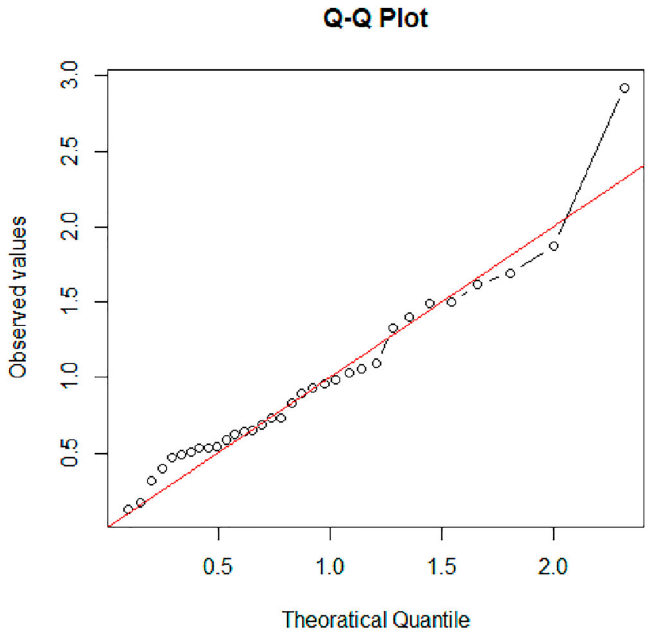


Figure 3. The Q–Q plot for real-life example.

efficiencies by $E_{1,h} = \mathrm{MSE}(\bar{y}_1)/\mathrm{MSE}(\bar{y}_h)$ where$h = 1, p_1, p_2$. The results are reported in Table 4. In Table 4, it may be noted that a larger value of the MSE of the traditional ratio estimator as compared to (22) and (23). It indicates that the traditional ratio estimator is adversely affected if an assumption of normality of the study variable is violated. While, the values of MSEs of the ratio estimators based on MMLE and GLSE methodology are consistent to deal with the issue of normality. Also, the MSE of the proposed robust ratio estimator is less than the MSEs of their competing estimators. So, the proposed robust ratio estimator based on GLSE may be considered as a slightly more efficient estimator (see Table 4) than the proposed robust ratio estimator via MMLE in this study. These results are expected because the conditions (35) and (36) are satisfied (see Table 4).

4.2. Simulation study for proposed estimators for case-II

In this Section, we studied the properties of the proposed ratio estimator $\bar{y}_{p3}$ in presence of the inliers. The inliers are the erroneous observations located to closer to the mean. There are several models to generate the outliers which are discussed in Section 4.1. However, there is less information about how to generate inliers. Tiku et al. [25] are proposed a mechanism for generating the inliers; reciprocate the smallest or largest order statistics in a random sample of size n from a normal population. We follow the same mechanism for generating the inliers for skewed distribution: reciprocate the r largest order statistics from the model (1) so that shifted observations get closer to the middle observations and erroneous. Where $r = \lceil 0.1 \times n \rceil$. To explain the procedure, we have generated a random sample of size n from the model (1) in which e_j and x_j are generated independently from the skewed population (4) with p, $\mu_y = 0$, $\sigma_y = 1$. In the sample size n, we have

Table 5. The simulated results of proposed robust estimator for case-II.

	$\mathrm{MSE}(\bar{y}_1)$	$\mathrm{MSE}(\bar{y}_{p3})$	$E_{1,p3}$
$n = 10$			
$p = 1.5$	0.0118	0.0106	1.1132
$p = 1.8$	0.0111	0.0095	1.1684
$p = 2$	0.0097	0.0082	1.1829
$p = 2.5$	0.0077	0.0062	1.2419
$n = 20$			
$p = 1.5$	0.0051	0.0043	1.1860
$p = 1.8$	0.0037	0.0028	1.3214
$p = 2$	0.0030	0.0027	1.1111
$p = 2.5$	0.0026	0.0019	1.3684

Table 6. Computational result of real-life application.

$N = 33$	$\mathrm{cov}(\bar{y}, \bar{x}) = 4.1467$	$\mathrm{MSE}(\bar{y}_{p3}) = 7.9311$
$n = 5$	$\mathrm{var}(\bar{x}) = 18.9858$	$\mathrm{MSE}(\bar{y}_1) = 15.9550$
$R = 0.3169$	$C_4 = 16.8065$	$E_{1,p3} = 2.0116$
$\mathrm{var}(\bar{y}) = 16.6765$	$\mathrm{cov}(\hat{\theta}_3, \bar{x}) = 1.1476$	

reciprocated the r largest observations to ensure that sample contains inliers. In calculating the $\bar{y}_{p3}$ using $n = 10$ and 20, we calculate the coefficients g_j^* using the (11) and (12), respectively, and use them in (10) and (24). To compare the efficiencies, we calculate the values of the MSEs of each estimator following the expressions $\mathrm{MSE}(\bar{y}_1) = \sum_{i=1}^{M} (\bar{y}_1 - \bar{Y})^2/M$, and $\mathrm{MSE}(\bar{y}_{p3}) = \sum_{i=1}^{M} (\bar{y}_{p3} - \bar{Y})^2/M$, respectively. Based on $M = 50000$ runs we have obtained the results of the calculated MSEs and the relative efficiencies with respect to $\bar{y}_1 (E_{1,p3} = \mathrm{MSE}(\bar{y}_1)/\mathrm{MSE}(\bar{y}_{p3}))$ are presented in Table 5.

In Table 5, it can be seen that the proposed estimator is more efficient than the traditional ratio estimator in presence of inliers. Thus, the middle order statistics which can be potential inliers receive small weights. This depletes the effect of discrepant observations (in the middle) which is instrumental in achieving robustness to inliers.

4.2.1. *Real-life application for case-II*

We extend the real-life example given in Section 4.1.1 for Case-II. To verify the effectiveness of the proposed ratio estimator in presence of the inliers, we generate the inliers by following the mechanism of Tiku et al. [25] with the reason that the expenditures reported by family are often not true due to avoiding the Taxes. Suppose 10% large expenditures are reciprocated and shifted to middle observations. The resulting data of expenditure of the study variable are fabricated thus, producing inliers in y-direction which clearly depict in Figure 4.

We calculate the solution of each Equations (10)–(13) and (28). The results are presented in Table 6. In Table 6, it reveals that proposed robust ratio estimator has less MSE as compared to the traditional ratio estimator. This is expected because the condition (37) is satisfied (see Table 6). The summary result is given in Table 6.

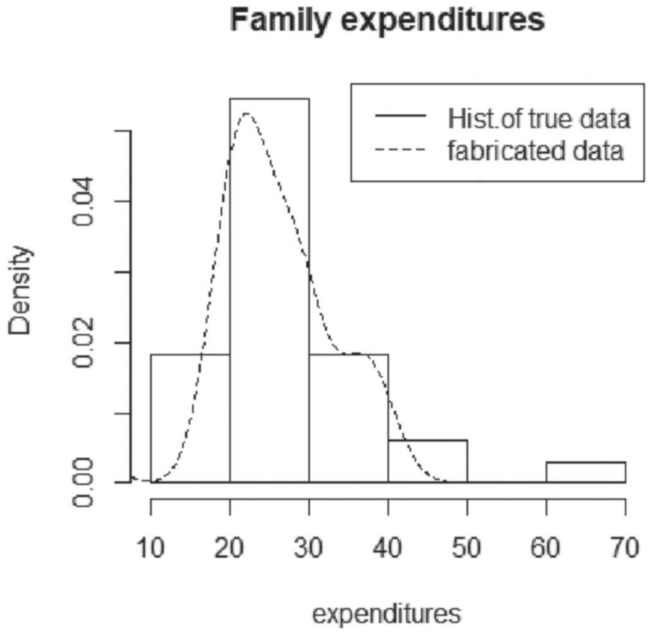


Figure 4. Histogram of true and density curve of fabricated data.

In Table 6, it clearly indicates that due to fabricated information mean square error of existing estimators is effected but proposed robust ratio estimator consistent to deal with issue.

5. Conclusion

The performance of the ratio estimator of population mean becomes poor when in many real-life applications, the assumption of normality is not realistic and the data can be fit by skewed distribution. For example, survey data about family expenditures, family size and household budget, reveal skewed distributions. Thus, in this study, we assume that $g(e_j)$ is characterized by the skewed distribution in Equation (1). Furthermore, the presence of outliers and inliers also adversely affected the performance of the traditional ratio estimator. For this situation, one needs to modify the estimation procedure in order to make the estimator robust to outliers or inliers. We integrate GLSE and MMLE into the traditional ratio estimator and proposed the robust ratio estimators, derive their MSEs, and provide the conditions for which they perform better than the traditional ratio estimator. We study their properties and robustness under both misspecification and contamination models. Furthermore, we construct the ratio estimator which is robust with respect to inliers and derive its properties. Through simulation study and real-life example, it has been shown that the proposed robust estimators are consistent to deal with the issue of non-normality and presence of outliers or inliers.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

[1] Cochran WG. The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. J Agri Sci. 1940;30:262–275.

[2] Cochran WG. Sampling techniques. New York: John Wiley and Sons; 1977.

[3] Farrella JP, Barrera MS. A comparison of several robust estimators for a finite population mean. J Stat Stud. 2006;26:29–43.

[4] Kadilar C, Candan M, Cingi H. Ratio estimators using robust regression. Hacettepe J Math Stat. 2007;36(2):181–188.

[5] Ali N, Ahmad I, Hanif M, et al. Robust-regression-type estimators for improving mean estimation of sensitive variables by using auxiliary information. Commun Stat Theory Method. 2019: 1–14.

[6] Zaman T, Bulut H. Modified ratio estimators using robust regression methods. Commun Stat - Theory Methods. 2019a;48(8):2039–2048.

[7] Zaman T, Bulut H. Modified regression estimators using robust regression methods and covariance matrices in stratified random sampling. Commun Stat Theory Method. 2019: 1–14.

[8] Zaman T. Improvement of modified ratio estimators using robust regression methods. Appl Math Comput. 2019;348:627–631.

[9] Grover LK, Kaur A. An improved regression type estimator of population mean with two auxiliary variables and its variant using robust regression method. J Comput Appl Math. 2021;382:113072.

[10] Oral E, Kadilar C. Robust ratio-type estimators in simple random sampling. J Korean Stat Soc. 2011a;40(4):457–467.

[11] Oral E, Kadilar C. Improved ratio estimators via modified maximum likelihood. Pak J Stat. 2011b;27(3):269–282.

[12] Oral E, Oral E. A robust alternative to the ratio estimator under non-normality. Stat Prob Lett. 2011;81(8):930–936.

[13] Tiku ML, Vellaisamy P. Improving efficiency of survey sample procedures through order statistics. J Ind Soc Agri Stat. 1996;49:363–385.

[14] Sanaullah A, Ahmed A, Hanif M. A new robust ratio estimator with reference to non-normal distribution. Commun Stat Theory Method. 2019: 1–18. doi:10.1080/03610926.2019.1646766.

[15] Tiku ML, Akkaya AD. Robust estimation and hypothesis testing. New Delhi: New Age International (P) Limited, Publishers; 2004; p. 337.

[16] Lloyd EH. Least-squares estimation of location and scale parameters using order statistics. Biometrika. 1952;39(1/2):88–95.

[17] Chan LK, Cheng SW. The best linear unbiased estimates of parameters using order statistics. Soochow J Math. 1982;8:1–13.

[18] David HA, Nagaraja HN. (2004). Order statistics. *Encyclopedia of statistical sciences*, Vol. 9. John Wiley and Sons.

[19] Watterson GA. Linear estimation in censored samples from multivariate normal populations. Ann Math Stat. 1959;30(3):814–824.

[20] Wang K. On concomitants of order statistics [Doctoral dissertation]. The Ohio State University; 2008.

[21] Loeve M. Probability theory I. New York: Springer; 1977.

[22] Mira A. Distribution-free test for symmetry based on Bonferroni's measure. J Appl Stat. 1999;26:959–972.

[23] Cabilio P, Masaro J. A simple test of symmetry about an unknown median. Canad J Stat. 1996;24:349–361.

[24] Miao W, Gel YR, Gastwirth JL. A new test of symmetry about an unknown median. In: A Hsiung, C-H Zhang, Z Ying, editor. Random walk, sequential analysis and related topics – a festschrift in honor of Yuan-Shih Chow. Singapore: World Scientific Publisher; 2006. p. 199–214.

[25] Tiku ML, Islam MQ, SelÃ§uk AS. Nonnormal regression. II. Symmetric distributions. Commun Stat Theory Method. 2001;30(6):1021–1045.