A robust unbiased dual to product estimator for population mean through modified maximum likelihood in simple random sampling

Sanjay Kumar* and Priyanka Chhaparwal

Abstract: In simple random sampling setting, the ratio estimator is more efficient than the mean of a simple random sampling without replacement if \( \rho_{yx} > \frac{1}{2} \frac{C_y}{C_x} \), provided \( R > 0 \), which is usually the case. This shows that if auxiliary information is such that \( \rho_{yx} < \frac{1}{2} \frac{C_y}{C_x} \) then we cannot use the ratio method of estimation to improve the sample mean as an estimator of population mean. So there is need of another type of estimator which also makes use of information on auxiliary variable \( x \). Product method of estimation is an attempt in this direction. Product-type estimators are widely used for estimating population mean when the correlation between study and auxiliary variables is negatively high. This paper is developed to the study of the estimation of the population mean using unbiased dual to product estimator by incorporating robust modified maximum likelihood estimators (MMLEs). Their properties have been obtained theoretically. For the support of the theoretical results, simulations studies under several super-population models have been made. We study the robustness properties of the modified estimators. We show that the utilization of MMLEs in estimating finite population mean results to robust estimates, which is very gainful when we have non-normality or common data anomalies, such as outliers.

ABOUT THE AUTHORS

Sanjay Kumar obtained his Ph.D from Banaras Hindu University, Varanasi, India. He has been working as an Assistant Professor since 2011 in the Department of Statistics, Central University of Rajasthan, Ajmer, Rajasthan, India. His research interests include estimation, optimization problems and robustness study in sampling theory.

Priyanka Chhaparwal is currently working as a research scholar at the Department of Statistics, Central University of Rajasthan, Ajmer, Rajasthan, India. Her research area includes estimating problems in sampling theory.

PUBLIC INTEREST STATEMENT

In sampling theory, for obtaining the estimators of parameters of interest with more precision is an important objective in any statistical estimation procedure in the field of agriculture, medicine and social sciences. For example: estimating quantity of fruits in a village. Supplementary information obtained from auxiliary variable helps in improving the efficiency of the estimators. For example: for estimating quantity of fruits in a village, size of plots can be used as supplementary information which will help in improving the estimators. Several authors have studied such problems under normality case. In this paper, we consider the case where the underlying distribution is not normal, which is a more realistic in real-life situations. We support the theoretical results with simulations under several super-population models and study the robustness property of the modified estimator.
1. Introduction

The use of additional information supplied by auxiliary variables in sample survey has been considered mainly in the area of actuarial, medicine, agriculture and social science at the stage of organization, designing, collection of units and developing the estimation procedure. The use of such auxiliary information in sample surveys has been studied by Cochran (1940), who used it for estimating yields of agricultural crops in agricultural sciences. Product method of estimation is a popular estimation method in sampling theory. In case of negative correlation between study variable and auxiliary variable, Robson (1957) defined a product estimator for the estimation of population mean which was revisited by Murthy (1967). The product estimator performs better than the simple mean per unit estimator under certain conditions. The use of auxiliary information in sample surveys is widely studied in the books written by Yates (1960), Cochran (1977) and Sukhatme, Sukhatme, and Asok (1984). Further, Jhajj, Sharma, and Grover (2006), Bouza (2008, 2015), Swain (2013) and Chanu and Singh (2014) studied the use of auxiliary information under different sampling designs for improving several estimators.

Let $\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ and $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ be the population means of the study variable $y$ and the auxiliary variable $x$, respectively for the population $U = \{U_1, U_2, \ldots, U_N\}$ of size $N$ with coefficient of variations $C_y = \frac{S_y}{\bar{y}}$ and $C_x = \frac{S_x}{\bar{x}}$ and correlation coefficient $\rho_{yx}$. where $S_y$ and $S_x$ are the population mean squares for the study variable ($y$) and the auxiliary variable ($x$). The traditional product estimator for population mean $Y$ proposed by Murthy (1964) is given by

$$\bar{y}_p = \frac{\bar{y}}{\bar{x}} \bar{x}.$$  \hspace{1cm} (1.1)

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ and $n$ is the size of the sample.

The bias and the mean square error (MSE) of the estimator $\bar{y}_p$ is given by

$$B(\bar{y}_p) = \left(1 - \frac{f}{n}\right) Y C_{yx}$$  \hspace{1cm} (1.2)

and

$$MSE(\bar{y}_p) = \left(1 - \frac{f}{n}\right) Y^2 \left(C_y^2 + C_x^2 + 2C_{yx}\right)$$  \hspace{1cm} (1.3)

where $C_y = \frac{S_y}{\bar{y}}$, $C_x = \frac{S_x}{\bar{x}}$, $C_{yx} = \frac{S_{yx}}{\bar{y}\bar{x}}$, $S_y = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y})^2$, $S_x = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$, $S_{yx} = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})$, $f = \frac{n}{N}$ and $S_{yx}$ is the covariance between the study variable and auxiliary variable.

An unbiased estimator $\bar{y}_{pu}$ of the population mean $Y$ after correcting the bias of $\bar{y}_p$ is given by

$$\bar{y}_{pu} \equiv \bar{y}_p - B(\bar{y}_p)$$  \hspace{1cm} (1.4)

To $O(1/n)$, $MSE(\bar{y}_{pu}) \equiv MSE(\bar{y}_p) = \left(1 - \frac{f}{n}\right) Y^2 \left(C_y^2 + C_x^2 + 2C_{yx}\right)$

By making transformation, $z_i = \frac{nx_i - nx}{n - n}$ $(i = 1, 2, \ldots, N)$, Bandopadhyay (1980) proposed a dual to product estimator, which is given by
where the sample mean of $z$ is $\bar{z} = \frac{N}{N-n} \bar{x}$, the population mean of $z$ is $\bar{Z} = \bar{X}$, and $x$ is negatively correlated and $y$ is positively correlated with transformed variable $z$.

Further, $V(z) = (\frac{1}{n} - \frac{1}{N}) \bar{y}^2 S_y^2$ and $\text{Cov}(\bar{y}, z) = -(\frac{1}{n} - \frac{1}{N}) \rho_{yx} S_{yx}$, where $\rho_{yx} = \frac{\bar{x} \bar{y}}{\sqrt{S_x S_y}}$. The bias and the MSE of the estimator $t_1$ are given by

$$B(t_1) = \left( \frac{1 - \frac{1}{n}}{n} \right) \gamma (k + 1) \bar{Y} C_x^2$$

and

$$\text{MSE}(t_1) = \left( \frac{1 - \frac{1}{n}}{n} \right) \bar{y}^2 \left( C_y^2 + \bar{y}^2 C_x^2 + 2 \rho_{yx} C_y C_x \right)$$

where $\rho_{yx} (< 0)$ is the correlation between $y$ and $x$, $k = \frac{C_y}{C_x} = \rho_{yx} \frac{C_y}{C_x}$.

The estimator $t_1$ is preferred to $\bar{y}_p$ when $k > -\frac{1}{2} (1 + \gamma)$, $(1 - \gamma) > 0$, $k$ being negative because $\rho_{yx} < 0$.

Further, by using this transformation and applying the technique of Hartley and Ross (1954), we have an unbiased dual to product estimator (see Singh, 2003) given by

$$t_2 = r_1 \bar{z} + \frac{n(N - 1)}{N(n - 1)} (\bar{y} - r_1 \bar{z}).$$

where $r_1 = \frac{1}{n} \sum_{i=1}^{N} \frac{y_i (N-n)}{N(n-n_i)}$.

The variance of $t_2$ to $O(1/n)$ is given by

$$V(t_2) = E(\bar{y} - \bar{Y})^2 + \bar{r}_1^2 \bar{r}^2 V(\bar{X}) + 2 \bar{r}_1 \gamma \text{Cov}(\bar{y}, \bar{X}) = \left( \frac{1}{n} - \frac{1}{N} \right) \left( S_y^2 + \bar{r}_1^2 \bar{r}^2 S_x^2 + 2 \bar{r}_1 \gamma \bar{r} S_{yx} \right).$$

where $\bar{r}_1 = \frac{1}{n} \sum_{i=1}^{N} \frac{y_i (N-n)}{N(n-n_i)}$.

However, in all of these studies mentioned above, the underlying distribution of $y$ is assumed to be from a normal population. In this paper, we consider the case where the underlying distribution is not normal, which is a more realistic in real-life situations.

Zheng and Al-Saleh (2002) and Islam, Shaibur, and Hossain (2009) have studied the effectivity of modified maximum likelihood estimators (MMLEs) which play a key role in increasing the efficiency of the estimators. Using modified maximum likelihood (MML) methodology (see Tiku, Tan, and Balakrishnan (1986)), we propose a new dual to product type estimator that is based on order statistics. We have shown that the proposed estimator has always smaller MSE with respect to the corresponding unbiased dual to product estimator (1.8), unless the underlying distribution is normal. When the underlying distribution is normal, both the estimators provide exactly the same MSE. We support the theoretical result with simulations under several super population models and study the robustness property of the modified dual to product estimator. We show that utilization of MMLE for estimating finite populations mean results in robust estimate, which is very gainful when we have non-normality or other common data anomalies such as outliers.
2. Long-tailed symmetric family

For the super population linear regression model, \( y_i = \theta x_i + e_i; \ i = 1, 2, \ldots, n \), let the underlying distribution of the study variable \( y \) follows the long-tailed symmetric family.

\[
f(y) = \text{LTS}(p, \sigma) = \frac{1^p}{\sigma \sqrt{K \Gamma(p/2)} \Gamma(p - \frac{1}{2})} \left[ 1 + \frac{1}{K} \left( \frac{y - \mu}{\sigma} \right)^2 \right]^{-\frac{p}{2}}, \quad -\infty < y < \infty,
\]

where \( K = 2p - 3 \) and \( p \geq 2 \) is the shape parameter (\( p \) is known) with \( E(y) = \mu \) and \( \text{Var}(y) = \sigma^2 \). Here it can be obtained that the kurtosis of (2.1) is \( \frac{\mu_4}{\mu_2^2} = \frac{3K}{K^2} \).

The coefficients of kurtosis of the LTS family that we consider in this family are \( \infty, 6, 4.5, 4.0 \) for \( p = 2 \), \( 3 \), \( 4 \), \( 5 \) respectively.

We realize that when \( p = 1 \), (2.1) reduces to a normal distribution. The likelihood function obtained from (2.1) is given by

\[
\log L = -n \log \sigma - p \sum_{i=1}^{n} \log \left[ 1 + \frac{1}{K} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right] z_i - \frac{1}{\sigma^2} \sum_{i=1}^{n} y_i \mu \quad (2.2)
\]

The MLE of \( \mu \) (assuming \( \sigma \) is known) is the solution of the likelihood equation

\[
\frac{d \log L}{d \mu} = 2p \frac{1}{\sigma^2} \sum_{i=1}^{n} g(z_i) = 0, \quad g(z_i) = z_i / \left[ 1 + \frac{1}{K}(z_i^2) \right].
\]

which do not have explicit solutions.

Vaughan (1992a) showed that Equation (2.2) is known to have multiple roots for all \( p < 1 \) but unknown and the number of roots increases as \( n \) increases.

The robust MMLE which is known to be asymptotically equivalent to the MLE is obtained in the following three steps: (see also, Oral 2010)

1. The likelihood equations are expressed in terms of the ordered variates:

\[ y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(n)} \]

2. The function \( g(z_i) \) is linearized by using the first two terms of a Taylor series expansion around \( t_{(i)} = E(z_{(i)}) \), \( z_{(i)} = \frac{y_{(i)} - \mu}{\sigma} \), \( 1 \leq i \leq n \).

3. The resulting equations are solved for the parameters which give a unique solution (MMLE).

The values of \( t_{(i)}; 1 \leq i \leq n \) are given in Tiku and Kumra (1981) for \( p = 2 \) (0.5)10 and Vaughan (1992b) for \( p = 1.5 \) when \( n \leq 20 \). For \( n > 20 \), the approximate values of \( t_{(i)} \) can be used which are obtained from the equations

\[
\frac{1}{\sigma \sqrt{K \Gamma(p/2)} \Gamma(p - \frac{1}{2})} \int_{-\infty}^{t_{(i)}} \left[ 1 + \frac{1}{K} z^2 \right]^{-\frac{p}{2}} dz = \frac{i}{n+1}; 1 \leq i \leq n
\]

We note that \( t = \sqrt{\frac{v}{K}} z \) follows a Student’s \( T \)-distribution with degrees of freedom \( v = 2p - 1 \).

We have now

\[
\frac{d \log L}{d \mu} = 2p \frac{1}{\sigma^2} \sum_{i=1}^{n} g(z_{(i)}) = 0, \quad \text{since} \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_{(i)}
\]

A Taylor series expansion of \( g(z_{(i)}) \) around \( t_{(i)} \) with first two terms of expansion gives

\[
\frac{d \log L}{d \mu} = 2p \frac{1}{\sigma^2} \sum_{i=1}^{n} g(z_{(i)}) = 0, \quad \text{since} \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_{(i)}
\]
\[
g(z_i) = g(t_i) + \left\{z_i - t_i\right\}\left\{\begin{array}{c}
df(g(z)) \\
dz \end{array}\right\}_{z=t_i} = \alpha_i + \beta_i z_i; \quad 1 \leq i \leq n, \quad (2.6)
\]

where
\[
\alpha_i = \left(\frac{2}{K}\right) \frac{t_i^3}{\left(1 + (1/K)t_i^2\right)^{2}} \quad \text{and} \quad \beta_i = \frac{1 - (1/K)t_i^2}{\left(1 + (1/K)t_i^2\right)^2}. \quad (2.7)
\]

Further, for symmetric distributions, it may be noted that \(t_i = -t_{(n-i+1)}\), and hence
\[
\alpha_i = -\alpha_{(n-i+1)}, \quad \sum_{i=1}^{n} \alpha_i = 0 \quad \text{and} \quad \beta_i = \beta_{(n-i+1)} \quad (2.8)
\]

Now, using (2.6) and (2.7) in (2.5), we have the modified likelihood equation which is given by
\[
\frac{d \log L}{d \mu} \approx \frac{2p}{K} \sigma \sum_{i=1}^{n} (\alpha_i + \beta_i z_i) = 0. \quad (2.9)
\]

Hence, the solution of (2.9) is the MMLE \(\hat{\mu}\) is given by
\[
\hat{\mu} = \frac{\sum_{i=1}^{n} \beta_i y_i}{m}, \quad (2.10)
\]

where \(m = \sum_{i=1}^{n} \beta_i\).

Tiku and Vellaisamy (1996) and Oral and Oral (2011) showed that
\[
E(\hat{\mu} - \bar{Y}) = 0, \quad (2.11)
\]

and
\[
E(\hat{\mu} - \bar{Y})^2 = V(\hat{\mu}) - \frac{2n}{N} \text{Cov}(\hat{\mu}, \bar{Y}) + \frac{\sigma^2}{N}. \quad (2.12)
\]

The exact variance of \(\hat{\mu}\) is given by
\[
(\hat{\mu}) = (\beta' \Omega)^{-1} \sigma^2 / m, \quad \text{where} \quad \beta' = (\beta_1, \beta_2, \beta_3, \ldots, \beta_n) \quad \text{and} \quad \Omega \quad \text{is the variance-covariance matrix of the standard variates} \quad z_i = \frac{y_i - \mu}{\sigma}; \quad 1 \leq i \leq n. \quad \text{The term} \quad \text{Cov}(\hat{\mu}, \bar{Y}) \quad \text{in} \quad (2.12) \quad \text{can be evaluated as}
\]
\[
\text{Cov}(\hat{\mu}, \bar{Y}) = (\beta' \Omega \omega)^{-1} \sigma^2 / m, \quad \text{where} \quad \omega \quad \text{is the} \quad 1 \times n \quad \text{row vector with elements} \quad 1/n. \quad \text{The elements of} \quad \Omega \quad \text{are tabulated in} \quad \text{Tiku and Kumra (1981)} \quad \text{and} \quad \text{Vaughan (1992b).}
\]

When \(\sigma\) is not known, the MMLE \(\hat{\sigma}\) can be obtained as given by Tiku and Suresh (1992) and Tiku and Vellaisamy (1996), i.e.
\[
\hat{\sigma} = \frac{F + \sqrt{F^2 + 4nC}}{2 \sqrt{n(n-1)}}, \quad (2.13)
\]

where
\[
F = \frac{2p}{\pi} \sum_{i=1}^{n} a_i y_i, \quad C = \frac{2p}{\pi} \sum_{i=1}^{n} \beta_i (y_i - \hat{\mu})^2.
\]

The methodology of MML is employed in those situations where maximum likelihood (ML) estimation is intractable as widely used by Puthenpura and Sinha (1986), Tiku and Suresh (1992), and Oral (2006, 2010), Oral and Oral (2011) and Oral and Kadilar (2011). Under some regularity conditions, MMLEs have exactly the same asymptotic properties as ML estimators (MLEs) as discussed in Vaughan and Tiku (2000), and for small \(n\) values they are known to be essentially as efficient as MLEs.
3. The proposed dual to product estimator and its variance

In the context of sampling theory, Tiku and Bhasin (1982) and Tiku and Vellaisamy (1996) used the MMLE (2.10) and showed that utilizing the MMLEs lead to improvements in efficiencies in estimating the finite population mean.

Motivated from such approach, we propose a new unbiased dual to product estimator which is given by

\[ T_1 = r_1 \tilde{Y} + \frac{n(N - 1)}{N(n - 1)} (\bar{Y} - r_1 \tilde{Y}), \quad (3.1) \]

assuming the population mean of the auxiliary variable \( \bar{X} \) is known.

The expression for the variance of the proposed estimator \( T_1 \), up to the terms of order \( n^{-1} \) is given as follows:

Let \( \bar{Y} = \bar{Y}(1 + \epsilon_0) \), \( \bar{Z} = \bar{Z}(1 + \epsilon_1) \), such that \( E(\epsilon_0) = 0 = E(\epsilon_1) \).

By using simple random sampling without replacement method of sampling, we have,

\[ E(\epsilon_0) = E(\epsilon_1) = \frac{1}{\pi} E(\tilde{y} - \bar{Y})^2 = \frac{1}{\pi} \left\{ V(\bar{Y}) - \frac{2\pi}{N} \text{Cov}(\bar{Y}, \tilde{Y}) + \frac{\pi^2}{N} \right\}, \]

\[ E(\epsilon_1) = \frac{1}{Z^2} V(z) = \frac{1}{X^2} \left( \frac{n}{N - n} \right)^2 V(x) = \frac{1}{X^2} \left( \frac{n}{N - n} \right)^2 \left( \frac{1}{n} - \frac{1}{N} \right) S_x^2 \]

\[ = \frac{1}{X^2} \left( \frac{n}{N - n} \right)^2 \left( \frac{1}{n} - \frac{1}{N} \right) \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \bar{X})^2 \]

\[ = \frac{1}{X^2} \frac{n}{(N - n)N(N - 1)} \sum_{i=1}^{N} (x_i - \bar{X})^2 \]

\[ E(\epsilon_0, \epsilon_1) = \frac{1}{rX} \text{Cov}(\bar{Y}, \tilde{Z}) = \frac{1}{rX} \text{Cov}(\bar{Y}, \tilde{X}) \]

Now, we have

\[ B(T_1) = 0. \]

\[ V(T_1) = E(\bar{Y} - \bar{Y})^2 + R_1^2 S_x^2 V(x) + 1R_1 Y \text{Cov}(\bar{Y}, \tilde{X}) \quad (3.2) \]

where, the term \( \text{Cov}(\bar{Y}, \tilde{X}) \) is calculated by Oral and Oral (2011) as

\[ \text{Cov}(\tilde{Y}, \tilde{X}) = (1/\pi) \left\{ \text{Cov}(\bar{Y}, \tilde{Y} - \bar{Y}) \right\} = (1/\pi) \left\{ \text{Cov}(\bar{Y}, \tilde{Y} - \bar{Y} + \bar{Y} - \bar{X}) \right\}, \]

\[ = (1/\pi) \left\{ \text{Cov}(\bar{Y}, \tilde{Y} - \bar{Y} + \bar{Y} - \bar{X}) \right\}, \]

\[ \bar{X} = \sum_{i=1}^{n} \beta_i x_i / m, \quad \bar{E}_i = \sum_{i=1}^{n} \beta_i e_i / m, \quad \bar{y}_i = y_i - \alpha x_i \] and \( x_i \) is the concomitant of \( y_i \), when \( (y_i, x_i) \) are ordered with respect to \( y_i; 1 \leq i \leq n \). Here \( x \) is assumed to be non-stochastic for the linear regression model \( y = \alpha x + e \). Hence, \( \text{Cov}(x_i, e_i) \) will not be affected by the ordering of the \( y \) values for \( 1 \leq i \leq n \) and \( 1 \leq j \leq n \); hence

\[ \text{Cov}(\bar{Y}, \tilde{X}) = (1/\pi) \left\{ \text{Cov}(\bar{Y}, \tilde{Y} - \bar{Y} + \bar{Y} - \bar{X}) \right\}, \]

where \( \text{Cov}(e_i, e_i) = (\beta' \Omega \alpha) S_x^2 / \pi^2 \)

In the case of 5% exceeding the sampling fraction \( n/N \), the finite population correction \( (N - n)/N \) can be introduced as

\[ \text{Cov}(\bar{Y}, \tilde{X}) = \{(N - n)/N\} \left\{ \text{Cov}(\bar{Y}, \tilde{Y} - \bar{Y} + \bar{Y} - \bar{X}) \right\} \]
4. Monte-Carlo simulation study

In this study for the simulation, we have used R-programming software in the similar way as Oral and Oral (2011) and Oral and Kadilar (2011) did simulation study through FORTRAN software. We use the following model in the generated super population models

\[ y_i = \theta x_i + e_i, \quad i = 1, 2, \ldots, N, \]

where we generate \( e_i \) and \( x_i \) independently and calculate \( y_i \) for \( i = 1, 2, \ldots, N \). Let the random observations \( e_i, i = 1, 2, \ldots, N \) be from (2.1) with \( E(e) = 0 \) and \( V(e) = \sigma^2_e \). Let the population \( U_n \) consists of \( N \) pairs \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\). To calculate the MSE of the proposed estimator in (3.1), we calculate \( T_1 \) for all possible simple random samples \( \binom{N}{n} \) of size \( n = (5, 11, 15) \) from \( U_n \).

Since \( \binom{N}{n} \) is extremely large, so we proceed as follows.

We consider \( N = 500 \) and from the generated finite population \( U_{500} \) pairs from an assumed super population, we have selected a sample of size \( n = (5, 11, 15) \) by simple random sampling without replacement. Now, we choose at random \( S = 10,000 \) samples for all the possible \( \binom{500}{n} \) samples of size \( n = (5, 11, 15) \), which gives 10,000 values of \( T_1 \). To compare the efficiency of the proposed estimator under different models for a given \( n \), we calculate the values of mean square errors as follows:

\[
\text{MSE}(T_1) = \frac{1}{S} \sum_{j=1}^{S} (T_{1j} - \bar{Y})^2, \quad \text{MSE}(t_1) = \frac{1}{S} \sum_{j=1}^{S} (t_{1j} - \bar{Y})^2,
\]

\[
\text{MSE}(\tilde{y}_{pu}) = \frac{1}{S} \sum_{j=1}^{S} (\tilde{y}_{puj} - \bar{Y})^2.
\]

For setting the population correlation \( \rho_{yx} \) sufficiently high as studied by Oral and Kedilar (2011) we choose the value of parameter \( \theta \) in the model \( y = \theta x + e \), such that the correlation coefficient between study variable \( y \) and auxiliary variable \( x \) is \( \rho_{yx} \). To determine the value of \( \theta \) that satisfies this condition, we follow a similar way given by Rao and Beegle (1967) and write the population correlation between the study variable \( y \) and the auxiliary variable \( x \). For example if \( X \sim U(0, 1) \), the value of \( \theta \) for which the population correlation between \( y \) and \( x \) becomes \( \theta^2 = \frac{12\rho_{yx}^2}{1 - \rho_{yx}^2} \) for the LTS family. Similarly, if \( x \) is generated from \( \text{Exp}(1) \), the value of \( \theta \) for which the population correlation becomes \( \theta^2 = \frac{\rho_{yx}^2}{1 - \rho_{yx}^2} \) for the symmetric family. In the same way, we can have \( x \sim \text{Exp}(0.5) \), \( x \sim \text{N}(0, 1) \), \( x \sim U(-1, 2.5) \) etc.

and the corresponding values of \( \theta \) can be calculated accordingly. Here we take \( \sigma^2 = 1 \), in all situations without loss of generality and calculate the required parameter \( \theta \) for which \( \rho_{yx} = -0.45 \).

5. Comparison of efficiencies of the proposed estimator

The conditions under which the proposed estimator \( T_1 \) is more efficient than the corresponding estimators \( \tilde{y}_{pu}, t_1 \) and \( t_2 \) are given as follows:

\[
\text{MSE}(T_1) \leq \text{MSE}(t_2) \quad \text{if}
\]

\[
\text{Cov}(\bar{y}, \bar{x}) \geq \frac{1}{2R_{1\gamma}} \left\{ E(\bar{y} - \bar{Y})^2 - E(\bar{X} - \bar{Y})^2 \right\} + \text{Cov}(\bar{y}, \bar{X}),
\]

\[
\text{MSE}(T_1) \leq \text{MSE}(t_1) \quad \text{if}
\]

\[
\text{Cov}(\bar{y}, \bar{x}) \geq \frac{1}{2R_{1\gamma}} \left\{ E(\bar{y} - \bar{Y})^2 - E(\bar{X} - \bar{Y})^2 \right\} + \frac{\gamma}{2R} V(\bar{X}) \left\{ R_1^2 - R^2 \right\} + \frac{R_1}{R} \text{Cov}(\bar{y}, \bar{X})
\]

\[
\text{MSE}(T_1) \leq \text{MSE}(\tilde{y}_{pu}) \quad \text{if}
\]
\[ \text{Cov}(\hat{y}, \hat{x}) \geq \frac{1}{2R} \left\{ E(\hat{\mu} - Y)^2 - E(\hat{\bar{y}} - \bar{y})^2 \right\} + \frac{1}{2R} \text{Cov}(\hat{\mu}, \bar{x}) \]  
\[ + \frac{R_{1Y}}{R_{11}} \text{Cov}(\hat{\bar{y}}, \bar{x}), \]  
\[ \text{(5.3)} \]

\[ \text{MSE}(T_1) \leq \text{MSE}(t_2) \leq \text{MSE}(t_3) \text{ if} \]
\[ \frac{1}{2R_{1Y}} \left\{ E(\hat{\mu} - Y)^2 - E(\hat{\bar{y}} - \bar{y})^2 \right\} + \text{Cov}(\hat{\mu}, \bar{x}) \leq \text{Cov}(\hat{\bar{y}}, \bar{x}) \]
\[ \leq \frac{R_{1Y}}{R_{11}} V(\bar{x}) \left\{ R^2 - R_{1Y}^2 \right\} + \frac{R}{R_{1Y}} \text{Cov}(\hat{\bar{y}}, \bar{x}) \]  
\[ \text{(5.4)} \]

\[ \text{MSE}(T_1) \leq \text{MSE}(t_2) \leq \text{MSE}(\hat{y}_{\mu}) \text{ if} \]
\[ \frac{1}{2R_{1Y}} \left\{ E(\hat{\mu} - Y)^2 - E(\hat{\bar{y}} - \bar{y})^2 \right\} + \text{Cov}(\hat{\mu}, \bar{x}) \leq \text{Cov}(\hat{\bar{y}}, \bar{x}) \]
\[ \leq \frac{1}{2R_{1Y}} V(\bar{x}) \left\{ R^2 - R_{1Y}^2 \right\} + \frac{R}{R_{1Y}} \text{Cov}(\hat{\bar{y}}, \bar{x}) \]  
\[ \text{(5.5)} \]

and

\[ \text{MSE}(T_1) \leq \text{MSE}(t_2) \leq \text{MSE}(t_3) \leq \text{MSE}(\hat{y}_{\mu}) \text{ if} \]
\[ \frac{1}{2R_{1Y}} \left\{ E(\hat{\mu} - Y)^2 - E(\hat{\bar{y}} - \bar{y})^2 \right\} + \text{Cov}(\hat{\mu}, \bar{x}) \leq \text{Cov}(\hat{\bar{y}}, \bar{x}) \]
\[ \leq \frac{R_{1Y}}{R_{11}} V(\bar{x}) \left\{ R^2 - R_{1Y}^2 \right\} + \frac{R}{R_{1Y}} \text{Cov}(\hat{\bar{y}}, \bar{x}) \]
\[ \leq \frac{1}{2R_{1Y}} V(\bar{x}) \left\{ R^2 - R_{1Y}^2 \right\} + \frac{R}{R_{1Y}} \text{Cov}(\hat{\bar{y}}, \bar{x}). \]  
\[ \text{(5.6)} \]

where \( \text{Cov}(\hat{\bar{y}}, \bar{x}) = \left( \frac{1}{2} - \frac{1}{R} \right) S_{xy} \).

We assume two different super-population models given below to see how much efficiency we gain with the proposed modified estimator, when the conditions given in Section 5 are satisfied under non-normality:

1. \( x \sim U(0, 1) \) and \( e \sim \text{LTS}(p, 1) \)
2. \( x \sim \exp(1) \) and \( e \sim \text{LTS}(p, 1) \)

For the models (1) and (2), the values of \( \theta \) which makes the population correlation \( \rho_{xy} = -0.45 \) are given below in Table 1.

Here, we note that for the LTS family (2.1), the value of \( \theta \) does not depend on the shape parameter \( p \).

To verify that the super-populations are generated appropriately, we provide a scatter graph and the underlying distribution of model for \( p = 3.5 \) for model (2) in Figure 1 and in Figure 2.

Relative efficiencies are calculated by \( RE = \frac{\text{MSE}(\hat{y}_{\mu})}{\text{MSE}(\bar{y})} \times 100 \),

where MSE (.) and relative efficiency are given in the Table 2 for the model (1) and (2).

**Table 1. Parameter values of \( \theta \) used in models (1)-(2) that give \( \rho_{xy} = -0.45 \)**

| Population | \( p \) | 2.5  | 4.5  | 5.5  |
|------------|---------|------|------|------|
| Model (1)  |         | -1.746 | -1.746 | -1.746 |
| Model (2)  |         | -0.504 | -0.504 | -0.504 |
From Table 2, we see that the proposed estimator $T_1$ is more efficient than the corresponding estimators $y_{pu}, t_3$ and $t_2$ because the theoretical conditions given in Section 5 are satisfied. We also observe that when sample size increases, mean square error decreases.

6. Robustness of the proposed estimator

The outliers in sample data are normally a focused problem for survey statistician. In practice, the shape parameters $p$ in $LTS(p, \sigma)$ might be mis-specified. Therefore, it is very important for estimators to have efficiencies of robustness estimates.

Here, we take $N = 500$ and $\sigma^2 = 1$ without loss of generality and we study the robustness property of proposed estimator under different outlier models as follows.

We assume $x \sim U(0, 1)$ as well as $x \sim \text{Exp}(1)$ and $y \sim LTS(p = 3.5, \sigma^2 = 1)$. We determine our super-population model as follows:

![Figure 1. A scatter graph of the study variable and auxiliary variable obtained from model (2) for $p = 3.5$.](image1)

![Figure 2. Underlying distribution of the study variable obtained from model (2) for $p = 3.5$.](image2)
Table 2. Mean square error and efficiencies of the estimators under super-populations (1–2)

| Est. | \( x \sim U(0, 1) \) and \( e \sim LTS(p, 1) \) | \( x \sim \exp(1) \) and \( e \sim LTS(p, 1) \) |
|------|---------------------------------|---------------------------------|
|      | \( n = 5 \) | \( n = 11 \) | \( n = 15 \) | \( n = 5 \) | \( n = 11 \) | \( n = 15 \) |
| \( p = 2.5 \) | \( T_1 \) | 229.74 (0.2333) | 215.50 (0.0961) | 195.85 (0.0747) | 214.54 (0.2152) | 202.80 (0.0963) | 213.27 (0.0678) |
|      | \( t_2 \) | 197.20 (0.2718) | 187.08 (0.1107) | 163.46 (0.0225) | 191.34 (0.2413) | 171.91 (0.1136) | 170.12 (0.0850) |
|      | \( t_3 \) | 197.13 (0.2719) | 186.91 (0.1108) | 163.28 (0.0896) | 191.18 (0.2415) | 171.62 (0.1138) | 169.62 (0.0851) |
|      | \( \hat{y}_{n0} \) | 100.00 (0.5336) | 100.00 (0.2071) | 100.00 (0.1463) | 100.00 (0.4617) | 100.00 (0.1953) | 100.00 (0.1446) |
| \( p = 4.5 \) | \( T_1 \) | 226.97 (0.2403) | 177.83 (0.1087) | 180.72 (0.0773) | 179.04 (0.2447) | 168.95 (0.1124) | 172.99 (0.0822) |
|      | \( t_2 \) | 224.35 (0.2431) | 174.14 (0.1110) | 178.42 (0.0783) | 172.28 (0.2543) | 161.21 (0.1178) | 164.58 (0.0864) |
|      | \( t_3 \) | 224.26 (0.2432) | 173.99 (0.1111) | 178.19 (0.0784) | 172.01 (0.2547) | 160.93 (0.1180) | 164.39 (0.0865) |
|      | \( \hat{y}_{n0} \) | 100.00 (0.5484) | 100.00 (0.1933) | 100.00 (0.1397) | 100.00 (0.4381) | 100.00 (0.1899) | 100.00 (0.1422) |
| \( p = 5.5 \) | \( T_1 \) | 209.83 (0.2513) | 180.92 (0.1174) | 191.68 (0.0793) | 192.57 (0.2490) | 176.75 (0.1161) | 158.31 (0.0842) |
|      | \( t_2 \) | 208.25 (0.2532) | 178.64 (0.1189) | 190.00 (0.0800) | 188.71 (0.2541) | 171.29 (0.1196) | 152.87 (0.0872) |
|      | \( t_3 \) | 208.17 (0.2533) | 178.49 (0.1190) | 189.76 (0.0801) | 188.48 (0.2544) | 171.00 (0.1200) | 152.69 (0.0873) |
|      | \( \hat{y}_{n0} \) | 100.00 (0.5273) | 100.00 (0.2126) | 100.00 (0.1520) | 100.00 (0.4795) | 100.00 (0.2052) | 100.00 (0.1333) |

(*Mean square errors are in the parenthesis)

(5). True model: \( LTS(p = 3.5, \sigma^2 = 1) \)

(6). Dixon’s outliers model: \( N \sim N_0 \) observations from \( LTS(3.5, 1) \) and \( N_0 \) (we do not know which) form \( LTS(3.5, 2.0) \)

(7). Mis-specified model: \( LTS(4.0, 1) \)

Here, we realize that the model (5), the assumed super population model is given for the purpose of comparison and the models (6) and (7) are taken as its plausible alternatives. Here, we have assumed the super population model \( LTS(3.5, 1) \). The coefficients \( (\alpha_i, \beta_i) \) from (2.7) are calculated with \( p = 3.5 \) and are used in models (5) and (6). \( N_0 \) in the model (6) is calculated from the formula \( \lfloor 0.5 + 0.1 \times N \rfloor = 50 \) for \( N = 500 \). We standardized the generated \( \varepsilon_i \)s, \( (i = 1, 2, \ldots, N) \) in all the models to have the same variance as that of \( LTS(3.5, 1) \) i.e. it should be equal to 1. The simulated values of MSE and the relative efficiency are given in Table 3. Here, theoretical conditions are satisfied for the models.

From Table 3, we see that the proposed estimator \( T_1 \) is more efficient than the corresponding estimators \( \hat{y}_{n0}, t_1 \) and \( t_2 \) because the theoretical conditions are satisfied. We also observe that when sample size increases, mean square error decreases.

7. Determination of the shape parameter

It may be possible that the shape parameter \( p \) is unknown, then in such a case in order to determine whether a particular density is appropriate for the underlying distribution of the study variable \( y \), a Q-Q plot is made by plotting the population quantiles for the density against the ordered values of \( y \).
The population quantiles \( t_i(\cdot) \) are determined from the equation

\[
\int_{-\infty}^{t_i(\cdot)} t(u) \, du = \frac{i}{n+1}; \quad 1 \leq i \leq n,
\]

where \( n \) is the sample size.

The Q-Q plot that closely approximates a straight line would be assumed to be the most appropriate. Using such procedure, we can also obtain a plausible value for the shape parameter simply.

8. Conclusions
In this study, we show that when the underlying distribution of the study variable is not normal (e.g., Logistic distribution, T-distribution, etc.), which is applicable in most of the areas, MML integrated estimators can improve the efficiency of the estimators. In the paper, we show when the underlying distribution of the study variable is a long-tailed symmetric distribution, MML integrated dual to product estimator \( T_1 \) can improve the efficiency of the unbiased dual to product estimator \( T_2 \). The proposed estimator is also more efficient than the product estimators \( T_1 \) and \( y_{pu} \). We also show that the MML integrated dual to product estimator \( T_1 \) is robust to outliers as well as other data anomalies.

Acknowledgements
The authors are grateful to the Editors and referees for their valuable suggestions which led to improvements in the article.

Funding
The authors received no direct funding for this research.

Author details
Sanjay Kumar

---

### Table 3. Mean square errors and efficiencies under super-populations (5)–(7) for LTS family

| Est. | \( n \) = 5 | \( n \) = 11 | \( n \) = 15 | \( n \) = 5 | \( n \) = 11 | \( n \) = 15 |
|------|------------|------------|------------|------------|------------|------------|
|      | True Model: \( x \sim \text{Uni}(0, 1) \) | Dixon outlier Model: \( x \sim \text{Uni}(0, 1) \) | Mis-specified Model: \( x \sim \text{Uni}(0, 1) \) | True Model: \( x \sim \text{Exp}(1) \) | Mis-specified Model: \( x \sim \text{Exp}(1) \) |
| \( T_1 \) | 220.58 (0.2473) | 200.19 (0.1037) | 191.73 (0.0810) | 275.37 (0.9490) | 233.97 (0.3441) | 194.52 (0.2374) |
| \( T_2 \) | 213.25 (0.2558) | 188.04 (0.1104) | 177.48 (0.0875) | 248.53 (1.0515) | 210.15 (0.3831) | 182.39 (0.2532) |
| \( T_3 \) | 213.17 (0.2559) | 187.87 (0.1105) | 177.28 (0.0876) | 248.25 (1.0527) | 209.93 (0.3835) | 182.24 (0.2534) |
| \( y_{pu} \) | 100.00 (0.5455) | 100.00 (0.2076) | 100.00 (0.1553) | 100.00 (2.6133) | 100.00 (0.8051) | 100.00 (0.4618) |

*Mean square errors are in the parenthesis.*
References

Bouza, C. N. (2004). Probability and statistics. Revista Investigación Operacional, 30, 1-66.

Chau, W. W. (2014). Improved class of ratio estimators for finite population. Communications in Statistics: Theory and Methods, 43, 2743-2752.

Chau, W. W., & Singh, B. K. (2014). Improved class of ratio estimators for finite population mean in two phase sampling. Global Journal of Science Frontier Research: Mathematics and Decision Sciences, 14(2-1), 56-64.

Cochran, W. G. (1946). Sampling techniques. New York: Wiley.

Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. Nature, 174, 270-271.

Islam, T., Shaibur, M. R., & Hossain, S. S. (2009). Effectiveness of modified maximum likelihood estimators using selected ranked set sampling data. Austrian Journal of Statistics, 38(2), 109-120. doi:10.17713/ajs.v38i2.265

Jhajj, H. S., Sharma, M. K., & Grover, L. K. (2006). Dual of ratio estimators of finite population mean obtained on using linear transformation to auxiliary variable. Journal of Japan Statistical Society, 36(1), 107-119. doi:10.1480/js.36.107

Murthy, M. N. (1964). Product method of estimation. Sankhya A, 26, 57-74.

Murthy, M. N. (1967). Sampling Theory and Methods. Calcutta: Statistical Publishing Society.

Ora, E. (2004). Binary regression with stochastic covariates. Communications in Statistics: Theory and Methods, 35, 1429-1447. doi:10.1080/0361092060037123

Ora, E. (2010). Improving efficiency of ratio-type estimators through order statistics. In: JSM proceedings, section on survey research methods (pp. 4231-4239). Alexandria, VA: American Statistical Association.

Ora, E., & Ora, E. (2011). A robust alternative to the ratio estimator under non-normality. Statistics and Probability Letters 81:930–936.

Ora, E., & Ora, E. (2013). Robust ratio-type estimators in simple random sampling. Journal of the Korean Statistical Society, 42, 457-467.

Pathenpura, S., & Sinha, N. K. (1966). Modified maximum likelihood method for the robust estimation of system parameters from noisy data. Automatica, 22, 231-235. doi:10.1016/0005-1098(86)90005-3

Rao, J. N. K., & Beegle, L. D. (1967). A Monte Carlo study of some ratio estimators. Sankhya: Series B, 47-56.

Robson, D. (1957). Applications of multivariate polykeys to the theory of unbiased ratio-type estimation. Journal of American Statistical Association, 52(280), 511-522. doi:10.1080/01621459.1957.10501607

Singh, S. (2003). Advanced sampling theory with applications, vol. 1. The Netherlands: Kluwer Academic Publishers.

Sukhatme, P. V., Sukhatme, B. V., & Asok, C. (1986). Sampling theory of surveys with applications. New Delhi: Indian Society Agricultural Statistics.

Swain, A. K. P. (2013). On some modified ratio and product type estimators-revisited. Revista Investigación Operacional, 34(1), 35-57.

Tiku, M. L., & Bhaskar, P. (1982). Usefulness of robust estimators in sample survey. Communications in Statistics: Theory and Methods, 11, 2597-2610. doi:10.1080/0361091820882843

Tiku, M. L., & Kumar, S. (2018). Expected values and variances of order statistics for a family of symmetric distributions (student’s t). Selected Tables in Mathematical Statistics, American Mathematical Society, 8, 141-270.

Tiku, M. L., & Suresh, R. P. (1992). A new method of estimation for location and scale parameters. Journal of Statistical Planning and Inference, 30, 281-292. doi:10.1016/0378-3758(92)90088-A

Tiku, M. L., Tan, M. Y., & Balakrishnan, N. (1986). Robust Inference. New York: Marcel Dekker Inc.

Tiku, M. L., & Vellaisamy, P. (1996). Improving efficiency of survey sample procedures through order statistics. Journal of Indian Society Agricultural Statistics, 49, 363-385.

Vaughan, D. C. (1992a). On the Tiku-Suresh method of estimation. Communications in Statistics-Theory and Methods, 21, 451-469. doi:10.1080/03610929208830788

Vaughan, D. C. (1992b). Expected values, variances and covariances of order statistics for student’s t-distribution with two degrees of freedom. Communications in Statistics-Simulation and Computation, 21, 391-404. doi:10.1080/03610919208831025

Vaughan, D. C., & Tiku, M. L. (2000). Estimation and hypothesis testing for nonnormal bivariate distribution with applications. Journal of Mathematical and Computer Modelling, 32, 53-67. doi:10.1016/S0895-7177(00)00119-9

Yates, F. (1960). Sampling methods in censuses and surveys. London: Charles Griffin and Co.

Zheng, G., & Al-Saleh, M. F. (2002). Modified maximum likelihood estimators based on ranked set samples. Annals of the Institute of Statistical Mathematics, 54 (3), 641-658. doi:10.1023/A:1022475413950
