Kumar & Chhaparwal, *Cogent Mathematics* (2016), 3: 1168070
http://dx.doi.org/10.1080/23311835.2016.1168070

cogent mathematics

STATISTICS | RESEARCH ARTICLE

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

Sanjay Kumar[1]* and Priyanka Chhaparwal[1]

Received: 30 September 2015
Accepted: 04 March 2016
Published: 06 April 2016

*Corresponding author: Sanjay Kumar, Department of Statistics, Central University of Rajasthan, Bandarsindri, Kishangarh, Ajmer, Rajasthan 305817, India
E-mail: sanjay.kumar@curaj.ac.in

Reviewing editor:
Guohua Zou, Chinese Academy of Sciences, China

Additional information is available at the end of the article

Abstract: In simple random sampling setting, the ratio estimator is more efficient than the mean of a simple random sampling without replacement (SRSWOR) if $\rho_{yx} > \frac{1}{2}\frac{C_x}{C_y}$, provided $R > 0$, which is usually the case. This shows that if auxiliary information is such that $\rho_{yx} < -\frac{1}{2}\frac{C_x}{C_y}$, then we cannot use the ratio method of estimation to improve the sample mean as an estimator of population mean. So there is need for 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 of unbiased dual to product estimator by incorporating robust modified maximum likelihood estimators (MMLE's). 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 MMLE's 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.

Subjects: Mathematics & Statistics; Science; Statistical Computing; Statistical Theory & Methods; Statistics; Statistics & Computing; Statistics & Probability

Priyanka Chhaparwal

ABOUT THE AUTHORS

Sanjay Kumar obtained his PhD 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.

cogent oa

© 2016 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

Keywords: product estimator; unbiased dual to product estimator; auxiliary variable; simulation study; modified maximum likelihood; transformed auxiliary variable

Mathematics subject classification: 62D05

1. Introduction

The use of additional information supplied by auxiliary variables in sample survey have 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 $\bar{Y}\left(=\frac{1}{N}\sum_{i=1}^{N}y_i\right)$ and $\bar{X}\left(=\frac{1}{N}\sum_{i=1}^{N}x_i\right)$ 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 ρ_{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 $\bar{Y}$ proposed by Murthy (1964) is given by

$$\bar{y}_p = \frac{\bar{y}}{\bar{X}}\bar{x}, \tag{1.1}$$

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 $\bar{y}_p$ are given by

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

and

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

where $C_y^2 = \frac{s_y^2}{\bar{Y}^2}, C_x^2 = \frac{s_x^2}{\bar{X}^2}, C_{yx} = \frac{S_{yx}}{\bar{Y}\bar{X}}, S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2, S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{X})^2$

$f = \frac{n}{N}$ and S_{yx} is the covariance between the study variable and auxiliary variable.

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

$$\bar{y}_{pu} \cong \bar{y}_p - B(\bar{y}_p) \tag{1.4}$$

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

By making transformation, $z_i = \frac{N\bar{X} - nx_i}{N-n}(i = 1, 2, \ldots, N)$, Bandopadhyay (1980) proposed a dual to product estimator, which is given by

$$t_1 = \frac{\bar{y}}{\bar{z}}\bar{X}, \tag{1.5}$$

where the sample mean of z is $\bar{z} = \frac{N\bar{X}-n\bar{x}}{N-n}$, the population mean of z is $\bar{Z} = \bar{X}$, y and x is negatively correlated and y is positively correlated with transformed variable z.

Further, $V(\bar{z}) = \left(\frac{1}{n} - \frac{1}{N}\right)\gamma^2 S_x^2$ and $Cov(\bar{y}, \bar{z}) = -\left(\frac{1}{n} - \frac{1}{N}\right)\gamma S_{yx}$,

where $\gamma = \frac{n}{N-n}$.

The bias and the MSE of the estimator t_1 are given by

$$B(t_1) = \left(\frac{1-f}{n}\right)\gamma(k+1)\bar{Y}C_x^2 \tag{1.6}$$

and

$$MSE(t_1) = \left(\frac{1-f}{n}\right)\bar{Y}^2\left(C_y^2 + \gamma^2 C_x^2 + 2\gamma p_{yx} C_y C_x\right) \tag{1.7}$$

where $\rho_{yx}(<0)$ is the correlation between y and x, $k = \frac{C_{yx}}{C_x^2} = \rho_{yx}\frac{C_y}{C_x}$.

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, 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

$$t_2 = \bar{r}_1\bar{Z} + \frac{n(N-1)}{N(n-1)}(\bar{y} - \bar{r}_1\bar{z}), \tag{1.8}$$

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

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

$$V(t_2) = E(\bar{y} - \bar{Y})^2 + \bar{R}_1^2\gamma^2 V(\bar{x}) + 2\bar{R}_1\gamma Cov(\bar{y}, \bar{x}) = \left(\frac{1}{n} - \frac{1}{N}\right)(S_y^2 + \bar{R}_1^2\gamma^2 S_x^2 + 2\bar{R}_1\gamma S_{yx}), \tag{1.9}$$

where $\bar{R}_1 = \frac{1}{N}\sum_{i=1}^{N}\frac{y_i(N-n)}{N\bar{X}-nx_i}$

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 (MMLE's) which plays a key role in increasing the efficiency of the estimators. Using modified maximum likelihood (MML) methodology (see Tiku, Tan, & 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 mean square error (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 mean square error. 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 to 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 follow the long-tailed symmetric family.

$$f(y) = LTS(p,\sigma) = \frac{\Gamma p}{\sigma\sqrt{K}\Gamma\left(\frac{1}{2}\right)\Gamma\left(p-\frac{1}{2}\right)}\left\{1+\frac{1}{K}\left(\frac{y-\mu}{\sigma}\right)^2\right\}^{-p}; -\infty < y < \infty, \tag{2.1}$$

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

The coefficients of kurtosis of the LTS family that we consider in this family are ∞, 6, 4.5, 4.0 for $p = 2.5, 3.5, 4.5, 5.5$, respectively.

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

$$LogL \propto -nlog\sigma - p\sum_{i=1}^{n}\log\left\{1+\frac{1}{K}z_i^2\right\}; z_i = \frac{y_i-\mu}{\sigma}. \tag{2.2}$$

The MLE of μ (assuming σ is known) is the solution of the likelihood equation

$$\frac{dLogL}{d\mu} = \frac{2p}{K\sigma}\sum_{i=1}^{n} g(z_i) = 0, g(z_i) = z_i/\left\{1+\frac{1}{K}(z_i^2)\right\}, \tag{2.3}$$

which does not have explicit solutions.

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

The robust MMLE which is known to be asymptotically equivalent to the MLE are obtained in following three steps:

(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 $g(z_i)$ are linearized using the first two terms of a Taylor series expansion around $t_{(i)} = E\left(z_{(i)}\right), z_{(i)} = \frac{y_{(i)}-\mu}{\sigma}; 1 \leq i \leq n$,

(3) The resulting equations are solved for the parameters which gives 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

$$\frac{\Gamma p}{\sigma\sqrt{K}\Gamma\left(\frac{1}{2}\right)\Gamma\left(p-\frac{1}{2}\right)}\int_{-\infty}^{t_{(i)}}\{1+\frac{1}{K}z^2\}^{-p}dz = \frac{i}{n+1}; 1 \leq i \leq n. \tag{2.4}$$

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

$$\frac{dLogL}{d\mu} = \frac{2p}{K\sigma}\sum_{i=1}^{n} g(z_{(i)}) = 0, since \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} y_{(i)} \tag{2.5}$$

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

$$g(z_{(i)}) \cong g\left(t_{(i)}\right) + \left\{z_{(i)} - t_{(i)}\right\}\left\{\frac{d\{g(z)\}}{dz}|_{z=t_{(i)}}\right\} = \alpha_i + \beta_i z_{(i)};\ 1 \le i \le n, \tag{2.6}$$

$$\text{where } \alpha_i = \left(\frac{2}{K}\right)\frac{t_{(i)}^3}{\{1+(1/K)t_{(i)}^2\}^2} \text{ and } \beta_i = \frac{1-(1/K)t_{(i)}^2}{\{1+(1/K)t_{(i)}^2\}^2}. \tag{2.7}$$

Further, for symmetric distributions, it may be noted that $t_{(i)} = -t_{(n-i+1)}$ and hence

$$\alpha_i = -\alpha_{(n-i+1)}, \sum_{i=1}^{n}\alpha_i = 0 \text{ and } \beta_i = \beta_{(n-i+1)}. \tag{2.8}$$

Now, using (2.6) and (2.7) in (2.5), we have the modified likelihood equation which is given by

$$\frac{dLogL}{d\mu} \cong \frac{dLogL^*}{d\mu} = \frac{2p}{K\sigma}\sum_{i=1}^{n}(\alpha_i + \beta_i z_{(i)}) = 0. \tag{2.9}$$

Hence, the solution of (2.9) is the MMLE $\hat{\mu}$ is given by

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\beta_i y_{(i)}}{m} \tag{2.10}$$

where $m = \sum_{i=1}^{n}\beta_i$

Tiku and Vellaisamy (1996) showed that

$$E(\hat{\mu} - \overline{Y}) = 0 \tag{2.11}$$

and

$$E(\hat{\mu} - \overline{Y})^2 = V(\hat{\mu}) - \frac{2n}{N}Cov(\hat{\mu}, \bar{y}) + \frac{\sigma^2}{N}. \tag{2.12}$$

The exact variance of $\hat{\mu}$ is given by $V(\hat{\mu}) = (\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\beta})\sigma^2/m^2$, where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \beta_3, \ldots., \beta_n)$ and $\boldsymbol{\Omega}$ is the variance–covariance matrix of the standard variates $z_{(i)} = \frac{y_{(i)}-\mu}{\sigma}$; $1 \le i \le n$. The term $Cov(\hat{\mu}, \bar{y})$ in (2.12) can be evaluated as

$Cov(\hat{\mu}, \bar{y}) = (\boldsymbol{\beta}'\boldsymbol{\Omega}\boldsymbol{\omega})\sigma^2/m$, where $\boldsymbol{\omega}'$ is the $1 \times n$ row vector with elements $1/n$. The elements of $\boldsymbol{\Omega}$ are tabulated in Tiku and Kumra (1981) and Vaughan (1992b).

When σ is not known, the MMLE $\hat{\sigma}$ can be obtained as given by Tiku and Suresh (1992) and Tiku and Vellaisamy (1996), i.e.

$$\hat{\sigma} = \frac{F + \sqrt{F^2 + 4nC}}{2\sqrt{n(n-1)}}, \tag{2.13}$$

where $F = \frac{2p}{K}\sum_{i=1}^{n}\alpha_i y_{(i)}$ and $C = \frac{2p}{K}\sum_{i=1}^{n}\beta_i(y_{(i)} - \hat{\mu})^2$

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). 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 Bhasln (1982) and Tiku and Vellaisamy (1996) used the MMLE (2.10) and showed that utilizing the MMLEs leads 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

$$T_1 = \bar{r}_1\overline{Z} + \frac{n(N-1)}{N(n-1)}\left(\hat{\mu} - \bar{r}_1\overline{Z}\right) \tag{3.1}$$

assuming the population mean of the auxiliary variable $\overline{X}$ 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 $\hat{\mu} = \overline{Y}(1+\in_0), \bar{z} = \overline{Z}(1+\in_1)$ such that $E(\in_0) = 0 = E(\in_1)$

Using simple random sampling without replacement method of sampling, we have,

$$E(\in_0^2) = \frac{1}{\overline{Y}^2}E(\hat{\mu}-\overline{Y})^2 = \frac{1}{\overline{Y}^2}\left\{V(\hat{\mu}) - \frac{2n}{N}Cov(\hat{\mu},\bar{y}) + \frac{\sigma^2}{N}\right\},$$

$$E\left(\in_1^2\right) = \tfrac{1}{\overline{Z}^2}V\left(\overline{Z}\right) = \tfrac{1}{\overline{X}^2}\left(\tfrac{n}{N-n}\right)^2 V(\bar{x}) = \tfrac{1}{\overline{X}^2}\left(\tfrac{n}{N-n}\right)^2\left(\tfrac{1}{n}-\tfrac{1}{N}\right)S_x^2$$

$$= \tfrac{1}{\overline{X}^2}\left(\tfrac{n}{N-n}\right)^2\left(\tfrac{1}{n}-\tfrac{1}{N}\right)\tfrac{1}{N-1}\textstyle\sum_{i=1}^{N}\left(x_i-\overline{X}\right)^2$$

$$= \tfrac{1}{\overline{X}^2}\tfrac{n}{(N-n)N(N-1)}\textstyle\sum_{i=1}^{N}\left(x_i-\overline{X}\right)^2$$

$$E(\in_0,\in_1) = \frac{1}{\overline{Y}\overline{X}}Cov(\hat{\mu},\overline{Z}) = -\frac{1}{\overline{Y},\overline{X}}\gamma Cov(\hat{\mu},\bar{x})$$

Now, we have

$$B\left(T_1\right) = 0,$$

$$V(T_1) = E(\hat{\mu}-\overline{Y})^2 + \bar{R}_1^2\gamma^2 V(\bar{x}) + 2\bar{R}_1\gamma Cov(\hat{\mu},\bar{x}) \tag{3.2}$$

where $Cov(\hat{\mu},\bar{x}) = \left(\frac{1}{\theta}\right)\left\{Cov(\hat{\mu},\bar{y}-\bar{e})\right\} = (1/\theta)\left\{Cov(\hat{\mu},\bar{y}) - Cov(\theta\bar{x}_{[\cdot]}+\bar{e}_{[\cdot]},\bar{e})\right\}$

$\bar{x}_{[\cdot]} = \sum_{i=1}^{n}\beta_i\bar{x}_{[\cdot]}/m, \bar{e}_{[\cdot]} = \sum_{i=1}^{n}\beta_i\bar{e}_{[\cdot]}/m, \bar{e}_{[\cdot]} = y_{(i)} - \theta x_{[i]}$, and $x_{[i]}$ is the concomitant of $y_{(i)}$, i.e. $x_{[i]}$ is that observation x_i which is coupled with $y_{(i)}$, when (y_i, x_i) are ordered with respect to y_i; $i \le i \le n$. Here, we realize that x is assumed to be non-stochastic in nature in the super-population linear regression model $y = \theta x + e$, $Cov\left(x_i, e_j\right)$ is not affected by the ordering of the y values for $1 \le i \le n$ and $1 \le j \le n$; hence

$$Cov(\hat{\mu},\bar{x}) = (1/\theta)\left\{Cov(\hat{\mu},\bar{y}) - Cov(\bar{e}_{[\cdot]},\bar{e})\right\},$$

where $Cov(\bar{e}_{[\cdot]},\bar{e}) = (\boldsymbol{\beta}'\Omega\boldsymbol{\omega})\frac{\sigma_e^2}{m}$,

Note that if the sampling fraction n/N exceeds 5%, the finite population correction $(N - n)/N$ can be introduced as

$$Cov(\hat{\mu},\bar{x}) = \{(N-n)/N\theta\}\left\{Cov(\hat{\mu},\bar{y})\right\} - Cov(\bar{e}_{[\cdot]},\bar{e})\}$$

4. Monte Carlo simulation study

In this study for the simulation, we have used R-programming software. In the super-population models generated, we use the model

$$y_i = \theta x_i + e_i, i = 1, 2, \dots, N, \tag{4.1}$$

where we generate e_i and x_i independently and calculate y_i for i = 1, 2, ..., N. Let the errors $e_1, e_2, \dots, e_N$ be the random observations from a super-population from (2.1) with $E(e) = 0$ and $V(e) = \sigma_e^2$. Let $U_N$denotes the corresponding finite population consists of N pairs (x_1, y_1), (x_2, y_2), ..., (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 conduct all Monte Carlo studies as follows.

We take N = 500 in each simulation and generate U_{500} pairs from an assumed super-population. From the generated finite population U_{500}, 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:

$MSE(T_1) = \frac{1}{s}\sum_{j=1}^{s}(T_{1j} - \overline{Y})^2$, $MSE(t_1) = \frac{1}{s}\sum_{j=1}^{s}(t_{1j} - \overline{Y})^2$, $MSE(\bar{y}_{pu}) = \frac{1}{s}\sum_{j=1}^{s}(\bar{y}_{pj} - \overline{Y})^2$ and $MSE(t_2) = \frac{1}{s}\sum_{j=1}^{s}(t_{2j} - \overline{Y})^2$.

For setting the population correlation ρ_{yx} sufficiently high, we choose the value of parameter θ in the model $y = \theta x + e$, such that the correlation coefficient between study variable (y) and auxiliary variable (x) is ρ_{yx}. To determine the value of θ 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 θ for which the population correlation between y and x becomes $\theta^2 = \frac{12\sigma^2\rho_{yx}^2}{1-\rho_{yx}^2}$ for the LTS family. Similarly, if x is generated from Exp (1), the value of θ for the population correlation becomes $\theta^2 = \frac{\sigma^2\rho_{yx}^2}{1-\rho_{yx}^2}$ for the symmetric family. In the same way, we can have $x \sim exp(0.5), x \sim N(0,1), x \sim U(-1, 2.5)$ etc. and the corresponding values of θ can be calculated accordingly. Here, we take σ^2 = 1, in all situations without loss of generality and calculate the required parameter θ for which ρ_{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 $\bar{y}_{pu}, t_1$ and t_2 are given as follows:

$MSE(T_1) \le MSE(t_2)$ if

$$Cov(\bar{y}, \bar{x}) \ge \frac{1}{2\bar{R}_1\gamma}\{E(\hat{\mu} - \overline{Y})^2 - E(\bar{y} - \overline{Y})^2\} + Cov(\hat{\mu}, \bar{x}), \tag{5.1}$$

$MSE(T_1) \le MSE(t_1) if$

$$Cov(\bar{y}, \bar{x}) \ge \frac{1}{2R\gamma}\{E(\hat{\mu} - \overline{Y})^2 - E(\bar{y} - \overline{Y})^2\} + \frac{\gamma}{2R}V(\bar{x})\{\bar{R}_1^2 - R^2\} + \frac{\bar{R}_1}{R}Cov(\hat{\mu}, \bar{x}) \tag{5.2}$$

$MSE(T_1) \le MSE(\bar{y}_{pu}) if$

$$Cov(\bar{y},\bar{x}) \geq \frac{1}{2R}\{E(\hat{\mu}-\overline{Y})^2 - E(\bar{y}-\overline{Y})^2\} + \frac{1}{2R}V(\bar{x})\{\bar{R}_1^2\gamma^2 - R^2\} + \frac{\bar{R}_1\gamma^2}{R}Cov(\hat{\mu},\bar{x}), \tag{5.3}$$

$MSE(T_1) \leq MSE(t_2) \geq MSE(t_1)$*if*

$$\frac{1}{2\bar{R}_1\gamma}\{E\left(\hat{\mu}-\overline{Y}\right)^2 - E\left(\overline{Y}-\overline{Y}\right)^2\} + Cov(\hat{\mu},\bar{x}) \leq Cov(\bar{y},\bar{x}) \leq \frac{\gamma}{2\bar{R}_1}V(\bar{x})\{R^2-\bar{R}_1^2\} + \frac{R}{\bar{R}_1}Cov(\bar{y},\bar{x}) \tag{5.4}$$

$MSE(T_1) \leq MSE(t_2) \leq MSE(\bar{y}_{pu})$ if

$$\frac{1}{2\bar{R}_1\gamma}\{E\left(\hat{\mu}-\overline{Y}\right)^2 - E\left(\bar{y}-\overline{Y}\right)^2\} + Cov(\hat{\mu},\bar{x}) \leq Cov(\bar{y},\bar{x}) \leq \frac{1}{2\bar{R}_1\gamma}V(\bar{x})\left\{R^2-\bar{R}_1^2\gamma^2\right\} + \frac{R}{\bar{R}_1\gamma}Cov(\bar{y},\bar{x}), \tag{5.5}$$

and

$MSE(T_1) \leq MSE(t_2) \leq MSE(t_1) \leq MSE(\bar{y}_{pu})$ if

$$\begin{aligned}\frac{1}{2\bar{R}_1\gamma}\{E\left(\hat{\mu}-\overline{Y}\right)^2 - E\left(\bar{y}-\overline{Y}\right)^2\} + Cov(\hat{\mu},\bar{x}) \leq Cov(\bar{y},\bar{x}) \leq \frac{\gamma}{2\bar{R}_1}V(\bar{x})\left\{R^2-\bar{R}_1^2\right\} + \frac{R}{\bar{R}_1}Cov(\bar{y},\bar{x}) \\ \leq \frac{1}{2\bar{R}_1\gamma}V(\bar{x})\{R^2-\bar{R}_1^2\gamma^2\} + \frac{R}{\bar{R}_1\gamma}Cov(\bar{y},\bar{x}),\end{aligned} \tag{5.6}$$

where $Cov(\bar{y},\bar{x}) = \left(\frac{1}{n}-\frac{1}{N}\right)S_{yx}$.

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) $x \sim U(-1, 2.5)$ and $e \sim LTS(p, 1)$.

(2) $x \sim exp(1)$ and $e \sim LTS(p, 1)$.

For the models (1) and (2), the values θ which makes the population correlation $\rho_{yx} = -0.45$ are given in Table 1.

Here, we note that for the LTS family (2.1), the value of θ 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 Figures 1 and 2.

Relative efficiencies are calculated by $RE = \frac{MSE(\bar{y}_{pu})}{MSE(\cdot)} * 100$,

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

Table 1. Parameter values of θ used in models (1)–(2) that give $\rho_{yx} = -0.45$

Population	*p*		
	2.5	**4.5**	**5.5**
Model (1)	−1.746	−1.746	−1.746
Model (2)	−0.504	−0.504	−0.504

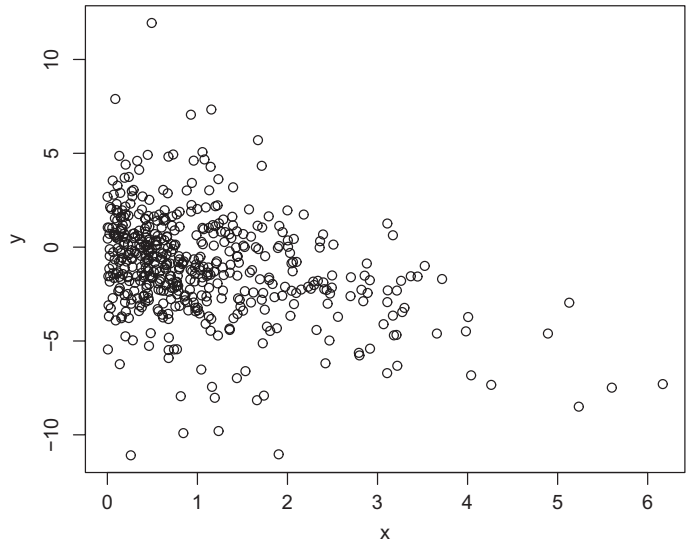

Figure 1. A scatter graph of the study variable and auxiliary variable obtained from model (2) for $p = 3.5$.

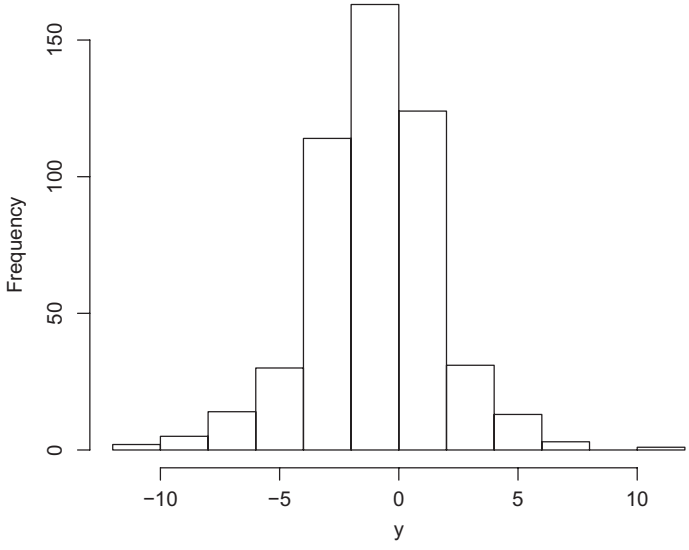

Figure 2. Underlying distribution of the study variable obtained from model (2) for $p = 3.5$.

From Table 2, we see that the proposed estimator T_1 is more efficient than the corresponding estimators $\bar{y}_{pu}, t_1$ 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 such as an estimator is full efficient or nearly so for an assumed model and maintains high efficiencies for plausible to the assumed model.

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 $x \sim U(-1, 2.5)$ as well as $x \sim Exp(1)$ and $y \sim LTS(p = 3.5, \sigma^2 = 1)$. We determine our super-population model as follow:

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			n		
		5	11	15	5	11	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_1	197.13 (0.2719)	186.91 (0.1108)	163.28 (0.0896)	191.18 (0.2415)	171.62 (0.1138)	169.62 (0.0851)
	$\bar{y}_{pu}$	100.00 (0.536)	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_1	224.26 (0.2432)	173.99 (0.1111)	178.19 (0.0784)	172.01 (0.2547)	160.93 (0.1180)	164.39 (0.0865)
	$\bar{y}_{pu}$	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 (0.1198)	152.87 (0.0872)
	t_1	208.17 (0.2533)	178.49 (0.1190)	189.76 (0.0801)	188.48 (0.2544)	171.00 (0.1200)	152.69 (0.0873)
	$\bar{y}_{pu}$	100.00 (0.5273)	100.00 (0.2124)	100.00 (0.1520)	100.00 (0.4795)	100.00 (0.2052)	100.00 (0.1333)

Note: Mean square errors are in the parenthesis.

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

(6) Dixon's outliers model: $N–N_o$ observations from LTS (3.5, 1)and N_o (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 (α_i, β_i) from (2.7) are calculated with $p = 3.5$ and are used in models (5) and (6). N_o in the model (6) is calculated from the formula $(|0.5 + 0.1 * N| = 50)$ for $N = 500$. We standardized the generated $e_i's, (i = 1, 2, \ldots, N)$ 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 the Table 3, we see that the proposed estimator T_1 is more efficient than the corresponding estimators $\bar{y}_{pu}, t_1$ 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.

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

Est.	*n*			*n*		
	5	**11**	**15**	**5**	**11**	**15**
	True Model (5): $x \sim Uni\ (0, 1)$			Dixon outlier Model (6): $x \sim Uni\ (0, 1)$		
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_1	213.17 (0.2559)	187.87 (0.1105)	177.28 (0.0876)	248.25 (1.0527)	209.93 (0.3835)	182.24 (0.2534)
$\bar{y}_{pu}$	100.00 (0.5455)	100.00 (0.2076)	100.00 (0.1553)	100.00 (2.6133))	100.00 (0.8051))	100.00 (0.4618)
	Mis - specified Model (7): $x \sim Uni\ (0, 1)$			True Model (5): $x \sim Exp\ (1)$		
T_1	231.94 (0.2683)	183.21 (0.1102)	185.40 (0.0822)	207.97 (0.2308)	184.50 (0.1026)	174.59 (0.0724)
t_2	226.62 (0.2746)	179.15 (0.1127)	180.57 (0.0844)	201.34 (0.2384)	167.97 (0.1127)	159.80 (0.0791)
t_1	226.62 (0.2747)	178.99 (0.1128)	180.36 (0.0845)	201.01 (0.2388)	167.82 (0.1128)	159.60 (0.0792)
$\bar{y}_{pu}$	100.00 (0.6223)	100.00 (0.2019)	100.00 (0.1524)	100.00 (0.4800)	100.00 (0.1893)	100.00 (0.1264)
	Dixon outlier Model (6): $x \sim Exp\ (1)$			Mis - specified Model (7): $x \sim Exp\ (1)$		
T_1	246.62 (0.3029)	198.58 (0.1354)	211.18 (0.0966)	188.42 (0.2349)	174.13 (0.1090)	174.04 (0.0782)
t_2	243.64 (0.3066)	188.68 (0.1404)	197.87 (0.1031)	183.19 (0.2416)	163.90 (0.1158)	169.28 (0.0804)
t_1	242.77 (0.3077)	188.14 (0.1408)	197.30 (0.1034)	182.82 (0.2421)	163.76 (0.1159)	169.07 (0.0805)
$\bar{y}_{pu}$	100.00 (0.7470)	100.00 (0.2649)	100.00 (0.2040)	100.00 (0.4426)	100.00 (0.1898)	100.00 (0.1361)

Note: Mean square errors are in the parenthesis.

The population quantiles $t_{(i)}$ are determined from the equation

$\int_{-\infty}^{t_{(i)}} t(u)du = \frac{i}{n+1};\ 1 \le i \le n$, 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. Pareto distribution etc.), which is applicable in most of 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 $\bar{y}_{pu}$ and t_1. We also show that the MML integrated dual to product estimator (T_1) is robust to outliers as well as other data anomalies.

Acknowledgment
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[1]
E-mail: sanjay.kumar@curaj.ac.in
Priyanka Chhaparwal[1]
E-mail: priyankachhaparwal4@gmail.com
[1] Department of Statistics, Central University of Rajasthan, Bandarsindri, Kishangarh, Ajmer, Rajasthan 305817, India.

Citation information
Cite this article as: A robust unbiased dual to product estimator for population mean through modified maximum likelihood in simple random sampling, Sanjay Kumar & Priyanka Chhaparwal, *Cogent Mathematics* (2016), 3: 1168070.

References

Bandopadhyay, S. (1980). Improved ratio and product estimators. *Sankhya, 42*, 45–49.
Bouza, C. N. (2008). Ranked set sampling for the product estimator. *Revista Investigación Operacional, 29*, 201–206.
Bouza, C. N. (2015). A family of ratio estimators of the mean containing primals and duals for simple random sampling with replacement and ranked set sampling designs. *Journal of Basic and Applied Research International, 8*, 245–253.
Chanu, W. W., & Singh, B. K. (2014). Improved class of ratio -cum-product estimators of finite population mean in two phase sampling. *Global Journal of Science Frontier Research: Mathematics and Decision Sciences, 14*, 69–81.
Cochran, W. G. (1940). The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science, 30*, 262–275. http://dx.doi.org/10.1017/S0021859600048012
Cochran, W. G. (1977). *Sampling techniques*. New York, NY: Wiley.
Hartley, H. O., & Ross, A. (1954). Unbiased ratio estimators. *Nature, 174*, 270–271. http://dx.doi.org/10.1038/174270a0
Islam, T., Shaibur, M. R., & Hossain, S. S. (2009). Effectivity of modified maximum likelihhod estimators using selected ranked set sampling data. *Austrian Journal of Statistics, 38*, 109–120.
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 THE JAPAN STATISTICAL SOCIETY, 36*, 107–119. http://dx.doi.org/10.14490/jjss.36.107
Murthy, M. N. (1964). Product method of estimation. *Sankhya A, 26*, 69–74.
Murthy, M. N. (1967). *Sampling theory and methods*. Calcutta: Statistical Publishing Society.
Oral, E. (2006). Binary regression with stochastic covariates. *Communications in Statistics: Theory and Methods, 35*, 1429–1447. http://dx.doi.org/10.1080/03610920600637123
Puthenpura, S., & Sinha, N. K. (1986). Modified maximum likelihood method for the robust estimation of system parameters from very noisy data. *Automatica, 22*, 231–235. http://dx.doi.org/10.1016/0005-1098(86)90085-3
Rao, J. N. K., & Beegle, L. D. (1967). A Monte Carlo study of some ratio estimators. *Sankhy¯a: Series B, 29*, 47–56.
Robson, D. (1957). Applications of multivariate polykays to the theory of unbiased ratio-type estimation. *Journal of the American Statistical Association, 52*, 511–522. http://dx.doi.org/10.1080/01621459.1957.10501407
Singh, S. (2003). *Advanced sampling theory with applications* (Vol. 1). Kluwer Academic Publishers, the Netherlands. http://dx.doi.org/10.1007/978-94-007-0789-4
Sukhatme, P. V., Sukhatme, B. V., & Asok, C. (1984). *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*, 35–57.
Tiku, M. L., & Bhasln, P. (1982). Usefulness of robust estimators in sample survey. *Communications in Statistics: Theory and Methods, 11*, 2597–2610. http://dx.doi.org/10.1080/03610918208828409
Tiku, M. L., & Kumra, S. (1981). Expected values and variances and covariances 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. http://dx.doi.org/10.1016/0378-3758(92)90088-A
Tiku, M. L., & Vellaisamy, P. (1996). Improving efficiency of survey sample procedures through order statistics. *Journal of Indian Society Agricultural Statistics, 49*, 363–385.
Tiku, M. L., Tan, M. Y., & Balakrishnan, N. (1986). *Robust inference*. New York, NY: Marcel Dekker.
Vaughan, D. C. (1992a). On the tiku-suresh method of estimation. *Communications in Statistics-Theory and Methods, 21*, 451–469. http://dx.doi.org/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. http://dx.doi.org/10.1080/03610919208813025
Vaughan, D. C., & Tiku, M. L. (2000). Estimation and hypothesis testing for a nonnormal bivariate distribution with applications. *Mathematical and Computer Modelling, 32*, 53–67. http://dx.doi.org/10.1016/S0895-7177(00)00119-9
Yates, F. (1960). *Sampling methods in censuses and surveys*. London: Charles Griffin.
Zheng, G., & Al-Saleh, M. F. (2002). Modified maximum likelihood estimators based on ranked set samples. *Annals of the Institute of Statistical Mathematics, 54*, 641–658. http://dx.doi.org/10.1023/A:1022475413950

cogent oa

© 2016 The Author(s). This open access article is distributed under a Creative Commons Attribution (CC-BY) 4.0 license.

You are free to:
Share — copy and redistribute the material in any medium or format
Adapt — remix, transform, and build upon the material for any purpose, even commercially.
The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:
Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made.
You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
No additional restrictions

You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.