Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Estimation of finite population mean using double sampling under probability proportional to size sampling in the presence of extreme values

Jing Wang [a], Sohaib Ahmad [b], Muhammad Arslan [c,h,*], Showkat Ahmad Lone [d], A.H. Abd Ellah [e], Maha A. Aldahlan [f], Mohammed Elgarhy [g]

[a] *School of Economics and Management, Taiyuan Normal University, Jinzhong 030619, China*
[b] *Department of Statistics, Abdul Wali Khan University, Mardan, Pakistan*
[c] *Department of Mathematics and Statistics, Institute of Southern Punjab, Pakistan*
[d] *Department of Basic Sciences, College of Science and Theoretical Studies, Saudi Electronic University, Riyadh 11673, Saudi Arabia*
[e] *Mathematics Department, Faculty of Science, Al-Baha University, Saudi Arabia*
[f] *Department of Statisitcs, College of Science, University of Jeddah, Jeddah 23218, Saudi Arabia*
[g] *Mathematics and Computer Science Department, Faculty of Science, Beni-Suef University, Beni-Suef 62521, Egypt*
[h] *School of Finance and Economics, Jiangsu University, Zhenjiang 212013, China*

ARTICLE INFO

Keywords:
PPS
double sampling
Auxiliary variable
Bias
MSE
PRE

ABSTRACT

Values that are too large or small enough can be found in many data sets. Therefore, the estimator can yield ambiguous findings if several of the incredible deals are picked for the sample. When such extreme values occur, we propose improved estimators to determine the finite population means using double sampling based on probability proportional to size sampling (PPS). The properties of estimators are obtained up to the first order of approximations. When the size of the units varies widely, the PPS sampling technique may be employed. To determine the values of *Pi* when using PPS, we must be acquainted with the aggregate of the auxiliary variable X_i. However the designs and estimation techniques we have looked at so far are unsuccessful and are less effective when this information is difficult to locate or when other information is missing. The two-phase approach is preferable and more feasible in these kinds of circumstances. To demonstrate how effectively the recommended estimators performed, we used three actual data sets. We show mathematically and theoretically that the suggested estimators outperform alternative estimators.

1. Introduction

The effective use of auxiliary variables in survey sampling may boost the precision of estimators of the population parameter. The best statistical property estimates for population quantities, like mean, total, median, etc., are frequently searched for by researchers. For this, an illustrative sample of the population is needed. If the aggregate of concern is equivalent, choosing the entities can be done

* Corresponding authorDepartment of Mathematics and Statistics, Institute of Southern Punjab, Pakistan.
E-mail addresses: wangjingjjx@tynu.edu.cn (J. Wang), sohaib_ahmad@awkum.edu.pk (S. Ahmad), arslanma76172@gmail.com (M. Arslan), s.lone@seu.edu.sa (S. Ahmad Lone), aabdulmotajali@bu.edu.sa (A.H.A. Ellah), maal-dahlan@uj.edu.sa (M.A. Aldahlan), m_elgarhy85@sva.edu.eg (M. Elgarhy).

https://doi.org/10.1016/j.heliyon.2023.e21418
Received 9 June 2023; Received in revised form 3 October 2023; Accepted 20 October 2023
Available online 21 October 2023

utilizing a SRS approach. It is necessary to know the aggregate constraints of the auxiliary variable in order to use the ratio, product, or regression methods of estimate. The ratio estimator plays an important role when there is a significant connection between the research and the auxiliary information. Apart from, the product estimator works effectively when there is a lack of association amongst the research and the auxiliary variable. By applicably adapting the auxiliary information, numerous researchers have developed various ratio estimators. Researchers can investigate this research by looking at [1] recommended on certail procedures of enlightening ratio and regression estimators [2]. recommended a better class of estimators for the mean of the population that use PPS sampling [3]. suggested that under linear transformation of the auxiliary variable, exponential estimators of the population mean be of the ratio type [4]. they reviewed a class of estimators of the population mean that hold satisfactorily against linear modification of the auxiliary information [5]. discussed a class of exponential ratio estimators consuming two auxiliary information [6]. studied mean estimate using quantile regression ratios under full and partial auxiliary information [7]. suggested robust quantile regression with two more variables for mean estimation [8]. recommended methods of enlightening estimators. The [9] discussed ameliorate estimation of mean using skewness and kurtosis of auxiliary character [10]. recommended a class of product estimators of population mean utilizing auxiliary information has been presented and questioned [11]. suggested an estimation of the population mean that was of the generalized exponential type and used auxiliary features [12]. recommended estimators of the mean of a population using simple random sampling that are based on robust ratios were proposed [13]. estimators for the mean of a population that make use of supplementary data and execute consecutive sampling on two occasions are recommended [14]. presented several imputation strategies for addressing missing information in two-sample consecutive sampling [15,16]. recommended estimation of population mean under probability proportional to size sampling with and without measurement errors.

In various situations, such as medical studies or surveys, it is common for the population sizes to diverge significantly. This can lead to variations in the probabilities or outcomes of different units within the population. For example, in a medical study examining a specific disease may be relatively small compared to the overall population size. This divergence in size can affect the probability of selecting individuals with the disease in a random sample. Researchers may need to account for this difference in population size and adjust their sampling methods or statistical analyses accordingly to ensure accurate representation and valid conclusions. Similarly, in surveys related to family income, the number of siblings within families can vary widely. This divergence in family size can influence the overall distribution of income levels within the survey population. It may be necessary to consider the different family sizes when analyzing the survey data or drawing conclusions about the relationship between family income and other variables. In situations like these, statistical techniques such as weighting, stratified random sampling, or other methods can be employed to address the divergent population sizes and account for the varying probabilities of units within the population. These techniques aim to provide accurate estimates and make valid inference despite the difference in population sizes. We utilize PPS sampling to deal with such an unequal probability. A PPS is an unequal random sampling in which, for each sampling component taken collectively, the chance of choices is proportional to an auxiliary variable. Let the context where we must evaluate the population in districts inside a province; we choose the auxiliary variable that has the determined relationship with the research variable.

For example.

(i) The aggregate of all districts inside the province (associated with research variable = 0.85).
(ii) The quantity of families in all societies inside the districts (association with the study variable = 0.98).

On the origin of these facts: (ii) more useful as an auxiliary variable.

Researchers can investigate this research by looking at [17] a discussion of using outliers to estimate the average of a population using a probability-based sampling design [18,19]. recommended PPS when outliers are present [20]. discussed combination of ratio and PPS estimators [21]. offered a more accurate estimation of the population size using PPS data [22]. discussed improved estimators in simple random sampling [23]. recommended on mixture of ratio and PPS estimators [24]. recommended substitute estimators in PPS sampling [25]. two auxiliary variables were suggested for improved estimate of the population mean using PPS.

Therefore when evidence like that is not readily accessible or when the auxiliary variable is not available, the earlier designs and estimating procedures do not produce capable results, and their efficiency decreases. Double-phase sampling is more beneficial and effective in this situation. The populations mean of the auxiliary information, which will be used in the evaluation or selection phase, can be estimated using an adequate initial sample.

For example:

On the condition of a single auxiliary information X, we take a sizeable investigative sample for estimating the population mean and only a subsample for computing the research variable Y because obtaining evidence on X is less expensive. This may imply allocating a portion of the assets to this large initial sample, resulting in a smaller sample size for computing the study variable. When the improvement in accuracy is significant compared to the rise in price due to the gathering of information on the auxiliary information for huge samples, this technique is favorable. The difficulty of calculating total buffalo milk production in a given region is an actual illustration of this situation. We use a community as the sampling element and the quantity of milk buffalo in a community as the auxiliary information in this study. Because the whole amount of milk buffalo in each community in the region may not be known, the investigator may choose a huge sample of communities and gather data on the number of milk buffalo in each village. This data is then utilized to calculate an estimate of X, the total number of milk buffalo in the area. The researchers are focused on an article regarding double-phase sampling at [26] who proposed the generalized regression estimator for two-phase tax record samples [27]. recommended the mean of a finite population can be estimated using linear regression and the ratio product [24]. presented double-sampling modified exponential estimators for the mean of a finite population [28]. recommended combining exponential functions for effective estimate when two-phase sampling is used [29]. in the context of stratified two-stage sampling, we talked about exponential chain ratio estimators [30]. consuming two auxiliary information in stratified two-phase sampling, a new, more accurate calibration estimator was presented [31]. recommended a family of estimators for predicting population mean from auxiliary proportions in single- and two-stage samples [32]. discussed a two-phase sampling method that uses a generalized methodology to estimate a finite population mean was suggested [33]. estimated the mean of a finite population using a mixed exponential-type estimator and a two-stage sampling design [34]. proposed that two-phase sampling could improve mean population estimates [35]. recommended an effective group of double-sampling estimators for the population mean [36]. for double-sampling the mean of a finite population, an exponential estimator of the chain-ratio type is proposed.

Our primary objectives are highlighted as follows.

1. In this paper, the primary objective of the contemporary effort is to estimate the finite population means using double sampling under PPS in the existence of extreme values (minimum and maximum values).
2. The numerical properties i.e. bias and MSE of the recommended estimator, are consequent up to the first order of approximation.
3. The application of the recommended estimator is highlighted through the use of real data sets from various domains.

2. Sampling methodology

Let a population $\Psi = \{\Psi_1, \Psi_2, \ldots, \Psi_N\}$ of size N unlike elements. In the first phase, we draw an initial large sample of size "m" ($m < N$) from Ψ by making use of the SRSWOR sampling design and estimating the auxiliary information x. In the second phase, we take out a sub-sample of size "n" from the first phase of size "m", i.e., ($n < m$) by SRSWOR or at first hand from Ψ, and notice both the study and auxiliary variables. Consider y_i, x_i and z_i to be the study and auxiliary variables, respectively.

Let $P_i = \frac{Z_i}{\sum_{i=1}^{N} Z_i}$, be the PPS to size for i^{th} units, where

$$\bar{u}_{11} = \frac{\sum_{i=1}^{n} u_{i11}}{n} = \bar{y}_{pps}, \bar{v}_{11} = \frac{\sum_{i=1}^{N} v_{i12}}{n} = \bar{x}_{pps},$$

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i, \text{ and } \bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i,$$

$$u_{i11} = \frac{1}{NP_i}\, y_i, v_{i12} = \frac{1}{NP_i}\, x_i, s_{u11}^2 = \sum_{i=1}^{N} P_i (u_{i11} - \bar{Y})^2, s_{v12}^2 = \sum_{i=1}^{N} P_i (v_{i12} - \bar{X})^2,$$

$$\rho_{uv} = \frac{\sum_{i=1}^{N} (u_{i11} - \bar{Y})(v_{i12} - \bar{X})}{s_{u11} s_{v12}}, s_{u11} = \sqrt{P_i (u_{i11} - \bar{Y})^2}, s_{v12} = \sqrt{P_i (v_{i12} - \bar{X})^2}.$$

Some real data sets include extreme values, e.g., when estimating the intelligence quotient (IQ), the brilliant students got (maximum) marks, and the weak students got (minimum) marks. If there are unexpectedly large or small elements in the population, the finite population mean is particularly delicate to unpredicted values. Furthermore, because the mean estimator is particularly delicate to such unpredicted findings, the population mean will either be ordinary or overstated depending on whether the sample contains large or small values. Consequently, if any of the surprising values are picked in the sample, the estimator can produce ambiguous conclusions. [37], suggested the following unbiased estimator to overcome this issue, which is given in equation (1).

$$\widehat{\bar{y}_{ss}} = \begin{cases} \bar{y} + s, \text{if the sample contains only minimum, not maximum values} \\ \bar{y} - s, \text{if the sample contains only maximum, not minimum values} \\ \qquad \bar{y}, \quad \text{if sample contains all} \\ \qquad\qquad \text{observation} \end{cases} \tag{1}$$

$$\text{Var}(\widehat{\bar{y}_{ss}}) = \varphi s_{u11}^2 - \frac{2\varphi n c_2}{N-1}[\sigma_{u11} - nc_1]$$

The MSE of $\widehat{\bar{y}_{ss}}$, at the unknown value of c, which is given in equation (2):

$$\text{Var}(\widehat{\bar{y}_{ss}})_{min} = \text{Var}(\bar{y}) \text{ - } \frac{\varphi {\sigma_{u11}}^2}{2(N-1)}, \tag{2}$$

where

$$\text{Var}(\bar{y}) = \varphi s_{u11}^2$$

[11]recommended population total under PPS, which are given in equation (3):

$$\widehat{\overline{y}_{pps}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i+\varphi)}{p_i} \text{ - N } \varphi, \tag{3}$$

where, $p_i = \frac{c_{xi+\varkappa}}{c_{xi+N\varkappa}}$.

For estimation of the population means, we can also write equation (3) as given by:

$$\widehat{\overline{y}_{pps}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i+\varphi)}{Np_i} - \varphi$$

$\frac{(c_{xi+N\varkappa})}{N\varkappa}\sum_{i=1}^{N}\frac{(y_i+\varphi)}{c_{xi+\varkappa}}$, when $c = 1$, $\varkappa = 0$, $\varphi = 0$

$$\widehat{\overline{y}_{pps}} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i)}{Np_i} = \frac{1}{n}\sum_{i=1}^{n}u_{i11} = \overline{u}_{11} = \overline{y}_{pps}$$

The variance of $\overline{y}_{pps}$ is given in equation (4):

$$\text{Var}\left(\widehat{\overline{y}_{T,PPS}}\right) = \varphi\overline{u}_{11}c_{u11}^2 \tag{4}$$

The ratio and product estimators [38,39] which are given in equations (5) and (6):

$$\overline{y}_{RT,PPS} = \overline{u}_{11}\left(\frac{\overline{X}^*}{\overline{v}_{11}}\right), \tag{5}$$

$$\overline{y}_{PT,PPS} = \overline{u}_{11}\left(\frac{\overline{v}_{11}}{\overline{X}^*}\right), \tag{6}$$

The MSE of $\overline{y}_{RT,PPS}$, and $\overline{y}_{PT,PPS}$ are given in equations (7) and (8):

$$\text{MSE}\left(\overline{y}_{RT,PPS}\right) = \overline{Y}^2\left\{\varphi c_{u11}^2 + \varphi_2 c_{v12}\left(c_{v12} - 2\rho_{u11v12}c_{u11}\right)\right\} \tag{7}$$

and

$$\text{MSE}\left(\overline{y}_{PT,PPS}\right) = \overline{Y}^2\left\{\varphi c_{u11}^2 + \varphi_2 c_{v12}\left(c_{v12} + 2\rho_{u11v12}c_{u11}\right)\right\} \tag{8}$$

The regression estimator is given in equation (9):

$$\overline{y}_{RegT,PPS} = \overline{u}_{11} + \beta(\overline{X}^* - \overline{v}_{11}). \tag{9}$$

The variance of regression is given in equation (10):

$$\text{Var}\left(\overline{y}_{RegT,PPS}\right) = \varphi s_{u11}^2 + \varphi_2 s_{u11}^2\left(1\text{- }\rho_{u11v12}^2\right). \tag{10}$$

Where.

$\varphi = \left(\frac{1}{n} - \frac{1}{N}\right)$, $\varphi' = \left(\frac{1}{m} - \frac{1}{N}\right)$, $\varphi_2 = \left(\frac{1}{n} - \frac{1}{m}\right)$.

3. Suggested estimators

Some real data sets included extreme values, either very large or small. The efficiency of estimators may suffer in the manifestation of these extreme values. For example, while measuring the average export of goods, China may produce a large number of goods for the international market due to new technology and improved skills of its people, compared to Pakistan's small amount of goods due to poor management and lack of technology. Similarly, if we wish to know the average yearly wheat production in our country, we can see that wheat production in Punjab is extremely large as compared to other provinces. To deal with such an extreme values taking motivation from Refs. [17,18], we suggested an improved ratio, product, and regression type estimator for double phase with PPS sampling in the occurrence of extreme values. The recommended improved estimators are presented in three different situations.

Situation-I: Mean per unit estimator, given in equation (11)

$$\overline{y}_{T,PPS} = \begin{cases} \overline{u}_{11} + c, \text{If the selected} \\ \text{observation included small value of } \mathrm{u}_{i11} \\ \overline{u}_{11} - c, \text{If the selected} \\ \text{observation included large value of } \mathrm{u}_{i11} \\ \overline{u}_{11}, \text{If the selected} \\ \text{observation included other values} \end{cases} \tag{11}$$

The optimal value of C, is given as:

$$C = \frac{\varphi\sigma^2_{u11}}{2(N-1)},$$

The least variance at the value of C are given in equation (12):

$$\mathrm{V}\left(\widehat{\overline{Y}_{T,PPS}}\right) = \mathrm{V}\left(\widehat{\overline{y}_{T,PPS}}\right) \text{ - } \frac{\varphi\sigma^2_{u11}}{2(N-1)} \tag{12}$$

Situation-II: When u and v are positively correlated.

When the correlation between u and v is positive, when the minimum cost of u is chosen, the collection of the minimum value of v is presumed. And for a maximum value of v, a maximum cost of u is assumed to be nominated. In such a scenario, we suggest the following improved ratio type estimator, which is given in equation (13).

$$\widehat{\overline{Y}_{RT,PPS}} = \overline{u}_{c11}\left(\frac{\overline{X}^*}{\overline{v}_{c21}}\right), \tag{13}$$

or

$$\widehat{\overline{Y}_{RT,PPS}} = \begin{cases} (\overline{u}_{11}+c_1)\dfrac{(\overline{X}^*+c_2)}{(\overline{v}+c_2)}, \text{If the sample included small value of } \mathrm{u}_{\mathrm{i}11} \text{ and } \mathrm{v}_{\mathrm{i}12} \\ (\overline{u}_{11}-c_1)\dfrac{(\overline{X}^*-c_2)}{(\overline{v}-c_2)}, \text{If the sample included large value of } \mathrm{u}_{\mathrm{i}11} \text{ and } \mathrm{v}_{\mathrm{i}12}\text{ ,} \\ \overline{u}_{11}\left(\dfrac{\overline{X}^*}{\overline{v}}\right), \text{for all other samples} \end{cases}$$

where $(\overline{u}_{c11} = \overline{u}_{11} + c_1, \overline{X}^*_{c21} = \overline{X}^* + c_2, \overline{v}_{c21} = \overline{v}_{11} + c_2)$. If the trial contains minimum values of u and v. $(\overline{u}_{c11} = \overline{u}_{11} - c_1, \overline{X}^*_{c21} = \overline{X}^* - c_2, \overline{v}_{c21} = \overline{v}_{11} - c_2)$. If a trial contains maximum values of u and v, and $(\overline{u}_{c11} = \overline{u}_{11}, \overline{X}^*_{c21} = \overline{X}^*, \overline{v}_{c21} = \overline{v}_{11})$, for all further samples. Where c_1 and c_2 are sustained, its value y be decisive for optimal conditions.

The regression estimator is given in equation (14):

$$\overline{y}_{T,Reg1,PPS} = \overline{u}_{c11} + \mathrm{b}\ (\overline{X}^* - \overline{v}_{c21}), \tag{14}$$

where $(\overline{u}_{c11} = \overline{u}_{11} + c_1, \overline{v}_{c21} = \overline{v}_{11} + c_2)$ if the trial comprises u and v minimum. $(\overline{u}_{c11} = \overline{u}_{11} - c_1, \overline{v}_{c21} = \overline{v}_{11} - c_2)$ if the trial comprises u and v maximum, and $(\overline{u}_{c11} = \overline{u}_{11}, \overline{v}_{c21} = \overline{v}_{11})$, for all other samples.

Situation-III: When u and v are negatively correlated.

While u and v are both negatively correlated with one another, the picking of a large assessment of v is expected to be accompanied by a small value of u. Similarly, when a small value of v is selected, it is expected to select a large value of u. Based on these situations, we suggested the following improved product type estimator, which is given in equation (15):

$$\widehat{\overline{Y}_{PT,PPS}} = \overline{u}_{c12}\left(\frac{\overline{v}_{c22}}{\overline{X}^*}\right), \tag{15}$$

or

$$\widehat{\overline{Y}_{PT,PPS}} = \begin{cases} (\overline{u}_{11}+c_1)\dfrac{(\overline{v}+c_2)}{(\overline{X}^*+c_2)}, \text{If the sample included small value of } \mathrm{u}_{\mathrm{i}11} \text{ and large values of } \mathrm{v}_{\mathrm{i}12} \\ (\overline{u}_{11}-c_1)\dfrac{(\overline{v}-c_2)}{(\overline{X}^*-c_2)}, \text{If the sample included large value of } \mathrm{u}_{\mathrm{i}11} \text{ and small values of } \mathrm{v}_{\mathrm{i}12} \\ \overline{u}\left(\dfrac{\overline{v}}{\overline{X}^*}\right), \text{for all other samples} \end{cases}$$

The regression estimator is given in equation (16):

$$\overline{y}_{T,Reg2,PPS} = \overline{u}_{c12} + \mathrm{b}\ (\overline{X}^* - \overline{v}_{c22}), \tag{16}$$

where $(\overline{u}_{c12} = \overline{u}_{11} + c_1, \overline{v}_{c22} = \overline{v}_{11} - c_2)$ if the sample comprises u and v minimum. $(\overline{u}_{c12} = \overline{u}_{11} - c_1, \overline{v}_{c22} = \overline{v}_{11} + c_2)$ if the sample comprises u and v maximum, and $(\overline{u}_{c12} = \overline{u}_{11}, \overline{v}_{c22} = \overline{v}_{11})$, for all other samples.

To find out biases and MSE we explain the relative error term and their expectation given as:

Let

$$\epsilon_0 = \frac{\overline{u}_{11} - \overline{U}}{\overline{U}}, \epsilon_1 = \frac{\overline{v}_{11} - \overline{V}}{\overline{V}}, \epsilon_1{}' = \frac{\overline{X}^* - \overline{V}}{\overline{V}},$$

$$E\ (\epsilon_0) = \mathrm{E}\ (\epsilon_1) = \mathrm{E}\ (\epsilon_1{}') = 0.$$

$$E\left({\epsilon_0}^2\right)=\left[\frac{\varphi}{\overline{Y}^2}\left(s_{u11}^2-\frac{2nc_2}{N-1}\left[\sigma_{u11}-nc_1\right]\right),E\left({\epsilon_1}^2\right)=\frac{\varphi}{\overline{X}^2}\ \left(s_{v12}^2-\frac{2mc_2}{N-1}\left[\sigma_{v12}-mc_2\right]\right)\right],$$

$$E\left({\epsilon_1}^{'2}\right)=\left[\frac{\varphi^{'}}{\overline{X}^2}\ \left(s_{v12}^2-\frac{2m^{'}c_2}{N-1}\left[\sigma_{v12}-mc_2\right]\right),E(\epsilon_1\epsilon_1^{'})=\frac{\varphi^{'}}{\overline{X}^2}\ \left(s_{v12}^2-\frac{2m^{'}c_2}{N-1}\left[\sigma_{v12}-mc_2\right]\right)\right],$$

$$E\left(\epsilon_0\epsilon_1\right)=\left[\frac{\varphi}{\overline{XY}}\left(s_{u11v12}-\frac{n}{N-1}\left[c_2\sigma_{u11}+c_1\sigma_v-2nc_1c_2\right]\right)\right],$$

$$E\left(\epsilon_0\epsilon_1^{'}\right)=\left[\frac{\varphi^{'}}{\overline{XY}}\left(s_{u11v12}-\frac{m}{N-1}\left[c_2\sigma_{u11}+c_1\sigma_{v12}-2mc_1c_2\right]\right)\right],$$

where

$\sigma_u=u_{max}$ - u_{min} and $\sigma_v\ =\ v_{max}$- v_{min}.

By simplifying (12), in terms of e's.
$\widehat{\overline{Y}_{RT,PPS}}\ =\ \overline{Y}(1+\epsilon_0)(1+\epsilon_1^{'})^{-1}$,or.
$\widehat{\overline{Y}_{RT,PPS}}\ =\ \overline{Y}(\epsilon_0\ -\ \epsilon_1^{'}\ -\ \epsilon_0\epsilon_1^{'}\ +\ \epsilon_1^{'2})$.
Taking expectations from both sides, we have

$$\text{Bias}\left(\overline{y}_{Tr,PPS}\right)=\overline{Y}\left(\varphi_2c_{v12}^2-\varphi_2c_{u11v12}\right)\text{-}\frac{R}{(N-1)}\left\{\frac{2c_2}{\overline{X}}\left\{(n\varphi-m\varphi^{'})\sigma_{v12}-c_2\left(n^2\varphi-m^2\varphi\right)\right\}-(n\varphi-m\varphi^{'})(c_2\sigma_{u11}+c_1\sigma_{v12})+2c_1c_2\left(n^2\varphi-m^2\varphi\right)\right\},$$

where $R=\frac{\overline{Y}}{\overline{X}}$.

Unique values of c_1 and c_2 are not possible, because we have one equation and two unknown values.

$$c_{1(optimal)}=\frac{\sigma_{u11}}{\sigma_{v12}},$$

$$c_{2(optimal)}=\left\{\frac{(N-1)\sigma_{u11}-mR\sigma_{v12}}{2mR(N-m-n)}\right\}$$

Putting the ideal values of $c_{1(optimal)}$ and $c_{2(optimal)}$, the least MSE of $\widehat{\overline{Y}_{RT,PPS}}$, is given in equation (17):

$$\text{MSE}\left(\widehat{\overline{Y}_{RT,PPS}}\right)=\text{MSE}\left(\overline{y}_{Tr,PPS}\right)\text{ - }\frac{1}{2mN(N-1)}\left[\begin{array}{c}\sigma_v\{(N-n)\sigma_{u11}+2R(n-m)\sigma_{v12}\}\\+\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] \tag{17}$$

Similarly, the bias of product type estimator is given:
$\text{B}\left(\widehat{\overline{Y}_{PT,PPS}}\right)\ =\ \left[\frac{\varphi}{\overline{XY}}\left\{s_{u11v12}\ -\ \frac{n}{N-1}\left[c_2\sigma_{u11}\ +\ c_1\sigma_{v12}\ -\ 2nc_1c_2\right]\right\}\right]$.
The MSE of product type estimator is given in equation (18):

$$\text{MSE}\left(\widehat{\overline{Y}_{PT,PPS}}\right)=\text{MSE}\left(\overline{y}_{PT,PPS}\right)\text{ - }\frac{1}{2mN(N-1)}\left[\begin{array}{c}\sigma_{v12}\{(N-n)\sigma_{u11}+2R(n-m)\sigma_{v12}\}\\+\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] \tag{18}$$

In circumstance of positive correlation, the variance of $\overline{y}_{T,Reg1,PPS}$, given in equation (19);

$$\text{Var}\left(\overline{y}_{T,Reg1,PPS}\right)=\text{MSE}\left(\overline{y}_{RegT,PPS}\right)\text{- }\frac{1}{2mN(N-1)}\left[\begin{array}{c}\sigma_{v12}\{(N-n)\sigma_{u11}-2\beta(n-m)\sigma_{v12}\}\\+\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] \tag{19}$$

In circumstance of negative correlation, the variance of $\overline{y}_{T,Reg2,PPS}$, given in equation (20):

$$\text{Var}\left(\overline{y}_{T,Reg2,PPS}\right)=\text{MSE}\left(\overline{y}_{RegT,PPS}\right)\text{ - }\frac{1}{2mN(N-1)}\left[\begin{array}{c}\sigma_{v12}\{(N-n)\sigma_{u11}+2\beta(n-m)\sigma_{v12}\}\\+\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] \tag{20}$$

Generally, we can write the variance of the regression estimator as given in equation (21):

$$\text{Var}\left(\widehat{\overline{Y}_{RegGT,PPS}}\right) = \text{MSE}\left(\bar{y}_{RegT,PPS}\right) \text{ - } \frac{1}{2mN(N-1)}\left[\begin{array}{l}\sigma_{v12}\{(N-n)\sigma_{u11}+2\,|\beta|(n-m)\sigma_{v12}\} \\ +\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] \tag{21}$$

4. Efficiency comparison

In this section, we equate theoretically the suggested estimators with existing counterparts.

(i) By taking (4) and (12)

$\text{Var}(\widehat{\overline{Y}_{T,PPS}}) < \text{Var}(\widehat{\bar{y}_{T,PPS}})$, or

$\text{Var}(\widehat{\bar{y}_{T,PPS}})$ - $\text{Var}(\widehat{\overline{Y}_{T,PPS}}) > 0$

$\frac{\varphi\sigma_{u11}^2}{2(N-1)} > 0$

(ii) By taking (7) and (17)

MSE $(\widehat{\overline{Y}_{RT,PPS}}) <$ MSE$(\bar{y}_{RT,PPS})$, or

MSE$(\bar{y}_{RT,PPS})$ - MSE $(\widehat{\overline{Y}_{RT,PPS}}) > 0$

$$\frac{1}{2mN(N-1)}\left[\begin{array}{l}\sigma_{v12}\{(N-n)\sigma_{u11}+2R(n-m)\sigma_{v12}\} \\ +\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] > 0$$

(iii) By taking (8) and (18)

MSE $(\widehat{\overline{Y}_{PT,PPS}}) <$ MSE$(\bar{y}_{PT,PPS})$, or

MSE$(\bar{y}_{PT,PPS})$ - MSE $(\widehat{\overline{Y}_{PT,PPS}}) > 0$

$$\frac{1}{2mN(N-1)}\left[\begin{array}{l}\sigma_{v12}\{(N-n)\sigma_{u11}+2R(n-m)\sigma_{v12}\} \\ +\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] > 0$$

(iv) By taking (10) and (19)

Var $(\bar{y}_{T,Reg1,PPS}) <$ Var$(\bar{y}_{RegT,PPS})$, or

Var$(\bar{y}_{RegT,PPS})$ - Var $(\bar{y}_{T,Reg1,PPS}) > 0$

$$\frac{1}{2mN(N-1)}\left[\begin{array}{l}\sigma_{v12}\{(N-n)\sigma_{u11}-2\beta(n-m)\sigma_{v12}\} \\ +\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right] > 0$$

(v) By taking (10) and (20)

Var $(\bar{y}_{T,Reg2,PPS}) <$ Var$(\bar{y}_{RegT,PPS})$, or

Var$(\bar{y}_{RegT,PPS})$ - Var $(\bar{y}_{T,Reg2,PPS}) > 0$

$$\frac{1}{2mN(N-1)}\left(\begin{array}{l}\sigma_{v}\{(N-n)\sigma_{u11}+2\beta(n-m)\sigma_{v12}\} \\ +\dfrac{(n-m)\{(N-n)\sigma_{u11}-mR\sigma_{v12}\}^2}{m(N-m-n)}\end{array}\right) > 0$$

5. Numerical investigation

We took three data sets to determine the suggested estimator's efficiency with existing counterparts. The summary statistics of these data sets are given below:

Data-I [Source: [24]]:

$Y =$ Expected fish caught throughout 1995,

$X =$ expected fish caught throughout 1994,

Z = expected fish caught throughout 1993.

Data-II [Source: [24]]:

Table 1
Summary statistic for Data-I.

$N = 69$	$\overline{X} = 4954.435$	$c_{u11} = 0.4720461$	$c_{v12} = 0.5049075$	$u_{max} = 11873.89$	$\mathrm{R} = 0.9112843$
$m = 36$	$\overline{Z} = 4591.072$	$c_{u11}^2 = 0.2228275$	$c_{v12}^2 = 0.2549316$	$u_{min} = 467.0002$	$c_{u11v12} = 0.063411$
$m_1 = 20$	$s_{u11}^2 = 2420387$	$s_{v12}^2 = 2007238$	$\rho_{u11v12} = 0.2660536$	$v_{max} = 18850.12$	$\beta = 0.292154$
$\overline{Y} = 4514.899$	$s_{u11} = 1555.759$	$s_{v12} = 1416.77$	$s_{u11v12} = 1{,}418{,}429$	$v_{min} = 894.5356$	

Table 2
Summary statistic for Data-II.

$N = 69$	$\overline{X} = 4591.072$	$c_{u11} = 0.8523346$	$c_{v12} = 1.509517$	$u_{max} = 20949.43$	$\mathrm{R} = 0.983408$
$m = 36$	$\overline{Z} = 4230.174$	$c_{u11}^2 = 0.7264743$	$c_{v12}^2 = 2.278641$	$u_{min} = 1002.527$	$c_{u11v12} = 0.05634$
$m_1 = 20$	$s_{u11}^2 = 3{,}361{,}469$	$s_{v12}^2 = 2{,}245{,}140$	$\rho_{u11v12} = 0.04379289$	$v_{max} = 54678.91$	$\beta = 0.05358$
$\overline{Y} = 4514.899$	$s_{u11} = 1833.488$	$s_{v12} = 1498.379$	$s_{u11v12} = 1{,}167{,}922$	$v_{min} = 1634.557$	

Table 3
Summary statistic for Data-III.

$N = 80$	$\overline{X} = 1126.463$	$c_{u11} = 0.775310$	$c_{v12} = 0.281887$	$u_{max} = 15586.8$	$\mathrm{R} = 4.600808$
$m = 45$	$\overline{Z} = 285.125$	$c_{u11}^2 = 0.6011057$	$c_{v12}^2 = 0.07946076$	$u_{min} = 2408.59$	$c_{u11v12} = 0.12676$
$m_1 = 25$	$s_{u11}^2 = 10{,}568{,}817$	$s_{v12}^2 = 63257.12$	$\rho_{u11v12} = 0.5800092$	$v_{max} = 1944.034$	$\beta = 7.497101$
$\overline{Y} = 5182.637$	$s_{u11} = 3250.972$	$s_{v12} = 251.5097$	$s_{u11v12} = 740038.3$	$v_{min} = 592.6127$	

Table 4
MSE of the existing and suggested estimators.

Estimators	Data-I MSE(.)	Data-II MSE(.)	Data-III MSE(.)
$\overline{y}_{T,PPS}$	2,079,853	8,604,841	12,067,836
$\overline{y}_{RT,PPS}$	219312.1	1,506,958	360886.7
$\widehat{\overline{Y}_{RT,PPS}}$	218,119	1,547,809	349856.2
$\overline{y}_{PT,PPS}$	334209.1	1,609,051	603003.9
$\widehat{\overline{Y}_{PT,PPS}}$	333016.1	1,649,902	591973.5
$\overline{y}_{RegT,PPS}$	115820.1	163927.6	358827.8
$\widehat{\overline{Y}_{RegGT,PPS}}$	102615.2	111587.4	351424.7

$Y =$ Expected fish caught throughout 1995,
$X =$ expected fish caught throughout 1993,
$Z =$ expected number of fish caught throughout 1992.
Data-III [Source: [40] **]:**
Y=Output for 80 yard,
$X =$ stable capital in a region,
$Z =$ number of labors.

6. Discussion

As previously mentioned, we evaluated the performance of our suggested estimators using three real data sets. The proposed estimators are numerically and mathematically related to their current equivalents. The actual data are summarised statistically in Tables 1–3. The MSE and PRE of our proposed and current counterparts are displayed in Tables 4 and 5. In phrase of MSE and PRE, it is detected that the suggested estimators are efficient than existing counterparts. The gain in data 2 is greater as compared to data 1 and data 3. Fig. 1 shows a comparison of estimators in terms of MSE. We plotted estimators on the X-axis and MSE values on the Y-axis. The estimator is more effective when you reduce the value of MSE. The efficiency of an estimator is directly related to the trend of lines. As the value of MSE is the minimum, the line graph shows the downward direction. Fig. 2 shows a comparison of estimators in terms of percentage relative estimators. When compared to their counterparts, our proposed estimators gain the most percentage relative efficiency. We plot estimators on the X-axis and values of PRE on the Y-axis. The higher the value of PRE, the better is the estimator. The trend line indicates an increasing path based on the PRE values.

Table 5
PRE of the existing and suggested estimators.

′Estimators	Data-I	Data-II	Data-III
$\overline{y}_{T,PPS}$	100	100	100
$\overline{y}_{RT,PPS}$	948.3534	571.0073	3343.941
$\widehat{\overline{Y}_{RT,PPS}}$	953.5406	555.9368	3449.37
$\overline{y}_{PT,PPS}$	622.3209	534.7774	2001.286
$\widehat{\overline{Y}_{PT,PPS}}$	624.5504	521.5365	2038.577
$\overline{y}_{RegT,PPS}$	1795.762	5249.171	3363.127
$\widehat{\overline{Y}_{RegGT,PPS}}$	2026.846	7711.299	3433.975

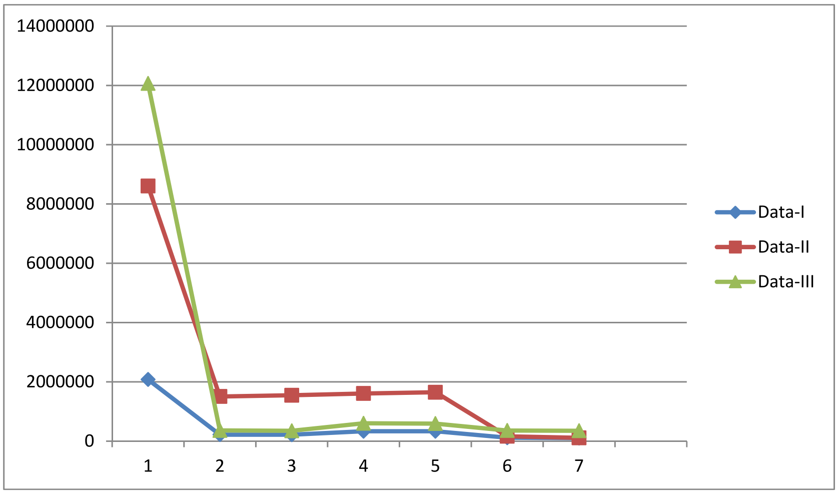

Fig. 1. MSE of the suggested and existing estimators
Fig. 1: On Y-axis, we put the values of mean square error, and on X-axis, we put the estimators.

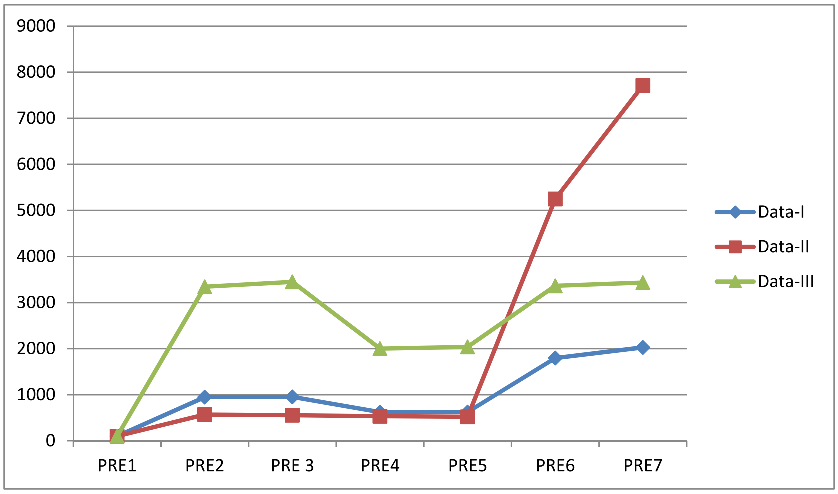

Fig. 2. PRE of the suggested and existing estimators
Fig. 2: On Y-axis, we put the values of mean square error, and on X-axis, we put the estimators.

7. Conclusion

In this paper, we have recommended an enhanced ratio, product, and regression type estimators for the estimation of finite population mean in double-phase with PPS sampling in the incidence of extreme values. The numerical expressions of properties are derived up to the first order of approximation. The purpose of this proposal is to enhance the accuracy and precision of mean estimation compared to existing estimators. To evaluate the efficiency of the recommended estimator, we conduct a comparative analysis with several existing counterparts. By comparing the performance of the proposed estimator against these alternatives, we aim to demonstrate its uniqueness and superiority. We used three actual data sets to obtain the MSEs and PRE. From the numerical results, recommended estimators perform well in terms of minimum mean square error and advanced PRE. It has been validated through empirical efficiency comparisons that our proposed estimators perform more effectively than the traditional estimators. The recommended estimators performed well, with the greatest gain in efficiency, and would perform well in applied surveys. The current work can be easily extended to yield an improved family of estimators under stratified random sampling and measurement error using the auxiliary information or attributes for estimation of population mean and variance. Additionally, it would be interesting to examine the efficiency of our recommended estimator in more complex survey settings, such as clustered and stratified sampling.

Data availability

Data will be made available on request.

Funding

There is no funding.

CRediT authorship contribution statement

Jing Wang: Formal analysis. **Sohaib Ahmad:** Conceptualization, Investigation, Writing – original draft. **Muhammad Arslan:** Formal analysis. **Showkat Ahmad Lone:** Resources, Validation. **A.H. Abd Ellah:** Methodology, Validation. **Maha A. Aldahlan:** Conceptualization, Data curation. **Mohammed Elgarhy:** Data curation, Resources.

Declaration of competing interest

The authors declare that they have no known cometing finincial intersts or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] S. Singh, Advanced Sampling Theory with Applications: How MichaelSelected Amy, ume 2, Kluwer Academic Publishers, 2003.

[2] S. Ahmad, J. Shabbir, E. Zahid, M. Aamir, Improved family of estimators for the population mean using supplementary variables under PPS sampling, Sci. Prog. 106 (2) (2023), 00368504231180085.

[3] L.K. Grover, P. Kaur, Ratio type exponential estimators of population mean under linear transformation of auxiliary variable: theory and methods, S. Afr. Stat. J. 45 (2) (2011) 205–230.

[4] L.K. Grover, P. Kaur, A generalized class of ratio type exponential estimators of population mean under linear transformation of auxiliary variable, Commun. Stat. Simulat. Comput. 43 (7) (2014) 1552–1574.

[5] R. Singh, P. Sharma, A class of exponential ratio estimators of finite population mean using two auxiliary variables, Pak. J. Statistics Oper. Res. (2015) 221–229.

[6] U. Shahzad, M. Hanif, I. Sajjad, M.M. Anas, Quantile regression-ratio-type estimators for mean estimation under complete and partial auxiliary information, Sci. Iran. 29 (3) (2022) 1705–1715.

[7] U. Shahzad, I. Ahmad, I.M. Almanjahie, N.H. Al-Noor, M. Hanif, Mean estimation using robust quantile regression with two auxiliary variables, Sci. Iran. 30 (2022) 1245–1254.

[8] T.J. Rao, On certail methods of improving ration and regression estimators, Commun. Stat. Theor. Methods 20 (10) (1991) 3325–3340.

[9] R.R. Sinha, Bharti, Ameliorate estimation of mean using skewness and kurtosis of auxiliary character, J. Stat. Manag. Syst. 25 (4) (2022) 927–944.

[10] S.K. Yadav, T. Zaman, A. Khokhar, S. Saha, Questing elevated family of product estimators of population mean using auxiliary varaibles, J. Sci. Arts 22 (2) (2022) 343–350.

[11] S. Ahmad, M. Arslan, A. Khan, J. Shabbir, A generalized exponential-type estimator for population mean using auxiliary attributes, PLoS One 16 (5) (2021), e0246947.

[12] T. Zaman, H. Bulut, S.K. Yadav, Robust ratio-type estimators for finite population mean in simple random sampling: a simulation study, Concurrency Comput. Pract. Ex. 34 (25) (2022), e7273.

[13] G.N. Singh, M. Khalid, Efficient class of estimators for finite population mean using auxiliary information in two-occasion successive sampling, J. Mod. Appl. Stat. Methods 17 (2) (2019) 14.

[14] G.N. Singh, M. Khalid, J.M. Kim, Some imputation methods to deal with the problems of missing data in two-occasion successive sampling, Commun. Stat. Simulat. Comput. 50 (2) (2021) 557–580.

[15] R.R. Sinha, Families of estimators for estimating mean using information of auxiliary variate under response and non-response, Journal of Reliability and Statistical Studies (2020) 21–60.

[16] R.R. Sinha, B. Khanna, Estimation of population mean under probability proportional to size sampling with and without measurement errors, Concurrency Comput. Pract. Ex. 34 (18) (2022), e7023.

[17] S. Ahmad, J. Shabbir, Use of extreme values to estimate finite population mean under pps sampling scheme, Journal of Reliability and Statistical Studies (2018) 99–112.

[18] S. Al-Marzouki, C. Chesneau, S. Akhtar, J.A. Nasir, S. Ahmad, S. Hussain, M. El-Morshedy, Estimation of finite population mean under PPS in presence of maximum and minimum values, AIMS Mathematics 6 (5) (2021) 5397–5409.

[19] S. Ahmad, E. Zahid, J. Shabbir, M. Aamir, R. Onyango, Enhanced estimation of the population mean using two auxiliary variables under probability proportional to size sampling, Math. Probl Eng. (2023) 2023.

[20] S. Agarwal, P. Kumar, Combination of ratio and pps estimators, J. Indian Soc. Agric. Stat. 32 (1980) 81–86.

[21] H.P. Singh, A.C. Mishra, S.K. Pal, Improved estimator of population total in PPS sampling, Commun. Stat. Theor. Methods 47 (4) (2018) 912–934.

[22] P. Sharma, R. Singh, Improved estimators in simple random sampling when study variable is an attribute, J. Stat. Appl. Pro. Lett 2 (1) (2015) 51–58.

[23] S. Pandey, R.K. Singh, On combination of ratio and PPS estimators, Biom. J. 26 (3) (1984) 333–336.

[24] J. Rao, Alternative estimators in PPS sampling for multiple characteristics, Sankhya: The Indian Journal of Statistics, Series A 28 (1966) 47–60.

[25] T. Srivenkataramana, D.S. Tracy, Transforming the study variate after PPS Sampling, Metron 37 (1) (1979) 175–181.

[26] J. Armstrong, H. St-Jean, Generalized regression estimator for two-phase sample of tax records, Surv. Methodol. 20 (1983) 91–105.

[27] H.P. Singh, M.R. Espejo, On linear regression and ratio-product estimation of a finite population mean, The Statistician 1 (2003) 59–67.

[28] Y. Hassan, M. Ismail, W. Murray, M.Q. Shahbaz, Efficient estimation combining exponential and functions under two phase sampling, AIMS Mathematics 5 (6) (2020) 7605–7623.

[29] A. Sanaullah, H.A. Ali, M.N. ul Amin, M. Hanif, Generalized exponential chain ratio estimators under stratified two-phase random sampling, Appl. Math. Comput. 226 (2014) 541–547.

[30] N. Ozgul, New improved calibration estimator based on two auxiliary variables in stratified two-phase sampling, J. Stat. Comput. Simulat. 91 (6) (2021) 1243–1256.

[31] G.M. Oyeyemi, I. Muhammad, A.O. Kareem, Combined exponential-type estimators for finite population mean in two-phase sampling, Asian Journal of Probability and Statistics 21 (2) (2023) 44–58.

[32] A. Tiwari, M. Kumar, S.K. Dubey, A generalized approach for estimation of a finite population mean in two-phase sampling, Asian Journal of Probability and Statistics 21 (3) (2023) 45–58.

[33] X. Liu, M. Arslan, A general class of estimators on estimating population mean using the auxiliary proportions under simple and two phase sampling, AIMS Mathematics 6 (12) (2021) 13592–13607.

[34] S. Bhushan, A. Kumar, Enhanced estimation of population mean under two-phase sampling, Int. J. Math. Model. Numer. Optim. 13 (1) (2023) 34–48.

[35] S.B. Bhushan, A.K. Kumar, S. Kumar, Efficient class of estimators of population mean under double sampling, Thailand Statistician 21 (3) (2023) 498–509.

[36] R. Janbandhu, N. Garg, R. Tailor, Chain ratio type exponential estimator for finite population mean in double sampling, Thailand Statistician 21 (3) (2023) 675–690.

[37] C.E. Sarndal, Sample Survey Theory vs General Statistical Theory: Estimation of the Population Mean, vol. 40, International Statistical Institute, 1972, pp. 1–12.

[38] M.N. Murthy, Sampling Theory and Methods, second ed., Statistical Publishing Society, Calcutta, 1967.

[39] Mankal Narasinha Murthy, Sampling Theory and Methods, Statistical Pub. Society, Calcutta, 1967.

[40] L.N. Upadhyaya, H.P. Singh, Use of transformed auxiliary variable in estimating the finite population mean, Biom. J. 41 (5) (1999) 627–636.