An Unbiased Estimator of Finite Population Mean Using Auxiliary Information

B. Mahanty1,*, G. Mishra2

1 P.G. Department of Statistics, Utkal University, Bhubaneswar-751004, India
2 P.G. Department of Statistics, Utkal University, Bhubaneswar-751004, India

ARTICLE INFO

Article History
Received 20 July 2020
Accepted 25 Dec 2020

Keywords
Simple random sampling
Auxiliary variable
Unbiased estimator
Optimality
Efficiency

ABSTRACT

In this paper, an unbiased estimator is constructed by using a linear combination of an estimator of study variable and mean per unit estimator of an auxiliary variable under simple random sampling without replacement scheme. The efficiency of the estimator under optimality compared with the mean per unit estimator, an almost unbiased ratio estimator, an unbiased product estimator, and a regression estimator both theoretically and with the numerical illustration.

© 2021 The Authors. Published by Atlantis Press B.V.
This is an open access article distributed under the CC BY-NC 4.0 license (http://creativecommons.org/licenses/by-nc/4.0/).

1. INTRODUCTION

In the survey sampling method, it is a common practice to utilize auxiliary information, which is frequently acknowledged to the higher precision of the estimators of population parameters. The classical ratio estimator, product estimator, and regression estimator are good examples in this context. When the correlation between the study variable \(y\) and auxiliary variable \(x\) is highly positively correlated, the ratio method of estimation is quite effective. Similarly when there is an existence of a high negative correlation between \(y\) and \(x\) then the product method of estimation is effectively used.

The property of unbiasedness is one of the important properties of an estimator. So it is desirable to construct unbiased estimators simultaneously keeping in mind its efficiency.

Tin [1] suggested an almost unbiased ratio estimator estimate the population means. Tin showed that this estimator is an almost unbiased and more efficient than the conventional ratio estimator suggested by Cochran [2]. Robson [3] suggested an estimator when there exists a negative correlation between study variable and auxiliary variable and is known as product estimator. Further, Robson constructed an unbiased estimator to estimate the population means by subtracting estimated bias from the product estimator.

In this paper, the unbiased estimator is suggested and its efficiency is compared with the mean per unit estimator, with an almost unbiased ratio estimator, suggested by Tin [1], unbiased product estimator suggested by Robson [3] and regression estimator by Watson [4].

2. UNBIASED ESTIMATORS

Let's consider a finite population \(U = (U_1, U_2, ..., U_N)\) size \(N\). Let \((y, x)\) be the study and auxiliary variables respectively. Now we consider \((y_i, x_i), i = 1, 2, 3, ..., n\). \(n\) denotes a sample of size “\(n\)” on the characteristics \(y\) and \(x\) have drawn from the population \(U\) with SRSWOR.

We denote \(\bar{Y}\) and \(\bar{X}\) are the population means of \(y\) and \(x\) respectively.

Let \(\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i\) and \(\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i\) are the sample mean of \(y\) and \(x\) respectively.

*Corresponding author. Email: bulu.mahanty@gmail.com
Further, we denote $s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$, $s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$, and $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$

Tin [1] suggested an almost unbiased ratio type estimator and it is given by

$$t_{Tin} = \frac{\bar{y} \bar{x}}{\bar{X}} \left[ 1 + \theta \left( \frac{s_{xy}}{s_y^2} - \frac{s_x^2}{s_x^2} \right) \right]$$

(1)

where, $\theta = \left( \frac{1}{n} - \frac{1}{N} \right)$

$$V(t_{Tin}) = \theta \bar{y}^2 \left[ C_y^2 + C_x^2 - 2C_{yx} \right] \text{ Considering up to } O \left( \frac{1}{n} \right).$$

(2)

Robson [3] suggested an unbiased product type estimator and it is given by

$$t_{Rob} = \frac{\bar{y}}{\bar{X}} - \theta \frac{s_{yx}}{s_x^2}$$

(3)

$$V(t_{Rob}) = \theta \bar{y}^2 \left[ C_y^2 + C_x^2 + 2C_{yx} \right] \text{ Considering up to } O \left( \frac{1}{n} \right).$$

(4)

Further, the regression estimator which is biased but in most cases, it is more precise is given by

$$t_r = \bar{y} + b (\bar{X} - \bar{x})$$

(5)

$$V(t_r) = \theta \bar{y}^2 C_y^2 (1 - \rho^2)$$

(6)

where, $b = \frac{s_{yx}}{s_y^2}$, the regression coefficient of $y$ on $x$.

### 2.1. Proposed Unbiased Estimator

Considering the linear combination of Robson’s an unbiased estimator and the mean per unit estimator of auxiliary variable $x$, we construct an unbiased estimator, given by

$$t_{MU} = \lambda_0 \left( \frac{\bar{y}}{\bar{X}} - \theta \frac{s_{yx}}{s_x^2} \right) + \lambda_1 \bar{x}$$

(7)

where, $\lambda_0$ and $\lambda_1$ are two suitable chosen constants such that estimator $t_{MU}$ is unbiased.

Hence, we put an unbiased condition

$$E(t_{MU}) = E \left[ \lambda_0 t_{Rob} + \lambda_1 \bar{x} \right] = \bar{y}$$

$$\Rightarrow \lambda_0 E(t_{Rob}) + \lambda_1 E(\bar{x}) = \bar{y}$$

$$\Rightarrow \lambda_0 \bar{y} + \lambda_1 \bar{x} = \bar{y}$$

$$\Rightarrow (\lambda_0 - 1) \bar{Y} + \lambda_1 \bar{X} = 0$$

(8)

Now, we derive the variance of the proposed unbiased estimator by considering the condition. The value of $\lambda_0$ and $\lambda_1$, which makes the estimator unbiased.
3. VARIANCE OF THE PROPOSED ESTIMATOR $t_{MU}$

The variance of the estimator $t_{MU}$ is given as

$$V(t_{MU}) = \left[ \lambda_0^2 V(t_{Rob}) + \lambda_1^2 V(\bar{x}) + 2\lambda_0\lambda_1 Cov(\bar{x}, t_{Rob}) \right]$$

$$= \lambda' S_2 \lambda$$

(9)

where, $\lambda' = [\lambda_0, \lambda_1]$, $S_2 = \begin{bmatrix} S_{00} & S_{01} \\ S_{10} & S_{11} \end{bmatrix}$, and $\lambda = [\lambda_0, \lambda_1]$

Further, $S_{00} = V(\bar{y}_{Rob})$, $S_{10} = S_{01} = Cov(\bar{x}, t_{Rob})$, and $S_{11} = V(\bar{x})$

$S_{00} = V(t_{Rob}) = 2C^2 + 2C^2_{xy}$, Considering up to $O(\frac{1}{n})$

$S_{10} = S_{01} = Cov(t_{Rob}, \bar{x}) = 2\bar{y} \bar{x} [C_{yx} + C^2_y]$, Considering up to $O(\frac{1}{n})$

$S_{11} = V(\bar{x}) = 2\bar{x}^2 C^2_y$, Considering up to $O(\frac{1}{n})$

The value of $\lambda_0$ and $\lambda_1$ which minimize the $V(t_{MU})$ given in (9) and subject to the condition of unbiasedness of $t_{MU}$ given in (7).

The optimum value of $\lambda$ is obtained as

$$\lambda = \frac{\bar{y} S^{-1}_2 Q_2}{Q' S^{-1}_2 Q_2}$$

(10)

where, $S^{-1}_2$ is the inverse of the matrix of $S_2$, i.e.,

$$S^{-1}_2 = \frac{1}{S_{00}S_{11} - S_{01}^2} \begin{bmatrix} S_{11} & -S_{10} \\ -S_{01} & S_{00} \end{bmatrix}$$

$$Q_2 = \begin{bmatrix} \bar{y} \\ \bar{x} \end{bmatrix}$$

where, $Q'_2$ is the transpose of $Q_2$ i.e.,

$$Q'_2 = \begin{bmatrix} \bar{y} \\ \bar{x} \end{bmatrix}$$

Hence, after the expansion

$$\lambda = \frac{\bar{y} \left( \bar{y} S_{11} - \bar{x} S_{01} \right)}{\bar{y}^2 S_{11} - 2\bar{y} \bar{x} S_{10} + \bar{x}^2 S_{00}}$$

(11)

From the Equation (11) we get the optimum values of $\lambda_0$ and $\lambda_1$

$$\lambda_0 = \frac{\bar{y} \left( \bar{y} S_{11} - \bar{x} S_{01} \right)}{\bar{y}^2 S_{11} - 2\bar{y} \bar{x} S_{10} + \bar{x}^2 S_{00}}$$

$$= R \frac{V(t_{Rob}) - RCov(\bar{x}, t_{Rob})}{V(t_{Rob}) + R^2 V(\bar{x}) - 2RCov(\bar{x}, t_{Rob})}$$

(12)

$$\lambda_1 = \frac{\bar{y} \left( \bar{x} S_{00} - \bar{y} S_{10} \right)}{\bar{y}^2 S_{11} - 2\bar{y} \bar{x} S_{10} + \bar{x}^2 S_{00}}$$

$$= R \frac{V(t_{Rob}) - RCov(\bar{x}, t_{Rob})}{V(t_{Rob}) + R^2 V(\bar{x}) - 2RCov(\bar{x}, t_{Rob})}$$

(13)

where, $R = \frac{\bar{y}}{\bar{x}}$
The optimum variance is obtained by substituting the optimum value of $\lambda_0$ and $\lambda_1$

\[
V(t_{MU})_{opt} = \frac{\bar{Y}^2}{Q' S_{y2}^{-1} Q} = \frac{\bar{Y}^2 (S_{00}S_{11} - S_{01}^2)}{\bar{Y}^2 S_{11} - 2\bar{Y} S_{10} + \bar{X}^2 S_{00}} = \theta^2 C^2 (1 - \rho^2)
\]

(14)

4. COMPARISON OF EFFICIENCY

i. Comparison of $t_{MU}$ under optimality with mean per unit estimator $\bar{Y}$.

The variance of mean per unit estimator is

\[
V(\bar{Y}) = \theta^2 C^2 P
\]

(15)

Comparing the variance of suggested unbiased estimator $t_{MU}$ under optimum value from the Equation (14) with the variance of Mean per unit estimator from Equation (15) we have,

\[
V(\bar{Y}) - V(t_{MU})_{opt} = \theta^2 C^2 C^2 - \theta^2 C^2 (1 - \rho^2) = \theta^2 [C^2 - C^2 (1 - \rho^2)]
\]

(16)

Hence, the proposed estimator $t_{MU}$ under optimality more efficient than the mean per unit estimator if

\[
C^2 > C^2 (1 - \rho^2).
\]

(17)

ii. Comparison of $t_{MU}$ under optimality with an almost unbiased estimator due to Tin $t_{Tin}$.

From the above Equations (2) and (14)

\[
MSE (t_{Tin}) - V(t_{MU}) = \theta^2 C^2 (C^2 + C^2 - 2\rho C^2 C^2) - \theta^2 C^2 (1 - \rho^2)
\]

\[
= \theta^2 C^2 (C^2 + \rho^2 C^2 - 2\rho C^2 C^2)
\]

\[
= \theta^2 C^2 (C^2 - \rho C^2) > 0
\]

(18)

Since, $(C^2 - \rho C^2)^2$ always positive the proposed estimator $t_{MU}$ is more efficient than the Tin estimator $t_{Tin}$.

iii. Comparison of $t_{MU}$ under optimality with an unbiased product estimator due to Robson $t_{Rob}$.

From the above Equations (4) and (14)

\[
MSE (t_{Rob}) - V(t_{MU}) = \theta^2 C^2 (C^2 + C^2 + 2\rho C^2 C^2) - \theta^2 C^2 (1 - \rho^2)
\]

\[
= \theta^2 C^2 (C^2 + \rho^2 C^2 + 2\rho C^2 C^2)
\]

\[
= \theta^2 C^2 (C^2 + \rho C^2) > 0
\]

(19)

Since, $(C^2 + \rho C^2)^2$ always positive the proposed estimator $t_{MU}$ is more efficient than the unbiased product estimator $t_{Rob}$.

iv. Comparison of $t_{MU}$ under optimality with regression estimator $t_r$

From the above Equations (6) and (14)

Now,

\[
MSE (t_r) - V(t_{MU}) = \theta^2 C^2 (1 - \rho^2) - \theta^2 C^2 (1 - \rho^2)
\]

\[
= \theta^2 C^2 (1 - \rho^2) [C^2 - C^2] > 0
\]

(20)

when, $C^2$ is greater than, $C^2$, $t_{MU}$ is more efficient than $t_r$.
5. NUMERICALLY ILLUSTRATIONS

To study the performance of the estimators, $\bar{y}$ (mean per unit estimator), $t_{\text{Tin}}$, $t_{\text{Rob}}$, $t_{\text{lr}}$, $t_{\text{MU}}$ numerically, we consider seven natural populations with positive correlation and seven natural population with negative correlation described in the Tables 1 and 2. The comparison is based on the computation of the variance of different estimators.

Remarks: Table 3, which shows that the variance of the proposed estimator $t_{\text{MU}}$ is the least when, $C_y^2$ is greater than $C_x^2$, followed by the regression estimator.

Remarks: Table 4, which shows that the variance of the estimator $t_{\text{MU}}$ is the least when, $C_y^2$ is greater than $C_x^2$, followed by the regression estimator.

Table 1 | Description of the population (Correlation coefficient is positive).

<table>
<thead>
<tr>
<th>Pop.ⁿ</th>
<th>Sources</th>
<th>Popⁿ size Y (N)</th>
<th>x</th>
<th>$C_y$</th>
<th>$C_x$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daniel and Cross [5], p. 474</td>
<td>25 Paired Serum</td>
<td>Dry Blood Spot Specimens</td>
<td>0.5177</td>
<td>0.4358</td>
<td>0.95</td>
</tr>
<tr>
<td>2</td>
<td>Gujarati [6], p. 598</td>
<td>10 Income (thousands of dollars)</td>
<td>No. of Families Owning a House</td>
<td>0.3096</td>
<td>0.1811</td>
<td>0.36</td>
</tr>
<tr>
<td>3</td>
<td>Gupta [7], p. 451</td>
<td>11 Fathers Height (cm)</td>
<td>Sons Height (cm)</td>
<td>0.0014</td>
<td>0.0015</td>
<td>0.55</td>
</tr>
<tr>
<td>4</td>
<td>Armitage and Berry [8], p. 161</td>
<td>17 Birth Weight</td>
<td>Increase in Weight</td>
<td>0.6076</td>
<td>0.2104</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>Sukhatme and Sukhatme [9], p. 166</td>
<td>20 No. of Banana Bunches</td>
<td>No. of Banana Pits</td>
<td>0.0605</td>
<td>0.0431</td>
<td>0.75</td>
</tr>
<tr>
<td>6</td>
<td>Daniel and Cross [5], pp. 455–456</td>
<td>20 Age (years)</td>
<td>Bilirubin Levels (mg/dl)</td>
<td>0.4518</td>
<td>0.0483</td>
<td>0.46</td>
</tr>
<tr>
<td>7</td>
<td>Cochran [10], p. 186</td>
<td>21 Number of Family Members</td>
<td>Number of Cars</td>
<td>0.5231</td>
<td>0.1156</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 2 | Description of the population (Correlation coefficient is negative).

<table>
<thead>
<tr>
<th>Pop.ⁿ</th>
<th>Sources</th>
<th>Popⁿ size Y (N)</th>
<th>x</th>
<th>$C_y$</th>
<th>$C_x$</th>
<th>$\rho$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Singh and Mangat [11], p. 71</td>
<td>10 Number of Tube wells</td>
<td>Samples of Villages</td>
<td>0.7412</td>
<td>0.3981</td>
<td>−0.19</td>
</tr>
<tr>
<td>2</td>
<td>Singh and Mangat [11], p. 227</td>
<td>36 Samples Student OGPA</td>
<td>Number of hours per Week Devoted to TV Viewing</td>
<td>0.2852</td>
<td>0.0190</td>
<td>−0.58</td>
</tr>
<tr>
<td>3</td>
<td>Singh and Mangat [11], p. 195</td>
<td>37 Collected in respect of OGPA</td>
<td>Said nonacademic Activities</td>
<td>0.1648</td>
<td>0.0443</td>
<td>−0.36</td>
</tr>
<tr>
<td>4</td>
<td>Armitage and Berry [8], p. 161</td>
<td>32 Increase in Weight after 70–100 days</td>
<td>Birth Weight,</td>
<td>0.1069</td>
<td>0.0256</td>
<td>−0.68</td>
</tr>
<tr>
<td>5</td>
<td>Maddala [12], p. 316</td>
<td>16 Veal (Consumption per capita LB)</td>
<td>Veal (Price per pound)</td>
<td>0.0519</td>
<td>0.0097</td>
<td>−0.68</td>
</tr>
<tr>
<td>6</td>
<td>Maddala [12], p. 316</td>
<td>16 Lamb (Consumption per capita LB)</td>
<td>Lamb (Price per pound)</td>
<td>0.011</td>
<td>0.0105</td>
<td>−0.75</td>
</tr>
<tr>
<td>7</td>
<td>Gujarati [6], p. 598</td>
<td>10 Hypothetical data on Income 1000$</td>
<td>No. of Families at Income</td>
<td>0.344</td>
<td>0.1423</td>
<td>−0.43</td>
</tr>
</tbody>
</table>

Table 3 | MSE of different estimators (sample size n = 4) (Correlation coefficient is positive).

<table>
<thead>
<tr>
<th>Popⁿ No.</th>
<th>$\bar{y}$</th>
<th>$t_{\text{Tin}}$</th>
<th>$t_{\text{Rob}}$</th>
<th>$t_{\text{lr}}$</th>
<th>$t_{\text{MU}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>137.4757</td>
<td>13.5067</td>
<td>492.9409</td>
<td>13.3607</td>
<td>11.2491</td>
</tr>
<tr>
<td>2</td>
<td>18.954</td>
<td>19.5847</td>
<td>40.4995</td>
<td>16.4883</td>
<td>9.6457</td>
</tr>
<tr>
<td>3</td>
<td>0.9545</td>
<td>0.8289</td>
<td>2.9216</td>
<td>0.6572</td>
<td>0.6340</td>
</tr>
<tr>
<td>4</td>
<td>3.2559</td>
<td>1.9426</td>
<td>6.8246</td>
<td>1.9350</td>
<td>0.6701</td>
</tr>
<tr>
<td>5</td>
<td>9842.1</td>
<td>4470.5</td>
<td>4438.3</td>
<td>6232.4</td>
<td>3128.54</td>
</tr>
<tr>
<td>6</td>
<td>4.4792</td>
<td>3.5936</td>
<td>6.3234</td>
<td>3.5075</td>
<td>0.3753</td>
</tr>
<tr>
<td>7</td>
<td>0.3111</td>
<td>0.0949</td>
<td>0.6648</td>
<td>0.0158</td>
<td>0.0035</td>
</tr>
</tbody>
</table>

Table 4 | MSE of different estimators (sample size n = 4) (Correlation coefficient is negative).

<table>
<thead>
<tr>
<th>Popⁿ No.</th>
<th>$\bar{y}$</th>
<th>$t_{\text{Tin}}$</th>
<th>$t_{\text{Rob}}$</th>
<th>$t_{\text{lr}}$</th>
<th>$t_{\text{MU}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2800.20</td>
<td>5087.90</td>
<td>3521.06</td>
<td>2698.20</td>
<td>1449.48</td>
</tr>
<tr>
<td>2</td>
<td>2.9360</td>
<td>4.0209</td>
<td>2.2425</td>
<td>1.9258</td>
<td>0.1283</td>
</tr>
<tr>
<td>3</td>
<td>8.0612</td>
<td>13.2546</td>
<td>7.2073</td>
<td>7.0078</td>
<td>1.8862</td>
</tr>
<tr>
<td>4</td>
<td>115.5169</td>
<td>221.0861</td>
<td>65.4413</td>
<td>60.9492</td>
<td>14.6398</td>
</tr>
<tr>
<td>5</td>
<td>0.568</td>
<td>0.3396</td>
<td>0.3412</td>
<td>0.7866</td>
<td>0.0568</td>
</tr>
<tr>
<td>6</td>
<td>10.657</td>
<td>5.4721</td>
<td>5.4356</td>
<td>8.3346</td>
<td>4.8318</td>
</tr>
<tr>
<td>7</td>
<td>18.954</td>
<td>37.359</td>
<td>16.2325</td>
<td>15.3967</td>
<td>6.3700</td>
</tr>
</tbody>
</table>
6. CONCLUSION

In this paper, an almost unbiased estimator $t_{MU}$ perform better than the estimators mean per unit estimator ($\bar{y}$), Tin estimator ($t_{Tin}$), Robson estimator ($t_{Rob}$), and regression estimator ($t_{lr}$) both theoretically and numerically when the coefficient variation of $y$ is greater than the coefficient variation of $x$.

CONFLICTS OF INTEREST

We the author(s) declare(s) that there is no conflict of interest.

AUTHORS’ CONTRIBUTIONS

B. Mahanty developed the theoretical formalism, performed the analytic calculations and performed the numerical illustrations. Both B. Mahanty and G. Mishra. authors contributed to the final version of the manuscript. G.Mishra. supervised the manuscript.

Funding Statement

The author(s) received no financial support to carry out this research in the manuscript.

ACKNOWLEDGMENTS

The authors are indebted to two anonymous referees and the Editor for their productive comments and suggestions, which led to improve the presentation of this manuscript.

REFERENCES