Dual to Ratio cum Product Estimators of Finite Population Mean Using Auxiliary Attribute(s) in Stratified Random Sampling

Sachin Malik and Rajesh Singh

Department of Statistics, Banaras Hindu University, Varanasi-221005, India

Abstract: In this paper some new improved estimators are suggested using auxiliary attribute(s) in stratified random sampling. The expressions of the mean square error of the proposed estimators are derived. Also, the theoretical findings are supported by a numerical example.

Key words: Auxiliary information · Attributes · Stratified sampling · Bias · Mean squared error

INTRODUCTION

The problem of estimating the population mean in the presence of an auxiliary variable has been widely discussed in finite population sampling literature [1], suggested a class of estimators of the population mean using one auxiliary variable in the stratified random sampling and examined the MSE of the estimators up to the kth order of approximation [2-8] suggested some ratio cum product estimators in simple random sampling [10, 11] suggested some exponential ratio type estimators. However in many practical situations instead of existence of auxiliary variables there exist some auxiliary attributes (say), which are highly correlated with the study variable y. Taking into consideration the point biserial correlation coefficient between auxiliary attribute \( \phi \) and the study variable y, several authors including [12-17] envisaged large number of improved estimators for the population mean \( \bar{y} \) of the study variable y. There are some situations when in place of one auxiliary attribute, we have information on two qualitative variables. For illustration, to estimate the hourly wages we can use the information on marital status and region of residence [18]. Here we assume that both auxiliary attributes have significant point bi-serial correlation with the study variable and there is significant phi-correlation [19] between the two auxiliary attributes. Using point biserial correlation and phi-correlation [20-24] have proposed improved estimators for the population mean \( \bar{y} \).

In planning surveys, stratified sampling has often proved as useful in improving the precision of un-stratified sampling strategies to estimate the finite population mean of the study variable, \( \bar{y} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} \).

Assume that the population U consist of L strata as U=U_1, U_2, ..., U_L. Here the size of the stratum U_h is N_h and the size of simple random sample in stratum U_h is n_h where h=1, 2, ..., L. In this study, under stratified random sampling without replacement scheme, we suggest estimators to estimate \( \bar{y} \) by considering the estimators in [25] and in [26]. Consider a finite population of size N and is divided into L strata such that \( \sum_{h=1}^{L} N_h = N \) where N_h is the size of the \( h^{th} \) stratum (h=1,2,...,L). We select a sample of size n_h from each stratum by simple random sample without replacement sampling such that \( \sum_{h=1}^{L} n_h = n \), where n_h is the stratum sample size. Let \( (y_{hi}, \Psi_{i}) \) (j=1,2) denote the values of the study variable (y) and the auxiliary attributes \( \Psi_j \), (j=1,2) respectively in the \( h^{th} \) stratum.

Suppose \( \Psi_{hi} = 1 \), if \( i^{th} \) unit in the stratum h possesses the attribute \( \Psi_j \);
\( \Psi_{hi} = 0 \), otherwise.
Let \( p_j = \sum_{h=1}^{L} w_h p_{jh} \) and \( p_{jst} = \sum_{h=1}^{L} w_h p_{jth} \) (\( j=1, 2 \)) be the population and sample proportions of units of the auxiliary attributes where \( p_{jh} = \frac{A_{jh}}{N_h} \), \( p_{jth} = \frac{a_{jth}}{n_h} \) and \( A_{jh} = \sum_{i=1}^{N_h} \psi_{jhi} \) and \( a_{jth} = \sum_{i=1}^{n_h} \psi_{jthi} \).

To estimate \( \bar{y} = \sum_{h=1}^{L} w_h y_h/N_h \), we assume that \( p_{jh} \) (\( j=1, 2 \)) are known.

To obtain the bias and MSE of the proposed estimators, we use the following notations in the rest of the article:

\[
\bar{y}_{st} = \sum_{h=1}^{L} w_h y_h = \bar{y} (1 + e_0)
\]

\[
p_{1st} = \sum_{h=1}^{L} w_h p_{1th} = \bar{p}_1 (1 + e_1)
\]

\[
p_{2st} = \sum_{h=1}^{L} w_h p_{2th} = \bar{p}_2 (1 + e_2)
\]

such that,

\[
E(e_0) = E(e_1) = E(e_2) = 0.
\]

Let

\[
S^2_{y_{1h}} = \sum_{i=1}^{N_h} \left( y_{hi} - \bar{y}_h \right)^2/N_h - 1, S^2_{\psi_{j}h} = \sum_{i=1}^{N_h} \left( \psi_{jhi} - p_{jh} \right)^2/N_h - 1
\]

be the population variances of the study variable and the auxiliary attributes in the \( h \)th stratum, where,

\[
\bar{y}_h = \sum_{i=1}^{N_h} y_{hi}/N_h, S_{\psi_{j}h} = \sum_{i=1}^{N_h} \left( \psi_{jhi} - \bar{y}_h \right) \left( \psi_{jhi} - p_{jh} \right) / N_h - 1
\]

\[
\rho_{\psi_{j}h} = \frac{S_{\psi_{j}h}}{S_{\psi_{j}h} S_{\psi_{j}h}}
\]

and the MSE of \( t_1 \) and \( t_2 \) to the first degree of approximation are, respectively, given by

\[
\text{MSE}(t_1) = \bar{y}^2 \left[ V_1 + V_2 - 2V_4 \right]
\]
The approximate MSE of this estimator is

$$\text{MSE}(t_7) = \sum \left[ \left( 1 + \epsilon_0 \right) \left( \epsilon_1 - 1 \right) \left( 1 + \epsilon_2 \right) \right]$$

**Suggested Estimators:** Motivated by [26], we propose the following product estimator for the stratified random sampling scheme as

$$t_8 = \bar{y}_{st} \left( \frac{u}{U} \frac{P_{2st}}{P_2} \right)$$

Expressing $t_8$ in terms of $\epsilon$'s, we can write (2.1) as

$$t_8 = \bar{y}(1 + \epsilon_0)(1 - \epsilon_1)(1 + \epsilon_2)$$

The MSE($t_8$) to the first order of approximation, is given as

$$\text{MSE}(t_8) = \sum \left[ \frac{\lambda^2 V_2 + V_3 + 2(V_5 - V_6)}{V_2} \right]$$

And this MSE equation is minimized for

$$\lambda = \frac{V_4 + V_6}{V_2}$$

Note that the corresponding $A$ is

$$A_{opt} = \frac{(1 - \lambda_{opt})P_1}{\lambda_{opt}}$$

By putting the optimum value of $\lambda$, in (12), we can obtain the minimum MSE equation for the first proposed estimator $t_8$.

Following [7], we proposed a family of dual to ratio cum estimators as

$$t_9 = \bar{y}_{st} \left( \frac{P_{1st}}{P_1} \right) \left( \frac{P_{2st}}{P_2} \right)$$

where,

$$P_{1st} = (1 + g_h P_{1h}) - g_h P_{1h}, \quad P_{2st} = (1 + g_h P_{2h}) - g_h P_{2h}$$

In that case, following [9] we use $t_{p1} = p_{1st} + \alpha_1(P_1 - P_{1st})$ and $t_{p2} = p_{2st} + \alpha_2(P_2 - P_{2st})$ instead of $p_{1st}$ and $p_{2st}$ respectively. Where, $\alpha_1$ and $\alpha_2$ are constants that makes the MSE minimum.

Adapting $t_6$ to the stratified random sampling, the ratio cum product estimator using two auxiliary attributes can be defined as
To obtain the MSE of the estimator $t_\psi$, we can write

$$\overline{\psi}_{st} = \overline{Y}(1 + e_0)$$

$$p_{1st}^* = P_1(1 + e_1^*)$$

$$p_{2st}^* = P_2(1 + e_2^*)$$

Expressing (2.5) in terms of $e$'s, we have

$$t_0 = \overline{Y}(1 + e_0)(1 + e_1^*)^0(1 + e_2^*)^{-0_2}$$

(2.6)

Expanding the right hand side of (2.6), to the first order of approximation, we get

$$t_0 = \overline{Y}\left[\psi_0 + \theta_1 e_0 e_1^* - \theta_2 e_0 e_2^* + \theta_1 e_1^* - \theta_2 e_1 e_2^* - \theta_2 e_1 e_2^* + \frac{\theta_1 (\theta_1 - 1)}{2} e_1^* e_2^* + \frac{\theta_2 (\theta_2 + 1)}{2} e_2^* e_2^*\right]$$

Squaring both sides of (16) and then taking expectation, we obtain the MSE of the second proposed estimator, $t_\psi$ to the first order approximation, as

$$\text{MSE}(t_0) = \overline{Y}^2\left[\psi_1^* + \theta_1^2 \psi_2^* + \theta_2^2 \psi_3^* + 2(\theta_1 \psi_4^* - \theta_2 \psi_5^* - \theta_2 \psi_6^*)\right]$$

(2.7)

This MSE equation is minimized for the optimum values of $\theta_1$ and $\theta_2$ given by

$$\theta_1 = \frac{\psi_2^* \psi_4^* - \psi_1^* \psi_5^*}{\psi_2^* \psi_3^* - \psi_6^*} \quad \text{and} \quad \theta_2 = \frac{\psi_2^* \psi_5^* - \psi_4^* \psi_6^*}{\psi_1^* \psi_3^* - \psi_6^*}$$

Putting these values of $\theta_1$ and $\theta_2$ in $\text{MSE}(t_0)$, given in (2.7), we can obtain the minimum MSE of the second proposed estimator, $(t_\psi)$.

**Theoretical Efficiency Comparisons:** In this section, we first compare the efficiency between the first proposed estimator $t_\psi$, with and the classical combined estimator, $t_\psi$, as follows:

$$\text{MSE}(t_\psi) < \overline{V}\left[\overline{\psi}_{st}\right]$$

$$\overline{Y}^2\left[\psi_1 + \lambda^2 \psi_2 + \psi_3 - 2(\lambda \psi_4 - \psi_5 + \lambda \psi_6)\right] < \overline{Y}^2 \psi_1$$

$$\frac{A_1}{2A_2} < 1$$

(3.1)

Where, $A_1 = \lambda^2 \psi_2 + \psi_1$ and $A_2 = \lambda \psi_4 - \psi_5 + \lambda \psi_6$.

If the condition (3.1) is satisfied, the first proposed estimator $t_\psi$ performs better than the classical combined estimator.
We also find the condition under which the second proposed estimator, \( t_n \), performs better than the classical combined estimator in theory as follows:

\[
\text{MSE}(t_n) < \sqrt{\text{var}(\bar{y}_{\text{st}})}
\]

\[
\bar{Y}^2 \left[ \left( \theta_1 V_1^* + \theta_2 V_2^* + \theta_3 V_3^* + 2(\theta_4 V_4^* - \theta_5 V_5^* - \theta_6 V_6^*) \right) \right] < \bar{Y}^2 V_1
\]

\[
\frac{A_3}{A_4} < 1 \quad (3.2)
\]

where, \( A_3 = \theta_1^2 V_2^* + \theta_2^2 V_3^* - 2 \theta_1 \theta_2 V_6^* \) and \( A_4 = \gamma_2 V_5^* - \gamma_1 V_4^* \)

**Empirical Study:** Source: [See (28)]

We randomly select a sample of size \( n_i \) from each stratum by using the Neyman allocation and consider the first 10%, 20% and 30% values in each stratum as non-response for \( W \leq 0.1 \), \( W = 0.2 \) and \( W = 0.3 \) respectively.

The population consists of village wise complete enumeration and data are obtained in 1951 and 1961 censuses for a Tehsil. The area of village is used to stratify the population into three strata.

Let \( y \) be the cultivated area in the village in hectares in 1951 and \( \Psi_i = 0, 1 \) are given below

Let

\( \Psi_{1i} = 1 \), if \( i^{th} \) village in the stratum 1 has area greater than 550 hectares =0, otherwise.

\( \Psi_{2i} = 1 \), if \( i^{th} \) village in the stratum 2 has area greater than 1300 hectares =0, otherwise and

\( \Psi_{3i} = 1 \), if \( i^{th} \) village in the stratum 3 has area greater than 2500 hectares =0, otherwise.

Let

\( \Psi_{1i} = 1 \), if \( i^{th} \) village in the stratum 1 has number of cultivators greater than 550 =0, otherwise.

\( \Psi_{2i} = 1 \), if \( i^{th} \) village in the stratum 2 has number of cultivators greater than 700 =0, otherwise and

\( \Psi_{3i} = 1 \), if \( i^{th} \) village in the stratum 3 has number of cultivators greater than 1500 =0, otherwise.

Data is presented in Table 4.1 and results are given in Table 4.2.

<table>
<thead>
<tr>
<th>Stratum no.</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( N_i )</td>
<td>43</td>
<td>45</td>
<td>40</td>
</tr>
<tr>
<td>( n_i )</td>
<td>10</td>
<td>12</td>
<td>18</td>
</tr>
<tr>
<td>( \bar{Y}_{hi} )</td>
<td>397.1425</td>
<td>760.1795</td>
<td>1234.6180</td>
</tr>
<tr>
<td>( P_{hi} )</td>
<td>0.5814</td>
<td>0.4444</td>
<td>0.4000</td>
</tr>
<tr>
<td>( P_{hi} )</td>
<td>0.39535</td>
<td>0.5333</td>
<td>0.4000</td>
</tr>
<tr>
<td>( \psi_i )</td>
<td>39975.0569</td>
<td>61455.990</td>
<td>172425.900</td>
</tr>
<tr>
<td>( \psi_{ihi} )</td>
<td>0.2492</td>
<td>0.2525</td>
<td>0.2462</td>
</tr>
<tr>
<td>( \psi_{ihi} )</td>
<td>0.2448</td>
<td>0.2545</td>
<td>0.2462</td>
</tr>
<tr>
<td>( \rho_{ihi} )</td>
<td>0.6922</td>
<td>0.3750</td>
<td>0.5057</td>
</tr>
<tr>
<td>( \rho_{ihi} )</td>
<td>0.3956</td>
<td>0.2847</td>
<td>0.3261</td>
</tr>
<tr>
<td>( \rho_{ihi} )</td>
<td>0.2040</td>
<td>0.3884</td>
<td>0.5833</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Estimators</th>
<th>( \psi_{ihi} )</th>
<th>( \psi_{ihi} )</th>
<th>( \psi_{ihi} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \psi_{ihi} )</td>
<td>100</td>
<td>309.854</td>
<td>15.125</td>
</tr>
<tr>
<td>( \psi_{ihi} )</td>
<td>15.125</td>
<td>125.232</td>
<td>31.578</td>
</tr>
<tr>
<td>( \psi_{ihi} )</td>
<td>31.578</td>
<td>58.694</td>
<td>26.482</td>
</tr>
<tr>
<td>( \psi_{ihi} )</td>
<td>26.482</td>
<td>71.753</td>
<td>428.773</td>
</tr>
</tbody>
</table>

We compute percent relative efficiencies as

\[
\text{PRE} = \frac{\text{var}(\bar{y}_{\text{st}})}{\text{MSE}(\bar{y}_{i})} \times 100, \quad (i=1, 2, 3)
\]

**CONCLUSION**

Superiority of the proposed estimator is established theoretically by the conditions derived in section 3. Results in Table 4.2 confirm this superiority numerically using the previously used data set.

**REFERENCES**