ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 39 (2019*)*, 131-141

An Efficient Class of Double Sampling Estimators for the Population Mean Using Auxiliary Information on First and Second Moments about Zero

Tarunpreet Kaur Ahuja, Peeyush Misra*, V. S. Singh and R. Karan Singh

[Received on May, 2019. Accepted on October, 2019]

ABSTRACT

This paper focuses on the estimation problem of population mean by using auxiliary information. By using auxiliary information on first and second moments about zero, a new efficient generalized class of estimators for estimating the finite population mean under two-phase or double sampling scheme have been proposed. The properties of the suggested class of estimators in terms of bias and mean squared error have been studied. The efficiency comparisons have been carried out with respect to the sample mean estimator to establish its superiority. An empirical study is also included as an illustration to justify the results through data.

1. Introduction

The use of auxiliary information in sample surveys is proved to be of immense importance and therefore statisticians very often make use of the information available on an auxiliary variable with the variable under study for improving the efficiency of an estimator. For better understanding one may see Cochran (1977), Murthy (1967) and Sukhatme et al. (1984).

It is well known fact that the auxiliary information in sample surveys results in substantial improvement in the precision of the estimators of the population parameters, and if, the parameters of the auxiliary variables are not known in advance, double or two-phase sampling technique is used. The prime advantage

Correspondence author:* Peeyush Misra, Department of Statistics, DAV (PG) College, Dehradun, Uttarakhand, India- 248001. dr.pmisra.dav@gmail.com, Tarunpreet Kaur Ahuja, Department of Statistics, D.A.V. (P.G.) College, Dehradun, Uttarakhand, V.S. Singh, Department of Statistics, HNB Garhwal University, SRT Campus, Badshahithaul, New Tehri, Uttarakhand, R.Karan Singh, Department of Statistics, University of Lucknow, Lucknow, Uttar Pradesh.

of double or two-phase sampling scheme is that it is more flexible and considerably cost effective. In double sampling or two-phase sampling technique, we first take a preliminary large sample of size n' (called first phase sample) from a population of size N and then a sub-sample of size n (called second phase sample) is drawn from the first phase sample of size n' using simple random sampling without replacement at both the phases. At first phase sample of size n', only the auxiliary variable X be observed but at the second phase sample of size n, the study variable Y and the auxiliary variable X both are observed.

Let $\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ be the population mean of study variable y and

$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the population mean of auxiliary variable x .

$\sigma_Y^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^2$, $\sigma_X^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2$ and

$\rho = \frac{\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sigma_Y \sigma_X}$ be the population correlation coefficient between y and x .

Also let $\mu_{rs} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \overline{Y}\right)^r \left(X_i - \overline{X}\right)^s$, $C_Y^2 = \frac{\sigma_Y^2}{\overline{Y}^2}$, $C_X^2 = \frac{\sigma_X^2}{\overline{X}^2} = \frac{\mu_{02}}{\overline{X}^2}$,

$\rho = \frac{\mu_{11}}{\sigma_Y \sigma_X}$, $\lambda = \frac{\mu_{12}}{\overline{Y}\sigma_X^2} = \frac{\mu_{12}}{\overline{Y}\mu_{02}}$, $\beta_2 = \frac{\mu_{04}}{\mu_{02}^2}$, $\beta_1 = \frac{\mu_{03}^2}{\mu_{02}^3}$ and $\gamma_1 = \sqrt{\beta_1}$.

Let the first phase sample of size n' be $\left(x_1', x_2', \ldots, x_n'\right)$ on x and the second phase sample of size n be $\left\{\left(y_1, x_1\right), \left(y_2, x_2\right), \ldots, \left(y_n, x_n\right)\right\}$ on variables $\left(y, x\right)$ with the first phase sample mean $\bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x'_i$ estimator of population mean $\overline{X}$ and the second phase sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ respectively on y and x .

With the first two moments about zero $\bar{x}' = \frac{1}{n'}\sum_{i=1}^{n'} x'_i$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{\theta}_x{}' = \frac{1}{n'}\sum_{i=1}^{n'} x_i'^2$, $\bar{\theta}_x = \frac{1}{n}\sum_{i=1}^{n} x_i{}^2$ of auxiliary variable x, the proposed generalized class of double sampling estimators of the population mean is $\bar{y}_g = \bar{y}\, g\left(\frac{\bar{x}}{\bar{x}'}, \frac{\bar{\theta}_x}{\bar{\theta}_x{}'}\right) = \bar{y}\, g(u_1,\ u_2)$, $g(u_1,\ u_2)$ satisfying the validity conditions of Taylors series expansion is a bounded function of $t = (u_1,\ u_2)$ such that at the point $T = (1,\ 1)$ we have

(i) $g(t = T) = 1$ (1.2)

(ii) The first order partial derivatives are

$$g_1 = \left(\frac{\partial g(u_1,\ u_2)}{\partial u_1}\right)_T \text{ and } g_2 = \left(\frac{\partial g(u_1,\ u_2)}{\partial u_2}\right)_T \tag{1.3}$$

(iii) The second order partial derivatives are

$$g_{11} = \left(\frac{\partial^2 g(u_1,\ u_2)}{\partial u_1^2}\right)_T ,\ g_{22} = \left(\frac{\partial^2 g(u_1,\ u_2)}{\partial u_2^2}\right)_T \text{ and } g_{12} = \left(\frac{\partial^2 g(u_1,\ u_2)}{\partial u_1 \partial u_2}\right)_T \tag{1.4}$$

2. Bias and Mean Squared Error of the Proposed Estimator

In order to obtain bias and mean square error of the proposed estimator, let us denote by

$$\bar{y} = \bar{Y}(1 + e_0)$$

$$\bar{x} = \bar{X}(1 + e_1)$$

$$\bar{x}' = \bar{X}(1 + e_1')$$

$$\bar{\theta}_x = \bar{\theta}_X(1 + e_2)$$

$$\bar{\theta}_x' = \bar{\theta}_X(1 + e_2') \tag{2.1}$$

So that ignoring finite population correction for simplicity we have

$$E(e_0) = E(e_1) = E(e_1') = E(e_2) = E(e_2') = 0 \qquad (2.2)$$

$$E(e_0^2) = \frac{1}{n} C_Y^2$$

$$E(e_1^2) = \frac{1}{n} C_X^2$$

$$E(e_1'^2) = \frac{1}{n'} C_X^2$$

$$E(e_2^2) = \frac{1}{n\bar{\theta}_X^2}\left(\mu_{04} + 4\bar{X}\mu_{03} + 4\bar{X}^2\mu_{02} - \mu_{02}^2\right)$$

$$E(e_2'^2) = \frac{1}{n'\bar{\theta}_X^2}\left(\mu_{04} + 4\bar{X}\mu_{03} + 4\bar{X}^2\mu_{02} - \mu_{02}^2\right)$$

$$E(e_0 e_1) = \frac{1}{n} \rho C_Y C_X$$

$$E(e_0 e_1') = \frac{1}{n'} \rho C_Y C_X$$

$$E(e_0 e_2) = \frac{1}{n\bar{Y}\bar{\theta}_X}\left(\mu_{12} + 2\bar{X}\mu_{11}\right)$$

$$E(e_0 e_2') = \frac{1}{n'\bar{Y}\bar{\theta}_X}\left(\mu_{12} + 2\bar{X}\mu_{11}\right)$$

$$E(e_1 e_1') = \frac{1}{n'} C_X^2$$

$$E(e_1 e_2) = \frac{1}{n\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)$$

$$E(e_1e_2') = \frac{1}{n'\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)$$

$$E(e_1'e_2) = \frac{1}{n'\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)$$

$$E(e_1'e_2') = \frac{1}{n'\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)$$

$$E(e_2e_2') = \frac{1}{n'\bar{\theta}_X^2}\left(\mu_{04} + 4\bar{X}\mu_{03} + 4\bar{X}^2\mu_{02} - \mu_{02}^2\right) \tag{2.3}$$

Expanding $g(u_1,\ u_2)$ about the point $T = (1,\ 1)$ in the third order Taylor's series, we have to the first degree of approximation

$$\bar{y}_g = \bar{y}\Big[g(1,1) + (u_1 - 1)g_1 + (u_2 - 1)g_2 + \frac{1}{2!}\left\{(u_1 - 1)^2 g_{11} + (u_2 - 1)^2 g_{22}\right\}$$

$$+ 2(u_1 - 1)(u_2 - 1)g_{12}\Big] \tag{2.4}$$

where $g_1,\ g_2,\ g_{12},\ g_{11},\ g_{22}$ are already defined and $u_1^* = 1 + h(u_1 - 1)$, $u_2^* = 1 + h(u_2 - 1)$ for $0 < h < 1$.

In terms of e_i's, $i = 0,1,2$; $\bar{y}_g$ from equation (2.4) to the first degree of approximation may be written as

$$\bar{y}_g - \bar{Y} = \bar{Y}\left(e_0 + e_1g_1 - e_1'g_1 + e_2g_2 - e_2'g_2\right) + \bar{Y}\Big\{\left(e_0e_1 - e_0e_1' + e_1'^2 - e_1e_2'\right)g_1$$

$$+ \left(e_0e_2 - e_0e_2' + e_2'^2 - e_2e_2'\right)g_2\Big\} + \frac{\bar{Y}}{2!}\Big\{\left(e_1^2 - e_1'^2 - 2e_1e_1'\right)g_{11}$$

$$+ \left(e_2^2 - e_2'^2 - 2e_2e_2'\right)g_{22} + 2\left(e_1e_2 - e_1e_2' - e_1'e_2 + e_1'e_2'\right)g_{12}\Big\} \tag{2.5}$$

Taking expectation on both the sides of equation (2.5), the bias of $\bar{y}_g$ up to terms of order $O(1/n)$ is given by

Bias $\left(\bar{y}_g\right)$

$$= \left\{E\left(\bar{y}_g\right) - \bar{Y}\right\} = \bar{Y}\left(\frac{1}{n} - \frac{1}{n'}\right)\left[\rho C_Y C_X g_1 + \frac{1}{\bar{Y}\bar{\theta}_X}\left(\mu_{12} + 2\bar{X}\mu_{11}\right)g_2 + \frac{1}{2}\left\{C_X^2 g_{11}\right.\right.$$

$$\left.\left. + \frac{1}{\bar{\theta}_X^2}\left(\mu_{04} + 4\bar{X}\mu_{03} + 4\bar{X}^2\mu_{02} - \mu_{02}^2\right)g_{22} + \frac{2}{\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)g_{12}\right\}\right] \tag{2.6}$$

Now squaring both sides of equation (2.5) and then taking expectation, the mean squared error of $\bar{y}_g$ up to terms of order $O(1/n)$ is given by

$$\text{MSE}\left(\bar{y}_g\right) = \left\{E\left(\bar{y}_g\right) - \bar{Y}\right\}^2$$

$$= \bar{Y}^2 E\left(e_0^2\right) + \bar{Y}^2 g_1^2\left\{E\left(e_1^2\right) + E\left(e_1'^2\right) - 2E\left(e_1 e_1'\right)\right\}$$

$$+ \bar{Y}^2 g_2^2\left\{E\left(e_2^2\right) + E\left(e_2'^2\right) - 2E\left(e_2 e_2'\right)\right\}$$

$$+ 2\bar{Y}^2 g_1\left\{E\left(e_0 e_1\right) - E\left(e_0 e_1'\right)\right\} + 2\bar{Y}^2 g_2\left\{E\left(e_0 e_2\right) - E\left(e_1 e_2'\right)\right\}$$

$$+ 2\bar{Y}^2 g_1 g_2\left\{E\left(e_1 e_2\right) - E\left(e_1 e_2'\right) - E\left(e_1' e_2\right) + E\left(e_1' e_2'\right)\right\}$$

using values of the expectations given in equation (2.2) and equation (2.3), the above expression reduces to

$\text{MSE}\left(\bar{y}_g\right)$

$$= \frac{\bar{Y}^2}{n}C_Y^2 + \bar{Y}^2\left(\frac{1}{n} - \frac{1}{n'}\right)\left\{C_X^2 g_1^2 + \frac{1}{\bar{\theta}_X^2}\left(\mu_{04} + 4\bar{X}\mu_{03} + 4\bar{X}^2\mu_{02} - \mu_{02}^2\right)g_2^2\right.$$

$$\left. + 2g_1\rho C_Y C_X + 2g_2\frac{1}{\bar{Y}\bar{\theta}_X}\left(\mu_{12} + 2\bar{X}\mu_{11}\right) + 2g_1 g_2\frac{1}{\bar{X}\bar{\theta}_X}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)\right\} \tag{2.7}$$

which attains the minimum for the optimum values

$$g_1 = -\frac{1}{C_X^2}\left\{\rho C_Y C_X - \frac{1}{\bar{Y}\Delta}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)\left(\delta_1 - \delta_2\right)\right\} \tag{2.8}$$

$$g_2 = -\frac{1}{\overline{Y}\Delta}\overline{X}\overline{\theta}_X(\delta_1 - \delta_2) \tag{2.9}$$

where $\delta_1 = C_X^2\overline{X}(\mu_{12} + 2\overline{X}\mu_{11})$, $\delta_2 = \rho C_Y C_X \overline{Y}(\mu_{03} + 2\overline{X}\mu_{02})$ and

$$\Delta = C_X^2\overline{X}^2(\mu_{04} + 4\overline{X}\mu_{03} + 4\overline{X}^2\mu_{02} - \mu_{02}^2) - (\mu_{03} + 2\overline{X}\mu_{02})^2$$

$$= \mu_{02}^3(\beta_2 - \beta_1 - 1) > 0$$

Substituting the values of g_1 and g_2 given by equation (2.8) and equation (2.9) in equation (2.7), the minimum mean squared error of $\bar{y}_g$ is given by

$$\text{MSE}(\bar{y}_g)_{\min} = \frac{\overline{Y}^2}{n}C_Y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\left\{\rho^2\overline{Y}^2C_Y^2 + \frac{1}{\Delta C_X^2}(\delta_1 - \delta_2)^2\right\} \tag{2.10}$$

3. Efficiency Comparison

The general estimator of Mean in case of SRSWOR is $\hat{\bar{y}}_{wor} = \bar{y}$ with

$$MSE(\hat{\bar{y}}) = \frac{\mu_{20}}{n} \tag{3.1}$$

It is clear from equations (2.10) and (3.1) that the proposed generalized class of estimators is more efficient than the estimator $\hat{\bar{y}}_{wor}$ based on simple random sampling when no auxiliary information is used.

4. Empirical Study

To illustrate the performance of the proposed estimator, let us consider the following data.

Population I: Cochran (1977, Page Number- 181)

y : Paralytic Polio Cases 'placebo' group
x : Paralytic Polio Cases in not inoculated group

μ_{02} = 71.8650173, μ_{20} = 9.889273356, μ_{11} = 19.4349481, μ_{12} = 346.3174191,

μ_{03} = 1453.077703, μ_{40} = 424.1846721, μ_{21} = 94.21286383, μ_{22} = 3029.312542,

μ_{30} = 47.34479951, μ_{04} = 46132.5679, $\bar{y}$ = 2.588235294, $\bar{x}$ = 8.370588235,

S_x = 8.477323711, S_y = 3.144721507, ρ =0.729025009, $\beta_2(y)$ = 4.337367369,

$\beta_2(x)$ = 8.932490454, C_X = 1.012751251, C_Y =1.215006037, β = 0.270436839,

n = 34, n' =50 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$ = 0.290860981 and $MSE\left(\bar{y}_g\right)\min$ = 0.245359035.

PRE of the proposed estimator $\bar{y}_g$ over $\hat{\bar{y}}_{wor}$ = 118.5450461.

Population II: Mukhopadhyay (2012, Page Number - 104)
y : Quality of raw materials (in lakhs of bales)
x : Number of labourers (in thousands)

μ_{02} = 9704.4475, μ_{20} = 90.95, μ_{11} = 612.725, μ_{12} = 93756.3475,

μ_{03} = 988621.5173, μ_{40} = 35456.4125, μ_{21} = 11087.635, μ_{22} = 2893630.349,
μ_{30} = 1058.55, μ_{04} = 341222548.2, $\bar{y}$ = 41.5, $\bar{x}$ = 441.95, S_x = 98.51115419,
S_y = 9.536770942, ρ =0.652197067, $\beta_2(y)$ = 4.286367314,

$\beta_2(x)$ = 3.623231573, C_X = 0.22290113, C_Y =0.229801709,

β =0.063138576, n = 20, n' =35 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$ = 4.5475 and $MSE\left(\bar{y}_g\right)\min$ = 3.941933079.

PRE of the proposed estimator $\bar{y}_g$ over $\hat{\bar{y}}_{wor}$ = 115.3621817.

Population III: Murthy (1967, Page Number - 398)
y : Number of absentees
x : Number of workers

μ_{02} = 1299.318551, μ_{20} = 42.13412655, μ_{11} = 154.6041103, μ_{12} = 5086.694392,

μ_{03} = 32025.12931, μ_{40} = 11608.18508, μ_{21} = 1328.325745, μ_{22} = 148328.4069,

μ_{30} = 425.9735118, μ_{04} = 4409987.245, $\bar{y}$ = 9.651162791, $\bar{x}$ = 79.46511628,

S_x = 36.04606151, S_y = 6.491080538, ρ = 0.660763765, $\beta_2(y)$= 6.53877409,

$\beta_2(x)$= 2.612197776, C_X = 0.453608617, C_X = 0.672569791,

β = 0.118988612, n = 43, n' =50 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 0.979863408 and $MSE\left(\bar{y}_g\right)$min = 0.623109375.

PRE of the proposed estimator $\bar{y}_g$ over $\hat{\bar{y}}_{wor}$ = 106.1481375.

Population IV: Singh and Chaudhary (1997, Page Number - 176)

y : Total number of guava trees

x : Area under guava orchard (in acres)

μ_{02} = 12.50056686, μ_{20} = 187123.9172, μ_{11} = 1377.39858, μ_{12} = 4835.465464,

μ_{03} = 37.09863123, μ_{40} = 1.48935E+11, μ_{21} = 712662.4414,

μ_{22} = 8747904.451, μ_{30} = 100476814.5, μ_{04} = 540.1635491, $\bar{y}$ = 746.9230769,

$\bar{x}$ = 5.661538462, S_x = 3.535614072, S_y = 432.5782209, ρ = 0.900596235,

$\beta_2(y)$= 4.253426603, $\beta_2(x)$= 3.456733187, C_X = 0.624497051,

C_Y = 0.579146949, β = 110.1868895, n = 13, n' =30 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 14394.14747 and $MSE\left(\bar{y}_g\right)$min = 7689.477763.

PRE of the proposed estimator $\bar{y}_g$ over $\hat{\bar{y}}_{wor}$ = 187.1927837.

Population V: Singh and Chaudhary (1997, Page Number: 154-155)

y : Number of milch animals in survey

x : Number of milch animals in census

μ_{02} = 431.5847751, μ_{20} = 270.9134948, μ_{11} = 247.3944637,

μ_{12} = 3119.839406, μ_{03} = 5789.778954, μ_{40} = 154027.4827, μ_{21} = 2422.297374, μ_{22} = 210594.3138, μ_{30} = 2273.46265, μ_{04} = 508642.4447, $\bar{y}$ =

1133.294118, $\bar{x}$ = 1140.058824, S_x = 20.77461853, S_y = 16.45945002, ρ = 0.723505104,

$\beta_2(y)$= 2.098635139, $\beta_2(x)$= 2.730740091, C_X = 0.018222409,

C_Y = 0.014523547, β = 0.573223334, n = 17, n' =30 (say).

$MSE\left(\hat{\bar{y}}_{wor}\right)$= 15.93609 and $MSE\left(\bar{y}_g\right)$min = 12.31714015.

PRE of the proposed estimator $\bar{y}_g$ over $\hat{\bar{y}}_{wor}$ = 129.3813964.

5. Conclusions

(i) Any estimator belonging to the generalized class of estimators represented by $\bar{y}_g$ cannot have mean square error less than the expression $\frac{\bar{Y}^2}{n}C_Y^2 - \left(\frac{1}{n} - \frac{1}{n'}\right)\left\{\rho^2\bar{Y}^2C_Y^2 + \frac{1}{\Delta C_X^2}\left(\delta_1 - \delta_2\right)^2\right\}$ (5.1)

(ii) From equation (2.8) and equation (2.9), the mean squared error of the estimator $\bar{y}_g$ i.e. MSE $\left(\bar{y}_g\right)$ is minimized for the optimum values

$$g_1 = -\frac{1}{C_X^2}\left\{\frac{\mu_{11}}{\bar{X}}\bar{Y} - \frac{1}{\Delta}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)\left(\delta_1 - \delta_2\right)\right\} \tag{5.2}$$

and $g_2 = -\frac{1}{\Delta}\bar{X}\bar{\theta}_X\left(\delta_1 - \delta_2\right)$ (5.3)

where $\delta_1 = C_X^2\bar{X}^2\left(\mu_{12} + 2\bar{X}\mu_{11}\right)$, $\delta_2 = \rho C_Y C_X \bar{Y}\left(\mu_{03} + 2\bar{X}\mu_{02}\right)$,

$$\Delta = \mu_{02}^3\left(\beta_2 - \beta_1 - 1\right),\ \beta_2 = \frac{\mu_{04}}{\mu_{02}^2} \quad \text{and} \quad \beta_1 = \frac{\mu_{03}^2}{\mu_{02}^3}.$$

The optimum values involving some unknown parameters may not be known in advance for practical purposes and hence the alternative is to replace the unknown parameters of the optimum values by their unbiased estimators giving a subclass of estimators depending upon estimated optimum values.

Acknowledgement

The authors are immensely grateful to the anonymous referees and the editor in chief for providing their useful suggestions on an earlier version of the paper and also for mentioning the words of encouragement towards the research work done.

References

Cochran, W.G. (1977): Sampling Techniques, 3rd edition, John Wiley and Sons, New York.

Murthy, M (1967): Sampling Theory and Methods, 1st edition, Calcutta Statistical Publishing Society, Kolkata, India.

Mukhopadhyay, Parimal (2012): Theory and Methods of Survey and Sampling, 2nd edition, PHI Learning Private Limited, New Delhi, India.

Singh, Daroga and Chaudhary, F. S. (1997): Theory and Analysis of Sampling Survey Designs, New Age International Publishers, New Delhi, India.

Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. And Asok, C. (1984): Sampling Theory of Surveys with Applications, 3rd Edition, Ames, Iowa (USA) and Indian Society of Agricultural Statistics, New Delhi, India.