A Two Phase Sampling Estimator of Population Mean Using Auxiliary Information

R. Karan Singh (Rtd)
Department of Statistics, University of Lucknow, Lucknow- 226007, Uttar Pradesh (India)
singhrkaran@rediffmail.com
Peeyush Misra
Department of Statistics, D.A.V.(P.G.) College, Dehradun- 248001, Uttarakhand (India)
dr.pmisra.dav@gmail.com
V. S. Singh
Department of Statistics, HNB Garhwal University, SRT Campus, Badshahithaul,
Tehri Garhwal- 249199, Uttarakhand (India)
vssingh1965@gmail.com

Abstract: This paper provides a two phase sampling estimator of population mean using auxiliary information. Its bias and mean square error are found. Comparative studies with some of the well known estimators have been done. An empirical study is also given as an illustration.

Keywords: Auxiliary Variable, Bias, Mean Squared Error and Efficiency.

1. Introduction:

One of the major developments in sample surveys over the last five decades is the use of auxiliary information and statisticians make use of this 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), Des Raj (1968), Murthy (1967), Mukhopadhyay (2012), Singh & Chaudhary (1997) and Sukhatme et, al. (1984). It is well known that the auxiliary information in sample surveys results in substantial improvement in the precision of the estimators of the population parameters and we know that sometimes parameters of the auxiliary variables are not known in advance then double or two phase sampling technique is used. 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 $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$ be the population mean of study variable y and $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ be the population mean of auxiliary variable x .

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

be the population correlation coefficient between y and x.

Also let $\mu_{rs} = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \bar{Y}\right)^r\left(X_i - \bar{X}\right)^s$, $C_Y^2 = \frac{\sigma_Y^2}{\bar{Y}^2}$, $C_X^2 = \frac{\sigma_X^2}{\bar{X}^2} = \frac{\mu_{02}}{\bar{X}^2}$, $\rho = \frac{\mu_{11}}{\sigma_Y \sigma_X}$, $\beta_2 = \frac{\mu_{04}}{\mu_{02}^2}$, $\beta_1 = \frac{\mu_{03}^2}{\mu_{02}^3}$, $\gamma_1 = \sqrt{\beta_1}$.

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 $\left\{(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)\right\}$ on variables (y, x) 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.

For simplicity, it is assumed that N is large enough as compared to n so that finite population correction terms may be ignored. A new improved two phase sampling estimator represented by $\hat{\bar{y}}$ for estimating the population mean is proposed as

$$\hat{\bar{y}} = \bar{y} + b(\bar{x}' - \bar{x}) + k\left(\frac{s_x^2}{\bar{x}^2} - C_X'^2\right) \tag{1.1}$$

where $C_X'^2 = \frac{s_x'^2}{\bar{x}'^2}$ and b is an estimate of the change in y when x is increased by unity.

2. Bias and Mean Square Error of the Proposed Estimator:

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

$$\bar{y} = \overline{Y}(1 + e_0)$$
$$\bar{x} = \overline{X}(1 + e_1)$$
$$\bar{x}' = \overline{X}(1 + e_1')$$
$$s_{yx} = e_2 + S_{YX}$$
$$s_x^2 = e_3 + S_X^2$$
$$s_x'^2 = e_3' + S_X^2 \tag{2.1}$$

so that ignoring finite population correction, for simplicity we have

$$E(e_0) = E(e_1) = E(e_1') = E(e_2) = E(e_3) = E(e_3') = 0 \tag{2.2}$$

$$E(e_0^2) = \frac{\mu_{20}}{n\overline{Y}^2} = \frac{1}{n}C_Y^2$$

$$E(e_1^2) = \frac{\mu_{02}}{n\overline{X}^2} = \frac{1}{n}C_X^2$$

$$E(e_1'^2) = \frac{\mu_{02}}{n'\overline{X}^2} = \frac{1}{n'}C_X^2$$

$$E(e_3^2) = \left(\frac{\beta_2(x) - 1}{n}\right)S_X^4 = \frac{\mu_{02}^2}{n}\left(\frac{\mu_{04}}{\mu_{02}^2} - 1\right)$$

$$E(e_3'^2) = \left(\frac{\beta_2(x) - 1}{n'}\right)S_X^4 = \frac{\mu_{02}^2}{n'}\left(\frac{\mu_{04}}{\mu_{02}^2} - 1\right)$$

$$E(e_0e_1) = \frac{\mu_{11}}{n\overline{Y}\overline{X}} = \frac{1}{n}\rho C_Y C_X$$

$$E(e_0e_1') = \frac{\mu_{11}}{n'\overline{Y}\overline{X}} = \frac{1}{n'}\rho C_Y C_X$$

$$E(e_0e_3) = \frac{\mu_{12}}{n\overline{Y}}$$

$$E(e_0 e_3') = \frac{\mu_{12}}{n'\bar{Y}}$$

$$E(e_1 e_1') = \frac{\mu_{02}}{n'\bar{X}^2} = \frac{1}{n'} C_X^2$$

$$E(e_1 e_2) = \frac{\mu_{12}}{n\bar{X}}$$

$$E(e_1 e_3) = \frac{\mu_{03}}{n\bar{X}}$$

$$E(e_1 e_3') = \frac{\mu_{03}}{n'\bar{X}}$$

$$E(e_1' e_2) = \frac{\mu_{12}}{n'\bar{X}}$$

$$E(e_1' e_3) = \frac{\mu_{03}}{n'\bar{X}}$$

$$E(e_1' e_3') = \frac{\mu_{03}}{n'\bar{X}}$$

$$E(e_3 e_3') = \left(\frac{\beta_2(x)-1}{n'}\right) S_X^4 = \frac{\mu_{02}^2}{n'}\left(\frac{\mu_{04}}{\mu_{02}^2} - 1\right) \quad (2.3)$$

The proposed two phase sampling estimator represented by $\hat{\hat{y}}$ for estimating the population mean given in (1.1) is

$$\hat{\hat{y}} = \bar{y} + b(x' - \bar{x}) + k\left(\frac{s_x^2}{\bar{x}^2} - C_X'^2\right) \quad where \quad C_X'^2 = \frac{s_x'^2}{\bar{x}'^2} \quad (2.4)$$

In terms of e_i's, $i = 0,1,2,3$; the above proposed two phase sampling estimator up to terms of order $O(1/n)$ reduces to

$$\hat{\hat{y}} - \bar{Y} = \bar{Y}e_0 - \beta\bar{X}e_1 + \beta\bar{X}e_1' + \frac{k}{\bar{X}^2}e_3 - \frac{k}{\bar{X}^2}e_3' - 2kC_X^2 e_1 + 2kC_X^2 e_1' - \frac{\beta}{\bar{X}C_X^2}(e_1'e_3 - e_1e_3) + \frac{\bar{X}}{S_X^2}(e_1'e_2 - e_1e_2)$$
$$- \frac{2k}{\bar{X}^2}(e_1e_3 - e_1'e_3') + 3kC_X^2(e_1^2 - e_1'^2) \quad (2.5)$$

where $\beta = \dfrac{S_{YX}}{S_X^2}$.

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

$$\text{Bias}\left(\hat{\hat{y}}\right) = \left\{E\left(\hat{\hat{y}}\right) - \bar{Y}\right\} = \left(\frac{1}{n} - \frac{1}{n'}\right)\left(\frac{\beta\mu_{03}}{\bar{X}^2C_X^2} - \frac{\mu_{12}}{S_X^2} - \frac{2k\mu_{03}}{\bar{X}^3} + \frac{3kC_X^2\mu_{02}}{\bar{X}^2}\right) \quad (2.6)$$

Now squaring both sides of (2.5) and taking expectation, the mean squared error up to terms of order $O(1/n)$ is given by

$$\text{MSE}\left(\hat{\hat{y}}\right) = \left\{E\left(\hat{\hat{y}}\right) - \bar{Y}\right\}^2$$
$$= \bar{Y}^2E(e_0^2) + \beta^2\bar{X}^2E(e_1^2) + \beta^2\bar{X}^2E(e_1'^2) - 2\beta^2\bar{X}^2E(e_1e_1') - 2\beta\overline{YX}E(e_0e_1) + 2\beta\overline{YX}E(e_0e_1')$$

$$+\frac{k^2}{\bar{X}^4}E\left(e_3^2\right)+\frac{k^2}{\bar{X}^4}E\left(e_3'^2\right)+4k^2C_X^4E\left(e_1^2\right)+4k^2C_X^4E\left(e_1'^2\right)+\frac{2k\bar{Y}}{\bar{X}^2}E\left(e_0e_3\right)-\frac{2k\bar{Y}}{\bar{X}^2}E\left(e_0e_3'\right)$$
$$-4\bar{Y}kC_X^2E\left(e_0e_1\right)+4\bar{Y}kC_X^2E\left(e_0e_1'\right)+\frac{2\beta k}{\bar{X}}E\left(e_1'e_3\right)-\frac{2\beta k}{\bar{X}}E\left(e_1'e_3'\right)-4\beta\bar{X}kC_X^2E\left(e_1e_1'\right)$$
$$+4\beta\bar{X}kC_X^2E\left(e_1'^2\right)-\frac{2\beta k}{\bar{X}}E\left(e_1e_3\right)+\frac{2\beta k}{\bar{X}}E\left(e_1e_3'\right)+4\beta\bar{X}kC_X^2E\left(e_1^2\right)-4\beta\bar{X}kC_X^2E\left(e_1e_1'\right)$$
$$-\frac{2k^2}{\bar{X}^4}E\left(e_3e_3'\right)-\frac{4k^2C_X^2}{\bar{X}^2}E\left(e_1e_3\right)+\frac{4k^2C_X^2}{\bar{X}^2}E\left(e_1'e_3\right)+\frac{4k^2C_X^2}{\bar{X}^2}E\left(e_1e_3'\right)$$
$$-\frac{4k^2C_X^2}{\bar{X}^2}E\left(e_1'e_3'\right)-8k^2C_X^4E\left(e_1e_1'\right)$$

using values of the expectation given in (2.2) and (2.3), we have

$$\text{MSE}\left(\hat{\bar{y}}\right)=\frac{\mu_{20}}{n}+\beta^2\mu_{02}\left(\frac{1}{n}-\frac{1}{n'}\right)-2\beta\mu_{11}\left(\frac{1}{n}-\frac{1}{n'}\right)+\left\{\frac{\mu_{02}^2}{\bar{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)+\frac{4C_X^4\mu_{02}}{\bar{X}^2}-\frac{4C_X^2\mu_{03}}{\bar{X}^3}\right\}\left(\frac{1}{n}-\frac{1}{n'}\right)k^2$$
$$+\left\{\frac{2\mu_{12}}{\bar{X}^2}-\frac{4C_X^2\mu_{11}}{\bar{X}}-\frac{2\beta\mu_{03}}{\bar{X}^2}+\frac{4\beta C_X^2\mu_{02}}{\bar{X}}\right\}\left(\frac{1}{n}-\frac{1}{n'}\right)k \qquad (2.7)$$

which attains the minimum for the optimum value

$$k=\frac{\left\{\frac{4C_X^2\mu_{11}}{\bar{X}}+\frac{2\beta\mu_{03}}{\bar{X}^2}-\frac{2\mu_{12}}{\bar{X}^2}-\frac{4\beta C_X^2\mu_{02}}{\bar{X}}\right\}}{2\left\{\frac{\mu_{02}^2}{\bar{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2}-1\right)+\frac{4C_X^4\mu_{02}}{\bar{X}^2}-\frac{4C_X^2\mu_{03}}{\bar{X}^3}\right\}} \qquad (2.8)$$

Substituting the value of k given by (2.8) in (2.7), we get the minimum mean squared error of $\hat{\bar{y}}$ to be

$$\text{MSE}\left(\hat{\bar{y}}\right)_{\min}=\frac{\mu_{20}}{n}+\left(\frac{1}{n}-\frac{1}{n'}\right)\beta^2\mu_{02}-2\left(\frac{1}{n}-\frac{1}{n'}\right)\beta\mu_{11}$$
$$-\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{\left\{2\bar{X}C_X^2\left(\mu_{11}-\beta\mu_{02}\right)-\left(\mu_{12}-\beta\mu_{03}\right)\right\}^2}{\left\{\left(\mu_{04}-\mu_{02}^2\right)-4\bar{X}C_X^2\left(\mu_{03}-\bar{X}C_X^2\mu_{02}\right)\right\}} \qquad (2.9)$$

3. Efficiency Comparison:

(i) General estimator of mean in case of SRSWOR:

The general estimator of mean in case of SRSWOR is $\hat{\bar{y}}_{wor}=\bar{y}$ with $MSE\left(\hat{\bar{y}}\right)=\frac{\mu_{20}}{n}$ (3.1)

It is clear that the proposed estimator is more efficient than the estimator $\hat{\bar{y}}_{wor}$ based on simple random sampling when no auxiliary information is used.

(ii) Usual double sampling regression estimator:

The usual double sampling regression estimator is $\bar{y}_{ld}=\bar{y}+b\left(\bar{x}'-\bar{x}\right)$ with

$$MSE\left(\bar{y}_{ld}\right)=\frac{\mu_{20}}{n}-\left(\frac{1}{n}-\frac{1}{n'}\right)\frac{\mu_{11}^2}{\mu_{02}} \qquad (3.2)$$

It is clear that the proposed estimator is more efficient than the usual double sampling regression estimator where the auxiliary information already is in use.

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).

$MSE\left(\hat{\hat{y}}_{wor}\right)$= 0.290860981, $MSE\left(\bar{y}_{ld}\right)$ = 0.241393443 and $MSE\left(\hat{\hat{y}}\right)\min$ = 0.239722289

PRE of the proposed estimator $\hat{\hat{y}}$ over $\hat{\hat{y}}_{wor}$ = 121.3324727.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\bar{y}_{ld}$ = 100.6971207.

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).

$MSE\left(\hat{\hat{y}}_{wor}\right)$= 4.5475, $MSE\left(\bar{y}_{ld}\right)$= 3.718501766 and $MSE\left(\hat{\hat{y}}\right)\min$ = 3.600902564.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\hat{\hat{y}}_{wor}$ = 126.2877826.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\bar{y}_{ld}$ = 103.2658257.

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).

$MSE\left(\hat{\hat{y}}_{wor}\right)$= 0.979863408, $MSE\left(\bar{y}_{ld}\right)$= 0.919969037 and $MSE\left(\hat{\hat{y}}\right)\min$ = 0.917340179.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\hat{\hat{y}}_{wor}$ = 106.8157081.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\bar{y}_{ld}$ = 100.2865739.

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).
$MSE(\hat{\hat{y}}_{wor})$= 14394.14747, $MSE(\bar{y}_{ld})$= 7778.476942 and $MSE(\hat{\hat{y}})\min$ = 7697.255082

PRE of the proposed estimator $\hat{\hat{y}}$ over $\hat{\hat{y}}_{wor}$ = 187.0036438.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\bar{y}_{ld}$ = 101.0552055.

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).
$MSE(\hat{\hat{y}}_{wor})$= 15.93609, $MSE(\bar{y}_{ld})$= 12.32127 and $MSE(\hat{\hat{y}})\min$ = 12.31804899.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\hat{\hat{y}}_{wor}$ = 129.3718505.

PRE of the proposed estimator $\hat{\hat{y}}$ over $\bar{y}_{ld}$ = 100.0261091.

5. Conclusions:

(i) From (2.9) it is clear that the proposed two phase sampling estimator is more efficient than the estimator $\hat{\hat{y}}_{wor}$ based on simple random sampling when no auxiliary information is used and is also more efficient than the usual double sampling regression estimator $\bar{y}_{ld}$ of mean where the auxiliary information already is in use.

(ii) From (2.8), the mean squared error of the estimator $\hat{\hat{y}}$ is minimized for the optimum value

$$k = \frac{\left\{\frac{4C_X^2\mu_{11}}{\bar{X}} + \frac{2\beta\mu_{03}}{\bar{X}^2} - \frac{2\mu_{12}}{\bar{X}^2} - \frac{4\beta C_X^2\mu_{02}}{\bar{X}}\right\}}{2\left\{\frac{\mu_{02}^2}{\bar{X}^4}\left(\frac{\mu_{04}}{\mu_{02}^2} - 1\right) + \frac{4C_X^4\mu_{02}}{\bar{X}^2} - \frac{4C_X^2\mu_{03}}{\bar{X}^3}\right\}} \quad (5.1)$$

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

Acknowledgement: The authors are thankful to the referees and the editor in chief for their valuable comments regarding improvement of the paper.

References:

1. Cochran, W.G. (1977): Sampling Techniques, 3rd edition, John Wiley and Sons, New York.
2. Des Raj (1968): Sampling Theory, McGraw- Hill, New York.
3. Murthy, M (1967): Sampling Theory and Methods, 1st edition, Calcutta Statistical Publishing Society, Kolkata, India.
4. Mukhopadhyay, Parimal (2012): Theory and Methods of Survey and Sampling, 2nd edition, PHI Learning Private Limited, New Delhi, India.
5. Singh, Daroga and Chaudhary, F. S. (1997): Theory and Analysis of Sampling Survey Designs, New Age International Publishers, New Delhi, India.
6. 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.

