ISSN 0974 - 9373

Vol. 18 No.1 (2014) **Journal of International Academy of Physical Sciences** pp. 01-10

An Improved Estimation Procedure of Population Mean in Two Phase Sampling

G. N. Singh and A. K. Sharma
Department of Applied Mathematics
Indian School of Mines, Dhanbad-826004, India
E-mails: gnsingh_ism@yahoo.com, aksharma.ism@gmail.com

(Received January 20, 2014)

Abstract: This paper presents an improved estimator of population mean under the frame work of two phase sampling scheme in presence of two auxiliary variables. Performance of the proposed estimator has been examined in terms of theoretical and empirical comparisons. Empirical studies suggest that the proposed estimator dominates over the several well known estimators under the similar situations. Results are analyzed and recommendations have been made.

Keywords: Two-phase, Study variable, Auxiliary variable, Bias, Mean square error.

Mathematics subject classification: 62D05

1. Introduction

In sample surveys it is an usual practice to make use of information on auxiliary variables to obtain more efficient estimators. It is well known that when the auxiliary information is to be used at the estimation stage, the ratio, product and regression estimators are widely utilized in many situations. If the information on the auxiliary variable is not available a large preliminary sample is usually taken from the population to furnish an estimate of the population mean of the auxiliary variable and a sub sample from the preliminary sample is drawn for collecting the information on study variable. This technique known as two-phase sampling is especially appropriate if the information on auxiliary variable is easily accessible and much cheaper to collect than the information on study variable. Utilizing the information on known population mean of another auxiliary variable in first phase sample, Chand[1] introduced chain-type ratio estimator of population mean of study variable. Further his work was extended by Kiregyera[2,3] Mukharjee et al.[4], Sirvastava et al.[5],Singh and Singh[6], Singh et al.[7], Singh and Upadhayaya[8] , Upadhayaya and Singh[9], Singh[10] and Pradhan[11] among others. Motivated with the preceding work the aim of present research is to propose a new chain-type estimator in two-phase sampling which may

estimate the population mean in more precise way in compare to the contemporary estimators of similar kind.

2. Two-Phase Sampling Set Up

Let $U=(U_1, U_2,..., U_N)$ be a finite population of size N and the variables associated with the unit U_i of population is (y_i, x_i, z_i) $(i=1, 2,..., N)$. We wish to estimate the population mean $\bar{Y}$ of study variable y in the presence of two auxiliary variables x and z. Let x and z be called as first and second auxiliary variables respectively such that y is highly correlated with x while in compare to x, it is remotely correlated with z (i.e. $\rho_{yx} > \rho_{yz}$). When the population mean $\bar{X}$ of x is unknown but information on z is available on all the units of the population, we use the following two-phase sampling scheme.

Let us now consider a two-phase sampling where in the first phase a large (preliminary) sample $s'(s' \subset U)$ of fixed size n' is drawn following SRSWOR and observe two auxiliary variables x and z to estimate $\bar{X}$, while in the second phase a sub-sample $s \subset s'$ of fixed size n is drawn by SRSWOR to observe the characteristic y under study.

3. Estimators Based on One Auxiliary Variable

Ratio and regression estimators in two-phase sampling are the traditional estimators utilize the information on one auxiliary variable and are reproduced below along with their respective mean square error up to the first order approximations.

(3.1) $$\bar{y}_{rd} = \frac{\bar{y}}{\bar{x}} \bar{x}'$$

(3.2) $$M(\bar{y}_{rd}) = \bar{Y}^2 \left[f_1 c_y^2 + f_3 (c_x^2 - 2\rho_{yx} C_y C_x) \right]$$

(3.3) $$\bar{y}_{lrd} = \bar{y} + b_{yx}(n)(\bar{x}' - \bar{x})$$

(3.4) $$M(\bar{y}_{lrd}) = S_y^2 \left[f_1 (1 - \rho_{yx}^2) + f_2 \rho_{yx}^2 \right]$$

where $b_{yx}(n)$ is the sample regression coefficient of y on x calculated from the data based on s and

$$\bar{y}=\frac{1}{n}\sum_{i\in s} y_i,\ \bar{x}=\frac{1}{n}\sum_{i\in s} x_i \text{ and } \bar{x}'=\frac{1}{n'}\sum_{i\in s'} x_i,\ f_1=\left(\frac{1}{n}-\frac{1}{N}\right),$$

$$f_2=\left(\frac{1}{n'}-\frac{1}{N}\right),\ f_3=(f_1-f_2)=\left(\frac{1}{n}-\frac{1}{n'}\right)$$

$$S_x^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\bar{X}\right)^2,\ S_y^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(y_i-\bar{Y}\right)^2,$$

$C_x=\frac{S_x}{\bar{X}}, C_y=\frac{S_y}{\bar{Y}}$ and ρ_{yx} be the correlation coefficient between variables y and x.

4. Estimators based on two auxiliary variables

Chand[1] introduced a chain-type ratio estimator under two-phase sampling using two auxiliary variables x and z when the population mean $\bar{X}$ of x is unknown but information on z is available on all the units of the population (i.e. population mean $\bar{Z}$ of auxiliary variable z is known), which is given as

$$\bar{y}_{rc}=\frac{\bar{y}}{\bar{x}}\frac{\bar{x}'}{\bar{z}'}\bar{Z} \tag{4.1}$$

The mean square error of the estimator $\bar{y}_{rc}$ up to the first order approximations is derived as

$$M(\bar{y}_{rc})=\bar{Y}^2\left[f_1C_y^2+f_3\left(C_x^2-2\rho_{yx}C_yC_x\right)+f_2\left(C_z^2-2\rho_{yz}C_yC_z\right)\right] \tag{4.2}$$

where $C_z=\frac{S_z}{\bar{Z}}$, $S_z^2=\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i-\bar{Z}\right)^2$ and ρ_{yz} be correlation coefficient between variables y and z.

Kiregyera[2,3] extended the work of Chand[1] and suggested chain-type ratio to regression, regression to ratio and regression to regression estimators of population mean of study variable y in two-phase sampling which utilize the information on two auxiliary variables. The suggested estimators are given below along with their respective mean square error up to the first order approximations.

$$(4.3) \qquad \bar{y}_{k1} = \frac{\bar{y}}{\bar{x}}\left[\bar{x}' + b_{xz}(n')\left(\bar{Z} - \bar{z}'\right)\right]$$

$$(4.4) \qquad M\left(\bar{y}_{k1}\right) = \bar{Y}^2\left[f_3\left(C_x^2 + C_y^2 - 2\rho_{yx}C_yC_x\right) + f_2C_y^2 + f_2\rho_{xz}C_x\left(\rho_{xz}C_x - 2\rho_{yz}C_y\right)\right]$$

$$(4.5) \qquad \bar{y}_{k2} = \bar{y} + b_{yx}(n)\left(\bar{x}'_{rd} - \bar{x}\right);\ \bar{x}'_{rd} = \frac{\bar{x}'}{\bar{z}'}\bar{Z}$$

$$(4.6) \qquad M\left(\bar{y}_{k2}\right) = \bar{Y}^2C_y^2\left[f_1\left(1 - \rho_{yx}^2\right) + f_2\left(\rho_{yx}^2 + \rho_{yx}^2\frac{C_z^2}{C_x^2} - 2\rho_{yx}\rho_{yz}\frac{C_z}{C_x}\right)\right]$$

$$(4.7) \qquad \bar{y}_{k3} = \bar{y} + b_{yx}(n)\left(\bar{x}'_{ld} - \bar{x}\right);\ \bar{x}'_{ld} = \left[\bar{x}' + b_{xz}(n')\left(\bar{Z} - \bar{z}'\right)\right]$$

$$(4.8) \qquad M\left(\bar{y}_{k3}\right) = \bar{Y}^2C_y^2\left[f_3\left(1 - \rho_{yx}^2\right) + f_2\left(1 + \rho_{yx}^2\rho_{xz}^2 - 2\rho_{yx}\rho_{yz}\rho_{xz}\right)\right]$$

where $b_{xz}(n')$ is the sample regression coefficient of x on z calculated from the data based on s′, $\bar{z}' = \frac{1}{n'}\sum_{i\in s'} z_i$ and ρ_{xz} is the correlation coefficient between x and z.

5. Proposed Estimator

Motivated with the work related to the proposition of chain-type estimators in two-phase sampling set up, we suggest below regression to exponential chain type estimator of population mean $\bar{Y}$ of the study variable y.

(5.1) $$T = \bar{y}^{*} + b_{yx}(n)\left(\bar{x}^{**} - \bar{x}^{*}\right)$$

where $\bar{y}^{*} = y + b_{yz}(n)\left(\bar{Z} - \bar{z}\right)$, $\bar{x}^{**} = \bar{x}^{'}\exp\left(\frac{\bar{Z} - \bar{z}^{'}}{\bar{Z} - \bar{z}^{'}}\right)$ and $\bar{x}^{*} = \bar{x}\exp\left(\frac{\bar{Z} - \bar{z}}{\bar{Z} + \bar{z}}\right)$

where $b_{yz}(n)$ is the sample regression coefficient of y on z calculated from the data based on s.

6. Properties of the Proposed Estimator

Theorem 6.1: Bias of the estimator T defined in (5.1) up to the first order approximations is obtained as

(6.1) $$B(T) = f_1\left(\frac{\alpha_{014}}{\alpha_{003}} - \frac{\alpha_{023}}{\alpha_{013}}\right) + f_3\left(\frac{1}{\bar{Z}}\left(\frac{1}{2}\frac{\alpha_{211}}{\alpha_{200}} - \frac{3}{8}\frac{\bar{X}}{\bar{Z}}\frac{\alpha_{112}}{\alpha_{200}}\right)\right)$$

where $\alpha_{rst} = E\left[\left(x_i - \bar{X}\right)^{r}\left(y_i - \bar{Y}\right)^{s}\left(z_i - \bar{Z}\right)^{t}\right]$; $(r, s, t) \geq 0$ are integers.

Theorem 6.2: Mean square error of the estimators T defined in (5.1) up to the first order approximations is derived as

(6.2) $$M(T) = \bar{Y}^{2}C_y^{2}\left[f_1\left(1 - \rho_{yz}^{2}\right) + f_3\left\{\rho_{yx}^{2}\left(\frac{1}{4}\left(\frac{C_z}{C_x}\right)^{2} - \rho_{xz}\frac{C_z}{C_x} - 1\right) + 2\rho_{yz}\rho_{yx}\rho_{xz}\right\}\right]$$

7. Efficiency Comparison

In this section we compare the proposed estimator T with respect to the estimators $\bar{y}_{rd}, \bar{y}_{lrd}, \bar{y}_{rc}, \bar{y}_{k1}, \bar{y}_{k2}$ and $\bar{y}_{k3}$. Preference zones of the estimator T are explored and shown below:

(i) T is better than $\bar{y}_{rd}$ if $M(T) \leq M(\bar{y}_{rd})$, which gives

$$\frac{A_1-\left(\rho_{yx}-\frac{C_x}{C_y}\right)^2}{\rho_{yz}^2}\le\frac{f_1}{f_2} \tag{7.1}$$

(ii) T is preferable over $\bar{y}_{lrd}$ if $M(T)\le M(\bar{y}_{lrd})$, which shows

$$\frac{A_1}{\rho_{yz}^2}\le\frac{f_1}{f_2} \tag{7.2}$$

(iii) T will dominate $\bar{y}_{rc}$ if $M(T)\le M(\bar{y}_{rc})$, subsequently we get

$$\frac{A_1-\left(\rho_{yx}-\frac{C_x}{C_y}\right)^2-\rho_{yz}^2}{A_1-\left(\rho_{yx}-\frac{C_x}{C_y}\right)^2+\left(\rho_{yz}-\frac{C_z}{C_y}\right)^2-\rho_{yz}^2}\le\frac{f_2}{f_1} \tag{7.3}$$

(iv) T is more efficient than $\bar{y}_{k1}$ if $M(T)\le M(\bar{y}_{k1})$, which gives

$$\frac{A_1-\left(\rho_{yx}-\frac{C_x}{C_y}\right)^2-\rho_{yz}^2}{A_1-\left(\rho_{yx}-\frac{C_x}{C_y}\right)^2-\left(\rho_{yz}-\rho_{xz}\frac{C_x}{C_y}\right)^2-\rho_{yz}^2}\le\frac{f_2}{f_1} \tag{7.4}$$

(v) T is more desirable over $\bar{y}_{k2}$ if $M(T)\le M(\bar{y}_{k2})$, which is shown below

$$\frac{A_1-\rho_{yz}^2}{A_1+\rho_{yx}^2\frac{C_z}{C_x}\left(\frac{C_z}{C_x}-2\rho_{yx}\rho_{yz}\right)^2-\rho_{yx}^2}\le\frac{f_2}{f_1} \tag{7.5}$$

(vi) T is favorable over $\bar{y}_{k3}$ if $M(T)\le M(\bar{y}_{k3})$, subsequently we get

$$(7.6) \qquad \frac{A_1 - \rho_{yz}^2}{A_1 + \rho_{yx}^2\left(\rho_{xz}^2 - 1\right) - 2\rho_{yx}\rho_{yz}\rho_{xz}} \leq \frac{f_2}{f_1}$$

where $A_1 = \rho_{yx}^2 \frac{C_z}{C_x}\left(\frac{1}{4}\frac{C_z}{C_x} - \rho_{xz}\right) + 2\rho_{yz}\rho_{yx}\rho_{xz}$

8. Empirical Studies

To examine the performance of the proposed estimator T, we have computed the percent relative efficiencies of T with respect to $\bar{y}_{rd}, \bar{y}_{lrd}, \bar{y}_{rc}, \bar{y}_{k1}, \bar{y}_{k2}$ and $\bar{y}_{k3}$ based on seven natural populations and presented in Table-1.The percent relative efficiencies of the estimator T with respect to an estimator δ is defined as

$$\text{PRE} = \left[\frac{\text{MSE}(\delta)}{\text{MSE}(T)}\right] x100$$

Population I: Source: Fisher et al.[12]

Y: Measurement of Petal width
X: Measurement of Sepal width
Z: Measurement of Petal length

$$N=50,\ n'=18,\ n=8,\ \bar{Y} = 2.026,\ \rho_{yx} = 0.5377,\ \rho_{yz} = 0.3221,$$
$$\rho_{xz} = 0.4010,\ C_y^2 = 0.018009, C_x^2 = 0.011524,\ C_z^2 = 0.009683.$$

Population II: Source: Fisher et al. [12]

Y: Measurement of Petal width
X: Measurement of Sepal width
Z: Measurement of Petal length

$$N=50,\ n'=20,\ n=10,\ \bar{Y} = 2.770,\ \rho_{yx} = 0.5259,\ \rho_{yz} = 0.5605,$$
$$\rho_{xz} = 0.7540,\ C_y^2 = 0.012566, C_x^2 = 0.007343,\ C_z^2 = 0.011924.$$

Population III: Source: Cochran [13]

Y: Number of 'placebo' children
X: Number of paralytic polio cases in the placebo group
Z: Number of paralytic polio cases in the 'not inoculated' group

$N=34,\ n'=15,\ n=10,\ \bar{Y}=4.92,\ \bar{X}=2.59,\ \bar{Z}=2.91,\ \rho_{yx}=0.7326,$

$\rho_{yz}=0.6430,\ \rho_{xz}=0.6837,\ C_y^2=1.0248,\ C_x^2=1.5175,\ C_z^2=1.1492.$

Population IV: Source: Shukla[14]

Y: measurement of yield of fiber
X: measurement of height
Z: measurement of base diameter

$N=50,\ n'=15,\ n=8,\ \bar{Y}=2.5840,\ \bar{X}=2.59,\ \bar{Z}=2.91,\ \rho_{yx}=0.4800,$

$\rho_{yz}=0.3700,\ \rho_{xz}=0.7300,\ C_y^2=0.0866,\ C_x^2=0.01163,\ C_z^2=0.0170.$

Population V: Source: Fisher *et al.* [12]

Y: Measurement of Petal width
X: Measurement of Sepal width
Z: Measurement of Petal length

$N=34,\ n'=15,\ n=10,\ \bar{Y}=2.770,\ \rho_{yx}=0.5605,\ \rho_{yz}=0.5259,$

$\rho_{xz}=0.7540,\ C_y^2=0.012566, C_x^2=0.007343,\ C_z^2=0.011924.$

Population VI: Source: Srivnstava, Srivastava, and Khare [5]

Y: Measurement of weight of children.
X: Mid-arm circumference of children
Z: Skull circumference of children

$N=82,\ n'=43,\ n=25,\ \bar{Y}=5.60\ \text{kg},\ \rho_{yx}=0.09,\ \rho_{yz}=0.12,$

$\rho_{xz}=0.86,\ C_y^2=0.0107, C_x^2=0.0052,\ C_z^2=0.0008.$

Population VII: Source: Anderson[15]

Y: Head length of second son.
X: Head length of first son
Z: Head breadth of first son

$N=25,\ n'=10,\ n=7,\ \bar{Y}=183.84,\ \rho_{yx}=0.7108,\ \rho_{yz}=0.6932,$

$\rho_{xz}=0.7346,\ C_y^2=0.002981, C_x^2=0.002766,\ C_z^2=0.002381.$

Table 1: Percentage relative efficiencies of $M(T)$ with respect to different estimators $M(\bar{y}_{rd})$, $M(\bar{y}_{lrd})$, $M(\bar{y}_{rc})$, $M(\bar{y}_{k1})$, $M(\bar{y}_{k2})$ and $M(\bar{y}_{k3})$.

Estimator	PRE FOR POPULATION I					
	$\overline{y}_{rd}$	$\overline{y}_{lrd}$	$\overline{y}_{rc}$	$\overline{y}_{k1}$	$\overline{y}_{k2}$	$\overline{y}_{k3}$
T	111.41	105.48	114.298	106.83	102.18	101.40
PRE FOR POPULATION II						
T	124.15	119.03	116.43	107.21	102.73	103.53
PRE FOR POPULATION III						
T	160.40	139.68	136.66	123.10	*	100.85
PRE FOR POPULATION IV						
T	108.14	107.26	100.95	101.23	102.20	100.00
PRE FOR POPULATION V						
T	124.00	121.20	118.31	103.42	103.09	101.23
PRE FOR POPULATION VI						
T	122.89	120.47	123.25	131.54	100.18	100.00
PRE FOR POPULATION VII						
T	159.26	154.07	109.13	104.58	100.00	102.70

(*) denotes no gain.

11. Conclusions

From Table-1, it is visible that the proposed estimator T is preferable over the estimators $\overline{y}_{rd}, \overline{y}_{lrd}, \overline{y}_{rc}, \overline{y}_{k1}, \overline{y}_{k2}$ and $\overline{y}_{k3}$ in almost all the populations. Therefore the proposed estimator may be recommended for its practical applications.

Acknowledgement

Authors are thankful to the UGC, New Delhi and Indian School of Mines, Dhanbad, for providing financial assistance and necessary infrastructure to carry out the present research work.

References

1. L. Chand, *Some ratio-type estimators based on two or more auxiliary variables*, Ph. D. dissertation, Iowa State University, Ames, Iowa (1975).
2. B. Kiregyera, A chain ratio-type estimator in finite population double sampling using two auxiliary variables*, Metrika*, **27**(1980) 217-223.
3. B. Kiregyera, Regression-type estimators using the two auxiliary variables and the model of double sampling from finite population*. Metrika*, **31(3-4)** (1984) 215-226.

4. R.Mukherjee, T. J. Rao and K.Vijayan, Regression type estimators using multiple auxiliary information, *Australian & New Zealand Journal of Statistics*, **29**(1987) 244-254.

5. S. R. Srivastava, S. P. Srivastava and B. B. Khare, Chain ratio type estimators for ratio of two population means using auxiliary characters, *Communications in Statistics-Theory and Methods*, **18**(1989) 3917-3926.

6. V. K. Singh, G.N. Singh, Chain type regression estimators with two auxiliary variables under double sampling scheme. *Metron*, **XLIX(1-4)** (1991) 279-289.

7. V. K. Singh, H. P. Singh and H. P. Singh, A general class of chain estimators for ratio and product of two means of a finite population, *Communications in Statistics-Theory and Methods*, **23**(1994) 1341-1355.

8. L. N. Upadhyaya, G. N. Singh, Chain type estimators using transformed auxiliary variable in two-phase sampling, *Advances in modeling and analysis*, **38 (1-2)** (2001) 1-10.

9. G. N Singh, L. N. Upadhyaya, A class of modified chain type estimators using two auxiliary variables in two-phase sampling, *Metro*n, **LIII** (1995) 117-125.

10. G. N. Singh, On the use of transformed auxiliary variable in the estimation of population mean in two-phase sampling, *Statistics in Transition*, **5(3)** (2001) 405-416.

11. B. K. Pradhan, A chain regression estimator in two-phase sampling using multi-auxiliary information, *Bulletin of the malaysian mathematical sciences society*, **28(1)** (2005) 81-86.

12. R. A. Fisher, The use of multiple measurements in taxonomic problems. *Ann. Eugenics*, **7**(1936)179-188.

13. W. G. Cochran, Sampling techniques, *New-York: JohnWiley and Sons,* (1977).

14. G. K. Shukla, An alternative multivariate ratio estimate for finite population, *Calcutta Statistical Association Bulletin,* **15** (1966) 127-134.

15. T. W. Anderson, An Introduction to Multivariate Statistical Analysis, *John Wiley & Sons, Inc., New York* (1958).