Open Journal of Mathematical Sciences

PSR Press

*Article*

# A modified efficient difference-type estimator for population mean under two-phase sampling design

**A. E. Anieting[1,*] and J. K. Mosugu[2]**

[1] Department of Statistics, University of Uyo, Uyo, Nigeria.
[2] National Open University of Nigeria, Abuja, Nigeria.
* Correspondence: akaninyeneanieting@uniuyo.edu.ng

Received:3 December 2019; Accepted: 30 March 2020; Published:9-June-2020.

**Abstract:** In this article, modified difference-type estimator for the population mean in two-phase sampling scheme using two auxiliary variables has been proposed. The mean squared error of the proposed estimator has also been derived using large sample approximation. The efficiency comparison conditions for the proposed estimator in comparison with other existing estimators in which the proposed estimator performed better than the other relevant existing estimators have been given.

**Keywords:** Difference-type estimator, efficiency, mean squared-error, two phase sampling.

**MSC:** 62D05.

## 1. Introduction

Auxiliary information is used either in the estimation stage or in the formation of an estimator to get improved designs and increase the efficiency of estimators in sampling technique. In [1], Laplace started the use of the auxiliary information in formulating ratio type estimation. The statisticians paid a lot of care towards the formation of new and efficient estimators for the population parameters estimation [2–13]. Khan and Al-Hossain [14] suggested a generalized chain ratio in regression estimator for mean of the population using two auxiliary variables. In this research work, a modified form of difference-type estimator for mean of the population using two-phase sampling is suggested [15].

Firstly, we give some definitions and notions. Consider a finite population of size $N$ of different units $U = \{U_1, U_2, U_3, \ldots, U_N\}$. Let $x$ and $y$ be the auxiliary and the study variables with corresponding values $x_i$ and $y_i$ respectively for the $i^{th}$ unit $i = \{1, 2, 3, \ldots, N\}$ defined in a finite population $U$ with means

$$\overline{Y} = (1/N)\sum_{i}^{N} y_i$$

and

$$\overline{X} = (1/N)\sum_{i}^{N} x_i$$

of the study as well as auxiliary variable respectively.

Also let

$$S_x^2 = \frac{1}{N-1}\sum_{i}^{N}(x_i - \overline{X})^2$$

and

$$S_y^2 = \frac{1}{N-1}\sum_{i}^{N}(y_i - \overline{Y})^2$$

be the population variances of the auxiliary and the study variables respectively and let $C_x$ and $C_y$ be the coefficient of variation of the auxiliary as well as study variable respectively, while $\rho_{yx}$ is the correlation coefficient between $x$ and $y$.

Let the sample mean of $x$ and $y$ be as

$$\overline{X} = \frac{1}{n-1}\sum_{i}^{n} x_i$$

and

$$\overline{y} = \frac{1}{n-1}\sum_{i}^{n} y_i$$

respectively. Also let

$$\widehat{S}_y^2 = \frac{1}{n-1}\sum_{i}^{n}(y_i - \overline{y})^2$$

and

$$\widehat{S}_x^2 = \frac{1}{n-1}\sum_{i}^{n}(x_i - \overline{x})^2$$

be the corresponding sample variances of the study as well as auxiliary variable respectively. Let

$$S_{yx} = \frac{\sum_i^N \left(y_i - \overline{Y}\right)\left(x_i - \overline{X}\right)}{N-1},$$

$$S_{yz} = \frac{\sum_i^N \left(y_i - \overline{Y}\right)\left(z_i - \overline{Z}\right)}{N-1}$$

and

$$S_{xz} = \frac{\sum_i^N \left(z_i - \overline{Z}\right)\left(x_i - \overline{X}\right)}{N-1}$$

be the co-variances between their respective subscripts. Similarly

$$b_{yx} = \frac{\hat{S}_{xy}}{\hat{S}_x^2}$$

is the corresponding sample regression coefficient of $y$ on $x$ based on a sample of size $n$. Also,

$$C_y = \frac{S_y}{\overline{Y}}, C_x = \frac{S_x}{\overline{X}} \text{ and } C_z = \frac{S_z}{\overline{Z}}$$

are the coefficient of variations of the study and the auxiliary variables respectively. Also, $\theta = \frac{1}{n} - \frac{1}{N}$, $\theta_1 = \frac{1}{n'} - \frac{1}{N}$ and $\theta_2 = \frac{1}{n} - \frac{1}{n'}$.

## 2. Some existing estimators

Consider a finite population of size N units. To estimate the mean of the population $\overline{Y}$, it is assumed that the correlation between y and x is greater than the correlation between y and z, (i.e$\rho_{yx} > \rho_{yz}$ ). When the mean of the population $\overline{X}$ of the auxiliary variable x is unknown, but information on the other cheaply auxiliary variable say z closely related to x but compared to x remotely to y, is available for all the units in a population. The usage of two phase sampling is imperative in such a situation. In double sampling scheme, a large initial sample of size n′ (n′ <N) is drawn from the population U using simple random sample without replacement sampling (SRSWOR) scheme and measure x and z to estimate $\overline{X}$ and $\overline{Z}$ . In the second phase, a sample (subsample) of size n from first phase sample of size n′, i.e. (n< n′) is drawn using (SRSWOR) or directly from the population U and observed the study variable $y$.

The usual variance of simple estimator $t_o = \overline{y} = \frac{1}{n}\sum_i^n y_i$ up to first order of approximation is given by

$$V(t_o) = \theta S_y^2. \tag{1}$$

The ratio and regression estimators in two-phase sampling and their mean square errors up to first order of approximation are given by

$$t_1 = \frac{\overline{y}\overline{x'}}{\overline{x}}, \tag{2}$$

$$\text{MSE}(t_1) = \overline{Y}^2\left[\theta C_y^2 + \theta_2(C_x^2 - 2C_{yx})\right], \tag{3}$$

$$t_2 = \overline{y} + b_{yx(n)}(\overline{x'} - \overline{x}), \tag{4}$$

$$\text{MSE}(t_2) = S_y^2\left[\theta(1-\rho_{yx}^2) + \theta_1(\rho_{yx}^2)\right]. \tag{5}$$

Chand in [5] proposed the following chain ratio-type estimator which is given by:

$$t_3 = \frac{\overline{y}\overline{x'}}{\overline{x}\overline{z'}}\overline{Z}, \tag{6}$$

$$\text{MSE}(t_3) = \overline{Y}^2\left[\theta C_y^2 + \theta_2\left(C_x^2 - 2C_{yx}\right) + \theta_1\left(C_z^2 - 2C_{yz}\right)\right]. \tag{7}$$

Singh and Majhi in [15] formulated a chain-type exponential estimators for $\overline{Y}$ given by

$$t_5 = \frac{\overline{y}\overline{x'}}{\overline{x}}\exp\left(\frac{\overline{Z} - \overline{z'}}{\overline{Z} + \overline{z'}}\right), \tag{8}$$

$$\text{MSE}(t_5) = \overline{Y}^2\left[\theta C_y^2 + \theta_2\left(C_x^2 - 2C_{yx}\right) + \theta_1/4\left(C_z^2 - 2C_{yz}\right)\right]. \tag{9}$$

Khan and Al-Hossain in [14] gave a difference-type estimator for the mean of the population in two-phase sampling scheme using two auxiliary variables as

$$t_m = \overline{y} + k_1\left(\overline{x'}\frac{\overline{Z}}{\overline{z'}} - \overline{x}\right) + k_2\left(\overline{Z}\frac{\overline{x'}}{\overline{x}} - \overline{z}\right), \tag{10}$$

$$\begin{aligned}\text{MSE}(t_m) = {} & \overline{Y}^2\theta C_y^2 + k_1^2\overline{X}^2(\theta_1 C_z^2 + \theta_2 C_x^2) + k_2^2\overline{Z}^2(\theta C_z^2 + \theta_2 C_x^2 + 2\theta_2 C_{xz}) + 2k_1k_2\overline{XZ}(\theta_2 C_x^2 + \theta_1 C_z^2 + \theta_2 C_{xz}) \\ & - 2k_1\overline{XY}(\theta_2 C_{yx} + \theta_1 C_{yz}) - 2k_2\overline{ZY}(\theta_2 C_{yx} + \theta C_{yz}).\end{aligned} \tag{11}$$

## 3. The proposed estimator

On the basis of Khan and Al-hossain [14], a modified difference-type estimator for the mean of the population in two-phase sampling scheme using two auxiliary variables is proposed as

$$t_{ae} = \overline{y} + k_1\left(\overline{x'} - \frac{\overline{Z}}{\overline{z'}}\overline{x}\right) + k_2\left(\overline{z} - \overline{Z}\frac{\overline{x'}}{\overline{x}}\right), \tag{12}$$

where $k_1$ and $k_2$ are unknown constants.

Let

$$\begin{cases} e_o = \frac{\overline{y}-\overline{Y}}{\overline{Y}}, \\ e_1 = \frac{\overline{x}-\overline{X}}{\overline{X}}, \\ e'_1 = \frac{\overline{x'}-\overline{X}}{\overline{X}}, \\ e_2 = \frac{\overline{z}-\overline{Z}}{\overline{Z}}, \\ e'_2 = \frac{\overline{z'}-\overline{Z}}{\overline{Z}}, \end{cases}$$

hence

$$\begin{cases} E(e_o) = E(e_1) = E(e'_1) = E(e_2) = E(e'_2) = 0 \\ E(e_0^2) = \theta C_y^2 E(e_1^2) = \theta C_x^2, \\ E(e_2^2) = \theta C_z^2, E(e'^2_1) = \theta_1 C_x^2, \\ E(e_1 e'_1) = \theta_1 C_x^2, E(e_o e'_2) = \theta_1 C_{yz}, \\ E(e_o e_1) = \theta C_{yx}, \\ E(e_o e'_1) = \theta_1 C_{yx}, \\ E(e_o e_2) = \theta C_{yz}, \\ E(e_1 e'_2) = E(e'_1 e'_2) = E(e'_1 e_2) = \theta_1 C_{xz}, \\ E(e_1 e_2) = \theta C_{xz}, \\ E(e'^2_2) = E(e_2 e'_2) = \theta_1 C_z^2. \end{cases}$$

Now, the MSE($t_{ae}$) is given as

$$\begin{aligned}\text{MSE}(t_{ae}) = {} & \overline{Y}^2\theta C_y^2 + k_1^2\overline{X}^2(\theta_1 C_z^2 + \theta_2 C_x^2) + k_2^2\overline{Z}^2(\theta C_z^2 + \theta_2 C_x^2 + 2\theta_2 C_{xz}) - 2k_1k_2\overline{XZ}(\theta_2 C_x^2 - \theta_1 C_z^2 + \theta_2 C_{xz}) \\ & - 2k_1\overline{XY}(\theta_2 C_{yx} - \theta_1 C_{yz}) + 2k_2\overline{ZY}(\theta_2 C_{yx} + \theta C_{yz}). \end{aligned} \tag{13}$$

To find the minimum mean squared error of the estimator $t_{ae}$, we differentiate (13) with respect to $k_1$ and $k_2$ respectively and putting it equal to zero, that is

$$\frac{\partial(\text{MSE}\,(t_{ae}))}{\partial k_1} = 0 \quad \text{and} \quad \frac{\partial(\text{MSE}\,(t_{ae}))}{\partial k_2} = 0,$$

$$k_{1(opt)} = \frac{\overline{Y}(\overline{X}^2 CB - \overline{Z}^2 DE)}{\overline{X}(\overline{X}^2 AB - \overline{Z}^2 E^2)} \text{ and } k_{2(opt)} = \frac{\overline{YZ}(EC - AD)}{(\overline{X}^2 AB - \overline{Z}^2 E^2)},$$

where

$$\begin{cases} A = \theta_1 C_z^2 + \theta_2 C_x^2, \\ B = \theta C_z^2 + \theta_2 C_x^2 + 2\theta_2 C_{xz}, \\ C = \theta_2 C_{yx} - \theta_1 C_{yz}, \\ D = \theta_2 C_{yx} + \theta C_{yz}, \\ E = \theta_2 C_x^2 - \theta_1 C_z^2 + \theta_2 C_{xz}. \end{cases}$$

When substituting the optimum values of $k_1$ and $k_2$ in Equation (13), the minimum MSE($t_{ae}$) is derived as:

$$\text{MSE}\,(t_{ae})_{min} = \overline{Y}^2 \left[\theta C_y^2 - \left(\frac{\overline{Z}^2 AD + \overline{X}^2 C^2 B - 2\overline{Z}^2 CED}{\overline{X}^2 AB - \overline{Z}^2 E^2}\right)\right].$$

## 4. Comparison of efficiency

In this section, the proposed estimator is compared with other existing estimators.

1. By (1) and (13)

$$\text{MSE}\,(t_{ae})_{min} < \text{MSE}(t_o) \text{ if } \left(\frac{\overline{Z}^2 AD + \overline{X}^2 C^2 B - 2\overline{Z}^2 CED}{\overline{X}^2 AB - \overline{Z}^2 E^2}\right) > 0.$$

1. By (11) and (13)

$$\text{MSE}\,(t_{ae})_{min} < \text{MSE}\,(t_m) \text{ if } \left(\frac{\overline{Z}^2 AD + \overline{X}^2 C^2 B - 2\overline{Z}^2 CED}{\overline{X}^2 AB - \overline{Z}^2 E^2}\right) + \frac{AD^2 + BC^2 - 2CDE}{AB - E^2} > 0.$$

1. By (3) and (13)

$$\text{MSE}\,(t_{ae})_{min} < \text{MSE}(t_1) \text{ if } \left(\frac{\overline{Z}^2 AD + \overline{X}^2 C^2 B - 2\overline{Z}^2 CED}{\overline{X}^2 AB - \overline{Z}^2 E^2}\right) + \theta_2(C_x^2 - 2C_{yx}) > 0.$$

1. By (7) and (13)

$$\text{MSE}\,(t_{ae})_{min} < \text{MSE}(t_3) \text{ if } \left(\frac{\overline{Z}^2 AD + \overline{X}^2 C^2 B - 2\overline{Z}^2 CED}{\overline{X}^2 AB - \overline{Z}^2 E^2}\right) + \theta_2\left(C_x^2 - 2C_{yx}\right) + \theta_1\left(C_z^2 - 2C_{yz}\right) > 0.$$

## 5. Numerical comparison

Utilizing the Data set given in [14], the mean square errors (MSE's) together with the percent relative efficiencies (PRE's) of the proposed estimator with respect to $t_0$ is given in Table 1.

## 6. Conclusion

Inferring from Table 1, it shows that the proposed estimator has smaller mean squared error and higher percent relative efficiency than the other existing estimators. Hence, the proposed estimator is efficient and highly recommended for use in practice with respect to difference type estimation.

**Table 1**

| Estimators | MSE's | PRE's |
|---|---|---|
| $t_0$ | 1.7525 | 100 |
| $t_1$ | 1.5032 | 116.59 |
| $t_3$ | 1.2793 | 137.00 |
| $t_5$ | 1.1312 | 154.92 |
| $t_m$ | 0.8206 | 213.56 |
| $t_{ae}$ | 0.6693 | 261.84 |

**Author Contributions:** All authors contributed equally in writing of this paper. All authors read and approved the final manuscript.

**Conflicts of Interest:** "The authors declare no conflict of interest."

## References


[1] Laplace, P. S. (1820). *Théorie analytique des probabilités.* Courcier.

[2] Hansen, M. H., & Hurwitz, W. N. (1943). On the theory of sampling from finite populations. *The Annals of Mathematical Statistics, 14*(4), 333-362.

[3] Sukhatme, B. V. (1962). Some ratio-type estimators in two-phase sampling. *Journal of the American Statistical Association, 57*(299), 628-632.

[4] Srivastava, S. K. (1970). A Two-Phase Sampling Estimator in Sample Surveys. *Australian Journal of Statistics, 12*(1), 23-27.

[5] Chand, L. (1975). *Some ratio-type estimators based on two or more auxiliary variables.* Unpublished Ph.D. dissertation, Iowa State University, Ames 1975.

[6] Cochran, W. G. (1977). *Sampling techniques.* New York: Wiley and Sons, 3.

[7] Kiregyera, B. (1980). A chain ratio-type estimator in finite population double sampling using two auxiliary variables. *Metrika, 27*(1), 217-223.

[8] Kiregyera, B. (1984). Regression-type estimators using two auxiliary variables and the model of double sampling from finite populations. *Metrika, 31*(1), 215-226.

[9] Khare, B. B., Srivastava, U., & Kumar, K. (2013). A generalized chain ratio in regression estimator for population mean using two auxiliary characters in sample survey. *Journal of Scientific Research, 57*, 147-153.

[10] Bahl, S., & Tuteja, R. (1991). Ratio and product type exponential estimators. *Journal of information and optimization sciences, 12*(1), 159-164.

[11] Singh, H. P., Singh, S., & Kim, J. M. (2006). General families of chain ratio type estimators of the population mean with known coefficient of variation of the second auxiliary variable in two phase sampling. *Journal of the Korean Statistical Society, 35*(4), 377-395.

[12] Singh, R., Chauhan, P., Sawan, N., & Smarandache, F. (2011). Improved exponential estimator for population variance using two auxiliary variables. *Italian Journal of Pure and Applied Mathematics, 28*, 101-108.

[13] Singh, B. K., & Choudhury, S. (2012). Exponential chain ratio and product type estimators for finite population mean under double sampling scheme. *Journal of Science Frontier Research in Mathematics and Design Sciences, 12*(6), 0975-5896.

[14] Khan, M., & Al-Hossain, A. Y. (2016). A note on a difference-type estimator for population mean under two-phase sampling design. *SpringerPlus, 5*(1), 1-7.

[15] Singh, G., & Majhi, D. (2014). Some chain-type exponential estimators of population mean in two-phase sampling. *Statistics in Transition new series, Glówny Urzad Statystyczny (Polska), 15*(2), 221-230.