DEVELOPMENT OF ALMOST UNBIASED RATIO TYPE ESTIMATOR USING THE STANDARD DEVIATION WHEN AUXILIARY VARIABLE IS UNKNOWN

Mmaduakor C. O[1]*, Ngwu B. N.[1], Akintunde O. A.[1], Ajewole, O. R.[1]

1. Department of Mathematics, Federal University Oye-Ekiti, Nigeria.
email: chika.mmaduakor@fuoye.edu.ng

Abstract

Some authors had proposed modified ratio estimators of population mean using the population parameter of the auxiliary when the information on the auxiliary is known. These estimators are biased though with smaller mean square error compared to the classical ratio estimator. However, the information on the auxiliary variable may not be available in all cases. In this paper, the use of double sampling strategy was employed to obtain more information on the auxiliary variable and then the almost unbiased ratio estimator of population mean using standard deviation $\hat{S}_x$ is proposed. The mean square error and bias of the developed estimator were derived, as well as the condition under which the developed estimator performs better than the classical ratio estimator. To validate the merits of the proposed estimator over other estimators, an empirical study carried out revealed that the proposed estimator has the smallest bias among the existing estimators considered (0.0645 and 0.0175) for population 1 and population 2 respectively. Similarly, with respect to mean square errors, the proposed estimator has the least among the existing estimators considered (1.076 and 15.6953) for population 1 and population 2 respectively.

Keywords: Population mean, Standard deviation, Double sampling, Bias and Mean squared error.

1. INTRODUCTION

In sample survey practice, it is often the case that the characteristic that is of principal interest to the research may be very expensive to measure. However, another characteristic can be identified which is highly correlated with the first one and relatively inexpensive to measure. These characteristics are referred to as the study variable y and the auxiliary variable x respectively. An auxiliary variable is any variable about which information is available prior to sampling. The use of auxiliary information in sample surveys dates back to Cochran (1977), who used it for estimation of yields of agricultural crops in the course of his researches in agricultural sciences. Since then, there have been tremendous advancements in sampling theory using auxiliary information both at selection and estimation stages respectively.

In practise, when the auxiliary variable and study variable are highly and positively related, and the regression line passes through the origin, ratio estimator is preferred. The aim of this method is to obtain increased precision by taking advantage of the correlation between y and x.

When an estimator is better than the sample mean $\bar{y}$ selected by simple random sampling with or without replacement, the estimator is said to be efficient. The use of classical ratio estimator

ISSN NUMBER: 1116-249X

is more efficient than the sample mean per unit estimator to estimate either the population mean or population total of study variable. The only setback is that the ratio estimator is biased.

Studies like Sisodia and Dwivvedi (1981), Singh et al. (1993) and Singh et al. (2003) have shown that when population moment of the auxiliary variable are known in advance such as coefficient of variation C_x, kurtosis $\beta_2(x)$ and correlation respectively, they can be used to modify the classical ratio estimator to reduce its bias. Before further discussion on the modified estimators and the developed estimator, notations used in this paper are described as follows:

n – Sample size
N – Population size
X - Auxiliary variable
Y - Study variable
$\bar{x}, \bar{y}$ - Sample means of the auxiliary variable and study variable
$\bar{Y}, \bar{X}$ - Population means of the auxiliary variable and study variable
S_Y, S_X - Population standard deviations of the study variable and auxiliary variable
ρ - Coefficient of correlation
C_X, C_Y - Coefficient of variations of the auxiliary variable and study variable

$\beta_1 = \dfrac{N\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^3}{(N-1)(N-2)S_x^{\ 2}}$ - Coefficient of skewness of the auxiliary variable

$\beta_2 = \dfrac{N(N-1)\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^4}{(N-1)(N-2)(N-3)S_x^{\ 2}}$ - Coefficient of kurtosis of the auxiliary variable

$B(.)$ - Bias of the estimator

$MSE(.)$ - Mean square error of the estimator

The classical Ratio estimator for population mean $\bar{Y}$ of the study variable Y is defined as

$$\bar{Y}_{CR} = \frac{\bar{y}}{\bar{x}}\bar{X} = R\bar{X} \qquad 1$$

where $R = \dfrac{\bar{y}}{\bar{x}} = \dfrac{y}{x}$

When the population coefficient of variation, coefficient of kurtosis, coefficient of skewness , population variance of auxiliary variable and the population correlation between X and Y are known, Sisodia and Dwivvedi (1981) , Singh and Kakran (1993), Yan and Tian (2010), Singh (2003) and Singh and Tailor (2003) respectively suggested a modified ratio estimators for $\bar{Y}$.

For the convenience of reading, the estimators, bias and mean square errors (MSE) mentioned above are represented in a class of modified ratio estimators as given below:

ISSN NUMBER: 1116-249X

$$\bar{y}_i = \bar{y}\frac{\bar{X}+\phi_i}{\bar{X}+\phi_i} \quad i = 1,2,3,4,5 \tag{2}$$

The bias and MSE of these estimates are given as follow:

$$B(\bar{y}_i) = \frac{1-f}{n}\left(\theta_i^2 S_x^2 - \theta_i R S_{xy}\right) \tag{3}$$

$$MSE(\bar{y}_i) = \frac{1-f}{n}\left(S_y^2 + \theta_i^2 R^2 S_x^2 - 2\theta_i R S_{xy}\right) \tag{4}$$

where $\phi_1 = C_x, \phi_2 = \beta_2, \phi_3 = \beta_1, \phi_4 = S_x, \phi_5 = \rho$

$$\theta_1 = \frac{\bar{X}}{\bar{X}+C_x},\ f = \frac{n}{N},\ \theta_2 = \frac{\bar{X}}{\bar{X}+\beta_2},\ \theta_3 = \frac{\bar{X}}{\bar{X}+\beta_1},\ \theta_4 = \frac{\bar{X}}{\bar{X}+S_x} \text{ and } \theta_5 = \frac{\bar{X}}{\bar{X}+\rho}$$

The estimators mentioned above are biased but have minimum mean squared error compared to classical ratio estimator. However, availability of the auxiliary variable is not applicable in all cases. Therefore, we use double sampling design to obtain the information on the auxiliary variable and their population parameters, and then use them to develop almost unbiased ratio estimator that are comparable to classical ratio estimator and a class of modified ratio estimator with the aim of minimizing bias and maximizing efficiency.

2. MATERIAL AND METHODS

When information on the auxiliary variable is lacking, and if is convenient and cheap to do, then on the auxiliary variable is collected from large a preliminary large sample, while information on the study variable, y is collected from a second sample which is smaller in size than the preliminary sample. The aim of the preliminary large sample is to furnish a good estimate of $\bar{X}$ or of the distribution of x_i .The sample may be a subsample of the preliminary sample or may be an independent sample selected from the entire population. When the second sample is independent of the preliminary sample, information on both auxiliary and study characters is obtained from the second phase. This procedure is called double or two phase sampling. Mmaduakor et al. (2022) shown that the reduction in variance due to the addition auxiliary worth extra cost required to obtained the auxiliary. This sampling procedure can be extended to more than two phases; giving rise to what is called multiphase sampling.

2.1 Proposed Almost Unbiased Ratio Estimator

Assuming that the population mean $\bar{X}$ and the standard deviation s_x of the auxiliary variable is unknown; we propose a ratio type estimator of $\bar{Y}_p$ as

$$\bar{Y}_p = \bar{y}\left(\frac{\bar{x}' + \hat{s}_x}{\bar{x} + \hat{s}_x}\right) \tag{5}$$

In order to derive the large sample approximation of the bias and MSE of $\bar{Y}_p$, we redefine the $\bar{y}$, $\bar{x}$ and $\bar{x}'$ as follows:

ISSN NUMBER: 1116-249X

$\bar{y} = \bar{Y}(1+e_0)$, $\bar{x} = \bar{X}(1+e_0)$ and $\bar{x}' = \bar{X}(1+e_0')$ such that

$E(e_0) = E(e_1) = E(e_1') = 0$

$E(e_0^2) = f_1 C_y^2$, $E(e_1') = f_2 C_x^2$, $E(e_1^2) = f_1 C_x^2$

$E(e_0 e_1) = f_1 \rho C_x C_y$, $E(e_0 e_1') = f_2 \rho C_x C_y$ and $E(e_1' e_1) = f_2 C_x^2$

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

Expressing (5) in terms of e_i 's, then we have

$$\bar{Y}_p = \bar{Y}(1+e_0)\left[\frac{(1+e_1') + \hat{s}_x}{(1+e_1) + \hat{s}_x}\right]$$

$$= \bar{Y}(1+e_0)\left[\frac{(1+e_1') + \hat{s}_x}{(1+e_1) + \hat{s}_x + e_1 - e_1'}\right]$$

Divide the numerator and denominator by $(1+e_1') + \hat{s}_x$

$$= \bar{Y}(1+e_0)\left[\frac{1}{1 + \frac{e_1 - e_1'}{(1+e_1') + \hat{s}_x}}\right]$$

$$= \bar{Y}(1+e_0)(1+t)^{-1} \qquad 6$$

where $t = \frac{e_1 - e_1'}{(1+e_1') + \hat{s}_x}$

Using the series expansion, $S_\infty = \frac{d}{1-r}, d = 1 \,\&\, r = t$ and $t > 1$.

$$\bar{Y}_p = \bar{Y}(1+e_0)(1+te_1')(1 - te_1' + t^2 e_1'^2 - t^3 e_1'^3 + \ldots) \qquad 7$$

Note that the expected values of all first –degree terms in (7) are zero

Expanding (7) and retaining terms up to second order of e's, we have

$$\bar{Y}_p = \bar{Y}(1+e_0)(1+te_1')(1+te_1)^{-1} \qquad 8$$

From (8) the expectation of $\bar{Y}_p$ becomes

$$E(\bar{Y}_p) = \bar{Y} f_2 (1 + t^2 C_x^2 - tC_{xy}) \qquad 9$$

By expanding $(1+te_1)^{-1}$ up to the first degree and second of approximations, bias and mean squared error of the proposed estimator $\bar{Y}_p$ are obtained respectively as

$$B(\bar{Y}_p) = \bar{Y} f_2 (t^2 C_x^2 - t\rho C_x C_y) \qquad 10$$

ISSN NUMBER: 1116-249X

$$MSE\left(\bar{Y}_p\right)=\bar{Y}^2\left[f_2C_y^2+f_3\left(C_y^2+\varphi^2C_x^2-2\varphi\rho C_xC_y\right)\right] \tag{11}$$

Where $\varphi=\dfrac{\bar{x}'}{\bar{x}'+s_x}$

3.1 OPTIMUM ALLOCATION AND COMPARISON OF $\bar{Y}_p$ WITH SINGLE SAMPLING

Considering the simple cost function,

$$C=c_1n'+c_2n \;;\; c_1<c_2 \tag{12}$$

c_1: is the cost per unit of obtaining information on the auxiliary variable from the first-phase sample.

c_2: is the cost per unit of measuring the study variable from the second-phase sample.

The MSE of $\bar{Y}_p$ in equation (11) can be written in this form

$$\frac{\bar{Y}^2C_y^2}{n_2}-\frac{\bar{Y}C_y^2}{N}+\frac{\bar{Y}^2\varphi C_x^2}{n_2}-\frac{\bar{Y}^2\varphi C_x^2}{n_1}-\frac{2\bar{Y}^2\lambda_i\rho C_xC_y}{n_2}+\frac{2\bar{Y}^2\lambda_i\rho C_xC_y}{n_1} \tag{13}$$

where

$$n_1=n' \text{ and } n_2=n \tag{14}$$

$$G\left(n_1,n_2,\lambda\right)=MSE\left(\hat{\bar{Y}}_p\right)+\lambda\left(c_1n_1+c_2n_2-C\right)$$

Differentiate G with respect to n_1 and n_2 respectively and equate to zero, to obtain

$$\frac{\partial G}{\partial n_1}=\frac{\bar{Y}^2\varphi^2C_x^2}{n_1^2}-\frac{2\bar{Y}^2\varphi\rho C_xC_y}{n_1^2}+\lambda c_1=0 \tag{15}$$

$$\frac{\partial G}{\partial n_2}=-\frac{\bar{Y}^2C_y^2}{n_2^2}-\frac{\bar{Y}^2\varphi^2C_x^2}{n_2^2}+\frac{2\bar{Y}^2\varphi\rho C_xC_y}{n_2^2}+\lambda c_2=0 \tag{16}$$

Solving equations (15) and (16) simultaneously gives

$$n_2=n_1\left[\frac{C_y^2+\varphi^2C_x^2-2\varphi\rho C_xC_y}{2\varphi\rho C_xC_y+\varphi^2C_x^2}\left(\frac{c_1}{c_2}\right)\right]^{1/2} \tag{17}$$

Substitute equation (17) into the cost function (12), to obtain the optimum values of n_1 and n_2 respectively

ISSN NUMBER: 1116-249X

$$n_1 = \frac{C}{c_1 + c_2\left[\frac{C_y^2 + \varphi^2 C_x^2 - 2\rho C_x C_y \varphi}{2\rho C_x C_y \varphi - \varphi^2 C_x^2}\left(\frac{c_1}{c_2}\right)\right]^{\frac{1}{2}}} \tag{18}$$

$$n_2 = \frac{C}{c_1 + c_2\left[\frac{2\rho C_x C_y \varphi - C_x^2 \varphi^2}{C_y^2 + \varphi^2 C_x^2 - 2\rho C_x C_y \varphi}\left(\frac{c_2}{c_1}\right)\right]^{\frac{1}{2}}} \tag{19}$$

Substitute the optimum values of n_1 and n_2 into (11) to obtain the optimum mean square error for $\bar{Y}_p$ as

$$MSE_0\left(\bar{Y}_p\right) = \frac{\bar{Y}^2}{C}\left\{c_2\left(C_y^2 + \varphi^2 C_x^2 - \varphi\rho C_x C_y\right) - c_1\left(2\varphi\rho C_x C_y - \varphi^2 C_x^2\right)\right\} \tag{20}$$

3.2 OPTIMUM FOR CLASSCAL RATIO ESTIMATOR AND A CLASS MODIFIED RATIO ESTIMATORS

The cost function is define as

$$C = c_2 n \tag{21}$$

From equation 3.10

$$n = \frac{C}{c_2} \tag{22}$$

Therefore the optimum mean square error for $\bar{Y}_{CR}$ becomes

$$MES_0\left(\bar{Y}_{CR}\right) = \left(\frac{c_2}{C} - \frac{1}{N}\right)\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right) \tag{23}$$

We assume that N is so large, thus the 1/N can be ignored, then

$$MES_0\left(\bar{Y}_{CR}\right) = \frac{c_2}{C}\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right) \tag{24}$$

For a class of modified ratio estimators, we have

$$MSE_0\left(\bar{Y}_i\right) = \frac{c_2}{C}\bar{Y}^2\left(C_y^2 + \theta_i^2 C_x^2 - 2\theta_i \rho C_x C_y\right) \tag{25}$$

3.3 CONDITIONS FOR OPTIMALITY

In comparison of two estimators e_1 and e_2, when V (e_1) < V (e_2) or MSE (e_1) < MSE (e_2), then we say that e_1 is more efficient than e_2. Hence, to establish the conditions when the developed estimator $\bar{Y}_p$ is better than other existing estimators, we compare the optimum mean squares

ISSN NUMBER: 1116-249X

errors of the developed estimators with the existing ones and classical ratio estimator $\bar{Y}_{CR}$. Therefore, we compare the mean square errors of $\bar{Y}_p$ with $\bar{Y}_{CR}$ and $\bar{Y}_i$.

Let $MSE_0\left(\bar{Y}_{CR}\right) \supseteq MSE_0\left(\bar{Y}_p\right)$ then from equations (20) and (25), we have

$$\frac{c_2}{C}\bar{Y}^2\left(C_y^2 + C_x^2 - 2\rho C_x C_y\right) \supseteq \frac{1}{C}\bar{Y}^2\left\{c_2\left(C_y^2 + \varphi^2 C_x^2 - 2\rho\varphi C_x C_y\right) - \left(2\rho\varphi C_x C_y - \varphi^2 C_x^2\right)\right\} \qquad 26$$

It is observed from (26) that the developed estimator $\bar{Y}_p$ is more efficient than classical ratio estimator $\bar{Y}_{CR}$ if

$$\rho \supseteq \frac{\left(1 - c_1\varphi_j^2 - c_2\varphi_j^2\right)}{2}\left(\frac{C_x}{C_y}\right) \qquad 27$$

4.1 **Empirical study**

In order to compare the performance of the developed ratio estimator with classical and a class of modified ratio estimators, two natural population data were considered. The populations were taken from Murthy (1967) in page 228. The population constants obtained from the data were shown below:

Population 1: MURTHY P (228)

Y= output for factories in a given region

X=fixed capital

$N = 80; n' = 50; n = 20;\ \bar{Y} = 51.8264;\ \bar{X} = 11.3100;\ \rho = 0.9496; s_y = 18.3569;\ s_x = 9.6237$
$c_x^2 = 0.7240;\ c_y^2 = 0.1255;\ \rho c_x c_y = 0.2862;\ \beta_1 = 1.0560;\ \beta_2 = -0.3644$

Population 2: MURTHYP (228)

Y= 0utput for factories in a given region

X= number of workers

$N = 80;\ n' = 50;\ n = 20;\ \bar{Y} = 51.8264;\ \bar{X} = 3.2618;\ \rho = 0.9793;\ s_y = 18.3569;\ c_x^2 = 0.9841$
$;\ c_y^2 = 0.1255;\ \rho c_x c_y = 0.3441;\ \beta_1 = 1.2004;\ \beta_2 = -0.1167$

For each population the mean square errors and efficiency were calculated. The outcomes of the performance of these estimators with respect to mean squared errors and efficiency are presented in tables below:

ISSN NUMBER: 1116-249X

TABLE 1: ESTIMATES OF THE MEAN SQUARE ERROR AND EFFICIENCY OF THE SEVEN ESTIMATORS FOR POPULATION 1

SAMPLE	ESTIMATOR	BIAS	MSE	EFICIENCY
20	Y_{CR}	0.8531	18.6071	1
20	Y_{SD}	0.2944	15.2438	1.22064
20	Y_{SK}	0.2361	13.9405	1.33475
20	Y_{YT}	0.3766	20.6313	0.90189
20	Y_{SS}	0.2995	14.3445	1.29716
20	Y_{T}	0.1009	1.85093	10.0529
20	Y_{P}	0.0645	1.076	17.2928

Table1 shows that the developed estimator $\overline{Y}_p$ is more efficient than the classical ratio estimator $\overline{Y}_{CR}$ and other modified ratio estimators $\overline{Y}_{SD}, \overline{Y}_{SK}$, $\overline{Y}_{YT}$, $\overline{Y}_{SS}$ and $\overline{Y}_T$ considered. In population I the developed estimator $\overline{Y}_p$ is the most efficient followed by $\overline{Y}_T$ while the least efficient is $\overline{Y}_{YT}$. This result agrees with Mmaduakor et al (2018) that double sampling is more efficient than single sampling.

TABLE 2: ESTIMATES OF THE MEAN SQUARE ERROR AND EFFICIENCY OF THE SEVEN ESTIMATORS FOR POPULATION 2

SAMPLE	ESTIMATOR	BIAS	MSE	EFICIENCY
20	Y_{CR}	0.128	43.6538	1
20	Y_{SD}	0.0695	27.2821	1.60009
20	Y_{SK}	0.0735	30.7633	1.41902
20	Y_{YT}	0.0699	46.9336	0.93012
20	Y_{SS}	0.0631	27.9213	1.56346
s20	Y_{T}	0.0293	17.2393	2.53224
20	Y_{P}	0.0175	15.6953	2.78133

ISSN NUMBER: 1116-249X

From the results shown in the table 2 above, it was observed that $\bar{Y}_p$ has the least MSE followed by $\bar{Y}_T$ while $\bar{Y}_{YT}$ has the most among the seven estimators considered under population II.

5.1 **CONCLUSIONS**

In this study, an almost unbiased ratio type estimator was developed using the standard deviation of the auxiliary variable x, when the auxiliary variable is unknown. The mean square error and bias of the developed estimator were derived, as well as the condition under which the developed estimator performs better than the classical ratio estimator. In addition, the performance of the developed estimator with other existing modified ratio estimators for two populations were assessed. The numerical comparisons reveal that with respect to bias, mean square errors and efficiency the proposed estimator is consistently the best among the existing estimators considered. Therefore, we conclude that the developed almost unbiased ratio estimator have been able to reduce bias and also lead gain in efficiency over the existing estimators considered. Hence, the proposed estimator is a good alternative in situations where auxiliary variable is not unavailable before the survey.

References

1. Murthy M.N. (1967): sampling theory and methods, Statistics Publishing Society, Calcutta, India.
2. Cochran, W.G. (1977): sampling Techniques New York John Wiley.
3. Sisodia B.V. S and Dwivedi V. K. (1981): A modified ratio estimator using coefficient of variation of auxiliary variable. Journal of Indian Society of Agricultural Statistics 33 (1), 13 -18.
4. Singh H.P and Kakran M. S. (1993): A modified ratio estimator using known coefficient of Kurtosis of an auxiliary character, revised version submitted to Journal of Indian Society of Agricultural Statistics.
5. Singh G.N. (2003): on the improvement of product method of estimation in sample surveys, Journal of Indian Society of Agricultural Statistics 56 (3), 265 - 267
6. Singh H. P. and Tailor R. (2003): Use of known correlation coefficient in estimating the finite population means Statistics in Transition 6 (4), 555 – 560.
7. Khoshnevisam M., Singh R., Chaauhan P., Sawan N. and Smarandache F. (2007): A general family of estimators for estimating population mean using known value of some population parameter(s), Far East Journal of Theoretical Statistics 22, 181 -191.
8. Yan Z and Tian B. (2010): Ratio Method to the mean Estimation Using Coefficient of Skewness of Auxiliary Variable. ICICA 2010, Part II, CCIS 106, pp. 103 110.
9. Mmaduakor C.O, Amahia G.O., Ajewole O.R., Ngwu B.A. (2018): On Efficiency of Double Sampling over Single Sampling. FUOYE Journal of Pure and Applied Science, 3(1): 278-280.
10. Mmaduakor C., Ibeji J.U., Amahia G., Ajewole O.R and Ngwu B. (2022): Regression Type Estimator with two Auxiliary Variable cost Function. Journal of Innovation Science and Technology, Vol.2 (1): 89 – 96.

ISSN NUMBER: 1116-249X