ISSN No:-2456-2165

# Comparative Study of Ratio and Regression Estimators using Double Sampling for Estimation of Population Mean

[1]Oke Samuel A, [2]Adesina Oluwaseun A, [3]Oladimeji Lukman A, [4]Akinade Oludayo O, [5]OguntolaToyin O , [6]Tijani Rokibat A and [7]Adegoke Maryam A
Department of Statistics, Ladoke Akintola University of Technology, Ogbomoso, Nigeria.

**Abstract:- This study aims to explore the preference order regarding the utilization of different estimation methods in sample surveys. Through empirical analysis, the research examines both the conventional simple random sampling without replacement estimator and the efficiency of double sampling for ratio and regression estimators. The objective is to identify the methods that the estimator is the most efficient.**

**Double sampling procedure was adopted, and comparing the minimum variances empirically which was used to obtain the most efficient estimator using the data collected from the variable of interest (expenditure) and the auxiliary variable (salary). In the first phase, a sample size of (150, 120, 80, and 60) was chosen from the population and in the second phase a subsample of size (70, 55, 45, and 30) was selected from the first phase, each at four different levels (1, 2, 3 and 4) without replacement.**

**Of the three sampling methods, namely double sampling for ratio estimator, simple random sample without replacement, and double sampling for regression estimator, the last one shows the least variability, making it the most effective estimator in terms of efficiency. Consequently, when the auxiliary variable is accessible, it is advisable to utilize the double sampling for regression method in order to enhance the accuracy of estimating the population parameter.**

***Keywords:-*** *Double sampling, ratio estimator, regression estimator, simple random sampling without Replacement, minimum variances.*

## I. INTRODUCTION

Sampling theory endeavors to achieve reasonably accurate outcomes regarding the population parameters of the study variable, denoted as "y," through the acquisition of random samples. Commonly, the incorporation of data from auxiliary variables is practiced to derive estimators with improved efficiency.

When there is a lack of supplementary information on the auxiliary variable, the Simple Random Sampling Without Replacement (SRSWOR) stands as the most straightforward method for estimating the population mean. However, if an auxiliary variable "x" that exhibits a close relationship with the study variable "y" is accessible, it becomes possible to enhance the estimator's performance by leveraging the known values of the population parameters associated with the auxiliary variable.

Double sampling is a design where a preliminary sample of units is chosen solely for the purpose of gathering supplementary data, followed by the selection of a second sample where the variable of interest is observed along with the auxiliary information. This technique is also known as two-phase sampling and is valuable for acquiring auxiliary variables to estimate ratios and regression. One of the benefits of double sampling is its ability to enhance the accuracy of an estimate by utilizing auxiliary information.

Sampling methods can be categorized into two primary types: probability sampling and non-probability sampling. Probability sampling ensures that each individual within the population has a recognized and nonzero likelihood of being chosen for the sample. On the other hand, non-probability sampling does not guarantee that every member of the population has a known probability of being selected for the sample.

Top of Form Double sampling was initially introduced by Neyman in 1938, while Rao explored its application in stratification and analytic studies in 1973. Cochran, in 1977, provided fundamental findings on two-phase sampling, encompassing the most straightforward regression estimators for this particular sampling design.

According to Lorh S.L. (2010), a straight line across the origin provides the best fit for the data when ratio estimation is used. Also, in the realm of ratio and regression estimation using double sampling, noteworthy contributions have been made by Okafor and Lee (2000), Sodipo and Obisesan (2007), and Kumar et al. (2011), among other authors.

Ogunyinka and Sodipo (2013) proposed the utilizing double sampling technique to find the most effective estimator among ratio and regression estimators, as well as SRSWOR estimators. They performed a statistical analysis by comparing the minimum variances, relative efficiency, and coefficients of variation in order to determine the estimators that are most efficient. The results revealed that if the origin of the regression line does not intercept, double sampling for linear regression estimation outperforms double sampling for ratio estimation and simple random sampling without replacement estimation in terms of efficiency.

Singh and Vishwakarma (2007) conducted a study focusing on the exponential ratio and product estimators for estimating the finite population mean using auxiliary information within the framework of double sampling. The authors analyzed the properties of these estimators and

compared their precision to that of the simple mean per unit estimator, as well as the usual double sampling ratio and product estimators. The results clearly indicate that the ratio and product estimators, denoted as ($\bar{Y}$)ReMd and ($\bar{Y}$)PeMd, exhibit greater efficiency compared to ("$\bar{Y}$"), ("$\bar{Y}$")Rd, and ("$\bar{Y}$")Pd. The study also observed that the performance of the proposed estimators surpasses that of Case-II, except for population V. Based on these findings, the authors recommend the adoption of the suggested estimators in practical applications.

Bhushan et al. (2011) introduced a more comprehensive class of double sampling regression estimators that incorporates the available auxiliary information from the initial phase, specifically the mean and variance of the auxiliary variable. The authors derived the bias and mean squared error (MSE) for this class of estimators. The concluding remarks highlight that the proposed class of estimators outperforms the previously suggested double sampling estimator.

Olanrewaju et al (2023) propose a new ratio type estimator for double sampling with two auxiliary variables. The terms for bias and mean square error (MSE) of the proposed estimator have been obtained. Some realistic conditions have been obtained under which the proposed estimator is more efficient than the usual unbiased well-known existing estimators of double sampling using two auxiliary variables for population mean and it was found to be more efficient in many situations.

Toluwalase and Okafor (2018) proposed a class of ratio type estimators in double sampling using an auxiliary variable with known population parameters. The regression type ratio estimator within the proposed class of double sampling, when used at its optimum value, stands out as the most efficient estimator due to its minimal Mean Squared Error (MSE) and highest Percent Relative Efficiency (PRE). Moreover, it exhibits an equal level of efficiency to that of the traditional linear regression in the context of double sampling.

This research aims to assess the hierarchical preference of these estimators through an empirical analysis of their variances in double sampling for ratio and regression estimation, along with the conventional simple random sampling without replacement. The objective is to identify the estimator that are most efficient among them.

## II. MATERIALS AND METHODOLOGY

The research design adopted for this study is double sampling for ratio and regression estimators and simple random sampling without replacement. This study utilizes data acquired from four hundred and seventeen (417) questionnaires. The questionnaires were distributed to the civil servants of the North East local government Ibadan, Oyo state of Nigeria. The double sampling technique involves using household monthly food expenditure consumption (in thousands of Naira) as primary study variable (y), and the household's earned salary as the auxiliary variable (x). The process of double sampling generate two separate sample sizes at four different levels, denoted as $n'$ for the first phase sample size and n for the second phase sample size, as outlined below:

Table 1: An overview of the sample sizes for both the first and second phase at different various levels

| *Levels* | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| n′ | 150 | 120 | 80 | 60 |
| $n$ | 70 | 55 | 45 | 30 |
| $\frac{n}{n'}$ | 0.4667 | 0.4583 | 0.5625 | 0.5000 |

A simple linear regression analysis on the second phase data (n=55) was performed using R software and the model obtained is presented in the equation below:

$$\hat{y} = 56430 + 0.2786x + e$$

From the estimated model above, the intercept $\beta_O$ is 56430 indicating that the y-axis intercept is non-zero. It implies that the data becomes appropriate for double sampling in regression- type estimation.

### A. Ratio Estimator

The ratio estimator, a statistical parameter, represents the ratio between the means of two random variables. However, when utilizing ratio estimates in experimental or survey research, adjustments are necessary due to their skewed nature. Given the asymmetry of ratio estimates, it is not advisable to construct confidence intervals using symmetric tests such as the t-test. Suppose we are required to estimate the ratio of y to x, then:

$$\hat{R} = \frac{\bar{y}}{\bar{x}} \quad (2.1)$$

Where:

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

But R is biased and hence, the mean square error (MSE) of R is giving as

$$m(\hat{R}) = E\left[\hat{R} - R\right]^2 \quad (2.2)$$

The estimated variance of the estimated ratio is given as;

$$v(\hat{R}) = \left(\frac{1-f}{n\bar{x}}\left[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}\right]\right) \quad (2.3)$$

Where $f = \frac{n}{N}$

$$S_y^2 = \frac{\sum_{i=1}^{n}(y-\overline{y})^2}{n-1}$$

$$S_x^2 = \frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1}$$

$$S_{xy} = \frac{\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})}{n-1}$$

Where $f =$ is the sampling fraction, $S_x^2$ is the variance of at the first phase, $S_y^2$ is the variance of at the second phase and $s_{xy}$ is the covariance of x and y.

According to Mukhopadhay P. (2007), the estimated ratio mean estimator (using SRSWOR) is

$$\hat{\overline{y}}_r = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{X} \quad (2.4)$$

The corresponding estimated variance of the $y_r$ is given as:

$$\hat{v}(\overline{y}) = \left(\frac{1-f}{n}\left[sy^2 + \hat{R}^2 sx^2 - 2\hat{R}s_{xy}\right]\right) \quad (2.5)$$

*B. Classical Linear Regression*

Let $y_i, x_i$(i $= 1, 2, \ldots,$ n) be the sample values of the main character y and the auxiliary character x respectively obtained with SRSWOR of sample size n from the population size N. The linear regression estimator of the mean as given by Mukhopadhyay P (2007).

$$\overline{y}_l = \overline{y} + \hat{\beta}(\overline{X} - \overline{x}) \quad (2.6)$$

Where

$$\hat{\beta} = \frac{s_{xy}}{sx^2}$$

$\overline{X} =$ population mean
$\overline{x} =$ mean of the auxiliary information
$\overline{y} =$ mean of the study variable

The estimated MSE of $\overline{y}_l$ is given as;

$$\hat{v}\left(\overline{y}_l\right) = \left(\frac{1-f}{n}\left[sy^2 + \hat{\beta}^2 sx^2 - 2\hat{\beta}s_{xy}\right]\right) \quad (2.7)$$

*C. Simple Random Sampling Without Replacement (Srswor)*

According to Thompson (1992), simple random sampling involves selecting n units from a total population of N in such a way that each possible distinct sample of size n has an equal chance of being chosen, given that $(N > n)$. If a selected item has an equal probability of being chosen again multiple times, it is referred to as simple random sampling with replacement (SRSWR). On the other hand, if an item cannot be selected again once it has been chosen, it is known as simple random sampling without replacement (SRSWOR).

The corresponding sample mean $\overline{y}$ is given as:

$$\overline{y} = \frac{\sum_{i=1}^{n} yi}{n} \quad (2.8)$$

And the estimated variance of $\overline{y}$ is given as:

$$\hat{v}(\overline{y}) = \left(\frac{1-f}{n}\right)s^2 \quad (2.9)$$

Where $s^2 = \frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1}$ and $f = \frac{n}{N}$

*D. Double Sampling For Ratio Estimator*

The general framework for two-phase sampling is presented by Sarndal and Swenson (1987) and Legg and Fuller (2009). If our objective is to estimate the population mean of a variable y using the ratio estimation procedure, we first take a large preliminary sample, denoted as n', using simple random sampling without replacement (SRSWOR) from a population of size N. Information obtained from this phase is used to estimate the population mean, $\overline{X}'$. In the second phase, a subsample of size n (where n $< n'$ ) is selected from the units identified in the first phase, again using SRSWOR. Information on both y and x is obtained in this second stage sampling, which allows us to estimate the population means $\overline{y}$ and $\overline{x}$ above as mentioned earlier. Hence the estimated population mean is given as:

$$\overline{y}_{dr} = \hat{R}\overline{x}' \quad (2.10)$$

Where $\overline{x}'$ is the first sample mean

The corresponding estimated variance of $\overline{y}_{dr}$ is given as;

$$\hat{v}(\overline{y}_{dr}) = \left(\left[\frac{1}{n'} - \frac{1}{N}\right]sy^2 + \left[\frac{1}{n} - \frac{1}{n'}\right]\left[sy^2 + \hat{R}^2 sx^2 - 2\hat{R}s_{xy}\right]\right) \quad (2.11)$$

*E. DOUBLE SAMPLING FOR REGRESSION ESTIMATOR*

If one intends to utilize the double sampling for regression estimation instead of the ratio estimation procedure, it is necessary for there to be a linear association between the study variable (y) and the auxiliary variable ($x'$).

The double sampling linear regression estimator of population mean is given as;

$$\overline{y}_{dl} = \overline{y} + \hat{\beta}(\overline{x}' - \overline{x}) \quad (2.12)$$

$\hat{\beta}$= estimated simple linear regression coefficient
$\overline{x}'$= sample mean at the first phase

Hence, the corresponding estimated variance for $\overline{y}_{dl}$ is given as:

$$\hat{v}(\overline{y}_{dl}) = \left(\left[\frac{1}{n'} - \frac{1}{N}\right]sy^2 + \left[\frac{1}{n} - \frac{1}{n'}\right]\left[sy^2 + \hat{\beta}^2 sx^2 - 2\hat{\beta}s_{xy}\right]\right) \quad (2.13)$$

ISSN No:-2456-2165

## III. RESULT AND INTERPRETATION

### *A. Descriptive Statistics Of The Data*

Table 2: Descriptive statistics for expenditure

| Minimum | 1st quartile | Median | Mean | 3rd quartile | Maximum |
|---|---|---|---|---|---|
| 28000 | 70000 | 78000 | 77267 | 85500 | 126500 |

Table 3: Descriptive statistics for salary

| Minimum | 1st quartile | median | Mean | 3rd quartile | Maximum |
|---|---|---|---|---|---|
| 38111 | 58851 | 79761 | 82976 | 98651 | 191371 |

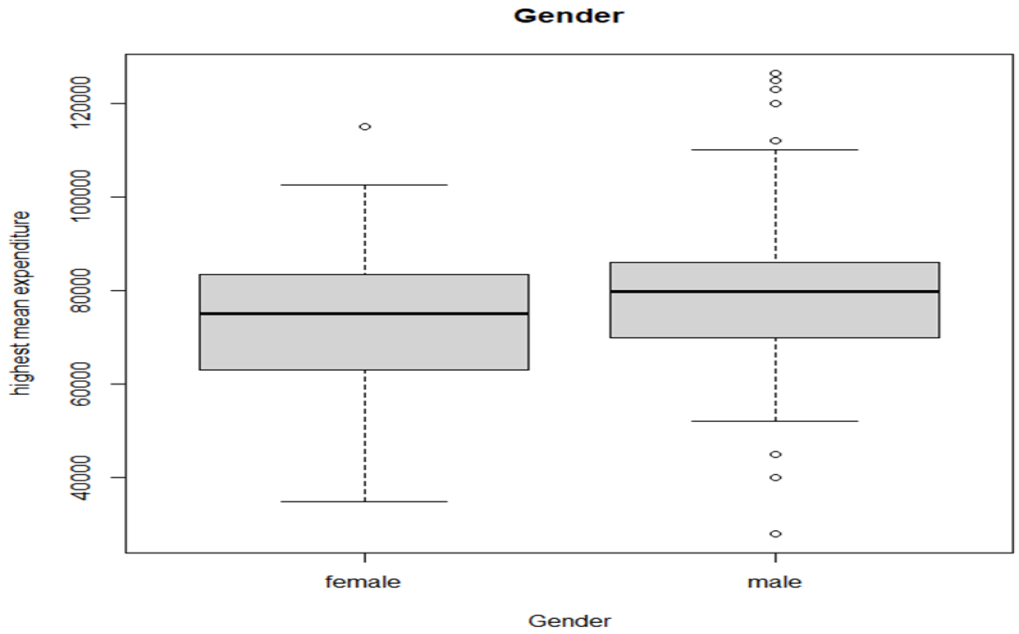


Fig. 1: Boxplot for salary and gender

Figure 1 illustrates the differences in expenditure between male and female individuals which indicate that there are more extreme values among males i.e., the males spend more than the female. Also, the distribution of monthly expenditure on food consumption for households in the data set is skewed to the right which means that there are more households that spend less on food consumption than households that spend more. The whiskers are also relatively long, which means that there is a lot of variation in the data where some households spend a lot on food consumption, while others spend very little.

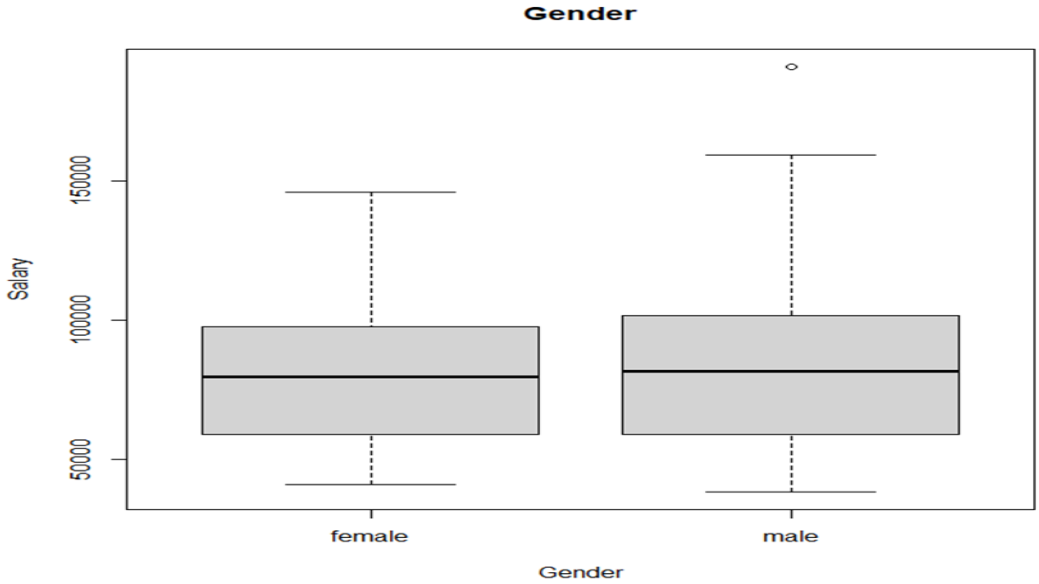


Fig. 2: Boxplot for salary and gender

ISSN No:-2456-2165

Figure 2 illustrates the disparities in salary between male and female individuals which indicates that there is no much significance between salaries being paid to male and female i.e., the males are being paid a more than the female. Also, the distribution of household for both male and female monthly salary earned, slightly skewed to the right. However, the whiskers are relatively long, which indicates that there is a significant amount of variation in the data, with some households earning significantly less than this amount (median) and some households earning significantly more than this amount (median)

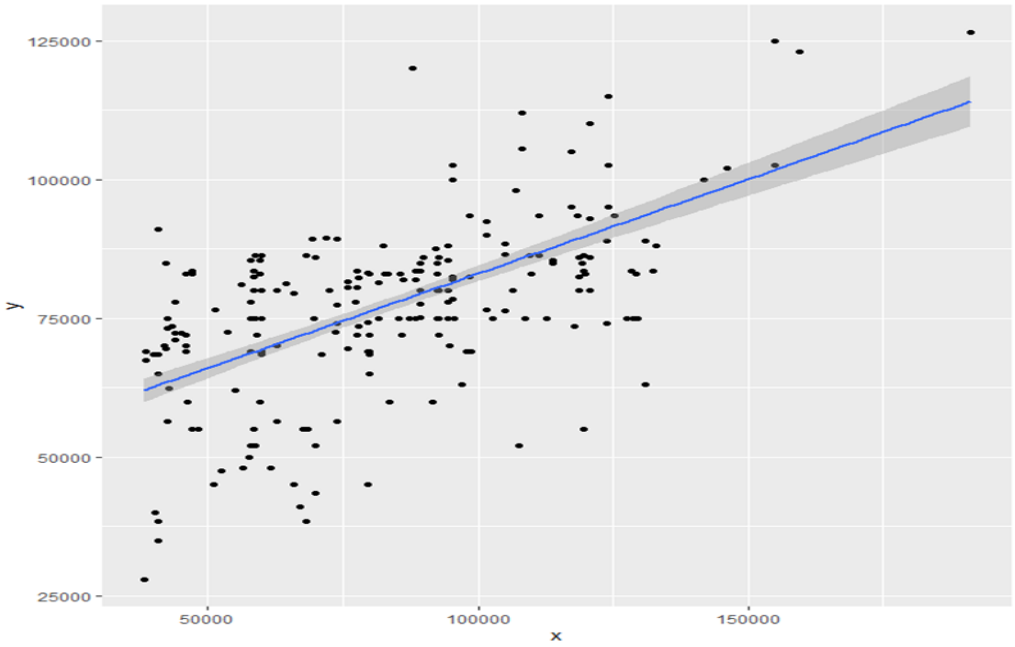


Fig. 3: The Scattered plot of study variable (expenditure) against the auxiliary variable (salary)

Figure 3 illustrates the relationship between the study variable (expenditure) and the auxiliary variable (salary) such that the regression line of y on x does not pass through the origin. Hence, therefore we use double sampling for regression estimation.

*B. Estimation Of The Variances For The Estimators*

Similarly, table 4 shows the summary of the variance obtained at each respective level for SRSWOR

Table 4: Summary of the variance obtained using SRSWOR

| Level | $n'$ | $\boldsymbol{n}$ | $v(\overline{y})$ |
|---|---|---|---|
| 1 | 150 | 70 | 8875121.8 |
| 2 | 120 | 55 | 9109391 |
| 3 | 80 | 45 | 12640870 |
| 4 | 60 | 30 | 13641870 |

Similarly, table 5 shows the summary of the variance obtained at each respective level for double sampling ratio type estimator.

Table 5: Summary of the variance obtained in double sampling for ratio estimator

| Level | $n'$ | $n$ | $v(y_r)$ |
|---|---|---|---|
| 1 | 150 | 70 | 4366305 |
| 2 | 120 | 55 | 5232727 |
| 3 | 80 | 45 | 6604767 |
| 4 | 60 | 30 | 7257414 |

Similarly, the table 6 shows the summary of the variance obtained at each respective level for regression type estimator

Table 6: Summary of the variance obtained in double sampling for regression

| Level | $n'$ | $n$ | $v(y_l)$ |
|---|---|---|---|
| 1 | 150 | 70 | 4031085 |
| 2 | 120 | 55 | 4679222 |
| 3 | 80 | 45 | 6805350 |
| 4 | 60 | 30 | 5400094 |

Also figure 4 below shows the Line graph for the variance of the combined Estimators.

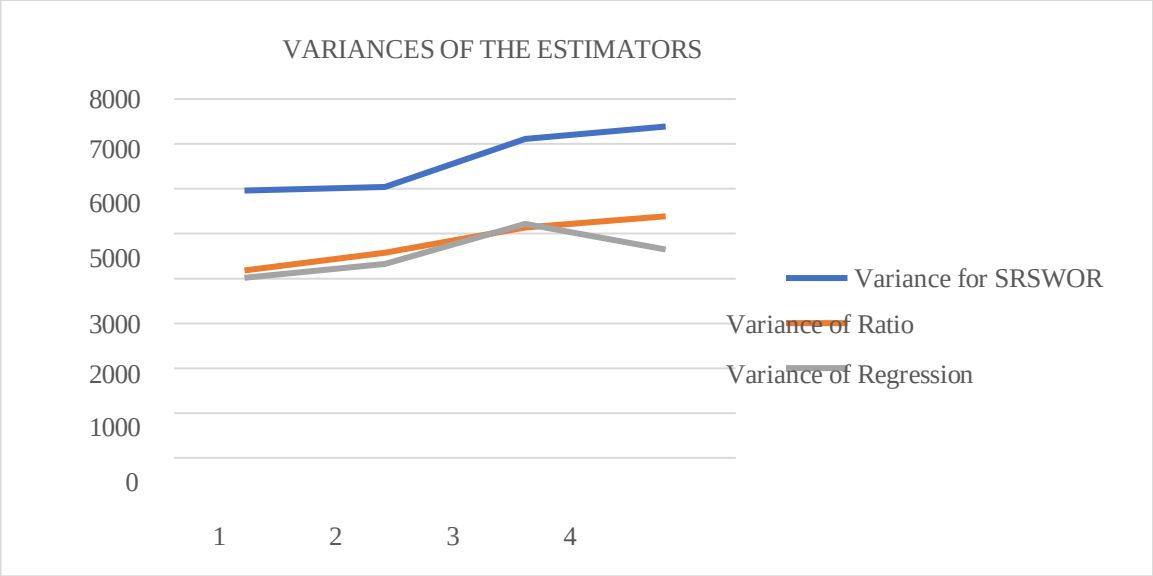


Fig. 4: Line graph for the variance of the combined Estimators

## IV. REMARKS

The primary aim of this study was to take into consideration the double sampling procedures to determine the efficient estimator among the double sampling of ratio estimators, double sampling of regression estimators and simple random sampling without replacement. The minimum variance was the criteria used to determine the efficient estimator.

The variances of the regression for double sampling estimates from table 6 at level 1, 2, 4 are (4031085,4679222,5400094) respectively. The variances of the ratio for double sampling estimates from table 5 at each level 1,2,4 are (4366305, 5232727,7257414). Hence the estimates of regression for double sampling gives the minimum variances when compared to that of the double sample for ratio estimates.

Also the variances of the simple random sampling without replacement estimates from table 4 at each level 1, 2, 4 are (8875121.8,9109391, 13641870) respectively. Also the estimates of regression for double sampling gives the minimum variances when compared to that of the

SRSWOR estimates.

## V. CONCLUSION

Since the double sampling for regression estimator gives the minimum variances when compared to the double sampling for ratio estimator and SRWWOR, it is therefore established that the double sampling for regression it is more efficient than the other estimators. For practical purposes, one would recommend the use of double sampling for regression to improve on the estimate of the population parameter when the auxiliary variable is available.

## REFERENCES

[1.] Cochran W. G (1977), "Sampling Technique", 3rd Edition, John Willey and sons Inc., New York.
[2.] Bhushan, S., Singh, R. K., & Pandey, A. (2011). Some generalized classes of double sampling regression type estimators using auxiliary information. *Science Vision*, *11*(1), 2-6.
[3.] Kumar, S., & Bhougal, S. (2011). Estimation of the population mean in presence of non-response. *Communications for Statistical Applications and Methods*, *18*(4), 537-548.
[4.] Sarndal, C.E. and Swensson, B. (1987), "A general view of estimation for two phases of selection with applications to two-phase sampling and non-response", International Statistical review, 55, 279-294.
[5.] Legg, J.C. and Fuller, W. A. (2009), "Two-phase Sampling", In D. Pfeffermann and C.R. Rao (Eds), Handbook of Statistics: Vol. 29A. Sample Survey: Design, Methods and aplications, 55-70.
[6.] Ogunyinka, P. I., & Sodipo, A. A. (2013). Efficiency of Ratio and Regression Estimators using Double sampling. *Journal of Natural Sciences Research. ISSN*, 2224-3186.
[7.] Mukhopadhyay, P. (2007), "Survey Sampling", Narosa Publishing House Pvt. Ltd, First Edition. 256.
[8.] Neyman J. (1938), "Contribution to the theory of Sampling Human Populations", Journal of the American Statistical Association. 33, 101-116
[9.] Okafor F. C., Lee H. (2000), "Double Sampling for Ratio and Regression Estimation with Sub-sampling the Non-respondents", Survey Methodology, Vol. 26, No 2, 183-188.
[10.] Sodipo A. A. and Obisesan K. O. (2007), "Estimation of the population mean using difference Cum Ratio estimator with full response on the auxiliary character", Research of Applied Sciences 2(6): 769-772.

[11.] Singh, H. P., & Vishwakarma, G. K. (2007). Modified exponential ratio and product estimators for finite population mean in double sampling. *Austrian journal of statistics*, *36*(3), 217- 225.
[12.] Olanrewaju, F., Esther, A., & Oluwadare, A (2023). A New Ratio Type Estimator for Double Sampling with Two Auxiliary Variables.
[13.] Thompson, S. K. (1992), "Sampling", John Wiley & Sons, New York
[14.] Sharon L. Lohr (2010), "Sampling Design and Analysis", Second Edition. Brooks/Cole Cengage Learning. 596.
[15.] Rao, C. R. (1973). Representations of best linear unbiased estimators in the Gauss-Markoff model with a singular dispersion matrix. *Journal of multivariate analysis*, *3*(3), 276-292.
[16.] Toluwalase J. Akingbade and Fabian C. Okafor (2018), " A class of ratio type estimators in double sampling using an auxiliary variable with some known population parameters", Journal of the Nigerian Statistical Association, Vol. 30, 14-29.