Journal of Natural Sciences Research www.iiste.org
ISSN 2224-3186 (Paper) ISSN 2225-0921 (Online)
Vol.3, No.7, 2013

Efficiency of Ratio and Regression Estimators Using Double Sampling

Ogunyinka, Peter I.[1*] Sodipo, A. A.[2]

1. Department of Mathematical Sciences, Olabisi Onabanjo University, Ago-Iwoye, Ogun State, Nigeria.
2. Department of Statistics, University of Ibadan, Ibadan, Oyo State, Nigeria.

* E-mail of the corresponding author: pixelgoldprod@yahoo.com

Abstract

This research work takes into consideration the Double sampling procedure to determine the efficient estimator among the double sampling for ratio and regression and Simple random sampling without replacement (SRSWOR) type estimators. The empirical comparison of the minimum variances, relative efficiency and the coefficient of variations were used in obtaining the most efficient estimator. It was established that, if the regression line does not pass through the origin, then double sampling for linear regression type estimation is efficient over double sampling for ratio and simple random sampling without replacement estimations.

Keywords: Ratio Estimator, Regression Estimator, Double Sampling and Simple Random Sampling Without Replacement

1. Introduction

Over the years, Samplers have been interested in methods to improve the precisions of estimators of population parameter both at the selection and estimation stages by making use of auxiliary information. In addition to estimating the population mean, total and proportion, population ratio of two characters is another interest. If it is known that the regression line of the variable of interest y on auxiliary variable does pass through the origin, the ratio type estimator may be used in estimating the population properties, otherwise the regression type estimator may be considered. However, if this auxiliary information is lacking and it is convenient and cheap to do, the information on auxiliary variable is collected from a preliminary large sample using simple random sampling without replacement (SRSWOR).While the information on the variable of interest y is collected from a second sample using SRSWOR which is smaller in size than the preliminary sample size. This sampling method is known as double sampling. One of the uses of double sampling is to make use of auxiliary information to improve the precision of an estimate. Neyman (1938) was first to propose double sampling, Rao (1973) studied double sampling in context of stratification and analytic studies and Cochran (1977) presented the basic result of two-phase sampling, including the simplest regression estimators for this type of sampling design. Lorh S.L. (2010) concluded that ratio estimation works best if the data are well fitted by a straight line through the origin. Among other authors that have contributed to the use of double sampling for ratio and regression estimators are Okafor and Lee (2000), Sodipo and Obisesan (2007) and Kumar et. al. (2011). This paper investigates the order of preference for the use of each of these estimators by empirically comparing the efficiency of double sampling for ratio and regression estimators and the usual Simple random sampling without replacement estimator to establish the most efficient estimator.

2. Methodology

2.1 The Ratio Estimator

Lorh S.L. (2010) concluded that ratio and regression estimation both take advantage of the correlation of x and y in the population; the higher the correlation, the better they work. Lorh S.L. (2010) similarly concluded that ratio estimation works best if the data are well fitted by a straight line through the origin. Suppose we are required to estimate the ratio of y to x, then.

$$\hat{R} = \frac{\overline{y}}{\overline{x}} \tag{1}$$

Where $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

$$\widehat{R} = \frac{Y}{X} \tag{2}$$

Where $Y = \sum_{i=1}^{n} y_i$ and $X = \sum_{i=1}^{n} x_i$. However, $\hat{R}$ is biased, hence, the mean square error (MSE) of $\hat{R}$ is giving as

$$M(\hat{R}) = E[\hat{R} - R]^2$$

$$M(\hat{R}) = E\left[\left(\frac{\overline{y} - R\overline{y}}{\overline{X}}\right)^2 (1 - 2\partial x + \partial^2 x + \cdots)\right]$$

$$V(\hat{R}) = \frac{1-f}{n\overline{X}^2}[S_y^2 + R^2 S_y^2 - 2RS_{xy}] \quad (3)$$

The estimated variance of the estimated ratio is giving as;

$$\hat{V}(\hat{R}) = \left(\frac{1-f}{n\overline{X}}[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}]\right) \quad (4)$$

Where $f = \frac{n}{N}$; $s_y^2 = \frac{\sum_{i=1}^{n}(y-\overline{y})^2}{n-1}$; $s_x^2 = \frac{\sum_{i=1}^{n}(x-\overline{x})^2}{n-1}$; and $s_{xy} = \frac{\sum_{i=1}^{n}(x-\overline{x})(y-\overline{y})}{n-1}$

Where f is the sampling fraction, s_x^2 is the variance of at the first phase, s_y^2 is the variance at the second phase and s_{xy} is the covariance of *x* and *y*. Mukhopadhyay P. (2007) estimated ratio mean estimator (using SRSWOR) as

$$\hat{\overline{y}}_r = \left(\frac{\overline{y}}{\overline{x}}\right)\overline{X} \quad (5)$$

The corresponding estimated variance of the $\hat{\overline{y}}_r$ is giving as:

$$\hat{V}(\overline{y}) = \left(\frac{1-f}{n}[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}]\right) \quad (6)$$

2.2 The Regression Estimator

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

$$\overline{y}_l = \overline{y} + \hat{\beth}(\overline{X} - \overline{x}) \quad (7)$$

Where $\hat{\beth} = \frac{s_{xy}}{s_x^2}$; $\overline{X}$ *is the population mean;* $\overline{x}$ *= mean of the auxiliary information; and* $\overline{y}$ *= mean of the study variable.*

Similarly, the estimated MSE of $\overline{y}_l$ is giving as:

$$\hat{V}(\overline{y}_l) = \left(\frac{1-f}{n}\left[s_y^2 + \hat{\beth}^2 s_x^2 - 2\hat{\beth}s_{xy}\right]\right) \quad (8)$$

2.3 Double Sampling For Ratio Estimator

A general framework for two-phase sampling is giving in Sarndal and Swensson (1987) and Legg and Fuller (2009). Suppose we are interested in obtaining the population mean of the character *y* using ratio estimation procedure then we take a large preliminary sample n' with SRSWOR from the population of size *N*, information on this phase is obtained to estimate the $\overline{X}'$, the population mean. A subsample (second phase sample) size *n* units (where $n < n'$) is selected from the first phase units by SRSWOR. We have obtained information on *y* and *x* at the second phase sampling to estimate $\overline{y}$ and $\overline{x}$ in equation (1). Hence, the estimated population mean is giving as:

$$\overline{y}_{dr} = \hat{R}\overline{x}' \quad (9)$$

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

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

$$\hat{V}(\overline{y}_{dr}) = \left(\left[\frac{1}{n'} - \frac{1}{N}\right]s_y^2 + \left[\frac{1}{n} - \frac{1}{n'}\right]\left[s_y^2 + \hat{R}^2 s_x^2 - 2\hat{R}s_{xy}\right]\right) \quad (10)$$

The corresponding optimum variance of the $\overline{y}_{dr}$ is giving as;

$$\hat{V}(\bar{y}_{dr})_{opt} = \frac{1}{c_0}\left[c^{\frac{1}{2}}s_{xy} + \left(c'\left(s_y^2 - s_{xy}\right)\right)^{\frac{1}{2}}\right]^2 \quad (11)$$

where

$c =$ Cost per unit for collecting information on the study variable y;

$c_0 =$ Total cost for the survey;

$c' =$ Cost per unit for collecting information on the auxuliary variable x' ; and

$for \;\; c > \; c'$

$$c_0 = cn + c'n' \quad (12)$$

2.4 Double Sampling For Regression Estimator

Assuming double sampling for regression estimation procedure is to be used in place of double sampling for ratio estimation procedure, then there must exists linear relationship between the study variable (y) and the auxiliary variable (x').

The double sampling linear regression estimator of population mean is giving as:

$$\bar{y}_{dl} = \bar{y} + \hat{\beth}\left(\bar{x}' - \bar{x}\right) \quad (13)$$

Where

$\hat{\beth} = estimated\ simple\ linear\ regression\ coefficient; and$;

$\bar{x}' = sample\ mean\ at\ the\ first\ phase$

Hence, the corresponding estimated variance for $\bar{y}_{dl}$ is giving as

$$\hat{V}\left(\bar{y}_{dl}\right) = \left(\left[\frac{1}{n'} - \frac{1}{N}\right]s_y^2 + \left[\frac{1}{n} - \frac{1}{n'}\right]\left[s_y^2 + \hat{\beth}^{\,2}s_x^2 - 2\hat{\beth}\, s_{xy}\right]\right) \quad (14)$$

Similarly, the optimum variance of double sampling for regression estimate is giving as:

$$\hat{V}\left(\bar{y}_{dl}\right)_{opt} = \frac{s_y^2}{c_0}\left[\left(\sqrt{c\left(1 - \hat{\Delta}^2\right)}\right) + \left(\sqrt{c'\hat{\Delta}^2}\right)\right]^2 \quad (15)$$

Where $\hat{\Delta}^2 = \left[\frac{s_{xy}}{s_x s_y}\right]^2$

Double sampling for regression mean will be more precise than SRSWOR sample mean, in terms of their variances) if

$$\hat{\Delta}^2 > \frac{4cc'}{(c+c')^2} \quad (16)$$

2.5 Simple Random Sampling Without Replacement (SRSWOR)

As discussed by Thompson (1992), simple random sampling is a method of selecting n units out of the N, such that, everyone of the $\binom{N}{n}$ distinctly sample has an equal chance of being drawn that is $(N > n)$. However, when a selected item still retains an equal probability of being reselected as many times as possible then it is called Simple random sampling with replacement (SRSWR) else it is known as Simple random sampling without replacement (SRSWOR). The corresponding sample mean $\bar{y}$ is giving as;

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad (17)$$

and the estimated variance of $\bar{y}$ is giving as:

$$\hat{V}(\bar{y}) = \left(\frac{1-f}{n}\right)s^2 \quad (18)$$

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

Similarly, the corresponding optimum variance that minimizes variance of this estimator is giving as:

$$\hat{V}(\overline{y})_{opt} = \frac{c}{c_0} s_y^2 \tag{19}$$

2.6 Relative Efficiency.

Relative Efficiency is a statistical tool that will be used to measure the efficiency of one estimator over another estimator.

i. The Relative Efficiency Of Double Sampling Of Ratio To Regression Estimator

The relative efficiency of double sampling of ratio estimate as compared to double sampling of regression estimate is giving as;

$$R.E_{.1} = \frac{V(\overline{y}_{dl})_{opt}}{V(\overline{y}_{dr})_{opt}} * 100\%$$

$$R.E_1 = \frac{cs_y^2\left[\left(\sqrt{1-\hat{\Delta}^2}\right)+\left(\hat{\Delta}\sqrt{\frac{c'}{c}}\right)\right]^2}{\left[\left(c^{\frac{1}{2}}s_{xy}\right)+\left(c'(s_y^2-s_{xy})\right)^{\frac{1}{2}}\right]^2} * 100\% \tag{20}$$

ii. The Relative Efficiency Of Double Sampling Of Regression To Ratio Estimator

The relative efficiency of double sampling of regression estimate as compared to double sampling of ratio estimate is giving as;

$$R.E_2 = \frac{V(\overline{y}_{dr})_{opt}}{V(\overline{y}_{dl})_{opt}} * 100\%$$

$$R.E_2 = \frac{\left[\left(c^{\frac{1}{2}}s_{xy}\right)+\left(c'(s_y^2-s_{xy})\right)^{\frac{1}{2}}\right]^2}{cs_y^2\left[\left(\sqrt{1-\hat{\Delta}^2}\right)+\left(\hat{\Delta}\sqrt{\frac{c'}{c}}\right)\right]^2} * 100\% \tag{21}$$

iii. The Relative Efficiency Of Double Sampling Of Regression To SRSWOR Estimator

The relative efficiency of double sampling of regression estimate as compared to SRSWOR estimate is giving as;

$$R.E_3 = \frac{V(\overline{y})_{opt}}{V(\overline{y}_{dl})_{opt}} * 100\%$$

$$R.E_3 = \frac{1}{\left[\left(\sqrt{1-\hat{\Delta}^2}\right)+\left(\hat{\Delta}\sqrt{\frac{c'}{c}}\right)\right]^2} * 100\% \tag{22}$$

iv. The Relative Efficiency Of Double Sampling Of Ratio To SRSWOR Estimator

The relative efficiency of double sampling of ratio estimate as compared to SRSWOR estimate is giving as;

$$R.E_3 = \frac{V(\overline{y})_{opt}}{V(\overline{y}_{dr})_{opt}} * 100\%$$

$$R.E_3 = \frac{cs_y^2}{\left[\left(c^{\frac{1}{2}}s_{xy}\right)+\left(c'(s_y^x-s_{xy})\right)^{\frac{1}{2}}\right]^2} * 100\% \tag{23}$$

b. Coefficient of Variation (CV)

Coefficient Of Variation is a statistical tool that will be used to know the level of variability in each of these estimates under consideration. Lohr L. S. (2010) defines the coefficient of variation (CV) of the estimator $(\overline{y})$ as the measure of relative variability and defined as

$$CV(\overline{y}) = \frac{\sqrt{V(\overline{y})}}{E(y)} \tag{24}$$

It is a measure that does not depend on unit of measurement. Lohr L.S. (2010) estimated the CV of an estimator using the standard error divided by the mean (defined only when mean is nonzero).

$$CV(\overline{y}) = \frac{SE(\overline{y})}{\overline{y}} \quad (25)$$

The estimated CV is the standard error expressed as a percentage of the mean.

i. *Coefficient Of Variation for Ratio*

$$CV_{dr} = \frac{SE(\overline{y}_{dr})_{min}}{(\overline{y}_{dr})_{min}} * 100\% \quad (26)$$

ii. *Coefficient Of Variation for Regression*

$$CV_{dl} = \frac{SE(\overline{y}_{dl})_{min}}{(\overline{y}_{dl})_{min}} * 100\% \quad (27)$$

iii. *Coefficient Of Variation for SRSWOR*

$$CV = \frac{SE(\overline{y})_{min}}{(\overline{y})_{min}} * 100\% \quad (28)$$

3. Empirical Comparison Of Estimators

This research work uses primary data obtained from five hundred and seventy four (574) questionnaires distributed to the staff and students of Nursing school, Perioperative Nursing School, School of mid-wifery and Occupational Health School, all in University College Hospital (UCH) in Oyo state of Nigeria. The double sampling uses the household monthly expenditure (in thousands of Naira) of household on food consumption as the study variable (y) and the household size as the auxiliary variable (x). The double sampling obtains the first and second sample sizes at five different levels as presented below. n' is the sample size at first phase and n is the sample size at the second phase.

Table 1: Summary of the first and second phase sample sizes at different levels.

Level	***1***	***2***	***3***	***4***	***5***
n'	*140*	*130*	*120*	*100*	*80*
n	*50*	*45*	*40*	*35*	*30*
n/n'	*0.3571*	*0.3462*	*0.3333*	*0.3500*	*0.3750*

SPSS software was used to perform simple linear regression analysis on second phase sample data ($n = 40$), the model obtained is presented in the equation below.

$$\hat{y} = -4.218 + 5.480\,x + e \quad (29)$$

This means that the intercept of y axis is not zero, hence, these data are suitable for double sampling for regression type estimation. Table 2 below shows the variances and standard errors of double sampling ratio type estimator at each respective level.

Table 2: Summary of the standard error obtained in double sampling for ratio

Level	***1***	***2***	***3***	***4***	***5***
n'	*140*	*130*	*120*	*100*	*80*
n	*50*	*45*	*40*	*35*	*30*
$V(\overline{y}_{dr})$	*11.7458*	*11.5189*	*8.4035*	*11.9659*	*25.4558*
$S.E.(\overline{y}_{dr})$	*3.4272*	*3.3940*	*2.8989*	*3.4592*	*5.0476*

Similarly, table 3 below shows the summary of the variance and standard error obtained at each respective level for double sampling regression type estimator.

Table 3: Summary of the standard error obtained in double sampling for regression.

Level	***1***	***2***	***3***	***4***	***5***
n'	*140*	*130*	*120*	*100*	*80*
n	*50*	*45*	*40*	*35*	*30*
$V(\overline{y}_{dr})$	*10.9917*	*10.2199*	*7.5598*	*8.0857*	*12.2398*
$S.E.(\overline{y}_{dl})$	*3.3154*	*3.1969*	*2.7495*	*2.8435*	*3.4958*

Similarly, table 4 below shows the summary of the variance and standard error obtained at each respective level for SRSWOR type estimator.

Table 4: Summary of the standard error obtained in SRSWOR.

Level	***1***	***2***	***3***	***4***	***5***
n'	*140*	*130*	*120*	*100*	*80*
$V(\overline{y})$	*12.5979*	*12.5251*	*10.2794*	*13.4830*	*13.9770*
$S.E.(\overline{y})$	*3.5494*	*3.5891*	*3.2062*	*3.6719*	*3.7386*

Similarly, obtaining the relative efficiency (RE) assuming even distribution cost of $C = C' = 1$, table 5 shows the relative efficiency for the corresponding estimators.

Table 5: Summary of the computed Relative Efficiency.

Description	***RE***	***Conclusion***
Double sampling for ratio to regression	*35%*	*Double sampling for Ratio is 35% efficient over regression*
Double sampling for regression to ratio	*289%*	*Double sampling for Regression is 289% efficient over ratio*
Double sampling for regression to SRSWOR	*51%*	*Double sampling for Regression is 51% efficient over SRSWOR*
Double sampling for ratio to SRSWOR	*18%*	*Double sampling for Ratio is 18% efficient over SRSWOR*

Table 6: Summary on the comparison of coefficient of variation for each estimator.

	R.E.	***Conclusion***
Double sampling for ratio	*10.73%*	*Higher Precision*
Double sampling for regression	*10.1604%*	*Highest Precision*
SRSWOR	*51%*	*High Precision*

4. Discussion

Table 7: Variances, Standard Errors and Coefficient of Variations at $(n/n' = 0.3333)$.

At $n/n' = 0.3333$	***Variance***	***S.E.***	***C.V.***
Double sampling for ratio	*8.4035*	*2.8989*	*10.73%*
Double sampling for regression	*7.5598*	*2.7495*	*10.16%*
SRSWOR	*10.2794*	*3.2062*	*51%*

This paper empirically investigates the efficiency of double sampling for ratio and regression estimators and the usual Simple random sampling without replacement estimator to establish the most efficient estimator. It was

observed that the minimum proportion is obtained at $n' = 120$ and $n = 40$ (as revealed in table one). Hence, $n' = 120$ and $n = 40$ are the optimum sample sizes for the first and second phases respectively. It was shown that the minimum variance of double sampling for ratio, regression and SRSWOR were obtained at the same corresponding minimum proportion where $n' = 120$ and $n = 40$ (as revealed in table two). It was also observed that the minimum standard errors for double sampling ratio and regression and SRSWOR estimates were obtained at this same point (as revealed in table two, three and four). It was shown that among the optimum variances, standard errors and coefficient of variations, double sampling for regression estimator has the minimum values (as revealed in table seven), hence, the coefficient of variation showed that double sampling for regression has the highest precision after which is double sampling for ratio. The relative efficiency showed that double sampling for regression estimate has the highest efficiency over double sampling for ratio type estimator (as revealed table five). Comparison was made on double sampling for ratio and regression estimators and Simple random without replacement estimator to know which of these estimators has the highest efficiency and precision.

5. Conclusion

It is, therefore, established that when there exists linear relationship between the study variable $(y)$and the auxiliary variable and such that the regression line of y on x does not pass through the origin, the double sampling for regression estimation is efficient over double sampling for ratio and simple random sampling without replacement.

References

Cochran W. G (1977), "Sampling Technique", 3rd Edition, John Willey and sons Inc., New York.
Kish, L (1965), "Survey Sampling". John Wiley and Sons, New York.
Kumar S., Sigh Housila P., Bhougal Sanddep and Gupta Rahul (2011), "A class of Ratio-Cum-Product type estimators under double sampling in the presence of Non-response"', Hacettepe Journal of Mathematics and Statistics 40(4), 589-599.
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.
Mukhopadhyay, P. (2007), "Survey Sampling", Narosa Publishing House Pvt. Ltd, First Edition. 256.
Neyman J. (1938), "Contribution to the theory of Sampling Human Populations", Journal of the American Statistical Association. 33, 101-116.
Okafor F. C. (2002), "Sample Survey theory with applications", Afro-Orbis publication Ltd.
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.
Rao JNK (1973), "On Double Sampling for Stratification and Analytical Surveys", Biometrika. 60, 125-133.
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.
Sharon L. Lohr (2010), "Sampling Design and Analysis", Second Edition. Brooks/Cole Cengage Learning. 596.
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.
Thompson, S. K. (1992), "Sampling", John Wiley & Sons, New York.