IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 1 Issue 8, October 2014.
www.ijiset.com
ISSN 2348 – 7968

Comparison of allocation procedures in a stratified random sampling of skewed populations under different distributions and sample sizes

[1]Adebola Femi Barnabas and [2]Ajayi Olusola Sunday

[1]Department of Statistics, Federal University of Technology, Akure, Ondo State, Nigeria and [2]Department of Statistics, Federal University of Technology, Akure, Ondo State, Nigeria

ABSTRACT

One of the ways of increasing precision in sampling is to classify the population of study into series of homogeneous groups and by this, the precision will be increased. In this paper, different methods are used to investigate the best design for the allocation of samples into strata. Stratified real life data from the Education and Meteorological sectors of Ondo state, Nigeria is subjected to analysis to obtain population characteristics using three allocation procedures. Having consideration to the analysis and estimates obtained, the variances under the three procedures are compared to make conclusion. The conclusion is that although the performances of the three allocation procedures (Equal, Proportional and Optimum/Neyman) vary under different conditions, the Optimum Allocation Procedure is the most efficient. The conditions considered in this work include Sample Sizes and Distributions.

Keywords*:* Stratified Random Sampling, Skewed population, Sample Distribution, Sample Size, Allocation Procedure

1. INTRODUCTION

Sampling is concerned with the selection of a subset of individuals from within a population to estimate characteristics of the whole population. Sampling methods are designed to provide valid, scientific and economical tools for research problems.

According to Kish (1965) and Hunt and Tyrell (2004), sampling plays a vital role in research design involving human population and commands increasing attention from social scientists, chemists, engineers, accountants, biologists and medical practitioners. Sampling problems are in general to practitioners engaged in marketing, commerce, industry, public health, biostatistics, education,public administration, economics, sociology, anthropology, psychology, political Science and even social workers.

Different sampling designs would result in different standard errors, and choosing the design with the smallest error is the principal aim of sampling design.

In general, there is need to devise a sampling scheme which is economical and easy to operate, that yields unbiased estimates, and minimizes the effects of sampling variation. Usually in sample surveys more than one population characteristics are estimated and these characteristics may be of conflicting nature. Stratified sampling has been designed to ensure that all important views are represented in samples. Stratification is a means of sample design by which the population of interest is divided into groups, called strata, according to some known characteristic(s). Stratified sample designs are employed for several reasons.

The precision of information obtained from a survey depends on many factors such as;

- The size of the sample
- Sampling technique
- Variation within the strata, among others.

The precision of information obtained from a survey depends on the size of the sample in that if the sample size is increased, the variance component of the estimate will decrease. Sampling technique relates to precision in that a technique will be efficient if it is related to the distribution of the population values while in variation within the strata precision is of considerable magnitude whenever variability within the strata is different .

It is observed that survey is rarely carried out without stratification where a heterogeneous population is divided into a set of nearly homogeneous subsets called strata and independent samples are drawn from each stratum. In stratified sampling, the values of sample size n_h in the respective strata are chosen by the sampler, which may be carried out with either the aim of minimizing the variance for a specified cost or to minimize the cost for a specified value of variance.

Moreover, the concept of skewed population and sample distributions are brought to fore in this work. This is with the aim of studying the conditions under which the allocation procedures are best suitable.

ISSN 2348 – 7968

2. AIM AND OBJECTIVES

This work is aimed at comparing some allocation procedures in the stratified sampling of skewed population from an empirical point of view.
The objectives are;

- To examine the skewnesses of data used in the research
- To examine the various method of allocation in stratified random sampling of skewed population;
- To examine which method of allocation is suitable for estimating some population characteristics of a skewed population;
- To verify the condition(s) under which one method of allocation is better than the other

3. METHODOLOGY

The research shall adopt the computations of population total of skewed populations. The data sets from some populations will be considered and distributions determined. Data shall be simulated and the allocation procedures shall be applied to each of the distributions to conclude on the best method. Since most real life data always follow skewed population, real life data shall be collected and population totals shall be computed. Using the concept of minimum variance, the work shall compute some variables such as the Mean, Variances and Mean Square Errors (MSEs). Using some allocation methods, the population characteristics shall be computed using empirical data. The Variances and Mean Squared Errors (MSEs) computed shall be compared and the allocation method with the minimum Variance and MSE shall be regarded as the best and most efficient. Several previous works have recommended the Optimum/Neyman Allocation Procedures. This is predicated upon the theorem that shows that the Optimum Allocation has the least Variance when compared with Proportional and Equal Allocation Procedures.

3.1 Theoretical Framework:

1. The average of stratified sampling is:

$$\bar{Y}_{st} = \sum_{h=1}^{k} w_h \, y_h$$

2. If the variance of stratified sampling (SRSWOR):

$$\mathrm{V}(\bar{y}\mathrm{st})\mathrm{SRSWOR} = \frac{L}{n}\sum(W_h^2 \, S_y^2) - \frac{1}{N}(\sum W_h \, S_y^2$$

The variance of stratified sampling (Proportional)

$$\mathrm{V}(\bar{\mathrm{y}}_{\mathrm{st}})_{\mathrm{prop}} = \frac{1-f}{n}\left(\frac{N_h}{N}\right)(S_y^2 W_h)$$

And the variance of stratified sampling (Optimal/Neyman)

$$\mathrm{V}(\bar{\mathrm{y}}_{\mathrm{op/ney}}) = \frac{(\sum W_h S_{yh})^2}{n} - \frac{(\sum W_h S_{yh}^2)}{N}$$

Then;

$$\text{var } (\bar{y}_{st})_{op} \leq var\ (\bar{y}_{st})_{prop} \leq var\ \ (\bar{y}_{st})_{SRSWOR}$$

Proof

Recall: $var\ \ (\bar{y}_{st})_{SRSWOR} = \frac{N-n}{N_n} S^2$

$$var\ (\bar{y}_{st})_{prop} = \frac{\sum_{h=1}^{k} W_h S_h^2}{n} - \frac{\sum_{h=1}^{h} W_h S_h^2}{N}$$

$$\text{var } (\bar{y}_{st})_{op} = (\frac{\sum_{h=1}^{k} W_h S_h}{n})^2 - \frac{\sum_{h=1}^{k} W_h S_h^2}{N}$$

From analysis of variance:

ISSN 2348 – 7968

$S^2 = \frac{\sum_{i=1}^{N}(Y_i - \bar{Y})^2}{N}$ If N → ∞

$$= NS^2 \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y}_h + \bar{y}_h)^2$$

$$= NS^2 \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y}_h)^2 + \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y})^2 + 2\sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y})^1\ (\bar{y} - \bar{y})$$

$$= NS^2 \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y}_h)^2 + \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_h - \bar{y})^2$$

Recall that: $S_h^2 = \frac{1}{N-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{y})^2$

$(N_h - 1)\, S_h^2 = \sum_{i=1}^{N_h}(y_{hi} - \bar{y})^2$

$$\therefore NS^2 \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_{hi} - \bar{y}_h)^2 + \sum_{h=1}^{k}\sum_{i=1}^{N_h}(y_h - \bar{y})^2$$

$$S^2 \sum_{h=1}^{k}\frac{(N_h - 1)S_h^2}{N} + \frac{\sum_{i=1}^{k}\sum_{i=1}^{N_h}(y_h - \bar{y})^2}{N}$$

If $N_h \rightarrow$ large $\Rightarrow N_h - 1 = N_h$

$$S^2 = \sum_{h=1}^{k}\frac{(N_h)S_h^2}{N} + \frac{\sum_{i=1}^{k}\sum_{i=1}^{N_h}(y_h - \bar{y})^2}{N}$$

$$S^2 = \sum_{h=1}^{k} W_h S_h^2 + \frac{\sum_{i=1}^{k}\sum_{i=1}^{N_h}(y_h - \bar{y})^2}{N}$$

Recall:

$$var\ (\bar{y}_{st})SRSWOR = \frac{N-n}{N_n}S_y^2$$

$$\frac{N-n}{N_n}S^2 = \frac{N-n}{N_n}\sum_{h=1}^{k} W_h S_h^2 + \frac{N-n}{N_n}\frac{\sum_{i=1}^{N_h}(y_h - \bar{y})}{N}$$

$$= \left(\sum_{h=1}\frac{W_h S_h^2}{n} - \sum_{h=1}^{k}\frac{W_h S_h^2}{n}\right) + \left(\frac{1-f}{n}\right)\sum_{h=1}^{k}\sum_{i=1}^{N_h}\frac{(y_h - \bar{y})}{n}$$

$$var\ (\bar{y}_{st})SRSWOR \geq v\ (\bar{y}_{st})prop$$

ISSN 2348 – 7968

Now compare the variance of optimal/Neyman allocation with the variance of proportional allocation under stratified sampling

$$\sum_{h=1}^{k} \frac{W_h S_h^2}{n} - \sum_{h=1}^{k} \frac{W_h S_h^2}{N} = \frac{(\sum_{h=1}^{k} W_h S_h)^2}{n} - \sum_{h=1}^{k} \frac{W_h S_h^2}{N} + \sum_{h=1}^{k} \frac{W_h S_h^2}{N} - \sum_{h=1}^{k} \frac{W_h S_h^2}{N} - \frac{(\sum_{h=1}^{k} W_h S_h)^2}{n} + \frac{W_h S_h^2}{N} = 0$$

$$\sum_{h=1}^{k} \frac{W_h S_h^2}{n} - \frac{(\sum_{h=1}^{k} W_h S_h)^2}{n} = 0$$

Multiply through by n

$$\sum_{h=1}^{k} W_h S_h^2 - (\sum_{h=1}^{k} W_h S_h)^2 - = 0$$

Since

$$\sum_{h=1}^{k} W_h S_h^2 - (\sum_{h=1}^{k} W_h S_h)^2 \geq 0$$

Hence, $\text{var}(\bar{y}_{st})prop \geq var\ (\bar{y}_{st})opt$

$\therefore \text{var}\ (\bar{y}_{st})opt \leq var\ (\bar{y}_{st})prop \leq var\ (\bar{y}_{st})SRSWOR$

Against the background of the above, attention is given to Skewed Population to ascertain the theorem

3.2 DETERMINATION OF DISTRIBUTION OF DATA

To determine the distributions of data sets used in this study, the Easy-Fit version 5.5 software was deployed. The software was used to determine the Distributions of Three Real Life Data and twelve other simulated data sets. The Kolmogorov Smirnov Test was used for the Goodness of Fit. The data contained both Discreet and Continuous Data. The variances for the different allocation procedures were calculated using the derived formulae in the Micro Soft Excel Software environment and the samples drawn by Simple Random Sampling through the use of Table of random Numbers.

3.3 ALLOCATION OF SAMPLE TO STRATA

There are ways of distributing sample size to strata. This study will consider three allocation methods. These are Equal Allocation, Proportional Allocation and Optimum/Neyman Allocation.

3.4 Equal Allocation:

$$n_h = \frac{L}{n}$$

where L = no of strata, n = total sample

Its corresponding variance of mean stratified is given as;

ISSN 2348 – 7968

$$\frac{L}{n}\sum_{h=1}^{L} w_h^2 S_h^2 - \frac{1}{N}\sum_{h=1}^{L} w_h S_h^2$$

3.5 Proportional Allocation

$$n_h = \frac{nN_h}{N}$$

Its corresponding variance of mean stratified is given as;

$$V(\bar{y}_{st(prop)}) = \frac{1-f}{n}\sum_{h=1}^{L} w_h S_h^2 - \frac{1}{n}\sum_{h=1}^{L} w_h S_h^2$$

3.6 Optimum/Neyman Allocation

$$n_h = \frac{nN_h S_{yh}}{\Sigma N_h S_{yh}}$$

Its corresponding variance of mean stratified is given as;

$$V(\bar{y}_{opt/ney}) = \frac{(\sum W_h S_{yh})^2}{n} - \frac{(\sum W_h S_{yh}^2)}{N}$$

4. RESULTS AND DISCUSSIONS

Three Real Life Data sets and other simulated data sets were used in this work. The data contained both Discreet and Continuous Data. The first real life data were Secondary data collected from the Education sector of Ondo State. It is the enrolment figures in the Public Secondary Schools during the 2012/2013 academic session. The second data set is on the Performance of Ondo State Candidates in the West Africa Secondary Certificate Examinations, 2012. Both the first and second data sets are Discrete Data with N = 304 and n = 132. The third data set is a continuous data of Average Daily Maximum Temperature recorded in Ondo State in eighteen years . The total population, N, was varied (100, 500, 1000 and 5000). The samples were selected with 5% margin of error. The summary of results generated is shown in table 1 (the summary table).

ISSN 2348 – 7968

Table 1:
Summary table:

Table					*Variance by Allocation Method*			
Table No	*Description of Data*	*Distribution*	*N*	*n*	*Equal*	*Proportional*	*Optimum*	*Preferred method*
1	Enrolments in Sec Schools	D. Uniform	304	132	2233.618700	1659.000000	1375.844000	Optimum
2	Performances of Schools in WASCE	Geometric	304	132	6.390310	6.265600	4.720960	Optimum
3	Average Annual Maximum Temperature in Ondo State	Gen. Pareto	101	78	0.0008998160	0.001024	0.000952	Equal
4	Simulated Data	D. Uniform	100	78	593.737200	598.801400	511.374400	Optimum
5	Simulated Data	D. Uniform	500	216	762.142500	620.365000	623.621700	Proportional
6	Simulated Data	D. Uniform	1000	270	820.593900	1375.278950	645.823428	Optimum
7	Simulated Data	D. Uniform	5000	366	651.184600	566.394300	595.486698	Proportional
8	Simulated Data	Geometric	100	78	0.005805	0.005140	0.005662	Proportional
9	Simulated Data	Geometric	500	216	6.215440	5.022175	4.724000	Optimum
10	Simulated Data	Geometric	1000	270	6.331240	4.808610	4.779000	Optimum
11	Simulated Data	Geometric	5000	366	6.457780	4.290500	5.452100	Proportional
12	Simulated Data	Gen. Pareto	100	78	0.005991	0.005194	0.004680	Optimum
13	Simulated Data	Beta	500	216	0.006208	0.005483	0.003810	Optimum
14	Simulated Data	Beta	1000	270	0.006455	0.005701	0.005720	Proportional
15	Simulated Data	Beta	5000	366	0.005805	0.005140	0.005662	Proportional

ISSN 2348 – 7968

CONCLUSION

A cursory look into the analysis indicates that the same structure of performances are noticed when the Distributions and coefficients of Skwenesses are considered i.e Optimum Allocation Procedure is preferred in most.
Earlier previous researches recommended the choice of Optimum allocation procedure.
In my results, it is revealed that if the sample is drawn by varying the sample sizes, it is observed that the performances also vary.
In the summary table, it is noted that out of Fifteen Data analysis carried out (Real and simulated), Eight (8) cases attracted the preference for Optimum Allocation. This is in consonant with the previous research. However, the Proportion allocation procedure is preferred in Six (6) cases. Only one case where using Equal Allocation attracted least variance .The different Distributions is suspected to be responsible for this differences. The same reason can also be advanced for the different performances of the allocation procedures when the coefficients of skewness are considered.
From that, we can draw the conclusion that for estimating the average and variances of parameters under stratified Random Sampling of Skewed population, the performance of Optimum Allocation Procedure considered in this work is the best when compared with two other allocation procedures considered. Also, the use of Proportional Allocation Procedures can be more efficient in some situations as highlighted in the results of the study. It is further concluded that the Distribution of such population has no effect on the performances of the allocation procedures.

RECOMMENDATION

Based on the findings of the study, it is recommended that different allocation procedures should be tested when dealing with Stratified Random sampling of skewed populations.

REFERENCES:

Hunt, Neville;Tyrrel, Sidney (2001) " Stratified Sampling" web page at Conventry University

Sarndal, Carl- Erik; et al (2003) "Stratified Sampling" Model Assisted Survey Sampling New York Springer. Pp. 100-109

Kish, L. 1965 Survey sampling. New York, Wiley

Hunt, N. and Tyrell, S. (2004), Stratified sampling. CoventryUniversity Press http://www.coventry.ac.uk/ec/¬nhunt/meths/strati.html(accessed February 28, 2011)

Arthanari, T.S and Dodge, Y. 1981. Mathematical programming on statistics. A Wiley-Interscience, Publication, John Wiley & Sons Inc.

Bethel, F. 1989. Bayes and Minimax prediction in finite population. Journal of Statistical Planning, 60, 127 – 135.

Chatterjee, S. 1972. A study of optimal allocation in multivariate stratified surveys. Skand Akt. 73, 55 – 57.

Cochran, W. G. 1977. Sampling Techniques (3rd Edition), New York, Wiley

Dalenius, T. 1957. Sampling in Sweden: Contributions to the methods and theories of sample survey practice, Almavist and wicksell, Stockholm.

Diaz-Garcia, J.A and Cortez, L.U. 2008. Multi-objective optimisation for optimum allocation

ISSN 2348 – 7968

in multivariate stratified sampling. Survey Methodology, Vol. 34, No 2, 215-222.

Ghosh, S.P., 1958. A note on Stratified Random Sampling with Multiple Characters. Col. Stat. Bull, 8, 81-89.

Hunt, N. And Tyrell, S. 2004. Stratified sampling. Coventry University Press. http://www.coventry.ac.uk/ec/~nhunt/meths/strati.html (accessed February 28, 2011).

Khan, M.G.M and Ahsan, M.J. 2003. A note on Optimum Allocation in Multivariate Stratified Sampling. South Pacific Journal Natural Science, 21, 91-95.

Khan, M.G.M, Jahan, N. and Ahsan, M.J. 1997. Determining the optimum cluster size. Journal of the Indian Society of Agricultural Statistics. Vol. 50 (2), 121-129.

Kish, L. 1965 Survey sampling. New York, Wiley.

Kokan, A.R and Khan, S.U., 1967. Optimum allocation in mutivariate surveys. An analytical solution. Journal of Royal Statistical Society. Series B, 29, 115-125.

Neymaan, J. 1934. On the two different aspects of the representative methods. The method stratified sampling and the method of purposive selection. Journal of Royal Statistical Society, 97, 558-606.

Sukhatme, P.V, Sukhatme, B.V, Sukhatme, S., and Asok, C. 1984 Sampling Theory of Survey with Applications. 3rd Edition. Ames, Iowa: Iowa State University Press.