FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 3 No. 2, March, 2019, pp 384 – 388
DOI: https://doi.org/10.33003/fjs-2019-0302-xxx

EXPONENTIAL MOVING AVERAGE STRATIFICATION ALGORITHM FOR STRATA BOUNDARY DETERMINATION IN STRATIFIED SAMPLING

[1]Olayiwola, O. M., [1]Apantaku, F. S., [1]Oguntolu, O. S., [2]Ajibade, F, B., [1]Akintunde, A. A., [3]Olawoore, S. A.

[1]Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria
[2]Department of General Studies, Mathematics units, Petroleum Training Institute, Effurun, Delta State Nigeria
[3]Department of Statistics, Oyo State College of Agriculture and Technology, Igboora, Oyo State, Nigeria

*Corresponding Author's Email: opeoguntolu@yahoo.com

ABSTRACT
Strata boundaries determination is one of the procedures in stratified sampling to ensure optimum stratification that makes strata internally non-overlapping and homogenous. Authors have provided sets of procedures for the determination of the strata boundaries, which are complex and time consuming. This study developed a method called Exponential Moving Average Stratification (EMAS) to break the complexity of previous approaches and reduce their implementation time. The EMAS uses exponential moving average and cumulative of mean deviation to group the population into number of required strata. The EMAS was compared with Moving Average Stratification (MAS) method. Strata boundaries were established with both methods. Variance of population mean, coefficient of variation between strata and relative efficiency for EMAS and MAS were compared. EMAS provided minimum variance of the population mean and minimum coefficient of variation. The relative efficiency of EMAS was greater than 100 percent, hence EMAS performed better than MAS and suggested for strata boundaries determination in stratified sampling.

Keywords: Strata, stratified sampling, exponential moving average, coefficient of variation, relative efficiency.

INTRODUCTION

Simple random sampling is a method of selecting n units from a population of N units such that every one of the n distinct samples has an equal chance of being drawn. However, other methods of sampling are often preferable to simple random sampling on the grounds of convenience or of increased precision. Stratification is one such methods (Gunning and Horgan, 2004). Stratified sampling design is a methodology in which the elements of a heterogeneous population are classified into mutually exclusive and exhaustive homogenous subgroups called strata based on one or more characteristics of importance, and samples are drawn from each stratum (Cochran, 1977).

Stratification is one of the most widely used techniques in sample survey design, serving the dual purpose of providing samples that are representative of major subgroups of the population and of improving the precision of estimators. Horgan (2006) stated that stratification technique is often used majorly to maximize the precision of some estimator $\hat{\theta}$ or equivalently to minimize the Mean Square Error MSE ($\hat{\theta}$). Depending on the sampling scheme employed in selecting the samples independently from each stratum, Stratified Sampling become Stratified Random Sampling when Simple Random is employed and when Systematic Sampling is used, it becomes Stratified Systematic Sampling (Kareem *et al*, 2015). Dalenius and Hodges (1959), Hess et al. (1966), Wang and Aggrawal (1984), Okafor (2002) and Horgan (2006)) itemized the following as specific design problems involved in stratification processes: the choice of a stratification variable, the choice of number of strata L to be formed, mode of stratification; that is, the way/manner in which strata boundaries are determined, the choice of sample size n_h to be taken from the h^{th} stratum; that is, the problem of allocation of sample size to strata; and choice of sampling design within strata. Cochran (1977) stated that for a single item or variable (Y), the best characteristic is clearly the frequency distribution of Y itself. The next best characteristic is presumably the frequency distribution of some other quantity highly correlated with Y (the study variate), that is, some auxiliary variable X, such as the value of Y at a previous census. On the number of Strata to be constructed, in most of the surveys, the number of strata is predetermined; while in others, optimum number of strata is believed to have been attained when there is no further gain in precision by increasing the number of strata. The simplest methods of obtaining boundaries are the quantile method which places the same number of units in each stratum and the equal range method suggested by Aoyama (1954) which divides the range by the number of strata. If the quantile method is applied to highly positively skewed populations, the strata at the lower end are too narrow and those at the upper end too wide for optimum estimation (Cochran, 1961). On the other hand, using the equal range method on positively skewed populations, the strata at the lower end are too wide and those at the upper end too narrow (Cochran, 1961). Another simple method (termed the equal aggregate method) was proposed by Mahalanobis (1952) and Hansen *et al*. (1953) where the total aggregate value is equal for all strata. Dalenius

and Gurney (1951) suggested that the formation of strata be on the basis of equalization of W_h however, since the calculation of T_h is required, and this depends on the stratum boundaries, this method is not convenient in practice (Cochran, 1961).

Cochran (1961) compared the cumulative root frequency method, the equal aggregate method of Mahalanobis (1952), Ekman's method and Durbin's method, for 2, 3 and 4 strata by applying them to eight real skewed populations. He found that both the cumulative root frequency method and Ekman's method performed consistently well, Durbin's method did fairly well except on the two most skewed populations. He also found that the equal aggregate method of Mahalanobis (1952) was relatively unsuccessful on the three least skewed populations, going on to explain that this result is not surprising since the method is not designed to work well for a rectangular distribution with the lower end at zero. For the other populations, the equal aggregate method behaved erratically. Hess *et al*. (1966) observed that "Sethi's method, to some extent, and the cumulative root frequency method to a greater extent, lead to the construction of top strata that are too wide, with the result that these strata contribute heavily to the total variance." Singh (1971) and Thomson (1976) recommended a method of obtaining stratum boundaries based on equal partitioning of the cumulative cubed root frequency of the density function. Singh's method requires prior knowledge of the regression model of the survey variable y on the auxiliary variable x, while Thomson (1976) assumes the regression model is linear. He also concluded that the cumulative cubed root frequency works better with proportional allocation than with equal allocation. He also claims this method compares favourably to the cumulative frequency method using proportional allocation. Another approach taken for determining optimum stratum boundaries is to formulate the problem as a mathematical programming problem. Khan *et al*. (2002) views the problem of stratum construction as a multistage decision where the optimum stratum widths are determined using dynamic programming to obtain the global minimum of the objective function using Neyman allocation for fixed sample size. Random search methods have also been suggested. One method proposed by Kozak (2004) iteratively increases or decreases one boundary by not more than 5 units while the other boundaries remain constant. He claims this algorithm is more efficient than the random search method proposed by Niemiro (1999) which changes a boundary by one unit which could result in the algorithm stopping at a local minimum and does not work well for large populations as it requires too many iteration steps. Model-based methods treat values in the population as random variables and derive inferences to the population from the model specified for the random variables. A model-based approach to stratification has also been suggested by researchers and is described in Sarndal *et al*. (1992). However, accuracy depends on the choice of model. There are several methods of constructing strata boundaries in the literature; the most recent of the method was moving average stratification (MAS) method by Kareem *et al* (2015). In this paper, we suggest a new approach to MAS method of stratification. The objective of this paper is to develop exponential moving average (EMAS) and compare with MAS and examine the performance of the new method. Optimum allocation was used while simple random sample was the choice scheme within the strata.

MATERIALS AND METHOD

Stratification of a finite population

Suppose there are L strata containing N_h units from which a sample of size n_h is to be chosen independently from each stratum ($1 \leq h \leq L$) using simple random sampling. We write the population size as

$$N = \sum_{h=1}^{L} N_h \quad (1)$$

And total sample size as

$$n = \sum_{h=1}^{L} n_h \quad (2)$$

The overall population mean is:

$$\bar{Y} = \frac{1}{N}\sum_{h-1}^{L}\sum_{i=1}^{N_h} y_{hi} \quad (3)$$

Where y_{hi} is the i^{th} unit in the h^{th} stratum. This population mean may also be written as:

$$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h \quad (4)$$

Where

$$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi} \quad (5)$$

is the mean of the units in the h^{th} stratum and

$$W_h = \frac{N_h}{N} \quad (6)$$

is the stratum weight, i.e. the proportion of population units falling in stratum h.

The overall population variance is

$$S^2 = \frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y})^2}{N-1} \quad (7)$$

An estimate of the population mean is formed by combining the separate stratum sample means using weights W_h. The stratified mean estimate is defined as:

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \quad (8)$$

Where

$$\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi} \quad (9)$$

is the mean of the sample units in the h^{th} stratum with y_{hi} being the i^{th} unit of the sample chosen in the h^{th} stratum. Note, it is easy to show that $\bar{y}_{st}$, defined in equation (9), is an unbiased estimator of the population mean $\bar{Y}$. Since

$E(\bar{y}_h) = \bar{Y}_h$

Then

$$E(\bar{y}_{st}) = \sum_{h=1}^{L} W_h E(\bar{y}_h) = \sum_{h=1}^{L} W_h \bar{Y}_h = \bar{Y} \quad (10)$$

The variance of the stratified mean is:

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \quad (11)$$

Now since $\bar{y}_h$ is the mean of a simple random sample drawn from the h_{th} stratum containing N_h units then

$$V(\bar{y}_h) = \left(1 - \frac{n_h}{N_h}\right)\frac{S_h^2}{n_h} \quad (12)$$

It follows that

$$V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 (1 - \frac{n_h}{N_h})\frac{S_h^2}{n_h} \quad (13)$$

Also

$$f_h = \frac{n_h}{N_h} \quad (14)$$

is the sampling fraction in stratum h and

$$f_{pc_h} = 1 - \frac{n_h}{N_h} \quad (15)$$

Is the finite population correction factor for stratum h
The coefficient of variation is a measure of dispersion relative to the mean, and is defined as

$$CV = \frac{S}{\bar{Y}} \quad (16)$$

The coefficient of variation of stratum h is written as:

$$CV_h = \frac{S_h}{\bar{Y}_h} \quad (17)$$

and the coefficient of variation of the stratified sample mean $\bar{y}_{st}$ is:

$$CV(\bar{y}_{st}) = \frac{\sqrt{V(\bar{y}_{st})}}{\bar{Y}} \quad (18)$$

Construction of stratum boundaries
First, we summarize the moving average stratification method as it was written by Kareem *et al.* (2015).

Moving Average stratification method (MAS)
Moving Average stratification procedure for stratum construction is carried out for the division of a population into L strata as follows:

a. Arrange the value of X in ascending order of magnitude
b. Obtain the moving average (MA) of order L: MA(X_L) = (X_i + X_{i+1})/L; i = 1,2…..N and L = 1,2,……h
c. Deviate the means of the data series from the MA of order L
d. Cumulate the absolute value to be G
e. Obtain the first boundary by dividing G by desired number of strata: k_h = G/L
f. The serial number I corresponding to the approximate value of k_h is the first boundary while other boundary are at the serial number corresponding to the approximate value of h*k_h for h = 1,2,……(L-1) depending on the number of strata required.

Where
X_i = first value of the stratification variable.
X_{i+1}= is the next value of the stratification variable. L is the number of predetermined strata.
K_h = is the boundary estimator which must be multiplied by corresponding value of h to determine other boundary.

Exponential moving average stratification method
The proposed method for stratification is based on the view that exponential moving average is more responsive to changes in trends, more sensitive and reactive to recent values and more efficient when comparing long term averages, also effect of lag in data may reduce the responsiveness of moving average indicator. Since moving average is based on prior data, they suffer a time lag before they reflect a change in trend.
Exponential moving Average stratification (EMAS) procedure for stratum construction is carried out for the division of a population into L strata as follows:
Step1: Arrange the values of X in ascending order of magnitude
Step2: Obtain the simple moving average (SMA) of order L
MA(X_L) = (X_i + X_{i+1})/L; i = 1,2,……N and L = 1,2,…….h
Step 3: calculate the weighting multiplier (WM) : $\frac{2}{L+1}$
Step 4: calculate the exponential moving average of order L
EMA = WM*(X_i - EMA_{i-1}) + EMA_{i-1}
Step6: Deviate the mean of the data series from the EMA of order L
Step7: Cumulate the absolute value to be G
Step8: Obtain the first boundary by dividing G by the desired number of strata: K_h = G/L
Step9: Other boundary are approximate value of h*k_h for h = 1,2,……L-1
This algorithm is similar to the MAS algorithm except for the introduction of step 3 and 4 where weighting multiplier is introduced and the value for EMA is determined.
In the next section, we examine how successful this algorithm is in term variance of the population mean, coefficient of variation between strata and relative efficiency.

Comparisons of methods of stratum construction
To test the performance of newly proposed algorithm, it was implemented on two real data, both of which are positively skewed (Rate of unemployment in 182 countries of the world as reported by National Bureau of Statistics (NBS) in December, 2018 and the number of conservative seats in municipal council of Sweden comes from Sarndal *et al.*'s book (1992).The two data described in Table 1 were divided into *L* = 3, 4, 5 and 6 strata using MAS and EMAS respectively, simple random samples of fixed sample sizes 40 and 60 are selected for data 1 and 2 respectively. Results obtained using optimum allocation are shown in the following tables respectively.

RESULTS AND DISCUSSION

Table 1: Summary statistics of the populations used in this study

SN	N	n	Range	Coefficient of Skewness	Mean	Variance
1	182	40	46.00	2.155	8.276	51.689
2	284	60	70.00	1.393	47.535	122.144

Table 1 presented the summary statistics of both the rate of unemployment data and the number of seats in municipal in Sweden. Data 1 has a population total of 182 with mean 8.276 and variance 51.689, with a sample size of 40 selected. Data 2 has a population total of 284 with mean 47.535 and variance 122.144 with a sample size of 60 selected. Both data are positively skewed with values of 2.155 and 1.393 respectively.

Table 2: Variance of the population mean between strata using optimum allocation.

Strata	Method	Variance of the population mean	
		Data 1	Data 2
3	EMAS	0.1684	0.3819
	MAS	0.1653	0.3832
4	EMAS	0.0786	0.1641
	MAS	0.0790	0.1678
5	EMAS	0.0598	0.1093
	MAS	0.0603	0.1135
6	EMAS	0.0383	0.0963
	MAS	0.0387	0.0848

Table 3: Coefficient of variation (CV) between strata using optimum allocation.

Strata	Method	Data 1	Data 2
3	EMAS	0.0496	0.0130
	MAS	0.0491	0.0130
4	EMAS	0.0339	0.0085
	MAS	0.0340	0.0086
5	EMAS	0.0295	0.0070
	MAS	0.0297	0.0071
6	EMAS	0.0236	0.0061
	MAS	0.0238	0.0065

Table 4: Relative efficiency of new method for data 1 and data 2

Strata	Data 1	Data 2
3	98.16	100.34
4	100.51	102.25
5	100.84	103.84
6	101.04	113.57

DISCUSSION

Table 2 presented the variance of the population mean $V(\bar{y}_{st})$ for rate of employment data and number of seats in municipal council of Sweden data using optimum allocation. In both data 1 and 2, when L= 3, 4, 5 and 6 exponential moving average stratification (EMAS) method performed better than MAS by producing lower variance of the population mean in all the strata. Table 3 presented the coefficient of variation between strata (CV), when L= 3, 4, 5 and 6. The result shows that

exponential moving average stratification method performed better than moving average stratification in both data and 1 and 2, though there were very close values between these two methods but it can be seen that exponential moving average stratification method provides more accurate estimates than moving average stratification method. Table 4 presented the relative efficiency of the new method EMAS to MAS and it can be seen that there was increase in the precision of the estimates as the number of strata increases. The result shows that gains are observed for EMAS in majority of the cases especially when L = 4, 5 and 6 both in data 1 and 2. EMAS performed better than MAS in most of the strata. This indicate that the precision of the proposed method is better than that of the comparator method.

CONCLUSION

This paper is an improvement on moving average stratification (MAS) algorithm for the construction of stratum boundaries in stratified sampling. This study reveals that exponential moving average stratification method is more efficient than moving average stratification method in minimizing the variance of the estimate of the population mean V($\bar{y}_{st}$). The results showed that the exponential moving average stratification (EMAS) method is more precise than the moving average stratification (MAS) method in most of the strata formations. Therefore, it is recommended for use in strata boundaries determination in stratified sampling.

REFERENCES

Aoyama, H. (1954). A study of the stratified random sampling. The Annals of Mathematical Statistics, VI(l):l-36.

Cochran, W.G. (1961). Comparisons of methods for determining stratum boundaries.Bulletin of the International Statistics Institutes, 32 (2): 345-358.

Cochran, W.G. (1977). Sampling Techniques, Third edition. John Wiley and Sons, New York.

Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification II. SkandinaviskAktuarietidskrift, (34):133-148.

Dalenius, T., and Hodges, J. L., Jr. (1959). Minimum Variance Stratification Journal of American Statistical Association 54: 88-101.

Durbin, J. (1959). Review of sampling in Sweden. Journal of the Royal Statistical Society (122):246-248.

Ekman,G.: (1959). An aproximation useful in univariate stratification, The Annals ofMathematicalStatistics 30, 219–229.

Gunning, P. and Horgan, J.: (2004). A new algorithm for the construction of stratum boundariesin skewed populations, Survey Methodology 30, 159–166.

Hansen, M., Hurwitz, W., and Madow, W. (1953). Sample Survey Methods and Theory. Wiley, New York.

Hess, I, Sethi, V.K. and Balakrishnan, T.R. (1966).“Stratification: A practical investigation”JASA, 61: 74-90.

Horgan, J. M. (2006). Stratification of Skewed Populations: A Review. International StatisticalReview, 74, (1): 67-76.

Kareem A. O., I, O, Oshugade, G, M, Oyeyemi and A. O. Adejumo (2015) “Moving Average Stratification algorithm for data boundary determination in skewed populations” CBN Journal of Applied Statistics Vol. 6 No. 1(b):205-217

Khan,E.A.,Khan,M.G. M. and Ahsan,M.J. (2002). Optimumstratification: A mathematical programming approach. Calcutta Statistical Association Bulletin 52(special):205–208.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. Statistics in Transition 6, 797–806.

Mahalanobis, P. (1952). Some aspects of the design of sample surveys. Sankhya,12:1-7.

Niemiro, W. (1999). Optimal stratification using random search method. WiadomosciStatystyczne 10:1–9.

Okafor, F. C. (2002). Sample Survey Theory with Applications. Afro-Orbis Publications Ltd.Nsukka, Nigeria.

Sarndal, C., Swensson, B. and Wretman, J. (1992). Model-Assisted Survey Sampling. New York: Springer Verlag.

Sethi, V. K. (1963). A note on optimum stratification of populations for estimating the populationmeans. Australian Journal of Statistics, 5:20-33.

Singh, R. (1971). Approximately optimal stratification of the auxiliary variable. Journal of the American Statistical Association. 66:829-833

Thomson, I. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. Metrika, 23:15-25.

Wang, W.C. and Aggarwal, V. (1984). Stratification under a particular pareto distribution.Commun. Statist. - Theory. Meth. 13 (6):711-35.

.