Moving Average Stratification Algorithm for Strata Boundary Determination in Skewed Populations

Akeem O. Kareem[1], Isaac O. Oshungade[2], Gafar M. Oyeyemi[3] and Adebowale O. Adejumo[4]

Moving Average Stratification (MAS) is a new competing and simple algorithm for strata boundary determination in Stratified Sampling. It eliminates arbitrary choice of class interval associated with cumulative square root of frequency method (Dalenius and Hodges Rule (DHR) 1959) and the inherent geometric gaps created within strata by Geometric Stratification (GMS) of Gunning & Horgan (2004). It competes favorably well with DHR and GMS in terms of its precision, simplicity and speeds and therefore recommended for use in strata boundaries determination especially in skewed populations.

Key Words: Class intervals, Homogeneity, Geometric gaps, Stratification, and Efficiency.

JEL Classification: C22, C42, C83.

1.0 Introduction

Strata boundary determination is one of the technical operations involved with the use of Stratified Sampling design in Survey Sampling. 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. In this study Stratified random sampling is used within the strata. Thus, for a population of Units U, divided into L strata, relation (1) below must be satisfied.

$$U = \bigcup_{h=1}^{L} U_h ; U_h \cap U_k = \varphi, \forall h \neq k, (h,k = 1,2,...,L) \quad (1)$$

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

[1]Institute for Security Studies, P. M. B. 493, Abuja, Nigeria. keemkareem@yahoo.com
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria.osungade@unilorin.edu.ng
[3]Department of Statistics, University of Ilorin, Ilorin, Nigeria. matanmi@unilorin.edu.ng
[4]Department of Statistics, University of Ilorin, Ilorin, Nigeria. aodejumo@unilorin.edu.ng

majorly to maximize the precision of some estimator $\hat{\theta}$ or equivalently to minimize the Mean Square Error MSE ($\hat{\theta}$). This study compares our new MAS algorithm with DHR and GMS using the minimum variance approach.

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:

(a) the choice of a stratification variable;
(b) the choice of number of strata L to be formed;
(c) mode of stratification; that is, the way/manner in which strata boundaries are determined;
(d) the choice of sample size n_h to be taken from the hth stratum; that is, the problem of allocation of sample size to strata; and
(e) 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. This study allows for optimum number of strata and the stratification process continued until when deep stratification occurs, that is, $N_h = 1, \forall h = 1, 2, ..., L$ (at least one population units in one or more stratum). There are several methods of constructing strata boundaries in the literature; the two most popularly used is compared in this study with our proposed MAS. Optimum and proportional allocations were used while simple random sample was the choice scheme within the strata.

2.0 Methods of Strata Boundaries Determination

Dalenius (1950) was credited with the first statistical research into the problem of strata boundary determination. He found the optimum stratification points for Neyman allocation to be those which satisfy the equation:

$$\frac{\sigma_h^2+(X_h-\mu_h)^2}{\sigma_h}=\frac{\sigma_{h+1}^2+(X_h-\mu_{h+1})^2}{\sigma_{h+1}}, h=1,2,...,L-1,L \quad (2)$$

and for proportional allocation to the number of units, the general expression for the simultaneous equation is:

$$X_h=\frac{(\mu_h+\mu_{h+1})}{2}; h=1,2,...,L-1 \quad (3)$$

while Cochran (1961) and Sethi (1963) reported that the general equation to be satisfied in order to obtain optimum stratification points for equal allocation is given as:

$$W_h[\sigma_h^2+(X_h-\mu_h)^2]=W_{h+1}[\sigma_{h+1}^2+(X_h-\mu_{h+1})^2], h=1,2,...,L-1,L \quad (4)$$

These equations are solved through various steps of iteration to obtain Optimum Points of Stratification (OPS) and these equations are derived on the assumptions that the variable of study is the stratification variable and that the frequency distribution is continuous. For details on derivation of these sets of general equations, see Dalenius and Hodges (1957, 1959), Murthy (1967, section 10.7a, pp.262), Sukhatme and Sukhatme (1970) section 3.11, pp. 108), Cochran (1977, section 5A.7, pp.127), Raj and Chandhok (1998, section 4.8, pp. 107) and Okafor (2002, section 4.6, pp. 120).

These sets of general equations had been greeted with lots of criticism in terms of their difficulty and time involved in solving the equations as well as their practical adaptability. Thus, for easy application, sets of approximate solutions have been suggested by various authors and these include: Mahalanobis (1952) suggested Equalization of Strata Totals (EST), Ekman (1959) on his part suggested that for a density over a finite range points$\{X_h\}$, Dalenius and Hodges (1959) presented a quick approximate method referred to here as Dalenius and Hodges Rule (DHR), Durbin (1959), while reviewing the DHR, proposed Durbin's Rule (DUR), Sethi (1963) solved the sets of general equations for some standard distributions (Normal, Beta, Gamma and various Chi-squares) referred to as Sethi's Rule (STR) and The Thomson Rule (TNR), Thomson (1976)

Other methods in the literature are those of Lavalle and Hidiroglou (1988) Method (LHM); Extended Ekman's Rule (EEKR) by Hedlin (2000), Random

Search method (RSM) was due to Kozak (2004); Geometric Stratification (GMS) by Gunning and Horgan (2004) and Genetic Algorithm (GA) by Keskinturk and Er (2007). At this juncture, it is pertinent to mention that of all the aforementioned methods and approximate solutions, DHR and GMS are the most commonly for their efficiency and forms the basis of comparison with MAS in this study.

3.0 Geometric Stratification

Gunning and Horgan (2004) introduced the new and now the most commonly used method of strata boundary determination called Geometric Stratification (GMS). It was applied to positively skewed populations and results compared with DHR. Stratum boundaries are automatically formed with GMS once the geometric ratio **r** is determined.

$$r = [\max Y_i / \min Y_i]^{1/L}$$
$$r = [Y_L / Y_0]^{1/L} \quad (5)$$

where Y_L is the highest value and Y_o is the smallest value of the study variate Y. The boundaries are at the points:

Minimum K_0 = a, ar, ar^2, . . . ,ar^L = Maximum K_L.

The general term is:

$$K_h = ar^h, h = 0,1,2,...,L-1 \quad (6)$$

Details of the GMS algorithm are in section 2 of Gunning and Horgan (2004). The simplicity of GMS had been extended to Pareto distribution by Gunning *et al.* (2006) and was found to be more efficient than DHR.

4.0 Moving Average Stratification (MAS)

The MAS technique was developed with the aim of tackling the problem of arbitrary choice of class interval associated with DHR (i.e. unavailability of a theory to guide the choice of class interval in the application of DHR) as observed by Hedlin (2000) as well as reducing the variability within strata and at the same time ensuring an approximately equal variability within strata. MAS technique could be likened to the EST in terms of its procedure.

Basically, moving Average technique is used in Time series analysis to smooth local fluctuations that may exist in a series of data. It is therefore employed as a stratification method in this study to distribute fluctuations (variations) that may exist within a given set of data on equal basis among the strata formations, thus achieving homogeneity of units within the strata. The MAS Algorithm is presented thus:

MAS Algorithm

Let X_1, X_2, ... , X_N be the values of the stratification variable X which is highly correlated with the study variable Y, and if the study variate Y itself is readily available, it could be applied directly. Dalenius (1959, Ghosh (1963), Hess *et al.* (1966) and Hedlin (2000) used study variable for the purpose of stratification.

i. Arrange the values of X in ascending order of magnitude and serially numbered.

ii. Obtain the Moving Averages (MA) of order L,

$$MA(\bar{X}_L) = (X_i + X_{i+1}) / L \qquad (7)$$

$$for, i = 1,2,..., N \text{ and } L = 1,2,.....h.$$

iii. Form the third column such that (L-1) gap are created in the first row of the third column. For example, when L = 2 relation (7) gives MA of order L = 2;

$MA(\bar{X}_2) = (X_i + X_{i+1})/2$, this could be likened to Dalenius equation (3) above.

iv. Let $\bar{X}_{Li}$ be the MA of order L obtained for $i = 1,2,..., N-(L-1)$

v. Deviate the mean of the data series $\bar{X}$ from the MA of order L, i.e. $\bar{X}_{Li} - \bar{X}$ to form the forth column.

vi. Cumulate the absolute value $\left|\bar{X}_{Li} - \bar{X}\right|$ such that $cum\left|\bar{X}_{Li} - \bar{X}\right| = G$

vii. Obtain the first boundary by dividing G by desired number of strata L i.e., $K_h = G / L$

viii. The serial number i corresponding to approximate value of K_h is the first boundary, while other boundaries are at the serial number i corresponding to approximate values of $h * K_{h,}\, for, h = 1,2,...(L-1)$ depending on the number of strata required.

Thus, $K_h = K_{h+1}$ (8)

This could be likened to Dalenius equation (2) or (4).

Remarkably, MAS is speedily accomplished even on Microsoft Excel sheet.

5.0 Estimation Procedures in Stratified Sampling

This section discusses estimation procedure in stratified random sampling. Symbols and notations of Cochran (1977, pp.90) were adopted in this study.

Notations

The subscript **h** denotes the stratum and **i** the unit within the stratum,

for h = 1, 2, . . . , L.

L	=	Number of strata.
N_h	=	Total number of population units in stratum h.
n_h	=	Number of sample units taken in stratum h.
N	=	Total number of population units in all the L strata
n	=	Sample size of the study
Y_{hi}	=	The observed value of the ith unit in the hth stratum
W_h	=	N_h/N = stratum weight (population units)
w_h	=	n_h/n = stratum weight (sample units)

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

True mean = $\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} Y_{hi}$ (10)

Sample Variance = $s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$ (11)

True Variance = $S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2$ (12)

Population Mean = $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h$

$$\text{Variance of the population mean} = V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\bar{y}_h) \qquad (13)$$

$$= \sum_{h=1}^{L} W_h^2 (1 - f_h) \frac{S_h^2}{n_h} \qquad (14)$$

where the subscript "st" denotes stratified.

6.0 Allocation of Samples to Strata

Once we select at least a sample from each stratum the procedure of stratified sampling is satisfied. However, this study requires at least two units in each stratum for estimation purposes. Thus, method of collapsed strata is not considered in this study, Cochran (1977), pp.138 5A.12.

After the sample size n is chosen, there are many ways of allocating n into individual stratum sizes $n_1, n_2, \ldots, n_L$ with the aim of using an allocation method that gives a specified amount of information at minimum cost. An allocation scheme is affected by the total number of units in each stratum, the variability of observations within each stratum and the cost of obtaining an observation from each stratum. The two popular employed are optimum and proportional allocations.

6.1 Optimum Allocation

With optimum allocation, samples may be selected to minimize the overall cost of the survey for a specified value of $V(\bar{y}_{st})$ or to minimize the $V(\bar{y}_{st})$ for a given overall cost of the survey.

Hence, $$n_h = \frac{nW_h S_h / \sqrt{C_h}}{\sum W_h S_h / \sqrt{C_h}} \qquad (15)$$

On the other hand, if the costs are unknown or constant (the same in each stratum i.e. $C_h = C$ for h = 1, 2, … , L, then expression (15) above reduces to:

$$n_h = \frac{nN_hS_h}{\sum N_hS_h} \tag{16}$$

This method of allocation of total sample size **n** to strata was due to Neyman (1934). Hence, it is often referred to as Neyman optimum allocation and its variance is given as:

$$V_{opt} = V_{\min}(\bar{y}_{st}) = \frac{(\sum W_hS_h)^2}{n} - \frac{\sum W_hS_h^2}{N} \tag{17}$$

6.2 Proportional Allocation

This allocation calls for making the stratum sample size n_h proportional to stratum size $\mathbf{N_h}$. It allocates large sampling units to large stratum and small sampling units to a small stratum, hence, a representative sample of the population units is obtained.

Proportional allocation is given by the expression:

$$n_h = \frac{nN_h}{N} = nW_h, h = 1,2,...L \tag{18}$$

And its variance is given as:

$$V(\bar{y}_{st})_{prop} = \frac{(1-f)}{n}\sum W_hS_h^2 = \frac{\sum W_hS_h^2}{n} - \frac{\sum W_hS_h^2}{N} \tag{19}$$

This allocation is often employed due to its simplicity. It is also good when S_h^2 is the same in all strata is and when there is no knowledge of relative size of within strata variances.

7.0 Results

The four (4) sets of data below, whose features are reflected in Table 1 are used for this study.

Data used in this study are;

i. Overall cumulative average scores of 145 students that graduated from the Faculty of Engineering University of Ilorin 1989/90 set.

ii. Data of Kano State Ministry of Commerce and Industry Survey (2008) on manpower strength of companies and industries in the six (6) industrial Estates of Kano.

iii. Grants allocation to 774 Local Government's Council in the country for the month of December, 2008 shared in January 2009. (See www.fmf.gov.ng)

iv. Population Census figures for the 774 Local Government Areas of Nigeria during the year 2006 census. (see www.nigeriastat.gov.ng)

Table 1: Summary Statistics of the data used in this study

S/N	N	N	Range	Coefficient of Skewness of the Population	Mean	Variance	Standard Deviation
1	145	48	44.7 - 68.8	0.712	55.48	20.05	4.48
2	171	57	3 – 3756	6.581	166	163923	405
3	774	258	72.2 - 365.0	3.239	108.96	700.61	26.47
4	774	258	11.7 - 1277.7	3.218	180	10281	101

Strata boundaries were obtained using DHR, GMS and MAS. Population units are placed in their respective stratum and simple random samples of fixed sample sizes 48, 57, 258 and 258 are selected for data 1 to 4 respectively in order to obtain relevant statistics for the purpose of estimating the population parameters using **R** packages (generating seed of 123). Results obtained for data 1 to 4 using optimum and proportional allocation is as shown in Tables 2 and 3 respectively.

Table 2: Variance of the Population Mean by Number of Strata and Methods of Strata Construction for DATA 1 to 4 Using Optimum Allocation.

	Data 1			Data2			Data 3			Data 4		
Strata	DHR	GMS	MAS	DHR	GMS	MAS	DHR	GMS	MAS	DHR	GMS	MAS
2	0.09869	0.089851	0.097375	34.78	128.76	154.02	0.5796	0.62554	0.62164	8.4273	18.2072	7.0855
3	0.048783	0.067589	0.074519	20.79	36.51	43.18	0.22953	0.36151	0.21933	3.0593	6.914	4.0212
4	0.024145	0.034797	0.035406	12.46	23.03	28.29	0.11481	0.24619	0.12244	2.1646	6.0417	2.1577
5	0.01523	0.023983	0.022959	13.11	3.43	8.23	0.09489	0.20642	0.11019	1.1201	3.4276	1.7057
6		0.016269	0.012268		5.33	9.9	0.08035		0.07036	0.8765	3.3028	0.9853

8.0 Discussion

Tables 2 and 3 above present the variance of the population mean of stratified random sampling for the three (3) methods of stratification considered.

In terms of number of strata formation, MAS produced more strata formations than DHR and GMS (8, 6, 10 and 10 Strata for data 1 to 4). DHR recorded 5strata formation for data 1 and 2 and 6strata formations for data 3 and 4. While GMS recorded 6strata formation data 1 and 4 and 5strata formation for data 2 and 3 respectively. However, performances of the three methods of strata boundaries determination studied is restricted to six strata formation i.e. L = 2, 3, 4, 5, and 6 using the four (4) data sets.

Table 3: Variance of the Population Mean by Number of Strata and Methods of Strata Construction for DATA 1 to 4 using Proportional Allocation.

	Data 1			Data2			Data 3			Data 4		
Strata	DHR	GMS	MAS	DHR	GMS	MAS	DHR	GMS	MAS	DHR	GMS	MAS
2	0.111925	0.099633	0.098897	1495.84	1715.5	1669.69	1.05755	0.6961	1.59444	11.3638	26.3685	12.017
3	0.060491	0.070774	0.08137	1260.81	1192.55	1315.27	0.5792	0.3778	0.67114	7.6474	8.3237	7.1949
4	0.040585	0.034715	0.044939	1264.85	1279.28	661.49	0.2854	0.29809	0.29168	4.8101	7.8909	5.197
5	0.015952	0.026129	0.034337	89.33	100.2	408.32	0.20067	0.22977	0.23858	2.516	8.7369	8.754
6		0.016136	0.014638			326.69	0.17339		0.19762	2.2992	7.9777	2.1701

When Optimum allocation is used, there is a strong competition by MAS with the existing methods. With Data 1, where coefficient of skewness of the population of study is less than1, beyond 2 strata formation, DHR is leading in terms of precision, followed by strong competition between GMS and MAS. Data 2 has the highest coefficient of skewness in this study, DHR leads in precision just as recorded with data 1 followed by GMS and MAS in this order. With Data 3 and 4 where the coefficient of skewness is moderate, that that is; < 4, MAS competes favourably well with DHR with consistent gain in precision for L = 2 to 6 and GMS performs poorly.

When proportional allocation is used MAS is most precise in 2 and 6 strata formation for data 1 while sustained strong competition were exhibited by DHR and GMS for the remaining strata formations. With data 2, DHR is most precise in 2 and 5 strata formations, GMS in 3 strata formation and MAS in 4strata formation. Data 3 shows that GMS is most precise in 2 and 3 strata formations and DHR most precise beyond 3 strata formations while MAS competes favourably with marginal differences in its precision when compared with DHR and GMS. Data 4 sustained strong competition between DHR and MAS.

9.0 Conclusion

This study provides MAS as strongly competing alternative to DHR and GMS. It can be speedily accomplished; it eliminates arbitrary choice of class interval associated with DHR as well as geometric gaps within strata by GMS. It is therefore recommended for strata boundary determination in stratified sampling most especially with moderately skewed data.

References

Cochran, W.G. (1961). "Comparison of Method for Determining Stratum Boundaries". *Bulleting of the International Statistical Institute* 38(2):345-358.

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

Dalenius, T. (1950). "The Problem of Optimum Stratification", *Skandinavisk Akturietidskrift,* 33:203-213.

Dalenius, T. and J.L. Hodges, Jr. (1957). "The Choice of Stratification Points. *Skandinavisk Akturietidskrift*, 198 -203.

Dalenius, T. and J.L. Hodges, Jr. (1959) "Minimum Variance Stratification" *JASA*, 54: 88 – 101.

Durbin, J. (1959). "Review of sampling in Sweden". *Journal of Royal Statistical Societies,* A(122):246-248.

Ekman, G. (1959). "An Approximation Useful in Univariate Stratification". *Annals of Mathematical Statistics*, 30: 210-229.

Ghosh, S.P. (1963) "Optimum Stratification with Two Characters". *Annals of Mathematical Statistics*, 34: 866-872.

Gunning, P, and Horgan, J.M. (2004) "A New algorithm for the construction of stratum boundaries in skewed population" *Survey Methodology*, 30(2):159-166.

Gunning, P., Horgan, J.M. and Keogh, G. (2006) "Efficient Pareto Stratification". Mathematical Proceedings of Royal Irish Academy 106A(2):131-138.

Hedlin, D. (2000) "A procedure for Stratification by an Extended Ekman rule". *Journal of Official Statistics*, 6(1):15-29.

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 Statistical Review,* 74(1):67-76.

Keskinturk, T. and Er, S. (2007) “A Genetic algorithm approach to determine boundaries and Sample size of each stratum in stratified Sampling”. *South Pacific Journal of Natural Sciences*, B(21):91-95.

Kozak, M. (2004) Optimal Stratification Using Random Search Method in Agricultural Surveys. *Statistics in Transition*, 6(5):797-806.

Lavalle, P. and Hidiroglou, M.A. (1988). On the Stratification of Skewed Populations. *Survey Methodology,* 14(1):33-43.

Mahalanobis, P.C. (1952). Some Aspects of the Design of Sample Surveys. *Sankhya*, 12:1-7.

Murthy, M.N. (1967). Sampling Theory and methods 2nd Edition Statistical Publishing Society, Calcutta - 35, India.

Neyman, J. (1934). “On the Two different aspects of the representative method: The method of Stratified sampling and the method of purposive selection”. *Journal of Royal Statistical Society*, 97:558-606.

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

Raj, D. and Chandhok, P. (1998). Sample Survey Theory; Narosa publishing House, 6, Community Centre, Panchsheel Park, New Delhi 100 017.

Sethi, V.K. (1963). “A Note on Optimum Stratification for Estimating the Population Means”. *The Australian Journal of Statistics*, 5:20-33.

Sukhatme, P.V. and Sukhatme, B.V. (1970). Sampling Theory with Applications. 2nd Edition, Iowa University Press, USA.

Thomson, J. (1976). “A comparison of an approximately optimal stratification given proportional allocation with other methods of stratification and allocation”. *Metrika*, 23(1):15-25.

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