Journal
of Science
Research

On linear stratification of skewed and normal populations

Kareem, A. O.,[1*] **Oshungade, I. O.**[2] and **Oyeyemi, G. M.**[2]
[1]Institute for Security Studies, Abuja, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria
[2]Department of Statistics, University of Ilorin, Ilorin, Nigeria
*Corresponding author: keemkareem@yahoo.com

Abstract

Strata boundary determination is one of the technical operations in Stratified Sampling. Maximized precision dominated the literature in the appraisal of the performance of methods of strata construction which fails to account for the bias associated with each method because the most precise method may not actually be the most efficient. This study develops Linear Stratification (LS) as a new and simple approach to strata boundary determination. Strata boundaries were established with LS, cumulative square root of frequency method and Geometric Stratification. Samples were selected randomly without replacement from each stratum and estimates of the population parameters obtained. These estimates were compared i.e. LS with that of the two existing methods using four sets of real life data with varying degrees of skewness. With the Mean Square Error (MSE) value rather than minimum variance commonly used for appraisal, the results show that LS provides minimum MSE value in both skewed and normal populations, hence the most efficient when compared with the two competing methods in strata boundary determination.

Keywords: Deep stratification; efficiency; linear progression; linear stratification; mean square error.

Introduction

Stratified Sampling is the design in which a heterogeneous population is divided into mutually exclusive and exhaustive subgroups called strata and independent samples drawn from each stratum. 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.

Stratification technique is often used majorly to maximize the precision of some estimator $\hat{\theta}$ or equivalently to minimize the Mean Square Error MSE ($\hat{\theta}$), and when is an unbiased estimate of θ (i.e. the estimate of the bias is zero), the MSE ($\hat{\theta}$) is the variance of [1]. Literature had continuously reported maximized precision (minimum variance) as a measure of appraisal of the performance of methods of stratification [1-4]. This approach fails to account for the bias associated with a particular method.

Stratification technique is employed in sample survey not only for its improved precision and provision of samples that are representative of the population units, but also for administrative convenience in its application. It is also very important when dealing with skewed population since greater weight will be given to the new extremely large units for reducing sampling variability [7].

This paper therefore, proposes the Linear Stratification (LS) as a new and simple approach to strata boundaries determination; appraises its performance based on the MSE value and compares its estimates with cumulative square root of frequency method herein referred to as Dalenius and Hodges Rule (DHR) [2] and Geometric Stratification (GMS) due to [3].

Specific design problems associated with stratified sampling as enumerated by [1, 2], [4-6] are; choice of a stratification variable, choice of number of strata L to be formed, mode of stratification, i.e. the way and manner in which strata boundaries are determined, choice of sample size n_h to be taken from the h^{th} stratum, i.e. the problem of allocation of sample size to strata; and the choice of sampling design within strata.

Solutions have also been suggested to most of these problems by different authors. On the choice of stratification variable, [7] enjoined the use of the frequency distribution of the study variate itself as stratification variable if available or that of an auxiliary variable X which is highly correlated with the study variate and perhaps the value of variable Y at a recent census. Same view was expressed by [8]. The significance of a highly correlated auxiliary variable as a choice stratification variable was examined by [4] while [9] took it to a logical conclusion in a multivariate stratification study. In some studies, the study variable is also used as the stratification variable; [2, 4, 10, 11, 15]. Therefore, this study also assumes the study variable for stratification.

On the choice of number of strata L to be formed, in most cases, it is predetermined in order to attain a specified level of precision. However [7], developed a model representing the approximate reduction in the precision of Stratified Sample mean compared to that obtained with Simple Random Sampling and concluded that beyond six strata $L \geq 6$ there is little or no further gain, in terms of precision. This was premised on the following two basic questions [7] said to be considered to efficiently determine the number of strata:

(a) at what rate does the variance of $V(\bar{y}_{st})$ decrease as L increases; and

(b) how is the cost of the survey affected by increase in L?

Thus, when there is no appreciable gain in precision with additional strata, then optimum number of strata has been reached and when the cost of sampling additional strata is already overshooting the survey budget, it is obvious that the number of strata to be surveyed should be limited to the one covered by the survey budget.

Furthermore, [4] confirmed that efficient number of strata L can be arrived at by observing the $Var(L)/Var(L-1)$ reduction in variance attained when additional stratum is considered, with the remark that in many multipurpose investigations, only marginal stratification gains could be expected from the use of more than six strata. Nevertheless, [3, 4], [10-14] have all studied methods of strata construction using predetermined number of strata.

In determining the optimum stratum boundaries, [6] stated that we are absolutely free to choose the number of strata we desired, which is opposed to a situation in which the strata have been predetermined e.g. geographically or/an administrative stratification. This study allows for optimum number of strata as suggested by [6], while deep stratification is attained when $N_h = 1$ in any or all of the strata.

The next operation is the mode of stratification, i.e. strata boundary determination. Various methods had been reported in the literature for determining strata boundaries. Dalenius, T. [15] took the lead; Equalization of strata Totals (EST) was developed by [2]. Others are Ekman's Rule (EKR) [17], Durbin's Rule (DUR) [18], Sethi's Rule (STR) [19], Thomson Rule (TNR) [20], Lavallée and Hidiroglou Method (LHM) [21], Extended Ekman's Rule (EEKR) by [11], Random Search method (RSM) was due to [13], Geometric Stratification (GMS) by [3] and Genetic Algorithm (GA) by [14]. Of all the aforementioned, DHR and GMS are popularly in use for ease of application and precision and therefore form the basis of comparison with LS in this study.

On the problem of allocation of sample to stratum, literature had extensively dwelt on the subject matter with the following result: Optimum allocation was due to [22], proportional and equal allocations have been traditionally long in use. Compromise allocation was by [23]; it was used and improved upon by [24]. Power allocation was used by [21] while Genetic Allocation (GA) was developed and used by [14]. This study makes use of the optimum allocations for its highest precision and yielding minimum MSE estimate for LS when compared with other competing methods of strata construction.

On the choice of sampling design within the stratum, literature concentrated on the use of Simple Random Sampling with or without replacement [1, 3, 11, 12]. While [26] report that systematic sampling is more ideal for sample selection within strata for its precision when units in each stratum are arranged in the order of magnitude before selection. This study makes use of Simple Random Sampling without replacement in each stratum.

Now, this paper is organized as follows: the second section deals on materials and methods, it describes the four sets of data used for the study, the algorithm of the three methods of strata construction studied as well as method of estimation. The third section presents the results and discussion of the analysis while the last section gives the concluding remarks on this study.

Materials and methods

In this section, we examine our Linear Stratification (LS) in comparison with other competing methods of strata construction, i.e. DHR and GMS. The data structure reflects varying degree of skewness. These four (4) sets of life data whose features are reflected in Table 1 are used for this study:

i. Overall Cumulative Average Scores (OCAS) 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, Nigeria.
iii. Grants allocation to 774 Local Government's Council in Nigeria 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).

Methods of stratification

Dalenius and Hodges Rule (DHR)

DHR evolves as an approximate solution to Dalenius equation [2] and [15]. It requires us to choose equal class interval, obtain the cumulative square root of the frequency (cum$\sqrt{(fy)}$) of the study variate and determine the strata boundaries by dividing the total cumulative square root of the frequency by the required number of strata L and the boundary is placed at this division point. In practice, the boundaries do not fall at the exact point of stratification hence the boundaries are established at the approximate boundary value (ABV). In a recent study, [26] developed an interpolation method that placed the stratification point at the exact boundary value (EBV) yielding more precise estimates than at the ABV. DHR is popularly in use for its precision and ease of application. However, it has been criticized for arbitrariness in the choice of the class interval and the absence of a theory to guide the best interval to use [11].

Geometric Stratification (GMS)

GMS was introduced by [3] as the new and the most frequently used stratification method in the recent past. It was applied to positively skewed populations and the results competes favourably well with DHR. Stratum boundaries are automatically formed with this method once the geometric ratio r is determined.

$$r = \left[\max X_i / \min X_i\right]^{1/L}$$

$$r = \left[x_L / X_o\right]^{1/L} \qquad .\ .\ .\ .\ 1$$

Where X_L is the largest variate and X_0 is the smallest value of X.

The boundaries are at the points K_h:

Minimum $K_0 = a, ar, ar^2 \ldots, ar^L =$ Maximum K_L

The general term is $K_h = ar^h$

$$h = (0, 1, 2, \ldots, L) \qquad .\ .\ .\ .\ 2$$

Using the relation (1) on data 1, e.g. for two strata situation:

$$X_0 = 44.7, \quad X_L = 68.8, \ L = 2$$

Then, $r = [68.8/44.7]^{1/2}$
$= 1.5391$

Strata boundaries are at the points $K_h = ar^h$, h = (0, 1, 2, . . . , L)

$a = X_0 = 44.7, \ r = 1.5391$ thus,
for $h = 0$, $K_h = K_0 = 44.7 * (1.5391)^0 = 44.7 * 1 = 44.7$
for $h = 1$, $K_1 = 44.7 * (1.5391)^1 = 55.5$
for $h = 2$, $K_2 = 44.7 * (1.5391)^2 = 68.8$

Using the relation (1) and (2) on data 1-4, strata boundaries were established by GMS for two through ten strata. It has been established that GMS has its limitations in the fact that it does not work for normal distributions. Also, it does not work well with variables that have very low starting points as this will lead to too many small strata [3], hence sample estimation is impossible in early strata formations by GMS.

Linear stratification (LS)

It has been stated that optimum boundaries are attained when the Coefficient of Variation (CV) are approximately equal in all the strata [27]. Subsequently, [3] developed a recurrence relation based on submission of [27] to derive the GMS. Our proposed LS algorithm toed the path of assumptions of [27] and [3].

Similarly, the algorithm of strata boundary determination of minimum sample size for a given precision developed by [12] and the LHM by [21] assumed equality of CV.

Linear stratification: mathematical background

To stratify a population of size N into strata of size N_1, $N_2, \ldots, N_L$. Let $X_1, X_2, \ldots, X_N$ be observations of the stratification variable *x* highly correlated with the study variate Y with observations $Y_1, Y_2, \ldots, Y_N$. Suppose the population size *N* is subdivided into intervals, with end points $b_0 < b_1 < \ldots < b_L$. In order to make the breaks $(b_0, b_1, \ldots, b_L)$ for any given b_0 and b_L we use the recommendation of [27] as adopted by [3].

We seek equality of co-efficient of variation CV for all the strata, i.e.

$$CV_h = \frac{\sigma_h}{\bar{X}_h} \quad \text{for, } h = 1, 2, \ldots, L \qquad \ldots .3$$

Where σ_h, the standard deviation of the stratification variable *X* in the h^th stratum is estimated by S_h and $\bar{X}_h$ is the mean of the stratification variable X for the hth stratum. To attain homogeneity of units within the stratum, we further assume that the probability distribution within each stratum is approximately uniform. Thus, if $X \sim U(a, b)$, the mean and the standard deviation are $\bar{X}_h = \frac{(b+a)}{2}$ and $S = \frac{1}{\sqrt{12}}$ $(b - a)$ respectively. Thus for boundaries $b_0, b_1, \ldots, b_L$. The mean and variance of hth stratum is given as:

$$\bar{X}_h = \frac{b_h + b_{h-1}}{2} \quad \text{and} \quad S_h = \frac{1}{\sqrt{12}}(b_h - b_{h-1})$$

therefore,

$$CV_h = \frac{(b_h + b_{h-1})/2}{(b_h + b_{h-1})/\sqrt{12}} \quad \text{for } h = 1, 2, \ldots, L$$

With equity of CV_h therefore, we have

$$\frac{b_{h+1} - b_h}{b_{h-1} + b_h} = \frac{b_h - b_{h-1}}{b_h + b_{h-1}} \qquad \ldots .4$$

as obtained by [3].

They stated further that this new and recurrence relation (4) reduces however to

$$b_h^2 = b_{h+1} * b_{h-1} \qquad \ldots .5$$

See relation (8) in [3] and therefore chosen the stratum boundaries K_h as terms of geometric progression.

$$k_h = ar^h, \quad h = 0, 1, 2, \ldots, L-1 \qquad \ldots .2$$

The recurrence relation (4) obtained by [3] based on the aforementioned assumption satisfy the requirement of our algorithm. However, we defer on the geometric progression of the strata boundaries, as our empirical investigation has shown that the most efficient boundaries are reached when the sequence is on linear progression.

Furthermore, it was stated by [3] and [29] that the GMS created wide gaps within strata and this makes one doubt the genuineness of homogeneity of units within the strata. With the wide gaps, there will be high variability among the units in each stratum. Literature has overstressed the need for homogenous units within the stratum. Stratified sampling is said to be at its best when the strata are internally homogenous [6].

Linear stratification algorithm is as follows:

A. For a normal population

Optimum Points of Stratification (OPS) for some standard distributions has been developed by [19] (Normal, Beta, Gamma and various *chi*-squares). He also tabulated the optimum boundaries for Neyman, equal and proportional allocations for $L \leq 6$ (for Gamma distribution) and $L \leq 10$ for normal distribution. He advised that using STR requires the knowledge of the shape of the distribution of the population units [19].

Similarly, coefficient of skewness speaks for the departure of any given sets of data from normality. Skewness of zero indicates that the distribution is balance hence symmetric. Coefficient of skewness not very far away from zero shows that the data set is approximately normal. The importance of visual display of data sets in determining it skewness and using the appropriate statistic has also being stressed by [30]. Therefore, when the coefficient of skewness is < 2 and the frequency distribution plot of the data sets reflect a normal or approximately normal distribution we apply the following procedure.

For a normal population, Let $X_1, X_2, \ldots, X_N$ be units of the stratification variable *X* which is highly correlated with the study variate *Y* with units $Y_1, Y_2, \ldots, Y_N$. We assume $(X = Y)$ like in [2, 4, 10, 11, 15]:

i. Arrange the variables in ascending order of magnitude.
ii. Take the least value in the series as b_0 and the Largest value as b_L in the (L+1) terms forming the population units.
iii. Obtain the range of the series $R = b_L - b_0$
iv. Obtain the common range difference d_r by dividing R by the number of strata L desired,

thus,

$$d_r = R/L \qquad \ldots.6$$

v. Obtain the strata boundaries using the relation.

$$b_h = b_0 + h * d_r \qquad \ldots.7$$

This is the general term for h = 0, 1, 2, . . . , L.

B. For a positively skewed data i.e. Coefficient of Skewness > 2

The following steps of the algorithm are applicable:

i. Arrange the variables in ascending order of magnitude.

ii. Plot the variables against their serial number.

iii. Identify the variable at which disjoint occur along the curve.

iv. Take the least value in the series as b_0 and the variable at the point of disjoint as b_K in the new $(L – 1)$ terms forming the population units.

v. Variables $X_1, \ldots, X_K$ form stratum 1 while variables $X_{K+1}, \ldots, X_N$ forms the second stratum in two strata case and remain constant as the last stratum for all other $L – 1$ strata formations.

vi. Obtain the range of the new series $R^* = b_K – b_0$.

vii. Obtain the common range difference d^*_r by dividing R^* by the number of $(L – 1)$ strata desired, thus,

$$d_r^* = R^*/(L-1) \qquad \ldots.8$$

viii. Obtain the strata boundaries using the relation.

$$b_h = b_0 + h * d_r^* \text{ for } h = 0, 1, 2, \ldots, L \qquad \ldots.9$$

The justification for this algorithm is that for a positively skewed data set, there is a great departure from normal distribution occasioned by some extraneous variables (outliers), hence deep stratification may occur within three strata formation if the procedure for normal population is applied.

Numerical examples

According to the data reported by [3] in a four strata formation where:

$L = 4, K_0 = b_0 = 5; K_L = b_L = 50{,}000.$

$R = b_L – b_0 = 50{,}000 – 5 = 49995.$

$d_r = R / L = 49995/4 = 12499.$

Thus, using Relation (7) above,

$b_0 = 5 + o * 12499 = 5.$

$b_1 = 5 + 1 * 12499 = 12504.$

$b_2 = 5 + 2 * 12499 = 25003.$

$b_3 = 5 + 3 * 12499 = 37502.$

$b_4 = 5 + 4 * 12499 = 50{,}000.$

We obtain the following strata boundaries:

5 – 12,504; 12,505 – 25,003; 25,004 – 37,502; 37,503 – 50,000.

Compared to boundaries obtained by [3] at:

5 – 50; 51 – 500; 501 – 5,000 and 5,001 – 50,000

Similarly, for the data reported by [29] for three strata formation, we obtain our strata boundaries as follows; where $a = K_0 = b_0 = 40$, $K_L = b_L = 28{,}000$ and $L = 3$.

$R = b_L – b_0 = 28{,}000 – 40 = 27960.$

$d_r = R/L = 27960/3 = 9320.$

Thus, $b_0 = b_0 + 0 * 9320 = 40.$

$b_1 = 40 + 1 * 9320 = 9360.$

$b_2 = 40 + 2 * 9320= 18680.$

$b_3 = 40 + 3 * 9320 = 28000.$

Our strata boundaries are:

40 – 9,360; 9,361 – 18,680; 18,681 – 28,000.

Compared to those of [29] at:

40 – 354; 355 – 3,152; 3,153 – 28,000.

It could be observed from the two examples that the strata formed by GMS are really moving in geometric order, creating too wide gaps within the strata. This fact was also acknowledged by [3] and [29]. The former states that "as the values of the variable increases, the stratum width increases geometrically" while the latter mentioned that "this makes it appropriate to take small intervals at the beginning and large intervals at the end". Therefore, these large intervals and geometric stratum width eliminate the concept of homogeneity of units within the stratum and that assumption of uniform distribution within the stratum may not hold any longer unlike our LS that maintains equidistant within the stratum. It further implies that with little population units N_h in the early stratum, zero sampling units may be allocated thereby forming stratum where no unit is sampled.

Strata formations

For the sets of data used in this study (see Table 1), the coefficient of skewness of data 1 = 0.712. This is not too far away from zero and the frequency distribution plot on its histogram tends to normal distribution, hence approximated as a normal distribution. Therefore relation (7) above is applied:

$K_0 = b_0 = 44.7; K_L = b_L = 68.8.$

$R = b_L - b_0 = 68.8 - 44.7 = 24.1.$

$d_r = R / L = 24.1 / L$

when two strata are required,

$d_r = R / L = 24.1 / L = 24.1/2 = 12.05$

Therefore, strata boundaries are:

$K_h = b_0 + h * d_r$

$K_0 = 44.7 + 0 * 12.05 = 44.7.$

$K_1 = b_0 + 1 * 12.05 = 44.7 + 12.05 = 56.75.$

$K_2 = b_0 + 2 * 12.05 = 44.7 + (2* 12.05) = 68.8.$

For data 2-4, their coefficient of skewness is greater than 2 (see Table 1), this indicate a positively skewed data, hence LS Procedure B is applied. The scatter plots of ordered observations of data 2-4 are as shown in Figures 1, 2 and 3. The points of disjoint on the curve are the observation whose labels appear on the curve. Thus, the new upper limit b_K are the observation before those shown on the plots. The common range difference is obtained using relation (8)

$$d_r^* = R^* /(L-1) \qquad \ldots.8$$

and the strata boundaries are obtained using relation (9)

Table 1. Summary statistics of the data used in this study.

Data	N	n	Range	Coefficient of Skewness	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

$$b_h = b_o + h * d_r^* \qquad \ldots.9$$

The outliers on the plots brings about deep stratification quickly, i.e. $N_h = 1, \forall \; L \leq 3$ when relation (7) is used. Hence, the procedure differs depending on the degree of skewness of a data set. The conclusion drawn from this method is that it creates equal intervals between the strata which is one of the features of DHR unlike GMS that creates geometric gaps between strata. Our last stratum when relation (9) is used could be likened to the "take-all stratum" of [21].

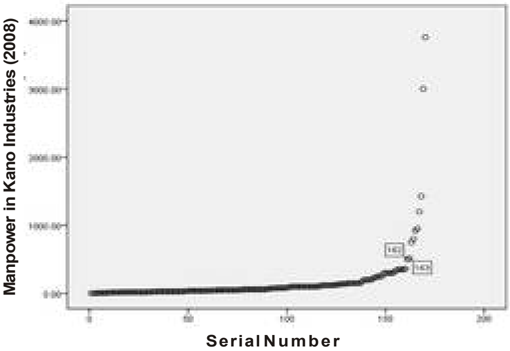

Figure 1. Simple scatter Plot of Ordered Observation Data 2.

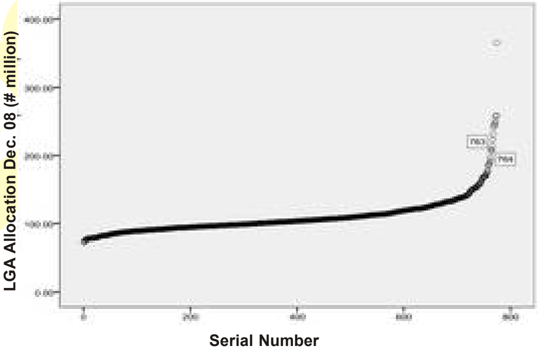

Figure 2. Simple scatter Plot of Ordered Observation Data 3.

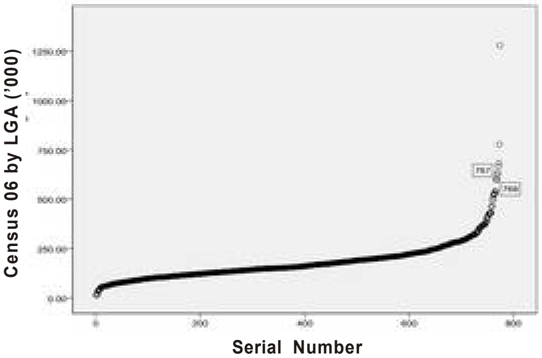

Figure 3. Simple scatter Plot of Ordered Observation Data 4.

Estimation procedure

This section discusses estimation procedure in stratified random sampling. Symbols and notations of [7] were adopted in this study.

Notations

The subscript h denotes the stratum and i the unit within the stratum.

L = Number of strata.
N_h = Total number of population units in stratum *h*.
n_h = Number of sampled units in stratum *h*.
N = Total number of population units in all the L strata.
n = Sample size of the study.
Y_{hi} = is the observation of the i^{th} unit in the h^{th} stratum.
W_h = N_h/N = stratum weight (population units).
w_h = n_h/n = stratum weight (sample units).

Optimum allocation is employed in this study to allocate fixed sample sizes into the strata.

The expression for the optimum allocation is given as:

$$n_h = \frac{nN_hS_h}{\sum N_hS_h} \quad \text{. . . . 10}$$

With the variance as:

$$V_{opt} = V_{\min}(\bar{y}_{st}) \frac{(\Sigma W_hS_h)^2}{n} - \frac{\Sigma W_hS_h^2}{N} \quad \text{. . . . 11}$$

$$MSE(\bar{y}_{st}) = V(\bar{y}_{st})_{opt} + [\sum(w_h - W_h)\bar{Y}_h]^2$$

$$= V(\bar{y}_{st}) + [Bias]^2 \quad \text{. . . . 12}$$

When optimum allocation is used,

$$MSE(\bar{y}_{st}) = V(\bar{y}_{st})_{opt} + \left[\sum(W_h - W_h)\bar{Y}_h\right]^2 \quad \text{. . . . 13}$$

It should be noted that the true stratum weight is known and applied in this study.

In stratum where optimum allocation produces n_h (stratum sample sizes) which are larger than the stratum size N_h. (i.e. when $n_h > N_h$) the revised optimum allocation is used [7].

$$\text{Ropt} = \tilde{n}_h = (n - N_i)\frac{N_hS_h}{\Sigma N_hS_h} \quad \text{. . . . 14}$$

Where *i* is the stratum in which $n_h > N_h$.

e.g. if $n_1 > N_1$ then, for $h \geq 2$ $\tilde{n}_h = (n - N_i)\frac{N_hS_h}{\Sigma N_hS_h}$.

If more than one stratum is involved, the entire affected strata where $n_h > N_h$ are deducted from sample size n to obtain Ropt allocation using relation (14) above. Expression for the variance of Ropt allocation is given as:

$$V_{Ropt}(\bar{y}_{st}) = \frac{(\Sigma' W_hS_h)^2}{n'} - \frac{\Sigma' W_hS_h^2}{N} \quad \text{. . . . 15}$$

where n' is the revised total sample size and Σ' is the summation over the strata in which $\tilde{n}_h < N_h$.

Thus, relation (15) fits back into relation (13) to obtain $MSE(\bar{y}_{st})$ for strata formations where Ropt allocation is used.

Results and discussions

Table 1 gives the descriptive statistics of the four sets of data used in this study. The population size of each data set is as reflected in the second column followed by the sample sizes. The coefficient of skewness of the data set is as shown in the fifth column which shows that the study makes use of positively skewed and normal population.

Number of strata

The numbers of strata formed depend on the structure of the data sets or the variability among the units. For the four (4) sets of data used, DHR formed five strata for data 1 and 2, six strata for data 3 and 4 while GMS and LS formed ten strata for each of the four sets of data except in data 3 where GMS formed five strata. This study allows for optimum number of strata and the stratification process were continued until when deep stratification occurred, i.e. $N_h = 1, \forall h =$ 1, 2, . . ., L (at least one population unit in one or more stratum). However, sample estimation was restricted to strata formation in which $n_h \geq 2$. With less than 10 population units in stratum I of four strata formations and above by GMS, optimum allocation gives zero sample units to the first stratum and to the first two strata in eight strata formation for data 2 and 4.

Sample estimation

Fixed sample of sizes of 48, 57, 258 and 258 were selected from data 1 to 4 respectively. Sample allocation to stratum employed optimum allocation while simple random sampling without replacement was the sampling scheme used within the strata. In order to obtain relevant statistics for the purpose of estimating the population parameters we use $\mathbf{R}_{2.6.1}$ packages (generating seed of 123).

Estimates from the selected samples were computed to obtain the stratified estimates of the population mean $(\bar{y}_{st})$ and its $MSE(\bar{y}_{st})$. Estimates of the MSE of the population mean are shown in Table 2.

Table 2. MSE of the population mean for the three approaches for the four data sets.

	Data 1			Data 2			Data 3			Data 4		
Strata	DHR	GMS	LS	DHR	GMS	LS	DHR	GMS	LS	DHR	GMS	LS
2	0.87617	0.64637	0.43835	41114.2	50980.6	22943.3	148.217	33.335	5.730	1702.04	533.63	108.33
3	0.10790	0.62326	0.11399	36373.7	34263.7	24691.1	206.897	54.794	25.781	3422.86	1432.74	330.14
4	1.78109	0.63904	0.03716	35700.5	36302.1	24045.0	176.650	74.482	29.928	1943.71	1651.97	714.43
5	0.03496	0.12031	0.02259	30470.8	45593.6	23747.0	115.776	48.911	33.774	2310.21	1705.40	444.76
6		0.06573	0.01882		43252.6	25350.9	52.655		27.577	1485.07	1655.82	311.05
7			0.09549		42345.5	23877.0			30.349		2339.13	294.73
8					41668.5				20.139		2199.67	194.73
9									22.054			148.17
10									26.507			125.62

Table 2 presents the MSE value for the three methods of strata construction studied. For data 1 to 4, MSE (LS) gives the minimum estimate when compared to the values obtained for DHR and GMS. Thus, LS gives the most accurate estimates among the methods of strata construction studied. LS is also estimable in all strata formed unlike the other methods. In terms of computational simplicity, LS and GMS could be accomplished at the same speed unlike DHR. LS also work for normal population (data 1) unlike GMS as stated by [3]. Both GMS and LS have their similarity in the fact that they both break down completely when the lower end point is zero, i.e., when X_0, the smallest value of the variable is zero.

Conclusion and recommendation

Statistical inference has suggested that most accurate estimators are those with minimum MSE and that erroneousness weight in stratified sampling leads to sample estimate that is biased [7]. Therefore, it is ideal to assess the performance of a procedure using the MSE criterion rather than the variance (precision). This study thus, identifies the best method among competing methods of stratification such that the method with the least (minimum) MSE is adjudged the best among the competing methods, that is: $MSE\ (T^*) \leq MSE\ (T)$, $i = 1, 2$, and 3.

Therefore, our new LS in terms of efficiency has the minimum MSE value irrespective of the coefficient of skewness of sets of data used when simple random sampling scheme without replacement is used within the strata and optimum allocation employed (Table 2). This implies that our new LS have the minimum estimates of Bias, i.e.:

i. $Bias(\bar{y}_{st})_{LS} < Bias(\bar{y}_{st})_{T_1}$ and

ii. $MSE(\bar{y}_{st})_{LS} < MSE(\bar{y}_{st})_{T}$.

Where T_i = DHR and GMS.

Therefore, the new LS is the most efficient of these existing methods of strata construction studied using optimum allocation in skewed and normal populations and hereby recommend its usage when accurate estimates are required.

References

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

[2] Dalenius, T., and Hodges, J. L., Jr. 1959. Minimum Variance Stratification *Journal of American Statistical Association 54*: 88-101.

[3] 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.

[4] Hess, I., Sethi, V. K., and Balakrishnan, T.R. 1966. Stratification: A practical investigation. *Journal of American Statistical Association 61*: 74-90.

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

[6] Okafor, F. C. 2002. *Sample Survey Theory With Applications.* Afro-Orbis Publications Ltd. Nsukka, Nigeria.

[7] Cochran, W.G. 1977. *Sampling Techniques, Third Edition.* John Wiley and Sons, New York.

[8] Sukhatme, P.V., Sukhatme, B.V., and Asok, C. 1984. *Sampling Theory with Applications. 3rd Edition*, Iowa University Press, USA.

[9] Kish, L., and Anderson, A. W. 1978. Multivariate and Multipurpose Stratification. *Journal of American Statistical Association 73, (361)*: 24-34.

[10] Ghosh, S. P. 1963. Optimum Stratification with Two Characters. *Annals of Mathematical Statistics, 34* : 866-872.

[11] Hedlin, D. 2000. A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics 6 (1)*: 15-29.

[12] Hidiroglou, M. A. 1986. The construction of self-representing stratum of large units of survey design. *The American Statistician 40 (1)*: 27-31.

[13] Kozak, M. 2004. Optimal stratification using random search method in agricultural surveys. *Statistics in Transition, 6(5)* : 797-806.

[14] 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.

[15] Dalenius, T. 1950. The problem of optimum stratification. *Skandinavisk Akturietidskrift*, *33*: 203-213.

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

[17] Ekman, G. 1959. An approximation useful in univariate stratification. *Annalsof Mathematical Statistics, 30*: 210-229.

[18] Durbin, J. 1959. Review of sampling in Sweden. *Journal of Royal Statistical Societies. A. 122*: 246-248.

[19] Sethi, V. K. 1963. A note on optimum stratification for estimating the population means. *The Australian Journal of Statistics, 5*: 20-33.

[20] 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.

[21] Lavallée, P. and Hidiroglou, M. A. 1988. On the stratification of skewed populations. *Survey Methodology 14 (1* : 33-43.

[22] 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.

[23] Chartterjee, S. 1968. Multivariate stratified surveys. *Journal of the American Statistical Association 63*: 530-535.

[24] Kharn, M. G. M., and Ahsan, M. J., 2003. A note on optimum allocation in multivariate sampling. *South Pacific Journal of Natural Science B(21):* 91-95.

[25] Kareem, A. O., Oyeyemi, G. M., and Adewara, A. A. 2015. On the choice of an efficient sampling scheme within strata ICASTOR. *Indian Journal of Mathematical Science 9 (1):* 15-32.

[26] Kareem, A. O., Oshungade, I. O. and Oyeyemi. G. M. 2016. An improvement on some approximate solutions to Dalenius equation. *Al-Hikmah Journal of Pure and Applied Sciences 2 (3):* 130-142.

[27] Cochran, W.G. 1961. Comparison of method for determining stratum boundaries. *Bulleting of the International Statistical Institute, 38(2)*: 345- 358.

[28] Gunning, P., Horgan, J. M., and Keogh, G., 2006. Efficient pareto stratification. *Mathematical Proceedings of Royal Irish Academy 106A (2)*: 131-138.

[29] Gunning, P., Horgan, J. M., and Yancey, W. 2004. Geometric stratification of accounting data. *J. de contaduria Y. Administration, 214, Septembrie - Diciembre*

[30] Doane, P.V. and Seward, L. E. 2011. Measuring skewness: A forgotten statistic? *Journal of Statistics Education, 19 (2):* 1-18.

Journal of Science Research
Volume 15, 2016, ISSN 1119 7333
Citation: Kareem, A. O., Oshungade, I. O. and Oyeyemi, G. M.
On linear stratification of skewed and normal populations.

Textflow Limited
Ibadan, Nigeria