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

1Adebola Femi Barnabas and 2Ajayi Olusola Sunday

1Department of Statistics, Federal University of Technology, Akure, Ondo State, Nigeria and 2Department 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.
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):

   \[ \text{V(\bar{y}_{st})}_{\text{SRSWOR}} = \frac{L}{n} \sum (W_k^2 S_y^2) - \frac{1}{N} \left( \sum W_h S_y^2 \right) \]

The variance of stratified sampling (Proportional)

\[ \text{V(\bar{y}_{st})}_{\text{prop}} = \frac{1-f}{n} \left( \frac{n}{N} \right) S_y^2 \]

And the variance of stratified sampling (Optimal/Neyman)

\[ \text{V(\bar{y}_{op/ney})} = \frac{(\sum W_h S_y h)^2}{n} - \frac{(\sum W_h S_y^2)^2}{N} \]

Then;

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

**Proof**

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

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

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

From analysis of variance:
\[ S^2 = \sum_{i=1}^{N}(Y_i - \bar{Y})^2 \] If \( N \to \infty \)

\[ = NS^2 \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h + \bar{Y})^2 \]

\[ = NS^2 \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2 + 2 \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^1 (\bar{Y} - \bar{Y}) \]

\[ = NS^2 \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_h - \bar{Y})^2 \]

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

\[(N_h - 1) S_h^2 = \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y})^2 \]

\[ : NS^2 \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^{k} \sum_{i=1}^{N_h} (Y_h - \bar{Y})^2 \]

\[ S^2 = \sum_{h=1}^{k} \frac{(N_h - 1) S_h^2}{N} + \sum_{i=1}^{k} \sum_{i=1}^{N_h} \frac{(Y_h - \bar{Y})^2}{N} \]

If \( N_h \to \text{large} \Rightarrow N_h - 1 = N_h \)

\[ S^2 = \sum_{h=1}^{k} \frac{(N_h) S_h^2}{N} + \sum_{i=1}^{k} \sum_{i=1}^{N_h} \frac{(Y_h - \bar{Y})^2}{N} \]

Recall:

\[ \text{var} (\bar{Y}_{st})_{SRSWOR} = \frac{N - n}{N_n} S^2 \]

\[ \frac{N - n}{N_n} S^2 = \frac{N - n}{N_n} \sum_{h=1}^{k} W_h S_h^2 + \frac{N - n}{N_n} \frac{N_h}{N} \sum_{i=1}^{N_h} (Y_h - \bar{Y}) \]

\[ = (\sum_{h=1}^{k} \frac{W_h S_h^2}{n} - \sum_{h=1}^{k} \frac{W_h S_h^2}{n}) + \left( \frac{1 - f}{n} \right) \sum_{h=1}^{k} \sum_{i=1}^{N_h} \frac{(Y_h - \bar{Y})}{n} \]

\[ \text{var} (\bar{Y}_{st})_{SRSWOR} \geq v (\bar{Y}_{st})_{prop} \]
Now compare the variance of optimal/Neyman allocation with the variance of proportional allocation under stratified sampling

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

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

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

\( \therefore \text{var}(\bar{y}_{st})^{opt} \leq \text{var}(\bar{y}_{st})^{prop} \leq \text{var}(\bar{y}_{st})^{SRSWR} \)

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

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

\[ n_h = \frac{nN_h S_{yh}}{\sum N_h S_{yh}} \]

Its corresponding variance of mean stratified is given as;

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

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).
## Table 1:
Summary table:

<table>
<thead>
<tr>
<th>Table No</th>
<th>Description of Data</th>
<th>Distribution</th>
<th>N</th>
<th>n</th>
<th>Variance by Allocation Method</th>
<th>Preferred method</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Enrolments in Sec Schools</td>
<td>D. Uniform</td>
<td>304</td>
<td>132</td>
<td>Equal</td>
<td>223.618700</td>
</tr>
<tr>
<td>2</td>
<td>Performances of Schools in WASCE</td>
<td>Geometric</td>
<td>304</td>
<td>132</td>
<td>Proportional</td>
<td>6.390310</td>
</tr>
<tr>
<td>3</td>
<td>Average Annual Maximum Temperature in Ondo State</td>
<td>Gen. Pareto</td>
<td>101</td>
<td>78</td>
<td>Equal</td>
<td>0.0008998160</td>
</tr>
<tr>
<td>4</td>
<td>Simulated Data</td>
<td>D. Uniform</td>
<td>100</td>
<td>78</td>
<td>Optimum</td>
<td>593.737200</td>
</tr>
<tr>
<td>5</td>
<td>Simulated Data</td>
<td>D. Uniform</td>
<td>500</td>
<td>216</td>
<td>Proportional</td>
<td>762.142500</td>
</tr>
<tr>
<td>6</td>
<td>Simulated Data</td>
<td>D. Uniform</td>
<td>1000</td>
<td>270</td>
<td>Optimum</td>
<td>820.593900</td>
</tr>
<tr>
<td>7</td>
<td>Simulated Data</td>
<td>D. Uniform</td>
<td>5000</td>
<td>366</td>
<td>Proportional</td>
<td>651.184600</td>
</tr>
<tr>
<td>8</td>
<td>Simulated Data</td>
<td>Geometric</td>
<td>100</td>
<td>78</td>
<td>Proportional</td>
<td>0.005805</td>
</tr>
<tr>
<td>9</td>
<td>Simulated Data</td>
<td>Geometric</td>
<td>500</td>
<td>216</td>
<td>Optimum</td>
<td>6.215440</td>
</tr>
<tr>
<td>10</td>
<td>Simulated Data</td>
<td>Geometric</td>
<td>1000</td>
<td>270</td>
<td>Optimum</td>
<td>6.331240</td>
</tr>
<tr>
<td>11</td>
<td>Simulated Data</td>
<td>Geometric</td>
<td>5000</td>
<td>366</td>
<td>Proportional</td>
<td>6.457780</td>
</tr>
<tr>
<td>12</td>
<td>Simulated Data</td>
<td>Gen. Pareto</td>
<td>100</td>
<td>78</td>
<td>Optimum</td>
<td>0.005991</td>
</tr>
<tr>
<td>13</td>
<td>Simulated Data</td>
<td>Beta</td>
<td>500</td>
<td>216</td>
<td>Optimum</td>
<td>0.006208</td>
</tr>
<tr>
<td>14</td>
<td>Simulated Data</td>
<td>Beta</td>
<td>1000</td>
<td>270</td>
<td>Optimum</td>
<td>0.006455</td>
</tr>
<tr>
<td>15</td>
<td>Simulated Data</td>
<td>Beta</td>
<td>5000</td>
<td>366</td>
<td>Optimum</td>
<td>0.005805</td>
</tr>
</tbody>
</table>
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 Coventry University


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


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


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


