ISSN:
Print - 2277 - 0593
Online - 2315 - 7461

Journal of Natural Science, Engineering and Technology

COMPARING THE OPTIMAL ALLOCATION IN STRATIFIED AND POST-STRATIFIED SAMPLING USING MULTI-ITEMS

F. S. APANTAKU

Department of Statistics, Federal University of Agriculture, Abeokuta, Nigeria
***Corresponding author:** fsapantaku@yahoo.com **Tel:** +2348037183877

ABSTRACT

Usually in sample surveys more than one population characteristics are estimated (multi-item). These characteristics may be of conflicting nature. Optimal allocation using a stratified random sample solves the statistical problem that may be found with proportional allocation, by ensuring that enough respondents are studied in each segment to provide the highest level of accuracy for the overall results. The study compared optimum allocation in stratified and post stratified sampling using multi-items and determined the variations of the components of the multi-items in the proposed model. The idea of optimum allocation based on the multi-items was approached using a linear programming problem that minimizes the covariance of the stratified variable subject to a fixed cost. The covariance matrix was defined based on the four socio-economic characteristics of 400 heads of household in Abeokuta South and Ijebu North Local Government Areas of Ogun State, Nigeria. The characteristics were occupation, income, household size and educational level. The data from the survey was transformed for each of the four characteristics. The estimates used in the computation were calculated using statistical analysis software Splus. From the analysis, it was seen that for both Abeokuta and Ijebu data sets, the variance based on the four characteristics as multivariate is less than that of the variables when considered as a univariate. From the results, it was seen that there was no difference in the percentage of the total variance accounted for by the different components from the merged sample when compared with the individual sample.

Key words: Optimum allocation, stratified sampling, post stratified sampling, multi-items, optimization

INTRODUCTION

In social research, special emphasis is placed on the comparative and analytical use of samples. Knowledge, attitudes, and actions in life everyday are based to a very large extent on samples (Cheang, 2011; Cochran, 1977). In survey, samples are used instead of population and most of these samples are prepared by Statisticians and one of the areas of Statistics that is most commonly used in all fields of scientific investigation is that of probabilistic sampling. Probability sample designs can be made better with features to assure representation of population subgroups in the samples and stratification is one of such features. Surveys used by social scientists are based on complex sampling designs (Lumley, 2004; Winship and Radbill, 1994).

One of the main problems in sampling survey is the optimal allocation of resources. Usually, the solution of such problem is rather arbitrary due to the fact that no best allocation is defined. In terms of this model, the allocation problem is to find the allocation of a sample to strata which minimizes cost of investigation subject to a given condition about the sampling error.

An effective sampling technique within a population represents an appropriate extraction of useful data which provides meaningful knowledge of the important aspects of the population (Diaz-Garcia and Cortez, 2006). Probability sampling methods are usually designed to be measurable, that is, so designed that statistical inference to population values can be based on measures of variability, usually standard errors, computed from the sample data.

Empirical research may be performed in different ways: by haphazard observations, controlled observations, experiments, or surveys. This study concentrates on sampling for surveys. Survey research is aimed at estimating specified population values. A population value is a numerical expression that summarizes the values of some characteristic(s) for all elements of an entire population. It is a summary measure of some features of the distribution of the variable(s) in the defined population. The population is defined jointly with the elements. The population is the aggregate of the elements, and the elements are the basic units that comprise and define the population. The population must be defined in terms of content, units, extent and time (Cheang, 2011).

The survey population achieved may differ somewhat from the desired target population. The major difference frequently arises from non responses and non coverage and only the survey population is represented in the sample. The sample provides statistical inference to the survey population. Behind every survey population stands some hypothetical universe, explicit or implicit, definite or indefinite. Characteristics of population elements are transformed to variables by the survey operations of measurement and a statistic based on the variables found in a sample results in a random variable which is called a variate. Any population value is determined by four factors (a) the defined survey population; (b) the nature of the survey variable (s) and their distributions in some cases; (c) the method of observations; and (d) the mathematical expression for deriving the population value from the individual element values (Kolenikov, 2010).

In random sampling, stratified random sampling can produce a more concentrated distribution of estimates. This suggests that stratification can produce sample statistics which are more precise or which have smaller error due to sampling than simple random samples. Stratification in this study is to allow the investigation of the characteristic of interest for particular subgroups by ensuring adequate representation from each subgroup of interest (Lumley, 2004; Sethna and Groeneveld, 1984).

Stratification by making use of existing knowledge concerning the population provides an effective means of reducing the measured variability of estimates. The population variability itself does not change but the technique for calculating variability via stratification offers the possibility, when done successfully, of reducing the variance estimate of the population (Sethna and Groeneveld, 1984). Lower measured variability

permits the researcher to make a more precise estimate or it enables one to make an interval estimate at a lower cost. The stratification technique is also an effective means of assuring representation in the sample from each stratum in the population. Simple random sampling techniques offer no assurances that each segment of the population is represented. Stratification and subsequent sampling with strata assure this representation.

Post-stratification may be used on the subclasses even if a proportionate sample of the entire population has been selected. Post stratification is an example of improving the estimator by the proper utilization of ancillary sources of information. The method of using stratification is to increase the precision of the sample mean as in contrast to proportionate sampling. It involves the deliberate use of widely different sampling rates for the various strata. Optimum stratification is used when the standard deviations of the population strata are known to differ substantially. This technique is a method of allocating larger size samples to those strata with larger standard deviations. The designation optimum allocation to disproportionate sampling refers to the aim of assigning sampling rates to the strata in such a way as to achieve the least variance for the overall mean per unit of cost (Diaz-Garcia and Cortez, 2006).

Usually in sample surveys more than one population characteristics are estimated (multi-item). These characteristics may be of conflicting nature. Stratification may produce a gain in precision in the estimates of characteristics of the whole population. It may possibly divide the heterogeneous population into sub-populations, each of which it internally homogeneous.

Optimal allocation using a stratified random sampling solves the statistical problem that may be found with proportional allocation, by ensuring that enough respondents are selected and studied in each stratum to provide the highest level of precision for the overall results.

The objectives of the study are to:

a. Compare optimum allocation in stratified and post stratified sampling using multi-items, and
b. Determine the variations of the components of the multi-items in the model improved tremendously on what I meant on ground, in terms of record keeping which are readily available for checking. proposed.

METHODOLOGY

The procedure for estimation from multiple frames was given by Hartley (1962, 1964). According to Hartley, choosing a simple cost function provides rules for optimal choices of subject to a given value. Saxens *et al*. (1986) considered the extension of Hartley's procedure to the case of two stage sampling of the multi-stage sampling. They worked out optimal choices of the variable of interest considering suitable cost functions and recommended replacement of unknown parameters occurring in the optimal solutions by sample analogues. Hence the problem of small domain statistics and a special method of estimation is needed for the parameters relating to small domains. Bankier (1996) discussed a few issues involved in small area or local area estimation. The problem is how to estimate the domain. These estimators make a minimal use of data that may be available. To improve upon the estimators, the database is broadened and strengths are borrowed from data available on similar domains and secondary external sources. Ac-

cording to Bankier (1996), post-stratified estimators of auxiliary data, is to be used. These post strata may stand for age, sex, or ethnic groups in usual practices.

The multivariate stratified sample design

The multivariate stratified sample design is used for multi-objective surveys in which there is difference among the importance of interest variables. All these method consider the computation of the stratum sample size, which can be computed by various procedures, but optimum allocation has been found to be a useful approach. Optimum allocation in multivariate stratified sampling can be seen as a multi-objective optimization problem and multi-objective optimization problem is a particular case of a matrix optimization.

Khan and Ahsan (2003, 1967) proposed a method in which they formulated multi-objective surveys as a nonlinear programming problem and use a dynamic programming technique to find a solution. One problem with this approach is how to weigh the variances. There is no single solution for doing this, and it is not always easy to predict what the consequences of a particular choice of weights are.

The multivariate optimum allocation

The problem of allocating sample to various strata may be viewed as minimizing the variances of various characters subject to the conditions of the given budget and tolerance limits on certain variances. The problem turns out to be nonlinear programming problem with several linear objective functions and single convex constraint. Pizada and Maqbool (2003), solved the resulting linear programming problem through Chebyshev approximation. The criteria behind the Chebyshev approximation are to find a solution that minimizes the single worst. Suppose that $p-$ characteristics are measured on each unit of a population which is partitioned into L strata. Let n_i, be the number of units drawn from the $i-th$ stratum ($(i=1,2,.....,L).$ For the $j-th$ character an unbiased estimate of the population mean, $\bar{Y}_j$, is $\bar{y}_{jst}$ which has the sampling variance.

$$Var(\bar{y}_{jst}) = \sum_{i=1}^{L} W_i^2 S_{ij}^2 X_i \qquad j = 1,2,...., p \tag{2.1}$$

where

$$W_i = \frac{N_i}{N}, \quad S_{ij}^2 = \frac{1}{N_i - 1}\sum_{h=1}^{N_i}(y_{ijth} - \bar{y}_{ij})^2$$

and

$$X_i = \frac{1}{n_i} - \frac{1}{N_i}, \quad a_{ij} = W_i^2 S_{ij}^2,$$

in usual notation.

Let c_{ij} be the cost of enumerating the $j-th$ characteristic in the $i-th$ stratum and let k be the upper limit on total cost of the survey. Then

$$\sum_{i=1}^{L}\sum_{j=1}^{p} c_{ij} n_i \le k \tag{2.2}$$

The multivariate allocation problem can be stated as

$$z_j = \sum_{i=1}^{L} a_{ij} X_i - \sum_{i=1}^{L} \frac{a_{ij}}{N_i}, \qquad j = 1,2,\ldots., p$$

Minimize

Subject to

$$\sum_{i=1}^{L}\sum_{j=1}^{p} \frac{c_{ij}}{X_i} \le k$$

$$\frac{1}{N_i} \le X_i \le 1, \quad i = 1,2,\ldots..,L \tag{2.3}$$

where $\frac{1}{X_i}$ is used for n_i. If (34) is considered separately for each character, by ignoring the constant term in the objective function, the problem for kth character becomes

$$Z_k = \sum_{i=1}^{L} \frac{a_{ik}}{X_i}$$

Minimize

Subject to

$$\sum_{i=1}^{L}\sum_{j=1}^{p} c_{ij} X_i \le k$$

$$1 \le X_i \le N_i, \qquad i = 1,2,\ldots.....,L. \tag{2.4}$$

By introducing a new variable x_{L+k}, the problem (3.75) transforms to

Minimize $Z_k = x_{L+k}$ (a)

Subject to

$$g_k(X) = \sum_{i=1}^{L} \frac{a_{ik}}{X_i} - X_{L+k} \le 0 \tag{b}$$

$$\sum_{i=1}^{L}\sum_{j=1}^{p} c_{ij} X_i \le k \tag{c}$$

$$1 \le X_i \le N_i, \quad i = 1,2,\ldots.L. \tag{d}$$

(2.5)

The constraints in (2.5b) are convex (Kokan and Khan, 1967) and the constraint (2.5c) and the bounds (2.5d) are linear. The problem (2.5a)-(2.5d) is therefore a convex programming problem with linear objective and can be solved by using any method of convex programming. The Chebyshev approximation formulation of the multiple objective allocation problems in (2.5) is the following linear programming problem (LPP):

Minimize δ

Subject to

$$
\begin{aligned}
&2\sum_{i=1}^{L}\frac{a_{ik}}{X_i^{k(0)}}-\sum_{i=1}^{L}\frac{a_{ik}X_i}{X_i^{k(0)^2}}-X_{L+k}\le 0\\
&\sum_{i=1}^{L}\sum_{j=1}^{p}c_{ij}X_i\le k \qquad l=1,2,.....,t_k\\
&\qquad\qquad\qquad\qquad k=1,2,....., p\\
&X_{L+k}-\delta\le z_k^0\\
&1\le X_i\le N_i \qquad i=1,2,.....,L
\end{aligned}
\tag{2.6}
$$

The p solutions $X_1^0, X_2^0,...., X_p^0$ have been obtained by minimizing the individual objective functions subject to the linearized constraints by letting the minimum values of Z_k to be found as $Z_k^0,\ \ k=1,2,..., p$ at the corresponding minimal points $X_k^0, k=1,2,...., p.$ This gives the aspiration levels being used in Chebyshev approximation.

Formally the problem of optimum allocation in stratified sampling can be presented as a multi-objective, nonlinear optimization as

$$
\min \hat{V}ar(\bar{y}_{st})=\min_{n}\begin{pmatrix}\hat{V}ar(\bar{y}_{st}^1)\\ :\\ :\\ \hat{V}ar(\bar{y}_{st}^G)\end{pmatrix}
$$

Subject to (2.7)

$$c'n+c_0=C$$

Where C is the total cost, c_0 is the fixed cost and $c'=(c_1,.....c_H)$ and $n'=(n_1,n_2,.....n_H)$

The solutions in (2.7) take real values and the sample sizes n_h must be integers. There is the problem of estimating the variance on the basis of the sample size in each stratum and also the problem of over sampling, that is, when $n_h \geq N_h$ for at least some h.

An alternative to (2.7), is given as

$$\min \hat{V}ar(\bar{y}_{st}) = \min_{n} \begin{pmatrix} \hat{V}ar(\bar{y}_{st}^{1}) \\ \vdots \\ \vdots \\ \hat{V}ar(\bar{y}_{st}^{G}) \end{pmatrix}$$

where G is number of characteristics

subject to (2.8)

$$c'n + c_0 = C$$

$$2 \leq n_h \leq N_h \, , h = 1,2,.....,H$$

$$n_h \in \mathrm{N}$$

Where N denotes the set of natural numbers. The methods for resolving a multi-objective optimization programme can be classified by considering the amount of information possessed concerning the study population, with three different scenarios, namely complete, partial or zero information (Steuer, 1986; Miettinen, 1999; Diaz-Garcia and Ulloa, 2006). Diaz-Garcia and Ulloa (2006) consider problem (2.8) from the stand-point of the multi-objective optimization methods by using complete information such as value function and lexicographic, partial information method such as $\varepsilon -$ constraint and also zero information such as the distances.

Optimum allocation via multi-objective optimization

The estimator of the population mean in multivariate stratified sampling for the j -th characteristic is defined as

$$\bar{y}_{st}^{j} = \sum_{h=1}^{H} W_h \bar{y}_h^j \qquad (2.9)$$

Where $\bar{y}_h^j = \frac{1}{n_h} \sum_{h=1}^{n_h} y_{hi}^j$ is the sample mean in stratum h of the j -th characteristic, and y_{hi}^j is the value obtained for the i -th unit in stratum h of the j -th characteristic. The Var $(\bar{y}_{st}^{j})$ is defined using the population variances $S_h^2, h = 1,2,.....,H,$ which are usually unknown, and therefore these are substituted by the sample variances

$s_h^2, h = 1,2,.....,H,$ defined as

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 \tag{2.10}$$

And thus $\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{\,j})$ is substituted by the estimated variance $\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{\,j})$, which is given by

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{\,j}) = \sum_{h=1}^{H}\frac{W_h^2 s_{hj}^2}{n_h} - \sum_{h=1}^{H}\frac{W_h s_{hj}^2}{N} \tag{2.11}$$

Value function

Under the value function technique, programme (2.8) is expressed as

$$\min_{n} v(\hat{V}ar(\bar{y}_{st}))$$
subject to
$$2 \le n_h \le N_h \qquad h = 1,2,.....,H$$
$$n_h \in N \tag{2.12}$$

Where $v(\cdot)$ is a scalar function that summarizes the importance of each of the variances of the G characteristics. Evidently for every problem the value function $v(\cdot)$ may take an infinite number of forms and this constitutes the difficulty for the evaluator in defining such a function. Some simple functions have given excellent results in applications and one of these particular forms is the weighting method. Under the weighting approach, (2.12) can be expressed as

$$\min_{n}\sum_{j=1}^{G}\lambda_j \hat{V}ar(\bar{y}_{st}^{\,j})$$
subject to
$$\sum_{h=1}^{H} c_h n_h + c_0 = C$$
$$2 \le n_h \le N_h, \qquad h = 1,2,.....,H$$
$$n_h \in N \tag{2.13}$$

Such that

$$\sum_{j=1}^{G}\lambda_j = 1, \qquad \lambda_j \ge 0 \ \forall\ j = 1,2,.....,G;$$

where λ_j weighs the importance of each characteristic. In the context of multi-objective optimization, (2.13) is without doubt the method that has been mostly thoroughly studied. Its popularity is due to the fact that the value function is unique. The value function method is utilized for recurrent studies in which over time, the results obtained using (2.13) help in reaching a better inference for future experiments, in which the appropriate weighting can be applied.

Lexicographic method

This method like value function requires complete information on the phenomenon in order to create an important ordered hierarchy of the variances evaluated. Unlike the value function method, it is not necessary to know what weight to allocate to each characteristic, but only the order of importance they represent in obtaining the sample. In this case, to optimize programme (2.8), the variances of the characteristics must be ordered by the evaluator, beginning with the

one presenting the most important characteristics, and then by descending order of importance, thus obtaining

$$\hat{V}ar(\bar{y}_{st}^{i_1}),\hat{V}ar(\bar{y}_{st}^{i_2}),.....,\hat{V}ar(\bar{y}_{st}^{i_G}) \tag{2.14}$$

where $i_1,.....,i_G$ is a permutation with the desired, descending order of the set of super indices 1,2,….,G. To reach the stages in the vector, it is necessary to resolve the following programme

By letting v_2 be the minimum of problem (2.16), for the third stage the problem is resolved as

$$\begin{aligned}
&\min_n \hat{V}ar(\bar{y}_{st}^{i_3}) \\
&\text{subject to} \\
&\sum_{h=1}^{H}\frac{W_h^2 S_{h1}^2}{n_h}-\sum_{h=1}^{H}\frac{W_h S_{h1}^2}{N}=v_1 \\
&\sum_{h=1}^{H}\frac{W_h^2 S_{h2}^2}{n_h}-\sum_{h=1}^{H}\frac{W_h S_{h2}^2}{N}=v_2 \\
&\sum_{h=1}^{H} c_h n_h + c_0 = C \\
&2 \le n_h \le N_h, \quad h=1,2,....., H \\
&n_h \in N
\end{aligned} \tag{2.17}$$

To reach stage G, the next problem to be resolved is as

$$\begin{aligned}
&\min_n \hat{V}ar(\bar{y}_{st}^{i_G}) \\
&\text{subject to} \\
&\sum_{h=1}^{H}\frac{W_h^2 S_{h1}^2}{n_h}-\sum_{h=1}^{H}\frac{W_h S_{h1}^2}{N}=v_1 \\
&\sum_{h=1}^{H}\frac{W_h^2 S_{h2}^2}{n_h}-\sum_{h=1}^{H}\frac{W_h S_{h2}^2}{N}=v_2 \\
&\quad\vdots \\
&\quad\vdots \\
&\sum_{h=1}^{H}\frac{W_h^2 S_{hG-1}^2}{n_h}-\sum_{h=1}^{H}\frac{W_h S_{hG-1}^2}{N}=v_{G-1} \\
&\sum_{h=1}^{H} c_h n_h + c_0 = C \\
&2 \le n_h \le N_h, \quad h=1,2,....., H \\
&n_h \in N
\end{aligned} \tag{2.18}$$

Hence the vector obtained in this stage is the optimum solution to the problem.

Optimal design for a multivariate stratified sampling adopted in the study
The idea of optimal allocation under a multivariate stratified sampling in the study is based on an alternative approach as in Diaz-Garcia and Ramos-Quiroga (2011).
The linear programming problem is assumed to be

$$\min_{n} \theta$$

Subject to

$$\sum_{h=1}^{H} C_h n_h + C_0 = C$$

(2.19)

$$2 \leq n_h \leq N_h$$

Where $\theta = Cov(\bar{y}_{st})$. This is the matrix of variance covariances of the vector
$\tilde{y}_{st} = (\tilde{y}_{st}, \ldots\ldots\ldots\ldots\ldots\ldots \tilde{y}_{st})$.
the sub index h = 1,2, ….. H denotes the stratum i = 1,2,…..N_h or n_h within stratum h and j = 1,2,…., G. denotes the characteristic (variable).

The covariance matrix of $\tilde{y}_{st}$ denoted as cov ($\tilde{y}_{st}$) is defined in matrix

(2.20)

$$cov(\tilde{y}_{st}) = \begin{matrix} Var\ (\tilde{y}^1_{st}) & Cov\ (\tilde{y}^1_{st}\ \tilde{y}^2_{st}) & Cov(\tilde{y}^1_{st}\ \tilde{y}^G_{st}) \\ Cov\ (\tilde{y}^2_{st,}\ \tilde{y}^1_{st}) & Var\ (\tilde{y}^2_{st}) & Cov(\tilde{y}^2_{st}\ \tilde{y}^G_{st}) \\ Cov\ (\tilde{y}^G_{st,}\ \tilde{y}^1_{st}) & Cov\ (\tilde{y}^G_{st}\ \tilde{y}^2_{st}) & Var(\tilde{y}^G_{st}) \end{matrix}$$

and the estimated covariance of $\tilde{y}^i_{st}$ and $\tilde{y}^j_{st}$ as Cov ($\tilde{y}^i_{st}$ and $\tilde{y}^j_{st}$). this Cov ($\tilde{y}^i_{st}$ and $\tilde{y}^j_{st}$) Cov ($\tilde{y}^i_{st}$ and $\tilde{y}^j_{st}$) =

(2.21)

$$Cov\ (\tilde{y}^i_{st},\ \tilde{y}^i_{st}) = \sum_{h=1}^{H} \frac{W_h^2\, S_{hij}}{n_h} - \sum_{h=1}^{H} \frac{W_h\, S_{hij}}{N}$$

and $$Cov\ (\tilde{y}^i_{st},\ \tilde{y}^i_{st}) = \sum_{h=1}^{H} \frac{W_h^2\, S^2_{hi}}{n_h} - \sum_{h=1}^{H} \frac{W_h\, S^2_{hi}}{N}$$

and C_h is the cost per G – dimensional sampling unit in stratum h and its vector
$C = (C_1, \ldots\ldots\ldots\ldots C_G)^1$.

Principal component analysis

Optimal allocation in multi-item is developed as a multivariate optimization problem by finding the principal components. This was done by determining the overall linear combinations that concentrates the variability into few variables. We then search for a set of mutually uncorrelated variables, $Y_1, Y_2, ..., Y_p$ each one being a linear combination of the original set of variables, $X_1, X_2, ..., X_p$. One of the motivations for determining such a collection is in of, if we derive a set that concentrates the overall variability into the first few variables, it is perhaps easier to see what accounts for the variation in the data.

Indeed, if a few of the $\{Y_i\}$ seem to account for most of the variation in the data, then it could be argued that the effective dimensionality is less than P and this could result in a simplified analysis based on a smaller set of variables (Jolliffe, 2002; Khan and Ahsan, 2003; Garcia and Cortez, 2006).

THE EMPIRICAL EXAMPLE (CASE)

Data on four socioeconomic characteristics of 400 heads of households in Abeokuta South and Ijebu North Local Government Areas (LGAs) of Ogun State, Nigeria were investigated. This comprised of 200 households from each LGA. The characteristics were occupation, income, household size and educational level.

RESULTS

The data from the survey was transformed for each of the four characteristics. Occupation was transformed into: (1) unemployed (2) paid employment (3) self employment, while income into: (1) 0 – < N10,000; (2) N10,000 - < N20,000; (3) N20,000 and above. Household size was transformed into: (1) small (1-3); (2) moderate (4-7); (3) large (7 and above), while educational level was transformed into: (1) primary; (2) secondary; (3) tertiary. The estimates used in the computation were calculated using statistical analysis software Splus.

Stratification by making use of existing knowledge concerning the population provided an effective means of reducing the measured variability of estimates. The population variability itself does not change but the technique for calculating variability via stratification offers the possibility. The stratification technique in this study divided up the population into sub-population or strata. The strata for the four characteristics are in Tables 3.1, 3.2, 3.3 and 3.4.

Table 3.1: Stratified Data on Occupation of Heads of Household in both Abeokuta South and Ijebu North

Strata	Occupation	Number in Abeokuta South population	Number in Ijebu North population
1	Unemployed	10	2
2	Paid employment	47	54
3	Self employment	143	144
		200	200

Source: Field Survey, 2012

Table 3.2: Stratified Data on Income of Heads of Household in both Abeokuta South and Ijebu North

Strata	Income N(000)	Number in Abeokuta South population	Number in Ijebu North population
1	0 to under N10,000	42	28
2	N10,000< N20,000	73	91
3	N20,000 and over	85	81
		200	200

Source: Field Survey, 2012

Table 3.3: Stratified Data on Dependant Size of Heads of Household in both Abeokuta South and Ijebu North

Strata	Dependant Size	Number in Abeokuta South population	Number in Ijebu North population
1	Small (1 to 3)	138	140
2	Mrate (4 to 7)	58	55
3	Large (7 and over)	4	5
		200	200

Source: Field Survey, 2012

Table 3.4: Stratified Data on Educational Level of Heads of Household in Abeokuta South and Ijebu North

Strata	Educational Level	Number in Abeokuta South population	Number in Ijebu North population
1	Primary	53	44
2	Secondary	74	85
3	Tertiary	73	71
		200	200

Source: Field Survey, 2012

The merged stratified data for the four socioeconomic characteristics of Abeokuta South and Ijebu North LGAs are shown in Table 3.5.

Table 3.5: Stratified Data on Occupation, Income, Dependant Size and Educational Level of Heads of Households in Abeokuta South and Ijebu North

Item No.	Name	Stratum No.	Name	Size of Stratum Abeokuta South and Ijebu-North
1	Occupation	1	Unemployed	12
		2	Paid employment	101
		3	Self employment	287
2	Income (in N'000)	1	0-10	70
		2	10-20	164
		3	20+	166
3	Dependant Size	1	Small (1-3)	278
		2	Moderate (4-7)	113
		3	Large (7+)	9
4	Educational Level	1	Primary	97
		2	Secondary	159
		3	Tertiary	144

Source: Field Survey, 2012

Using the data set for Abeokuta and Ijebu, the general multi-objective optimisation programme as in (2.8) is

$$\min \hat{V}ar(\bar{y}_{st}) = \min_{n} \begin{pmatrix} \hat{V}ar(\bar{y}_{st}^{1}) \\ \\ \hat{V}ar(\bar{y}_{st}^{2}) \end{pmatrix} \tag{2.22}$$

Subject to

$$\sum_{h=1}^{4} n_h = 200$$

$$2 \le n_h \le N_h \,, h = 1,2,3$$

$$n_h \in \mathrm{N}$$

Furthermore, we consider the following two programmes for the non linear minimizing of integers:

$$\min_{n} \hat{V}ar(\bar{y}_{st}^{1}) \tag{2.23}$$

Subject to

$$\sum_{h=1}^{4} n_h = 200$$

$$2 \le n_h \le N_h \, , h = 1,2,3$$

$$n_h \in \mathrm{N}$$

and

$$\min_{n} \hat{V}ar(\bar{y}_{st}^{2})$$

Subject to

$$\sum_{h=1}^{4} n_h = 200$$

$$2 \le n_h \le N_h \, , h = 1,2,3 \qquad (2.24)$$

$$n_h \in \mathrm{N}$$

The study adopted an approach based on the fact that its methodology is more realistic under the ambit of multivariate analysis. To extend the idea of this approach, the first step is to compute the matrix of variance-covariances of the vector $\bar{y}_{st} = (\bar{y}_{st}^{1}, ..., \bar{y}_{st}^{G})'$. The Eigenvalues of the covariance matrix of Abeokuta and Ijebu data set is as shown in Table 3.6.

Table 3.6: Eigenvalues of the Covariance Matrix of Abeokuta and Ijebu Data Set

Eigenvalues (λ_i)	Abeokuta	Ijebu
1	0.7593	0.7788
2	0.3970	0.3391
3	0.2297	0.2089
4	0.1539	0.1266

Source: Field Survey, 2012

Analysis based on principal component analysis

The principal component analysis ensured that the variance-covariance matrix was decomposed and the eigenvalues and eigenvectors calculated from the multivariate data representing information from the households. The principal component on the basis of the sample covariance matrix for the merged sample data sets for Abeokuta South and Ijebu North are:

$$Y_1 = -0.283X_1 + 0.428X_2 + 0.278X_3 + 0.812X_4$$

$$Y_2 = -0.069X_1 - 0.0169X_2 - 0.948X_3 + 0.309X_4$$

$$Y_3 = -0.667X_1 - 0.729X_2 - 0.010X_3 + 0.118X_4$$

$$Y_4 = -0.686X_1 + 0.534X_2 - 0{,}116X_3 - 0.481X_4$$

with corresponding sample variance 0.7788, 0.3391, 0.2089 and 0.1266 respectively. The total variance is 1.4534 and the principal components $\ddot{Y}_1, \ddot{Y}_2, \ddot{Y}_3, \ddot{Y}_4$ accounts for 53,6%, 23.3%, 14.4% and 8.7% of the total variance. Similarly, the principal components based on the merged sample correlation matrix are given by

$$\tilde{Y}_1 = 1.000X_1 - 0.151X_2 - 0.131X_3 - 0.425X_4$$

$$\tilde{Y}_2 = -0.151X_1 + 1.000X_2 + 0.211X_3 + 0.505X_4$$

$$\tilde{Y}_3 = -0.131X_1 + 0.211X_2 + 1.000X_3 + 0.158X_4$$

$$\tilde{Y}_4 = -0.425X_1 + 0.505X_2 + 0.158X_3 + 1.000X_4$$

The sample variance of the new principal components $\ddot{Y}_1, \ddot{Y}_2, \ddot{Y}_3, \ddot{Y}_4$ are 1.8381, 0.9244, 0.8323 and 0.4052 respectively while the total variance is 4. The principal components account for 44.6%, 23.1%, 20.8% and 10.1% of the total variance. Using the Eigen function, Eigen values of the merged sample covariance matrix were 0.76516, 0.36722, 0.21742 and 0.14319 with standard deviations 0.8747, 0.6060, 0.4663 and 0.3784 respectively.

CONCLUSION

In this study, optimal allocation in multi-item is developed as a multivariate optimization problem by finding the principal components. This was done by determining the overall linear combinations that concentrates the variability into few variables. It was demonstrated that the stratified samples are no longer independent. Post Stratified and Stratified Sampling approach have been presented as an appropriate research design and data collection instruments.

With each unit in the population having equal chances of being chosen, the individual observations from the sample were treated as being of equal weight. Since the standard deviations of the sample strata computed differed substantially, optimum stratification was used as a method of allocating larger size samples to those strata with larger standard deviations.

From the principal component analysis, it was seen that for both Abeokuta and Ijebu data sets, the variance based on the four characteristics as multivariate is less than that of the variables when considered as a univariate. From the results, it was seen that there was no difference in the percentage of the total variance accounted for by the dif-

ferent components from the merged sample when compared with the individual sample. The dispersion, $D(\hat{\beta}) = (X^1V^{-1}X)^{-1}\sigma^2$ which was replicated on all the characteristics considered, showed that optimum allocation was achieved when there was stratification as can be seen from the analysis.

REFERENCES

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

Bankier, M. D. 1996. Estimators based on several stratified samples with applications to multiple frame survey. *Journal of American Stat. Assoc*. **81**: 1074 – 1079.

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

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

Cheang, C. 2011. *Sampling strategies and their advantages and disadvantages*. http://www.2.hawaii.edu/~cheang/Sampling%20Strategies%20Advantages%20and%20Disadvantages.htm (Accessed March 5, 2011).

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

Diaz-Garcia, J.A and Ramos-Quiroga, R. 2011. Multivariate Stratified Sampling by Stochastic Multi Objective Optimization. Xiv: 1106.0773VI. *Statistical Methodology. XIV*, 1116-1123.

Diaz-Garcia, J. A., Cortez, L. U. 2008. Multivariate sampling techniques in science. *Communicacion Technica: Communicaciones Del CIMAT*. 8 (9), 76-83.

Diaz-Garcia, J. A., Cortez, L. U. 2006. Optimum allocation in multivariate stratified sampling: multi-objective programming. *Communicacion Technica: Communicaciones Del CIMAT*. 6(7), 28-33.

Diaz-Garcia, J.A., Cortez, L. U. 2008. Multi-objective optimisation for optimum allocation in multivariate stratified sampling. *Survey Methodology,* Vol. 34, No 2, 215-222.

Hartley, H. O. 1962. *Multiple frame surveys*: Proceeding of the social statistics section of American Statistical Association, 205 – 215.

Hartley, H. O. 1964. A new estimation theory for sampling surveys. *Biometrics*, 55, 545 – 557.

Jolliffe, I.T. 1986. Principal Component Analysis. *Springer Series in Statistics*. Springer Publishers.

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

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

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

Kolenikov, S. 2010. *Re: about multi-stage stratified sampling designing*. http://www.stata.com/statalist/archive/2010-04/msg01453.html

(Accessed February 27, 2011).

Lumley, T. 2004. *Analysis of complex survey samples*. Department of Biostatistics. University of Washington Press.

Miettinen, K. M. 1999. *Non linear multi-objective optimization*. Kluwer Academic Publishers, Boston.

Pirzada, S., Maqbool, S. 2003. Optimal Allocation in Multivariate sampling Through Chebyshev's Approximation. *Bulletin of the Malaysian Mathematical Science Society,* 2, (26), 221 – 230.

Saxens, J., Narain, M. U., Srivastava, S. 1986. The maximum likelihood method for non-response in Surveys. *Survey Methodology*. 12, 50 – 62.

Sethna, B. N., Groeneveld, L. 1984. *Research Methods in Marketing and Management*. Tata, Mcgraw-Hill, publishing, New-Delhi.

Steuer, R.E. 1986. *Multiple criteria optimization: Theory, Computation and applications*. John Wiley, New York.

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

Winship, C., Radbill, L. 1994. Sampling weights and regression analysis. *Sociological Methods and Research,* 23, (2), 230 – 257.

(Manuscript received: 16th January, 2014; accepted: 10th September, 2014).