American Journal of Theoretical and Applied Statistics
2013; 2(5): 142-148
Published online September 10, 2013 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20130205.14

Optimum allocation of multi-items in stratified random sampling using principal component analysis

Apantaku Fadeke Sola.[1], Olayiwola Olaniyi Mathew[1], Adewara Amos Adedayo[2]

[1]Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria
[2]Department of Statistics, University of Ilorin, Nigeria

Email address:
laniyimathew@yahoo.com(O. M. Olayiwola)

To cite this article:
Apantaku Fadeke Sola., Olayiwola Olaniyi Mathew, Adewara Amos Adedayo. Optimum Allocation of Multi-Items in Stratified Random Sampling Using Principal Component Analysis. *American Journal of Theoretical and Applied Statistics.* Vol. 2, No. 5, 2013, pp. 142-148.
doi: 10.11648/j.ajtas.20130205.14

Abstract: The problem of allocation with more than one characteristic in stratified sampling is conflicting in nature, as the best allocation for one characteristic will not in general be best for others. Some compromise must be reached to obtain an allocation that is efficient for all characteristics. In this study, the allocation of a sample to strata which minimizes cost of investigation, subject to a given condition about the sampling error was considered. The 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. These comprised of 200 households from each LGA. The characteristics were occupation, income, household size and educational level. Optimal allocation in multi-item was 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. 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 different components from the merged sample when compared with the individual sample. Optimum allocation was achieved when there was stratification

Keywords: Stratified Sampling, Optimum Allocation, Stratification, Optimization

1. Introduction

In social research, special emphasis is placed on the comparative and analytical use of samples. Knowledge, attitudes, and actions in everyday life 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. 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 this problem is rather arbitrary due to the fact that no best allocation is defined. In this study, the allocation of a sample to strata which minimizes cost of investigation, subject to a given condition about the sampling error was considered.

2. The Data

The 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. These comprised of 200 households from each LGA. The characteristics were occupation, income, household size and educational level.

3. Methodology

3.1. Introduction

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. According 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.

3.2. 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 ith stratum $(i=1,2,.....,L)$. For the jth character an unbiased estimate of the population mean, $\overline{Y}_j$, is $\overline{y}_{jst}$ which has the sampling variance.

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

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

and $X_i=\frac{1}{n_i}-\frac{1}{N_i}$, $a_{ij}=W_i^2S_{ij}^2$, in usual notation.

Let c_{ij} be the cost of enumerating the jth characteristic in the ith stratum and let k be the upper limit on total cost of the survey. Then

$$\sum_{i=1}^{L}\sum_{j=1}^{p}c_{ij}n_i\le k \qquad (2)$$

The multivariate allocation problem can be stated as

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

Subject to

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

$$\frac{1}{N_i}\le X_i\le 1, \quad i=1,2,.....,L \qquad (3)$$

where $\frac{1}{X_i}$ is used for n_i. If (3) is considered separately for each character, by ignoring the constant term in the objective function, the problem for kth character becomes

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

Subject to

$$\sum_{i=1}^{L}\sum_{j=1}^{p}c_{ij}X_i\le k \qquad (4)$$

$1\le X_i\le N_i, \qquad i=1,2,.....,L.$

By introducing a new variable x_{L+k}, the problem (4) transforms to

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

Subject to

$$g_k(X)=\sum_{i=1}^{L}\frac{a_{ik}}{X_i}-X_{L+k}\le 0 \qquad (b) \qquad (5)$$

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

$$1\le X_i\le N_i, \quad i=1,2,....L. \qquad (d)$$

The constraints in (5b) are convex (Kokan and Khan, 1967) and the constraint (5c) and the bounds (5d) are linear. The problem (5a)-(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 (5) is the following linear programming problem (LPP):

Minimize δ

Subject to

$$2\sum_{i=1}^{L}\frac{a_{ik}}{X_i^{k(0)}}-\sum_{i=1}^{L}\frac{a_{ik}X_i}{X_i^{k(0)^2}}-X_{L+k}\le 0$$

$$\sum_{i=1}^{L}\sum_{j=1}^{p}c_{ij}X_i\le k \qquad l=1,2,.....,t_k \qquad (6)$$

$$k=1,2,.....,p$$

$$X_{L+k}-\delta\le z_k^0$$

$$1\le X_i\le N_i \qquad i=1,2,.....,L$$

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

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

Subject to (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 (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 (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 characterristics

subject to (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 (3.79) 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.

3.3. Optimum Allocation via Multi-objective Optimization

The estimator of the population mean in multivariate stratified sampling for the jth characteristic is defined as

$$\bar{y}_{st}^j = \sum_{h=1}^{H} W_h \bar{y}_h^j \quad (9)$$

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 jth 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

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2 \quad (10)$$

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

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^j) = \sum_{h=1}^{H} \frac{W_h^2 s_{hj}^2}{n_h} - \sum_{h=1}^{H} \frac{W_h s_{hj}^2}{N} \quad (11)$$

3.4. Value Function

Under the value function technique, programme (8) is expressed as

$$\begin{aligned} &\min_{\mathrm{n}} v(\hat{V}ar(\bar{y}_{st})) \\ &\text{subject to} \\ &2 \leq n_h \leq N_h \qquad h = 1,2,....., H \\ &\qquad n_h \in N \end{aligned} \quad (12)$$

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, (12) can be expressed as

$$\begin{aligned} &\min_n \sum_{j=1}^{G} \lambda_j \hat{V}ar(\bar{y}_{st}^j) \\ &\text{subject to} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \\ &2 \leq n_h \leq N_h, \qquad h = 1,2,....., H \\ &\qquad n_h \in N \end{aligned} \quad (13)$$

Such that $\sum_{j=1}^{G} \lambda_j = 1, \quad \lambda_j \geq 0 \ \forall \ j = 1,2,....., G;$ where λ_j

weighs the importance of each characteristic. In the context of multi-objective optimization, (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 (13) help in reaching a better inference for future experiments, in which the appropriate weighting can be applied.

3.5. Optimal Design for a Multivariate Stratified Sampling Adopted in this Study

The idea of optimal allocation under a multivariate stratified sampling in this 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

$$\sum_{h=1}^{H} C_h n_h + C_0 = C \qquad (14)$$

$$2 \le n_h \le N_h$$

Where $\theta = Cov(\bar{y}_{st})$. This is the matrix of variance covariances of the vector

$\tilde{y}_{st} = (\tilde{y}_{st}, \ldots\ldots\ldots\ldots\ldots \tilde{y}_{st})$.

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

$$\operatorname{cov}(\tilde{y}_{st}) = \begin{Bmatrix} \operatorname{Var}(\tilde{y}^1_{st}) & \operatorname{Cov}(\tilde{y}^1_{st}\tilde{y}^2_{st}) & \operatorname{Cov}(\tilde{y}^1_{st}\tilde{y}^G_{st}) \\ \operatorname{Cov}(\tilde{y}^2_{st}\tilde{y}^1_{st}) & \operatorname{Var}(\tilde{y}^2_{st}) & \operatorname{Cov}(\tilde{y}^2_{st}\tilde{y}^G_{st}) \\ \operatorname{Cov}(\tilde{y}^G_{st}\tilde{y}^1_{st}) & \operatorname{Cov}(\tilde{y}^G_{st}\tilde{y}^2_{st}) & \operatorname{Var}(\tilde{y}^G_{st}) \end{Bmatrix} \qquad (15)$$

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 ($\tilde{y}^j_{st}$ and $\tilde{y}^i_{st}$)

$$\text{Cov}(\tilde{y}^i_{st}, \tilde{y}^j_{st}) = \sum_{h=1}^{H} \frac{W_h^2 S_{hij}}{n_h} - \sum_{h=1}^{H} \frac{W_h S_{hij}}{N} \qquad (16)$$

$$\text{and Cov}(\tilde{y}^i_{st}, \tilde{y}^j_{st}) = \sum_{h=1}^{H} \frac{W_h^2 S^2_{hi}}{n_h} - \sum_{h=1}^{H} \frac{W_h S^2_{hi}}{N} \qquad (17)$$

and C_h is the cost per G – dimensional sampling unit in stratum h and its vector

$C = (C_1, \ldots\ldots\ldots\ldots C_G)^1$.

3.6. 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, $Y_1, Y_2, \ldots\ldots, Y_p$ each one being a linear combination of the original set of variables, $X_1, X_2, \ldots\ldots, X_p$. 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 (Khan and Ahsan, 2003; Garcia and Cortez, 2006).

3.7. Finding Principal Components

Suppose that $X = (X_1, X_2, \ldots\ldots, X_p)'$ is a random vector with mean μ and covariance matrix Σ. Then the principal components of X, defined by $Y_1, Y_2, \ldots\ldots, Y_p$ satisfies the following conditions:

i. $Y_1, Y_2, \ldots\ldots, Y_p$ are mutually uncorrelated.

ii. $Var(Y_1) \ge Var(Y_2) \ge \ldots\ldots\ldots \ge Var(Y_p)$.

iii. $Y_j = a_{1j}X_1 + a_{2j}X_2 + \ldots\ldots + a_{pj}X_p = a'_j X$.

where $a_j = (a_{1j}, a_{2j}, \ldots\ldots, a_{pj})'$ is a vector of constants satisfying

$$\left\| a_j \right\|^2 = a'_j a_j \qquad (22)$$

$$= \sum_{k=1}^{p} a_{kj}^2 \qquad (23)$$

$$= 1 \qquad \text{for all } j = 1,2,\ldots\ldots, p .$$

In addition, the j^{th} principal component $Y_j = a'_j X$

is the linear compound of X that maximizes $Var(Y_j)$, subject to being uncorrelated with the preceding components $Y_1, Y_2, \ldots\ldots\ldots Y_{j-1}$.

Since $Y_j = a'X$ is a linear compound,

then

$$Var(Y_j) = Var(a'_j X) \qquad (24)$$

$$= a'_j \Sigma a_j \qquad j = 1,2,\ldots\ldots, p$$

To derive the first principal component of Y_1 , we have

$Var(Y_1) = a'_1 \Sigma a_1$

The idea is to select a_1 in such a way that $Var(Y_1)$ is as large as possible, subject to the constraint $a'_1 a_1 = 1$

This is a standard problem in constrained optimization and may be solved using the method of LaGrange multipliers.

To use this method the LaGrangian is formed as

$$L_1(a) = a'\Sigma a - \delta(a'a - 1) \qquad (25)$$

The required a_1 is the value of a that is a stationary point of (13).

Now define

$$\nabla a(\bullet) = \left(\frac{\partial}{\partial a_1}, \frac{\partial}{\partial a_2}, ..., \frac{\partial}{\partial a_p} \right)' \qquad (18)$$

It may be shown that

$\nabla a(a'\Sigma a) = 2\Sigma a$

$\nabla a(a'a) = 2a$

A stationary point of (3.122) must satisfy:

$\nabla a(L_1(a)) = 0$

Since

$$\begin{aligned} \nabla a(L_1(a)) &= \nabla a(a'\Sigma a) - \delta \nabla a(a'a - 1) \\ &= 2\Sigma a - 2\delta a \end{aligned} \qquad (19)$$

It follows that a_1 satisfies

$2\Sigma a_1 - 2\delta a_1 = 0$

That is,

$$(\Sigma - \delta \mathrm{I}) a_1 = 0 \qquad (20)$$

A non-trivial solution $(a_1 \neq 0)$ to the above exists if, and only if

$| \Sigma - \delta \mathrm{I} | = 0$

Where $|\bullet|$ is the determinant operator.

Thus δ must be an Eigen value of ∑, with a_1 being its corresponding Eigen vector:

Since ∑ is a $p \times p$ symmetric matrix, then there can be up to p distinct Eigen values. Since ∑ is positive (semi) definite, then all of its Eigen values are non-negative.

Assume, for the moment, that the Eigen values of Σ, $\lambda_1, \lambda_2, ..., \lambda_p$ are all distinct,

That is

$\lambda_1 > \lambda_2 > ..... > \lambda_p \geq 0$

$$Var(Y) = Var(a_1 X) \qquad (21)$$

$= a_1'\Sigma a_1$

$= a_1'(\delta \mathrm{I} a_1)$

Using (14), which is equal to $\delta a_1' a_1 = \delta$ will take it largest value at $\delta = \lambda_1$, since this is the value of the largest Eigen value, with a_1 being the Eigen vector corresponding to λ_1.

4. Results

The data from the survey were grouped for each of the four characteristics. Occupation was grouped into unemployed, paid employment and self employment. Income was grouped into 0 – < N10,000, N10,000 - < N20,000, N20,000 and above. Household size was grouped into small (1-3), moderate (4-7), large (7 and above) and educational level was grouped into primary, secondary and tertiary. S-plus was for the analysis.

The stratification technique in this study divided up the population into sub-population or strata. The strata for the four characteristics are in Table 1, 2, 3, and 4

Table 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

Table 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

Table 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

Table 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

The merged stratified data for the four socioeconomic characteristics of Abeokuta South and Ijebu North LGAs are shown in Table 5.

Table 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	Unemployed	12
1	Occupation	2	Paid employment	101
		3	Self employment	287
		1	0-10	70
2	Income (in ₦'000)	2	10-20	164
		3	20+	166
		1	Small (1-3)	278
3	Dependant Size	2	Moderate (4-7)	113
		3	Large (7+)	9
		1	Primary	97
4	Educational Level	2	Secondary	159
		3	Tertiary	144

Using the data set for Abeokuta and Ijebu, the general multi-objective optimisation programme as in (8) is

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

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:

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

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

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

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

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

To extend the idea of this approach, the matrix of variance-covariances of the vector $\bar{y}_{st} = (\bar{y}_{st}^{1}, ..., \bar{y}_{st}^{G})'$. was computed. The Eigen-values of the covariance matrix of Abeokuta and Ijebu data sets are as shown in Table 6.

Table 6: *Eigen-values 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

The summary estimates of the sample statistics for Abeokuta South and Ijebu North samples are as shown in Tables 7 and 8

Table 7: *Summary Estimates of Abeokuta South Sample Statistics*

	Income	Dependant Size
Mean $(\bar{y}_{st})$	2.067	1.3000
$\mathrm{V}_{\mathrm{srs}}(\bar{y})$	0.0214	0.0062
Var(post)	0.0073	0.0036
$\mathrm{V}_{\mathrm{mod}}(\bar{y}_{st})$	0.0045	0.0023

Table 8: *Summary Estimates of Ijebu North Sample Statistics*

	Income	Dependant Size
Mean $(\bar{y}_{st})$	2.033	1.333
$\mathrm{V}_{\mathrm{srs}}(\bar{y})$	0.0206	0.0016
Var(post)	0.0067	0.0014
$\mathrm{V}_{\mathrm{mod}}(\bar{y}_{st})$	0.0038	0.0011

The variance-covariance matrix for Abeokuta and Ijebu data sets are shown in tables 9 and 10 respectively

Table 9: *Variance-Covariance Matrix of Abeokuta Data Set*

	Occupation	Income	Dependant Size	Educational Level
Occupation	0.2361	-0.0272	-0.0391	-0.1333
Income	-0.0272	0.2924	0.0677	0.2052
Dependant Size	-0.0391	0.0677	0.4046	0.0447
Educational Level	-0.1333	0.2052	0.0447	0.6068

Table 10: *Variance-Covariance Matrix of Ijebu Data Set*

	Occupation	Income	Dependant Size	Educational Level
Occupation	0.2197	-0.0508	-0.0392	-0.1744
Income	-0.0508	0.3020	0.0761	0.2132
Dependant Size	-0.0392	0.0761	0.3832	0.1059
Educational Level	-0.1744	0.2132	0.1059	0.5484

The principal component analysis ensured that the variance-covariance matrix was decomposed and the eigen-values 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 $\vec{Y}_1, \vec{Y}_2, \vec{Y}_3, \vec{Y}_4$ 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

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

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

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

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

The sample variance of the new principal components $\vec{Y}_1, \vec{Y}_2, \vec{Y}_3, \vec{Y}_4$ 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.

5. 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.

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 different components from the merged sample when compared with the individual sample. Optimum allocation was achieved when there was stratification.

References

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

[2] 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).

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

[4] 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.

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

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

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

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

[9] 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.

[10] Lumley, T. 2004. Analysis of complex survey samples. Department of Biostatistics. University of Washington Press.

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

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

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

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

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

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