Using Principal Components in Multivariate Stratification Scheme

Apantaku Fadeke Sola
Department of Statistics
Federal University of Agriculture
PMB 2240, Abeokuta, Nigeria

Abstract

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. In multivariate stratified sample design, correlation is considered among interest variables. A variation of one variable with lower correlation is more important than others. Optimal allocation in multi-item is developed as a multivariate optimization problem by finding the principal components. A search was made 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,* $X_1, X_2, ..., X_p$ *. An empirical study from a household survey conducted in Abeokuta South and Ijebu North local government areas were used. The data about the households are available for four characteristics or variables that are related to the survey. These characteristics include occupation, income, number of dependants and the educational level. Each of the two local government areas with a sample size of 200 households each were randomly selected using simple random sampling technique making a total of 400 households. The heads of the households were interviewed. The study adopted an approach based on the fact that its methodology is more realistic under the ambit of multivariate analysis. Using Splus software, the variance- covariances matrices were computed. 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 were computed.*

Keywords: Principal components, multivariate stratification, optimal designs, sampling, allocation.

1.0 Introduction

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 equally material to practitioners engaged in marketing, commerce, industry, public health, biostatistics, education, public administration, economics, sociology, anthropology, psychology, political Science and even social workers.

Sampling methods are developed as means to an end originating in substantive research problems especially in the social sciences and their applications (Kish, 1965; Hunt and Tyrell, 2004). A working knowledge of practical sampling methods with an understanding of their theoretical background need to be a requirement for quantitatively oriented students in the social sciences as well as in allied fields. It is helpful although difficult to separate sampling design from the related activities involved in survey research. The sample design covers the tasks of selection and estimation for making inference from sample value to the population value. Beyond this are the problems of making inferences from one survey population to another and generally broader population, with measurements free from error. Different sampling designs would result in different standard errors, and choosing the design with the smallest error is the principal aim of sampling design.

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 (Garcia and Cortez, 2006). Probability samples 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.

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. These include: 1) to increase the precision of estimates for the whole population for one or more key data items being collected in the survey; 2) to obtain more precise estimates for interesting domains; 3) to allow the use of different sampling, non-response adjustment, editing, or estimation methods for domains with differing characteristics affecting the choice of method, and 4) to facilitate administration of the survey. Stratified sampling is always more restrictive than simple random sampling.

1.1 The Multivariate Stratification Scheme

Moreover, in the context of stratified sampling, some multivariate approaches have been proposed whereby the sample size and its allocation within strata take diverse characteristics into consideration (Sukhatme *et al*., 1984 and Arthanari and Dodge 1981). The multivariate stratified sample design is different with two steps from the univariate stratified sample design. The first step is to decide strata using stratified variables which are multivariate. The second step is to decide a sample size and the optimal allocation, that is, to decide a sample size of each stratum using interest variables which are multivariate.

There are many methods for allocation to strata in the multivariate stratified sample design. The first method is the proportional allocation, and the second is the multivariate allocation using one interest variable which is selected of multivariate interest variables. The third is a compromise allocation which is a weighted average of sample size of strata using individual allocation (Cochran, 1977; Chatterjee, 1972). The fourth is the optimal allocation for a loss function of characteristic values which combine variances of all variables (Kish, 1976; Sukhatme *et al*, 1984; Bethel, 1989; Khan and Ahsan, 2003; and Diaz-Garcia and Cortez, 2008).

In multivariate stratified sample design, correlation is considered among interest variables. The first method is a compromise allocation weighted by correlation coefficients or covariances, and the second method is the optimal allocation for a loss function of characteristic values of variance-covariance matrix. A variation of one variable with lower correlation is more important than others. The third method is to use weighted object function by the importance of interest variables in a mathematical programming and of importance is the error of estimation, the correlation coefficient or the covariances. The multivariate stratified sample design is used for multi-objective surveys in which there is difference among the importance of interest variables.

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. This problem was first considered by Neyman (1934), Dalenius (1957), Ghosh (1958), Kokan and Khan (1967), Khan and Ahsan (2003), Khan, Jahan and Ahsan (1997). Attempts were made for an acceptable allocation by either suggesting new criteria or exploring existing criteria further.

One of the problems of stratification is that loss in precision in the estimate of a characteristic increases if the characteristic in a stratum is not internally homogenous. To refrain from the increase in loss of precision is to assign a maximum weight to the j^{th} characteristic. Optimal allocation in multi-item is developed as a multivariate optimization problem by finding the principal components. This could be done by determining the overall linear combinations that concentrate the variability into few variables.

The objectives of the study are to:

1. Find allocation in multi-item stratified sampling using principal components,
2. Determine which of the components accounts for most of the variation in the model proposed, and
3. Compare the allocation in the two zones where the empirical data were collected.

2.0 Methodology

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.

2.1 Optimum Allocation via Multi-objective Optimization

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

$$\bar{y}_{st}^{j} = \sum_{h=1}^{H} W_h \bar{y}_h^{j} \qquad (2.1)$$

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 $\mathrm{Var}(\bar{y}_{st}^{j})$ 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

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2 \qquad (2.2)$$

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

$$\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{j}) = \sum_{h=1}^{H}\frac{W_h^2 s_{hj}^2}{n_h} - \sum_{h=1}^{H}\frac{W_h s_{hj}^2}{N} \qquad (2.3)$$

2.2 Optimal Designs in Multivariate Stratified Sampling

The problem of optimal allocation of a stratified sample when several variables are measured on each unit can be summarized as follows.

Suppose that

$$\theta_i = \theta_i(Y_{i1}, Y_{i2}, ..., Y_{iN_i}) \qquad i = 1,2,....., k$$

is a specified parametric function of the unknown values in the i[th] stratum, an objective is to estimate p linearly independent functions from the stratified sample

$$L_j(Y) = \sum_{i=1}^{k} a_{ij}\theta_i \qquad j = 1,2,.....p$$

To proceed with an allocation, Kokan and Khan (1967), suggested the minimization of the total sampling cost

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

such that constraints on each variance

$$\mathrm{var}(L_j | Y) \qquad j = 1,2,....., p$$

The conditions Kokan and Khan (1967) imposes on each variance are such that all $\hat{L}_j$ satisfies the proportional closeness requirement namely

$$\mathrm{P}\{|\hat{L}_j - L_j| < \lambda |L_j|\} \geq 1 - \alpha \qquad \text{for } j = 1,...p, \;\; 0 < \lambda < 1 \text{ and } 0 < \alpha < 1$$

Draper and Guttman (1968) remarked that it could be of interest to consider specific linear combination $L'\mu$ of elements of μ. This leads to the following:

Let the p – vector Y^{i} be normally distributed with mean μ^{i} and Covariance $\Sigma^{i}. i = 1,2,.....,k$ where μ^{i} is a $p-$vector and Σ^{i} is a $p \times p$ non singular matrix. Draper and Guttman, (1968) suggested that the first principal component is to be chosen so as to make its variance as large as possible.

2.2.1 Distance-based Method

This method is a vector of ideal goals, which is determined with the null information. On many occasions, the investigator comes up against the problem that no antecedents are available with which to address it. With this method, it is possible to obtain the optimum values, minimizing the distance between the optimum and the vector of targets simultaneously. By letting v_j to be the ideal point or goal for the objective $\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{j})$, $j = 1,2,.....,G$, the vector of targets V is given as

$$\mathrm{V} = \begin{pmatrix} \mathrm{v}_1 \\ : \\ : \\ \mathrm{v}_G \end{pmatrix}$$

A great advantage of this method is that this vector of targets V can be calculated without additional information. This is done by minimizing, separately, each objective $\hat{\mathrm{V}}\mathrm{ar}(\bar{\mathrm{y}}_{\mathrm{st}}^{\mathrm{j}})$, $j = 1,2,.....,G$, such that the vector V is defined as the vector of its individual minima, which is achieved on resolving the following G nonlinear minimization programmes for integers (Rao, 1985, 2003, 2005):

$$\begin{aligned} &\min_{\mathrm{n}} \hat{V}ar(\bar{y}_{st}^{j})) \\ &\text{subject to} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.4) \\ &2 \le n_h \le N_h \qquad h = 1,2,....., H \\ &n_h \in N \end{aligned}$$

For j=1,2,……,G. When the vector V has been established, one proceed to examine the problem to be optimized with the new objective function, namely

$$\begin{aligned} &\min_{n} d\ (\hat{\mathrm{V}}\mathrm{ar}(\bar{y}_{st}^{j}), \mathrm{V}) \\ &\text{subject to} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.5) \\ &2 \le n_h \le N_h, h = 1,2,....., H \\ &n_h \in N. \end{aligned}$$

Where d corresponds to a weighted norm. More generally, when the weighted norm L_q, is considered, the problem to be optimized takes the following form

$$\begin{aligned} &\min_{n}\left[\sum_{j=1}^{G} \lambda_j \mid \mathrm{Var}(\bar{y}_{st}^{j}) - v_j \mid^{q}\right]^{\frac{1}{q}} \\ &\text{subject to} \\ &\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.6) \\ &2 \le n_h \le N_h, \quad h = 1,2,....., H \\ &n_h \in N, \end{aligned}$$

with $1 \le q \le \infty$ and $\lambda_j \ge 0,$ which is the weight or priority given to each objective j. By taking $\lambda_j = 1$ with $q = 1,$ we have the following problem

$$\min_n \left[\sum_{j=1}^{G} | \operatorname{Var}(\bar{y}_{st}^{j}) - v_j | \right]$$

subject to

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.7)$$

$$2 \le n_h \le N_h, \quad h = 1,2,....., H$$

$$n_h \in N,$$

As v_j are constant for all j=1,2,……,G, the problem is reduced to

$$\min_n \sum_{j=1}^{G} \operatorname{Var}(\bar{y}_{st}^{j})$$

subject to

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.8)$$

$$2 \le n_h \le N_h, \quad h = 1,2,....., H$$

$$n_h \in N,$$

With $q = \infty$, one need only to take into account the maximum deviation, and so the problem to be optimized is

$$\min_n \max_{j=1,2,...,G} \left[\operatorname{Var}(\bar{y}_{st}^{j}) - v_j \right]$$

subject to

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.9)$$

$$2 \le n_h \le N_h, \quad h = 1,2,....., H$$

$$n_h \in N,$$

And for q=2 the problem is

$$\min_n \left[\sum_{j=1}^{G} | \operatorname{Var}(\bar{y}_{st}^{j}) - v_j |^2 \right]^{\frac{1}{2}}$$

subject to

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.10)$$

$$2 \le n_h \le N_h, \quad h = 1,2,....., H$$

$$n_h \in N,$$

Alternatively, another distance has been proposed by Khuri and Cornell (1987):

$$\min_n \sum_{j=1}^{G} \left[\frac{(\operatorname{Var}(\bar{y}_{st}^{j}) - v_j)^2}{v_j^2} \right]$$

subject to

$$\sum_{h=1}^{H} c_h n_h + c_0 = C \qquad (2.11)$$

$$2 \le n_h \le N_h, \quad h = 1,2,....., H$$

$$n_h \in N,$$

In all these optimization methods, the cost restriction $\sum_{h=1}^{H} c_h n_h + c_0 = C$ has been utilized. However, on some occasions the restrictions do not apply to the costs but to the availability of man-hours for carrying out a survey, or simply to the total time available for performing the survey.

2.2.2 Linear Compounds

A linear compound $Y_1, Y_2, ..., Y_p$ or $\underline{Y}$ of a $p \times 1$ random vector X is a linear combination of its compounds such that

$$Y = a'X$$

where $a = (a_1, a_2, ...a_p)'$ is a vector of real constants. Suppose that X is a $p \times 1$ random vector, a is a $p \times 1$ vector of constants and $Y = a'X$, then the mean of the linear compound is

$$\begin{aligned} E(Y) &= E(a'X) \\ &= E(\sum_{i=1}^{p} a_i X_i) \\ &= \sum_{i=-1}^{p} a_i E(X_i) \\ &= \sum_{i=1}^{p} a_i \mu_i \\ &= a'\mu \end{aligned}$$

Thus,

$$E(\underline{Y}) = a'\mu$$

The variance is defined as

$$Var(\underline{Y}) = E[(a'X - a'\mu)^2].$$

where $a'(X - \mu)$ is a scalar and so is also equal to $(X - \mu)'a.$

Thus,

$$\begin{aligned} \text{var}(\underline{Y}) &= E[a'(X - \mu)^2] \\ &= E[a'(X - \mu)(X - \mu)'a] \\ &= a'[E(X - \mu)(X - \mu)']a \\ &= a'\Sigma a \end{aligned}$$

For a $p \times 1$ random vector X, and a $p \times n$ matrix of real constants A

Then

$$E[A'X] = A'\mu$$

and

$$Var(A'X) = A'\Sigma A$$

2.3 Principal Component Analysis

We 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.

2.3.1 Finding Principal Components

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

(i) $Y_1, Y_2, ..., Y_p$ are mutually uncorrelated.

(ii) $Var(Y_1) \geq Var(Y_2) \geq ..... \geq Var(Y_p)$.

(iii) $Y_j = a_{1j}X_1 + a_{2j}X_2 + ... + a_{pj}X_p = a_j'X$.

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

$$\|a_j\|^2 = a_j'a_j$$

$$= \sum_{k=1}^{p} a_{kj}^2$$

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

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 $Y_1, Y_2, ... Y_{j-1}$.

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

then

$$Var(Y_j) = Var(a_j'X)$$

$$= a_j'\Sigma a_j \qquad j = 1,2,....., p$$

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

$$L_1(a) = a'\Sigma a - \delta(a'a - 1) \qquad (2.12)$$

The required a_1 is the value of a that is a stationary point of (2.12).

Now define

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

It may be shown that

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

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

A stationary point of (2.12) must satisfy:

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

Since

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

$$= 2\Sigma a - 2\delta a$$

It follows that a_1 satisfies

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

That is,

$$(\Sigma - \delta \mathrm{I})a_1 = 0 \qquad (2.13)$$

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

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

$$Var(Y) = Var(a_1 X)$$

$$= a_1'\Sigma a_1$$
$$= a_1'(\delta \mathrm{I} a_1) \qquad (2.14)$$

Using (2.13), which is equal to $\delta a_1' a_1 = \delta$ will take its 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.

3.0 The Empirical Study

An empirical data from a household survey conducted in Abeokuta South and Ijebu North local government areas were used. The data about the households are available for four characteristics or variables that are related to the survey. These characteristics include occupation, income, number of dependants and the educational level. Each of the two local government areas with a sample size of 200 households each were randomly selected using simple random sampling technique making a total of 400 households. The heads of the households were interviewed.

Results

3.1 Parameter Estimation

The primary concern in all sample surveys is the derivation of point estimates for the parameters of main interest. The point estimate for the proportionately allocated sample data are obtained by adding up the observation over all the strata and dividing by the sample size. However, equally important is the derivation of the variances of the above estimates. The sampling variance is indeed one of the key indicators of quality in sample surveys and estimation. Variance estimation is crucial issue in the assessment of the survey results. Statistical software Splus was used in the analysis of data.

The summary estimates of the sample statistics for Abeokuta South and Ijebu North samples are as shown in Tables 3.1 and 3.2.

Table 3.1: Summary Estimates of Abeokuta South Sample Statistics

	Occupation	**Income**	**Dependant Size**	**Educational Level**
Mean $(\bar{y}_{st})$	1.833	2.067	1.3000	1.700
$V_{srs}(\bar{y})$	0.008	0.0214	0.0062	0.018
Var(post)	0.0036	0.0073	0.0036	0.0069
$V_{mod}(\bar{y}_{st})$	0.0027	0.0045	0.0023	0.0058

Table 3.2: Summary Estimates of Ijebu North Sample Statistics

	Occupation	**Income**	**Dependant Size**	**Educational Level**
Mean $(\bar{y}_{st})$	1.833	2.033	1.333	2.033
$V_{srs}(\bar{y})$	0.0079	0.0206	0.0016	0.0019
Var(post)	0.0047	0.0067	0.0014	0.0015
$V_{mod}(\bar{y}_{st})$	0.0013	0.0038	0.0011	0.0012

3.2 Multivariate Stratified Sampling

The problem of optimum allocation in multivariate stratified sampling has been examined in statistical literature, but the solutions proposed have been particular cases of a multi-objective optimisation technique. Using our data set for Abeokuta and Ijebu, the general multi-objective optimisation programme as in (3.79) 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

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

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

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

The study adopted an approach based on the fact that its methodology is more realistic under the ambit of multivariate analysis. The first step is to compute the matrix of variance-covariances of the vector $\bar{y}_{st} = (\bar{y}_{st}^1, ..., \bar{y}_{st}^G)'$. Using Splus software, the variance- covariances matrix is as in Tables 3.3 and 3.4. The Eigenvalues of the covariance matrix of Abeokuta and Ijebu data set is as shown in Table 3.5 while the Eigen vectors are as shown in Tables 3.6 and 3.7.

Table 3.3: 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 3.4: 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

Table 3.5: 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

Table 3.6: Eigen Vectors of Abeokuta Data Set

	1	2	3	4
1	-0.2532	0.0117	-0.7562	-0.6033
2	0.4176	-0.0713	-0.6469	0.6341
3	0.2143	-0.9569	0.0617	-0.1858
4	0.8459	0.2811	0.0774	-0.4466

Table 3.7: Eigen Vectors of Ijebu Data Set

	1	2	3	4
1	-0.3064	-0.1546	0.5704	0.7463
2	0.4344	0.0398	0.7952	-0.4212
3	0.3236	-0.9387	-0.1156	0.0266
4	0.7828	0.3054	-0.1703	0.5148

3.3 Results 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 components were computed from the study on the basis of the sample covariance matrix, and the result for Abeokuta samples are

$$Y_1 = -0.253X_1 + 0.418X_2 + 0.214X_3 + 0.846X_4$$
$$Y_2 = 0.117X_1 - 0.0713X_2 - 0.95X_3 + 0.281X_4$$
$$Y_3 = -0.756X_1 - 0.647X_2 + 0.062X_3 + 0.073X_4$$
$$Y_4 = -0.603X_1 + 0.634X_2 - 0.186X_3 - 0.447X_4$$

with corresponding sample variances 0.7593, 0.3970, 0.2297 and 0.1539 respectively.

Thus, the total variance is 1.5399 and the principal components, $\tilde{\tilde{Y}}_1, \tilde{\tilde{Y}}_2, \tilde{\tilde{Y}}_3, \tilde{\tilde{Y}}_4$ successively accounts for 49.3%, 25.8%, 14.9% and 10.0% of the total variance. Similarly, the principal components, based on the sample correlation matrix for Abeokuta are given by

$$\tilde{\tilde{Y}}_1 = -0.438X_1 + 0.566X_2 + 0.307X_3 + 0.628X_4$$
$$\tilde{\tilde{Y}}_2 = 0.333X_1 + 0.109X_2 + 0.887X_3 + 0.300X_4$$
$$\tilde{\tilde{Y}}_3 = -0.752X_1 - 0.565X_2 + 0.297X_3 - 0.161X_4$$
$$\tilde{\tilde{Y}}_4 = -0.363X_1 + 0.591X_2 - 0.173X_3 - 0.700X_4$$

Where $\overline{X}_i = \frac{X_i}{S_i}$ for i = 1,2,3,4

The sample variances of the new principal components $\tilde{\tilde{Y}}_1, \tilde{\tilde{Y}}_2, \tilde{\tilde{Y}}_3, \tilde{\tilde{Y}}_4$ are 1.7280, 0.9463, 0.8971 and 0.4286 respectively. In this case, the total variance is 4 and the principal components account successively for 43.2%, 23.7%, 22.4% and 10.7% of the total variance.

By using the Eigen function, we found that the Eigen values of the sample covariance matrix were 0.7593, 0.3970, 0.2297 and 0.1539. The square root of these values is the standard deviations of the principal components and is 0.8714, 0.6301, 0.4792 and 0.3924 respectively.

The principal components on the basis of the sample covariance matrix for Ijebu samples are

$$Y_1 = -0.306X_1 + 0.434X_2 + 0.324X_3 + 0.783X_4$$
$$Y_2 = -0.155X_1 + 0.040X_2 - 0.939X_3 + 0.305X_4$$
$$Y_3 = 0.570X_1 + 0.795X_2 - 0.116X_3 - 0.170X_4$$
$$Y_4 = 0.746X_1 - 0.421X_2 + 0.027X_3 + 0.515X_4$$

with corresponding sample variances 0.7788, 0.3391, 0.2089 and 0.1266 respectively.

Thus, the total variance is 1.4534 and the principal components, successively accounts for 53.4%, 23.3%, 14.4% and 8.7% of the total variance. Similarly, the principal components, based on the sample correlation matrix for Abeokuta are given by

$$\tilde{\tilde{Y}}_1 = -0.481X_1 + 0.519X_2 + 0.340X_3 + 0.620X_4$$
$$\tilde{\tilde{Y}}_2 = 0.509X_1 + 0.158X_2 + 0.825X_3 - 0.189X_4$$
$$\tilde{\tilde{Y}}_3 = -0.557X_1 - 0.689X_2 + 0.452X_3 - 0.104X_4$$
$$\tilde{\tilde{Y}}_4 = -0.446X_1 + 0.481X_2 + 0.010X_3 - 0.755X_4$$

Where $\overline{X}_i = \frac{X_i}{S_i}$ for i = 1,2,3,4

The sample variances of the new principal components $\tilde{\tilde{Y}}_1, \tilde{\tilde{Y}}_2, \tilde{\tilde{Y}}_3, \tilde{\tilde{Y}}_4$ are 1.9552, 0.9065, 0.7726 and 0.3658 respectively. In this case, the total variance is 4 and the principal components account successively for 48.9%, 22.7%, 19.3% and 9.1% of the total variance.

By using the Eigen function, we found that the Eigen values of the sample covariance matrix were 0.7788, 0.3391, 0.2089 and 0.1266. The square root of these values is the standard deviations of the principal components and is 0.8825, 0.5823, 0.4571 and 0.3558 respectively.

References

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

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.

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

Dalenius, T. 1957. Sampling in Sweden: Contributions to the methods and theories of sample survey practice. Almavist and Wicksell, Stockholm.

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

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.

Draper, N.R and Guttman, I. 1968. Some Bayesian Stratified Two Phase Sampling Results. *Biometrika* 55, 58-78

Ghosh, S.P., 1958. A note on Stratified Random Sampling with Multiple Characters. Col.Stat. Bull, 8, 81-89.

Hunt, N. and Tyrell, S. 2004. Stratified sampling. Coventry University Press. http://www.coventry.ac.uk/ec/~nhunt/meths/strati.html (accessed February 28, 2011).

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

Khan, M.G.M, Jahan, N. and Ahsan, M.J. 1997. Determining the optimum cluster size. *Journal of the Indian Society of Agricultural Statistics*. Vol. 50 (2), 121-129.

Khuri, A.I. and Cornell, J. 1987. Response Surfaces; Design and Analysis. Marcel Dekker, New York, NY.

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

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

Neymaan, J. 1934. On the two different aspects of the representative methods. The method stratified sampling and the method of purposive selection. *Journal of Royal Statistical Society*, 97, 558-606.

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.

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

Rao, J. N. K. 2005. Inference from stratified samples, *Journal of American Statistical Association*, 109, 650 – 660.

Rao, J. N. K. 2003. Small area estimation. New York, Wiley.

Rao, J. N. K. 1985. Conditional inference in survey sampling. *Survey Methodology*. II, 15 – 31.