THE UNIVERSITY OF THE SOUTH PACIFIC
LIBRARY

Author Statement of Accessibility

Name of Candidate : Sushita Sharma

Degree : MSc

Department/School : FSTE / SCIMS

Institution/University : USP

Thesis Title : Use of Mathematical Programming In Sampling

Date of completion of requirements for award : _____

1. This thesis may be consulted in the Library without the author's permission. (Yes/No)
2. This thesis may be cited without the author's permission providing it is suitably acknowledged. (Yes/No)
3. This thesis may be photocopied in whole without the author's written permission. (Yes/No)
4. This thesis may be photocopied in proportion without the author's written permission. (Yes/No)
   Part that may be copied:

   Under 10% _____ 40-60% ✓

   10-20% _____ 60-80% _____

   20-40% _____ Over 80% _____
5. I authorise the University to produce a microfilm or microfiche copy for retention and use in the Library according to rules 1-4 above (for security and preservation purposes mainly). (Yes/No)
6. I authorise the Library to retain a copy of this thesis in e-format for archival and preservation purposes. (Yes/No)
7. After a period of 5 years from the date of publication, the USP Library may issue the thesis in whole or in part, in photostat or microfilm or e-format or other copying medium, without first seeking the author's written permission. (Yes/No)
8. I authorise the University to make this thesis available on the Internet for access by authorised users. (Yes/No)

Signed: Sharma

Date: 4/7/17

Contact Address

c/o Pacific TAFE
Statham. Campus, USP

Permanent Address

8 Vinod Karsanji St
Rifle Range

USE OF MATHEMATICAL PROGRAMMING IN SAMPLING

by

Sushita Sharma

A thesis submitted in partial fulfilment of the

requirements for the degree of

Masters of Science in Mathematics

School of Computing, Information and Mathematical Sciences

Faculty of Science, Technology and Environment

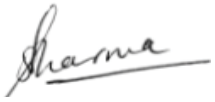
The University of the South Pacific

June, 2017

Declaration of Originality

Statement by Author:

I hereby declare that the work presented in this thesis, which I now submit for assessment of the award of Master of Science is entirely my own work and has not been taken from the work of others; except in citations and references which have been acknowledged and results in this thesis has not been previously submitted for a University degree either in whole or in part elsewhere.



.....
Sushita Sharma
May 2017

Statement by Supervisor:

The research work in this thesis was accomplished under my supervision and to the best of my knowledge is the sole work of Ms Sushita Sharma.

.....
Dr. M.G.M. Khan (Supervisor)
Associate Professor of Statistics and Deputy Head of School
School of Computing, Information and Mathematical Science
Faculty of Science, Technology and Environment
The University of the South Pacific
Suva, Fiji

Acknowledgements

As I complete this thesis, I would like to acknowledge and dedicate this work to some very special people who have enabled me to undertake this enormous task and to see it through its completion.

Firstly, I must appreciate Dr Bibhya Sharma, my husband and an academic, for his continuous drive for me to finish my research write up. Without his encouragement and understanding, this task would never have been possible.

Secondly, my heartfelt thanks go my supervisor, Dr M.G.M. Khan. Without his expert knowledge in this field; his guidance and professional advice, this thesis would not have been successful. I deeply express my sincere appreciation for his valuable suggestions, motivation, time and help in every phase of this mammoth task.

Thirdly, it goes without saying that I want to mention Sid for all his unselfish ways and beliefs that have made me a stronger person, and has empowered me complete this task with a positive mind.

Also, I acknowledge my special friends who have always encouraged me to keep on and keeping on; in those not so easy days of researching and putting everything together.

Finally, I thank my parents and siblings for those childhood days of dreaming, that keeps me going and to keep trying to do things better.

Table of Contents

Abstract

Stratified random sampling is one of the most commonly used techniques in sample surveys as it increases the precision of the estimate of the population parameters. In this technique, two basic problems, that is, the problem of constructing optimum stratum boundaries (OSB) and the problem of determining sample allocation to different strata are well known in the sampling literature. To increase the precision in the estimates of population parameters these problems are to be addressed by the sampler while using stratified sampling. There are several methods available to determine the OSB when the frequency distribution of the study (or a correlated) variable is available. However, many of these attempts have been made with an unrealistic assumption that the frequency distribution and the stratum variances of the target population are known prior to conducting the survey. On the other hand, the problem of determining optimum allocation to strata was addressed in the literature mostly as a separate problem assuming that the strata are already formed and the stratum variances are known. As both the problems are not addressed simultaneously, the OSB and the sample allocations so obtained may not be feasible or may be far from optimum. In this thesis, these two problems are discussed simultaneously when the population mean of the study variable y is of interest and a frequency distribution $f(y)$ or the frequency distribution $f(x)$ of its auxiliary variable x is available. The problem is formulated as a Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter of the target population, which is subjected to a fixed total sample size. The formulated MPP is then solved by executing a program coded in a user's friendly optimization software, LINGO. Two numerical examples, when the study variable or the auxiliary variable follows a uniform and a right-triangular distribution in the population, are presented to demonstrate the practical application of the proposed method or its computational details. The proposed technique can easily be applied to other frequency distributions.

In sample surveys, cluster sampling is another widely used sampling technique, which is employed when the target population is spread across region and natural groupings are evident in the population. In this thesis, a problem of determining optimum allocation of sample size in multivariate surveys is also studied. When a cluster sampling design is to be used and more than one characteristic are under study, usually

it is not possible to use the individual optimum cluster size and sampling unit for one reason or the other. In such situations some criterion is needed to work out an acceptable cluster size and sampling unit which are optimum for all characteristics in some sense. Moreover, for practical implementation of sample size, we need integer values of the cluster size and sampling unit. The present thesis addresses the problem of determining integer optimum compromise cluster size and sampling unit when the population means of various characteristic are of interest. The problem is formulated as an All Integer Nonlinear Programming Problem (AINLPP) and a solution procedure is proposed using evolutionary algorithm implemented in Excel Solver. The result shows that evolutionary algorithm can be efficiently applied in determining the sample size in multivariate cluster sampling design. A numerical example is presented to illustrate the practical application of the solution procedure.

Preface

This thesis entitled “Use of Mathematical Programming in Sampling” is submitted to The University of the South Pacific, Suva, Fiji to supplicate the Master of Science in Mathematics. All of the work presented henceforth is carried out by me in the School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Suva, Fiji.

Samplings are widely used as means of gathering statistical data or information on an extensive range of subjects for both research and administrative purposes. Numerous surveys are conducted to develop hypotheses in different disciplines such as demography, political science, health, economics and education. Governments make considerable use of surveys to inform them of the conditions of their populations in terms of unemployment, income, expenditure, education, health and others.

Sampling is used to reduce the cost and the time spent on the survey. Thus, the decision on the right sample size is very important because a large sample would increase the cost and time, and also the result may not be precise. Similarly, too small a sample may diminish the purpose of the survey. To overcome this problem, sampling design can be formulated and solved as a Mathematical Programming Problem (MPP) to minimize the cost of the survey and to prevent the loss of precision in the estimates. This thesis studies some problems that are occurred in sample surveys and can be dealt by formulating as MPP in two commonly used sampling techniques: stratified random sampling and cluster sampling.

In sample surveys stratified random sampling is one of the most widely used sampling techniques, which is employed to increase the precision of the information or estimate. The use of this technique needs the solution of some basic problems: (1) the determination of optimum number of strata, (2) the determination of optimum strata boundaries (OSB) and (3) the determination of optimum sample size from each strata. This thesis is an attempt to provide solutions to the problems (2) and (3) using a mathematical programing approach, which is discussed in Chapter 2.

The cluster sampling is another sampling technique in surveys, which frequently used by the researchers. When the information is to gather from more than one characteristic using cluster sampling, a foremost problem to the surveyor is to determine the optimum size of sampling units and the size of cluster that increases the precision of estimates of all characteristics. In this thesis, an attempt is also made to develop a multivariate cluster sampling design to solve this problem using a mathematical programming technique, which is discussed in Chapter 3.

This thesis consists of four chapters. **Chapter 1** begins with an introduction to survey sampling and describes briefly about the stratified random sampling, the allocation of sample size, the stratification, the use of auxiliary variable, the cluster sampling, the optimum cluster size, mathematical programming technique, mathematical programming in sampling and objective of the study. This is followed by a brief review of the literature and studies about optimum strata boundaries as well as optimum cluster size multivariate cluster sampling.

Chapter 2 discusses the problems of determining optimum strata boundaries (OSB) and optimum sample size allocation in stratified random sampling. In this thesis, the author addresses a solution of both the problems simultaneously. In sampling literature, the first problem has been addressed without considering the later assuming that the optimum sample allocation is made by an optimum or Neyman allocation and the later problem was addressed as a separate problem assuming that the OSB are already formed. Such solutions, when both the problems are not addressed simultaneously, may not be feasible or may be far from optimum. In this chapter, both the problems considered simultaneously and formulated as a Mathematical Programming Problem (MPP) in order to minimize the variance of the estimated population of a given target population whilst subjected to a fixed total sample size. The formulated problem is then solved through the execution of a program coded in a user's friendly optimization software, LINGO. Numerical examples are presented to demonstrate the computational details of the solution procedure using two populations when the study variable follows respectively a uniform and a right-triangular distribution.

In **Chapter 3**, a problem of determining optimum cluster size and sampling unit in multivariate cluster sampling is discussed. When more than one characteristic are

under study, individual optimum cluster size and sampling unit for one characteristic is not necessarily optimum for all characteristics. Moreover, for practical utility of the results the integer values of the cluster size and sampling unit are required. In this chapter, the problem of determining integer optimum compromise cluster size and sampling unit is formulated as an All Integer Nonlinear Programming Problem (AINLPP) and a solution procedure is proposed using an evolutionary algorithm implemented in Excel Solver. A numerical example illustrates the computational details.

Chapter 4 forms the final chapter of this thesis. It provides a brief conclusion to our research studies and offers suggestions for further future work. A comprehensive list of references is presented in Bibliography at the end of the thesis.

Chapter 1

Introduction

1.1 Survey Sampling

In this age and era of technology, where every claim must be validated by evidence, we see more and more use of sample surveys. People do not merely take a person's or institution's word for granted unless it is backed up by some authentic information; hence data gathering and information search is on the uprise. What we are seeing now is sample surveys, which are widely accepted means of gathering information, are carried out in all works of life from governmental departments to market research; statistical data are being collected through sample surveys for various reasons ranging from research to administrative purpose.

A sample survey can be defined as a scientific and methodical way of selecting a range of elements from a target population in order to conduct a survey. The results of a survey need to be unbiased and reliable, hence mathematical thought and planning is needed. Also, the sample size must be considered as a very important factor to ensure the validity of results obtained from survey. Some of the advantages of sample survey are as follows:

i) Easy and cost efficient.

ii) Reduces time needed to collect and process the data and produce results as it requires a smaller scale of operation.

iii) Convenient data gathering.

iv) Enables characteristics to be tested which could not otherwise be assessed.

v) Importantly, a goal in the design of sample surveys is to obtain a sample that is representative of the population so that precise inference is made.

In surveys, since the important decisions made or the efficient estimates of the characteristics obtained are based on the sample results, it becomes imperative that the best sampling methods are used to collect data.

Sampling methods are classified as either probability or nonprobability. In probability sampling, each member of the population has a known non-zero probability of being selected. Probability methods include random sampling, systematic sampling, stratified sampling, etc. In nonprobability sampling, members are selected from the population in some nonrandom manner. These include convenience sampling, judgment sampling, quota sampling, and snowball sampling. The advantage of probability sampling is that sampling error, that is, the degree to which a sample might differ from the population, can be calculated. When inferring to the population, results are reported plus or minus the sampling error. Whereas, in nonprobability sampling, the sampling error remains unknown.

1.2 Stratified Random Sampling

A variety of sampling techniques from simple to complex, such as simple random sampling, stratified random sampling, cluster sampling, multistage sampling and double sampling, have been developed to provide efficient estimates of the characteristics under study. Out of these, the most common and popular technique used in sample surveys is the **stratified random sampling technique**.

It has been seen that in simple random sampling the precision of an estimator depends not only on the sample size but also on the variability or heterogeneity among the units of the population. However, when the units vary considerably (heterogeneous), one possible way to estimate the population mean (or population total) with greater precision is to divide the population into several non-overlapping mutually exclusive groups, each of which is more relatively homogenous than the entire population before sampling is carried out. These groups are called strata and the sampling procedure is called stratified random sampling. Then a simple random sample of pre-determined size is independently drawn from each stratum.

The advantages of stratified random sampling over other sampling designs are:

i) Increases precisions and accuracy: That the total, mean, and other parameters of the entire population can be estimated with high precision and accuracy. If each stratum is internally homogenous, the measurement vary little from one unit to another, then a precise estimate of any stratum total or mean can be obtained from a small sample in that stratum.

ii) Time and cost effective: If the stratification variable were equal to the survey variable, each element of the stratum would be a perfect representative of that characteristic. It would then be sufficient to take a handful element out of each stratum to get the actual distribution of the characteristic in the parent population. Therefore, there are frequently savings in time, cost and resources needed for sampling the units.

iii) Provides estimate of strata: The estimates for each stratum can be obtained separately. This ensures that all important subgroups are represented in the sample.

iv) Administrative convenience: The administrative convenience may dictate the use of stratified sampling as it is easier to sample separately from the strata rather than from the entire population (especially, if it is too large). That is, from the standpoint of the agency conducting the survey, each subpopulation can be supervised separately. This can also allow separate analysis of each stratum, therefore, the differences among the strata can be evaluated.

v) Increases sampling efficiency: Stratification increases sampling efficiency if a sub-division of the population is made so that the variability between units within a stratum is reduced as compared to the variability within the entire population.

vi) Increases validity of inference: Since the units selected for inclusion within the sample are chosen using probabilistic methods, stratified random sampling allows us to make statistical conclusions from the data collected that will be considered to be valid.

1.2.1 Study and Auxiliary Variable

A variable is an attribute or characteristic under study that assumes different values for different elements. The variable which is to be investigated is called a survey or study variable, whereas another variable which has some kind of relationship with the study variable is called an auxiliary variable. For example, if the income of a person is a study variable, the tax paid by the person can be considered as an auxiliary variable. Some advantages of the use of auxiliary variables are:

i) The cost of measurements and effort is less when an auxiliary variable is used. Because the auxiliary information may be availability from the past survey, books or journals, whereas the information for study variable obtained from current surveys or experiments.
ii) The auxiliary variable has less error in measurement as compared to the study variable.

1.2.2 Notations

Consider that a population of size N is divided into L non-overlapping strata each of size $N_h;\ h=1,2,...,L$ from which a probability sample of size n_h is drawn using a simple random sample without replacement such that $N=\sum_{h=1}^{L}N_h$ and $n=\sum_{h=1}^{L}n_h$, where n is total sample size. Let y_{hi} and x_{hi} denote the values of ith unit of the study variable y and the auxiliary variable x in the hth stratum, respectively.

In estimating a characteristic in a stratified random sampling (SRS), let us define the notations used as in Cochran (1977):

N : Population Size

N_h : Stratum size of the hth stratum; $h=1,2,...,L.$

$W_h=\frac{N_h}{N}$: Stratum weight, where $\sum_{h=1}^{L}W_h=1$.

y_{hi} : Value of the ith unit of the study variable y in the hth stratum.

x_{hi} : Value of the ith unit of the auxiliary variable x in the hth stratum.

$Y_h=\sum_{i=1}^{N_h}y_{hi}$: Population total of y in the hth stratum.

$X_h = \sum_{i=1}^{N_h} x_{hi}$: Population total of x in the hth stratum.

$\bar{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}$: Population mean of y in the hth stratum.

$\bar{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{hi}$: Population mean of x in the hth stratum.

$S_{hy}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(y_{hi} - \bar{Y}_h \right)^2$: Population variance of y in the hth stratum.

$S_{hx}^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(x_{hi} - \bar{X}_h \right)^2$: Population variance of x in the hth stratum.

$Y = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}$ = Population total of y.

$X = \sum_{h=1}^{L} \sum_{i=1}^{N_h} x_{hi}$ = Population total of x.

$\bar{Y} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{N}$ = Population mean of y.

$\bar{X} = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} x_{hi}}{N}$ = Population mean of x.

n_h : Size of the hth stratum

$f_h = \frac{n_h}{N_h}$: Sampling fraction in hth stratum.

$n = \sum_{h=1}^{L} n_h$ = Total sample size.

$\bar{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi}$: Sample mean of y in the hth stratum.

$\bar{x}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} x_{hi}$: Sample mean of x in the hth stratum.

$s_{hy}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(y_{hi} - \bar{y}_h \right)^2$: Sample variance of y in the hth stratum.

$s_{hx}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(x_{hi} - \bar{x}_h \right)^2$: Sample variance of x in the hth stratum.

1.2.3 Stratified Estimators

Let $\bar{y}_{st}$ be a stratified estimator of the population mean $\bar{Y}$. Then, an unbiased estimate of $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \tag{1.1}$$

and the variance of $\bar{y}_{st}$ is given by

$$V\left(\bar{y}_{st}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_{hy}^2 \,. \tag{1.2}$$

An unbiased estimate of the population total Y is given as

$$\hat{Y}_{st} = N\,\bar{y}_{st} \tag{1.3}$$

with variance

$$V\left(\hat{Y}_{st}\right) = N^2 \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_{hy}^2 \,. \tag{1.4}$$

The unbiased estimate of $V\left(\bar{y}_{st}\right)$ is given by

$$v\left(\bar{y}_{st}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 s_{hy}^2 \,. \tag{1.5}$$

1.2.4 Allocation of Sample Size in Stratified Sampling

In stratified sampling the allocation of the sample size to different strata is done by the consideration of three factors:

i) The total number of the units in the stratum i.e. stratum size.

ii) The variability within the stratum.

iii) The cost of taking observations per sampling unit in the stratum.

Then, the various methods of allocating sample to L strata are given as:

1. **Equal Allocation:** This allocation method suggests the sample size in each stratum is equal, i.e.

$$n_h = \frac{n}{L} \tag{1.6}$$

The variance of the estimator $\bar{y}_{st}$ using equal allocation is obtained by

$$V\left(\bar{y}_{st}\right) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_{hy}^2 - \frac{1}{N}\sum_{h=1}^{L} W_h^2 S_{hy} \tag{1.7}$$

2. **Proportional Allocation**: When no information except the stratum size N_h is available, in this method the allocation of a given sample of size n to different strata is done in proportion to their strata sizes, that is,

$$n_h \propto N_h$$

or $$n_h = n \cdot W_h \tag{1.8}$$

The variance of the estimator $\bar{y}_{st}$ using proportional allocation is obtained by

$$V\left(\bar{y}_{st}\right) = \frac{1}{n}\sum_{h=1}^{L} W_h S_{hy}^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_{hy}^2 \tag{1.9}$$

3. **Neyman Allocation:** The allocation of samples to different strata, which minimizes $V\left(\bar{y}_{st}\right)$ defined in (1.2) for a fixed total sample size $\sum_{h=1}^{L} n_h = n$, is known as Neyman allocation and is given by:

$$n_h = n \cdot \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \tag{1.10}$$

The variance of the estimator $\bar{y}_{st}$ using Neyman allocation is obtained by

$$V\left(\bar{y}_{st}\right) = \frac{1}{n}\left(\sum_{h=1}^{L} W_h S_{hy}\right)^2 - \frac{1}{N}\sum_{h=1}^{L} W_h S_{hy}^2 \tag{1.11}$$

4. **Optimum Allocation:** If c_h be the cost of collecting information from a unit in *h*th stratum, the allocation of samples to different strata, which minimizes $V\left(\bar{y}_{st}\right)$ defined in (1.2) for a fixed a fixed cost $C = c_o + \sum_{h=1}^{L} c_h n_h$, is obtained as:

$$n_h = \frac{C - c_0}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \cdot \frac{W_h S_h}{\sqrt{c_h}} \tag{1.12}$$

where c_0 is the overhead cost and *C* is the total cost of the survey.

1.2.5 Stratification

'Stratification' is known as the construction of non-overlapping homogeneous groups, called strata, of population units with respect to a stratification variable so that the maximum precision of the estimates is achieved.

In many surveys, the stratification is made based on a convenient manner such as the use of geographical regions (such as North, Central, West, etc.), administrative regions (such as provinces, districts, etc.), demographic or other natural characteristics (such as gender, age, race, income, etc.). For example, for an income and expenditure survey of a state, province, and districts may be considered as strata. For business surveys on employee size, production and sales, the stratification is usually based on industrial classifications. For agricultural surveys, villages and geographical regions may compose the strata. In survey of auditing financial transactions, the transactions may be grouped into strata on the basis of their nominal values such as high, medium, low, etc. While in marketing studies, where the target consumer population is defined, strata can be formed by sex, age, income or other demographic variables.

However, the stratification by convenient manner is not a reasonable criterion as the strata formed may not be internally homogeneous with respect to the variable of interest, which may end up with reduction of the precision of survey estimates. Thus, one has to look for another way of stratification, which is to determine optimum strata boundaries (OSB) using the study (or stratification) variable when its frequency distribution is known. Such strata are so obtained where they are more internally homogeneous and the precision of the estimate of population parameters (mean or total) are optimized. This stratification method, known as optimum stratification, is one of the focuses of this thesis and will be discussed more detail in Chapter 2.

1.2.6 Use of Auxiliary Variable for Stratification

Indisputably, optimum stratification discussed in previous section could be achieved effectively by having the distribution of the main study variable known, and create strata by cutting the range of the distribution at suitable points. It could be noted that thus far, research in this area have been made with an unrealistic assumption that stratification is based on the frequency distribution of study variable (y) itself. Thus, optimum stratification on the study variable is not feasible in practice since the study

variable is unknown prior to conducting the survey. The non-availability of knowledge about the main study variable forces one to substitute for it the distribution of another known closely related variable (x), called auxiliary variable. Often y is highly correlated with x such that the regression of y upon x has homoscedastic errors. In situations like this, stratification can be achieved using the auxiliary variable. By and large, auxiliary data are readily available or can be made available easily with minimum cost and effort.

If the stratification is made based on x, it may lead to substantial gains in precision in the estimate, although it will not be as efficient as the one based on y. However, if the regression of y on x fits well within all strata, the boundary points for both the variables should be nearly the same.

1.3 Cluster Sampling

One of the objectives of sample survey design is to obtain a specified amount of information about a population parameter at minimum cost. Since different sampling design may incur varying cost of surveys, the employment of appropriate sampling design is also very important to meet this objective. The cluster sampling, which is frequently used in surveys to estimate the parameters of a population, is such sampling design that sometimes gives more information per unit cost than many other designs.

In cluster sampling the population units are divided into non-overlapping groups, called ***cluster*** and these clusters are used as sampling units. Then, a simple random sample of n clusters from N clusters in the population is obtained and every single unit in these selected clusters are measured and recorded.

Cluster sampling is less costly as compared to simple or stratified random sampling if the cost of obtaining a sampling frame is very high or if the cost of measuring units increases as the distance among the units increases. For example, if we wish to estimate the average income per household in a large city, the use of simple random sampling will need a frame listing all households in the city, and this frame may be very costly or impossible to obtain. Similarly, the use of stratified random sampling is still required a frame for each stratum in the population and again getting such frame may

be costly or impossible. In such situation, if we could divide the city into regions such as blocks (or clusters of elements) and select a simple random sample of blocks from the population, the task is easily completed by using a frame that lists all city blocks. Then the income of every household within each sampled block could be measured (Scheaffer et al. 2012).

Thus, the two major reasons for using cluster sampling on a population are (Cochran, 1977):

1. Usually when a complete list of the population units (sampling frame) is not available and hence, this limits the use as sampling units.
2. And in those cases when a complete list (sampling frame) is available economic restrictions may compel researchers to take larger sampling units. For a given size of the sample, usually smaller sampling units give more precise results as compared to larger sampling units, but this heightens the costs incurred while locating and approaching smaller units to measure them.

There are two possibilities of clustering the sampling units:

1) Non-overlapping clusters of equal size, or
2) Non-overlapping clusters of unequal sizes

It is obvious that, if every cluster has the same number of units, the chance of selecting each cluster in the sample will be the same. Whereas, if the number of units in the clusters is different and known, then different probabilities proportional to the number of units in the clusters can be assigned before taking the sample. In the following sections, we discuss both the situations.

1.3.1 Non-overlapping Clusters of Equal Size

Let the population of NM units (elements) be divided into N clusters each of size M. Let a simple random sampling of n cluster be drawn and hence nM units are selected from the population. Let us define:

y_{ij} = Measurement on the jth element of the ith cluster in population/sample.

$y_i = \sum_{j=1}^{M} y_{ij}$ = Total of *i*th cluster.

$\bar{Y} = \frac{\sum_{i=1}^{N} y_i}{N} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}}{N}$ = Mean per cluster (or cluster mean) in the population.

$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N} y_i}{NM} = \frac{\bar{Y}}{M}$ = Mean per element in the population (or population mean).

$S_b^2 = \frac{1}{M(N-1)}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$ = Variance among the cluster totals.

$S^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ij} - \bar{\bar{Y}}\right)^2}{MN-1}$ = Population variance between elements.

$\rho = \frac{E\left(y_{ij} - \bar{\bar{Y}}\right)\left(y_{ik} - \bar{\bar{Y}}\right)}{E\left(y_{ij} - \bar{\bar{Y}}\right)^2} = \frac{2\sum_{i=1}^{N}\sum_{j<k}^{M}\left(y_{ij} - \bar{\bar{Y}}\right)\left(y_{ik} - \bar{\bar{Y}}\right)}{(M-1)(MN-1)S^2}$ = Intra-cluster correlation coefficient

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ = Sample mean per cluster.

$\bar{\bar{y}} = \frac{\bar{y}}{M} = \frac{\sum_{i=1}^{n} y_i}{nM}$ = Sample mean per element.

Then, in this situation, the sample mean per element

$$\bar{\bar{y}} = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ij} \qquad (1.13)$$

is an unbiased estimate of population mean per element $\bar{\bar{Y}}$ with variance

$$V\left(\bar{\bar{y}}\right)=\frac{1-f}{n}\cdot\frac{NM-1}{M^2\left(N-1\right)}S^2\left[1+\left(M-1\right)\rho\right] \tag{1.14}$$

where ρ is the intra-cluster correlation coefficient. When the population size NM is large, that is $NM-1\cong NM-M$, the variance of $\bar{\bar{y}}$ in (1.14) is equivalent to:

$$V\left(\bar{\bar{y}}\right)=\frac{1-f}{nM}S^2\left[1+\left(M-1\right)\rho\right] \tag{1.15}$$

When S_b^2 denotes the variance among the cluster totals on a single unit basis, then an alternative expression for $V\left(\bar{\bar{y}}\right)$ is given by

$$V\left(\bar{\bar{y}}\right)=\frac{1-f}{nM}S_b^2 \tag{1.16}$$

1.3.2. Non-overlapping Clusters of Unequal Sizes

Let the population units (elements) be divided into N clusters of sizes $M_i\ (i=1,2,\cdots,N)$ and $M_0=\sum_{i=1}^{N}M_i$ be the total number of units in the population. Let a SRS of n clusters be drawn from N clusters. Let us define:

M_i = Number of units in ith cluster; $i=1,2,\cdots,N$.

y_{ij} = Value of jth observation in ith cluster.

$y_i=\sum_{j=1}^{M_i}y_{ij}$ = Total of ith cluster.

$\bar{y}_i=\frac{1}{M_i}\sum_{j=1}^{M_i}y_{ij}=\frac{y_i}{M_i}$ = Mean for ith cluster.

$Y=\sum_{i=1}^{N}\sum_{j=1}^{M_i}y_{ij}=\sum_{i=1}^{N}y_i$ = Population total.

$\bar{Y} = \frac{Y}{N}$ = Population mean per cluster.

$\bar{\bar{Y}} = \frac{\sum_{i=1}^{N} y_i}{M_0}$ = Mean per element in the population (or population mean).

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ =Sample mean per cluster.

$\bar{\bar{y}} = \frac{1}{m_0}\sum_{i=1}^{n} M_i y_i$ = Sample mean per element. Where, $m_0 = \sum_{i=1}^{n} M_i$.

$\bar{\bar{y}}^* = \frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i y_i$ = Overall sample mean per element. Where, $\bar{M} = \frac{1}{N}\sum_{i=1}^{M} M_i$.

Then, in this situation, the estimator $\bar{\bar{y}}$ is a biased estimate of population mean $\bar{\bar{Y}}$ with variance

$$V\left(\bar{\bar{y}}\right) = \frac{1-f}{n} S_b^2, \tag{1.17}$$

where $S_b^2 = \frac{1}{\bar{M}^2\left(N-1\right)}\sum_{i=1}^{N} M_i^2\left(\bar{y}_i - \bar{\bar{Y}}\right)^2$.

However, the estimator $\bar{\bar{y}}^*$ is an unbiased estimate of population mean $\bar{\bar{Y}}$ with variance

$$V\left(\bar{\bar{y}}\right) = \frac{1-f}{n} S_b^{*2}, \tag{1.18}$$

where $S_b^{*2} = \frac{1}{\left(N-1\right)}\sum_{i=1}^{N}\left(\frac{M_i}{\bar{M}}\bar{y}_i - \bar{\bar{Y}}\right)^2$.

1.3.3 Optimum Cluster Size

Since the size of the clusters or sampling units influences the efficiency of the sampling design, the practitioner has to decide its optimum size in using the cluster sampling technique effectively. In sampling literature, when a single characteristic is under study, many authors have proposed various techniques to determine the cluster size that gives maximum precision within the available resources. However, when more than one characteristic are under study, the procedures for determining optimum cluster size are not well defined in the literature. In this thesis, the problem of determining optimum compromise cluster size and sampling unit of equal size for multivariate cluster sampling, which is formulated as All Integer Nonlinear Programming Problems (AINLPP), is discussed.

1.4 Mathematical Programming Problem (MPP)

The Mathematical Programming is a technique used to determine the optimum value (maximum or minimum) of a function of several decision variables which are subjected to a number of constraints. The technique is commonly used to solve many decision making problem in the areas such as sociology, social psychology, demography, political science, economics, education, public health and many others. The essential components of an MPP model are the values of the decision variables, which describe the solutions; the objective function which measures the quality of solutions; the constraints which presents the relationships between decision variables. A general form of an MPP can be stated as follows:

$$\begin{aligned} &\text{Maximize (or Minimize)} \qquad Z = f(x_1, x_2, x_3, ..., x_n) \\ &\text{subject to} \qquad g_i(x_1, x_2, x_3, ..., x_n) \{\leq, \equiv, \geq\} 0; \quad i = 1, 2, 3, ..., m \qquad (1.19) \\ &\text{and} \qquad x_j \geq 0; \quad j = 1, 2, 3, ..., n. \end{aligned}$$

Where in (1.19) only one sign among $\leq, \equiv, \geq$ holds true for each *i.* Usually, unless specified otherwise, in an MPP all the involved functions are assumed to be continuously differentiable.

The variables $x_j; j = 1, 2, 3, ..., n$ are called decision variables. If all the functions in an MPP are linear functions of the decision variables, the MPP is called a Linear Programming Problem (LPP). Similarly, if some or all the functions are nonlinear, the MPP is called a Nonlinear Programming Problem (NLPP). In an NLPP, when all the variables are restricted to be integers, the problem is known be an All Integer Nonlinear Programming Problem (AINLPP).

Depending on the nature of the functions involved and the restrictions on the objective function and the decision variables, one customarily distinguishes the MPP into the following branches:

- Integer Programming Problem (IPP)
- Quadratic Programming Problem (QPP)
- Convex Programming Problem (CPP)
- Separable Programming Problem (SPP)
- Multi-Objective Programming Problem (MOPP)
- Fractional Programming Problem (FPP)
- Geometric Programming Problem (GPP)

1.5 Mathematical Programming in Sampling

Sampling, which is the selection of 'part' (sample) of an aggregate to present the 'whole' (population), is used most frequently in surveys. The purpose of the sample survey is to obtain information about the population which is defined according to the aims and objectives of the survey. Since the information on the population is based on sample, in planning of sample survey, a stage is always reached at which a decision must be made about the size of the sample, the size of the sampling unit, the sampling scheme, the scope of the survey, the number of strata and strata boundaries (in case stratified sampling is used), etc. These decisions are very important. For example, the decision regarding the size of a sample to be selected is important because too large a sample implies a waste of resources and too small a sample diminishes the utility of the results obtained. Therefore, the problem of deriving the statistical information on population characteristics can be formulated as an MPP by minimizing the cost of survey subject to the restriction that the loss of precision is within a certain prescribed

limit or alternatively minimizing the loss in precision subject to the restriction that the cost of the survey remains within the given budget.

The problem studied in this thesis such as, in stratified random sampling, the problem of determining the number of strata, the problem of cutting the stratum boundaries, the problem of optimum allocation of sample sizes to various strata are treated as MPPs and solved by using MP technique by several authors. Chapter 2 is devoted to study these problems.

Similarly, in cluster sampling, the problem of choosing the optimum cluster size is also an MPP that can be solved by using MP techniques. In multivariate cluster sampling where more than one characteristics are to be measured on every selected elements, the above problems become more complicated because of the non-availability of a single optimality criterion which is suitable for all the characteristics. In such situations, a compromise is essential and one has to device a criterion which is optimum in some sense for all characteristics under study. The problem arising in multivariate cluster sampling can also be formulated as MPPs and is solved by using MP technique. The Chapter 3 of this manuscript is devoted to study this problem.

1.6 Objectives of the Study

As discussed in earlier sections, sample survey is a method of gathering information that select a sample of elements from the target population in order to estimate population attributes. Since the estimate is based on the sample, it is very important that the information obtained in surveys is as accurate as possible by employing appropriate sampling technique and determining appropriate sample size. Otherwise conducting the survey will be a waste of time, money and resources. Due to this, there has been a wide range of research done to try and improve ways of conducting the sample survey.

Stratified sampling is one of the most widely used sampling techniques in survey as it increases the precision of the estimate of the survey variable. However, in order to achieve maximum precision in the survey estimates using stratified sampling, the three important issues that must be addressed by the surveyors are: the determination of the

number of strata, the determination of strata boundaries and the determination of the size of samples from the strata.

In this research the problem of determining optimum strata boundaries (OSB) and the problem of determining sample size to stratum are considered. Many authors have studied both of these problems separately. However, studying separately the problem, one may not achieve the goal that the precision of the estimate is maximized. Thus, the research carried in this thesis is an attempt to determine the OSB and the allocation of samples to strata simultaneously that minimizes the variance of the estimate of the survey variable.

Another very common sampling method for survey is cluster sampling. This technique is widely used, when the complete list of the population units is not available and/or the reduction of the cost of survey is taken into consideration. One of the major problems in using cluster sampling is the determination of the optimum number of clusters and the cluster size. When a univariate population is studied, these problems were considered by several authors and are well known in sampling literature. However, in multivariate surveys where more than one characteristic are to be measured on each sampled unit, the optimum number of clusters and cluster size for the individual characteristics are of not much practical use. This is because the number of clusters that is optimum for one characteristic will generally be far from optimum for others. Thus, this research is also an attempt to develop a technique to determine the optimum number of clusters of equal size for multivariate surveys, which is optimum in some sense for all characteristics.

The problems arising in stratified sampling and cluster sampling as discussed above are usually nonlinear programming problems with nonlinear objective function and linear/nonlinear restriction on decision variables. The researches that are carried out and presented in this thesis are some attempts to effectively develop some techniques of MPP to solve these problems.

If stratified random sampling and multivariate cluster sampling are used in a survey to estimate population parameters in a survey, the specific objectives in this study are:

In stratified sampling:

1) To formulate the problem of determining OSB and optimum allocation of sample to strata as a Mathematical Programming Problem (MPP), when the study variable in the population has a known frequency distribution.

2) To formulate the problem of determining OSB and optimum allocation of sample to strata as a Mathematical Programming Problem (MPP), when the auxiliary variable in the population has a known frequency distribution.

3) To develop appropriate solution procedure for solving the problems discussed in objectives (1) and (2) to stratify the population, respectively using study or auxiliary variable.

In multivariate cluster sampling:

1) To formulate the problem of determining the optimum number of cluster (n) and the cluster size (M) as a Mathematical Programming Problem (MPP), when more than one characteristic are under study.

2) To develop appropriate solution procedure for solving the problem discussed in objective (1).

1.7 Review of Literature and Studies

This section presents the literature review and studies on the problem of determining optimum strata boundaries and optimum cluster size, respectively in Sections 1.7.1 and 1.7.2.

1.7.1 Optimum Strata Boundaries: Review of Literature and Studies

'Stratification' is known as the problem of constructing homogeneous groups of sampling units with respect to stratification variable so that the maximum precision of the estimates is achieved.

Thus, it is important while determining the optimum strata boundaries (OSB) that the strata are as internally homogeneous as possible. Accurate optimum strata boundaries

(OSB) will ensure maximum precision, which also means that stratum variance σ_h^2 needs to be as minimum as possible for a specified sample distribution.

Dalenius (1950) was one of the pioneer researchers to work on the problem of determining OSB, where both the estimation and the stratification variables were the same. In his study, he produced a set of minimal equations that were usually difficult to solve for OSB because of their implicit nature.

Drawing on from his work, it can be said that given a single characteristic is the focus of the study, then stratification process in principle, will be based on this study variable, and the idyllic condition would be that the distribution of the study variable is identified and the OSB can be determined by cutting the range of this distribution at suitable points.

Dalenius and Gurney (1951) were also interested in studies where the survey/study variable and the stratification variable were different. The approach taken in this study involved taking iterative steps starting from a conveniently chosen set of values, to arrive at an optimum set of values. Also, this study implicated that the strata boundaries could be determined if $W_h\sigma_h$ remained constant, where W_h is the weight of the h^{th} stratum. Other researchers also showed similar interest in determining approximately optimum strata boundaries with varying variables but the problem proves to be complex and elaborate.

Some of the researchers that have the worked on OSB when $W_h\sigma_h$ was kept as constant were Mahalanobis (1952) and Hansen, Hurwitz and Madow (1953). One of the conditions that they all applied was that the stratum totals were equal. This meant that within the strata, coefficients of variation were equal and remained the same even when the strata sizes were altered or adjusted; to note if there were any changes on the OSB. It was noted that when the coefficient of variation was constant in all strata, both the rules resulted in the same solution. This overall governing rule, where $W_h\sigma_h$ remains constant, proves significantly simple and at the same time it has been argued that the conditions under which it gives optimum set of stratification points are satisfied by a large number of real populations. However, studies conducted by Sethi (1963) claimed otherwise, he showed that the rule advocated by Hansel et al. (1953) did not lead to solutions of optimal points of stratification in all types of population.

He proposed that the boundaries could be obtained if minimal equations are solved using calculus methods, that is, optimum boundaries could be arrived at by solving the minimal equation: $\left(x_h-\mu_h\right)+\sigma_h^2/\sigma_h=\left(x_{h+1}-\mu_{h+1}\right)+\sigma_{h+1}^2/\sigma_{h+1}$. Aoyama (1954) continued with the studies and proposed an approximate rule and highlighted that the strata needed to be of equal width x_h-x_{h-1}, where x_h and x_{h-1} specifies the boundaries of the h^{th} stratum. Ekman (1959) further refined this condition to state the equation as: $W_h\left(x_h-x_{h-1}\right)=\text{Constant}$.

When the frequency distribution $f(x)$ of auxiliary variable x is known, many authors have made contributions suggesting near exact or approximate solution to the problem of determining the strata boundaries. A popular and much more convenient approximation solution was first proposed by Dalenius and Hodges (1959). This method involved constructing the strata by taking equal intervals on the cumulative of $\sqrt{f(x)}$. Later, Serfling (1968) extended this cum $\sqrt{f}$ method to obtain the optimum strata boundaries for both the estimation and auxiliary variables. Consequently, Cochran (1977) further supported the above cum $\sqrt{f(x)}$ rule about its worthiness especially given the condition when the regression of *y* on *x* is linear and ρ (correlation coefficient) is nearly perfect. However, some disadvantages of cumulative root frequency model have also been cited. It has been noted that in the cumulative root frequency model the number of initial class intervals and the intervals are arbitrary, and there is no theory on how to choose the optimum number of classes (see Heldin, 2000). Many other authors also used auxiliary information and developed stratification methods, such as Singh and Sukhatme (1969, 1972, 1973) extended the technique of minimal equations under Neman allocation for fixed sample size *n*. Singh (1971) continued working on Dalenius and Hodges (1959) method in an effort to improve its results, and again Singh and Prakash (1975) suggested modifications to Dalenius and Hodges' cum $\sqrt{f}$ stratification rule when auxiliary data are used. Many other authors such as Taga (1967), Mehta et al (1996), Rizvi et al (2002) and Gupta et al. (2005) have come up with a yet other approximation methods to determine strata boundaries.

While comparing different stratification method of with other classical approximate method, it is noted that the Ekman method and the Delanius and Hodges method have shown consistent results (see Cochran 1961; Hess, Sethi and Balakrishnan 1966;

Murthy 1967) but the latter is more convenient and easier to apply , (see Nicoloni 2001).

Unnithan (1978) put forward an iterative method using Shanno's modified newton method for determining the strata boundaries that resulted to a local minimum of the variance for Neyman allocation, however, a suitable initial solution had to be selected. This modified version proved to be faster when compared to a Dalenius and Hodges iterative procedure. After further studies Unnithan and Nair (1995) came up with a method of selecting an appropriate starting point for modified Newton method showing promising results leading to a global minimum of the variance.

Many times, in real-world applications data are skewed, therefore a certain stratum is required while stratification to increase the precision. For such situation Lavallée and Hidiroglou (1988) proposed an algorithm that worked to construct stratum boundaries for a power allocated stratified sample of non-certainty sample units. Following on this research study, Hidiroglou and Srinath (1993) showcased a more general form of algorithm, where by assigning different values to operating parameters modifies to a power allocation, a Neyman allocation, or a combination of these allocations.

Further studies on Lavallée and Hidiroglou algorithm carried out by Sweet and Sigman (1995) and Rivest (2002) for several strata showed that the algorithm convergence was slow or non-existent. Moreover, their studies showed that the different starting points could lead to different OSBs for the same population, and that the boundaries differed considerably.

Neimiro (1999) studied the effectiveness of a random search method in the stratification problem, but studies did not show any conclusive results that the algorithm would guarantee a global optimum. Another downfall of this method was that in a case of a large population, too many iteration steps (see Kozak 2004) would be necessary and thus incorrect results could be obtained.

Yet again, another alternative to the most popular Delanius and Hodges method was suggested by Nicolini (2001), which became known as the Natural Class Method (NCM), but its use remained only for use when a large number of strata was needed.

Lednicki and Wieczorkowski (2003) pursued on a method of stratification using the simplex method of Nelder and Mead (1965). Again this suggestion by Lednicki and Wieczorkowski (2003) had its own disadvantages, namely, it did not come up with the optimal values when a large number of variables were studied in addition to, being recognized as a significantly a slow method.

A year later, Kozak (2004) further the enhanced the random search algorithm of the optimal stratification. This time Kozak's algorithm displayed itself to provide faster and efficient results in relation to Rivest, and Lednicki and Wieczorkowski, however, it failed to assure users that the algorithm would lead to the global optimum.

Gunning and Horgan (2004) came up with another method, known as geometric method, to achieve approximate stratification for skewed population and they claimed that their method is better than the CRF method. Horgan (2006) carried out tests to compare this method to the Dalenius and Hodges (1959), Ekman (1959) and Lavallée and Hidiroglou (1988). He concluded that the geometric progression was a better method compared to other methods. This claim however, was refuted by Kozak and Verma (2006). Their research led them to believe that the algorithm put forward by Lavallée and Hidiroglou's was in fact a preferred method when compared to geometric progression method (see Kozak et al. 2007). Some disadvantages in implementing the geometric method are that it does not work well in normal and symmetric distributions. The method also performs poorly when extreme outliers are present (see Kozak and Verma, 2006; Kozal et al., 2007; Keskintürk & Er, 2007; Brito et al., 2010; Baillargeon & Rivest, 2009; Horgan, 2011).

Later, a genetic algorithm was suggested by Keskintürk and Er (2007) to solve the combined problem of finding strata boundaries and optimum allocation for finite populations. They compared the performance of their algorithm with CRF, geometric and modified geometric using some real and simulated populations and concluded that the best results are obtained using the genetic algorithm. Brito et al. (2010) proposed another algorithm, known as an iterative local search (ILS) metaheuristic algorithm, to determine strata boundaries which is designed to work for variables with any distribution. The method was tested for a variety of skewed populations and it was found that the algorithm works better than Kozak algorithm in most cases.

Recently, Yong et al. (2016) developed a technique for determining optimal stratification using baseline information in the area of predictive medicine with a desirable stratification scheme that would not only have small intra-stratum variation but also have a clinically meaningful discriminatory capability. They used the proposed technique in an AIDS clinical trial data with binary outcomes and a cardiovascular clinical data for censored event time outcomes.

In Khan et al (2008), there is mention of a computational technique developed by Bühler and Deutler (1975). Bühler and Deutler developed the problem of determining OSB as an optimisation problem and used dynamic programming to find solutions. This technique was tried out by Lavallée (1987, 1988) for determining OSB where the population domain was categorised into two stratification variables into distinct subsets, in an effort to maximise the precision of the variables of interest.

Khan et al (2002, 2005, 2008, 2009) and Nand and Khan (2009) worked on this procedure further to apply the dynamic programming technique in determining OSB when the frequency function of the study variable is known. They took the initiative to connect the problem of finding OSB as an equivalent problem of determining optimum strata width (OSW), which was formulated as an MPP and solved through dynamic programming technique. They made use of this to technique to define the OSB to determine the OSB for the populations having study variable that follows uniform, right triangular, exponential, triangular, normal, Cauchy and power distribution. The advantage of this technique is that it does not require any initial solution and it can be used even when the complete dataset of the stratification variable in unavailable. The technique requires only the frequency distribution and their parameters.

1.7.2 Multivariate Cluster Sampling: Review of Literature and Studies

The cluster sampling technique is widely practiced in sample surveys. In this technique, the total population is divided into groups (or clusters) of smaller units of the population, called sampling units. Once a sample unit has been defined, a simple random sample of the sampling units is selected. Since the cluster sampling helps in reducing the cost of surveys by not requiring the entire sampling frame, it is regarded as a very reasonable method of gathering information. The size of these clusters or

sampling units is the decisive cornerstone; as this decision is very crucial to the efficacy of sampling and thus giving rise to the problem of determining the optimum cluster size and the number of sampling units. In such scenarios it becomes essential to choose the cluster size which is optimum with respect to a carefully selected criterion.

In cluster sampling, for a given sample size, the sampling variance increases with the increase in cluster size and decreases with the number of sampling units (clusters). On the other hand, the cost of the survey decreases with the increase in cluster size and increases with the number of sampling units. Thus, several authors have successfully studied to compromise between the cluster size and the sampling units, when a single characteristic is under study, have worked out to determine the optimum cluster size and sampling units that minimizes the sampling variance for a fixed cost of the survey or to minimizes the cost for the given precision.

Jessen (1942) presented a comparative study of the relative standard errors of the estimates of various characters in a sample survey of farms taking seven different sampling units. Mahalanobis (1940, 1942, 1944) considered in details the question of determining the optimum cluster size in case of crop surveys from the point of view of both cost and variance on the basis of the extensive empirical studies carried out during the period 1937-1941. Sukhatme (1950) and Sukhatme and Panse (1951) experienced the same findings as Jessen (1942) regarding the use of the villages as unit of sampling in agricultural surveys. Singh (1956) also studied on the efficiency of cluster sampling and cluster size. Hansen, Hurwitz and Madow (1953) obtained the optimum cluster size and sampling unit size by substituting different values of sample size in in the cost function and working out the corresponding values of cluster size for fixed cost and selecting that pair which gives the minimum variance. Cochran (1963) gave an algebraic solution for continuous variations in the cluster size to the problem and suggested a trial and error method to obtain the optimum value of the cluster size. Murthy (1967) used graphical method for obtaining the optimum cluster size. Jessen (1978) presented a solution for the same problem with a simple cost function and for known or estimated value of measure of homogeneity (intra class correlation coefficient). Later, Khan, Jahan and Ahsan (1997) discussed a problem of determining the optimum cluster size and sampling units in without replacement

simple random sampling (SRSWOR) of clusters of equal sizes. They formulated the problem as a nonlinear programming problem (NLPP) that seeks minimization of a convex objective function for a specified range of intra class correlation coefficient subject to a cost constraint. The NLPP is then solved using a Lagrange multiplier technique to obtain explicit formulae for optimum cluster size and sampling units.

However, when more than one characteristic are under study, the procedures for determining optimum cluster size are not well defined in sampling literature. It is not very recently but Sheela and Unnithan (1992) extended a problem of determining the optimum plot (cluster) size in agriculture surveys to the multivariate case. Thus, in this thesis, an attempt is made to study the problem and to develop a technique that may address the problem of determining optimum compromise cluster size and sampling unit of equal size for multivariate cluster sampling.

Chapter 2

Determining Optimum Strata Boundaries and Optimum Allocation in Stratified Sampling

2.1 Introduction

The construction of homogenous strata, known as the problem of determining *optimum stratum boundaries* (OSB), can be achieved efficiently when the frequency distribution of (y) is known. This problem was first discussed by Dalenius (1950). As discussed in previous chapter, he presented a set of minimal equations for finding the OSB. Unfortunately these equations could not usually be solved because of their implicit nature. Thus, attempts have been made by several authors to obtain the strata boundaries including Dalenius and Gurney (1951), Mahalanobis (1952), Hansen, et al. (1953), Aoyama (1954), Dalenius and Hodges (1959), Ekman (1959), Sethi (1963), Unnithan (1978), Lavallée and Hidiroglou (1988), Hidiroglou and Srinath (1993), Sweet and Sigman (1995), Hedlin (2000), Rivest (2002), Lednicki and Wieczorkowski (2003), Gunning and Horgan (2004), Kozak (2004), Keskintürk and Er (2007), etc. They used the frequency distribution of the main study variable and proposed different techniques for determining the strata boundaries under various allocations. Most of these authors achieved the calculus equations for the strata boundaries which are not suitable for practical computations. They obtained only the approximate solutions of an OSB under certain assumptions. Also, when the frequency distribution of the auxiliary variable (x) is known, many authors such as Dalenius (1957), Taga (1967), Singh and Sukhatme (1969, 1972, 1973), Singh and Prakash (1975), Singh (1971, 1975), Mehta et al. (1996), Rizvi et al. (2002), and Gupta et al. (2005) have suggested different approximation method of determining OSB.

The main problem for many of these techniques discussed above is that the authors disregarded the sample allocation problem while constructing the OSB and kept it

separate assuming that either the proportional, Neyman or optimum allocation will be used. However, one must be mindful that such OSB may not always be feasible or may be non-optimal, especially when the populations are small and/or skewed. Thus, while stratifying a population the problem of optimum allocation of sample size to the strata should also be considered simultaneously.

In this chapter, we consider both the problems and attempt to determine both the OSB and sample allocation simultaneously when the population mean of the study variable y is of interest and a frequency distribution $f(y)$ or the frequency distribution $f(x)$ of its auxiliary variable x is available.

Section 2.2 provides the formulation of the problem of finding OSB and sample sizes as a Mathematical Programming Problem (MPP) that seeks minimization of the variance of the estimated population parameter of the target population, which is subjected to a fixed total sample size. In Section 2.3, a solution procedure using LINGO is discussed to solve the MPP.

Section 2.4 illustrates the computational details to demonstrate the practical application of the proposed method using two numerical examples, when the study variable or the auxiliary variable follows a uniform and a right-triangular distribution in the population. The proposed technique can easily be applied to other frequency distributions. Finally Section 2.5 provides a conclusion of the chapter.

The work presented in this chapter has already been presented in the VII International Symposium on Optimization and Statistics & III National Conference on Statistical Inference held at Aligarh Muslim University, India during Dec 21-23, 2012. A paper entitled "Determining Optimum Strata Boundaries and Optimum Allocation in Stratified Sampling" based on this chapter has also been published in *Aligarh Journal of Statistics*, Vol. 35, 23-40.

2.2 Formulation of the Problem as an MPP

2.2.1 Problem of Determining Optimum Sample Sizes:

Let the population be stratified into L strata based on a study variable y and the estimate of population mean $\bar{Y}$ is of interest. If n_h be a simple random sample drawn independently from h^{th} stratum and $\bar{y}_h$ denotes an unbiased sample estimate of $\bar{Y}_h$ for the characteristic in stratum h; $h = 1, 2, ...L$, then an unbiased estimate of the population mean $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h, \tag{2.1}$$

$W_h = \frac{N_h}{N}$, where N_h = size of the stratum h, and $N = \sum_{h=1}^{L} N_h$.

The sampling variance of the estimate $\bar{y}_{st}$ is given by

$$\text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N} . \tag{2.2}$$

Where, S_{hy}^2 denotes the stratum variance of the characteristic *y* in the h th stratum.

If the total sample size *n* for a stratified survey is predetermined, a reasonable criterion for obtaining the optimum allocation n_h is to minimize the variance, $\text{var}(\bar{y}_{st})$, of the stratified sample mean $\bar{y}_{st}$ given in (2.2).

Further, for practical application of an allocation the integer values of the sample sizes are required. They could be obtained by simply rounding off non-integer sample sizes

to their nearest integer values. However, in many situations the rounded-off sample allocation may be infeasible or non-optimal.

Considering the above facts, the problem of finding the allocation $(n_1,...,n_L)$ for a fixed sample size *n* may be given as the following MPP:

$$\text{Minimize } \operatorname{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N}$$

$$\text{subject to} \quad \sum_{h=1}^{L} n_h = n;$$

$$1 \le n_h \le N_h \text{ and } n_h \text{ are integers}, (h = 1,2,...,L). \tag{2.3}$$

Note that the restrictions $n_h \le N_h$ are imposed to avoid oversampling and the restrictions $1 \le n_h$ are imposed to ensure the representation of every stratum in the sample at least by one unit.

2.2.2 Problem of Determining OSB:

Let the population be stratified into L strata based on the study variable *y* and $f(y)$ denotes frequency function of *y*. If y_0 and y_L be the smallest and largest values of *y*, then a problem of determining the strata boundaries is to cut up the range,

$$y_L - y_0 = d \text{ (say)}, \tag{2.4}$$

at intermediate points $y_1 \le y_2 \le,...,\le y_{L-1}$ such that the variance of the stratified sample mean, $\operatorname{var}(\bar{y}_{st})$, given in (2.2) is minimum.

Thus, the problem of determining the OSB is to be stated as:

$$\text{Minimize } \operatorname{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N}$$

$$\text{subject to } y_0 \le y_1 \le y_2 \le, ..., \le y_{L-1} \le y_L . \tag{2.5}$$

It can be noted that, if the target population is small, one cannot expect that the sampling fraction n_h / N_h is negligible and hence the finite population correction in the objective function of the (2.5) cannot be ignored. Also, when frequency function of the study variable $f(y)$ is known, the values of W_h and S_{hy}^2 can be expressed as a function of boundary points (y_{h-1}, y_h) of h th stratum by

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy \tag{2.6}$$

$$S_{hy}^2 = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y^2 f(y)dy - \mu_{hy}^2 \tag{2.7}$$

Where,
$$\mu_{hy} = \frac{1}{W_h} \int_{y_{h-1}}^{y_h} y f(y)dy \tag{2.8}$$

Let $d_h = y_h - y_{h-1} \ge 0$ denotes the width of the h th ($h = 1, 2, ..., L$) stratum.

With the above definition, (2.4) is expressed as

$$\sum_{h=1}^{L} d_h = \sum_{h=1}^{L} (y_h - y_{h-1}) = y_L - y_0 = d$$

Thus, the k th stratification point y_k; $k = 1, 2, ..., L-1$ can be expressed as:

$$\begin{aligned} y_k &= y_0 + d_1 + d_2 + ... + d_k \\ &= y_{k-1} + d_k \end{aligned}$$

Then, the problem of determining OSB in (5) can also be considered as the problem of determining optimum strata widths and may be expressed as the following MPP:

$$\left.\begin{array}{ll} \text{Minimize} & \text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N} \\ \text{subject to} & \sum_{h=1}^{L} d_h = d, \\ \text{and} & d_h \geq 0;\ h = 1, 2, ..., L. \end{array}\right\} \qquad (2.9)$$

2.2.3 Problem of Determining OSB and Sample Sizes:

If the two problems discussed in Sections 2.2.1 and 2.2.2 are to be solved simultaneously, then one way of achieving this by merging them into a single problem. Therefore, merging (2.3) and (2.9), the problem of determining OSB and the optimum allocation of sample size is formulated as an MPP as follows:

$$\left.\begin{array}{ll} \text{Minimize} & \text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2 S_{hy}^2}{n_h} - \sum_{h=1}^{L} \frac{W_h S_{hy}^2}{N} \\ \text{subject to} & \sum_{h=1}^{L} d_h = d, \\ & \sum_{h=1}^{L} n_h = n, \\ \text{and} & d_h \geq 0;\ 1 \leq n_h \leq N_h;\ n_h \text{ integers}, h = 1, 2, ..., L. \end{array}\right\} \qquad (2.10)$$

2.2.4 Problem of Determining OSB and Sample Sizes using Auxiliary Variable:

Indisputably, optimum stratification could be achieved effectively by having the distribution of the main study variable known, and create strata by cutting the range of the distribution at suitable points. In Section 2.2.3, the problem of determining OSB and sample allocation is formulated based on the of study variable (y) itself and its frequency distribution $f(y)$ is assumed to be known. However, this may not be possible in practice since in many situations the study variable is unknown prior to conducting the survey, which leads to use of the distribution of closely related variable (x), called auxiliary variable. Often y is highly correlated with x such that the regression of y upon x has homoscedastic errors. In situations like this, stratification can be achieved using the auxiliary variable. By and large, auxiliary data are readily available or can be available easily with minimum cost and effort.

Moreover, if the stratification is made based on x, it may lead to substantial gains in precision in the estimate, although it will not be as efficient as the one based on y. However, if the regression of y on x fits well within all strata, the boundary points for both the variables should be nearly the same.

Consider the regression model:

$$y = \lambda(x) + \varepsilon, \tag{2.11}$$

where $\lambda(x)$ is a linear or nonlinear function of x and ε is an error term such that $E[\varepsilon|x] = 0$ and $\text{var}[\varepsilon|x] = \phi(x)$ for all x.

Under the model (2.11), the stratum mean μ_{hy} and the stratum variance S_{hy}^2 can be expressed as (see Singh and Sukhatme, 1969):

$$\mu_{hy} = \mu_{h\lambda} \tag{2.12}$$

and

$$S_{hy}^2 = S_{h\lambda}^2 + \mu_{h\phi} \tag{2.13}$$

where $\mu_{h\lambda}$ and $\mu_{h\phi}$ are the expected values of $\lambda(x)$ and $\phi(x)$ respectively, and $S^2_{h\lambda}$ denotes the variance of $\lambda(x)$ in the *h*th stratum.

If λ and ε are uncorrelated, from the model (2.11), S^2_{hy} can also be expressed as (see Dalenius and Gurney 1951):

$$S^2_{hy} = S^2_{h\lambda} + S^2_{h\varepsilon}, \tag{2.14}$$

where $S^2_{h\varepsilon}$ is the variance of ε in the *h*th stratum. It is, therefore, minimizing (2.2) is equivalent to minimizing

$$\text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2\left(S^2_{h\lambda} + \mu_{h\phi}\right)}{n_h} - \sum_{h=1}^{L} \frac{W_h\left(S^2_{h\lambda} + \mu_{h\phi}\right)}{N}. \tag{2.15}$$

Let $f(x);\ a \le x \le b$ be the frequency function of the auxiliary variable *x*, which is used for determining OSB by cutting its range $d = b - a$ at $(L-1)$ intermediate points $a = x_0 \le x_1 \le x_2 \le,...,\le x_{L-1} \le x_L = b$ such that (2.15) is minimum.

Thus, from (2.10), the OSB and the optimum allocation can be formulated as the following MPP:

$$\left.\begin{aligned} \text{Minimize} \quad & \text{var}(\bar{y}_{st}) = \sum_{h=1}^{L} \frac{W_h^2\left(S^2_{h\lambda} + \mu_{h\phi}\right)}{n_h} - \sum_{h=1}^{L} \frac{W_h\left(S^2_{h\lambda} + \mu_{h\phi}\right)}{N} \\ \text{subject to} \quad & \sum_{h=1}^{L} d_h = d, \\ & \sum_{h=1}^{L} n_h = n \\ \text{and} \quad & d_h \ge 0;\ 1 \le n_h \le N_h;\ n_h \text{ integers}, h = 1,2,...,L. \end{aligned}\right\} \tag{2.16}$$

For a known $f(x)$, the values of $S_{h\lambda}^2$ and $\mu_{h\phi}$ can be expressed as a function of boundary points (x_{h-1}, x_h) of h th stratum by

$$S_{h\lambda}^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda(x)^2 f(x) dx - \mu_{h\lambda}^2 \tag{2.17}$$

and $$\mu_{h\phi} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \phi(x) f(x) dx \tag{2.18}$$

Where,

$$W_h = \int_{x_{h-1}}^{x_h} f(x) dx \tag{2.19}$$

and $$\mu_{h\lambda} = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} \lambda(x) f(x) dx \tag{2.20}$$

2.3 The Solution Procedure Using LINGO

When the number of strata (L) and the total sample size (n) are predetermined, the MPP (2.10) and (2.16) may be solved by executing a program developed in the LINGO software package for a known $f(y)$ or $f(x)$ as the case may be.

2.4 Numerical Illustrations

In order to demonstrate the computational details of the proposed technique, two sets of populations that follow respectively uniform and right-triangular distributions are considered.

2.4.1 Population 1: Uniform Distribution

The uniform distribution is a family of continuous probability distributions. It is frequently a probability model of many events of items that has equal probability of occurrence over a given range. The distribution is defined by the two parameters, *a* and *b*, which are its minimum and maximum values.

The general formula for the probability density function of the uniform distribution is

$$f(x) = \begin{cases} \frac{1}{b-a}; & a \le x \le b \\ 0; & \text{otherwise} \end{cases} \qquad (2.21)$$

Some continuous variables in the engineering, industry, management, and biological sciences have uniform probability distributions. For example, in a survey of telecom industry, the actual time of occurrence of one telephone call arrived at switchboard within one interval, say $(0,t)$ is distributed uniformly over this interval. Similarly, the delivery time of equipment in an interval, or selecting a location to observe the work habit of workers in a certain assembly line, etc. are uniformly distributed (see Wackerly et al. 2008).

2.4.1.1 Estimating the distribution of the population:

The data set of Population 1 of size $N = 1000$ provides the information of the study variable *y* that has the minimum value y_0 = 0.000391 and the maximum value y_L = 0.998604. Thus, the dataset gives the range of the distribution as $d = y_L - y_0 = 0.998213$. To determine the distribution, we construct a relative frequency histogram of *y* as shown in Figure 2.1.

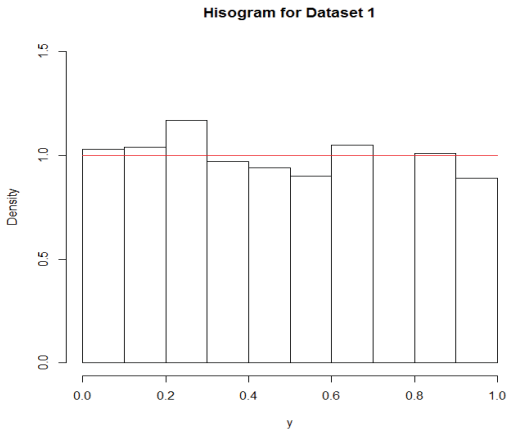


Figure 2.1: Frequency Histogram for *y*.

The Q-Q plot for *y* in Figure 2.2 and the Kolmogrov-Smirnov test (D=0.0291, *p*-value = 0.3653) also confirm that *y* has a uniform distribution.

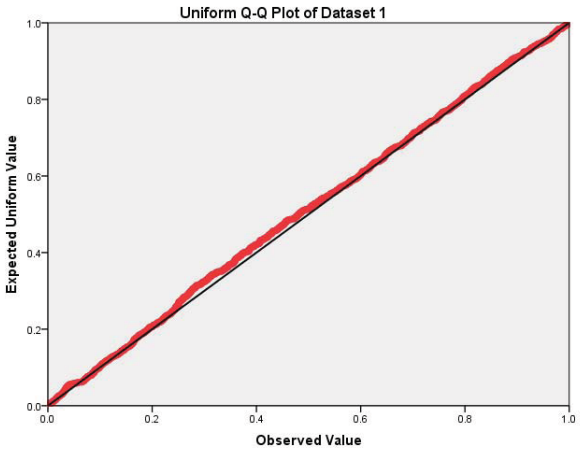


Figure 2.2: Q-Q plot for *y*.

For testing whether a particular distribution, if needed, one can use some method of goodness of fit. Also there are some software (such as, EasyFit at mathwave.com: http://www.mathwave.com/easyfit-distribution-fitting.html), which allow to automatically or manually fit a large number of distributions including uniform distribution to the data.

2.4.1.2 Formulation of the Problem Determining OSB and Sample sizes as an MPP:

As the study variable *y* has uniform distribution with density function $f(y)$ given in (2.21), using (2.6), (2.8) and (2.7), we obtain W_h (stratum weight), μ_{hy} (stratum mean) and S_{hy}^2 (stratum variance), respectively, as follows:

From (2.6), the **stratum weight** (W_h) is obtained as

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy$$
$$= \int_{y_{h-1}}^{y_h} \frac{1}{b-a}dy$$

The integral gives

$$W_h = \frac{d_h}{b-a}, \quad \text{where } d_h = y_h - y_{h-1} \tag{2.22}$$

Similarly, from (2.8) the **stratum mean** (μ_{hy}) is obtained as follows:

$$\mu_{hy} = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} yf(y)dy$$
$$= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y\cdot\frac{1}{b-a}dy$$
$$= \frac{1}{W_h}\cdot\frac{1}{b-a}\int_{y_{h-1}}^{y_h} ydy$$

Thus, substituting the value of W_h form (2.22), we get

$$\mu_{hy} = \frac{1}{d_h}\int_{y_{h-1}}^{y_h} ydy$$

Then, the integral reduces to

$$\mu_{hy} = \frac{1}{2d_h}\left(\left(y_h\right)^2 - \left(y_{h-1}\right)^2\right)$$
$$= \frac{1}{2d_h}\left[\left(y_h - y_{h-1}\right)\left(y_h + y_{h-1}\right)\right]$$

Finally, using the relation $d_h = y_h - y_{h-1}$, we get μ_{hy} as follows:

$$\mu_{hy} = \frac{d_h + 2y_{h-1}}{2}. \tag{2.23}$$

The **stratum variance** (S_{hy}^2) for Uniform Distribution is found as follows:

$$S_{hy}^2 = \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 f(y)dy - \mu_{hy}^2$$
$$= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 \cdot \frac{1}{b-a}dy - \mu_{hy}^2$$

Thus, taking the constant in front of the integral and using from (2.22) and (2.23), it reduces to

$$S_{hy}^2 = \frac{1}{y_h}\int_{y_{h-1}}^{y_h} y^2 dy - \left(\frac{y_h + 2y_{h-1}}{2}\right)^2$$

Integrating the function, it becomes

$$S_{hy}^2 = \frac{1}{y_h}\cdot\frac{\left(y_h\right)^3 - \left(y_{h-1}\right)^3}{3} - \left(\frac{d_h + 2y_{h-1}}{2}\right)^2$$

Using $y_h = d_h + y_{h-1}$, we get

$$S_{hy}^2 = \frac{1}{d_h}\cdot\frac{\left(d_h + y_{h-1}\right)^3 - \left(y_{h-1}\right)^3}{3} - \left(\frac{d_h + 2y_{h-1}}{2}\right)^2$$

Finally, by simplifying the function we get

$$S_{hy}^2 = \frac{d_h^2}{12} \tag{2.24}$$

Substituting (2.22) and (2.24), the MPP (2.10) is solved for constructing a predetermined number of strata, say $L = 4$, with a fixed total sample size $n = 100$ for Population 1 by executing a program coded in LINGO (see Appendix A) for determining the OSB and the optimum allocation. Also note that, for this population $a = y_0 = 0.000391$, $b = y_L = 0.998604$ and $d = y_L - y_0 = 0.998213$.

Then, the OSB $\left(y_h^*\right)$ and the optimum sample sizes $\left(n_h^*\right)$ using the proposed method as discussed in previous sections are obtained as shown in Table 2.1. The stratum width $\left(d_h^*\right)$, stratum weight $\left(W_h\right)$, stratum variance $\left(S_h^2\right)$ and the sampling variance, $\text{var}(\bar{y}_{st})$, are also presented in the table.

Table 2.1: Results for Uniform Distribution using Study Variable for *n* = 100

Optimum Strata Widths (OSW) d_h^*	**Optimum Strata Boundaries (OSB)** $y_h^* = y_{h-1}^* + d_h^*$	**Sample Size** n_h^*	**Stratum Weight** W_h	**Stratum Variance** S_h^2	$\text{var}(\bar{y}_{st})$
$d_1^* = 0.2521$ $d_2^* = 0.2483$ $d_3^* = 0.2483$ $d_4^* = 0.2495$	$y_1^* = 0.2525$ $y_2^* = 0.5008$ $y_3^* = 0.7491$ $y_4^* = 0.9986$	$n_1^* = 25$ $n_2^* = 25$ $n_3^* = 25$ $n_4^* = 25$	$W_1 = 0.2521$ $W_2 = 0.2487$ $W_3 = 0.2487$ $W_4 = 0.2499$	$S_1^2 = 0.0728$ $S_2^2 = 0.0717$ $S_3^2 = 0.0717$ $S_4^2 = 0.0720$	0.00004671

Suppose that the information on the study variable is not available but its auxiliary variable *x*. For a sample data, it has been seen that *y* has a linear regression model with *x*, that is:

$$\lambda(x) = \alpha + \beta x \qquad (2.25)$$

For this population, the auxiliary variable follows uniform distribution $f(y)$ of the form (2.21) with $a = x_0$ = 0.002877, $b = x_L$ = 1.999727 and $d = x_L - x_0$ = 1.998685. The estimated regression coefficients are found as $\hat{\alpha}$ = -0.026 and $\hat{\beta}$ = 0.505.

Now, using (2.17), (2.21) and (2.24) we obtain:

$$S_{h\lambda}^2 = \frac{\beta_h^2 d_h^2}{12}$$

Assuming that the regression model is common across the strata, that is, $\beta_h = \hat{\beta}$ = 0.505, the expected stratum variance of the error is obtained as:

$$\mu_{h\phi} = \text{MSE} = 0.000101,$$

where MSE is the mean square of residuals for the regression model. Then, solving the MPP (2.16) by executing a program coded in LINGO (see Appendix B), the OSB of the auxiliary variable $\left(x_h^*\right)$ and hence the OSB of study variable $\left(y_h^*\right)$ along with the optimum sample sizes $\left(n_h^*\right)$ and variance of the estimate of study variable are obtained as shown in Table 2.2.

Table 2.2: Results for Uniform Distribution using Auxiliary Variable for n = 100

Optimum Strata Widths (OSW) d_h^*	**OSB for x** $x_h^* = x_{h-1}^* + d_h^*$	**OSB for y** $y_h^* = \hat{\alpha}_h + \hat{\beta}_h x_h^*$	**Sample Size** n_h^*	$\text{var}(\bar{y}_{st})$
$d_1^* = 0.4752$	$x_1^* = 0.4781$	$y_1^* = 0.2154$	$n_1^* = 22$	0.00004868
$d_2^* = 0.5104$	$x_2^* = 0.9885$	$y_2^* = 0.4732$	$n_2^* = 27$	
$d_3^* = 0.5103$	$x_3^* = 1.4988$	$y_3^* = 0.7309$	$n_3^* = 26$	
$d_4^* = 0.5009$	$x_4^* = 1.9997$	$y_4^* = 0.9838$	$n_4^* = 25$	

2.4.2 Population 2: Right-Triangular Distribution

The right-triangular distribution is a family of continuous probability distribution, which models many observable phenomena that shows the number of successes when the most likely success falls at the maximum and the least likely success falls at the minimum values. For example; less income earned by a larger portion of families in a society, whereas a very few families earn large income.

The distribution is defined by two parameters a and b, which are its minimum and maximum values where respectively the most likely and the least likely number of items fall.

The general formula for the probability density function of a right-triangular distribution is given by

$$f(x) = \begin{cases} \dfrac{2(b-x)}{(b-a)^2}; & a \le x \le b \\ 0; & \text{otherwise.} \end{cases} \qquad (2.26)$$

2.4.2.1 Estimating the distribution of the population:

The data set of Population 2 of size $N = 800$ provides the information of the study variable y that follows a right-triangular distribution with the minimum value y_0 = 0.003716 and the maximum value y_L = 1.943307. Thus, the dataset gives the range of the distribution as $d = y_L - y_0 = 1.939591$.

2.4.2.2 Formulation of the Problem Determining OSB and Sample sizes as an MPP:

As the study variable y has right-triangular distribution with density function $f(y)$ given in (2.26), then from (2.6) the stratum weight (W_h) is obtained as

$$W_h = \int_{y_{h-1}}^{y_h} f(y)dy$$
$$= \int_{y_{h-1}}^{y_h} \frac{2(b-y)}{(b-a)^2}dy$$

Taking the constants out of the integral and applying simple integration yields

$$W_h = \frac{2}{(b-a)^2} \int_{y_{h-1}}^{y_h} (b-y)\,dy$$
$$= \frac{2}{(b-a)^2}\left[by - \frac{y^2}{2}\right]_{y_{h-1}}^{y_h}$$

Thus, simplifying the equation above, we get

$$W_h = \frac{2}{(b-a)^2}.\left[bd_h - \frac{d_h^2}{2} - d_h y_{h-1}\right]$$
$$= \frac{d_h}{(b-a)^2}.\left[(2b - d_h - 2y_{h-1})\right]$$

Finally, substituting $a_h = b - y_{h-1}$ into the equation yields

$$W_h = \frac{d_h\left(2a_h - d_h\right)}{\left(b-a\right)^2}, \quad \text{where } a_h = b - y_{h-1} = b - \left(a + \sum_{l=1}^{h-1} d_l\right) \tag{2.27}$$

From (2.8), the stratum mean (μ_{hy}) is obtained as follows:

$$\begin{aligned}\mu_{hy} &= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} yf\left(y\right)dy \\ &= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y\cdot\frac{2\left(b-y\right)}{\left(b-a\right)^2}dy\end{aligned}$$

Substituting the value of W_h from (2.27) and performing the simple integration, it gives us

$$\begin{aligned}\mu_{hy} &= \frac{\left(b-a\right)^2}{d_h\left(2a_h - d_h\right)}\cdot\frac{2}{\left(b-a\right)^2}\int_{y_{h-1}}^{y_h}\left(by - y^2\right)dy \\ &= \frac{2}{d_h\left(2a_h - d_h\right)}\left[\frac{by^2}{2} - \frac{y^3}{3}\right]_{y_{h-1}}^{y_h} \\ &= \frac{2}{d_h\left(2a_h - d_h\right)}\left[\frac{3by_h^2 - 2y_h^3 - 3by_{h-1}^2 + 2y_{h-1}^3}{6}\right]\end{aligned}$$

Substituting $y_h = d_h + y_{h-1}$ and expanding the equation, we get

$$\begin{aligned}\mu_{hy} &= \frac{1}{d_h\left(2a_h - d_h\right)}\left[\frac{3b\left(\left(d_h + y_{h-1}\right)^2 - y_{h-1}^2\right) - 2\left(\left(d_h + y_{h-1}\right)^3 - y_{h-1}^3\right)}{3}\right] \\ &= \frac{1}{d_h\left(2a_h - d_h\right)}\left[\frac{3b\left(d_h^2 + 2d_h y_{h-1}\right) - 2\left(d_h^3 + 3d_h^2 y_{h-1} + 3d_h y_{h-1}^2\right)}{3}\right]\end{aligned}$$

Finally, simplifying the expression, it yields

$$\mu_{hy} = \frac{3b\left(d_h + 2y_{h-1}\right) - 2\left(d_h^2 + 3d_h y_{h-1} + 3y_{h-1}^2\right)}{3\left(2a_h - d_h\right)} \tag{2.28}$$

Similarly, the stratum variance (S_{hy}^2) for Right-Triangular distribution is given by

$$\begin{aligned} S_{hy}^2 &= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 f\left(y\right) dy - \mu_{hy}^2 \\ &= \frac{1}{W_h}\int_{y_{h-1}}^{y_h} y^2 . \frac{2\left(b-y\right)}{\left(b-a\right)^2} dy - \left[\frac{3b\left(d_h + 2y_{h-1}\right) - 2\left(d_h^2 + 3d_h y_{h-1} + 3y_{h-1}^2\right)}{3\left(2a_h - d_h\right)}\right]^2 \end{aligned}$$

Substituting the value of W_h from (2.27) and simplifying the equation, it leaves us:

$$\begin{aligned} S_{hy}^2 &= \frac{\left(b-a\right)^2}{d_h\left(2a_h - d_h\right)} . \frac{2}{\left(b-a\right)^2}\int_{y_{h-1}}^{xy} y^2\left(b-y\right) dy - \left[\frac{3b\left(d_h + 2y_{h-1}\right) - 2\left(d_h^2 + 3d_h y_{h-1} + 3y_{h-1}^2\right)}{3\left(2a_h - d_h\right)}\right]^2 \\ &= \frac{2}{d_h\left(2a_h - d_h\right)}\int_{y_{h-1}}^{y_h}\left(y^2 b - y^3\right) dy - \left[\frac{3b\left(d_h + 2y_{h-1}\right) - 2\left(d_h^2 + 3d_h x_{h-1} + 3y_{h-1}^2\right)}{3\left(2a_h - d_h\right)}\right]^2 \end{aligned}$$

Performing simple integration gives

$$S_{hy}^2 = \frac{2}{d_h\left(2a_h - d_h\right)} . \left[\left(\frac{d_h^3 b}{3} - \frac{y_h^4}{4}\right) - \left(\frac{y_{h-1}^3 b}{3} - \frac{y_{h-1}^4}{4}\right)\right] - \left[\frac{3b\left(d_h + 2y_{h-1}\right) - 2\left(d_h^2 + 3d_h y_{h-1} + 3y_{h-1}^2\right)}{3\left(2a_h - d_h\right)}\right]^2$$

Substituting $y_h = d_h + y_{h-1}$, expanding and simplifying further yields

$$S_{hy}^2 = \frac{2}{d_h\left(2a_h - d_h\right)} \cdot \left[\left(\frac{4\left(d_h + y_{h-1}\right)^3 b - 3\left(d_h + y_{h-1}\right)^4}{12}\right) - \left(\frac{4y_{h-1}^3 b - 3y_{h-1}^4}{12}\right)\right] - \left[\frac{3b\left(l_h + 2x_{h-1}\right) - 2\left(l_h^2 + 3l_h x_{h-1} + 3x_{h-1}^2\right)}{3\left(2a_h - l_h\right)}\right]^2$$

Finally, we get

$$S_{yh}^2 = \frac{d_h^2\left(d_h^2 - 6a_h d_h + 6a_h^2\right)}{18\left(2a_h - d_h\right)^2} \tag{2.29}$$

Substituting (2.27) and (2.28), the MPP (2.10) is solved for constructing $L = 4$ with a fixed total sample size $n = 150$ using LINGO (see Appendix C).

Then, the OSB $\left(y_h^*\right)$ and the optimum sample sizes $\left(n_h^*\right)$ using the proposed method as discussed earlier are obtained as shown in Table 2.3. The stratum width $\left(d_h^*\right)$, stratum weight $\left(W_h\right)$, stratum variance $\left(S_h^2\right)$ and the variance of the estimate, $\text{var}(\bar{y}_{st})$, are also presented in the table.

Table 2.3: Results for Right-Triangular Distribution using Study Variable for *n* = 150

Optimum Strata Widths (OSW) d_h^*	**Optimum Strata Boundaries (OSB)** $y_h^* = y_{h-1}^* + d_h^*$	**Sample Size** n_h^*	**Stratum Weight** W_h	**Stratum Variance** S_h^2	$\text{var}(\bar{y}_{st})$
$d_1^* = 0.2807$ $d_2^* = 0.3237$ $d_3^* = 0.4028$ $d_4^* = 0.9324$	$y_1^* = 0.2844$ $y_2^* = 0.6081$ $y_3^* = 1.0109$ $y_4^* = 1.9433$	$n_1^* = 36$ $n_2^* = 36$ $n_3^* = 36$ $n_4^* = 42$	$W_1 = 0.2685$ $W_2 = 0.2576$ $W_3 = 0.2428$ $W_4 = 0.2311$	$S_1^2 = 0.1535$ $S_2^2 = 0.1614$ $S_3^2 = 0.1742$ $S_4^2 = 0.2122$	0.000171

When the information on the study variable is not available, an auxiliary variable *x* is considered, which is found to be linearly regressed with *y* for a sample data, that is:

$$\lambda(x) = \alpha + \beta x$$

The estimated regression coefficients are found as: $\hat{\alpha}$ = -0.023 and $\hat{\beta} = 0.675$.

Using (2.17), (2.25) and (2.24) we obtain:

$$S_{yh}^2 = \beta_h^2 \frac{d_h^2\left(d_h^2 - 6a_h d_h + 6a_h^2\right)}{18\left(2a_h - d_h\right)^2}$$

Assumed that the regression model is common across the strata, that is, $\beta_h = \hat{\beta} =$ 0.675.

The expected stratum variance of the error is obtained as:

$$\mu_{h\phi} = \text{MSE} = 0.000157,$$

where MSE is the mean square of residuals for the regression model. Then, solving the MPP (2.16) using LINGO (see Appendix D), the OSB of the auxiliary variable $\left(x_h^*\right)$ and hence the OSB of study variable $\left(y_h^*\right)$ along with the optimum sample sizes $\left(n_h^*\right)$ and variance of the estimate of study variable are obtained as shown in Table 2.4.

Table 2.4: Results for Right-Triangular Distribution using Auxiliary Variable for n = 150

Optimum Strata Widths (OSW) d_h^*	**OSB for x** $x_h^* = x_{h-1}^* + d_h^*$	**OSB for y** $y_h^* = \hat{\alpha}_h + \hat{\beta}_h x_h^*$	**Sample Size** n_h^*	$\text{var}(\bar{y}_{st})$
$d_1^* = 0.4278$	$x_1^* = 0.4323$	$y_1^* = 0.2688$	$n_1^* = 36$	0.00002773
$d_2^* = 0.4935$	$x_2^* = 0.9258$	$y_2^* = 0.6019$	$n_2^* = 36$	
$d_3^* = 0.6139$	$x_3^* = 1.5397$	$y_3^* = 1.0163$	$n_3^* = 36$	
$d_4^* = 1.4208$	$x_4^* = 2.9605$	$y_4^* = 1.9753$	$n_4^* = 42$	

2.5. Conclusion

The problem of determining OSB and the allocation of sample size are discussed by many authors mostly either separately or they determined the OSB under a particular allocation. The OSB so obtained may be infeasible or non-optimum, especially for small and skewed populations.

As seen in this chapter, an attempt has been made to propose a technique to solve both the problems simultaneously. The problems are formulated as an MPP, which has been solved by developing a program coded in a user friendly optimization software LINGO.

The advantage of the proposed technique is that it is a parametric base and has a wide scope of application since the complete dataset of the study variables is unknown, in

practice. The technique requires only the values of parameters of the stratification variables, which can easily be available from the past studies.

A numerical examples are presented to show the computational details and the applications of proposed method using LINGO.

Chapter 3

Determining Optimum Cluster Size and Sampling Unit for Multivariate Study Using Evolutionary Algorithm

3.1 Introduction

Cluster sampling is a sampling technique used frequently when the target population is spread across region and natural groupings are evident in the population. The technique is widely practiced in sample surveys. In this technique, the total population is divided into groups (or clusters) of smaller units of the population, called sampling units. Then, a simple random sample of the sampling units is selected. Cluster sampling helps in reducing the cost of surveys by not requiring the entire sampling frame, in other words, cluster sampling minimises the costs involved as well as saves time while carrying out surveys. The size of these clusters or sampling units influences the efficiency of sampling and thus the problem of determining the optimum cluster size and number of sampling units arises. It is, therefore, advisable to use that cluster size which is optimum with respect to a carefully chosen criterion.

It is worth bringing to the reader's attention that when a single characteristic is under study, many authors have attempted to determine the cluster size that gives maximum precision within the available resources. As per earlier discussion, Jessen (1942, 1978), Homeyer and Black (1946), Mahalanobis (1944), Sukhatme (1947, 1950), Cochran (1948), Sukhatme and Panse (1951), Hansen et al. (1953), Murthy (1967), Sheela and Unnithan (1992), etc. worked on determining the optimal cluster size. They proposed commonly used techniques such as substitution, trial and error and graphical methods. Later, Khan, Jahan and Ahsan (1997) formulated the problem as a mathematical programming problem (MPP) and developed a technique of determining the optimum cluster size and sampling unit in explicit forms.

However, when more than one characteristic are under study, the procedures for determining optimum cluster size are not well defined in sampling literature. The

traditional approach is to obtain optimal solution for each characteristic individually and then choose the final sampling design from among the individual solutions. In practice, it is not possible to use this approach of individual optimum solution because a solution, which is optimum for one characteristic, may not be optimum for other characteristics. Moreover, in the absence of a strong positive correlation between the characteristics under study, the individual optimum allocations may differ a lot and there may be no obvious compromise. In such situations some criterion is needed to work out an acceptable sampling design which is optimum, in some sense, for all characteristics (Cochran (1977), Khan et al. (1997, 2003)). Further, for practical application of a sampling design, the integer values of the cluster size and sampling unit are required. They could be obtained by simply rounding off non-integer solutions to their nearest integer values. However, in many situations the rounded-off solution may be infeasible or non-optimal. There may also be other restrictions that must be met in order to satisfy constraints on available budget, and minimum and maximum size of cluster and sampling unit.

In this chapter, the problem of determining optimum compromise cluster size and sampling unit for multivariate cluster sampling of equal size is deliberated and formulated as All Integer Nonlinear Programming Problems (AINLPP), which is discussed in Section 3.2. A solution procedure is discussed in Section 3.3. Given that the functions involved in the problem are non-smooth, the formulated AINLPP is solved using an Evolutionary algorithm implemented in Evolutionary Solver in Excel. The AINLPP has a convex objective function for a specified range of measure of homogeneity and a constraint function which is indefinite, that is, neither convex nor concave, which restrict the use of suitable nonlinear programming technique. The problem has also bounded variable restrictions on the decision variables. Section 3.4 presents a numerical example to illustrate the practical application of the solution procedure.

The studies carried out in this chapter have been presented by the author at the IEEE 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE 2015) conference held in Fiji during December 2-4, 2015. The paper has also

been published in *IEEE Proceedings of 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering.*

3.2 The Problem in Multivariate Cluster Sampling Design

In a multivariate cluster sampling, where p characteristics are under study, let n denotes a simple random sample of clusters drawn from a population of N clusters of equal size M. Let y_{ijk}, $\bar{y}_k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ijk}$, and $\bar{\bar{y}}_k = \frac{1}{nM}\sum_{i=1}^{n}\sum_{j=1}^{M} y_{ijk}$ denote, respectively, the value obtained from j th unit in the i th cluster, the sample mean per cluster, and the sample mean per unit for k th characteristic. It could be shown that $\bar{\bar{y}}_k$ is an unbiased estimate of the population mean $\bar{\bar{Y}}_k = \frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ijk}$ of k th characteristic with variance for a large population (see Cochran, 1977) given by

$$V(\bar{\bar{y}}_k) = \frac{S_k^2}{nM}\left[1+(M-1)\rho_k\right], \qquad \left(k = 1,2,...,p\right),$$

(3.1)

where

$$\rho_k = \frac{\sum_{i=1}^{N}\sum_{j\neq j'=1}^{M}\left(y_{ijk}-\bar{\bar{Y}}_k\right)\left(y_{ij'k}-\bar{\bar{Y}}_k\right)}{NM(M-1)S_k^2}$$

is the intra-cluster correlation coefficient, that is, measure of homogeneity, for k th characteristic and

$$S_k^2 = \frac{1}{NM-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{ijk}-\bar{\bar{Y}}_k\right)^2$$

is the total population variance among unit for k th characteristic. From (3.1), it is obvious that the value of ρ_k must satisfy $\rho_k \geq -1/(M-1)$, otherwise the RHS of (3.1) becomes negative.

Hansen et al. (1953) expressed ρ_k is a function of the cluster size M as:

$$\rho_k = aM^b$$

Where the constants *a* and *b* are to be estimated by using the above relation in ρ_k and M. Usually, intra-cluster correlation coefficient ρ_k is positive and decreases as M increases but the rate of decrease is small relative to the rate of increase in M. Jessen (1978) showed that ρ_k do not vary much with the small changes in M. This suggests that over a short but relevant range of M, ρ_k can be treated as a constant.

Ignoring the over-head cost of planning and analysis, the total cost function C of survey may be given as:

$$C = c_1 nM + c_2 d, \qquad (3.2)$$

where, $c_1 = \sum_{k=1}^{p} c_{1k}$ denotes the cost of enumeration per unit including the travel cost within clusters, and c_2 denotes the cost of travelling unit distance between clusters, *d* denotes the total distance between *n* randomly selected clusters. Also c_{1k} are the per unit costs of measuring the k th characteristic of a cluster.

From an emperical study, Mahalanobis (1940) showed that the total distance '*d*' between *n* points located at random is proportional to $n^{1/2} - n^{-1/2}$. Jessen (1942) also showed that $n^{1/2}$ is a fairly good approximation to '*d*' for practical purposes. Thus, using the Jessen's approximation the total cost of the survey may be expressed as

$$C = c_1 nM + c_2 \sqrt{n}, \qquad (3.3)$$

The optimum choice of n and M for an individual characteristic can thus be determined by minimizing the variance in (3.1) for the given cost in (3.2), or by minimizing the cost for fixed variance.

In multivariate cluster sample surveys usually a compromise criterion is needed to work out an acceptable choice of the number of clusters and sampling units which is optimum, in some sense, for all characteristics. However, if the total cost of the survey is predetermined, using the compromise criterion suggested by Khan, et al. (2003), an optimal choice may be one that minimizes the weighted sum of the sampling variances of the estimates of various characteristics within the available budget. It is, therefore, in a cluster sampling, if the population means of p characteristics are of interest, it may be a reasonable criterion for determining the optimal choice of n and M is to minimize a weighted sum of the variances of the cluster sample means of all the p characteristics, that is,

$$\sum_{k=1}^{p} a_k V\left(\bar{\bar{y}}_k\right), \tag{3.4}$$

where a_k is the weights assigned to the k th characteristic in proportion to its importance as compared to other characteristics and $V(\bar{\bar{y}}_k)$ as given in (3.1).

Thus, from (3.1), minimizing (3.4) is equivalent to minimize

$$V(n,M) = \frac{a_k S_k^2}{nM} + \frac{a_k S_k^2 (M-1)\rho_k}{nM}. \tag{3.5}$$

For a fixed budget C_0 given by (3.3) the optimum values of n and M are those which minimize (3.5) within the specified budget. Then the problem of determining n and M may be expressed as the following AINLPP:

$$\left.\begin{array}{ll} \text{Minimize} & V(n,M)=\sum_{k=1}^{p}\frac{a_k S_k^2}{nM}+\frac{a_k S_k^2(M-1)\rho_k}{nM} \\ \text{subject to} & c_1 nM + c_2\sqrt{n} \le C_0, \\ & 1 \le n \le N, 1 \le M \le NM, \\ \text{and} & n \text{ and } M \text{ are integers} \end{array}\right\} \quad (3.6)$$

The bounded variable restrictions $1 \le n \le N$ and $1 \le M \le NM$ are included to ensure the minimum and maximum possible selection of n and M in a cluster sampling design. Further, we need n and M to be integers in order to implement the solution practically.

The values of S_k^2, ρ_k, c_1 and c_2, if not known, may be estimated on some previous census or by conducting a pilot survey.

3.3 Solution Using Evolutionary Method

It has been seen that the function (3.1) is a convex function of n and M for $\rho_k \ge -3/(4M-3)$ and hence the objective function $V(n,M)$ in (3.6) as a sum of convex function is also a convex. However, the constraint $C(n,M)=c_1 nM + c_2\sqrt{n}$ is indefinite, that is, neither convex nor concave, which restrict the use of suitable nonlinear programming technique to solve AINLPP (3.6). However, in such situations, when all the functions involved are not smooth, one may use an evolutionary algorithm (see Back 1996).

Thus, in this chapter, we propose an evolutionary algorithm to solve the AINLPP (3.6). There are some software programs to use evolutionary algorithm. We use the algorithm implemented in Evolutionary Solver in Excel. For details of evolutionary algorithm one may refer to Harmon (2011).

3.4 Numerical Example

In order to demonstrate the computational details of the proposed technique, the following exercise from Singh (2003) has been worked out, in which the problem is to select n clusters of M states in USA from N =10 clusters. The data for the amount of agricultural loan in two types of farm (Y_1 = real estate farm loan, and Y_2 = nonreal estate farm loan) outstanding in 50 states in 1997 are given. The total population variances for both the variables are given as: $S_1^2 = 342021.5$ and $S_2^2 = 1176526.0$.

Dividing the states into 10 clusters for estimating Y_1, Singh (2003) obtain the value of intra-cluster correlation coefficient $\rho_1 = 0.349809$. In addition to this information, we assume that the intra-cluster correlation coefficient for Y_2 is $\rho_2 = 0.2734$.

Suppose one is interested to use multivariate cluster sampling to estimate the mean of both Y_1 and Y_2. We assume that $c_1 = \$15,\ c_2 = \$725$ and the available budget for the survey is $C_0 = \$2000$.

If $a_1 = a_2 = 1$, then the problem of determining optimum sampling unit (n) and cluster size (M) in AINLPP (3.6) can be expressed as:

$$\left.\begin{array}{ll} \text{Minimize} & V(n,M) = \dfrac{342021.5}{nM}\left[1+0.349809(M-1)\right] + \dfrac{1176526.0}{nM}\left[1+0.2734(M-1)\right] \\ \text{subject to} & 15nM + 725\sqrt{n} \le 2000, \\ & 1 \le n \le 10, 1 \le M \le 50, \\ \text{and} & n \text{ and } M \text{ are integers} \end{array}\right\} \tag{3.7}$$

Using an evolutionary algorithm implemented in Evolutionary Solver in Excel, the problem (3.7) is solved. The compromise optimum sampling unit and cluster size are obtained as:

$n = 5$ and $M = 5$ with minimum $V(n, M) = 131350.61$.

3.5 Conclusion

In a survey when more than one characteristic are under study and when a cluster sampling design is to be used, the best sample size for one characteristic will not be best for another and therefore, some compromise must be reached. In this chapter, we have discussed a problem of determining compromise optimum cluster size and sampling unit of equal size in multivariate cluster sampling. The problem was formulated as AINLPP, which was then solved using an evolutionary algorithm.

The main aim of this study was to determine optimum compromise sample size when more than one characteristic are of interest in a cluster sampling design, which is a widely used technique in surveys. We proposed a solution procedure using evolutionary algorithm, which may be useful for the selection of sample size in multivariate cluster sampling. Based on our results, it can be concluded that evolutionary algorithm should be considered as method of determining optimum cluster size and sampling unit in multivariate cluster sampling design.

Chapter 4

Conclusion and Future Research

Applied mathematics is the 21st Century's momentum. Applied mathematical knowledge combined with technology has seen tremendous advances in globalization, medicine, education, military, science, economics and industries. This digital revolution implicates that technology is the norm of the millennium and whether we like it or not, computers and digital devices are here to stay; hence the next best thing for us to do is to use it use wisely and meaningfully, which has resulted in this thesis where we make use of mathematical programming to come up with best solutions, especially in the area of sampling.

In Chapter 2, we endeavoured to determine OSB and sample allocation simultaneously when the estimate population mean of the study variable y is of interest and a frequency distribution $f(y)$ or the frequency distribution $f(x)$ of its auxiliary variable x was known. In Chapter 3, we discussed the problem of determining optimum compromise of cluster size and sample unit for multivariate cluster sampling. The problems addressed in both chapters have been formulated as Mathematical Programming Problems and their solutions are discussed.

This research confirms that use of mathematical programming techniques can help us achieve more precise and accurate results in determining samples, within the given constraints.

We proposed a solution procedure to determine OSB and sample allocation simultaneously using LINGO. Whereas, for the selection of sample size in multivariate cluster sampling we propose a solution procedure using evolutionary algorithm. We find that the evolutionary algorithm is very useful and may be considered as method of determining optimum cluster size and sampling unit in multivariate cluster sampling design.

Future research that maybe carried out in some other problems in the area of sampling that can formulated as mathematical programming problems and may be attempted to solve by using suitability chosen or especially developed techniques.

(i) To develop a method of determining optimum strata boundaries and sample allocation when a single study variable is of interest in the case where a multivariable auxiliary information is available.
(ii) To develop a method of multivariate stratification, that is, the problem of determining optimum strata boundaries and sample allocation when more than one study variable is of interest.
(iii) The problem of optimum allocation in multivariate stratified sampling when double sampling is used for stratification.

Bibliography

[1] Aoyama, H. (1954). A Study of Stratified Random Sampling. *Ann. Inst. Stat. Math.*, 6, 1-36.

[2] Back, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York,

[3] Baillargeon, S. and Rivest, L. P. (2009). A general algorithm for univariate stratification. *International Statistical Review*, 77(3):331–344.

[4] Brito, J., Ochi, L. S., Montenegro, F., and Maculan, N. (2010). An ILS approach applied to the optimal stratification problem. *International Transaction in Operational Research*, 17(6), 753–764.

[5] Bühler, W. and Deutler, T. (1975). Optimum Stratification and Grouping by Dynamic Programming. *Metrika*, Vol. 22, pp. 161-175.

[6] Cochran, W. G. (1948). *Notes on Sampling Survey Techniques (Mimeographed).* Raleigh, North Carolina: Istitute of Statistics.

[7] Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bull. Int. Stat. Inst.*, 38(2):345–358.

[8] Cochran, W. G. (1963). *Sampling Techniques* (2nd ed.). New York: John Wiley & Sons.

[9] Cochran, W. G. (1977). *Sampling Techniques* (3rd ed.). New York: John Wiley & Sons.

[10] Dalenius, T. (1950). The problem of Optimum Stratification-II. *Skand. Aktuartidskr*, 33, 203-213.

[11] Dalenius, T. (1957). *Sampling in Sweden*. Almqvist & Wiksell, Stockholm.

[12] Dalenius, T. and Gurney, M. (1951). *The problem of optimum stratification-II,* Skand.Akt., 34, 133-148.

[13] Dalenius, T. and Hodges, J. L. (1959): Minimum variance stratification. *J. Amer. Statist. Assoc.* 54, 88-101.

[14] Ekman, G. (1959). Approximate expression for conditional mean and variance over small intervals of a continuous distribution. *Ann. Inst. Stat. Math.*, 30, 1131-1134.

[15] Gunning, P. and Horgan, J. M. (2004). A New Algorithm for the Construction of Stratum Boundaries in Skewed Populations. *Survey Methodology*, 30(2), 159-166.

[16] Gupta, R. K., Singh, R. and Mahajan, P. K. (2005). Approximate Opimumum Strata Boundaries for Ratio and Regression Estimators. *Aligarh Journal of Statistics*, 25, 49-55.

[17] Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953): *Sample Survey Methods and Theory.* Vol. I & II, John Wiley and Sons, Inc., New York.

[18] Harmon, M. (2011). Step-By-Step Optimization with Excel Solver. (http:// excelmasterseries.com/ClickBank/New_Manuals/Optimization_With_Excel_Solver.php).

[19] Hedlin, D. (2000). A procedure for stratification by an extended Ekman rule. *Journal of Official Statistics*, 16(1), 15-29

[20] Hess, I., Sethi, V., and Balakrishnan, T. (1966). Stratification: A practical investigation. *Journal of the American Statistical Association*, 61(313):74–90.

[21] Hidiroglou, M.A. and Srinath, K.P. (1993). Problems associated with designing subannual business surveys. *Journal of Business & Economics Statistics*, 11, 397-405.

[22] Homeyer, P. G and Black, C. A. (1946). Sampling replicated field experiments on oats for yield determinations. *Proc. Soil Sci. Soc. America*. 11, 341-344.

[23] Horgan, J. M. (2006). Stratification of skewed populations: A review. *International Statistical Review*, 74(1):67–76.

[24] Horgan, J. M. (2011). Geometric stratification revisited. *In Proceedings of 58th World Statistical Congress*, Dublin, pages 3319–3328.

[25] Jessen, R. (1942). *Statistical Ivestigation of a sample survey for obtaining facts.* Agricultural Experiment Station Research. Lowa: lowa Agricultural Experiment Station Research.

[26] Jessen, R. J. (1978). *Statistical Survey Techniques.* New York: John Wiley and Sons Inc.

[27] Keskintürk, T. and Er, S. (2007). A genetic algorithm approach to determine stratum boundaries and sample sizes of each stratum in stratified sampling. *Computational Statistics & Data Analysis*, 52(1):53–67.

[28] Khan, E. A., Khan, M. G. M. and Ahsan, M. J. (2002). *Optimum Stratification: A Mathematical Programming Approach,* Calcutta Statistical Association Bulletin, 52 (special Volume), 323-333.

[29] Khan, M. G. M., Ahmad, N. and Khan, Sabiha (2009). Determining the Optimum Stratum Boundaries using Mathematical Programming. *Journal of Mathematical Modelling and Algorithms*", Springer, Netherland, DOI 10.1007/s10852-009-9115-3, 8(4), 409-423.

[30] Khan, M.G.M., Ahsan, M.J. and Jahan, N. (1997). Compromise allocation in multivariate stratified sampling: an integer solution. *Naval Research Logistics.* 44**,** 69-79.

[31] Khan M.G.M., Jahan N. and Ahsan M. J. (1997). Determining the optimum Cluster Size. *Jour. Ind. Soc. Ag. Stat*. 50 (2), 121-129.

[32] Khan, M. G. M., Khan, E. A. and Ahsan, M. J. (2003). An optimal multivariate stratified sampling design using dynamic programming. *Australian & New Zealand J. Statist*. 45 (1):107-113.

[33] Khan, M. G. M., Najmussehar and Ahsan, M. J. (2005). Optimum Stratification for Exponential Study Variable under Neyman Allocation. *Journal of Indian Society of Agricultural Statistics*, 59(2), 146-150.

[34] Khan, M. G. M., Nand, N. and Ahmad, N. (2008). Determining the Optimum Strata Boundary Points Using Dynamic Programming. *Survey Methodology*, 34(2), 205-214.

[35] Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, Vol. 6, No. 5, pp. 797-806.

[36] Kozak, M. and Verma, M. R. (2006). Geometric versus optimization approach to stratification: A comparison of efficiency. *Survey Methodology*, 32(2):157-163.

[37] Kozak, M., Verma, M. R., and Zielinski, A. (2007). Modern approach to optimum stratification: Review and perspectives. *Statistics in Transition*, 8(2):223–250.

[38] Lavallée, P. (1987). Some contributions to optimal stratification. *Master's thesis*, Carleton University, Ottawa, Canada.

[39] Lavallée, P. (1988). Two-way optimal stratification using dynamic programming. In *Proceedings of the Survey Research Methods-Section*, American Statistical Association, pages 646–651.

[40] Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, 14, 33-43.

[41] Lednicki, B., Wieczorkowski, R. (2003). Optimal Stratification and Sample Allocation between Subpopulations and Strata. *Statistics in Transition*, 6, 287-306.

[42] Mahalanobis, P.C. (1940). A sample of the acreage under jute in Bengal. *Sankhya*, 4, 511-530.

[43] Mahalanobis, P.C. (1942). *General report on the sample census of area under jute in Bengal.* Indian Central Jute Committee.

[44] Mahalanobis, P.C. (1944). On large-scale sample surveys. *Phil. Transac. Roy. Soc.*, *London* B, 231, 324-351.

[45] Mahalanobis, P. C. (1952). Some Aspects of the Design of Sample Surveys. *Sankhya*, 12, 1-7.

[46] Mehta, S. K., Singh, R. and Kishore, L. (1996). On Optimum Stratification for Allocation Proportional to Strata Totals. *Journal of Indian Statistical Association*, 34, 9-19.

[47] Murthy, M.N. (1967). *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta.

[48] Nand, N. and Khan, M. G. M. (2009). Optimum stratification for cauchy and power type study variable. *Journal of Applied Statistical Science*, 16(4):453.

[49] Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. The computer journal, 7(4):308–313.

[50] Neyman, J. (1934). On the Two Different Aspects of the Representatives Methods: *the Method Stratified Sampling and the Method of Purposive Selection.* J. Roy. Stat. Soc. 97, 558-606.

[51] Nicolini, G. (2001). A method to define strata boundaries. Departmental Working Paper. Department of Economics University of Milan, Italy. www.stat.fi/isi99/proceedings/arkisto/varasto/nico0784.pdf.

[52] Niemiro, W. (1999). Optimal stratification using random search method. *Wiadomosci Statystyczne*, 10:1–9.

[53] Rivest, L.P. (2002). A Generalization of Lavallée and Hidiroglou algorithm for stratification in business survey. *Survey Methodology*, 28, 191-198.

[54] Rizvi, S. E. H., Gupta, J. P. and Bhargava, M. (2002). Optimum Stratification based on Auxiliary Variable for Compromise Allocation. Metron, 28(1), 201-215.

[55] Scheaffer, R.L., Mendenhall III, W., Ott, R.L. and Gerow, K.G. (2012). *Elementary Survey Sampling* (7th ed.). Brook/Cole Cengage Learning, Boston, USA.

[56] Serfling, R. J. (1968). Approximately optimal stratification. *Journal of the American Statistical Association*, 63(324), 1298–1309.

[57] Sethi, V. K. (1963). A Note on Optimum Stratification of Population for Estimating the Population Mean. *Aust. J. Statist.,* 5, 20-33.

[58] Sheela, M. A. and Unnithan, V. K. G. (1992). Optimum size of plots in multivariate case. *J. Indian Soc. Agric. Statist.* 44 (3), 236-240.

[59] Singh, D. (1956). On efficiency of cluster sampling. J. Indian Soc. Agric. Statist., 8, 45-55.

[60] Singh, R. and Parkash, D. (1975). Optimum Stratification for Equal Allocation. *Annals of the Institute of Statistical Mathematics*, 27, 273-280.

[61] Singh, R. (1971). Approximately Optimum Stratification on the Auxiliary Variable. J. Amer. Stat. Assc., 66, 829-833.

[62] Singh, R. (1975). An Alternate Method of Stratification on the Auxiliary Variable. Sankhya. C, 37, 100-108.

[63] Singh, R. and Sukhatme, B. V. (1969). Optimum Stratification for Equal Allocation. Ann. Inst. Stat. Math., 27, 273-280.

[64] Singh, R. and Sukhatme, B. V. (1972). Optimum Stratification in Sampling with Varying Probabilities. Ann. Inst. Stat. Math., 24, 485-494.

[65] Singh, R. and Sukhatme, B. V. (1973). Optimum Stratification with Ratio and Regression Methods of Estimation. *Annals of the Institute of Statistical Mathematics*, 25, 627-633.

[66] Singh, S. (2003). *Advanced Sampling theory with Applications* (Vol. 2). Dordrecht: Kluwer Academic.

[67] Sukhatme, P. V. (1947). The problem of plot size in large-scale yield surveys. *J. Amer. Statist. Assoc.* 42, 297-310.

[68] Sukhatme, P. V. (1950). *Sample surveys in agriculture*. Presidential Address to the Section of Statistics, 37th Session, Indian Science Congress, Poona.

[69] Sukhatme, P.V. and Panse, V.G. (1951). Crop survey in India. *Jour. Ind. Soc. Ag. Stat.* 3, 97-168.

[70] Sweet, E.M., and Sigman, R.S. (1995). Evaluation of model-assisted procedures for stratifying skewed populations using auxiliary data, *Proceedings of the Survey Research Methods Section,* ASA, pp. 491-496.

[71] Taga, Y. (1967). On Optimum Stratification for the Objective Variable Based on Concomitant Variables using Prior Information. *Annals of the Institute of Statistical Mathematics*, 19, 101-129.

[72] Unnithan, V.K.G. (1978). The Minimum Variance Boundary Points of Stratification. *Sankhya*, 40(C), 60-72.

[73] Unnithan, V. K. G. and Nair, N. U. (1995). Minimum variance stratification. *Communications in Statistics-Simulation and Computation*, 24(1):275–284.

[74] Wackerly, D.W., Mendenhall, W. and Scheaffer, R. (2008). *Mathematical Statistics with Applications* (8th Edition), Thomson Learning, Inc., USA.

[75] Yong, F.H., Tian, L., Yu, S., Cai, T. and Wei, L.J. (2016). Optimal stratification in outcome prediction using baseline information. *Biometrika*, 103, 4, 817–828.

Appendix A:

LINGO code for determining the OSB and Sample Sizes for Uniform Distribution using Study Variable (y)

```
!Population size;
N=1000;
!Sample size;
n=100;
!Smalest value of Y;
a=0.000391;
!Largest value of Y;
b=0.998604;
! Range of Y;
d=b-a;
!Weights;
w1=d1/(b-a);
w2=d2/(b-a);
w3=d3/(b-a);
w4=d4/(b-a);
!Variance of Y variable;
s1^2=d1^2/12;
s2^2=d2^2/12;
s3^2=d3^2/12;
s4^2=d4^2/12;
!Objective;
min=(w1^2*s1^2/n1+w2^2*s2^2/n2+w3^2*s3^2/n3++w4^2*s4^2/n4)-
(w1*s1^2+w2*s2^2+w3*s3^2+w4*s4^2)/N;
!Constraints;
d1+d2+d3+d4=d;
n1+n2+n3+n4=n;
n1>=1;
n2>=1;
n3>=1;
n4>=1;
n1<=w1*N;
n2<=w2*N;
n3<=w3*N;
n4<=w4*N;
@gin(n1);
@gin(n2);
@gin(n3);
@gin(n4);
!Stratum boundaries;
y1=a+d1;
y2=y1+d2;
y3=y2+d3;
y4=y3+d4;
```

Appendix B:

LINGO code for determining the OSB and Sample Sizes for Uniform Distribution using Auxiliary Variable (x)

```
!Population size;
N=1000;
!Sample size;
n=100;
!Smalest value of X;
a=0.00287723;
!Largest value of X;
b=1.999727;
! Range of X;
d=b-a;
!c is regression coefficient (Beta) and m is MSE
c=0.505;
m=0.000101;
!Weights;
w1=d1/(b-a);
w2=d2/(b-a);
w3=d3/(b-a);
w4=d4/(b-a);
!Variance of Y variable;
s1^2=c^2*d1^2/12;
s2^2=c^2*d2^2/12;
s3^2=c^2*d3^2/12;
s4^2=c^2*d4^2/12;
!Objective;
min=(w1^2*(s1^2+m)/n1+w2^2*(s2^2+m)/n2+w3^2*(s3^2+m)/n3++w4^2*(s4^2+m)/n4)-
(w1*(s1^2+m)+w2*(s2^2+m)+w3*(s3^2+m)+w4*(s4^2+m))/N;
!Constraints;
d1+d2+d3+d4=d;
n1+n2+n3+n4=n;
n1>=1;
n2>=1;
n3>=1;
n4>=1;
n1<=w1*N;
n2<=w2*N;
n3<=w3*N;
n4<=w4*N;
@gin(n1);
@gin(n2);
@gin(n3);
@gin(n4);
!Stratum boundaries;
x1=a+d1;
x2=x1+d2;
x3=x2+d3;
x4=x3+d4;
```

Appendix C:

LINGO code for determining the OSB and Sample Sizes for Right-Triangular Distribution using Study Variable (y)

```
!Population size;
N=1000;
!Sample size;
n=150;
!Smalest value of Y;
a=0.000391;
a=0.003716;
!Largest value of Y;
b=1.943307;
! Range of Y;
d=b-a;
!Weights;
w1=d1*(2*a1-d1)/(b-a)^2;
w2=d2*(2*a2-d2)/(b-a)^2;
w3=d3*(2*a3-d3)/(b-a)^2;
w4=d4*(2*a4-d4)/(b-a)^2;
!Variance of Y variable;
a1=b-a;
a2=b-(a+d1);
a3=b-(a+d1+d2);
a4=b-(a+d1+d2+d3);
s1^2=d1^2*(d1^2-6*a1*d1+6*a1^2)/(18*(2*a1-d1));
s2^2=d2^2*(d2^2-6*a2*d2+6*a2^2)/(18*(2*a2-d2));
s3^2=d3^2*(d3^2-6*a3*d3+6*a3^2)/(18*(2*a3-d3));
s4^2=d4^2*(d4^2-6*a4*d4+6*a4^2)/(18*(2*a4-d4));
!Objective;
min=(w1^2*s1^2/n1+w2^2*s2^2/n2+w3^2*s3^2/n3++w4^2*s4^2/n4)-
(w1*s1^2+w2*s2^2+w3*s3^2+w4*s4^2)/N;
!Constraints;
d1+d2+d3+d4=d;
n1+n2+n3+n4=n;
n1>=1;
n2>=1;
n3>=1;
n4>=1;
n1<=w1*N;
n2<=w2*N;
n3<=w3*N;
n4<=w4*N;
@gin(n1);
@gin(n2);
@gin(n3);
@gin(n4);
!Stratum boundaries;
y1=a+d1;
y2=y1+d2;
y3=y2+d3;
y4=y3+d4;
```

Appendix D:

LINGO code for determining the OSB and Sample Sizes for Right-Triangular Distribution using Auxiliary Variable (x)

```
!Population size;
N=1000;
!Sample size;
n=150;
!Smalest value of X;
a=0.00454;
!Largest value of X;
b=2.960607;
! Range of X;
d=b-a;
!c is regression coefficient (Beta) and m is MSE
c=0.675;
m=0.000157;
!Weights;
w1=d1*(2*a1-d1)/(b-a)^2;
w2=d2*(2*a2-d2)/(b-a)^2;
w3=d3*(2*a3-d3)/(b-a)^2;
w4=d4*(2*a4-d4)/(b-a)^2;
!Variance of Y variable;
a1=b-a;
a2=b-(a+d1);
a3=b-(a+d1+d2);
a4=b-(a+d1+d2+d3);
s1^2=c^2*d1^2*(d1^2-6*a1*d1+6*a1^2)/(18*(2*a1-d1));
s2^2=c^2*d2^2*(d2^2-6*a2*d2+6*a2^2)/(18*(2*a2-d2));
s3^2=c^2*d3^2*(d3^2-6*a3*d3+6*a3^2)/(18*(2*a3-d3));
s4^2=c^2*d4^2*(d4^2-6*a4*d4+6*a4^2)/(18*(2*a4-d4));
!Objective;
min=(w1^2*(s1^2+m)/n1+w2^2*(s2^2+m)/n2+w3^2*(s3^2+m)/n3++w4^2*(s4^2+m)/n4)-
(w1*(s1^2+m)+w2*(s2^2+m)+w3*(s3^2+m)+w4*(s4^2+m))/N;
!Constraints;
d1+d2+d3+d4=d;
n1+n2+n3+n4=n;
n1>=1;
n2>=1;
n3>=1;
n4>=1;
n1<=w1*N;
n2<=w2*N;
n3<=w3*N;
n4<=w4*N;
@gin(n1);
@gin(n2);
@gin(n3);
@gin(n4);
!Stratum boundaries;
x1=a+d1;
x2=x1+d2;
x3=x2+d3;
x4=x3+d4;
```

Publications:

1. Khan, M.G.M. and Sharma, Sushita (2012). On constructing optimum strata and determining optimum allocation. Presented at the VII International Symposium on Optimization and Statistics & III National Conference on Statistical Inference, Sampling Techniques and Related Areas, held at Aligarh Muslim University, India during Dec 21-23, 2012.

2. Khan, M.G.M. and Sharma, Sushita (2015). "Determining Optimum Strata Boundaries and Optimum Allocation in Stratified Sampling", *Aligarh Journal of Statistics*, Vol. 35, 23-40.

3. Sharma, Sushita and Khan, M.G.M. (2015). Determining optimum cluster size and sampling unit for multivariate study, *IEEE Proceedings of 2015 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE),* 1-4, DOI: 10.1109/APWCCSE.2015.7476238.