THE UNIVERSITY OF THE SOUTH PACIFIC
LIBRARY

Author Statement of Accessibility

Name of Candidate Gurmindar Kaur Singh

Degree MSc

Department/School School of Computing, Information and Mathematical S

Institution/University University of the South Pacific

Thesis Title Calibration Approach for Stratified Sampling Design

Date of completion of requirements for award December, 2015

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.
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 Kaur

Date 01/03/16

Contact Address
Mobile · 9918078

Permanent Address
Catalina Street
Suva Point

CALIBRATION APPROACH FOR STRATIFIED SAMPLING DESIGN

by

Gurmindar Kaur Singh

A thesis submitted in fulfillment of the

requirements for the degree of

Master of Science

School of Computing, Information and

Mathematical Sciences

Faculty of Science, Technology and Environment

The University of the South Pacific

December, 2015

Declaration of Originality

I certify to the best of my knowledge, that this thesis is my original work and has been organized and submitted to the utmost best of my perception and expertise. I have faithfully acknowledged all the citations and references and the results presented have not been previously submitted for a University degree either in whole or in part elsewhere.

………………..
Gurmindar Kaur Singh
November, 2015

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

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

Acknowledgements

I would never have been able to finish my dissertation without the guidance of my supervisor, help from friends and support from my family and husband.

Foremost, I would like to express my sincere gratitude to my supervisor, Dr. M.G.M. Khan, for the continuous support of my Masters study for his patience, motivation, enthusiasm, immense knowledge, time and help in every phase of this meritorious task. I could not have imagined a better advisor and mentor for my Masters study.

My sincere thanks also goes to Mr Dinesh Rao, Assistant Lecturer of the School of Computing, Information and Mathematical Sciences, who always took out his valuable time for me whenever I was stuck in my thesis. He's comprehensive knowledge and expertise on the topic "Calibration" is worth appreciating and acknowledging. His skills in this specialized area allowed me to grasp more deeper understanding of the topic thus I was able to continue with my paper work confidently with his passion and support throughout my research. This attitude of him gave me positive energy to complete my thesis. He is a very good mentor for MSC students.

I would also like to express my deepest gratitude to the University of the South Pacific, Suva, Fiji for providing me with an excellent atmosphere for doing research.

My thesis would not have successfully completed without the unceasing support and attention of my parents, Mr and Mrs Jasvir Singh, my brother, Davindar Singh and my sister, Tasvindar

Kaur. I am also undauntedly thankful to my husband, Mr Navin Raaj who had the faith in me throughout this venture.

Finally, I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lent their hand in completing this thesis.

Table of Contents

Declaration of Originality i
Acknowledgements ii
Abstract 1
Preface 2
Chapter 1 6
Introduction 6
1.1 Sample Surveys 6
1.2 Stratified Random Sampling 7
1.3 Stratified Estimators 9
1.4 Allocation of Sample Sizes to Different Strata 11
1.4.1 Proportional Allocation 11
1.4.2 Neyman Allocation 11
1.4.3 Optimum Allocation 12
1.5 Study Variable and Auxiliary Information 12
1.6 Calibration Estimation 13
1.7 Distance Function 15
1.8 Nonlinear Programming Problem 16
1.9 Lagrange Multiplier Technique 17
1.10 Calibration Approach: Review of Literature and Studies 19
Chapter 2 23
A Calibration Estimator of Population Mean in Stratified Random Sampling using a Single Auxiliary Variable 23
2.1 Introduction 23
2.2 Formulation of the Problem as a NLPP 25
2.3 Determining Optimum Calibrated Weights 27
2.4 Numerical Examples 28
2.5 Comparison Study 33
A. Real Data 36
B. Simulated Data 38
Chapter 3 41

A Calibration Estimator of Population Mean in Stratified Random Sampling using Multiple Auxiliary Variable 41
3.1 Introduction 41
3.2 A General Formulation of the Problem as a NLPP 42
3.3 Determining Optimal Calibrated Weights 45
3.4 Numerical Example 47
3.5 Comparison Study 51
A. Real Data 52
B. Simulated Data 54
Chapter 4 57
Conclusion 57
Bibliography 59
Appendix A 64
Mathematica Code for finding $\lambda_0, \lambda_1, \varphi_1$ for univariate calibration estimator 64
Appendix B: 68
Mathematica Code for finding $\lambda_0, \lambda_1, \lambda_2, \varphi_1, \varphi_2$ for multivariate calibration estimator 68

Abstract

Stratified Sampling is one of the most widely used sampling techniques as it increases the precision of the estimate of the survey variable. On the other hand, calibration estimation is a method of adjusting the original design weights to improve survey estimates by using auxiliary information such as the known population total (or mean) of the auxiliary variables. A calibration estimator uses calibrated weights that are determined to minimize a given distance measure to the original design weights while satisfying a set of constraints related to the auxiliary information. In this thesis, we propose calibration estimators of population mean in stratified sampling design using the auxiliary information available respectively, for univariate and multivariate, which incorporates not only the population mean but also the stratum variances of the auxiliary variables. The problems of determining the optimum calibrated weights in both cases are formulated as Nonlinear Programming Problems (NLPP) that are solved using Lagrange multiplier technique. Numerical examples with real and simulated data are presented to illustrate the computational details of the solution procedure. Comparison studies are also carried out to evaluate the performance and the usefulness of the proposed calibration estimators. The results reveal that the proposed calibration estimators are more efficient than the other existing calibration estimators of the population mean in stratified sampling as the estimators gives least sampling error (SE) and higher gain in relative efficiency (RE) as compared to other estimators.

Preface

This thesis entitled "Calibration Approach for Stratified Sampling Design" 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.

Sampling is an essential part of most research since it is used as means of providing statistical data or information on an extensive range of subjects for both research and administrative purposes. Various surveys are conducted to develop hypotheses in diverse fields of human activities such as economics, operation research, demography, political science, education, statistical analysis, medical research, etc. A considerable use of survey is used by governments to look into the conditions of their population in terms of unemployment, income, expenditure, health, education, etc.

To gain precision in the estimate of a characteristic in sample survey, stratified sampling is a widely used sampling technique. Also, when the information on some auxiliary variables is available, the precision of the estimates can further be improved, if the auxiliary information is efficiently during estimation stage. The ratio estimation, regression estimation and the calibration estimation are such techniques that use the auxiliary information.

In this thesis, a research is carried out to improve the precision of the estimates in stratified sampling using calibration estimation. The calibration estimation is a technique that uses auxiliary information to improve the precision and accuracy of the estimators of population parameter in survey sampling. This technique chooses the adjusted weights that minimize a distance between the original weights and the adjusted weights, while satisfying a set of constraints related to the auxiliary information. The problem of determining the optimum calibrated weights may be formulated as a nonlinear programming problem (NLPP) which can be solved using Lagrange Multiplier technique. With the growing significant attention of calibration approach, several national statistical agencies are inculcating this technique since it leads to consistent estimates, provides an important class of technique for the efficient combination of data sources and has computational advantage to calculate estimates.

The use of calibration estimation was first introduced by Deville and Särndal (1992) in survey sampling. Since then many authors such as Dupont (1995), Singh et al. (1998, 1999, 2006, 2011), Singh (2001, 2003 , 2006, 2011, 2012), Farrell and Singh (2002, 2005), Wu and Sitter (2001), Sarndal (2007), Estevao and Sarndal (2001, 2003), Kott (2003), Montanari and Ranalli (2005), Rueda et al. (2010), Kim (2009, 2010) and many others have contributed to the study of calibrated estimation in survey sampling. In stratified random sampling, Singh et al. (1998) introduced the calibration approach. Later, many authors such as Tracy et al. (2003), Singh (2003), Kim et al. (2007), Rao et al. (2012) proposed different calibration estimators for population mean in stratified sampling.

This thesis is an attempt to propose calibration estimators of population mean in stratified random sampling incorporating population mean and stratum variances available for single and multiple auxiliary variables. Depending upon the availability and the use of single and multiple auxiliary variables, the problem of determining optimum weights for the proposed calibrated estimators, respectively, are considered two different kinds of optimization problems. The first kind of problem, which is the determination of calibrated weight in stratified random sampling when the information on a single auxiliary variable available, is discussed in Chapter 2. Whereas, the second kind of problem, which is the determination of calibrated weight in stratified random sampling when the information on more than one auxiliary variable is discussed in Chapter 3.

This thesis consists of four chapters. **Chapter 1** provides an introduction to survey sampling and describes the stratified random sampling, the use of auxiliary variable, calibration estimation, distance function, nonlinear programming and Lagrange multiplier technique. A brief review of the literature and studies about the calibration approach is also presented in this chapter.

In **Chapter 2**, a calibration estimator of population mean in stratified sampling design is proposed when the information on a single auxiliary variable is available. The estimator incorporates the information of the population mean but also the stratum variances available for the auxiliary variable. Then, the problem of determining optimum calibrated weights is formulated as a Nonlinear Programming Problem (NLPP) that seeks minimization of the chi-square type distance function subject to some calibration constraints. A new calibration constraint using the known stratum variances is introduced in our formulation of the NLPP, in addition to the constraints on weight and population mean. A solution procedure using Lagrange Multiplier technique is discussed to solve the NLPP. Two numerical examples are presented to illustrate the application and the computational details of the proposed method. Finally, using real and simulated data; a comparison study is also made to investigate the efficiency of the proposed calibration estimator.

In **Chapter 3**, a calibration estimator of population mean in stratified sampling design is proposed when the information on multiple auxiliary variables is available. The approach of determining the calibrated estimator is as similar as discussed in Chapter 2. The problem of determining optimum calibrated weights is formulated as an NLPP that seeks the sum of chi-square type distance of design weight of each auxiliary variable. Then, the formulated NLPP is solved using a Lagrange Multiplier technique. Numerical examples and a comparison study are presented, respectively, to illustrate the computational details and to investigate the efficiency of the proposed calibration estimator.

Finally, **Chapter 4** provides a brief conclusion to this thesis. A comprehensive list of references is presented in Bibliography at the end of the thesis.

Chapter 1

Introduction

1.1 Sample Surveys

The word 'survey' is used often to describe a method of gathering information on a target population. A **sample survey** is a method of gathering information that selects a sample of elements from the target population in order to estimate population attributes.

Sample surveys are nowadays widely accepted means of providing information on an extensive range of subjects for both research and administrative purposes. Numerous surveys are conducted to develop, test and refine research hypothesis in such disciplines as sociology, social psychology, demography, political science, economics, education and public health. Governments make considerable use of surveys to inform them of the conditions of their populations in terms of employment and unemployment, income and expenditure, housing conditions, education, nutrition, health, travel patterns, and many other subjects. They also conduct surveys of organizations such as manufacturers, retail outlets, farms, schools, and hospitals. Market researchers carry out surveys to identify market for products, to discover how the products are used and how they perform in practice, and to determine customer's reactions. Opinion polls keep track of the popularity of political leaders and their parties and measure public opinion on a variety of topical issues.

Some of the advantages of sample surveys 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.

A variety of sampling techniques have been developed to provide efficient estimates of the characteristics. Among the most widely used are simple random sampling, stratified sampling, systematic sampling, cluster sampling, multistage sampling and double sampling. **Stratified random sampling** is a common and popular technique among various sampling designs that are extensively used in sample surveys.

1.2 Stratified Random Sampling

Stratified random sampling is a technique which attempts to restrict the possible sample to those which are "less extreme" by ensuring that all parts of the population are represented in the sample in order to increase the efficiency (that is to decrease the error in the estimation). In stratified sampling a population of size N is first divided into L disjoint groups of sizes $N_1, N_2, ..., N_h, ..., N_L,$ respectively. These subgroups, called strata, together they compromise the whole population, so that $N_1 + N_2 + N_3 + ...N_L = N.$ For each stratum a simple random sample of pre-specified size, $n_h,$ is drawn independently from hth strata, such that $h = 1, 2, ,..., L$ such that

$n_1 + n_2 + ... + n_L = n.$

In general, stratification of population units depends on the purpose of the survey. 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.

Some of the advantages of stratified random sampling are as follows:

i. The overall population total, mean, and other parameters of the entire population can be estimated more efficiently, and produces results that are unbiased and accurate. 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. 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 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. The estimates for each stratum can be obtained separately. This ensures that all important subgroups are represented in the sample.

iv. It is easier to sample separately from the strata rather than from the entire population (especially, if the population 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. 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. 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.

1.3 Stratified Estimators

In estimating a characteristic in a stratified random sampling (SRS), let us define that:

N : Population Size

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

y_{hj} : Value of the *j*th unit in the *h*th stratum; $j = 1, 2, \cdots, N_h$ and $h = 1, 2, \cdots, L$.

$Y_h = \sum_{j=1}^{N_h} y_{hj}$: Total of the *h*th stratum.

$\bar{Y}_h = \frac{1}{N_h} \sum_{j=1}^{N_h} y_{hj}$: Stratum mean.

$W_h = \frac{N_h}{N}$: Stratum weight = Proportion of the units falling in *h*th stratum. $\left(\sum_{h=1}^{L} W_h = 1 \right)$

$S_h^2 = \frac{1}{N_h - 1} \sum_{j=1}^{N_h} \left(y_{hj} - \bar{Y}_h \right)^2$: Stratum variance.

$Y = \sum_{h=1}^{L} \sum_{j=1}^{N_h} y_{hj}$ = Population total.

$$\bar{Y} = \frac{\sum_{h=1}^{L}\sum_{j=1}^{N_h} y_{hj}}{N} = \text{Population mean.}$$

n_h : Size of SRS from the hth stratum

$$f_h = \frac{n_h}{N_h} = \text{Sampling fraction in } h\text{th stratum.}$$

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

$$\bar{y}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} y_{hj} = \text{Sample mean of } h\text{th stratum.}$$

$$s_h^2 = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(y_{hj} - \bar{y}_h\right)^2 = \text{Sample variance in } h\text{th stratum.}$$

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

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad \ldots(1.1)$$

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

$$V\left(\bar{y}_{st}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_h^2 \qquad \ldots(1.2)$$

An unbiased estimate of the population total Y is given as

$$\hat{Y}_{st} = N\,\bar{y}_{st} \qquad \ldots(1.3)$$

with variance

$$V\left(\hat{Y}_{st}\right) = N^2 \sum_{h=1}^{L}\left(\frac{1}{n_h} - \frac{1}{N_h}\right) W_h^2 S_h^2 \qquad \ldots(1.4)$$

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

$$v\left(\bar{y}_{st}\right)=\sum_{h=1}^{L}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)W_h^2 s_h^2 \qquad \ldots(1.5)$$

1.4 Allocation of Sample Sizes to Different Strata

In stratified sampling the allocation of the sample 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.

1.4.1 Proportional Allocation

When no information except N_h is available, the allocation of a given sample of size n to different strata is done in proportion to their strata sizes, that is,

$$n_h \;\alpha\; N_h$$

or

$$n_h = n \cdot W_h \qquad \ldots(1.6)$$

1.4.2 Neyman Allocation

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

$$n_h = n \cdot \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \qquad \ldots(1.7)$$

1.4.3 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 $V\left(\bar{y}_{st}\right)$ for a fixed a fixed cost $C = c_o + \sum_{h=1}^{L} c_h n_h$, is obtained as:

$$n_h = \frac{C - c_0}{\sum_{h=1}^{L} W_h S_h \sqrt{c_h}} \cdot \frac{W_h S_h}{\sqrt{c_h}} \qquad \ldots(1.8)$$

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

1.5 Study Variable and Auxiliary Information

In sample surveys, the variable of interest or the variable about which we want to draw some inference is called a study variable. In many surveys, it is possible to collect information about some variable(s) in addition to the study variable. This information is known as auxiliary information or variable, which is accurately known from many sources and is cheaper to obtain than the study variable. Sometimes auxiliary variables are closely related to the study variable. The problem of estimating the population parameters in the presence of an auxiliary variable has been widely discussed in finite population sampling literature. Out of many ratio, product and regression methods of estimation are good examples in this context.

An efficient use of the auxiliary population information provided by one or several auxiliary variables at the estimation stage is also a very useful technique in order to obtain more precise and reliable estimates. Powerful auxiliary information can produce a successful reduction of the bias and the sampling error Khan et al. (2010). Over the last couple of years, a new method, called calibration approach, has appeared to add auxiliary information with the aim to obtain new estimators of the population total or mean.

1.6 Calibration Estimation

Calibration estimation is a method of adjusting the original design weights to improve survey estimates by using auxiliary information such as the known population total (or mean) of the auxiliary variables. A calibration estimator uses calibrated weights that are determined to minimize a given distance measure to the original design weights while satisfying a set of constraints related to the auxiliary information.

Calibration is commonly used in survey sampling to increase the precision of the estimators of population parameter when auxiliary information is available. The method works by modifying the original design weights incorporating the known population characteristics, in practice population totals or population means, of the auxiliary variables. Deville and Särndal (1992) first used the calibration estimators in survey sampling.

The basic idea in calibration estimation is to use auxiliary information to obtain a better estimate of a population parameter. Consider a population of size N from which a probability sample of size n is drawn. Let y_i be the value of the study variable y for the ith population unit and x_i be the value of the ith unit of an associated auxiliary variable x. In a simple case, it is assumed that population total $X = \sum_{i=1}^{N} x_i$ is accurately known.

If the objective is to estimate the population total $Y=\sum_{i=1}^{N} y_i$, Deville and Särndal (1992) used the auxiliary information on the known population total X, to modify the basic sampling design weights, $d_i = \frac{1}{\pi_i}$, that appear in Horvitz and Thompson (1952) estimator

$$\hat{Y}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i} = \sum_{i=1}^{n} d_i y_i. \qquad \ldots(1.9)$$

Then Deville and Särndal (1992) proposed a new estimator

$$\hat{Y}_{DS} = \sum_{i=1}^{n} w_i y_i \qquad \ldots(1.10)$$

that minimize a chi-square type distance function given by

$$\sum_{i=1}^{n} \left(\frac{w_i - d_i}{d_i q_i} \right)^2 \qquad \ldots(1.11)$$

subject to the calibration constraint

$$\sum_{i=1}^{n} w_i x_i = X \qquad \ldots(1.12)$$

where, w_i is the new calibrated weights.

Note that q_i are appropriately chosen weights which determines the estimator. In most situations, $q_i = 1$.

Using a Lagrange multiplier technique, the optimum calibrated weights (w_i) are determined by

$$w_i = d_i + \frac{d_i q_i x_i}{\sum_{i=1}^{n} d_i q_i x_i^2} \left(X - \sum_{i=1}^{n} d_i x_i \right) \qquad \ldots(1.13)$$

Now substituting the value of w_i from (1.13) into (1.10) giving the traditional calibrated regression estimator of total as:

$$Y_{DS} = \sum_{i=1}^{n} d_i y_i + \sum_{i=1}^{n} \frac{d_i q_i x_i y_i}{d_i q_i x_i^2} \left(X - \sum_{i=1}^{n} d_i x_i \right) \qquad \ldots(1.14)$$

1.7 Distance Function

In calibration estimation discussed in Section 1.6, the optimum calibrated weights (w_i) are determined by formulating an optimization problem that seeks minimization of a distance function. Thus, distance functions are often used as error or cost functions to be minimized in an optimum problem. Deville and Särndal (1992) discussed about a class of distance functions for calibration estimation.

In stratified sampling design, some distance functions that are frequently used in practice are:

(a) Chi-square type:

$$\sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h Q_h}, \qquad \ldots(1.15)$$

where, Q_h are suitability chosen weights which determine the form of estimator.

(b) Mean-squared error:

$$\sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{L Q_h} \qquad \ldots(1.16)$$

(c) Euclidean:

$$\sqrt{\sum_{h=1}^{L}\frac{\left(W_h^* - W_h\right)^2}{Q_h}} \qquad \ldots(1.17)$$

(d) Hellinger:

$$\sum_{h=1}^{L}\frac{\left(\sqrt{W_h^*} - \sqrt{W_h}\right)^2}{Q_h} \qquad \ldots(1.18)$$

(e) Minimum entropy:

$$\sum_{h=1}^{L}\frac{\left(W_h^* - W_h\right)^2}{Q_h} - \sum_{h=1}^{L}\frac{W_h}{Q_h}\log\left(\frac{W_h^*}{W_h}\right) \qquad \ldots(1.19)$$

1.8 Nonlinear Programming Problem

The Nonlinear Programming Problem (NLPP) 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. In an NLPP, some or all the functions either the objective function or the left-hand side of the constraints are nonlinear. This technique is widely used to solve many decision making problems in areas including economics, education, public health, demography, psychology, sociology, political science and many others. For a non-linear programming model the main components are the values of the decision variables which describe the solutions; the objective function which measures the nature of solutions; the constraints which presents the relationships between decision variables.

A general Nonlinear Programming Problem (NLPP) can be expressed as follows:

Find the values of the decision variables $x_1, x_2, ..., x_n$ that

$$\text{Maximize (or Minimize)} \qquad Z = f(x_1, x_2, ..., x_n) \qquad \text{...(1.20)}$$

$$\text{Subject to} \qquad g_1(x_1, x_2, ..., x_n)(\leq, =, or \geq) b_1$$

$$g_2(x_1, x_2, ..., x_n)(\leq, =, or \geq) b_2$$

$$\vdots$$

$$g_m(x_1, x_2, ..., x_n)(\leq, =, or \geq) b_m \qquad \text{...(1.21)}$$

$$\text{and} \qquad x_j \geq 0;\ j = 1, 2, \cdots, n. \qquad \text{...(1.22)}$$

The function $f(x_1, x_2, ..., x_n)$ is known as objective function, the functions g_i in (1.21) are the constraints in which only one sign among $\leq, =, \geq$ holds true for each $i; (i = 1, 2, ..., m)$. The variables $x_j;\ j = 1, 2, \cdots, n$ are called decision variables.

The problems of calibrated estimation discussed in the thesis are usually non-linear programming problems with non-linear objective function and linear/nonlinear constraints.

1.9 Lagrange Multiplier Technique

Lagrange multiplier technique can be used to solve an NLPP in which all the constraints are equality constraints, that is,

$$\text{Maximize (or Minimize)} \quad Z = f(x_1, x_2, \cdots, x_n),$$

$$\text{Subject to} \qquad g_1(\mathrm{x}_1, x_2, ..., \mathrm{x}_n) = b_1$$

$$g_2(x_1, \mathrm{x}_2, ..., x_n) = b_2$$
$$\vdots$$

$$g_m(x_1, x_2, ..., x_n) = b_m \qquad \ldots(1.23)$$

To solve (1.23), we associate a multiplier λ_1 with the ith constraint in (1.23). Then, the Lagrangian function ϕ is formed as:

$$\phi(x_1, x_2, ..., x_{n,} \lambda_1, \lambda_2, ..., \lambda_m) = f(x_1, x_2, ..., x_n) + \sum_{i=1}^{m} \lambda_i [g_i(x_1, x_2, ..., x_n) - b_i] \qquad \ldots(1.24)$$

Then, we attempt to find a stationary point $(x_1, x_2, ..., x_{n,} \lambda_1, \lambda_2, ..., \lambda_m)$ that maximizes (or minimizes) the Lagrangian function $\phi(x_1, x_2, ..., x_{n,} \lambda_1, \lambda_2, ..., \lambda_m)$.

The necessary conditions for the solution of the problem (1.23) are:

$$\frac{\partial \phi}{\partial x_1} = \frac{\partial \phi}{\partial x_2} = ... = \frac{\partial \phi}{\partial x_m} = \frac{\partial \phi}{\partial \lambda_1} = \frac{\partial \phi}{\partial \lambda_2} = ... = \frac{\partial \phi}{\partial \lambda_m} = 0 \qquad \ldots(1.25)$$

If the objective function $f(x_1, x_2, ..., x_n)$ is a concave function in a maximization problem or a convex function in a minimization problem, then any point $(x_1, x_2, ..., x_n, \lambda_1, \lambda_2, ..., \lambda_m)$ that satisfies (1.25) yields an optimum solution $(x_1, x_2, ..., x_n)$.

In this thesis, the problems of determining the optimum weights in calibrated estimation is formulated as NLPPs that minimizes a distance, subject to the available calibration constraints. The NLPPs are then solved using Lagrange multiplier technique.

1.10 Calibration Approach: Review of Literature and Studies

Calibration estimation, on which the current research is conducted, dates back to 1992. A large amount of literature is being devoted to it, gaining significant attention in the field of survey methodology and survey practice. It is a technique that uses auxiliary information to improve the precision and the accuracy of the estimators of population parameter at survey variable in sampling. The technique works by modifying the design weights to improve the survey estimates by using the available auxiliary information. The notion of calibration estimators was first introduced by Deville and Särndal (1992) in survey sampling as discussed in Section 1.6.

Since then several survey statisticians such as Dupont (1995), Singh et al. (1998, 1999, 2006, 2011), Singh (2001, 2003 , 2006, 2011, 2012), Farrell and Singh (2002, 2005), Wu and Sitter (2001), Sarndal (2007), Estevao and Sarndal (2001, 2003), Kott (2003), Montanari and Ranalli (2005), Rueda et al. (2010), Kim (2009, 2010) and many others have contributed to the study of calibrated estimation in survey sampling.

In stratified random sampling, Singh et al. (1998) introduced the calibration approach. For the combined generalized regression estimator they proposed the calibration estimators by using a single auxiliary variable by making assumption that the population consists of L strata with N_h units in the *h*th stratum whereby a simple random of size n_h is drawn without replacement such that the population size $N=\sum_{h=1}^{L} N_h$ and the sample size is $n=\sum_{h=1}^{L} n_h,$. Let y_{hj} and x_{hj} be the values of the survey and the auxiliary variables, respectively, associated with *i*th units of the *h*th stratum. Also, let $W_h=\frac{N_h}{N}$ be the stratum weights, $f_h=\frac{n_h}{N_h}$ the sampling fraction and

$\bar{y}_h, \bar{x}_h, \bar{Y}_h, \bar{X}_h$ the sample and population means respectively. It is assumed that the population mean $\bar{X} = \sum_{h=1}^{L} W_h \bar{x}_h$ is completely known.

If the purpose is to estimate the population mean $\bar{Y}$ by using the auxiliary information X, the usual estimator of $\bar{Y}$ is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad \ldots(1.26)$$

Then, Singh et al. (1998) proposed a new estimator given by

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h \qquad \ldots(1.27)$$

with new weights W_h^{*} called the calibrated weights which are determined by minimizing the chi-square (CS) distance function

$$D = \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h q_h} \qquad \ldots(1.28)$$

subject to the calibration constraint

$$\sum_{h=1}^{L} W_h^{*} \bar{x}_h = \bar{X} .$$

Solving the problem stated above using the Lagrange multiplier technique, the optimum calibrated weight W_h^{*} is determined as:

$$W_h^{*} = W_h + W_h Q_h \bar{x}_h \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2 \right)^{-1} \left(\bar{X} - \bar{x}_{st} \right) \qquad \ldots(1.29)$$

Substituting (1.29) in (1.27) the calibrated estimator is obtained as:

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}\left(\bar{X} - \bar{x}_{st}\right). \qquad \ldots(1.30)$$

This leads to Singh et al. (1998) generalized linear regression (GREG) estimator, where $\hat{\beta} = \sum_{h=1}^{L} W_h Q_h \bar{x}_h \bar{y}_h \left(\sum_{h=1}^{L} W_h Q_h \bar{x}_h^2\right)^{-1}$.

Later, many authors proposed different calibration estimators for population mean in stratified sampling. Tracy et al. (2003) introduced a new calibration equation by making use of second order moments of auxiliary variable for estimating the population mean in stratified sampling. They proposed the estimator by considering the problem that seeks minimization of the chi-square (CS) distance function in (1.28) subject to the following two calibration constraints:

$$\sum_{h=1}^{L} W_h^{*} \bar{x}_h = \sum_{h=1}^{L} W_h \bar{X}_h$$

$$\sum_{h=1}^{L} W_h^{*} s_{hx}^2 = \sum_{h=1}^{L} W_h S_{hx}^2$$

where $s_{hx}^2 = \sum_{i=1}^{n_h} \frac{\left(x_{hi} - \bar{x}_h\right)^2}{n_h - 1}$ and $S_{hx}^2 = \sum_{i=1}^{N_h} \frac{\left(X_{hi} - \bar{X}_h\right)^2}{N_h - 1}$.

Singh (2003) proposed a calibration estimator of the population mean in stratified sampling by introducing a new calibration equation that is sum of calibrated weight be equal to sum of design weights, that is,

$$\sum_{h=1}^{L} W_h^* = 1 .$$

Kim et al. (2007) proposed various calibration approach ratio estimators and derived the estimator of the variance of the calibration approach for combined ratio estimators in stratified sampling. Rao et al. (2012) proposed a multivariate calibration estimator for the population mean using two auxiliary variables in stratified random sampling.

In this thesis, we propose new calibration estimators of population mean in stratified random sampling design using univariate and multivariate auxiliary information, which incorporates not only the population mean but also the population variance available for one and more than one auxiliary variable respectively. The problem of determining the optimum calibrated weights is formulated as a nonlinear programming problem (NLPP) that minimizes the chi-square type distance, subject to the available calibration constraints. A solution procedure is developed to solve the NLPPs using Lagrange multiplier technique.

Chapter 2

A Calibration Estimator of Population Mean in Stratified Random Sampling using a Single Auxiliary Variable

2.1 Introduction

In survey sampling, the use of auxiliary information greatly improves the precision of estimates of population parameters. Thus, the calibration technique that provides a systematic way to incorporate auxiliary information is becoming a widely used procedure of estimation in survey sampling. Since calibration is regarded as an important methodological instrument in large scale-production of statistics, several statisticians usually make large efforts to project surveys that are based on the efficient use of all the available auxiliary information, whether it is univariate or multivariate.

In stratified random sampling, when the information on a single auxiliary is available, many authors proposed different calibration estimators for population mean in stratified sampling. They used the mean of the auxiliary variable in their calibrations. Later, Tracy et al. (2003) introduced a calibration estimator by incorporating second order moments of the auxiliary variable. Singh (2003) also proposed a calibration estimator of the population mean by introducing a new calibration equation that is the sum of the calibrated weight which should be equal to sum of the design weights.

However, in addition to population mean, when the variance of the stratified mean of the auxiliary variable is accurately known, the precision of the estimate can further be increased by adjusting the design weights. Thus, in this chapter, we attempt to determine calibrated weights for estimating population mean in stratified sampling when the stratum variances of the auxiliary variable together with the population mean are available. The problem of determining the optimum calibrated weights is formulated as a nonlinear programming problem (NLPP) that minimizes the chi-square type distance, subject to the available calibration constraints. A solution procedure is developed to solve the NLPP using Lagrange multiplier technique.

Two numerical examples with real data are presented to illustrate the computational details of the solution procedure to determine the optimum calibrated weights and the optimum calibrated estimator. Finally, a comparison study with real and simulated data is carried out to demonstrate the practical application and the performance of the proposed estimator. The discoveries reveal that the proposed calibration estimator is more efficient than the calibrated estimators proposed by Singh et al. (1998), Singh (2003) and Tracy et al. (2003) of the population mean in stratified random sampling.

The work presented in this chapter has already been published in the Asia World Congress on Computer Science and Engineering (APWC on CSE), 2014, Conference Proceeding.

2.2 Formulation of the Problem as a NLPP

Let the population be divided into L non-overlapping strata and n_h be the number of units drawn by simple random sampling without replacement (SRSWOR) from the *h*th stratum consisting of N_h units, where $\sum_{h=1}^{L} n_h = n$ and $\sum_{h=1}^{L} N_h = N$ gives the total sample size and the population size respectively. For the *h*th strata, let $W_h = N_h / N$ be the strata (design) weights, and $\bar{y}_h$ and $\bar{Y}_h$ be the sample and population means, respectively, for the study variable.

Let the estimation of unknown population means $\bar{Y}$ be of interest using the information from an auxiliary variable X. Let y_{hi} and x_{hi} denote the values of the *i*th population (sampled) unit of the study variable (Y) and the auxiliary variable (X) respectively, in the *h*th stratum. Assume that the population mean $\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$ and also the stratum variances $S_h^2 = \frac{1}{N_h - 1} \sum_{i=1}^{N_h} \left(X_{hi} - \bar{X}_h \right)^2$ of the auxiliary variable are accurately known.

The purpose is to propose a calibration estimator of the population mean $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ by using the auxiliary information X, which improves the estimates. Let

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \quad . \qquad \ldots(2.1)$$

be the stratified sampling estimator of population mean $\bar{Y}$.

In the presence of an auxiliary information, the new calibration estimator of the population mean $\bar{Y}$ under stratified sampling given by (see Singh et al. 1998; Singh, 2003; Tracy et al. 2003)

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h, \qquad \ldots(2.2)$$

with calibrated weights W_h^*. When an auxiliary variable X is available, the new weights W_h^* are so chosen such that the sum of the chi-square type distances given by

$$\sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h q_h} \qquad \ldots(2.3)$$

is minimum, subject to the calibration constraints:

$$\sum_{h=1}^{L} W_h^* = 1, \qquad \ldots(2.4)$$

$$\sum_{h=1}^{L} W_h^* \bar{x}_h = \bar{X}\,. \qquad ...(2.5)$$

Note that $q_h > 0$ in (2.3) are suitability chosen weights which will determine the form of estimator.

When the stratum variances (S_h^2) of the auxiliary variable are known, we may introduce a new calibration equation:

$$\sum_{h=1}^{L} W_h^* d_h s_h^2 = \sum_{h=1}^{L} W_h d_h S_h^2\,. \qquad \ldots(2.6)$$

where $d_h = \left(1/n_h - 1/N_h\right)$ and $s_h^2 = \dfrac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_h)^2$.

Thus, the problem of determining the optimum calibrated weights W_h^* may be formulated as a nonlinear programming problem (NLPP) as given below:

Minimize $\quad Z\left(W_1^*, W_2^*, \ldots, W_L^*\right) = \sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h q_h}$

subject to $\quad \sum_{h=1}^{L} W_h^* = 1,$

$$\sum_{h=1}^{L} W_h^* \bar{x}_h = \bar{X},$$

$$\sum_{h=1}^{L} W_h^* d_h s_{hx}^2 = \sum_{h=1}^{L} W_h d_h S_{hx}^2,$$

and $\quad W_h^* \geq 0;\ h=1,2,\ldots,L. \quad \ldots(2.7)$

2.3 Determining Optimum Calibrated Weights

Ignoring the restriction $W_h^* \geq 0$, we can use Lagrange multiplier technique to solve the NLPP (2.7) for determining the optimum values of W_h^*, since the constraints are of inequality form. If the values W_h^* satisfy the ignored restrictions, the NLPP in (2.7) is solved completely.

Defining λ_1, λ_2 and λ_3 as Lagrange multipliers, the Lagrange function is given as:

$$\phi = \sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h q_h} - 2\lambda_0 \left(\sum_{h=1}^{L} W_h^* - 1\right) - 2\lambda_1 \left(\sum_{h=1}^{L} W_h^* \bar{x}_h - \bar{X}\right) - 2\varphi_1 \left(\sum_{h=1}^{L} W_h^* d_h s_{hx}^2 - \sum_{h=1}^{L} W_h d_h S_{hx}^2\right). \quad \ldots(2.8)$$

To find the optimum value of W_h^* the necessary and sufficient conditions for solving W_h^* are:

$$\frac{\partial \phi}{\partial W_h^*} = \frac{2\left(W_h^* - W_h\right)}{W_h q_h} - 2\lambda_0 - 2\lambda_1 \bar{x}_h - 2\varphi_1 d_h s_{hx}^2 = 0, \quad \ldots(2.9)$$

$$\frac{\partial \phi}{\partial \lambda_0} = -2\left(\sum_{h=1}^{L} W_h^* - 1 \right) = 0, \qquad \ldots(2.10)$$

$$\frac{\partial \phi}{\partial \lambda_1} = -2\left(\sum_{h=1}^{L} W_h^* \bar{x}_h - \bar{X} \right) = 0, \qquad \ldots(2.11)$$

and $$\frac{\partial \phi}{\partial \varphi_1} = -2\left(\sum_{h=1}^{L} W_h^* d_h s_{hx}^2 - \sum_{h=1}^{L} W_h d_h S_{hx}^2 \right) = 0. \qquad \ldots(2.12)$$

From (2.9) we have

$$W_h^* = W_h + W_1 q_h (\lambda_0 + \lambda_1 \bar{x}_h + \varphi_1 d_h s_{hx}^2) \qquad \ldots(2.13)$$

where λ_0, λ_1 and φ_1 will be obtained by using (2.13) and solving a system of nonlinear equations obtained by (2.9) - (2.12) using MATHEMATICA.

If $W_h^* > 0$ in NLPP (2.7) are satisfied, the above problem is solved completely using the Lagrange multiplier technique.

2.4 Numerical Examples

Example 1: In order to illustrate and to demonstrate the performance of the proposed calibration estimator in (2), we use the data obtained from the 2002 and 1997 Agricultural Censuses in Iowa State conducted by the National Agricultural Statistics Service, USDA, Washington D.C. (source: http://www. agcensus.usda.gov/). The data were used in Khan et al. (2010) in which $N = 99$ counties in Iowa State are divided into $L = 4$ strata. Suppose that an estimate of average production of corn harvested in 2002 ($\bar{Y}$) is of interest using single auxiliary variable X, which is the quantity of corn harvested in 1997 for each county. From the dataset a sample of $n = 28$ counties using proportional allocation was selected from four strata as shown in Table 2.1.

Table 2.1: Sample selected from agricultural censuses

Stratum No.	Country	Corn Yield in 1997 (X)	Corn Yield in 2002 (Y)
1	Sioux	869342.7	769073
	O'Brien	556474.4	539878.8
2	Clay	485421.2	406514.5
	Emmet	423227.3	353411.3
	Hancock	649740.1	598189.3
	Hamilton	735860.1	605422.1
	Cerro Gordo	642734.1	508673.6
	Hardin	679905.5	526762.2
	Floyd	532043.9	473622.5
	Bremer	519584.4	386810.9
	Winneshiek	493184.9	411742
	Clayton	545491.3	535903.6
3	Boone	579518.7	532223
	Cedar	645777.1	496703.5
	Clarke	80095.24	73191.5
	Dallas	463447.5	447856.1
	Davis	128108.8	103785.3
	Henry	316577.7	290002.9
	Jones	527055	486803.5
	Marshall	616432.3	453389.9
	Monroe	99940.57	86834.17
	Poweshiek	495272.3	386640.9
	Union	166518.6	121141
	Warren	251910.1	241285.8
	Washington	487297.3	404288.1
4	Pottawattamie	745552.2	735025.4
	Crawford	681493.3	511032.3
	Page	252826.5	302827.1

To compute the calibrated weights for the above data, the information needed is summarized in Table 2.2.

Table 2.2: Information from agricultural censuses

h	N_h	W_h	n_h	$\bar{x}_h$	d_h	$\bar{y}_h$	s_h^2
1	8	0.08081	2	654476	0.375	712909	26265020000
2	34	0.34343	10	480705	0.071	570719	7848598920
3	45	0.45455	13	317242	0.055	373689	29862000000
4	12	0.12121	3	559957	0.250	559957	46720000000

For this population the known population mean of the auxiliary variable is $\bar{X} = 405654.19$ and the known stratum variances (S_h^2) are presented in Table 2.3.

Table 2.3: Stratum variances of X

h	S_h^2
1	21601503189.82
2	19734615816.65
3	27129658749.96
4	17258237358.45

We assume that $q_h = 1;\ h = 1, 2, ..., 4$.

Solving the equations (2.10) - (2.13) using MATHEMATICA, the values for λ_0, λ_1 and φ_1 were determined. For this example, the values of λ_0, λ_1 and φ_1 are obtained as:

$\lambda_0 = 4.66889,$

$\lambda_1 = -7.62326 \times 10^{-6}$ and

$\varphi_1 = -2.72456 \times 10^{-9}.$

Finally, equation (2.13) gives the proposed optimum calibrated weights $W_h^*;\ h = 1, ..., 4$ as presented in Column 2 of Table 2.4.

Table 2.4: Calibrated Weights using agricultural censuses data

h	W_h^*	$W_h^{*(1)}$	$W_h^{*(2)}$	$W_h^{*(3)}$
1	0.01733	0.07555	0.05026	0.07520
2	0.45245	0.32701	0.31181	0.32657
3	0.48697	0.44020	0.53497	0.43799
4	0.04325	0.11499	0.10295	0.11407
Total	1	0.95774	1	0.95383

Using (2.2) an estimate of the average production of quantity of corn harvested using the proposed calibration weights is given by

$$\bar{y}_{st}^* = \sum_{h=1}^{L} W_h^* \bar{y}_h = 476769.69. \qquad \text{…(2.14)}$$

Example 2: In this example, we use a tobacco population data of $N = 106$ countries with two variables: area (in hectares) and production (in metric tons). The data are obtained from the Agriculture Statistics 1999 reported in Singh (2003). The countries were divided into $L = 4$ strata and a sample of $n = 28$ countries using proportional allocation was selected as shown in Table 2.1. Suppose that an estimate of average production ($\bar{Y}$) of tobacco crop is of interest using a single auxiliary variable X = area. Assume that $\bar{X}$ in different counties are known. To compute the univariate calibrated weights in stratified sampling and the value of the estimate of $\bar{Y}$, we use the same sample units as obtained in Singh (2003).

To compute the calibrated weights for the above data, the information needed is summarized in Table 2.5.

Table 2.5: Information from tobacco population

h	N_h	W_h	n_h	$\bar{x}_h$	d_h	$\bar{y}_h$	s^2_{hx}
1	6	0.05660	3	1304.667	0.167	2592	722185.33
2	6	0.05660	3	2907.5	0.167	26763	839008125
3	8	0.07547	3	5191.667	0.208	14559.67	74387858.3
4	10	0.09434	3	21700	0.233	29900	6070000
5	12	0.11321	4	6808	0.167	12462.5	63572981.3
6	4	0.03774	2	1800	0.25	3375	1620000
7	30	0.28302	11	24481.55	0.058	38411.82	1801653230
8	17	0.16038	6	294809.2	0.108	477961.83	3227740000
9	10	0.09434	3	6303.667	0.233	7480.33	59939890.3
10	3	0.02830	2	350	0.167	822.5	125000

For this population the known population mean of the auxiliary variable is $\bar{X} = 34438.61$ and the known stratum variances (S_h^2) are presented in Table 2.6.

Table 2.6: Stratum variances of X for tobacco population

H	S_h^2
1	10899652.70
2	584984730.00
3	635958094.84
4	209817189.17
5	27842810.52
6	5876666.67
7	760238523.44
8	124004506112.85
9	8340765245.43
10	2963333.33

We assume that $q_h = 1; h = 1, 2, ..., 10$. Then, solving the equations (2.10) - (2.13) using MATHEMATICA, the values of λ_0, λ_1 and φ_1 are obtained as:

$$\lambda_0 = -0.0442712,$$

$$\lambda_1 = 0.0000108073$$

and $$\varphi_1 = -1.07112 \times 10^{-10}.$$

Finally, equation (2.13) gives the proposed optimum calibrated weights $W_h^*; h = 1, ..., 10$ as presented in Column 2 of Table 2.7 with the following notations in Columns 3, 4, and 5 respectively:

- $W_h^{*(1)}$ represents Singh et al. (1998)
- $W_h^{*(1)}$ represents Singh (2003)
- $W_h^{*(1)}$ represents Tracy et al. (2003)

Table 2.7: Calibrated Weights using tobacco population

h	W_h^*	$W_h^{*(1)}$	$W_h^{*(2)}$	$W_h^{*(3)}$
1	0.0548916	0.0564722	0.0645010	0.0564719
2	0.0710314	0.0536658	0.0607508	0.05365402
3	0.0762381	0.0747725	0.0853052	0.0747708
4	0.112274	0.0906852	0.1029180	0.09068467
5	0.116399	0.1118320	0.1275270	0.1118298
6	0.0368017	0.0376147	0.0429636	0.0376146
7	0.342226	0.2706490	0.3068770	0.2705246
8	0.0665353	0.0759668	0.0704579	0.0633362
9	0.0964492	0.0932724	0.1063840	0.0932766
10	0.027154	0.0282843	0.0323149	0.0282842
Total	1	0.89321	1	0.8804474

Using (2.1) an estimate of the average production of the tobacco crop using the proposed calibration weights is given by

$$\overline{y}_{st}^* = \sum_{h=1}^{L} W_h^* \overline{y}_h = 53775.78. \qquad \ldots(2.15)$$

2.5 Comparison Study

In this section, a comparison study is conducted to demonstrate the practical application and the performance of the proposed estimator.

The comparison is carried out on the efficiency of the following estimators with the stratified sampling estimator under proportional allocation, $\overline{y}_{st}$, using the real and simulated data:

1. Singh et al. (1998):

   As discussed in Section 1.10, Singh et al. (1998) proposed calibrated estimator

$$\bar{y}_{st}^{(1)} = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}^{(1)} \left(\bar{X} - \bar{x}_{st} \right) \qquad \ldots(2.16)$$

where $\hat{\beta}^{(1)} = \sum_{h=1}^{L} W_h q_h \bar{x}_h \bar{y}_h \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \right)^{-1}$ and the optimum calibrated weight is given by

$$W_h^{*(1)} = W_h + W_h q_h \bar{x}_h \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \right)^{-1} \left(\bar{X} - \bar{x}_{st} \right) \qquad \ldots(2.17)$$

2. Singh (2003):

The calibrated estimator proposed by Singh (2003) is

$$\bar{y}_{st}^{(2)} = \sum_{h=1}^{L} W_h \bar{y}_h + \hat{\beta}^{(2)} \left(\bar{X} - \bar{x}_{st} \right), \qquad \ldots(2.18)$$

where $\hat{\beta}^{(2)} = \dfrac{\sum_{h=1}^{L} W_h q_h \bar{x}_h \bar{y}_h \left(\sum_{h=1}^{L} W_h q_h \right) - \sum_{h=1}^{L} W_h q_h \bar{y}_h \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h \right)}{\left(\sum_{h=1}^{L} W_h q_h \right) \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \right) - \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h \right)^2}$ and the optimum

calibrated weight is given by

$$W_h^{*(2)} = W_h + \left\{ \frac{W_h q_h \bar{x}_h \left(\sum_{h=1}^{L} W_h q_h \right) - W_h q_h \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h \right)}{\left(\sum_{h=1}^{L} W_h q_h \right) \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \right) - \left(\sum_{h=1}^{L} W_h q_h \bar{x}_h \right)^2} \right\} \left(\bar{X} - \bar{x}_{st} \right) \qquad \ldots(2.19)$$

3. Tracy et al. (2003): $\bar{y}_{st}^{(3)}$

Introducing the auxiliary information on second order moment Tracy et al. (2003) proposed a calibrated estimator given by

$$\bar{y}_{st}^{(3)} = \bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h \left[\bar{y}_h + \hat{\beta}_1 \left(\bar{X}_h - \bar{x}_h \right) + \hat{\beta}_2 \left(S_{hx}^2 - s_{hx}^2 \right) \right] \qquad \ldots(2.20)$$

where

$$\hat{\beta}_1 = \left\{ \frac{\sum_{h=1}^{L} W_h q_h \bar{x}_h \bar{y}_h \left[\sum_{h=1}^{L} W_h \left(\left(\bar{X}_h - \bar{x}_h \right) \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \sum_{h=1}^{L} W_h \left(\left(S_{hx}^2 - s_{hx}^2 \right) \sum_{h=1}^{L} W_h q_h \bar{x}_h s_{hx}^2 \right) \right) \right]}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h q_h s_{hx}^2 \right)^2} \right\}$$

and $$\hat{\beta}_2 = \left\{ \frac{\sum_{h=1}^{L} W_h q_h s_{hx}^2 \bar{y}_h \left[\sum_{h=1}^{L} W_h \left(S_{hx}^2 - s_{hx}^2 \right) \sum_{h=1}^{L} W_h q_h \bar{x}_h^2 - \sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{x}_h \right) \sum_{h=1}^{L} W_h q_h \bar{x}_h s_{hx}^2 \right]}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h q_h s_{hx}^2 \right)^2} \right\}$$

The optimum calibrated weight is given by

$$W_h^* = W_h + \left\{ \frac{\left(W_h q_h \bar{x}_h \right) \left(\sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{x}_h \right) \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \sum_{h=1}^{L} W_h \left(S_{hx}^2 - s_{hx}^2 \right) \sum_{h=1}^{L} W_h q_h s_{hx}^2 \right)}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h q_h s_{hx}^2 \right)^2} \right\} +$$

$$\left\{ \frac{\left(W_h q_h s_{hx}^2 \right) \left(\sum_{h=1}^{L} W_h \left(S_{hx}^2 - s_{hx}^2 \right) \sum_{h=1}^{L} W_h q_h \bar{x}_h^2 - \sum_{h=1}^{L} W_h \left(\bar{X}_h - \bar{x}_h \right) \sum_{h=1}^{L} W_h q_h \bar{x}_h s_{hx}^2 \right)}{\sum_{h=1}^{L} W_h q_h \bar{x}_h^2 \sum_{h=1}^{L} W_h q_h s_{hx}^4 - \left(\sum_{h=1}^{L} W_h q_h s_{hx}^2 \right)^2} \right\} \quad \ldots(2.21)$$

4. Proposed estimator:

The proposed calibrated estimator $\bar{y}_{st}^*$ and the optimum calibrated weights W_h^* are given respectively, in (2.2) and (2.13).

To compare the efficiency of the estimators we have considered the following measures:

1. The sampling error (SE) of an estimator $\hat{\bar{Y}}$, given by $SE = |\bar{Y} - \hat{\bar{Y}}|$, where $\bar{Y}$ is the true average value of the parameter.

2. The relative efficiency (RE) of an estimator $\hat{\bar{Y}}$ with respect to the stratified sampling estimator $(\bar{y}_{st})$ is given by $RE = \hat{v}(\bar{y}_{st}) / \hat{v}(\hat{\bar{Y}}) \times 100\%,$ where $\hat{v}(\hat{\bar{Y}})$ is the estimated variance of the estimators under comparison and $\hat{v}(\bar{y}_{st})$ is the estimated variance of the estimator $\bar{y}_{st}$ under proportional allocation.

The estimated variance for estimators under comparison is obtained by:

- Stratified sampling estimator under proportional allocation:

$$\hat{v}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_h}{n_h}\right) \cdot \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

where $f_h = \dfrac{n_h}{N_h}$

- Singh et al. (1998):

$$\hat{v}\left(\bar{y}_{st}^{(1)}\right) = \sum_{h=1}^{L} \left(W_h^{*(1)}\right)^2 \left(\frac{1-f_h}{n_h}\right) \cdot \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

- Singh (2003):

$$\hat{v}\left(\bar{y}_{st}^{(2)}\right) = \sum_{h=1}^{L} \left(W_h^{*(2)}\right)^2 \left(\frac{1-f_h}{n_h}\right) \cdot \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

- Tracy et al. (2003):

$$\hat{v}\left(\bar{y}_{st}^{(3)}\right) = \sum_{h=1}^{L} \left(W_h^{*(3)}\right)^2 \left(\frac{1-f_h}{n_h}\right) \cdot \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

- Proposed estimator:

$$\hat{v}\left(\bar{y}_{st}^{*}\right) = \sum_{h=1}^{L} \left(W_h^{*}\right)^2 \left(\frac{1-f_h}{n_h}\right) \cdot \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2.$$

A. Real Data

Based on the real data used in Examples 1 and 2 in Section 2.4 we compare the performance of the proposed estimator with the estimator under proportional allocation and the other three calibration estimators given in (2.16), (2.18) and (2.20). The calibrated weights for all the different methods are presented in Table 2.4 and 2.7. The true average productions $\left(\overline{Y}\right)$ of the corn population in Example 1 and tobacco population in Example 2 are found to be:

Corn Population: $\overline{Y} = 474973.90$

Tobacco Population: $\overline{Y} = 52444.56$

The Columns 2 and 3 in Table 2.8 show, respectively, the estimated average production $\left(\hat{\overline{Y}}\right)$ of corn for different estimators and the sampling error of the estimators. Columns 4 and 5 show the estimated variance $\hat{v}\left(\hat{\overline{Y}}\right)$ and the RE of the estimators. The results reveal that the proposed calibrated estimator is most efficient among all other estimators as it gives least SE and higher gain in RE. The gain in efficiency of the proposed estimator over the stratified sampling estimator is 128.38%.

Table 2.8: Analysis of corn data

Estimator	$\hat{\overline{Y}}$	$SE = \left\lvert \overline{Y} - \hat{\overline{Y}} \right\rvert$	$\hat{v}\left(\hat{\overline{Y}}\right)$	RE
Stratified estimator, $\overline{y}_{st}$	491344.93	16371.03	954078803.6	100.0
Singh et al. (1998), $\overline{y}_{st}^{(1)}$	469372.16	5601.74	874526686.5	109.1
Singh (2003), $\overline{y}_{st}^{(2)}$	471351.22	3622.68	850031397.0	112.2
Tracy et al. (2003), $\overline{y}_{st}^{(3)}$	467536.93	7436.97	865022433.4	110.3
Proposed estimator, $\overline{y}_{st}^{*}$	476769.69	1795.79	743156887.2	128.4

Similarly the results for the tobacco data presented in Table 2.9, reveal that the proposed calibrated estimator is most efficient among all other estimators as it produces least SE and higher RE. The gain in efficiency of the proposed estimator over the stratified sampling estimator is 545.14%.

Table 2.9: Analysis of tobacco population data

Estimator	$\hat{\bar{Y}}$	$SE = \left\lvert \bar{Y} - \hat{\bar{Y}} \right\rvert$	$\hat{v}\left(\hat{\bar{Y}}\right)$	RE
Stratified estimator, $\bar{y}_{st}$	95373.90	42929.34	2822731121.0	100.0
Singh et al. (1998), $\bar{y}_{st}^{(1)}$	54329.76	1885.20	650863727.5	433.69
Singh (2003), $\bar{y}_{st}^{(2)}$	54132.89	1688.33	569781129.3	495.41
Tracy et al. (2003), $\bar{y}_{st}^{(3)}$	48287.10	4156.85	609479471.9	463.14
Proposed estimator, $\bar{y}_{st}^{*}$	53775.78	1331.22	517799741.5	545.14

B. Simulated Data

In this subsection the performance of the proposed estimator was compared with the other calibration estimators using a simulated data. The data for the auxiliary variable (X) was randomly generated using the R software for a normal population of size $N = 1000$ with a mean of 50 and a standard deviation of 5.2. Then, the data for the survey variable (Y) was generated from X using the linear function of $Y_i = AX_i + B$, where the values of the coefficients A and B are 1.2 and 701.0, respectively.

This simulated population is divided into $L = 4$ strata. Suppose that an estimate of population mean $\bar{Y}$ is of interest using single auxiliary variable, X. Using proportional allocation a sample of $n = 100$ was selected from the four strata. To compute the calibrated weights, the information needed is summarized in Table 2.10 below.

Table 2.10: Information from simulated data

h	N_h	W_h	n_h	$\bar{x}_h$	d_h	$\bar{y}_h$	s_h^2
1	319	0.319	27	43.95871	0.034	753.75050	7.32511
2	362	0.362	25	49.86476	0.037	761.03770	0.07830
3	220	0.220	21	54.45950	0.043	766.32650	0.07703
4	99	0.099	27	59.09469	0.027	771.90000	0.07709

The known mean of the auxiliary variable for this population is found to be $\bar{X} = 49.89973$ and the stratum variances are found as shown in Table 2.11.

Table 2.11: Stratum variances of X for simulated data

H	S_h^2
1	7.30214979
2	2.14808052
3	1.78813814
4	2.87378679

Then, solving equations (2.10)-(2.13) using MATHEMATICA, $\lambda_0 = -12.9801, \lambda_1 = 0.239806$, and $\varphi_1 = 12.4334$. Hence, equation (2.13) gives the optimum calibrated weight of the proposed estimator, which is presented in Column 2 in Table 2.12. The last three columns of the table show the weights of other three calibrated estimators.

Table 2.12: Calibrated Weights using simulated data

h	W_h^*	$W_h^{*(1)}$	$W_h^{*(2)}$	$W_h^{*(3)}$
1	0.52892	0.31897	0.31943	0.31900
2	0.00500	0.36196	0.36200	0.36195
3	0.24658	0.21997	0.21977	0.21996
4	0.21949	0.09899	0.09870	0.09898
Total	1	0.99989	1	0.99989

Using (2.2) the proposed calibration estimator is obtained as:

$$\bar{y}_{st}^* = \sum_{h=1}^{L} W_h^* \bar{y}_h = 760.87151. \qquad \ldots(2.22)$$

Moreover, the true population mean of the simulated data is $\bar{Y} = 760.87968$. Table 2.13 presents the results for the simulated data. It reveals the similar results as found in the real data. Thus, the proposed calibrated estimator is most efficient as compared to other estimators and the gain in efficiency of the proposed estimator over the stratified estimator is 386.20%.

Table 2.13: Analysis of simulated data

Estimator	$\hat{\bar{Y}}$	$SE = \left\lvert \bar{Y} - \hat{\bar{Y}} \right\rvert$	$\hat{v}\left(\hat{\bar{Y}}\right)$	RE
Stratified estimator, $\bar{y}_{st}$	760.95199	0.07231	45844670263.9	100
Singh et al. (1998), $\bar{y}_{st}^{(1)}$	760.86745	0.01222	45835552384.3	100.02
Singh (2003), $\bar{y}_{st}^{(2)}$	760.86902	0.01065	45745087574.3	100.22
Tracy et al. (2003), $\bar{y}_{st}^{(3)}$	760.91479	0.03511	44382712304.2	103.29
Proposed estimator, $\bar{y}_{st}^*$	760.87151	0.00817	11870824590.0	386.20

Chapter 3

A Calibration Estimator of Population Mean in Stratified Random Sampling using Multiple Auxiliary Variable

3.1 Introduction

In addition to population mean, when the variance of the stratified mean of the auxiliary variable is accurately known like in Chapter 2, we proposed a calibrated estimator in stratified sampling when the information is available for a single auxiliary variable.

In surveys, when the information on more than one auxiliary information is available, the precision of the estimate can further be increased by adjusting the design weights based on all the auxiliary information. In this chapter, we propose new calibration estimator of population mean in stratified random sampling design with the aid of several auxiliary information, in particular, the variance of stratified mean together with the population mean of the auxiliary variables for improving the precision of the estimate. Therefore, the proposed estimator incorporates not only the population mean but also the variance of stratified mean available for the multiple auxiliary variables. The problem of determining the optimum calibrated weights is formulated as a nonlinear programming problem (NLPP) that seeks minimization of the sum of chi-square type distance of design weight of each auxiliary variable, subject to some calibration constraints on design weights, means and variances. A solution procedure is developed to solve the NLPP using

Lagrange multiplier technique. The computational details of the procedure are illustrated in the presence of two auxiliary variables.

Numerical examples with real and simulated data are presented to demonstrate the practical application of the proposed estimator. To compare the efficiency gain of the proposed multivariate calibration estimator with the other existing calibration estimators a comparison study is carried out. The study reveals that the proposed multivariate calibrated estimator is more efficient than other estimators.

3.2 A General Formulation of the Problem as a NLPP

Consider that a finite population U of size N is stratified into L strata $U_1,...,U_L$ containing N_h units in *h*th stratum ($h=1,2,...,L.$) such that $\sum_{h=1}^{L} N_h = N$ and $W_h = N_h/N$ be the strata weights. A simple random sample of size n_h is drawn without replacement from the *h*th stratum such that $\sum_{h=1}^{L} n_h = n$.

Let Y be the study variable and the estimation of unknown population means $\bar{Y}$ be of interest using the information from p auxiliary variables X_j, $j=1,2,...,p$. Let y_{ih} and x_{ihj} denote the observed values of the *i*th population (sampled) unit of the study variable (Y) and the *j*th auxiliary variable (X_j) respectively, in the *h*th stratum. For the *h*th strata, $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{ih}$ is the sample mean of the study variable. Assume that population means $\bar{X}_j = \sum_{h=1}^{L} W_h \bar{X}_{hj}$ and the population variances $S_{hj}^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}\left(x_{ihj} - \bar{X}_{hj}\right)^2$ of all the p auxiliary variables are accurately known, where $\bar{X}_{hj}$ is the population mean of jth auxiliary variable in hth stratum and.

The purpose is to propose a calibration estimator of the population mean $\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$ by using the information from p auxiliary variables $X_j;\ j = 1, 2, ..., p.$ The classical unbiased estimator of the population mean is given by

$$\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h \qquad \ldots(3.1)$$

In the presence of p auxiliary information, the new calibrated estimate of the population mean $\bar{Y}$ under stratified sampling given by (see Singh et al. (1998); Singh, 2003; Tracy et al. 2003)

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h \qquad \ldots(3.2)$$

with new weights W_h^{*}. When more than one auxiliary variables $X_j;\ j = 1, 2, ..., p$ are available, the new weights W_h^{*} are so chosen such that the sum of chi-square type distance of design weight of each auxiliary variable given by

$$\sum_{j=1}^{p} \sum_{h=1}^{L} \frac{\left(W_h^{*} - W_h\right)^2}{W_h q_{hj}} \qquad \ldots(3.3)$$

is minimum, subject to the calibration constraints

$$\sum_{h=1}^{L} W_h^* = 1 , \qquad \ldots(3.4)$$

$$\sum_{h=1}^{L} W_h^* \bar{x}_{hj} = \bar{X}_j ; j = 1, 2, ..., p \ . \qquad \ldots(3.5)$$

Note that $q_{hj} > 0$ in (3.3) are suitability chosen weights which will determine the form of estimator.

When the variances (S_{hj}^2) of the auxiliary variable are known, we may introduce new calibration equations:

$$\sum_{h=1}^{L} W_h^* d_h s_{hj}^2 = \sum_{h=1}^{L} W_h d_h S_{hj}^2 ; j = 1, 2, ..., p \qquad \ldots(3.6)$$

where $$s_{hj}^2 = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} (x_{ihj} - \bar{x}_{hj})^2 ,$$

$$d_h = \left(1/n_h - 1/N_h \right) .$$

One of the challenges in calibration approach of estimation is that sometimes the calibrated weights do not satisfy the desired constraint of weights being non-negative. To avoid such situation one needs to impose the non-negativity restrictions:

$$W_h^* \geq 0 ; \ h = 1, 2, ..., L. \qquad \ldots(3.7)$$

Thus, the problem of determining the optimum calibrated weights W_h^* may be formulated as a nonlinear programming problem (NLPP) as given below:

$$\text{Minimize } Z\left(W_1^*, W_2^*, ..., W_L^*\right) = \sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h}$$

subject to

$$\sum_{h=1}^{L} W_h^* = 1,$$

$$\sum_{h=1}^{L} W_h^* \bar{x}_{hj} = \bar{X}_j; j = 1, 2, ..., p,$$

$$\sum_{h=1}^{L} W_h^* d_h s_{hj}^2 = \sum_{h=1}^{L} W_h d_h S_{hj}^2; j = 1, 2, ..., p,$$

and $$W_h^* \geq 0;\ h = 1, 2, ..., L\ . \qquad \text{…(3.8)}$$

where $$Q_h = \left(\sum_{j=1}^{p} \frac{1}{q_{hj}}\right)^{-1}.$$

3.3 Determining Optimal Calibrated Weights

Ignoring the restrictions $W_h^* \geq 0$, we can use Lagrange multiplier technique to solve the NLPP (3.8) when information on two auxiliary variables $X_j; (j = 1, 2)$ is available for determining the optimum values of W_h^*, since the constraints are inequality form. If the values W_h^* satisfy the ignored restrictions, the NLPP in (3.8) is solved completely.

To illustrate the solution procedure using a Lagrange multiplier technique, let us consider that the information on two auxiliary variables $X_j; (j = 1, 2)$ is available. Then, ignoring the restrictions $W_h^* \geq 0$, the NLPP (3.8) can be expressed as:

Minimize $Z\left(W_1^*, W_2^*, ..., W_L^*\right) = \sum_{h=1}^{L} \frac{(W_h^* - W_h)^2}{W_h Q_h}$

subject to $\sum_{h=1}^{L} W_h^* = 1,$

$$\sum_{h=1}^{L} W_h^* \bar{x}_{h1} = \bar{X}_1,$$

$$\sum_{h=1}^{L} W_h^* \bar{x}_{h2} = \bar{X}_2,$$

$$\sum_{h=1}^{L} W_h^* d_h s_{hx1}^2 = \sum_{h=1}^{L} W_h d_h S_{hx1}^2,$$

$$\sum_{h=1}^{L} W_h^* d_h s_{hx2}^2 = \sum_{h=1}^{L} W_h d_h S_{hx2}^2,$$

Defining $\lambda_0, \lambda_1, \lambda_2, \varphi_1, \varphi_2$ as Lagrange multipliers, the Lagrange function is given as:

$$\phi = \sum_{h=1}^{L} \frac{\left(W_h^* - W_h\right)^2}{W_h Q_h} - 2\lambda_0 \left(\sum_{h=1}^{L} W_h^* - 1\right) - 2\lambda_1 \left(\sum_{h=1}^{L} W_h^* \bar{x}_{h1} - \bar{X}_1\right) - 2\lambda_2 \left(\sum_{h=1}^{L} W_h^* \bar{x}_{h2} - \bar{X}_2\right)$$
$$- 2\varphi_1 \left(\sum_{h=1}^{L} W_h^* d_h s_{h1}^2 - \sum_{h=1}^{L} W_h d_h S_{h1}^2\right) - 2\varphi_2 \left(\sum_{h=1}^{L} W_h^* d_h s_{h2}^2 - \sum_{h=1}^{L} W_h d_h S_{h2}^2\right) \quad \text{…(3.9)}$$

The necessary and sufficient conditions for solving optimum values of W_h^* are:

$$\frac{\partial \phi}{\partial W_h^*} = \frac{2\left(W_h^* - W_h\right)}{W_h Q_h} - 2\lambda_0 - 2\lambda_1 \bar{x}_{h1} - 2\lambda_2 \bar{x}_{h2} - 2\varphi_1 d_h s_{h1}^2 - 2\varphi_2 d_h s_{h2}^2 = 0 \quad \text{…(3.10)}$$

$$\frac{\partial \phi}{\partial \lambda_0} = -2\left(\sum_{h=1}^{L} W_h^* - 1\right) = 0 \quad \text{…(3.11)}$$

$$\frac{\partial \phi}{\partial \lambda_1} = -2\left(\sum_{h=1}^{L} W_h^* \bar{x}_{h1} - \bar{X}_1\right) = 0 \quad \text{…(3.12)}$$

$$\frac{\partial \phi}{\partial \lambda_2} = -2\left(\sum_{h=1}^{L} W_h^* \bar{x}_{h2} - \bar{X}_2\right) = 0 \quad \text{(3.13)}$$

$$\frac{\partial \phi}{\partial \varphi_1} = -2\left(\sum_{h=1}^{L} W_h^* d_h s_{h1}^2 - \sum_{h=1}^{L} W_h^* d_h S_{h1}^2\right) = 0$$ ...(3.14)

$$\frac{\partial \phi}{\partial \varphi_2} = -2\left(\sum_{h=1}^{L} W_h^* d_h s_{h2}^2 - \sum_{h=1}^{L} W_h^* d_h S_{h2}^2\right) = 0$$ ...(3.15)

From (3.10) we have

$$W_h^* = W_h + W_h Q_h \left(\lambda_0 + \lambda_1 \bar{x}_{h1} + \lambda_2 \bar{x}_{h2} + \varphi_1 d_h s_{hx1}^2 + \varphi_2 d_h s_{hx2}^2\right)$$ ...(3.16)

which could be generalized for p auxiliary variables as:

$$W_h^* = W_h + W_h Q_h \left(\lambda_0 + \lambda_1 \bar{x}_{h1} + \lambda_2 \bar{x}_{h2} + ... + \lambda_p \bar{x}_{hp} + \varphi_1 d_h s_{hx1}^2 + \varphi_2 d_h s_{hx2}^2 + ... + \varphi_p d_h s_{hp}^2\right)$$

where $\lambda_0, \lambda_1, \lambda_2, \varphi_1$ and φ_2 will be obtained by using (3.16) and solving a system of nonlinear equations obtained by (3.10)-(3.15) using MATHEMATICA.

If the calibrated weights W_h^* obtained by (3.16) are non-negative then the NLPP (3.8) is solved completely.

3.4 Numerical Example

In order to illustrate and to demonstrate the performance of the proposed multivariate calibration estimator, we use the tobacco population data discussed in Example 2 in Chapter 2. The population consists of data of $N = 106$ counties with three variables: area (in hectares), yield (in metric tons) and production (in metric tons). The counties were divided into $L = 10$ strata and a sample of $n = 40$ counties using proportional allocation was selected as shown in Table 3.1. Suppose that an estimate of average production $(\bar{Y})$ of tobacco crop is of interest using the two auxiliary variables $X_1 = area$ and $X_2 = yield.$ Assume that X_1 and X_2 in different counties are known. To compute the multivariate calibrated weights in stratified sampling and the value of the estimate of $\bar{Y},$ we use the same sample units as obtained in Singh (2003).

Table 3.1: Sample selected from agricultural censuses with two auxiliary variables

Stratum No	Unit	Area (X_1)	Yield (X_2)	Production (Y)
1	2	580	1.79	1038
	5	2240	2.03	4550
	6	1094	2	2188
2	1	5900	0.63	37000
	2	27050	1.51	40950
	4	1175	1.99	2339
3	1	105	1.9	199
	2	320	3.69	1180
	8	15150	2.79	42300
4	2	24000	0.63	15000
	7	22000	1.36	30000
	8	19100	2.34	44700
5	2	4304	0.26	1100
	7	18600	2.05	38150
	9	3228	2.48	8000
	10	1100	2.36	2600
6	1	2700	1.96	5300
	2	900	1.61	1450
7	1	3950	0.99	3900
	5	3700	1.11	4110
	6	3400	1.62	5500
	7	750	0.87	650
	11	10000	0.26	2600
	12	8805	2.51	22120
	15	116700	1.22	142300
	20	10000	2.1	21000
	22	103110	2.06	212050
	26	4000	0.5	2000
	30	4882	1.29	6300
8	2	36000	1.22	44000
	5	1445000	1.75	2524500
	6	206625	0.85	175631
	10	25730	2.02	52040
	11	4000	0.75	3000
	15	51500	1.33	68600
9	1	161	1.5	241
	5	3750	1.33	5000
	7	15000	1.15	17200
10	2	100	0.95	95
	3	600	2.58	1550

To compute the calibrated weights for the above data, the information needed is summarized in Table 3.2.

Table 3.2: Information from tobacco population with two auxiliary variables

h	N_h	W_h	n_h	$\bar{x}_{h1}$	$\bar{x}_{h2}$	d_h	$\bar{y}_h$	s_{h1}^2	s_{h2}^2
1	6	0.05660	3	1304.667	1.940	0.167	2592	722185.33	0.0171
2	6	0.05660	3	2907.5	1.377	0.167	26763	839008125	0.475733
3	8	0.07547	3	5191.667	2.793	0.208	14559.67	74387858.3	0.801033
4	10	0.09434	3	21700	1.443	0.233	29900	6070000	0.736233
5	12	0.11321	4	6808	1.788	0.167	12462.5	63572981.3	1.06983
6	4	0.03774	2	1800	1.785	0.25	3375	1620000	0.061250
7	30	0.28302	11	24481.55	1.323	0.058	38411.82	1801653230	0.48245
8	17	0.16038	6	294809.2	1.320	0.108	477961.83	3227740000	0.24616
9	10	0.09434	3	6303.667	1.327	0.233	7480.33	59939890.3	0.03063
10	3	0.02830	2	350	1.765	0.167	822.5	125000	1.32845

Assuming that $Q_h = 1,$ the known population means ($\bar{X}_j$) of the auxiliary variables are obtained as:

$$\bar{X}_1 = 34438.61 \text{ and } \bar{X}_2 = 1.56.$$

The known population stratum variances (S_{hj}^2) of the auxiliary variables are presented in Table 3.3.

***Table 3.3: Stratum* variances of X_1 and X_2**

H	S_{h1}^2	S_{h2}^2
1	10899652.70	0.0268
2	584984730.00	0.2181
3	635958094.84	0.3470
4	209817189.17	0.2346
5	27842810.52	0.5821
6	5876666.67	0.1531
7	760238523.44	0.3439
8	124004506112.85	0.3786
9	8340765245.43	2.0183
10	2963333.33	0.9746

Solving the equations (3.10) - (3.15) using MATHEMATICA, the values for $\lambda_0, \lambda_1, \lambda_2,\ \varphi_1$ and φ_2 were determined. For this example, the values of $\lambda_0, \lambda_1, \lambda_2,\ \varphi_1$ and φ_2 are obtained as: $\lambda_0 = 0.626375$, $\lambda_1 = 2.63547\times10^{-6}$, $\lambda_2 = -0.501509$, $\varphi_1 = -4.04381\times10^{-11}$, $\varphi_2 = 3.11699$.

Finally, equation (3.16) gives the proposed optimum calibrated weights $W_h^*; h = 1,...,10$ as presented in Column 2 of Table 3.4 with the following notations in Columns 3, 4, 5 and 6 respectively:

- $W_h^{*(1)}$ represents Singh et al. (1998)
- $W_h^{*(2)}$ represents Singh (2003)
- $W_h^{*(3)}$ represents Tracy et al. (2003)
- $W_h^{*(4)}$ author's univariate.

Table 3.4: Calibrated weights using tobacco population with two auxiliary variables

h	W_h^*	$W_h^{*(1)}$	$W_h^{*(2)}$	$W_h^{*(3)}$	$W_h^{*(4)}$
1	0.03768	0.0548916	0.0564722	0.0645010	0.0575757
2	0.07101	0.0710314	0.0536658	0.0607508	0.0776931
3	0.05720	0.0762381	0.0747725	0.0853052	0.0805716
4	0.14098	0.112274	0.0906852	0.1029180	0.121396
5	0.14766	0.116399	0.1118320	0.1275270	0.123299
6	0.02957	0.0368017	0.0376147	0.0429636	0.0386372
7	0.31456	0.342226	0.2706490	0.3068770	0.367502
8	0.06649	0.0665353	0.0759668	0.0704579	0.0627263
9	0.09428	0.0964492	0.0932724	0.1063840	0.202123
10	0.04057	0.027154	0.0282843	0.0323149	0.0164
Total	1	1	0.89321	1	1.14792

Using (3.2) an estimate of the average production of the tobacco crop using the proposed calibration weights is given by

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h^{*} \bar{y}_h = 53585.53. \qquad \ldots(3.17)$$

3.5 Comparison Study

In this section, using the real and simulated data, a comparison study is carried out on the efficiency of the proposed multivariate calibrated estimator $\left(\bar{y}_{st}^{*}\right)$ with following estimators as discussed in Chapter 2, which are based on the single auxiliary variable:

1. Singh et al. (1998) estimator, $\bar{y}_{st}^{(1)}$, given by (2.16)
2. Singh (2003) estimator, $\bar{y}_{st}^{(2)}$, given by (2.18)
3. Tracy et al. (2003) estimator, $\bar{y}_{st}^{(3)}$, given by (2.20)
4. Author's univariate estimator, $\bar{y}_{st}^{(4)}$, given by (2.13) proposed in Chapter 2.

To compare the efficiency of the estimators we have considered the following measures as discussed in Section 2.5, Chapter 2:

1. The sampling error, $SE = |\,\bar{Y} - \hat{\bar{Y}}\,|$,
2. The relative efficiency, $RE = \hat{v}\left(\bar{y}_{st}\right) / \hat{v}\left(\hat{\bar{Y}}\right) \times 100\%$,

where $\hat{v}\left(\hat{\bar{Y}}\right)$ is the estimated variance of the estimators under comparison and $\hat{v}\left(\bar{y}_{st}\right)$ is the estimated variance of the estimator $\bar{y}_{st}$ under proportional allocation.

The estimated variance for estimators under comparison is obtained by:

- Stratified sampling estimator under proportional allocation:

$$\hat{v}\left(\bar{y}_{st}\right)=\sum_{h=1}^{L}W_h^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

- Singh et al. (1998):

$$\hat{v}\left(\bar{y}_{st}^{(1)}\right)=\sum_{h=1}^{L}\left(W_h^{*(1)}\right)^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

- Singh (2003):

$$\hat{v}\left(\bar{y}_{st}^{(2)}\right)=\sum_{h=1}^{L}\left(W_h^{*(2)}\right)^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

- Tracy et al. (2003):

$$\hat{v}\left(\bar{y}_{st}^{(3)}\right)=\sum_{h=1}^{L}\left(W_h^{*(3)}\right)^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

- Author's univariate estimator proposed in (2.13):

$$\hat{v}\left(\bar{y}_{st}^{(4)}\right)=\sum_{h=1}^{L}\left(W_h^{*(4)}\right)^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

- Proposed multivariate estimator:

$$\hat{v}\left(\bar{y}_{st}^{*}\right)=\sum_{h=1}^{L}\left(W_h^{*}\right)^2\left(\frac{1-f_h}{n_h}\right)\cdot\frac{1}{n_h-1}\sum_{i=1}^{n_h}\left(y_{hi}-\bar{y}_h\right)^2.$$

A. Real Data

Based on the real data used in Section 3.4 we compare the performance of the proposed estimator obtained based on the information of the two auxiliary variables ($\bar{X}_1$ = Area and $\bar{X}_2$ = Yield) with the estimator under proportional allocation and the other calibration estimators obtained based on the information of the single auxiliary variable ($\bar{X}_1$ = Area) as mentioned above. Table 3.4 in Section 3.4 presents the calibrated weights for all the different methods. As discussed earlier the true average production $\left(\bar{Y}\right)$ of the tobacco crop for this population is found to be: $\bar{Y} = 52444.56$.

In Table 3.5, the Columns 2 and 3 show, respectively, the estimated average production $\left(\hat{\bar{Y}}\right)$ of tobacco for different estimators and the sampling error of the estimators. Whereas, the Columns 4 and 5 show the estimated variance $\hat{v}\left(\hat{\bar{Y}}\right)$ and the RE of the estimators. The results reveal that the proposed multivariate calibrated estimator further improves the result as compared to the results of all univariate calibrated estimators. The gain in efficiency of the proposed estimator over the stratified sampling estimator is 550.96%.

Table 3.5: Analysis of tobacco population data

Estimator	$\hat{\bar{Y}}$	$SE = \left\lvert\bar{Y} - \hat{\bar{Y}}\right\rvert$	$\hat{v}\left(\hat{\bar{Y}}\right)$	RE
Stratified estimator, $\bar{y}_{st}$	95373.90	42929.34	2822731121.0	100.0
Singh et al. (1998), $\bar{y}_{st}^{(1)}$	54329.76	1885.20	650863727.5	433.69
Singh (2003), $\bar{y}_{st}^{(2)}$	54132.89	1688.33	569781129.3	495.41
Tracy et al. (2003), $\bar{y}_{st}^{(3)}$	54321.02	1876.46	609889505.3	462.83
Proposed univariate estimator, $\bar{y}_{st}^{(4)}$	53775.78	1331.22	517799741.5	545.14
Proposed multivariate estimator, $\bar{y}_{st}^{*}$	53585.53	1140.97	512333228.2	550.96

B. Simulated Data

In this subsection the performance of the proposed multivariate estimator was compared with the other calibration estimators using a simulated data. The data for the auxiliary variable (X_1) and (X_2) was randomly generated using the R software for a normal population of size N = 1000 with a mean of 50 and a standard deviation of 5.2. Then, the data for the survey variable (Y) was generated from X using the linear function of $Y_i = AX_i + B$, where the values of the coefficients A and B are 1.2 and 701.0, respectively.

This simulated population is divided into $L = 4$ strata. Suppose that an estimate of population mean $\overline{Y}$ is of interest using two auxiliary variables, X_1 and X_2. Using proportional allocation a sample of $n = 100$ was selected from the four strata. To compute the calibrated weights, the information needed is summarized in Table 3.6 below.

Table 3.6: Information from simulated data

h	N_h	W_h	n_h	$\overline{x}_{h1}$	$\overline{x}_{h2}$	d_h	$\overline{y}_h$	s_{h1}^2	s_{h2}^2
1	319	0.319	27	43.95871	122.36840	0.034	753.75050	7.32511	33.15712
2	362	0.362	25	49.86476	125.31160	0.037	761.03770	0.07830	29.64958
3	220	0.220	21	54.45950	127.96450	0.043	766.32650	0.07703	32.95809
4	99	0.099	27	59.09469	122.41080	0.027	771.90000	0.07709	26.40835

The known means $(\overline{X}_j)$ of the auxiliary variables for this population are found to be $\overline{X}_1 = 49.89973$ and $\overline{X}_2 = 124.77390$, and the known stratum variances (S_{hj}^2) are given as shown in Table 3.7.

***Table 3.7: Stratum* variances of X_1 and X_2 for simulated data**

H	S_{h1}^2	S_{h2}^2
1	7.3021	938.4589
2	2.1481	851.6338
3	1.7881	784.3973
4	2.8738	914.7062

Then, solving equations (3.10)-(3.15) using MATHEMATICA the Lagrange multipliers are obtained as:

$$\lambda_0 = -14.32942, \lambda_1 = -0.00330, \lambda_2 = 0.12444, \varphi_1 = 1.49439, \varphi_2 = -0.000146.$$

Hence, equation (3.16) gives the optimum calibrated weight of the proposed estimator, which is presented in Column 2 in Table 3.8. The last three columns of the table show the weights of other calibrated estimators.

Table 3.8: Calibrated Weights using simulated data

h	W_h^*	$W_h^{*(1)}$	$W_h^{*(2)}$	$W_h^{*(3)}$	$W_h^{*(4)}$
1	0.52928	0.52892	0.31897	0.31943	0.31900
2	0.00400	0.00500	0.36196	0.36200	0.36195
3	0.24683	0.24658	0.21997	0.21977	0.21996
4	0.21990	0.21949	0.09899	0.09870	0.09898
Total	1	1	0.99989	1	0.99989

Using (2) the proposed calibration estimator is obtained as:

$$\bar{y}_{st}^* = \sum_{h=1}^{L} W_h^* \bar{y}_h = 760.88240. \qquad \text{…(3.18)}$$

Moreover, as discussed in Chapter 2, the true population mean of the simulated data $\overline{Y} = 760.87968$. Table 3.9 presents the results for the simulated data. It reveals the similar results as found in the real data. Thus, the proposed multivariate calibrated estimator is most efficient as compared to other estimators and the gain in efficiency of the proposed estimator over the stratified estimator is 386.85%.

Table 3.9: Analysis of simulated data

Estimator	$\hat{\overline{Y}}$	$SE = \left\|\overline{Y} - \hat{\overline{Y}}\right\|$	$\hat{v}\left(\hat{\overline{Y}}\right)$	RE
Stratified estimator, $\overline{y}_{st}$	760.95199	0.07231	45844670263.9	100
Singh et al. (1998), $\overline{y}_{st}^{(1)}$	760.86745	0.01222	45835552384.3	100.02
Singh (2003), $\overline{y}_{st}^{(2)}$	760.86902	0.01065	45745087574.3	100.22
Tracy et al. (2003), $\overline{y}_{st}^{(3)}$	760.86787	0.01181	45832943300.4	100.03
Proposed univariate estimator, $\overline{y}_{st}^{(4)}$	760.87151	0.00817	11870824590.0	386.20
Proposed multivariate estimator, $\overline{y}_{st}^{*}$	760.88240	0.00272	11850624683.1	386.85

Chapter 4

Conclusion

In surveys, the surveyors are often interested in the greater precision of the estimate of the characteristics that involves a continuous research carried out by the researchers to improve the precision in the estimates. Stratified random sampling is one of the most systematic and frequently used sampling techniques in surveys to estimate a population statistic. Moreover, the precision of the estimate can further be improved, if the auxiliary information of the study variable available is used. Thus, calibration estimation that incorporates the auxiliary information is a widely used technique in survey sampling to improve the precision and the accuracy of the estimators of population parameter.

Therefore, in this thesis, we designed a technique to determine the optimum calibrated weights and the optimum calibrated estimators using the information available, respectively, from single and multiple auxiliary variables in stratified random sampling. In the proposed calibration estimators, the auxiliary information used is not only the population means but also the population stratum variances available for auxiliary variables. The problem is formulated as Nonlinear Programming Problem (NLPP) that seeks minimization of the chi-square distance function, subject to the available calibration constraints. The NLPP is then solved by developing solution procedure using Lagrange multiplier technique.

Numerical examples with real data using the univariate and multivariate auxiliary cases are presented to illustrate the computational details of the proposed techniques. A comparison study with real and simulated data is carried out to determine the performance of the proposed estimators based on univariate and multivariate auxiliary information. The results for univariate case reveal that the calibrated estimator using the proposed technique may be useful for the surveyors as it performs better than the other calibration estimators in stratified sampling. Furthermore, as it has been discussed in literature that the precision of the survey estimates is always improved further when multiple auxiliary information is available, it is very evident in this thesis.

Therefore, this research will be very useful for survey researchers to get consistent estimates and provide more precise estimates of population parameters. However, one of the major problems encountered along this area of research was dealing with negative weights. Future researchers can address this problem whereby positive weights are being guaranteed.

Bibliography

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

[2] Arnab, R. and Singh, S., (2005). *A note on variance estimation for the generalized regression predictor.* Australia and New Zealand Journal of Statistics. 47(2): 231-234

[3] Baillargeon, S. and Rivest, L.P. (2011). *The construction of stratified designs in R with the package stratification.* Survey Methodology, 37 (1), 53-65.

[4] Bazaraa, M. S., Sherali, H. D. and Shetty, C. M. (1993). *Nonlinear Programming: Theory and Algorithms (Second Edition).* New York: John Wiley and Sons, Inc.

[5] Briedt, F.J. and Opsomer, J.D. (2000). *Local polynomial regression estimators in survey sampling.* Ann. Statist., 28, 1026-1053.

[6] Chatterji, S. (1968). *Multivariate Stratified Surveys.* J. Amer. Stat. Assc, 63, 530-534.

[7] Chen, J. and Qin, J. (1993). *Empirical likelihood estimation for finite populations and the effective usage of auxiliary information.* Biometrika 80, 107-116.

[8] Cochran, W. G. (1977). *Sampling techniques (3rd ed.).* John Wiley and Sons, Inc., New York.

[9] Dalenius, T. (1957). *Sampling in Sweden: Contribution to the Methods and theories of sample survey practice.* Almqvist och Wiksell, Stockholm.

[10] Deville, J.C, and Särndal, C.E.(1992). *Calibration Estimators in Survey Sampling.* J.Amer.Statist.Assoc., 87, 376-382.

[11] Durbin, J. (1959): *Review of Sampling in Sweden. J. Roy. Statist.* Soc. (A) 122, 146-148.

[12] Estevao, V.M. and Sarndal, C.E., *Survey estimates by calibration on complex auxiliary information.* International Statistical Review, 74(2006): 127-147.

[13] Fuller, W.A. (2009). *Sampling Statistics*. John Wiley & Sons, Inc.

[14] Ghosh, S. P. (1958). *A Note on Stratified Random Sampling with Multiple Characters.* Cal. Stat. Bull. 8, 81-89.

[15] 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.

[16] Hartley, H.O. & Rao, J.N.K. (1968). *A new estimation theory for sample surveys.* Biometrika, 55, 547-555.

[17] Hess, I., Sethi, V. K. and Balakrishnan, T. R. (1966). *Stratification: A practical investigation.* Jour. Amer. Stat. Assoc., 61, 74-90.

[18] Horvitz, D. G. & Thomas, D. J. (1952). *A Generalization of Sampling Without Replacement From A Finite Universe.* Journal of the American Statitical Association, Vol. 47, No. 260, 663 – 685.

[19] Khan, M.G.M., Maiti, T. and Ahsan, M.J. (2010). *An Optimal Multivariate Stratified Sampling Design using Auxiliary Information: An Integer Solution using Goal Programming Approach*. Journal of Official Statistics, Vol. 26(4), 695-708.

[20] Kim, J.K. (2009) *Calibration estimation using empirical likelihood in unequal probability sampling*. Statist. Y Sinnica., 19, 145-157.

[21] Kim, J.K., (2010). *Calibration estimation using exponential tilting in sample surveys*. Statistics Canada, Catalogue No. 12-001, Vol. 36, 2, 145-155.

[22] Kim, J.M., Sungur, E.A. and Heo, T.Y. (2007). *Calibration Approach Estimators in Stratified Sampling.* Statistics & Probability Letters; Vol. 77, 1, 99-103.

[23] Koyuncu, N., Kadilar, C. (2014). *Calibration Weighting In Stratified Random Sampling.* Communications in Statistics-Simulation and Computation, DOI:10.1080/03610918.2014.901354.

[24] Kuhn, H. W. and Tucker, A. W. (1951). *Nonlinear Programming.* Proceedings of the Second Berkley symposium on Mathematical Statistics and Probability, University of California Press, Berkley, 481-492.

[25] Lavallee, P. (1987). *Some contributions to optimal stratification.* Master Thesis, Carleton University, Ottawa, Canada.

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

[27] Montanari,G. E. & Ranalli, M. G. (2005). *Journal of the American Statistical Association.* Vol. 100, N0. 472, 1429 – 1442.

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

[29] Ranalli, M.G., (2008). *Recent Developments in Calibration Estimation.* Recenti sviluppi nella stima a ponderazione vincolata, (2008), 355-362.

[30] Rao, D., Khan, M.G.M. and Khan, S. (2012). *Mathematical Programming on Multivariate Calibration Estimation in Stratified Sampling*. World Academy of Science, Engineering and Technology 72, 58-62.

[31] Rao, J.N.K. (1994). *Estimating totals and functions using auxiliary information at the estimation stage.* J. Official Stat., 10, 153-165.

[32] Richard, L. S., William, M. and R. Layman. (2006). *Elementary Survey Sampling, (6^{th} Edition).* Thomson Brooks, 117-153.

[33] Rueda, M., Martinez, S., Arcos, A., Martinez, H. & Munoz, J.F. (2010). *On Determining the Calibration Equations to Construct Model – Calibration estimators of the Distribution function.* Vol. 25, 1, 87 – 95.

[34] Sarndal, C.-E., (2007). *The calibration approach in survey in theory and practice*, Statistics Canada, Catalogue No. 12-001, Vol. 33, 2, 99-119.

[35] Singh, A.C. & Mohl, C.A. (1996). *Understanding calibration estimators in survey sampling.* Sur. Methodol., 22, 107-115.

[36] Singh, Gurmindar; Rao, Dinesh and Khan, M. G. M. (2014) "*Calibration Estimator of Population Mean in Stratified Sampling*", IEEE Proceedings of 2014 Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE), 1-5, DOI: 10.1109/APWCCSE.2014.7053875

[37] Singh, S. (2003). *Advanced Sampling Theory with Applications*. Dordrecht: Kluwer Academic Publishers.

[38] Singh, S., Horn, S. and Yu, F. (1998). *Estimation of variance of the general regression estimator: higher level calibration approach,* Survey Methodology 24, 41–50.

[39] Sukhatme, P. V., Sukhatme, B. V., Sukhatme, S. & Asok, C. (1984). *Sampling Theory of Surveys with Applications.* Iowa State University Press, Ames, IA.

[40] Tracy, D.S., Singh, S. and Arnab, R., (2003). *Note on Calibration in Stratified and double Sampling*. Survey Methodology 29, 99-104.

[41] Theberge, A. (1999). *Extension of Calibration Estimators in Sampling.* JASA, 94, 635-644. http://dx.doi.org/10.1080/01621459.1999.10474157.

[42] Theberge, A. (2000). *Calibration and restricted weights.* Survey Methodology, 26, 99-107.

[43] Wittenberg, M., (2009). *Sample Survey Calibration: An Information-theoretic perspective.* A Southern Africa Labour and Development Research Unit Working Paper, 41.

[44] Wu, C. & Sitter, R.R. (2001). *A model-calibration approach to using complete auxiliary information from survey data.* J. Amer. Statist. Assoc., 96, 185-193.

Appendix A

Mathematica Code for finding $\lambda_0, \lambda_1, \varphi_1$ for univariate calibration estimator

```
L=10;

(*h=1*)

xbar[1]=1304.7;

w[1]=0.05660;

s[1]=722185.3;

d[1]=0.1667;


(*h=2*)

xbar[2]=29075.0;

w[2]=0.05660;

s[2]=839008125.0;

d[2]=0.1667;

(*h=3*)

xbar[3]=5191.7;

w[3]=0.07547;

s[3]=74387858.3;

d[3]=0.2083;

(*h=4*)
```

```
xbar[4]=21700.0;

w[4]=0.09434;

s[4]=6070000.0;

d[4]=0.2333;

xbar[5]=6808.0;

w[5]=0.11321;

s[5]=63572981.3;

d[5]=0.1667;

xbar[6]=1800.0;

w[6]=0.03774;

s[6]=1620000;

d[6]=0.2500;


xbar[7]=24481.5;

w[7]=0.28302;

s[7]=1801653230.3;

d[7]=0.0576;


xbar[8]=294809.2;

w[8]=0.16038;

s[8]=322774101004.2;

d[8]=0.1078;

xbar[9]=6303.7;
```

w[9]=0.09434;

s[9]=59939890.3;

d[9]=0.2333;

xbar[10]=350;

w[10]=0.02830;

s[10]=125000;

d[10]=0.1667;

XBAR=34438.61;

V=2364494472;

Find Root[{

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar[h]+\varphi_1 w[h]d[h]s[h])-1,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar[h]+\varphi_1 w[h]d[h]s[h])xbar[h]-XBAR,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar[h]+\varphi_1 w[h]d[h]s[h])d[h]s[h]-V,$$

a = 0.001;

b = 0.001;

c = 0.001;

Table[$w[h]+\lambda_0 w[h]+\lambda_1 w[h] xbar[h]+\varphi_1 w[h] d[h] s[h], \{h,1,L\}$]

Appendix B:

Mathematica Code for finding $\lambda_0, \lambda_1, \lambda_2, \varphi_1, \varphi_2$ for multivariate calibration estimator

```
L=10;

(*h=1*)

xbar1[1]=1304.66667;

s1[1]=722185.33333;

xbar2[1]=1.94;

s2[1]=0.0171;

w[1]=0.05660;

d[1]=0.167;


(*h=2*)

xbar1[2]=29075.0;

s1[2]=839008125.0;

xbar2[2]=1.37667;

s2[2]=0.47573;
```

```
w[2]=0.05660;
d[2]=0.167;
(*h=3*)
xbar1[3]=5191.66667;
s1[3]=74387858.33333;
xbar2[3]=2.79333;
s2[3]=0.801033;
w[3]=0.07547;
d[3]=0.208;

(*h=4*)
xbar1[4]=21700.0;
s1[4]=6070000.0;
xbar2[4]=1.44333;
s2[4]=0.736233;
w[4]=0.09434;
d[4]=0.233;

xbar1[5]=6808.0;
s1[5]=63572981.3;
```

```
xbar2[5]=1.78750;

s2[5]=1.06983;

w[5]=0.11321;

d[5]=0.167;


xbar1[6]=1800.0;

s1[6]=1620000;

xbar2[6]=1.7850;

s2[6]=0.061250;

w[6]=0.03774;

d[6]=0.250;


xbar1[7]=24481.5;

s1[7]=1801653230.3;

xbar2[7]=1.32091;

s2[7]=0.48245;

w[7]=0.28302;

d[7]=0.058;


xbar1[8]=294809.2;
```

s1[8]=322774101004.2;

xbar2[8]=1.32000;

s2[8]=0.24616;

w[8]=0.16038;

d[8]=0.108;

xbar1[9]=6303.7;

s1[9]=59939890.3;

xbar2[9]=1.32667;

s2[9]=0.03063;

w[9]=0.09434;

d[9]=0.233;

xbar1[10]=350;

s1[10]=125000;

xbar2[10]=1.765;

s2[10]=1.32845;

w[10]=0.02830;

d[10]=0.167;

XBAR1=34438.613;

XBAR2=1.5508;

V1 = 2364494472;

V2=0.086540631;

FindRoot[{

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h])-1,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h])xbar1[h]-XBAR1,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h])xbar2[h]-XBAR2,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h])d[h]s1[h]-V1,$$

$$\sum_{h=1}^{L}(w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h])d[h]s2[h]-V2,$$

a = 0.001;

b = 0.001;

c = 0.001;

Table[

$$w[h]+\lambda_0 w[h]+\lambda_1 w[h]xbar1[h]+\lambda_2 w[h]xbar2[h]+\varphi_1 w[h]d[h]s1[h]+\varphi_2 w[h]d[h]s2[h],\{h,1,L\}]$$