Stratification of Skewed Populations

Patricia Gunning

B.Sc.

A thesis submitted in fulfillment of the

requirements for award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University

Faculty of Engineering and Computing

School of Computing

Supervisor: Dr. Jane M. Horgan

September, 2006

DECLARATION

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Patricia Gunning

(Patricia Gunning)

Student ID: 96153539

Date: 22nd September 2006

Acknowledgements

A whole host of people provided inspiration and motivation during this study. Top of the list is my supervisor Dr. Jane M. Horgan for her constant advice and guidance, considered insights and enthusiasm. She has taught me that scientific research can not only be rewarding but also fun. I know of few people as intelligent, caring, ethical, patient or energetic as Jane. She has always given generously of her time. I feel deeply grateful and privileged to have been her student. Thank you so much.

Of course, I could never have completed this work without funding and I would like to acknowledge the Irish Research Council for Science, Engineering and Technology. This work was also part funded by DCU School of Computing during the first year.

I wish to thank Mr. Gary Keogh for his advice and ideas. I also wish to thank Dr. William Yancey, CPA, a tax audit sampling consultant from Texas, who alerted us to the problem of stratification in statistical auditing in the first instance.

Special thanks goes to my hugely supportive colleague and friend Dr. Noreen Quinn for her constant encouragement and advice. A warm thanks to Niall and Karl, for their words of wisdom and inspiring discussions. A debt of gratitude is also extended to Dr. Adel Sharkasi and Dr. Yaw Bimpeh for their endless support and advice. I would like to thank my other fellow postgraduate students Ana, Puspita, Grainne, Ashley, Dimitri, Tommy, George, Justin, Georgina, David, Noel, Michelle, Claire and Fabrice to name but a few! I would also like to thank all the staff and postgraduates at the School of Computing for their helpful comments whenever I presented my work. Also a big thank you to Martina, Marita, Barbara

and Sebastian for your encouragement.

I would like to extend my love and gratitude to my family. To my sister Deirdre and my brothers Brendan and Seamus, thanks for all your support. Special thanks and appreciation must go to my mother for her good advice, constant support and encouragement without which I would not be where I am today. Despite my father having passed away 13 years ago, I still treasure the values he has instilled in me.

St. Anthony and St. Jude for answered prayers.

Thank you all so much.

Patricia Gunning, 22/09/2006

Dedicated to

my mother, Kathleen and the memory of my father, James

my mentors.

CONTENTS

List of Figures

List of Tables

Publications Related to this Study

Journal Publications

Gunning, P., Horgan, J.M. and Keogh, G. (2006). Efficient Pareto Stratification. *The Mathematical Proceedings of the Royal Irish Academy*, 2 (to appear).

Gunning, P. and Horgan, J.M. (2006). Improving the Lavallée-Hidiroglou Algorithm for Stratification of Skewed Populations. *Journal of Statistical Computation and Simulation* (to appear).

Gunning, P., Horgan, J.M. and Yancey, W. (2004). Geometric Stratification of Accounting Data. *Contaduria y Administracion*, 214, 11-21.

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

Gunning, P. and Horgan J.M. (2004). Un Nouvel Algorithme pour la Construction de Bornes de Stratification dans les Populations Asymetriques. *Techniques dEnquete*, 30, 2, 177-185.

Conference Proceedings

Gunning, P. and Horgan, J.M. (2004). An Algorithm for Obtaining Strata with Equal Coefficients of Variation. *Proceedings of Computational Statistics 2004, 16th Symposium of IASC, Prague, Czech Republic*, Physica-Verlag. 1123-1129.

Gunning, P., Horgan, J.M. and Keogh, G. (2004). Obtaining Stratum Boundaries in Skewed Populations. *Proceedings of the Joint Statistical Meeting*, Toronto, Canada. 3620-3626.

Stratification of Skewed Populations

Patricia Gunning

Dublin City University,

Glasnevin,

Dublin 9.

Abstract

In this research an algorithm is derived for stratifying skewed populations which is much simpler to implement than any of those currently available. It is based on the suggestion by numerous researchers in the field that it is desirable when stratifying skewed populations to arrange for equal coefficients of variation in each subinterval. Our new algorithm makes the breaks in geometric progression and achieves near-equal stratum coefficients of variation when the populations are skewed. Simulation studies on real skewed populations have shown that the new method compares favourably to those commonly used in terms of precision of the estimator of the mean.

We also apply the geometric method to the Lavallée-Hidiroglou (1988) algorithm, an iterative method designed specifically for skewed populations. We show that by taking geometric boundaries as the starting points results in most cases in quicker convergence of the algorithm and achieves smaller sample sizes than the default starting points for the same precision.

Finally, geometric stratification is applied to the Pareto distribution, a typical model of skewed data. We show that if any finite range of this distribution is broken into a given number of strata, with boundaries obtained using geometric progression, then the stratum coefficients of variation are equal.

CHAPTER 1

INTRODUCTION

1.1 Introduction

A study in which every unit of the population is examined is time-consuming, expensive, often impossible and inaccurate. Summary statistics from a sample are often used to make extrapolations concerning the entire population. The main challenges in sampling are:

- how to select sample units which are cost-effective and representative of the population of interest;
- how to process the raw data into estimates of population parameters of interest and evaluate the precision of these estimates.

There are many sampling methods. Simple random sampling is a method of selecting n units from a population of N units such that every one of the ${}_NC_n$ distinct samples has an equal chance of being drawn. However, other methods of sampling are often preferable to simple random sampling on the grounds of convenience or of increased precision. Stratification is one such method, and this is the focus of this research.

In the remainder of this chapter:

(i) Stratified sampling is overviewed (1.2)

(ii) The choice of stratification variable is outlined (1.3)

(iii) The number of strata is discussed (1.4)

(iv) Sample allocation is overviewed (1.5)

(v) The construction of stratum boundaries is outlined (1.6)

(vi) The objectives of the study are stated (1.7)

(vii) The limitations of the study are explained (1.8)

(viii) An overview of the remaining chapters is provided (1.9).

1.2 Stratified Sampling

A stratified random sampling design is a sampling plan in which a population is divided into mutually exclusive strata or subgroups and simple random samples are drawn from each stratum independently.

Stratification is a commonly used sampling technique which:

1. allows separate estimates for each stratum.

2. improves precision. As Cochran (1977, p89) points out "it may be possible to divide a heterogenous population into subpopulations, each of which is internally homogeneous $\cdots$ If each stratum is homogeneous, in that the measurements vary little from one unit to another, a precise estimate of any stratum mean can be obtained from a small sample in that stratum. These estimates can then be combined into a precise estimate for the whole population."

The main objective of stratification is to construct strata to allow for efficient estimation of the quantity to be measured in the survey. For example, in the case of the stratified mean estimate, to minimise its variance for a fixed sample size or to minimise the sample size for a fixed variance of the stratified mean estimate.

1.2.1 Stratification of a Finite Population

Suppose there are L strata containing N_h units from which a sample of size n_h is to be chosen independently from each stratum ($1 \leq h \leq L$) using simple random sampling. We write the population size as $N = \sum_{h=1}^{L} N_h$ and total sample size as $n = \sum_{h=1}^{L} n_h$. The values obtained for any specific unit in the N units that comprise the population are denoted by $y_1, y_2, \ldots, y_N$. The corresponding values for the units in the sample are denoted by $y_1, y_2, \ldots, y_n$, or if we wish to refer to a typical sample member by y_i ($i = 1, 2, \ldots, n$). Note that the sample will not consist of the first n units in the population, except in the rare instance in which these units happen to be drawn.

The overall population mean is:

$$\overline{Y} = \frac{1}{N} \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} \tag{1.2.1}$$

where y_{hi} is the i^{th} unit in the h^{th} stratum. This population mean may also be written as:

$$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h \tag{1.2.2}$$

where

$$\overline{Y}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} y_{hi}, \tag{1.2.3}$$

is the mean of the units in the h^{th} stratum and

$$W_h = \frac{N_h}{N} \tag{1.2.4}$$

is the stratum weight, i.e. the proportion of population units falling in stratum h.

The overall population variance is

$$S^2 = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} \left(y_{hi} - \overline{Y}\right)^2}{N-1} \tag{1.2.5}$$

and the variance of the units in the h^{th} stratum is

$$S_h^2 = \frac{\sum_{i=1}^{N_h} \left(y_{hi} - \overline{Y}_h\right)^2}{N_h - 1}. \tag{1.2.6}$$

An estimate of the population mean is formed by combining the separate stratum sample means using weights W_h. The stratified mean estimate is defined as:

$$\overline{y}_{st} = \sum_{h=1}^{L} W_h \overline{y}_h, \tag{1.2.7}$$

where

$$\overline{y}_h = \frac{1}{n_h} \sum_{i=1}^{n_h} y_{hi} \tag{1.2.8}$$

is the mean of the sample units in the h^{th} stratum with y_{hi} being the i^{th} unit of the sample chosen in the h^{th} stratum.

Note, it is easy to show that $\overline{y}_{st}$, defined in equation (1.2.7), is an unbiased estimator of the population mean $\overline{Y}$. Since

$$E(\overline{y}_h) = \overline{Y}_h,$$

then

$$E(\overline{y}_{st}) = \sum_{h=1}^{L} W_h E(\overline{y}_h) = \sum_{h=1}^{L} W_h \overline{Y}_h = \overline{Y}.$$

The variance of the stratified mean is:

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 V(\overline{y}_h). \tag{1.2.9}$$

Now since $\overline{y}_h$ is the mean of a simple random sample drawn from the h^{th} stratum containing N_h units then

$$V(\overline{y}_h) = \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}. \tag{1.2.10}$$

It follows that

$$V(\overline{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}. \tag{1.2.11}$$

Also,

$$f_h = \frac{n_h}{N_h} \tag{1.2.12}$$

is the sampling fraction in stratum h and

$$fpc_h = 1 - \frac{n_h}{N_h} \tag{1.2.13}$$

is the finite population correction factor for stratum h.

When the population is finite, the finite population correction factor (1.2.13) is used in the variance. Some researchers such as Dalenius and Hodges (1959); Ekman (1959); Sethi (1963) and Serfling (1968) have made the assumption that the finite population correction can be ignored. This assumption is plausible provided the sampling fractions in the strata (1.2.12) are low, making (1.2.13) close to unity, and so the size of the population as such has no affect on the variance of the sample

estimate. The variance (1.2.11) can then be written as:

$$V(\overline{y}_{st}) \approx \sum_{h=1}^{L} \frac{W_h^2 S_h^2}{n_h}. \tag{1.2.14}$$

The coefficient of variation is a measure of dispersion relative to the mean, and is defined as:

$$cv = \frac{S}{\overline{Y}}. \tag{1.2.15}$$

The coefficient of variation of stratum h is written as:

$$cv_h = \frac{S_h}{\overline{Y}_h}, \tag{1.2.16}$$

and the coefficient of variation of the stratified sample mean $\overline{y}_{st}$ is:

$$cv(\overline{y}_{st}) = \frac{\sqrt{V(\overline{y}_{st})}}{\overline{Y}}. \tag{1.2.17}$$

The coefficient of skewness measures the degree of asymmetry of a distribution. The overall population coefficient of skewness is:

$$\eta_3 = \sqrt{N-1}\frac{\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}\right)^3}{\left(\sum_{h=1}^{L}\sum_{i=1}^{N_h}\left(y_{hi}-\overline{Y}\right)^2\right)^{3/2}}. \tag{1.2.18}$$

The skewness for a normal distribution is zero and any symmetric data should have a skewness near zero. If the coefficient of skewness > 1, the distribution is said to be positively skewed. If the coefficient of skewness < 1, the distribution is said to be negatively skewed.

1.2.2 Stratification of Continuous Data

In addressing the problem of stratification, some researchers such as Dalenius (1950) have assumed for convenience that the discrete distribution can be approximated by a continuous distribution with density $f(y)$. With continuous variables, it is the

convention to designate the population parameters with Greek letters.

The overall population mean is defined as:

$$\mu = \int_{-\infty}^{\infty} yf(y)dy. \tag{1.2.19}$$

This is also referred to as the first moment about zero where the r^{th} moment about zero is defined as:

$$\mu_r^{'} = \int_{-\infty}^{\infty} y^r f(y)dy. \tag{1.2.20}$$

The mean of the units in the h^{th} stratum is written as:

$$\mu_h = \int_{k_{h-1}}^{k_h} \frac{yf(y)dy}{W_h} \tag{1.2.21}$$

where k_h $(1 \leq h \leq L)$ are the stratum boundaries, and the stratum weight is

$$W_h = \int_{k_{h-1}}^{k_h} f(y)dy. \tag{1.2.22}$$

The overall population variance is defined as:

$$\sigma^2 = \int_{-\infty}^{\infty} (y-\mu)^2 f(y)dy \tag{1.2.23}$$

and is the second moment about the mean.

The variance for y values in stratum h is

$$\sigma_h^2 = \int_{k_{h-1}}^{k_h} \frac{(y-\mu_h)^2 f(y)dy}{W_h}. \tag{1.2.24}$$

The overall population coefficient of variation is defined as:

$$cv = \frac{\sigma}{\mu} \tag{1.2.25}$$

and the coefficient of variation of stratum h is:

$$cv_h = \frac{\sigma_h}{\mu_h}. \tag{1.2.26}$$

Generally, the r^{th} moment about the mean is defined as:

$$\mu_r = \int_{-\infty}^{\infty} (y - \mu)^r f(y) dy. \tag{1.2.27}$$

The third standardized moment about the mean is the coefficient of skewness and is defined as:

$$\eta_3 = \frac{\mu_3}{\sigma^3}. \tag{1.2.28}$$

This study concentrates on populations with high positive skewness. As Hess et al. (1966) pointed out, the importance of stratification increases as asymmetry and the variability in stratum sizes and stratum means increases.

1.3 Choice of Stratification Variable

Ideally the division of the population into strata should be based on the survey variable $y = y_1, y_2, \ldots, y_N$. Such a construction is of course not possible since y is unknown; if it were known we would not need to estimate it. Therefore, stratification needs prior knowledge of an auxiliary variable, $x = x_1, x_2, \ldots, x_N$ which is strongly correlated with the survey variable, y, and in business situations, such a variable is often readily available. For example, in auditing, book values may be used as the auxiliary variable which is highly correlated to the survey variable, the unknown audit values. Assuming that the values of x and y are strongly correlated, the simplest model to use is $x = y$. Although this assumption is unrealistic and researchers such as Rivest (2002) have attempted to account for the discrepancy between x and y using a regression model, it is widely used in practice (Hedlin, 1998). This is the model used in this study.

1.4 Number of Strata

Regarding the number of strata L to be constructed, in some cases the number is predetermined as with categorical variables such as geographic subdivisions, gender, classes in a university, etc. With continuous variables, on the other hand, such as wages, height, financial data, etc., it is necessary to decide on break points, k_h, along a range of the variable.

There are two issues to consider regarding the number of strata. One is the rate of decrease in the $V(\overline{y}_{st})$ given in (1.2.11) when L is increased, that is the ratio of the variance for L strata to the variance for $L-1$ strata i.e.

$$\frac{V_L(\overline{y}_{st})}{V_{L-1}(\overline{y}_{st})},$$

and how the cost of the survey is affected by an increase in the number of strata (Cochran, 1977, p132). It is expected that $V(\overline{y}_{st})$ decreases as the number of strata increases. However, this decrease, though substantial for initial increases in the number of strata, becomes marginal after a certain stage. Cochran (1977, p133) concluded that unless the correlation between x and y exceeds 0.95, little reduction in variance is to be expected beyond $L = 6$. With regard cost, it is often the case that little is gained from increasing L beyond 6 if the increase necessitates any substantial decrease in n in order to keep the cost constant (Cochran, 1977, p134).

1.5 Sample Allocation

There are various ways of allocating the sample of size n among L strata.

1.5.1 Equal Allocation

A very simple way to allocate the sample is to take an equal number of units from each stratum where

$$n_h = \frac{n}{L}. \tag{1.5.1}$$

For this equal allocation, the variance given in equation (1.2.14) becomes:

$$V_{eq}(\overline{y}_{st}) = \frac{L}{n}\sum_{h=1}^{L} W_h^2 S_h^2. \tag{1.5.2}$$

This allocation method takes no account of the number of units, N_h, or the variability S_h in each stratum and may be inefficient if the N_h or S_h differ substantially.

1.5.2 Proportional Allocation

A more logical allocation would be to allocate n_h proportional to stratum size N_h where

$$n_h = n\left(\frac{N_h}{N}\right). \tag{1.5.3}$$

Proportional allocation has the advantage that each unit in the sample has the same weight, that is each unit in the sample represents the same number of units in the population (Lohr, 1999, p104), and so the mean is self weighting, i.e.

$$\overline{y}_{st} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} y_{hi}}{n}.$$

For proportional allocation, the variance given in equation (1.2.14) becomes:

$$V_{prop}(\overline{y}_{st}) = \sum_{h=1}^{L} \frac{W_h S_h^2}{n}. \tag{1.5.4}$$

Proportional allocation simplifies the amount of bookkeeping involved in data processing and reduces computational expenses (Levy and Lemeshow, 1999, p154). "If the variances S_h^2 are more or less equal across all the strata, proportional allo-

cation is probably the best allocation for increasing precision" (Lohr, 1999, p106).

1.5.3 Optimum Allocation

In cases where the S_h^2 vary greatly, as with skewed populations, proportional allocation is an inefficient allocation of resources. As the larger units are likely to be more variable than the smaller units, these larger units should be sampled at a higher rate.

Since the objective of sampling is to gain the most information for the least cost, units should be allocated to strata in order to minimise $V(\overline{y}_{st})$ for a given total cost C or equivalently to minimise C for a fixed $V(\overline{y}_{st})$. The simplest form of the cost function would be for example,

$$C = c_0 + \sum_{h=1}^{L} c_h n_h \tag{1.5.5}$$

where c_0 is the fixed overhead cost and c_h is the cost of sampling a unit in the h^{th} stratum (Cochran, 1977, p96).

It is easy to show that $V(\overline{y}_{st})$ is minimised for fixed C when sample sizes n_h are choosen so that they are directly proportional to N_h and S_h and inversely proportional to the square root of cost c_h i.e.

$$n_h = \left(\frac{N_h S_h / \sqrt{c_h}}{\sum_{i=1}^{L} N_i S_i / \sqrt{c_i}} \right) n. \tag{1.5.6}$$

This type of allocation is called optimum allocation.

From (1.5.6) we see that optimum allocation leads to taking a large sample from a given stratum if the stratum is larger, more variable internally or sampling is cheaper in the stratum. One disadvantage of optimum allocation compared to proportional allocation is that the sample mean is not self weighting.

1.5.4 Neyman Allocation

For the special case where the cost of sampling a unit is the same for each stratum, optimum allocation of n sample units is given by

$$n_h = \left(\frac{N_h S_h}{\sum_{i=1}^{L} N_i S_i} \right) n. \tag{1.5.7}$$

This allocation is sometimes called Neyman allocation, after Neyman (1934). For optimum allocation, the approximate variance given in equation (1.2.14) is:

$$V_{opt}(\overline{y}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h \right)^2}{n}. \tag{1.5.8}$$

One problem that may be encountered with optimal or Neyman allocation is that the optimal sample size n_h may be greater than N_h. When this occurs, the standard solution, (Levy and Lemeshow, 1999, p163), is to set n_h equal to N_h for each stratum having optimal allocation greater than N_h. The remaining sample is then reallocated to other strata as specified by the algorithm for obtaining optimal allocation (Levy and Lemeshow, 1999, p163).

1.5.5 Power Allocation

Power allocation has been used in the design of several surveys at Statistics Canada (Bankier, 1988). Lavallée and Hidiroglou (1988) used power allocation which allocates stratum sample sizes as:

$$n_h = \left(\frac{(N_h \overline{Y}_h)^p}{\sum_{i=1}^{L} (N_i \overline{Y}_i)^p} \right) n, \tag{1.5.9}$$

where $0 < p \leq 1$ is the power of the allocation. According to Lavallée and Hidiroglou (1988)

"power allocations have the particularity that under relatively simple assumptions and for a suitable choice of p, the coefficients of variation for ... strata tend to be equalised without a significant increase in the overall coefficient of variation. This equality of coefficients of variation is often asked by the users of the survey data."

1.6 Boundaries

While Dalenius (1950) derived equations for determining boundaries so that the variance of the sample mean is minimised, these equations proved troublesome to solve because of dependencies among the components. Since then there have been many attempts to obtain efficient approximations to this optimum solution, for example, Dalenius and Hodges (1959); Ekman (1959) and Lavallée and Hidiroglou (1988), but all have implementation problems which make them difficult to use. For example, the well-known cumulative square root frequency method of Dalenius and Hodges (1959), referred to in this study as the cum $\sqrt{f(x)}$ method, depends on the arbitrary choice of initial class divisions of the frequency distribution. The Lavallée-Hidiroglou algorithm, an iterative method specifically for skewed populations, has convergence problems. In the next chapter we examine some of the available methods for obtaining stratum boundaries.

1.7 Objectives of the Study

The main objective of this research is to develop a stratum construction method that is both easy to use and efficient for positively skewed populations. Such an algorithm would be of benefit to users who encounter highly positively skewed populations such as audit, income and bank resources data. The specific objectives are:

1. To develop a new method for stratifying skewed populations which overcomes the problems of existing methods;

2. To investigate the efficiency of the new method compared to currently used methods;

3. To investigate if an improvement can be made to the performance of the Lavallée-Hidiroglou (1988) method;

4. To stratify the Pareto distribution using the new method.

A more detailed description of these objectives is given below.

1.7.1 Objective 1 - New Method

Various authors (Dalenius and Hodges, 1959; Cochran, 1961 and Lavallée and Hidiroglou, 1988) have suggested that in skewed populations near-optimum stratification can be achieved when each stratum has equal coefficients of variation. The first objective is to investigate if stratum breaks can be made such that near equal stratum coefficients of variation are achieved and to develop such a stratification method.

1.7.2 Objective 2 - Comparison of New Method with Methods Used in Practice

The second objective is to investigate the efficiency of the new stratification method compared to two currently used methods, the cum $\sqrt{f(x)}$ method of Dalenius and Hodges (1959) and the Lavallée-Hidiroglou (1988) method. The stratification methods are compared in terms of stratum breaks, stratum sizes and stratum sample sizes as well as equality of stratum coefficients of variation and precision of the estimates. The comparative performance of the methods is tested on four real positively skewed populations, an accounting population of debtors from a commercial entity in the Irish Public Sector detailed in Horgan (1996) and three populations used by Cochran (1961).

1.7.3 Objective 3 - Improving the Lavallée-Hidiroglou Method

The Lavallée-Hidiroglou (1988) procedure, specifically for skewed populations, starts with arbitrary initial boundaries and replaces them iteratively. Rivest (2002) reported numerical difficulties with the algorithm, failure to reach the global minimum sample size and non-convergence of the algorithm. The algorithm's starting values are of paramount importance as resulting boundaries depend on where the initial boundaries are set (Detlefsen and Veum, 1991). The third objective is to improve the convergence of the Lavallée-Hidiroglou algorithm by using initial boundaries created by the new method.

1.7.4 Objective 4 - Stratifying the Pareto Distribution

Many business surveys encounter highly positively skewed populations. These populations can naturally be modelled by distributions such as the log-normal, the exponential, the Pareto and others. The fourth objective investigates the stratification of the Pareto distribution using the new method.

1.8 Limitations of the Study

This study is an investigation of univariate stratification with respect to the construction of strata under the assumption that the stratification variable and the survey variable are the same. It does not deal with:

1. Stratification algorithms that take account of differences between the stratification variable and the survey variable;

2. Multivariate stratification. Surveys are often designed for estimating means and totals of many variables and several stratifying variables are available. The usual approach is to use some multivariate stratification scheme that represents a compromise solution for the different purposes.

1.9 Structure of the Thesis

The remainder of the thesis is structured as follows:

Chapter 2 gives an overview of stratum construction methods.

Chapter 3 details the methodology used in this study.

Chapter 4 develops a new easy-to-use stratum construction method. Using four real positively skewed populations, the performance of the new stratification method is compared with:

(i) the cum $\sqrt{f(x)}$ method;

(ii) the more recently developed method for skewed distributions, the Lavallée-Hidiroglou method.

The performance of the iterative Lavallée-Hidiroglou algorithm is compared using different starting points for the initial boundaries in Chapter 5.

Chapter 6 investigates the stratification of the Pareto distribution using the new method.

Finally, Chapter 7 gives a summary of the results and provides suggestions for future research.

Chapter 2

Stratification Bounds: An Overview

2.1 Introduction

The problem of obtaining break points that minimise the variance of the stratified mean has been studied theoretically by Dalenius (1950). He demonstrated that for fixed total sample size under Neyman allocation (1.5.7), the set (k_h) of cutting points satisfying the relation

$$\frac{\sigma_h^2 + (k_h - \mu_h)^2}{\sigma_h} = \frac{\sigma_{h+1}^2 + (k_h - \mu_{h+1})^2}{\sigma_{h+1}}, \; 1 \leq h \leq L-1 \qquad (2.1.1)$$

corresponds to minimum variance stratification when stratifying variables on the survey variable itself. However, as Dalenius pointed out, the above equation (2.1.1) is troublesome to solve due to the dependencies among the components: the stratum mean, μ_h, and stratum standard deviation, σ_h, cannot be computed until the boundaries are determined.

Numerous attempts have been made to develop procedures which would approximate optimum stratification. In this chapter we look at some of these

procedures. Section 2.2 discusses stratification procedures for stratifying general finite populations and Section 2.3 looks at stratifying skewed finite populations. A summary is given in Section 2.4.

2.2 Stratification Methods for General Finite Populations

The simplest methods of obtaining boundaries are the quantile method which places the same number of units in each stratum and the equal range method suggested by Aoyama (1954) which divides the range by the number of strata. If the quantile method is applied to highly positively skewed populations, the strata at the lower end are too narrow and those at the upper end too wide for optimum estimation (Cochran, 1961). On the other hand, using the equal range method on positively skewed populations, the strata at the lower end are too wide and those at the upper end too narrow (Cochran, 1961). Another simple method (termed the equal aggregrate method) was proposed by Mahalanobis (1952) and Hansen et al. (1953) where the total aggregrate value is equal for all strata i.e.

$$T_h = \frac{\sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi}}{L}.$$

Sethi (1963) demonstrated that the equal aggregrate method does not necessarily lead to efficient stratification when applied to normal, gamma or beta distributions. Raj (1964) also tested this rule on four theoretical distributions, three belonging to the exponential class and a right triangular distribution, and found that it was not optimum or near-optimum when L was large. The explanation given was that the lowest stratum made by this method was always too large compared with the corresponding stratum in the optimum case.

Dalenius and Gurney (1951) suggested that the formation of strata be on

the basis of equalisation of $W_h\sigma_h$. However, since the calculation of σ_h is required, and this depends on the stratum boundaries, this method is not convenient in practice (Cochran, 1961). Ekman (1959) suggested equalising the product of stratum weight and stratum range making $W_h(k_h - k_{h-1})$ constant. Although the method appears fairly simple, it is troublesome to apply in practice because the value of $\sum_{h=1}^{L} W_h(k_h - k_{h-1})$ is not constant, depending on both L and on the position of the boundaries (Cochran, 1961). Also the iterations become laborious (Hess, 1966) and require rather ominous calculations (Slanta and Krenzke, 1996). Hedlin (2000) took a geometric interpretation of Ekman's rule representing a population by a step function of cumulated frequencies. Strata are represented by rectangles and Hedlin attempted to "minimise the difference between the largest and smallest of the areas of the rectangles" which he stated would approximate Ekman's rule "as well as possible". However, Hedlin cautioned that convergence of the iterative process may be slow for large populations.

Durbin (1959) proposed obtaining stratum boundaries by taking equal intervals on the cumulative of $\frac{1}{2}(f(y) + r(y))$ where $r(y)$ is a rectangular distribution over the same range and with the same total frequency as $f(y)$.

Sethi (1963) suggested a method for finding optimum or near-optimum points of stratification for the normal and various chi-square distributions for 2 to 6 strata using equal (1.5.1), proportional (1.5.3) and Neyman (1.5.7) allocation, and tabulated these points. Then for any real population which resembles one of these standard distributions in shape, the corresponding points can be taken directly from the table. However, since this method calculates the optimum stratification points for certain distributions, the study population has to resemble one of these.

Dalenius and Hodges (1959) proposed constructing equal intervals on the cu-

mulative of the square root of the frequencies, the cum $\sqrt{f(x)}$ rule. This method is still the most commonly used in practice (Hedlin, 2000) and will be used in this study for comparison purposes. We will discuss this method in detail in the next chapter.

Cochran (1961) compared the cum $\sqrt{f(x)}$ method, the equal aggregrate method of Mahalanobis (1952), Ekman's method and Durbin's method, for 2, 3 and 4 strata by applying them to eight real skewed populations. He found that both the cum $\sqrt{f(x)}$ method and Ekman's method performed consistently well. Durbin's method did fairly well except on the two most skewed populations. He also found that the equal aggregrate method of Mahalanobis (1952) was relatively unsuccessful on the three least skewed populations, going on to explain that this result is not surprising since the method is not designed to work well for a rectangular distribution with the lower end at zero. For the other populations, the equal aggregrate method behaved erratically. Hess et al. (1966) observed that

> *"Sethi's method, to some extent, and the cum $\sqrt{f(x)}$ rule to a greater extent, lead to the construction of top strata that are too wide, with the result that these strata contribute heavily to the total variance."*

Singh (1971) and Thomsen (1976) recommended a method of obtaining stratum boundaries based on equal partitioning of the cumulative cubed root frequency of the density function. Singh's method requires prior knowledge of the regression model of the survey variable y on the auxiliary variable x, while Thomsen (1976) assumes the regression model is linear. Thomsen (1976) concluded that the cumulative cubed root frequency works better with proportional allocation (1.5.3) than with equal allocation (1.5.1). He also claims this method compares favourably to the cum $\sqrt{f(x)}$ method using proportional allocation (1.5.3).

Another approach taken for determining optimum stratum boundaries is to

formulate the problem as a mathematical programming problem. Khan et al. (2002) views the problem of stratum construction as a multistage decision where the optimum stratum widths are determined using dynamic programming to obtain the global minimum of the objective function using Neyman allocation (1.5.7) for fixed sample size.

Random search methods have also been suggested. One method proposed by Kozak (2004) iteratively increases or decreases one boundary by not more than 5 units while the other boundaries remain constant. He claims this algorithm is more efficient than the random search method proposed by Niemiro (1999) which changes a boundary by one unit which could result in the algorithm stopping at a local minimum and does not work well for large populations as it requires too many iteration steps.

Model-based methods treat values in the population as random variables and derive inferences to the population from the model specified for the random variables. A model-based approach to stratification has also been suggested by researchers and is described in Sarndal et al. (1992, sec. 12.4). However, accuracy depends on the choice of model.

2.3 Stratification Methods for Skewed Populations

Positively skewed populations with long tails to the right are characteristic of many business applications such as auditing, income and bank resources. In such populations stratification can greatly improve the precision of the sample estimates.

An approach for stratifying a skewed population is to create a certainty or take-all stratum which contains some of the largest units in the population, and take-some strata containing the remaining units, where the final break point k_{L-1} is

the boundary between the take-all and take-some strata (Hidiroglou, 1986). Units in the top take-all stratum are selected with certainty whereas a sample of units is taken from the take-some strata. The goal in defining a certainty stratum is to identify the extreme values within a population that heavily influence the estimate and its variance. Taking all of the N_L units in a certainty stratum reduces the sampling error to zero for this stratum.

It is common in practice for k_{L-1} to be judgementally selected (McCarthy and Clickner, 1985). For example, k_{L-1} can be taken at the point where data is sparse and no longer clustered (Falk and Rotz, 2003), or six times the population mean (Newman, 1976). Other methods create a certainty stratum containing a percentage of the total value. For example, Roshwalb et al. (1987) suggest taking 35% of the total. Alternatively, the certainty stratum could contain "outliers" identified using subject matter knowledge (Sigman and Monsour, 1995).

Approximate cut-off rules for optimally determining k_{L-1} in such a way that the variance of the estimate is minimised have been given by Dalenius (1952), Glasser (1962) and Hidiroglou (1986). Glasser (1962) expressed k_{L-1} as a function of the mean, the sampling interval and the population variance and attempted to minimise the variance for a fixed sample size. Hidiroglou (1986) proposed exact and approximate rules for determining k_{L-1} which minimises the sample size for a desired level of precision.

Chen (1989) applied the cum $\sqrt{f(x)}$ method to obtain the take-some stratum boundaries after determining a certainty stratum using Hidiroglou's method and found this to be an improvement over the sole use of the cum $\sqrt{f(x)}$ method on skewed distributions.

Lavallée and Hidiroglou (1988) presented an iterative procedure for stratifying skewed populations into a take-all stratum and $L-1$ take-some strata. The Lavallée-Hidiroglou method will be used in this study for comparison purposes. We will also attempt to improve its convergence. A more detailed description of its implementation is given in Chapter 3.

2.4 Summary

While the equations of Dalenius (2.1.1) give an exact solution, they are difficult to solve and various approximation methods have been developed. In this chapter we overviewed these approximations. The two methods that are frequently used are the cum $\sqrt{f(x)}$ method and the Lavallée-Hidiroglou method. However, both these methods have some worrying implementation problems. The cum $\sqrt{f(x)}$ method has an inbuilt arbitrariness while the Lavallée-Hidiroglou method, which takes a top take-all stratum, has convergence problems. These methods will be discussed in the next chapter.

Chapter 3

The Methodology

3.1 Introduction

This chapter details two frequently used methods for obtaining stratum boundaries. The cum $\sqrt{f(x)}$ method is described in Section 3.2.1 and the Lavallée-Hidiroglou iterative method is given in Section 3.2.2. These methods will be used as comparators for our new method. All three algorithms will be applied to four real positively skewed populations which are described in Section 3.3. A summary is given in Section 3.4.

3.2 The Stratification Methods Used as Comparators

The two benchmark methods will be described in this section.

3.2.1 The Cumulative Square Root Frequency Method

Dalenius and Hodges (1959) proposed constructing equal intervals on the cumulative of the square root of the frequencies, i.e. the cum $\sqrt{f(x)}$ rule. They showed that if

$$H = \int_{k_0}^{k_L} \sqrt{f(x)}dx$$

and the bounds k_h are chosen so that

$$\int_{k_{h-1}}^{k_h} \sqrt{f(x)}dx = \frac{H}{L}, \quad \forall h,$$

then $V(\overline{y}_{st})$ is approximately minimised for fixed n when Neyman allocation is used to allocate n among the strata. The argument of Dalenius and Hodges (1959) is as follows:

If the strata are numerous and narrow, the value of $f(x)$ is approximately constant within a given stratum i.e. x is uniformly distributed. Hence

$$S_h \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1}). \tag{3.2.1}$$

Also with

$$W_h = \int_{k_{h-1}}^{k_h} f(x)dx,$$

there exists by the mean value theorem, f_h in $(k_{h-1},\ k_h)$ so that

$$W_h = f_h(k_h - k_{h-1}). \tag{3.2.2}$$

Also by the mean value theorem there exists $f_h^{'}$,

$$\int_{k_{h-1}}^{k_h} \sqrt{f(x)}dx = \sqrt{f_h^{'}}(k_h - k_{h-1}) \approx \sqrt{f_h}(k_h - k_{h-1}), \tag{3.2.3}$$

assuming

$$f_h^{'} \approx f_h.$$

Recall from (1.5.8), the variance of the stratified mean with Neyman allocation ignoring the finite population correction factors for the strata (1.2.13) may be written as

$$V_{opt}(\overline{x}_{st}) = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{n}.$$

Clearly to minimise the variance of the stratified sample mean, it is sufficient to minimise

$$\sum_{h=1}^{L} W_h S_h. \tag{3.2.4}$$

Substituting approximations (3.2.1) and (3.2.2) into (3.2.4), we have

$$\sum_{h=1}^{L} W_h S_h \approx \frac{1}{\sqrt{12}} \sum_{h=1}^{L} f_h (k_h - k_{h-1})^2 = \frac{1}{\sqrt{12}} \sum_{h=1}^{L} \left(\sqrt{f_h} (k_h - k_{h-1}) \right)^2.$$

Therefore minimising

$$\sum_{h=1}^{L} W_h S_h$$

is equivalent to minimising

$$\sum_{h=1}^{L} \left(\sqrt{f_h} (k_h - k_{h-1}) \right)^2,$$

subject to

$$\sum_{h=1}^{L} \sqrt{f_h} (k_h - k_{h-1}) \approx \int_{k_0}^{k_L} \sqrt{f(x)} dx = H.$$

Using Lagrange multipliers, the minimum is achieved when

$$\sqrt{f_h} (k_h - k_{h-1}) = \int_{k_{h-1}}^{k_h} \sqrt{f(x)} dx = \frac{H}{L}, \quad \forall h.$$

So it follows that the minimum variance is approximately achieved when:

$$\int_{k_0}^{k_1} \sqrt{f(x)} dx = \int_{k_1}^{k_2} \sqrt{f(x)} dx = \ldots = \int_{k_{L-1}}^{k_L} \sqrt{f(x)} dx.$$

3.2.1.1 Implementation Details

Cochran (1961) showed how this algorithm may be used on finite data as follows:

1. The population values $x_1, x_2, \ldots, x_N$ are sorted in ascending order, grouped into an arbitrary number of classes, J, and the frequency for each class f_j,

$1 \leq j \leq J$, is determined.

2. The square root of the frequencies in each class is determined and then cumulated $\sum_{j=1}^{J} \sqrt{f_j}$.

3. This sum is then divided by the desired number of strata, L i.e.

$$Q = \frac{\sum_{j=1}^{J} \sqrt{f_j}}{L}. \tag{3.2.5}$$

4. The nearest available points to

$$Q, 2Q, \ldots, LQ$$

on the cumulated square root of the frequencies scale are selected. The upper boundaries of each stratum are the corresponding upper interval value on the class interval scale.

3.2.1.2 Implementation Problems

The cum $\sqrt{f(x)}$ method has some worrying drawbacks. The final strata depend on the number of J initial class divisions, and there is no theory that gives the best number of classes (Hedlin, 2000). However, Hedlin (2000) admits that this problem of arbitrariness in division breaks and the number of initial classes

> *"might not be severe, as the estimator variance regarded as a function of the stratum boundaries is usually flat around its minimum, which makes minor deviations from the minimum negligible."*

A related and more important drawback is the intricateness of developing an algorithm to deal with this arbitrariness. For most applications there is no way of obtaining an ideal J, so that the cum $\sqrt{f(x)}$ in each stratum is exactly equal to Q (3.2.5). Hedlin (2000) points out that it is difficult to construct an algorithm to

achieve this and to determine when the process should be repeated with a different J, and which new J should be used.

3.2.2 The Lavallée and Hidiroglou (1988) Algorithm

The Lavallée-Hidiroglou algorithm starts with $(L-1)$ arbitrary initial boundaries and replaces them iteratively, using a procedure suggested by Sethi (1963), until the sample size required to obtain the given precision is minimised; the precision is usually stated by requiring the $cv(\overline{x}_{st})$ to be a specified level between 1% and 10%. Lavallée and Hidiroglou used the quantile method, placing an equal number of units in each stratum to obtain initial boundaries. Sample sizes n_h in the take-some strata, $1 \leq h \leq L-1$, are determined using power allocation. Taking all of the N_L units in the take-all stratum reduces the sampling error to zero for this stratum. It is obviously assumed that $n > N_L$. With stratum L as the take-all stratum, the variance in (1.2.11) may be written as

$$V(\overline{x}_{st}) = \sum_{h=1}^{L-1} \left(1 - \frac{n_h}{N_h}\right) \frac{W_h^2 S_h^2}{(n - N_L)a_h} \tag{3.2.6}$$

where a_h is the proportion of the $n - N_L$ sampling units allocated to the h^{th} take-some stratum. Note the allocation rule a_h satisfies $\sum_{h=1}^{L-1} a_h = 1$.

Equation (3.2.6) can be expressed in terms of the sample size n as follows:

$$n = N_L + \frac{\sum_{h=1}^{L-1} W_h^2 S_h^2 / a_h}{V(\overline{x}_{st}) + \sum_{h=1}^{L-1} W_h S_h^2 / N}. \tag{3.2.7}$$

Writing $V(\overline{x}_{st}) = \overline{X}^2 cv^2(\overline{x}_{st})$ with power allocation

$$a_h = \frac{(W_h \overline{X}_h)^p}{\sum_{i=1}^{L-1} (W_i \overline{X}_i)^p}, \tag{3.2.8}$$

(3.2.7) becomes

$$n = N_L + \frac{\sum_{h=1}^{L-1} (W_h S_h)^2 (W_h \overline{X}_h)^{-p} \sum_{i=1}^{L-1} (W_i \overline{X}_i)^p}{\overline{X}^2 cv^2(\overline{x}_{st}) + \sum_{h=1}^{L-1} W_h S_h^2 / N}. \tag{3.2.9}$$

In (3.2.9) we can treat n as a function of the stratum boundaries $k_1, k_2, \ldots, k_{L-1}$, and the optimum k_h are those that minimise n for a given $cv(\overline{x}_{st})$, i.e.

$$\frac{\partial n}{\partial k_1} = \frac{\partial n}{\partial k_2} = \cdots = \frac{\partial n}{\partial k_{L-1}} = 0. \tag{3.2.10}$$

From (3.2.10) we apparently obtain a series of quadratic equations in k_h:

$$\alpha_h k_h^2 + \beta_h k_h + \gamma_h = 0, \quad 1 \leq h \leq L-1. \tag{3.2.11}$$

However the coefficients α_h, β_h and γ_h, as well as being functions of W_h, S_h, and $\overline{X}_h$, are also functions of k_h, and so the k_h can only be solved iteratively.

3.2.2.1 Implementation Details

The iterative procedure for solving (3.2.11) is described in detail in Lavallée and Hidiroglou (1988) and summarised below:

1. Sort the population values $x_1, x_2, \ldots, x_N$ in ascending order.
2. Choose the initial boundaries $k_1, k_2, \ldots, k_{L-1}$, so that each stratum has the same number of units.
3. Based on these boundaries, calculate the weights W_h, the means $\overline{X}_h$ and the variances S_h^2 $(h = 1, 2, \ldots, L)$.
4. Choose the N_L units in the top stratum, and allocate the remaining $n - N_L$ units among the $L-1$ remaining strata according to the power allocation method given in (3.2.8).
5. Replace the initial set of boundaries by taking the larger root of (3.2.11):

$$k_h^{new} = \frac{-\beta_h + \sqrt{\beta_h^2 - 4\alpha_h\gamma_h}}{2\alpha_h}. \tag{3.2.12}$$

6. Repeat steps 3, 4 and 5 with the new set of boundaries, continuing until two consecutive sets are either identical or differ by negligible quantities.

The SAS code used for implementing the algorithm is available on the web at *http://www.mat.ulaval.ca/pages/lpr/*.

3.2.2.2 Implementation Problems

Users of the Lavallée-Hidiroglou algorithm have highlighted some serious implementation problems:

- Slanta and Krenzke (1994, 1996) encountered numerical difficulties when using the algorithm with Neyman allocation. They found convergence of the algorithm slow, and that sometimes it did not converge to the true minimum sample size n. Because of the possible convergence problems with the default starting points where each stratum has the same number of units, they used the cum $\sqrt{f(x)}$ method to obtain the starting points in the Annual Capital Expenditures Survey (ACES) of the U.S. Bureau of the Census. However, as discussed in (3.2.1.2), the cum $\sqrt{f(x)}$ method has an inbuilt arbitrariness.

  Slanta and Krenzke (1994) attempted to address the convergence problem by setting up constraints to be met after each iteration. Under the assumption that the marginal gain achieved by further iterations is not worth the extra effort, they stopped the program when:

  (i) the difference between the new upper (lower) boundary and the previous iteration's upper (lower) boundary is less than one. Slanta and Krenzke (1996) stated

"the whole number, one, was used in our case since payroll values are only available to us in whole number values and any shifting of boundaries of a value less than one does not affect any companies;"

(ii) the improvement in sample size from iteration to iteration is marginal or nonexistent;

(iii) the program goes into the 30th iteration.

They concluded convergence should be determined on the basis of the sample size instead of the boundary values, as the boundaries vary greatly in the neighbourhood of the minimum sample size while sample size varies only slightly.

- Rivest (2002) reported similar numerical difficulties, failure to reach the global minimum sample size, and non-convergence of the algorithm when the number of strata is large. Rivest observed that using the algorithm with power allocation is generally more stable than using Neyman allocation.

- Detlefsen and Veum (1991) found that convergence occurs faster for a smaller number of strata. They modified the algorithm to carry out Neyman allocation (1.5.7) in the take-some strata. However, in applying the modified algorithm to the redesign of the U.S. Monthly Retail Trade Survey, they found that convergence of the algorithm was slow (often 50 - 100 iterations) or nonexistent. They also found that the resulting boundaries depend on where the initial boundaries are set (many times the boundaries differed substantially), so that the minimum sample size attained is a local but not necessarily a global minimum.

- Chen (1989) noted that the values in the square roots of (3.2.12) may be negative mwhich usually happens when the target precision $cv(\overline{x}_{st})$ is really

small causing the program to naturally terminate.

3.2.3 Summary of Section

As can be seen from the above algorithms, the cum $\sqrt{f(x)}$ method and the Lavallée and Hidiroglou method have serious implementation problems. Despite these problems, they are still being used possibly as the "best available".

3.3 The Data

In order to examine the performance of our new algorithm, and to compare it to the cum $\sqrt{f(x)}$ and the Lavallée and Hidiroglou methods, we implement them on four real positively skewed populations. The first is an accounting population of debtors from Horgan (1996) and the other three are from Cochran (1961). We detail these populations next.

3.3.1 Population 1

Population 1 consists of debtor accounts from a state scientific consultancy firm audited by the office of the Comptroller and Auditor General and detailed in Horgan (1996). The firm is responsible for a number of national standards and also provides various technical services to industry. The population consists of all positive balances of debtors. The main descriptive parameters of the population are given in Table 3.1. Table 3.2 contains the frequency table.

Table 3.1: Population 1 Parameters

Total Book Value	Ir£2,825,374.00
Mean	Ir£838.64
Standard Deviation	1,874.00
Skewness	6.44
Kurtosis	59.13
Minimum	Ir£40.00
First Quartile	Ir£117.00
Median	Ir£290.00
Third Quartile	Ir£700.00
Maximum	Ir£28,000.00
Number of Items	3,369

Table 3.2: Population 1 Frequency Table

Amount (Ir£s)	No. of Line Items	% Line Items
0 - 500	2,259	67.1
500 - 1,000	523	15.5
1,000 - 1,500	168	5.0
1,500 - 2,000	95	2.8
2,000 - 2,500	67	2.0
2,500 - 3,000	56	1.7
3,000 - 3,500	34	1.0
3,500 - 4,000	25	0.7
4,000 - 4,500	19	0.6
4,500 - 5,000	23	0.7
5,000 - 10,000	74	2.2
10,000 - 20,000	21	0.6
>20,000	5	0.1
Total	3,369	100

3.3.2 Population 2

Population 2 is one of the populations used by Cochran (1961) to test the efficiency of the cum $\sqrt{f(x)}$ method. This population shows the number of inhabitants (in thousands) of U.S. cities in 1940. The main descriptive parameters of the population are given in Table 3.3. Table 3.4 contains the frequency table.

Table 3.3: Population 2 Parameters

Mean	32.57
Standard Deviation	30.40
Skewness	2.88
Kurtosis	9.19
Minimum	10.00
First Quartile	16.00
Median	23.00
Third Quartile	33.00
Maximum	198.00
Number of Items	1,038

Table 3.4: Population 2 Frequency Table

No. of Inhabitants	No. of Cities	% Cities
0 - 20	434	41.8
20 - 30	315	30.4
30 - 40	89	8.6
40 - 50	49	4.7
50 - 60	27	2.6
60 - 70	28	2.7
70 - 80	17	1.6
80 - 90	25	2.4
90 - 100	11	1.1
100 - 150	20	1.9
>150	23	2.2
Total	1,038	100

3.3.3 Population 3

Population 3 is a population of the number of students in four-year U.S. colleges in 1952-1953 (Cochran, 1961). Table 3.5 gives the main descriptive parameters of the population. Table 3.6 contains the frequency table.

Table 3.5: Population 3 Parameters

Mean	1,563.00
Standard Deviation	1,799.06
Skewness	2.46
Kurtosis	5.88
Minimum	200.00
First Quartile	567.00
Median	911.00
Third Quartile	1,682.00
Maximum	9,623.00
Number of Items	677

Table 3.6: Population 3 Frequency Table

No. of Students	No. of Colleges	% Colleges
0 - 500	372	55.0
1,000 - 1,500	118	17.4
1,500 - 2,000	54	8.0
2,000 - 2,500	19	2.8
2,500 - 3,000	28	4.1
3,000 - 4,000	24	3.5
4,000 - 5,000	12	1.8
5,000 - 6,000	15	2.2
6,000 - 8,000	20	3.0
>8,000	15	2.2
Total	677	100

3.3.4 Population 4

Population 4 represents the resources in millions of dollars in 1957 of large commerical banks in the U.S. (Cochran, 1961). The main descriptive parameters of the population are given in Table 3.7. Table 3.8 contains the frequency table.

Table 3.7: Population 4 Parameters

Mean	US$ 225.62
Standard Deviation	190.46
Skewness	2.08
Kurtosis	4.18
Minimum	US$ 70
First Quartile	US$ 108
Median	US$ 144
Third Quartile	US$ 268
Maximum	US$ 977
Number of Items	357

Table 3.8: Population 4 Frequency Table

Resources of Banks ($millions)	No. of Banks	% Banks
0 - 150	187	52.4
150 - 300	89	24.9
300 - 400	27	7.6
400 - 500	24	6.7
500 - 800	20	5.6
>800	10	2.8
Total	357	100

3.3.5 Summary of Data

A summary of the four populations used in this study is given in Table 3.9 and illustrated in Figure 3.1.

Table 3.9: Percentage of Total Frequency Falling in Successive Tenths of the Range for the Four Populations

Range %	Population 1 % Line Items	Population 2 % Cities	Population 3 % Colleges	Population 4 % Banks
0 - 10	93.53	70.2	67.4	57.9
10 - 20	3.98	14.1	14.4	16.2
20 - 30	1.28	5.9	6.5	8.8
30 - 40	0.50	3.8	2.7	4.9
40 - 50	0.36	2.1	1.9	4.8
>50	0.36	3.9	7.1	7.4

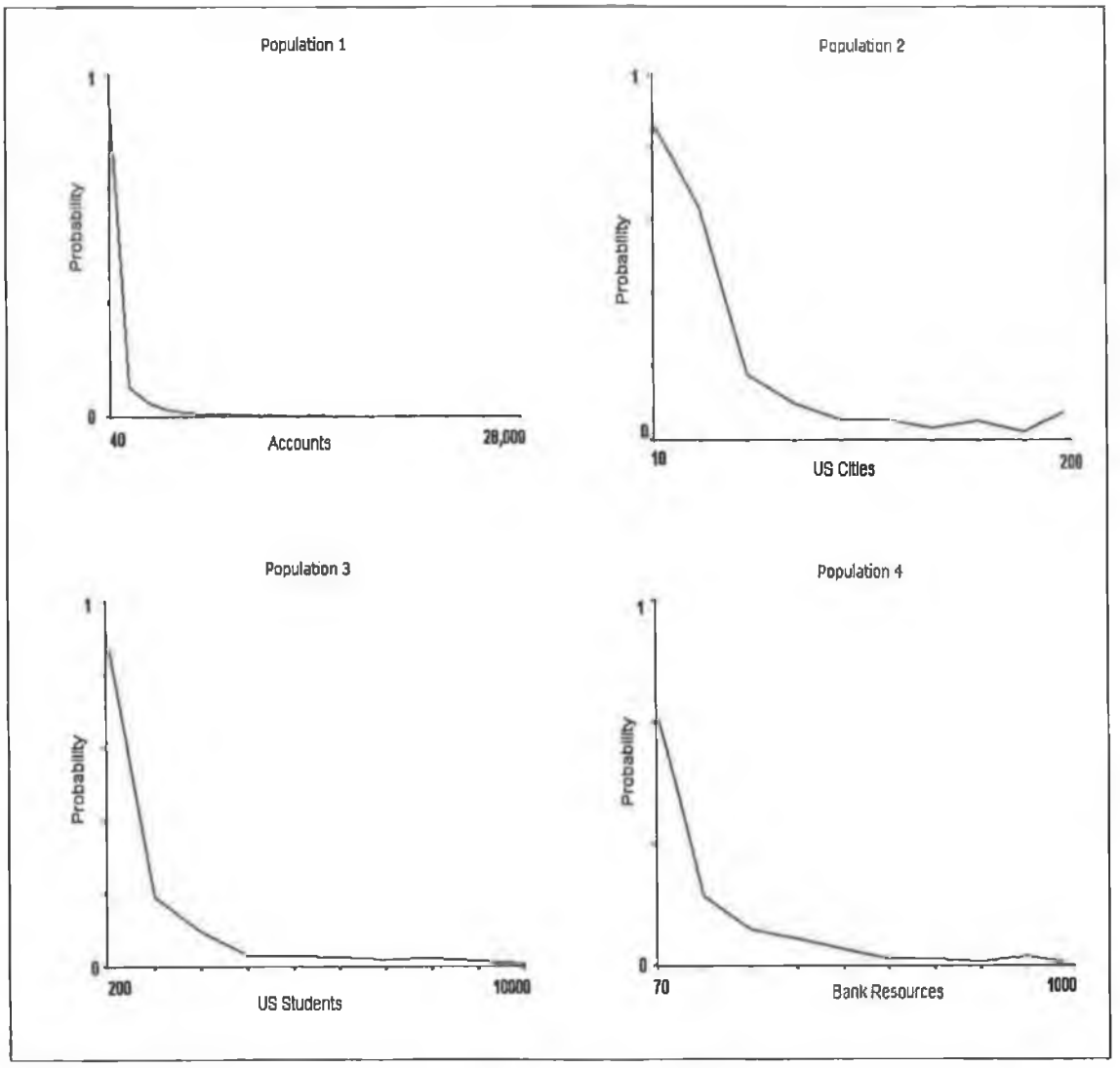

Figure 3.1: The Four Real Positively Skewed Populations used in this Study

From Table 3.9 and Figure 3.1 it can be seen that all four populations are positively skewed. There were five other populations in the paper of Cochran (1961) which turned out to be unsuitable for use with our algorithm. In three cases the variable was a proportion: agricultural loans, real estate loans and independent loans expressed as a percentage of the total amount of bank loans. Another, a population of farms in which the variable ranged from 1 to 18, was essentially discrete. Yet another, a population of income tax returns, was not sufficiently skewed: it owed its skewness to the top 0.05% of the population, and when this was removed, or put in a take-all stratum, the skewness disappeared.

As can be seen from Table 3.9, population 1 contains the greatest percentage of low valued items with the lowest 10% of the range containing over 93% of the items. Populations 2, 3 and 4 contain a lesser percentage in the lowest 10% of the range but all three have over 55% of the items in this range. In the upper 50% of the range for the four populations, the percentage of items is relatively low. Population 1 has the lowest percentage with only 0.36%. The other three populations have a higher percentage ranging from just under 4% to just over 7%.

The populations are all highly postively skewed and in each case, a small proportion of the items account for a large proportion of the total. The skewness of these populations ranges in decreasing order from 6.44 for population 1 down to 2.08 for population 4. These populations provide an opportunity of investigating the stratification methods on data of varying degrees of skewness typically found in business.

3.4 Chapter Summary

This chapter has given an overview of the cum $\sqrt{f(x)}$ method and the Lavallée-Hidiroglou method for stratum construction and discussed some of their implemen-

tation problems. We have also described the data on which the stratification algorithms will be applied. In what follows we will use these methods as comparators for our new algorithm.

Chapter 4

A New Stratum Construction Method

4.1 Introduction

This chapter derives an algorithm for constructing stratum boundaries which is much simpler to implement than any of those currently available. It is based on an observation made by a number of researchers:

Dalenius and Hodges (1959) stated that when the number of strata is large

> *"for many populations, and for reasonable location of the stratum boundaries, the relative variance does not vary much from stratum to stratum."*

Cochran (1961) examined the stratification of skewed populations and also noted that

> *"with near-optimum boundaries the coefficients of variation are often found to be approximately the same in all strata."*

Recall that the desire to equalise stratum coefficients of variation cv_h is often asked by the users of survey data (Lavallée and Hidiroglou, 1988). However, Cochran (1961) concluded that computing and setting equal the standard deviations of the strata would be too complicated to be feasible in practice.

We derive an algorithm, designed specifically for skewed distributions, which equalises cv_h in Section 4.2. The algorithm attempts to overcome the limitations of those currently available. To assess the performance of the new method, it is compared with the cum $\sqrt{f(x)}$ method of Dalenius and Hodges (1959) and the Lavallée-Hidiroglou method (1988) in Section 4.3. A summary and discussion is given in Section 4.4.

4.2 A New Stratum Construction Method

4.2.1 The Algorithm

For any given minimum and maximum data points, k_0 and k_L, we assume that the stratum breaks $(k_1, \ldots, k_{L-1})$ which divide the population into L strata are made so that the cv_h are the same for $h = 1, 2, \ldots, L$. We stratify a known auxiliary variable x and we wish to determine the stratum breaks so that

$$\frac{S_1}{\overline{X}_1} = \frac{S_2}{\overline{X}_2} = \ldots = \frac{S_L}{\overline{X}_L} .$$

Here S_h is the standard deviation in stratum h of the x variable,

$$S_h = \sqrt{\frac{\sum_{i=1}^{N_h} (x_{hi} - \overline{X}_h)^2}{N_h - 1}} ,$$

and $\overline{X}_h$ the mean in stratum h of the x variable,

$$\overline{X}_h = \frac{1}{N_h} \sum_{i=1}^{N_h} x_{hi} \ .$$

Making the assumption that the distribution within each stratum is approximately uniform, the mean of stratum h is

$$\overline{X}_h \approx \frac{k_h + k_{h-1}}{2}$$

and the standard deviation of stratum h is

$$S_h \approx \frac{1}{\sqrt{12}}(k_h - k_{h-1}).$$

As an approximation to the coefficient of variation of stratum h, this gives

$$cv_h \approx \frac{(k_h - k_{h-1})/\sqrt{12}}{(k_h + k_{h-1})/2} \ .$$

With $cv_h = cv_{h+1}$, this gives

$$\frac{k_{h+1} - k_h}{k_{h+1} + k_h} = \frac{k_h - k_{h-1}}{k_h + k_{h-1}} \ .$$

Cross multiplication gives:

$$(k_{h+1} - k_h)(k_h + k_{h-1}) = (k_h - k_{h-1})(k_{h+1} + k_h)$$

which reduces to

$$k_h^2 = k_{h+1} k_{h-1}$$

i.e.

$$\frac{k_h}{k_{h-1}} = \frac{k_{h+1}}{k_h}, \quad h = 1, 2, \ldots, L-1. \tag{4.2.1}$$

Thus the ratio

$$r = \frac{k_{h+1}}{k_h}$$

is independent of h, so that the stratum boundaries are the terms of a geometric progression:

$$k_h = ar^h,\ 1 \leq h \leq L.$$

In particular $a = k_0$, the minimum value of the variable, $ar^L = k_L$, the maximum value of the variable, and hence the constant ratio is given by

$$r = \left(\frac{k_L}{k_0}\right)^{1/L}.$$

4.2.2 A Numerical Example

The following example illustrates the geometric progression algorithm.

Taking

$$L = 4\ ;\ k_0 = 5\ ;\ k_4 = 50,000$$

gives

$$r = \left(\frac{50,000}{5}\right)^{1/4} = 10.$$

Thus $k_h = 5(10)^h$ $(h = 0, 1, 2, 3, 4)$ and the strata form the ranges

$$5 - 50;\ 50 - 500;\ 500 - 5,000;\ 5,000 - 50,000.$$

Clearly this is an extremely simple method of obtaining stratum breaks.

4.2.3 Uniform Distribution within Strata

The relationship (4.2.1) depends critically on the assumption that the distribution within each stratum is uniformly distributed. The assumption may be justified by the following heuristic argument.

When the distribution from which the sample is drawn is positively skewed, then the low values of the variable have a high incidence, which decreases as the variable values increase. This makes it appropriate to take small intervals at the beginning and large intervals at the end. This is what happens with a geometric series of constant ratio greater than one. In the lower range of the variable, the strata are narrow so that an assumption of rectangular distribution in them is not unreasonable. As the value of the variable increases, the stratum width increases geometrically. This coincides with the decreased rate of change of the incidence of the positively skewed variable, so here also the assumption of uniformity is reasonable. It should be noted that Dalenius and Hodges (1959) assumed uniformity within each stratum when developing the cum $\sqrt{f(x)}$ method.

4.3 Performance of New Method

We now compare the performance of this new algorithm with methods detailed in Chapter 3:

- Dalenius and Hodges (1959) cum $\sqrt{f(x)}$ method and
- Lavallée-Hidiroglou (1988) method.

For each of the four populations described in Chapter 3, the units are sorted in increasing order of size and stratified into 3, 4 and 5 strata. The number of strata is based on findings by Cochran (1977), and discussed in Section 1.4, who concluded that more than 5 or 6 strata produce very little additional variance reduction. Comparisons are made in terms of the following:

- Stratum breaks;
- Stratum sizes;
- Stratum sample sizes;
- Equality of stratum coefficients of variation;
- Relative efficiency.

4.3.1 Comparison with the Cum $\sqrt{f(x)}$ Method

The geometric method and the cum $\sqrt{f(x)}$ method are used to make the breaks with a sample of size $n = 100$ allocated using Neyman allocation for each method. When applying the cum $\sqrt{f(x)}$ method, the frequency distributions were divided into $J = 20$ equal class intervals. For population 1, the two lowest classes were each subdivided into 5 subclasses. Tables 4.1, 4.2 and 4.3 show the stratum breaks k_h, the stratum sizes N_h, stratum sample sizes n_h and the stratum coefficients of variation, cv_h. The precision expressed in terms of the coefficient of variation of the stratified sample mean $cv(\overline{x}_{st})$ obtained with each stratification method is given in the third column.

Table 4.1: The Geometric vs the Cum $\sqrt{f(x)}$: Stratum Breaks with $L = 3$ and $n = 100$

Population	Stratification Method	$cv(\bar{x}_{st})$		Stratum 1	2	3
1	Geometric	.0615	k_h	355	3153	
			N_h	1892	1288	189
			n_h	9	46	45
			cv_h	.59	.68	.64
	Cum $\sqrt{f}$	.0630	k_h	599	1997	
			N_h	2387	646	336
			n_h	19	12	69
			cv_h	.71	.35	.80
2	Geometric	.0270	k_h	27	73	
			N_h	701	243	94
			n_h	36	29	35
			cv_h	.28	.28	.33
	Cum $\sqrt{f}$	.0269	k_h	28	66	
			N_h	729	208	101
			n_h	40	22	38
			cv_h	.28	.25	.34
3	Geometric	.0316	k_h	727	2645	
			N_h	253	321	103
			n_h	9	38	53
			cv_h	.32	.37	.39
	Cum $\sqrt{f}$	.0282	k_h	1142	3498	
			N_h	438	170	69
			n_h	34	32	34
			cv_h	.39	.33	.27
4	Geometric	.0184	k_h	168	405	
			N_h	211	93	53
			n_h	27	27	46
			cv_h	.23	.24	.30
	Cum $\sqrt{f}$	.0195	k_h	160	432	
			N_h	202	109	46
			n_h	24	38	38
			cv_h	.22	.29	.28

Table 4.2: The Geometric vs the Cum $\sqrt{f(x)}$: Stratum Breaks with $L = 4$ and $n = 100$

Population	Stratification Method	$cv(\bar{x}_{st})$		Stratum 1	2	3	4
1	Geometric	.0439	k_h	205	1058	5443	
			N_h	1416	1382	483	88
			n_h	6	22	40	32
			cv_h	.45	.44	.48	.50
	Cum $\sqrt{f}$	.0461	k_h	319	1158	2836	
			N_h	1793	1046	312	218
			n_h	10	16	10	64
			cv_h	.56	.35	.26	.68
2	Geometric	.0192	k_h	21	44	93	
			N_h	459	398	130	51
			n_h	22	31	25	22
			cv_h	.21	.20	.22	.22
	Cum $\sqrt{f}$	.0199	k_h	19	38	85	
			N_h	393	428	155	62
			n_h	15	26	30	29
			cv_h	.19	.17	.24	.26
3	Geometric	.0216	k_h	526	1387	3653	
			N_h	138	343	127	69
			n_h	5	27	26	42
			cv_h	.27	.26	.26	.27
	Cum $\sqrt{f}$	.0228	k_h	671	2084	4911	
			N_h	224	326	74	53
			n_h	12	43	18	27
			cv_h	.30	.32	.22	.20
4	Geometric	.0141	k_h	135	261	505	
			N_h	156	109	63	29
			n_h	20	23	29	28
			cv_h	.18	.19	.19	.20
	Cum $\sqrt{f}$	.0142	k_h	160	296	523	
			N_h	202	73	54	28
			n_h	33	16	24	27
			cv_h	.22	.16	.17	.20

Table 4.3: The Geometric vs the Cum $\sqrt{f(x)}$: Stratum Breaks with $L = 5$ and $n = 100$

Population	Stratification Method	$cv(\bar{x}_{st})$		Stratum 1	2	3	4	5
1	Geometric	.0359	k_h	148	549	2037	7553	
			N_h	1054	1267	732	265	51
			n_h	2	14	27	33	24
			cv_h	.37	.38	.40	.37	.41
	Cum $\sqrt{f}$	.0357	k_h	319	599	1717	4234	
			N_h	1793	594	602	246	134
			n_h	12	4	16	14	54
			cv_h	.56	.17	.30	.25	.57
2	Geometric	.0141	k_h	18	33	59	108	
			N_h	364	418	130	87	39
			n_h	18	28	17	20	17
			cv_h	.18	.14	.15	.16	.15
	Cum $\sqrt{f}$	.0149	k_h	19	28	57	104	
			N_h	393	336	181	88	40
			n_h	21	15	26	20	18
			cv_h	.19	.10	.20	.16	.16
3	Geometric	.0179	k_h	433	941	2043	4434	
			N_h	94	255	198	74	56
			n_h	2	16	27	20	35
			cv_h	.22	.21	.24	.21	.21
	Cum $\sqrt{f}$	.0180	k_h	671	1613	3026	5853	
			N_h	224	279	90	48	36
			n_h	14	30	18	20	18
			cv_h	.30	.22	.19	.20	.14
4	Geometric	.0107	k_h	118	200	340	576	
			N_h	114	116	64	39	24
			n_h	12	20	26	18	24
			cv_h	.14	.14	.17	.12	.16
	Cum $\sqrt{f}$	.0110	k_h	115	206	342	568	
			N_h	110	127	57	39	24
			n_h	13	26	20	17	24
			cv_h	.14	.16	.16	.12	.16

(i) The Relative Efficiency

This section examines the precision of the geometric method and the cum $\sqrt{f(x)}$ method. Comparisons are made in terms of the relative efficiency defined as

$$eff = \frac{V_{geom}(\overline{x}_{st})}{V_{cum}(\overline{x}_{st})}\ , \tag{4.3.1}$$

where $V_{geom}(\overline{x}_{st})$ and $V_{cum}(\overline{x}_{st})$ are the variances of the stratified mean respectively with the geometric method and the cum $\sqrt{f(x)}$ method. The relative efficiency has two primary uses:

(i) In appraising the precision of two stratification methods;

This involves measuring the accuracy of the geometric method compared to the cum $\sqrt{f(x)}$ method. If eff is less than one, the geometric method is more precise than the cum $\sqrt{f(x)}$ method. If eff is greater than one, the accuracy of the geometric method is less than the cum $\sqrt{f(x)}$ method and if eff is equal to one, the accuracy of the two stratification methods are the same.

(ii) In sample size planning;

The relative efficiency may be interpreted as the proportional increase or decrease in the sample size of the geometric method to obtain the same precision as the cum $\sqrt{f(x)}$ method. For example, if the relative efficiency of the geometric method compared to the cum $\sqrt{f(x)}$ method is 0.8 with a sample of size $n = 100$, then the cum $\sqrt{f(x)}$ method needs a sample of size $n = 125$ (i.e. 100 / 0.8) to give the same precision. Similarly, if the relative efficiency is 1.25 based on a sample of size $n = 100$, the geometric method with $n = 100$ will give the same precision as the cum $\sqrt{f(x)}$ method with a sample of size $n = 80$ (i.e. 100 / 1.25).

Table 4.4 gives the efficiency for 3, 4 and 5 strata for each population.

Table 4.4: Efficiency of Geometric Relative to Cum $\sqrt{f(x)}$

	Stratum		
Population	3	4	5
1	0.95	0.90	1.01
2	1.01	0.93	0.89
3	1.26	0.89	0.98
4	0.89	0.98	0.94

The results in Table 4.4 show that gains are observed for the geometric method in the majority of cases. It should be noted that while the geometric method is not always more efficient than the cum $\sqrt{f(x)}$ method when it is, it is substantially better and when it is not, it is only marginally worse. For example, the values that are greater than 1 are, with one exception, within 1.05. The exception is population 3 with $L = 3$ which gives a value of 1.26.

Note, the efficiency may also be written in terms of the coefficients of variation as:

$$eff = \left(\frac{cv_{geom}(\overline{x}_{st})}{cv_{cum}(\overline{x}_{st})} \right)^2 \tag{4.3.2}$$

where $cv_{geom}(\overline{x}_{st})$ and $cv_{cum}(\overline{x}_{st})$ are the coefficients of variation of the stratified sample mean respectively with the geometric method and the cum $\sqrt{f(x)}$ method.

Recall that with Neyman allocation,

$$V_{opt}(\overline{x}_{st}) = \frac{\left(\sum_{h=1}^{L} W_h S_h \right)^2}{n}$$

assuming the finite population correction factor can be ignored. It is clear that the relative efficiency defined in (4.3.1) is independent of sample size n, therefore it can be deduced that the relative efficiency calculated in Table 4.4 for $n = 100$ pertains to any sample size. The results in Table 4.4 can be interpreted as the proportional increase or decrease in the sample size using the cum $\sqrt{f(x)}$ method to obtain the same precision as that obtained with geometric stratification.

(ii) Stratum Breaks, Stratum Sizes, Stratum Sample Sizes and Variability of Stratum Coefficients of Variation

From Tables 4.1, 4.2 and 4.3 it can be seen that the two methods define very different stratum breaks k_h, leading to different stratum sizes N_h and stratum sample sizes n_h for the two methods in all cases.

A cursory examination of the coefficients of variation in Tables 4.1, 4.2 and 4.3 suggests that the geometric method is more successful than the cum $\sqrt{f(x)}$ method in obtaining near-equal cv_h in most cases. For example, in population 1, which has the highest skewness, the cv_h differ substantially from each other when the cum $\sqrt{f(x)}$ method is used to make the breaks, while the geometric method appears to achieve near-equal cv_h in all cases of 3, 4 and 5 strata: the best results in terms of equality of cv_h are obtained with $L = 5$. In the other three populations, the cv_h are not as diverse with the cum $\sqrt{f(x)}$ method, but they still appear more variable than those obtained with the geometric method of stratum construction.

The homogeneity of cv_h between strata is better when L = 4 or 5 than when L = 3; this is to be expected since the validity of the assumption of uniformity of the distribution of units within strata is strengthened with increased number of strata.

Figure 4.1 gives a graphical display of the variability of the cv_h between strata. With just three exceptions, the standard deviations of the cv_h are substantially lower with the geometric method of stratum construction than with the cum $\sqrt{f(x)}$ method. In some cases the difference is of the order of 10. For example, with 4 strata in population 3, the standard deviation of the cv_h is 0.006 with the geometric compared with 0.059 with the cum $\sqrt{f(x)}$ method. The exceptions occur with L = 5 in population 4 and with L = 3, in populations 2 and 4. However the differences between them are not great. It can therefore be concluded that the geometric algorithm is more successful than the cum $\sqrt{f(x)}$ method in breaking the strata such that the cv_h are near-equal.

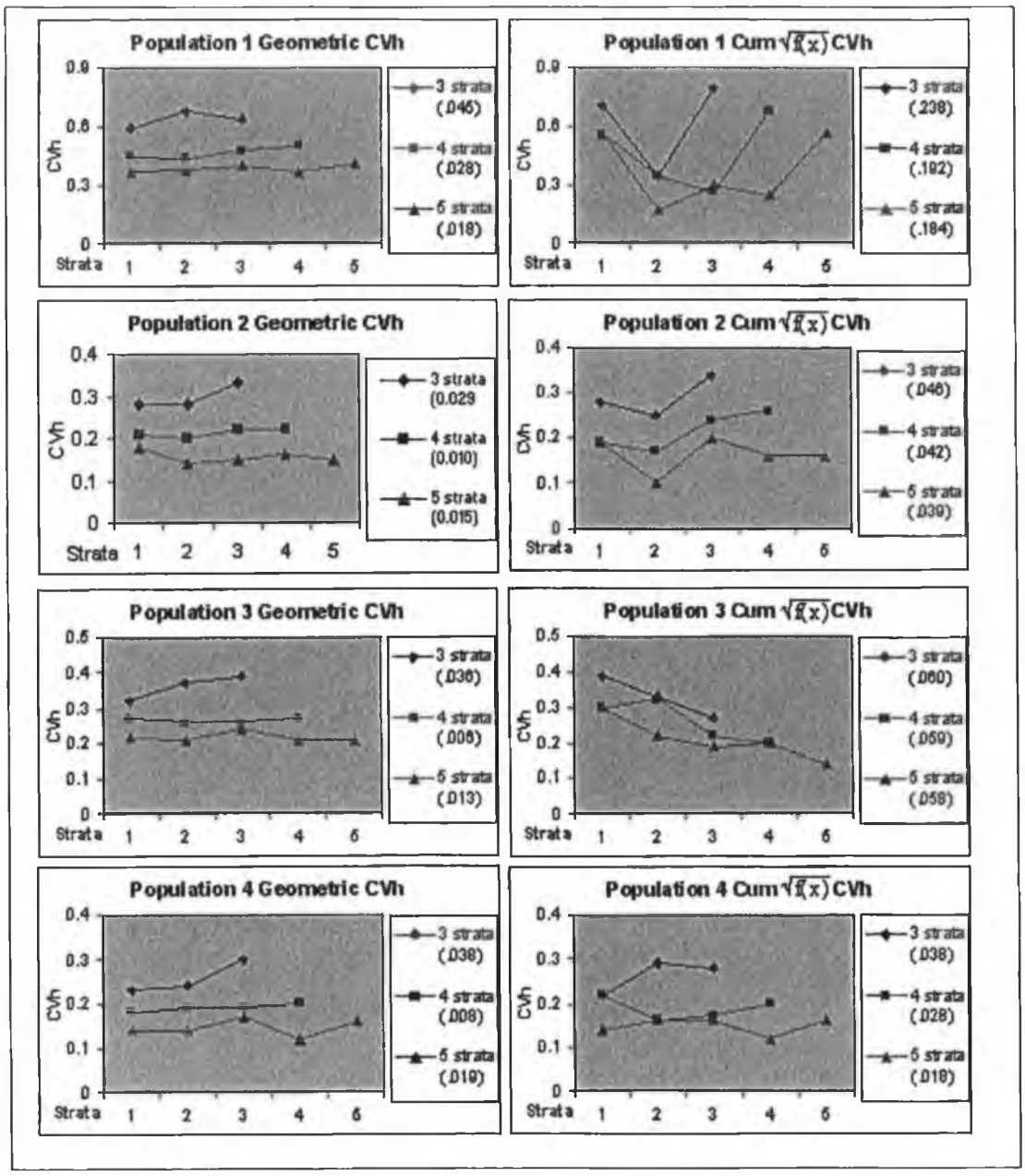

Figure 4.1: Strata Coefficients of Variation for Geometric and Cum $\sqrt{f(x)}$ Methods
* Note: Values in the legends represent the standard deviations of the strata coefficients of variation for each design.

4.3.2 Comparison with the Lavallée-Hidiroglou Method

Our final comparison relates to how the geometric method compares to the Lavallée-Hidiroglou method for stratifying skewed populations. Recall, the Lavallée-Hidiroglou method described in Chapter 3, uses an iterative procedure to obtain the minimum sample size for a given $cv(\bar{x}_{st})$. Using the $cv(\bar{x}_{st})$ given in the third column of Tables 4.1, 4.2 and 4.3 as input for the Lavallée-Hidiroglou algorithm, the sample sizes required to obtain the same precision as the geometric method with $n = 100$ are computed. The results are given in Table 4.5.

Table 4.5: Boundaries and Sample Size Required with the Lavallée-Hidiroglou Method to Obtain the Same $cv(\overline{x}_{st})$ as the Geometric Method when $n = 100$

		3 Strata					4 Strata						5 Strata						
Pop		n	$cv(\overline{x}_{st})$	1	2	3	n	$cv(\overline{x}_{st})$	1	2	3	4	n	$cv(\overline{x}_{st})$	1	2	3	4	5
1	k_h			1248	8676				442	1828	8411				342	1153	3431	10301	
	N_h			2867	464	38			2086	915	327	41			1846	993	357	147	26
	n_h			42	41	38			16	21	35	41			12	14	17	21	26
	cv_h			.87	.57	.37			.64	.41	.45	.38			.58	.34	.31	.31	.32
		121	.0600				113	.0430					90	.0360					
2	k_h			35	102				19	37	95				14	21	35	80	
	N_h			795	202	41			393	420	176	49			189	270	336	164	79
	n_h			47	35	41			13	21	34	49			4	7	16	30	79
	cv_h			.31	.31	.17			.19	.16	.28	.21			.12	.10	.12	.24	.30
		123	.0270				117	.0194					136	.0144					
3	k_h			1398	4197				740	1505	3819				512	869	1577	3675	
	N_h			481	135	61			256	234	118	69			133	180	185	110	69
	n_h			28	18	61			9	10	15	69			4	5	10	17	69
	cv_h			.41	.30	.24			.32	.18	.25	.27			.27	.15	.16	.23	.27
		107	.0317				103	.0214					105	.0184					
4	k_h			172	361				117	188	359				99	130	189	339	
	N_h			212	85	60			111	112	74	60			70	68	85	71	63
	n_h			22	18	60			7	9	17	60			4	4	8	20	63
	cv_h			.23	.21	.32			.14	.12	.19	.32			.10	.08	.10	.18	.33
		100	.0184				93	.0142					99	.0110					

(i) The Relative Efficiency

The results in Table 4.5 show the sample size n required with the Lavallée-Hidiroglou method to obtain the same precision as the geometric method using a sample size of 100. In all but four cases, the sample size required with the Lavallée-Hidiroglou method is greater than 100 and in many cases substantially greater. For example, population 2 needs sample sizes of 123, 117 and 136 for 3, 4 and 5 strata, respectively. When the sample size required falls below $n = 100$, the drop is not large. In population 4, with 4 and 5 strata, $n = 93$ and $n = 99$ respectively, and in population 1 with 5 strata, a sample size of $n = 90$ will suffice with the Lavallée-Hidiroglou algorithm to obtain the same precision as the geometric method. These results might appear to indicate that the geometric method compares favourably with the Lavallée-Hidiroglou method. However, it should be noted that the geometric method, unlike the Lavallée-Hidiroglou method, does not give a take-all stratum.

(ii) Stratum Breaks, Stratum Sizes, Stratum Sample Sizes and Equality of Stratum Coefficients of Variation

From Table 4.5, it can be seen that the stratum breaks are very different between the geometric method and the Lavallée-Hidiroglou method giving different stratum sizes and stratum sample sizes, with the Lavallée-Hidiroglou algorithm deriving a take-all stratum. The stratum coefficients of variation cv_h given in Table 4.5 are illustrated in Figure 4.2 showing how the cv_h vary for each method for 3, 4 and 5 strata. It can be seen that the variability of the cv_h of the geometric method are less than those of the Lavallée-Hidiroglou method, where the standard deviations are, in all cases, substantially lower with the geometric method than with the Lavallée-Hidiroglou method.

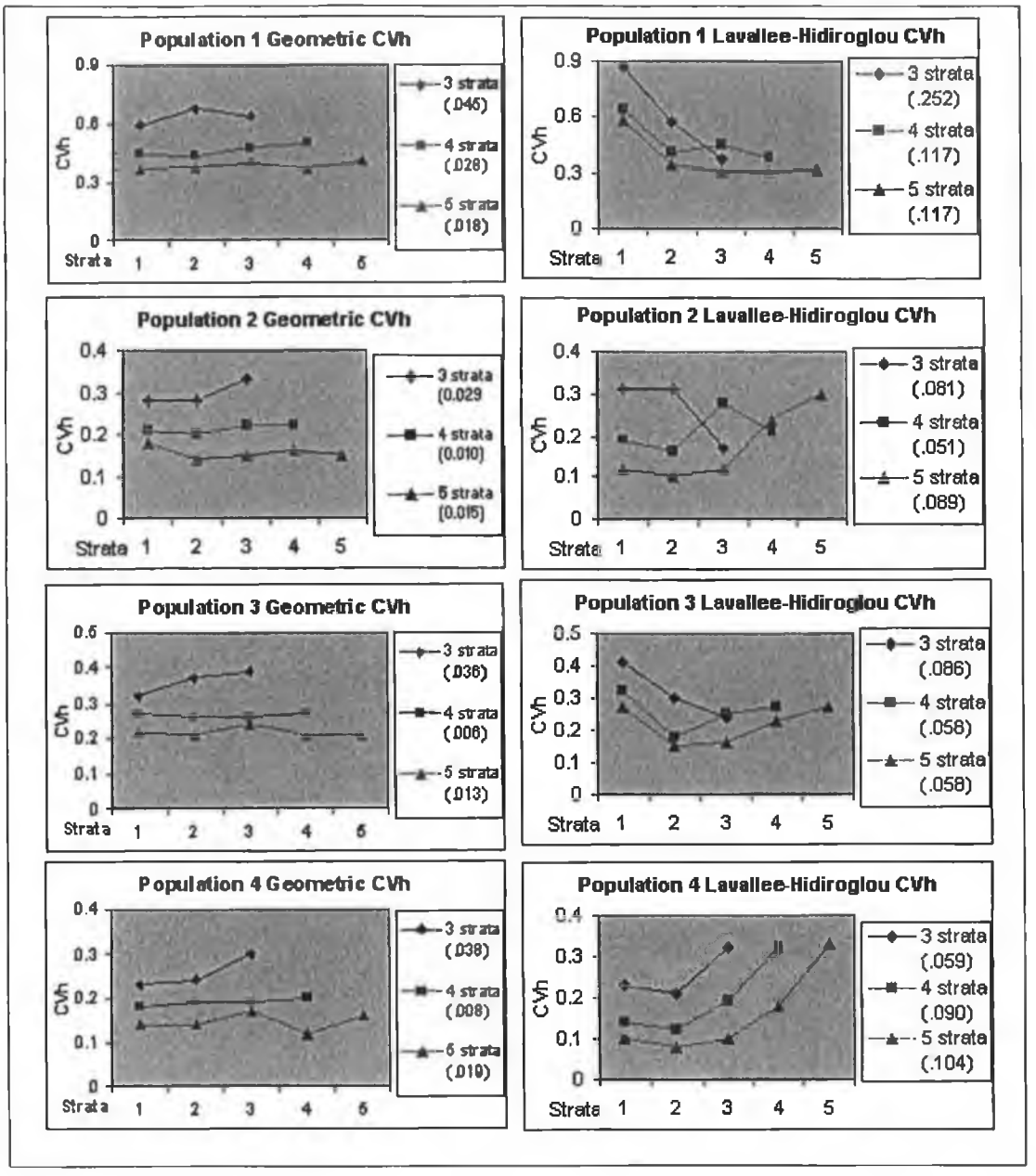

Figure 4.2: Strata Coefficients of Variation for Geometric and Lavallée-Hidiroglou Methods

* Note: Values in the legends represent the standard deviations of the strata coefficients of variation for each design.

4.4 Summary

This chapter derived an algorithm for the construction of stratum boundaries in positively skewed populations. The new method is based on equalising stratum coefficients of variation. It has been shown that near-equal stratum coefficients of variation can be achieved by taking the breaks in geometric progression with positively skewed populations. The proposed method is much easier to implement than the cum $\sqrt{f(x)}$ method or the Lavallée-Hidiroglou (1988) method, and avoids the arbitrariness of these two methods.

A comparison was carried out between the geometric method and the cum $\sqrt{f(x)}$ method. The four positively skewed real populations described in Chapter 3 were divided into 3, 4 and 5 strata. The precision of the stratified sample mean with the geometric method was in many cases as good as, and in some cases better than that of the cum $\sqrt{f(x)}$ method.

Comparisons with the Lavallée-Hidiroglou method indicate the geometric method is more precise. A greater sample size is required to obtain the same precision as the geometric method in most cases. One limitation of the geometric algorithm compared to the Lavallée-Hidiroglou method of stratum construction is that it does not determine a take-all top stratum. This issue is dealt with in the next chapter.

Chapter 5

Improving the Lavallée-Hidiroglou Method

5.1 Introduction

As discussed in Chapter 3, there are some serious implementation problems with the Lavallée-Hidiroglou (1988) algorithm. The boundaries obtained using the algorithm can depend on where the initial starting boundaries are set, so that the minimum sample size attained may be a local but not necessarily a global minimum. The initial starting boundaries can also affect convergence of the iterative process and in some cases the algorithm may not converge at all.

This chapter looks at the initial starting points and the convergence problem of the algorithm. In Section 5.2 we describe the design of the experiments carried out to compare the performance of the algorithm with geometric starting points to those currently in use. In Section 5.3 we look at the problem of convergence of the algorithm. Section 5.4 gives the overall results of the experiments and discusses the number of iterations, sample sizes and boundaries obtained. A summary is given in Section 5.5.

5.2 The Empirical Experiments

The following are the inputs required by the algorithm provided by Rivest (2002) and the values we used in this study.

5.2.1 Coefficients of Variation of the Stratified Sample Mean $cv(\overline{x}_{st})$

A requirement of the algorithm is to specify the coefficient of variation of the stratified sample mean $cv(\overline{x}_{st})$. In this study, three different values of $cv(\overline{x}_{st})$ are used, 0.01, 0.025 and 0.05. These values are based on those used in previous studies of this algorithm (Lavallée and Hidiroglou, 1988 and Chen, 1989).

5.2.2 Number of Strata

As the Lavallée-Hidiroglou algorithm creates a take-all stratum, it was decided to use 4, 5 and 6 strata, creating 3, 4 and 5 take-some strata, respectively.

5.2.3 Starting Points

The Lavallée-Hidiroglou algorithm requires the user to specify the starting points or to use those given with the algorithm. In this chapter we examine the effect of using different starting points on the performance of the algorithm. This is investigated using:

(i) the default starting points given in the algorithm which places the same number of units in each stratum;

(ii) cum $\sqrt{f(x)}$ starting points;

(iii) geometric starting points.

Table 5.1 shows the percentages in each stratum with each of the above set of starting boundaries.

Table 5.1: Percentage of Population in Each Stratum with Each Set of Starting Boundaries

Pop.	Starting Point	1	2	3	4	1	2	3	4	5	1	2	3	4	5	6
1	Default	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	Cum $\sqrt{f(x)}$	53%	31%	9%	7%	53%	18%	18%	7%	4%	53%	18%	13%	6%	8%	2%
	Geometric	42%	41%	14%	3%	31%	38%	22%	8%	1%	25%	31%	27%	11%	5%	1%
2	Default	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	Cum $\sqrt{f(x)}$	38%	41%	15%	6%	38%	32%	17%	9%	4%	38%	32%	9%	11%	6%	4%
	Geometric	44%	38%	13%	5%	35%	40%	13%	8%	4%	26%	41%	15%	9%	6%	3%
3	Default	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	Cum $\sqrt{f(x)}$	33%	48%	11%	8%	33%	41%	13%	7%	6%	33%	32%	16%	9%	5%	5%
	Geometric	20%	51%	19%	10%	14%	38%	29%	11%	8%	11%	26%	34%	14%	8%	7%
4	Default	25%	25%	25%	25%	20%	20%	20%	20%	20%	17%	17%	17%	17%	17%	17%
	Cum $\sqrt{f(x)}$	57%	20%	15%	8%	31%	35%	16%	11%	7%	31%	26%	20%	10%	8%	5%
	Geometric	44%	30%	18%	8%	32%	32%	18%	11%	7%	25%	34%	15%	11%	10%	5%

The first set of starting points are the ones given by the algorithm. As can be seen from Table 5.1, the default method gives an equal percentage in each stratum, putting 25%, 20% and 17% of the population respectively in each stratum for $L = 4, 5$ and 6. For skewed populations this is unlikely to be anywhere near an optimum: it is much more likely that there will be a large percentage of the population in the lower strata and a smaller percentage in the higher.

The second set of initial boundaries follows Slanta and Krenzke (1994) who used the cum $\sqrt{f(x)}$ method to obtain starting points. From Table 5.1, it can be seen that the cum $\sqrt{f(x)}$ starting points place a large percentage of the population in the lowest stratum and a smaller percentage in the higher strata.

Recall that a number of researchers observed that stratum coefficients of variation tend to be equalised with optimum design. This was discussed in Section 4.1. In Chapter 4 we stratified the four skewed populations detailed in Chapter 3 using the geometric method and obtained near-equal stratum coefficients of variation. We use geometric breaks as our third set of starting points to get us close to the optimum at the first stage of the iterative process. We can see from Table 5.1 that geometric starting points, like the cum $\sqrt{f(x)}$ starting points, place a large percentage of the population in the lower strata and a smaller percentage in the higher strata; there is never more than 10% of the population in the top stratum and always a large proportion in the lower strata, which is appropriate for skewed populations.

5.2.4 Allocation Methods

As we have already noted, it has been found that using the algorithm with Neyman allocation results in a less stable algorithm than when used with power allocation (Rivest 2002). Lavallée and Hidiroglou used power allocation of sample units instead

of the optimum method, Neyman allocation. Both power and Neyman are used in this study.

5.2.5 Sampling Strategies

In this experiment we examined the performance of the algorithm with different starting points and different allocation methods. The following are the five sampling strategies used:

(i) geometric starting points with Neyman allocation (Geometric);

(ii) cum $\sqrt{f(x)}$ starting points with Neyman allocation (Cum $\sqrt{f(x)}$);

(iii) default starting points with Neyman allocation (Default);

(iv) default starting points with power allocation using $p = 0.7$. Lavallée and Hidiroglou used this option and showed that for any given level of accuracy the value of the power "p" has only a minor impact on the resulting sample sizes. We follow Rivest (2002) and use $p = 0.7$ (p-Default);

(v) The boundaries are first obtained with default starting points and power allocation with $p = 0.7$ (option (iv)). These k_h are then used as starting points in a second application of the algorithm with Neyman allocation of the sample units. This is a suggestion by Rivest (2002) who proposed running the algorithm in several intermediate designs to get the final sampling design, with the boundaries obtained at one step used as starting points for the next step (Two-stage).

We started the experiment by first applying the program provided by Rivest (2002), which sets the maximum number of iterations to 30, with the above inputs to the four populations described in Chapter 3. We encountered convergence problems and in an attempt to solve these, we modified the program by increasing the maximum number

of iterations and ran it a second time. The convergence problems encountered are discussed in the next section.

5.3 Convergence Problems

The number of iterations required by the algorithm may not be all that important, and indeed even go unnoticed by the user, since this work is done by the computer. However, when non-convergence occurs or the algorithm converges to a non-optimal sample size, the number of iterations may be important. These are discussed next.

5.3.1 Non-Convergence

Non-convergence is where a sample size is not returned within the maximum number of iterations set by the program. In our experiment, there are four cases that did not converge to a sample size within 30 iterations, the maximum number allowable by the program. Allowing the program to run, we obtained the results given in Table 5.2.

Table 5.2: Cases that did not converge within 30 iterations

L	Population	Starting Point	$cv(\bar{x}_{st})$	n	Iterations	Boundaries
5	3	Default	·025	70	53	740, 1505, 3566, 7204
6	1	Default	·010	315	52	190, 438, 849, 1722, 3551
	3	Default	·025	58	35	512, 869, 1580, 3643, 7789
	4	Default	·050	10	33	116, 172, 289, 567, 968

As can be seen from Table 5.2, all cases of non-convergence occur with the default starting boundaries and the larger number of strata ($L = 5$ and 6), with three out of the four cases occurring for $L = 6$. By increasing the number of iterations, all four non-convergence cases successfully converge.

5.3.2 Convergence to Non-Optimal Sample Size

While Table 5.2 shows the cases when a sample size was not returned for the given inputs within a maximum number of iterations allowed of 30, allowing the program to run, we discover six cases where the sample size returned within 30 iterations could be reduced. Table 5.3 shows these cases. The first row of each case gives the sample size obtained at the 30^{th} iteration and the second row shows the reduced sample size.

Table 5.3: Cases that did not return an optimum sample size within 30 iterations

L	Population	Starting Point	$cv(\bar{x}_{st})$	n	Iterations	Boundaries
5	1	Default	·025	154	29	286, 870, 2389, 6859
				146	48	339, 1092, 2972, 7514
		Default	·010	386	29	230, 572, 1262, 2977
				384	37	236, 589, 1287, 2995
		$p-$Default	·025	152	29	281, 924, 2611, 7176
				150	45	317, 1067, 2972, 7852
	3	Geometric	·025	73	29	735, 1432, 3049, 6485
				70	36	740, 1505, 3566, 7204
6	1	Geometric	·025	110	29	247, 668, 1609, 3668, 8876
				109	43	267, 732, 1688, 3700, 8894
		Default	·025	119	29	198, 494, 1200, 3046, 8004
				109	65	267, 732, 1688, 3700, 8893

From Table 5.3 we observe that all the cases that failed to obtain optimal sample size occur with the larger number of strata ($L = 5$ and 6) and with just two exceptions, with the default starting points. The cases with the greatest improvement in sample size occurs with population 1 with default starting points for $cv(\bar{x}_{st}) = .025$ for $L = 5$ and 6, where an extra 19 and 36 iterations reduce the sample size by 8 and 10 units, respectively. For the other cases, the decrease in sample size was 3 units or less.

5.4 The Overall Results

The complete set of results obtained when the program is allowed to run for all four populations divided into 4, 5 and 6 strata with $cv(\bar{x}_{st}) = .05$, .025 and .01 for the five sampling strategies are given in Tables 5.4, 5.5 and 5.6. We examine the number of iterations, samples sizes and boundaries obtained in Sections 5.4.1, 5.4.2 and 5.4.3.

Table 5.4: Boundaries, Sample Sizes and Iterations with 4 Strata

		$cv(\bar{x}_{st}) = \cdot 05$			$cv(\bar{x}_{st}) = \cdot 025$			$cv(\bar{x}_{st}) = \cdot 01$		
Pop.	Starting Point	n	Iterations	Boundaries	n	Iterations	Boundaries	n	Iterations	Boundaries
1	Geometric	92	25	498, 2216, 10133	212	16	387, 1476, 5382	497	12	333, 1029, 2563
	Cum $\sqrt{f(x)}$	92	24	498, 2216, 10133	212	16	387, 1476, 5382	496	9	333, 1030, 2564
	Default	92	29	498, 2216, 10133	212	25	387, 1476, 5382	498	29	284, 845, 2238
	p-Default	93	22	485, 2221, 10142	213	21	373, 1493, 5395	501	29	267, 837, 2280
	Two-stage	92	5	498, 2216, 10133	212	5	387, 1476, 5382	499	8	285, 848, 2254
2	Geometric	36	10	21, 53, 195	88	4	20, 41, 112	213	11	20, 33, 63
	Cum $\sqrt{f(x)}$	36	11	21, 53, 195	90	5	19, 39, 110	212	8	19, 33, 63
	Default	36	14	21, 53, 195	90	12	19, 39, 110	247	7	15, 23, 45
	p-Default	33	20	30, 74, 195	88	13	21, 44, 111	214	15	19, 32, 45
	Two-stage	34	3	31, 72, 195	88	3	21, 43, 113	212	3	20, 32, 59
3	Geometric	37	25	1366, 3757, 9466	98	11	744, 1574, 4171	188	13	731, 1328, 2350
	Cum $\sqrt{f(x)}$	37	21	1366, 3757, 9466	98	9	807, 1764, 4432	188	17	731, 1328, 2350
	Default	37	21	1366, 3757, 9446	98	14	744, 1574, 4171	187	25	722, 1297, 2300
	p-Default	39	22	1260, 3704, 9446	98	15	734, 1653, 4118	192	21	665, 1268, 2404
	Two-stage	37	5	1367, 3758, 9446	97	7	769, 1607, 4190	187	10	723, 1298, 2300
4	Geometric	24	18	174, 387, 968	55	9	150, 277, 566	124	8	141, 245, 359
	Cum $\sqrt{f(x)}$	24	13	174, 387, 968	55	6	157, 282, 566	125	9	149, 250, 360
	Default	24	26	174, 387, 968	55	24	150, 277, 566	113	10	116, 171, 279
	p-Default	25	20	174, 389, 919	55	19	148, 286, 562	114	8	115, 173, 279
	Two-stage	24	4	175, 388, 968	55	3	151, 278, 567	113	2	117, 172, 280

Table 5.5: Boundaries, Sample Sizes and Iterations with 5 Strata

		$cv(\bar{x}_{st}) = \cdot 05$			$cv(\bar{x}_{st}) = \cdot 025$			$cv(\bar{x}_{st}) = \cdot 01$		
Pop.	Starting Point	n	Iterations	Boundaries	n	Iterations	Boundaries	n	Iterations	Boundaries
1	Geometric	57	24	367, 1248, 3757, 13226	146	29	339, 1090, 2970, 7513	384	12	249, 670, 1565, 3288
	Cum $\sqrt{f(x)}$	57	24	367, 1248, 3757, 13226	146	29	339, 1092, 2971, 7513	383	8	260, 688, 1606, 3335
	Default	57	29	360, 1238, 3752, 13226	146	48	339, 1092, 2972, 7514	384	37	236, 589, 1287, 2995
	p-Default	58	29	339, 1246, 3974, 13555	150	45	317, 1067, 2972, 7852	387	29	218, 582, 1344, 3080
	Two-stage	57	7	368, 1276, 3955, 13562	147	7	339, 1093, 2992, 7632	383	9	243, 619, 1383, 3130
2	Geometric	20	8	19, 34, 73, 195	62	6	19, 31, 58, 132	171	6	19, 31, 55, 91
	Cum $\sqrt{f(x)}$	20	8	19, 34, 73, 195	62	6	19, 31, 58, 132	172	6	18, 30, 54, 90
	Default	20	20	19, 34, 73, 195	77	12	14, 22, 42, 116	179	16	14, 21, 33, 66
	p-Default	20	22	21, 42, 94, 195	62	19	19, 33, 61, 128	183	10	15, 22, 34, 60
	Two-stage	18	8	21, 42, 104, 195	62	3	20, 33, 59, 133	179	7	15, 22, 34, 67
3	Geometric	23	18	742, 1534, 3807, 9446	70	36	740, 1505, 3566, 7204	159	18	579, 925, 1440, 2673
	Cum $\sqrt{f(x)}$	23	8	742, 1534, 3807, 9446	70	7	740, 1505, 3566, 7204	156	10	731, 1324, 2234, 3434
	Default	23	23	742, 1534, 3807, 9446	70	53	740, 1505, 3566, 7204	160	14	511, 857 , 1370, 2456
	p-Default	24	20	735, 1658, 4111, 9446	78	28	670, 1287, 2491, 5181	160	15	488, 839 , 1377, 2453
	Two-stage	23	6	769, 1621, 4127, 9446	70	24	740, 1505, 3567, 7204	160	6	512, 857 , 1370, 2456
4	Geometric	17	9	118, 195, 405, 968	41	5	118, 189, 353, 651	103	5	117, 185, 348, 503
	Cum $\sqrt{f(x)}$	17	9	118, 195, 405, 968	41	6	118, 189, 356, 652	103	6	118, 185, 348, 503
	Default	18	19	117, 195, 405, 968	43	16	116, 172, 289, 599	105	7	99, 129, 178, 298
	p-Default	15	29	149, 288, 553, 968	41	25	119, 198, 353, 646	106	7	101, 134, 183, 283
	Two-stage	14	3	152, 282, 567, 968	41	6	119, 190, 357, 653	105	5	103, 134, 182, 298

Table 5.6: Boundaries, Sample Sizes and Iterations with 6 Strata

		$cv(\overline{x}_{st}) = \cdot 05$			$cv(\overline{x}_{st}) = \cdot 025$			$cv(\overline{x}_{st}) = \cdot 01$		
Pop.	Starting Point	n	Iter.	Boundaries	n	Iter.	Boundaries	n	Iter.	Boundaries
1	Geometric	43	29	269, 741, 1767, 4378, 14915	109	43	267, 732, 1688, 3700, 8894	318	16	199, 484, 1044, 2125, 3936
	Cum $\sqrt{f(x)}$	43	26	269, 741, 1767, 4378, 14915	109	29	267, 732, 1687, 3700, 8893	313	14	233, 566, 1127, 2183, 4040
	Default	43	29	240, 639, 1619, 4295, 14829	109	65	267, 732, 1688, 3700, 8893	313	52	190, 438, 849, 1722, 3551
	p-Default	43	29	241, 703, 1818, 4782, 14764	112	29	217, 589, 1415, 3431, 8464	320	29	158, 383, 803, 1670, 3496
	Two-stage	40	8	270, 743, 1808, 4683, 15574	110	15	268, 733, 1688, 3700, 8894	315	11	191, 439, 850, 1722, 3551
2	Geometric	11	23	19, 31, 57, 110, 195	53	4	16, 25, 40, 69, 144	146	6	16, 25, 40, 67, 99
	Cum $\sqrt{f(x)}$	11	17	19, 31, 57, 110, 195	53	7	18, 27, 42, 69, 144	145	5	18, 27, 39, 65, 98
	Default	16	18	14, 21, 34, 73, 195	55	18	13, 20, 31, 58, 139	163	11	13, 17, 22, 34, 68
	p-Default	12	26	19, 32, 58, 108, 195	56	14	15, 22, 34, 61, 126	171	8	13, 18, 24, 35, 60
	Two-stage	11	3	19, 32, 58, 111, 195	54	3	15, 22, 33, 60, 140	162	5	13, 18, 24, 35, 69
3	Geometric	20	19	523, 909, 1665, 4133, 9446	58	16	512, 869, 1580, 3643, 7789	126	16	511, 857, 1363, 2240, 3496
	Cum $\sqrt{f(x)}$	16	10	723, 1311, 2303, 4605, 9446	52	11	723, 1304, 2234, 3782, 7857	148	13	722, 1295, 2226, 3555, 5332
	Default	20	27	523, 909, 1665, 4133, 9446	58	35	512, 869, 1580, 3643, 7789	143	16	428, 683, 969, 1480, 2839
	p-Default	17	29	667, 1278, 2403, 4800, 9446	58	29	520, 941, 1746, 3659, 7436	146	29	425, 695, 1012, 1565, 2666
	Two-stage	17	14	732, 1334, 2362, 4718, 9446	57	7	614, 1019, 1801, 3713, 7795	141	6	432, 707, 997, 1528, 2873
4	Geometric	10	12	116, 172, 289, 567, 968	32	6	116, 170, 257, 387, 680	74	6	116, 170, 256, 380, 516
	Cum $\sqrt{f(x)}$	10	10	116, 172, 289, 567, 968	32	8	116, 171, 257, 387, 680	74	7	116, 170, 256, 380, 516
	Default	10	33	116, 172, 289, 567, 968	39	9	93, 120, 172, 289, 607	81	10	93, 120, 170, 256, 387
	p-Default	11	29	118, 195, 341, 599, 968	32	26	115, 171, 259, 401, 661	81	10	94, 125, 173, 257, 383
	Two-stage	9	4	118, 190, 352, 602, 968	32	4	117, 171, 258, 388, 681	81	2	95, 125, 173, 257, 388

5.4.1 Number of Iterations

There is a huge difference in the number of iterations required with different starting points, as can be seen from Tables 5.4, 5.5 and 5.6. To establish whether or not the differences in iterations between geometric and each of the other starting points are significant, pairwise comparison t-tests are used for 4, 5 and 6 strata. Table 5.7 gives the mean for each design, the mean of the differences ($diff$), the standard error of the mean differences (SE), the value of Student's t-statistic for testing differences in pairs of observations (t) and the significance of the t-test (sig).

Table 5.7: Significance of the Mean Iterations

	Mean	Diff	SE	t	Sig
4 Strata					
Geometric	13.50				
Cum $\sqrt{f(x)}$	12.33	1.17	0.757	1.541	0.076
Default	20.25	-6.75	1.728	-3.906	0.001
p-Default	18.75	-5.25	1.702	-3.085	0.005
Two-stage	4.83	8.67	1.831	4.733	0.001
5 Strata					
Geometric	14.58				
Cum $\sqrt{f(x)}$	10.58	4.00	2.510	1.593	0.070
Default	24.50	-9.92	2.268	-4.373	0.001
p-Default	23.17	-8.58	2.718	-3.158	0.005
Two-stage	7.58	7.00	2.250	3.112	0.005
6 Strata					
Geometric	16.33				
Cum $\sqrt{f(x)}$	13.08	3.25	1.382	2.351	0.019
Default	26.92	-10.58	3.452	-3.066	0.006
p-Default	23.92	-7.58	2.656	-2.856	0.008
Two-stage	6.83	9.50	2.551	3.724	0.002

From Table 5.7 it can be observed that the mean number of iterations with cum $\sqrt{f(x)}$ and geometric starting points do not differ significantly from one another for $L = 4$ and 5. The two-stage method has the lowest number of iterations, however, recall that this represents the second stage only; to get to this stage, $p-$default was implemented at the first stage. Thus the true number of iterations is the sum of the two stages, making the mean number of iterations for this method higher than all the others. The mean number of iterations with geometric starting points is significantly less than the mean with the default methods in all cases ($p < .05$). Figures 5.1 and 5.2 illustrate further these significant differences in the number of iterations required to obtain optimum sample sizes using the geometric starting points compared with the default starting points for the four populations with $cv(\bar{x}_{st}) = 0.05, 0.025$ and 0.01 for 4, 5 and 6 strata.

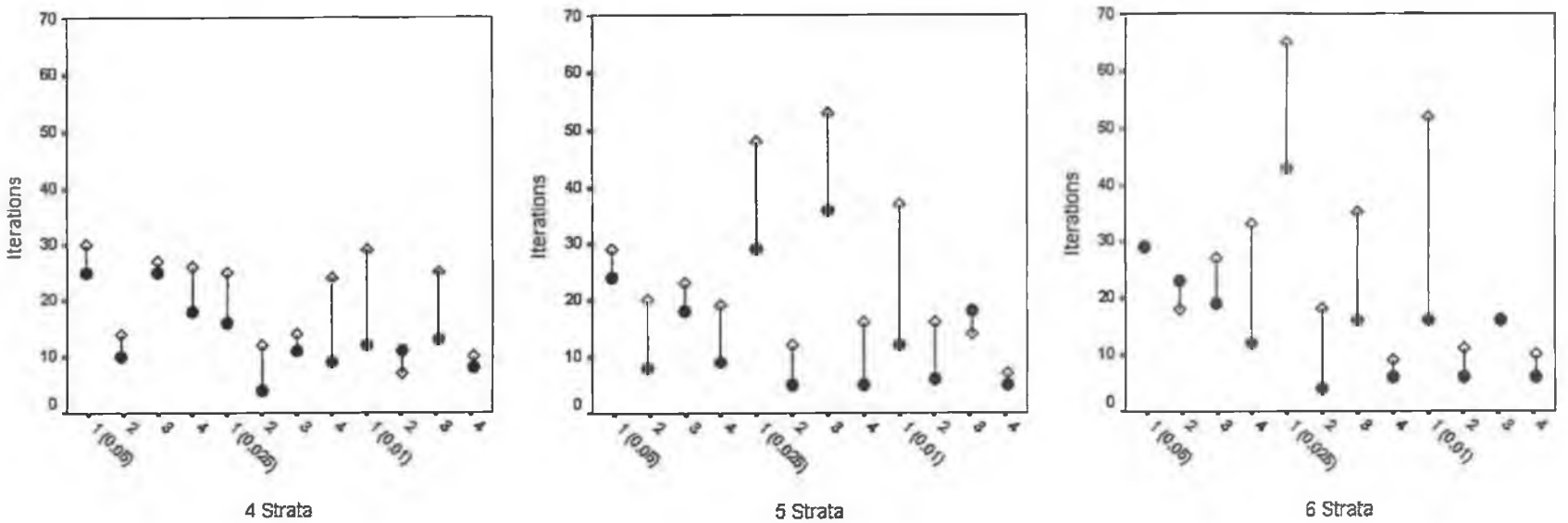

Figure 5.1: Iterations for $cv(\overline{x}_{st}) = .05$, .025 and .01 with Geometric ($\bullet$) and Default ($\Diamond$) starting boundaries

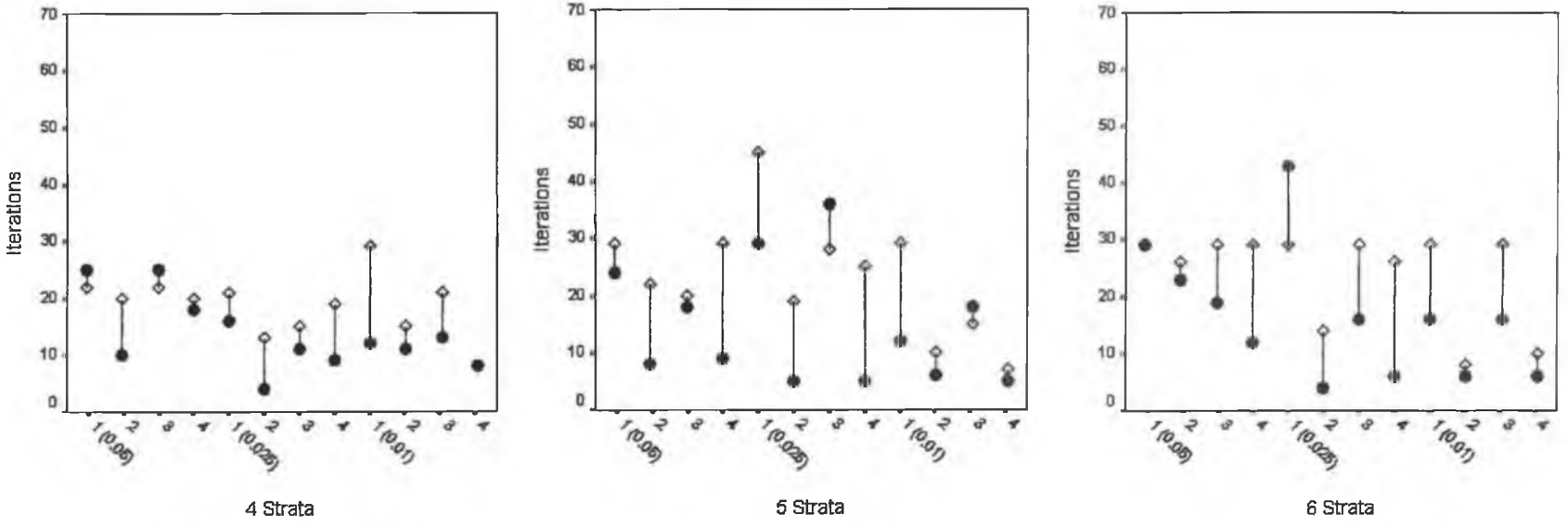

Figure 5.2: Iterations for $cv(\overline{x}_{st}) = .05$, .025 and .01 with Geometric ($\bullet$) and p-Default ($\diamondsuit$) starting boundaries

From Figures 5.1 and 5.2, it is clear that the lower points of the lines are occupied by the geometric in most cases indicating that this strategy converges faster. The largest differences in the number of iterations occur with $L = 5$ and $L = 6$. The following are the exceptions:

Default starting points converge faster than the geometric starting points for:

- $L = 4$ with population 2 for $cv(\overline{x}_{st}) = 0.01$;
- $L = 5$ with population 3 for $cv(\overline{x}_{st}) = 0.01$;
- $L = 6$ with population 2 for $cv(\overline{x}_{st}) = 0.05$.

However, the differences are within 5 iterations.

The $p-$default starting points converge faster than the geometric starting points for:

- $L = 4$ with populations 1 and 3 for $cv(\overline{x}_{st}) = 0.05$;
- $L = 5$ with population 3 for $cv(\overline{x}_{st}) = 0.025$ (difference of 8 iterations) and $cv(\overline{x}_{st}) = 0.01$;
- $L = 6$ with population 1 for $cv(\overline{x}_{st}) = 0.025$ (difference of 14 iterations).

The differences are within 3 iterations except in the two cases specified. It should be noted that in all of the above cases, the increased number of iterations resulted in smaller sample sizes for geometric starts.

5.4.2 Sample Sizes

A preliminary inspection of Tables 5.4, 5.5 and 5.6 indicates that the sample sizes needed to obtain a given coefficient of variation $cv(\overline{x}_{st})$ vary across starting points.

To look at the overall picture relating to sample size, the differences in sample sizes between geometric and each of the other starting points for 4, 5 and 6 strata are examined using pairwise comparison t-tests to investigate if the differences are significant. Table 5.8 gives the mean for each design, the mean of the differences ($diff$), the standard error of the mean differences (SE), the value of Student's t-statistic for testing differences in pairs of observations (t) and the significance of the t-test (sig).

Table 5.8: Significance of the Mean Sample Sizes

	Mean	Diff	SE	t	Sig
4 Strata					
Geometric	138.67				
Cum $\sqrt{f(x)}$	138.75	-0.080	0.229	-0.364	0.362
Default	140.83	-2.167	3.049	-0.710	0.246
p-Default	138.67	0.000	1.059	0.000	0.500
Two-stage	137.50	1.167	0.936	1.246	0.119
5 Strata					
Geometric	104.42				
Cum $\sqrt{f(x)}$	104.17	0.250	0.279	0.897	0.195
Default	106.83	-2.417	1.317	-1.835	0.047
p-Default	107.00	-2.583	1.131	-2.284	0.021
Two-stage	104.92	-0.500	0.783	-0.638	0.268
6 Strata					
Geometric	83.33				
Cum $\sqrt{f(x)}$	83.83	-0.500	2.058	-0.243	0.407
Default	87.75	-4.417	1.897	-2.328	0.020
p-Default	88.25	-4.917	2.487	-1.976	0.037
Two-stage	85.75	-2.417	1.928	-1.253	0.118

From Table 5.8 it can be observed that the mean sample size with geometric starts is less than or equal to the mean with the other methods in all cases except for 4 strata with the two-stage and for 5 strata with the cum $\sqrt{f(x)}$ but these are not significant.

The following significant results are observed:

- With 5 strata, the mean with geometric starting points is significantly less than the mean with default ($p = 0.047$) and p-default starts ($p = 0.021$).
- With 6 strata, the geometric method returns samples sizes significantly less than default ($p = 0.020$), and p-default ($p = 0.037$).

These significant differences are discussed next.

5.4.2.1 Geometric versus Default

As we have seen from Table 5.8, there are significant differences between the mean sample size with geometric starts and default starts for $L = 4$ and 5. The boxplot in Figure 5.3 illustrates the differences in sample sizes between the two strategies (geometric - default). A negative difference in sample size indicates that the sample size obtained using geometric starts is less than that with default starts while a positive difference indicates that default starts give a smaller sample than geometric starts.

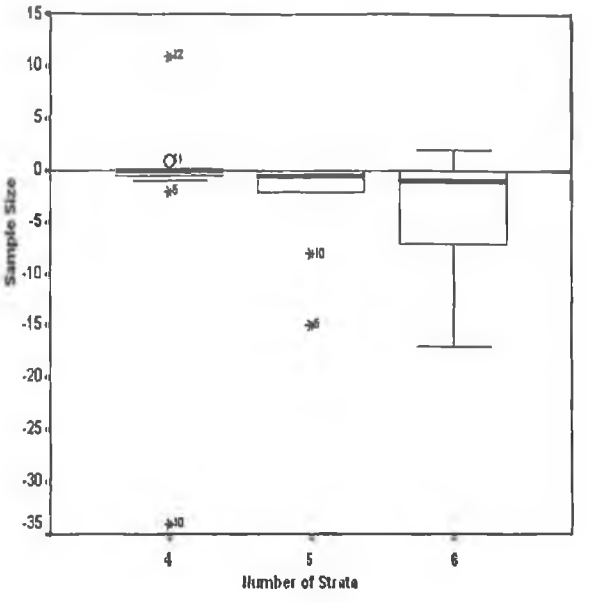

Figure 5.3: Differences in Sample Sizes (Geometric - Default)

From Figure 5.3, we observe that the geometric method yields sample sizes less than or equal to that obtained with default in most cases; in approximately 50% of the cases, the sample sizes are less than default, and sometimes substantially less. The greatest improvements in the sample sizes for geometric starting points occur with the larger number of strata. Most of the boxplot can be seen to be within 1 and 2 units of the zero-line for 4 and 5 strata. For 6 strata the lower quartile is -7, which indicates that 25% of the sample sizes with geometric are 7 units less than those with default. The following observations are made:

For 4 strata, the sample sizes coincide or are within one or two units of each other in all except two cases:

- in population 2 with $cv(\overline{x}_{st}) = \cdot 01$, $n = 247$ with default starts compared to $n = 213$ when the starting points are in geometric progression, an increase of 16%.
- The only major discrepancy in favour of default starts occurs with population

4 where the default starting points needed 9% less sampling units to attain $cv(\overline{x}_{st}) = 0.01$ than the geometric start method; $n = 124$ with geometric starting points, compared to $n = 113$ with default starts.

With 5 strata

- geometric starting points yielded sample sizes less than default in half of the cases. The greatest decrease is in population 2 with $cv(\overline{x}_{st}) = .025$, where $n = 62$ with geometric starts increased to $n = 77$ when the starts were default, a 24% increase.

In the case of 6 strata

- with $cv(\overline{x}_{st}) = \cdot 01$, $n = 146, 126$ and 74 with geometric starts in populations 2, 3 and 4 respectively compared to $n = 163, 143$ and 81 with default starts. This represents a percentage increase in sample sizes of 12%, 13% and 9% respectively when default starts are used.

- with $cv(\overline{x}_{st}) = \cdot 025$, $n = 32$ with the geometric method in population 4 compared with $n = 39$ with default starts, an increase of 22% when the starting points are default.

- with $cv(\overline{x}_{st}) = \cdot 05$, $n = 11$ with geometric starts in population 2 compared to $n = 16$ with default starts, an increase of 45%.

5.4.2.2 Geometric versus p-Default

As we have seen from Table 5.8, there are significant differences between the mean sample size with geometric starts and $p-$default starts for $L = 4$ and 5. The boxplot in Figure 5.4 illustrates the differences in sample sizes between the two strategies.

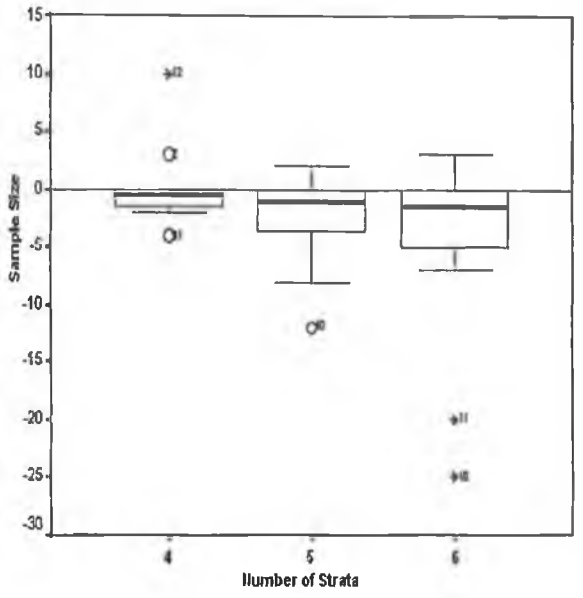

Figure 5.4: Differences in Sample Sizes (Geometric - $p-$Default)

The sample sizes obtained with geometric starts are less than or equal to those obtained with the $p-$default method in most cases. Figure 5.4 shows that for $p-$default most of the boxplot is within 2 units of the zero-line for 4 strata. For 6 strata the lower quartile is -5, which indicates that 25% of the sample sizes with geometric are 5 units less than those with $p-$default. With just a few exceptions, the $p-$default starting points yield greater sample sizes than the geometric.

For 4 strata there is one case in which the p-default yielded a sample size substantially less than the geometric:

- population 4, $n = 124$ for the geometric when compared with $n = 114$ for p-default when $cv(\overline{x}_{st}) = .01$; 10 units less with the p-default method.

At first glance this decrease appears a surprising result since the p-default method of obtaining boundaries uses power allocation which is not optimal, and should therefore not yield sample sizes smaller than optimal allocation which is used with the geometric method. Further examination of Table 5.4 indicates that the

boundaries obtained are quite different for each method. It is possible that in this case the method with the geometric starting boundaries led to a local rather than a global optimum. Notice also that this is the least skewed of the populations with the lowest number of strata. The geometric method works best on populations which are highly skewed and for large L.

For 5 strata, the major differences are:

- $cv(\overline{x}_{st}) = .01$ for population 2 $n = 171$ for geometric starts compared to $n = 183$ for the p-default method; 12 units or 7% increase;
- $cv(\overline{x}_{st}) = .025$ for population 3 $n = 70$ for geometric starts compared to $n = 78$ for the p-default method; 8 units less for the geometric.

For 6 strata, three major differences are:

- $cv(\overline{x}_{st}) = \cdot 01$, $n = 146$, 126, and 74 for geometric starts in populations 2, 3, and 4 compared to $n = 171$, 146, and 81 for the p-default method, 17%, 16% and 9% respective increases.

5.4.3 Boundaries

It can be seen from Tables 5.4, 5.5 and 5.6 that the boundaries are not always the same when different starting points are used: the discrepancies between them are greatest for the highest number of strata ($L = 6$) and the lowest coefficient of variation ($cv(\overline{x}_{st}) = \cdot 01$).

5.5 Summary

In this chapter geometric starting points are used as initial values for the Lavallée-Hidiroglou algorithm and compared with starting points determined using the default method, the cum $\sqrt{f(x)}$ method and a two-stage process suggested by

Rivest (2002). The first thing we noticed is that non-convergence of the algorithm or convergence to a non-optimal sample size is more likely to occur when the number of strata is large and default starting points are used. Default starting points are the only ones that did not converge to a sample size within 30 iterations. On two occasions when geometric starting points did not return an optimal sample size, it was close to the optimum sample size at the 30^{th} iteration. The mean number of iterations required by geometric starting points is less than that required by the default methods and similar to the cum $\sqrt{f(x)}$ starting points. The mean number of iterations for the two-stage process is higher than all the others as it is the sum of the two stages.

Geometric starting points give a mean sample size significantly less than the default starting points for $L = 5$ and 6. Comparisons with cum $\sqrt{f(x)}$ and the two-stage process starting points indicate that mean sample sizes were not significantly different. However, geometric starting points are preferable as it avoids the implementation problems of the cum $\sqrt{f(x)}$ method, discussed in Section 3.2.1.2. It was also observed that using geometric breaks as the initial boundaries is closer to optimal final boundaries than the default starting points as the geometric places a larger proportion in the lower strata and less in the top stratum.

Chapter 6

The Pareto Distribution

6.1 Introduction

As pointed out earlier, the geometric method for obtaining optimum boundaries relies on the same assumption made by Dalenius and Hodges (1959) in deriving their cum $\sqrt{f(x)}$ method that the density function of each stratum has an approximately uniform distribution. However, this is a rough approximation as there is usually only a small number of strata and this type of step function would not occur in practice. In this chapter, we take a different approach and use the Pareto distribution as a model of our skewed data.

In the remainder of the chapter,

(i) the properties of the Pareto distribution are detailed in Section 6.2;

(ii) the moments for the Pareto distribution are given in Section 6.3;

(iii) it is demonstrated that, for a Pareto distribution, taking break points in geometric progression gives equal cv_h in Section 6.4;

(iv) a summary is given in Section 6.5.

6.2 Properties of the Pareto Distribution

The Pareto distribution, a highly positively skewed distribution, is named after the 19th century Italian economist Vilfredo Pareto, who used it to model the considerable skewness in the distribution of wealth. It is often described on the basis of the "80-20 rule". For example, 20% of the population own 80% of the wealth: this was Pareto's empirical observation in Italy at the time. It is also known as the "power law". Applications of the Pareto distribution include the distribution of income and the classification of stock in a warehouse on the basis of frequency of movement (Evans et al., 2000). The generalised Pareto distributions are given by taking

$$f(x) = \begin{cases} \lambda\beta^{\lambda}x^{-\lambda-1}, & x \geq \beta \\ 0, & x < \beta \end{cases} \tag{6.2.1}$$

where $\beta \geq 1$ is the location parameter, $\lambda > 0$ is the shape parameter and $\beta \leq x < \infty$.

The cumulative distribution function is defined as

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{6.2.2}$$

and for the Pareto distribution

$$F(x) = 1 - \beta^{\lambda}x^{-\lambda}. \tag{6.2.3}$$

Figures 6.1 and 6.2 show the Pareto probability density function and cumulative distribution function for $\lambda = 1, 2, 3$ and $\beta = 1$.

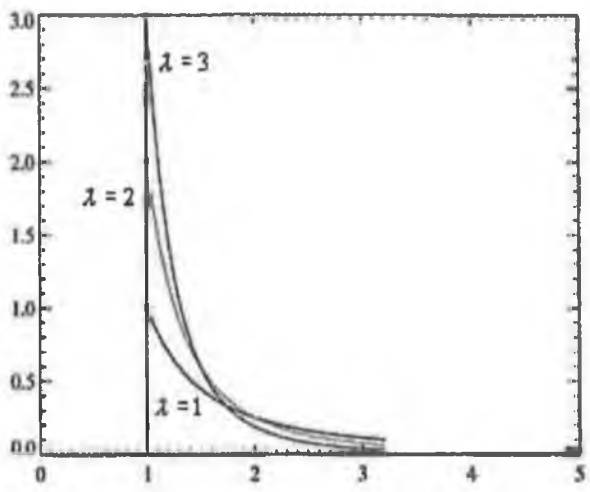

Figure 6.1: Pareto Probability Density Function ($\lambda = 1, 2, 3, \beta = 1$)

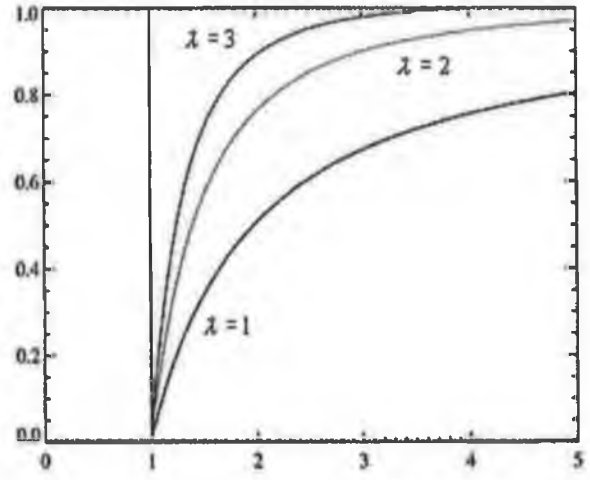

Figure 6.2: Pareto Cumulative Distribution Function ($\lambda = 1, 2, 3, \beta = 1$)

Source: *http://en.wikipedia.org/wiki/Pareto_distribution.*

6.3 Moments of the Distribution

For the Pareto distribution, it can be shown that the mean is:

$$\mu = \frac{\lambda\beta}{\lambda - 1} \tag{6.3.1}$$

which only exists when $\lambda > 1$.

The variance for the Pareto distribution is:

$$\sigma^2 = \frac{\lambda\beta^2}{(\lambda-1)^2(\lambda-2)} \tag{6.3.2}$$

which only exists when $\lambda > 2$.

The skewness for the Pareto distribution is

$$\eta_3 = \sqrt{\frac{\lambda-2}{\lambda}}\ \frac{2(\lambda+1)}{\lambda-3}, \quad \lambda > 3.$$

The coefficient of variation for the Pareto distribution is

$$cv = \frac{1}{\sqrt{\lambda(\lambda-2)}}, \quad \lambda > 2. \tag{6.3.3}$$

6.3.1 Distribution Restricted to an Interval

The area under a density function $f(.)$ restricted to an interval $[a, b]$ where $-\infty < a < b < \infty$, can be written as

$$P(a \leq X \leq b) = \int_{x=a}^{b} f(x)dx. \tag{6.3.4}$$

For the Pareto distribution

$$P(a \leq X \leq b) = \int_{x=a}^{b} f(x)dx = \beta^{\lambda}(a^{-\lambda} - b^{-\lambda}). \tag{6.3.5}$$

6.3.2 The Mean Restricted to an Interval $[a, b]$

The first moment about zero, i.e. the mean, restricted to an interval $[a, b]$ is defined as:

$$\mu_{a,b} = \frac{\int_a^b xf(x)dx}{\int_a^b f(x)dx}. \tag{6.3.6}$$

For the Pareto distribution

$$\int_{x=a}^{b} xf(x)dx = \frac{\lambda}{\lambda - 1}\beta^{\lambda}(a^{1-\lambda} - b^{1-\lambda}), \qquad \lambda > 1. \tag{6.3.7}$$

So the mean restricted to $[a, b]$ is

$$\mu_{a,b} = \frac{\lambda(a^{1-\lambda} - b^{1-\lambda})}{(\lambda - 1)(a^{-\lambda} - b^{-\lambda})}. \tag{6.3.8}$$

6.3.3 The Variance Restricted to an Interval $[a, b]$

The second moment about the mean, i.e. the variance, restricted to an interval $[a, b]$ is defined as:

$$\sigma^2_{a,b} = \frac{\int_a^b (x - \mu_{a,b})^2 f(x)dx}{\int_a^b f(x)dx}$$

which can be written as:

$$\sigma^2_{a,b} = \frac{\int_a^b x^2 f(x)dx}{\int_a^b f(x)dx} - \mu^2_{a,b}. \tag{6.3.9}$$

For the Pareto distribution

$$\int_{x=a}^{b} x^2 f(x)dx = \frac{\lambda}{\lambda - 2}\beta^{\lambda}(a^{2-\lambda} - b^{2-\lambda}), \qquad \lambda > 2. \tag{6.3.10}$$

So

$$\sigma^2_{a,b} = \frac{\lambda(a^{2-\lambda} - b^{2-\lambda})}{(\lambda-2)(a^{-\lambda} - b^{-\lambda})} - \left(\frac{\lambda(a^{1-\lambda} - b^{1-\lambda})}{(\lambda-1)(a^{-\lambda} - b^{-\lambda})}\right)^2. \tag{6.3.11}$$

6.3.4 The Coefficient of Variation Restricted to an Interval $[a, b]$

Using the expression for $\sigma^2_{a,b}$ (6.3.11) and $\mu_{a,b}$ (6.3.8), we may write $cv^2_{a,b}$ as

$$cv^2_{a,b} = \frac{(a^{-\lambda} - b^{-\lambda})\frac{\lambda}{\lambda-2}(a^{2-\lambda} - b^{2-\lambda}) - \left(\frac{\lambda}{\lambda-1}(a^{1-\lambda} - b^{1-\lambda})\right)^2}{\left(\frac{\lambda}{\lambda-1}(a^{1-\lambda} - b^{1-\lambda})\right)^2} \tag{6.3.12}$$

which simplifies to:

$$cv^2_{a,b} = \frac{(a^{-\lambda} - b^{-\lambda})\frac{\lambda}{\lambda-2}(a^{2-\lambda} - b^{2-\lambda})}{\left(\frac{\lambda}{\lambda-1}(a^{1-\lambda} - b^{1-\lambda})\right)^2} - 1. \tag{6.3.13}$$

6.4 Geometric Breaks

In this section we demonstrate that, for a Pareto distribution, taking break points in geometric progression gives equal coefficients of variation in each stratum.

Suppose

$$k_0 < k_1 < \ldots k_L,$$

points in a finite range $[k_0, k_L]$ of a Pareto variable, are in geometric progression so that successive ratios are the same:

$$\frac{k_1}{k_0} = \frac{k_2}{k_1} = \ldots = \frac{k_L}{k_{L-1}},$$

we show that for a Pareto distribution

$$cv_{k_0,k_1} = cv_{k_1,k_2} = \ldots = cv_{k_{L-1},k_L},$$

where

$$cv_{k_h,k_{h+1}}$$

is the coefficient of variation restricted to the interval $[k_h, k_{h+1}]$.

Theorem 6.1.

If f is a Pareto density as given in equation (6.2.1), and if the endpoints a, b, c of successive intervals $[a, b]$, $[b, c]$ form a geometric progression then the coefficients of variation in $[a, b]$ and $[b, c]$ are equal i.e.

$$cv_{a,b} = cv_{b,c}. \tag{6.4.1}$$

Proof:

Writing the endpoints for the two intervals $[a, b]$ and $[b, c]$ where

$$a < b < c$$

as multiples of the boundary break b then

$$[sb, b]$$

is the first interval and

$$[b, rb]$$

is the second interval where

$$0 < s < 1 < r.$$

Assuming the breaks are made in geometric progression i.e.

$$(sb)(rb) = b^2,$$

then

$$rs = 1. \tag{6.4.2}$$

For a Pareto distribution, substituting in the endpoints $[sb, b]$ for $[a, b]$ into equation (6.3.13),

$$cv^2_{sb,b} = \frac{\left(b^{-\lambda}(s^{-\lambda}-1)\right)\frac{\lambda}{\lambda-2}\left(b^{2-\lambda}(s^{2-\lambda}-1)\right)}{\left(\frac{\lambda}{\lambda-1}b^{1-\lambda}(s^{1-\lambda}-1)\right)^2} - 1. \tag{6.4.3}$$

Similarly, substituting in the endpoints $[b, rb]$ for $[a, b]$ into equation (6.3.13),

$$cv^2_{b,rb} = \frac{\left(b^{-\lambda}(1-r^{-\lambda})\right)\frac{\lambda}{\lambda-2}\left(b^{2-\lambda}(1-r^{2-\lambda})\right)}{\left(\frac{\lambda}{\lambda-1}b^{1-\lambda}(1-r^{1-\lambda})\right)^2} - 1. \tag{6.4.4}$$

For these to be equal i.e.

$$cv^2_{sb,b} = cv^2_{b,rb} \tag{6.4.5}$$

then the following equality must hold

$$\frac{(s^{-\lambda}-1)(s^{2-\lambda}-1)}{(s^{1-\lambda}-1)^2} = \frac{(1-r^{-\lambda})(1-r^{2-\lambda})}{(1-r^{1-\lambda})^2}. \tag{6.4.6}$$

Note $\lambda > 2$ for stable variance. Letting $\lambda = \ell + 2$ where $\ell > 0$, (6.4.6) becomes:

$$\frac{(s^{-\ell-2}-1)(s^{-\ell}-1)}{(s^{-\ell-1}-1)^2} = \frac{(1-r^{-\ell-2})(1-r^{-\ell})}{(1-r^{-\ell-1})^2}. \tag{6.4.7}$$

Multiplying above and below the left hand side of (6.4.7) by s^{2l+2} and similarly, multiplying above and below the right hand side of (6.4.7) by r^{2l+2}, equation (6.4.7) can be written as:

$$\frac{(1-s^{\ell+2})(1-s^{\ell})}{(1-s^{\ell+1})^2} = \frac{(r^{\ell+2}-1)(r^{\ell}-1)}{(r^{\ell+1}-1)^2}. \tag{6.4.8}$$

Cross multiplying equation (6.4.8)

$$(1-s^{\ell+2})(1-s^{\ell})(r^{\ell+1}-1)^2 = (r^{\ell+2}-1)(r^{\ell}-1)(1-s^{\ell+1})^2, \qquad (6.4.9)$$

the left hand side of (6.4.9) becomes

$$r^{2\ell+2} - r^{2\ell+2}s^{\ell} - r^{2\ell+2}s^{\ell+2} + r^{2\ell+2}s^{2\ell+2}$$

$$-2r^{\ell+1} + 2r^{\ell+1}s^{\ell} + 2r^{\ell+1}s^{\ell+2} - 2r^{\ell+1}s^{2\ell+2} - s^{\ell} - s^{\ell+2} + s^{2\ell+2} + 1 \qquad (6.4.10)$$

and the right hand side of (6.4.9) becomes

$$s^{2\ell+2} - s^{2\ell+2}r^{\ell} - s^{2\ell+2}r^{\ell+2} + s^{2\ell+2}r^{2\ell+2}$$

$$-2s^{\ell+1} + 2s^{\ell+1}r^{\ell} + 2s^{\ell+1}r^{\ell+2} - 2s^{\ell+1}r^{2\ell+2} - r^{\ell} - r^{\ell+2} + r^{2\ell+2} + 1. \qquad (6.4.11)$$

The assumption (6.4.2) that $rs = 1$, reduces each of the expressions (6.4.10) and (6.4.11) to the same expression, namely

$$2 + 2(s+r) - (s^{\ell} + r^{\ell}) - 2(s^{\ell+1} + r^{\ell+1}) - (s^{\ell+2} + r^{\ell+2}) + (s^{2\ell+2} + r^{2\ell+2}). \qquad (6.4.12)$$

This gives equality of (6.4.5) *Q.E.D.*

Theorem 6.1. shows that for any two intervals in any finite range of the Pareto distribution, taking the boundaries in geometric progression gives equal coefficients of variation in each stratum. The extension to three or more intervals is obvious.

6.5 Summary

In this chapter we considered the Pareto distribution as a model of skewed data. We examined its conditional mean, variance and coefficient of variation restricted to a finite interval along the range. We showed that if any finite range is broken into a given number of strata by using geometric progression, then the stratum coefficients of variation are equal.

Recall that in Chapter 4, we needed to assume uniformity within strata to show that geometric breaks resulted in equal stratum coefficients of variation. The results obtained in this chapter suggest that such an assumption is not necessary if the data can be modelled with a Pareto distribution, a typical distribution for modelling skewed data.

Chapter 7

Conclusions and Future Research

7.1 Introduction

This chapter reviews how the objectives stated in Chapter 1 have been achieved (7.2), presents a summary of the findings and draws conclusions from the results (7.3). Some areas of future research are suggested (7.4).

7.2 Achievement of the Objectives

Many stratification methods have been developed. However, those that are simple to implement are inappropriate for skewed populations while those currently used in practice suffer from implementation problems.

As previously stated in Chapter 1, the specific objectives of this study are to:

1. develop a new stratification method for positively skewed populations to overcome the implementation problems of those currently used while maintaining the same efficiency;

2. investigate the efficiency of the new method compared to currently used methods;

3. improve the performance of the Lavallée-Hidiroglou stratification method;

4. stratify the Pareto distribution using the new method.

7.2.1 The Methodology Used to Achieve the Objectives

For the first objective, we used the idea of equalising stratum coefficients of variation to develop a new stratification method. This has been suggested by numerous researchers in the field as a desired goal when stratifying skewed populations, implying near optimal design would be achieved.

To investigate the efficiency of the new method, two benchmark methods were used. The cum $\sqrt{f(x)}$ method of Dalenius and Hodges (1959) was selected as it is the most commonly used one in practice. The second method, the Lavallée-Hidiroglou (1988) method, was chosen as it is designed specifically for skewed populations. The methods were applied to four real positively skewed populations stratified into 3, 4 and 5 strata. One was an accounting population of debtors from a commercial entity in the Irish Public Sector (Horgan, 1996) and the other three populations were used by Cochran (1961) in his comparative study on methods for determining stratum boundaries. Comparisons were made in terms of stratum breaks, stratum sizes, stratum sample sizes, equality of stratum coefficients of variation and precision of the stratified sample mean.

The third objective of this research looked at improving the Lavallée-Hidiroglou method. The sample sizes and convergence rates obtained with this iterative algorithm for different levels of precision, $cv(\overline{x}_{st}) = 0.01, 0.025$ and 0.05 with different starting points using the four populations divided into 4, 5 and 6 strata

were compared.

To fulfil the final objective, we stratified the Pareto distribution, a highly positively skewed model that typically arises in business situations, using the new method.

7.3 Summary of the Findings

7.3.1 A New Method

We found that by assuming a uniform distribution within each stratum, equal stratum coefficients of variation can be achieved by simply making the breaks in geometric progression.

7.3.2 Efficiency of New Method

A comparison of the geometric method and the cum $\sqrt{f(x)}$ method showed that in the majority of cases, the geometric method was more efficient in terms of minimising the variance of the stratified mean. While the geometric method is not always more efficient than the cum $\sqrt{f(x)}$ method, when it is, it is substantially better and when it is not, it is only marginally worse.

In the majority of cases, the geometric method was more efficient than the Lavallée-Hidiroglou method. With a few exceptions, the geometric method showed a trend of increased efficiency over the Lavallée-Hidiroglou method as the populations increased in skewness.

The geometric method, the cum $\sqrt{f(x)}$ method and the Lavallée-Hidiroglou method gave different stratum boundaries, stratum sizes and stratum sample sizes. It was also found that the geometric method achieved near-equal stratum

coefficients of variation while those of the cum $\sqrt{f(x)}$ and Lavallée-Hidiroglou methods were much more variable.

The implications for practitioners is that they can achieve approximately the same, and in some cases better, precision with the geometric method as they currently achieve with the cum $\sqrt{f(x)}$ method without the need for arbitrary initial class divisions or with the Lavallée-Hidiroglou algorithm, without the need for arbitrary starting points for the initial boundaries.

7.3.3 Alternative Initial Boundaries for the Lavallée-Hidiroglou Method

With just a few minor exceptions, we found that by starting the iterative process of the Lavallée-Hidiroglou algorithm using a set of boundaries in geometric progression there was faster convergence than using default boundaries. Geometric starting points achieve convergence within the maximum 30 iterations given by the algorithm in all cases and when they are slow to converge, it was found that the sample size returned at the 30^{th} iteration was already close to the optimum sample size obtainable with this algorithm. We also found that the average number of iterations required with geometric starting points is similar to the average number required with the cum $\sqrt{f(x)}$ starting points. Slow or non-convergence of the algorithm is more likely to occur when the number of strata is large and with the default starting points.

The average sample sizes obtained with geometric starting points were less than those obtained with other methods and significantly less than those obtained with default starting points with the larger number of strata. Comparisons with cum $\sqrt{f(x)}$ and the two stage process starting points indicate that average sample sizes were not significantly different.

Using geometric breaks as the initial boundaries for the iterative Lavallée-Hidiroglou algorithm tend to avoid non-convergence as they are closer to the optimum than breaks determined using the same number of units in each stratum (the default method), which are inappropriate for skewed populations. Users of the algorithm have experienced instability problems when the algorithm is used with Neyman allocation. However, we found that optimal (Neyman) allocation can be maintained by taking the initial boundaries in geometric progression and so avoiding the need to use the non-optimal option of power allocation.

7.3.4 Stratifying the Pareto Distribution

It was shown that if any finite range of the Pareto distribution is broken into a given number of strata with breaks made in geometric progression, then the stratum coefficients of variation are equal. The results obtained also show that the assumption of uniformity within strata is not necessary in order to obtain equal stratum coefficients of variation if the data can be modelled with a Pareto distribution.

7.4 Recommendations for Future Research

Since this study derived a new univariate stratification method used to create L take-some strata, assuming the auxiliary variable and the survey variable are the same, future research in this area might involve:

(i) adapting the algorithm for multivariate stratification problems for the case where the number of survey variables is greater than one;

(ii) adapting the algorithm to allow for a take-all stratum;

(iii) developing models to account for the discrepancy between the auxiliary and survey variables and to use these with the algorithm.

BIBLIOGRAPHY

Aoyama, H. (1954). A study of the stratified random sampling. *The Annals of Mathematical Statistics*, VI(1):1–36.

Bankier, M. D. (1988). Power allocations: determining sample sizes for subnational areas. *The American Statistician*, 42(3):174–177.

Chen, W. (1989). Stratification of a population: programming of Lavallée and Hidiroglou's algorithm. *American Statistical Association Proceedings of the Section on Survey Research Methods*, pages 620–624.

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

Cochran, W. G. (1961). Comparison of methods for determining stratum boundaries. *Bulletin of the International Statistical Institute*, 38(2):345–358.

Dalenius, T. (1950). The problem of optimum stratification. *Skandinavisk Aktuarietidskrift*, (3-4):203–213.

Dalenius, T. (1952). The problem of optimum stratification in a special type of design. *Skandinavisk Aktuarietidskrift*, (35):61–70.

Dalenius, T. and Gurney, M. (1951). The problem of optimum stratification II. *Skandinavisk Aktuarietidskrift*, (3-4):133–148.

Dalenius, T. and Hodges, J.L., J. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54:88–101.

Detlefsen, R. E. and Veum, C. S. (1991). Design issues for the retail trade sample surveys of the U.S. Bureau of the Census. *American Statistical Association Proceedings of the Survey Research Methods Section*, pages 214–219.

Durbin, J. (1959). Review of sampling in Sweden. *Journal of the Royal Statistical Society*, (A. 122):246–248.

Ekman, G. (1959). An approximation useful in univariate stratification. *The Annals of Mathematical Statistics*, 30:219–229.

Evans, M., Hastings, N., and Peacock, B. (2000). *Statistical Distributions.* Wiley, New York, third edition.

Falk, E. and Rotz, W. (2003). Stratified sampling for sales and use tax highly skewed data - determination of the certainty stratum cut-off amount. *Proceedings of the American Statistical Association Section on Statistical Computing.*

Glasser, G. J. (1962). On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, 30(1):28–32.

Hansen, M., Hurwitz, W., and Madow, W. (1953). *Sample Survey Methods and Theory.* Wiley, New York.

Hedlin, D. (1998). On the stratification of highly skewed populations. *R and D report. Statistics Sweden*, 3:1.

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

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

Hidiroglou, M. (1986). The construction of a self-representing stratum of large units in survey design. *The American Statistician*, 40(1):27–31.

Horgan, J. (1996). The moment bound with unrestricted random, cell and sieve sampling of monetary units. *J. of Acc. and Bus. Res.*, 26(3):215–223.

Khan, E. A., Khan, M. G. M., and Ahsan, M. J. (2002). Optimum stratification: a mathematical programming approach. *Calcutta Statistical Association Bulletin*, 52.

Kozak, M. (2004). Optimal stratification using random search method in agricultural surveys. *Statistics in Transition*, 6(5):797–806.

Lavallée, P. and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14:33–43.

Levy, P. and Lemeshow, S. (1999). *Sampling of populations.* Wiley.

Lohr, S. (1999). *Sampling: design and analysis.* Duxbury Press.

Mahalanobis, P. (1952). Some aspects of the design of sample surveys. *Sankhya*, 12:1–7.

McCarthy, P. and Clickner, R. (1985). Optimum sample design for skewed populations. *Proceedings of the American Statistical Association Section on Survey Research Methods.*

Newman, M. (1976). *Financial accounting estimates through statistical sampling by computer.* New York: Wiley.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97:558–606.

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

Raj, D. (1964). On forming strata of equal aggregrate size. *Journal of the American Statistical Association*, 59(306):481–486.

Rivest, L. (2002). A generalization of the Lavallée -Hidiroglou algorithm for stratification in business surveys. *Survey Methodology*, 28:191–198.

Roshwalb, A., Wright, R., and Godfrey, J. (1987). A new approach for stratified sampling in inventory cost estimation. *Auditing: A journal of practice and theory*, 7:54–70.

Sarndal, C., Swensson, B., and Wretman, J. (1992). *Model-Assisted Survey Sampling*. New York: Springer Verlag.

Serfling, R. (1968). Approximately optimum stratification. *Journal of the American Statistical Association*, 63:1298–1309.

Sethi, V. K. (1963). A note on optimum stratification of populations for estimating the population means. *Australian Journal of Statistics*, 5:20–33.

Sigman, R. and Monsour, N. (1995). Selecting samples from list frames of businesses. *Business Survey Methods*, pages 133–152.

Singh, R. (1971). Approximately optimal stratification of the auxiliary variable. *Journal of the American Statistical Association*, 66:829–833.

Slanta, J. and Krenzke, T. (1994). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditures Survey. *American Statistical Association Proceedings of the Section on Survey Research Methods*, pages 693–698.

Slanta, J. and Krenzke, T. (1996). Applying the Lavallée and Hidiroglou method to obtain stratification boundaries for the Census Bureau's Annual Capital Expenditures Survey. *Survey Methodology*, 22:65–75.

Thomsen, I. (1976). A comparison of approximately optimal stratification given proportional allocation with other methods of stratification and allocation. *Metrika*, 23:15–25.