SYSTEMATIC SAMPLING FROM FINITE POPULATIONS

by

Llewellyn Reeve

Naidoo

Submitted in fulfilment of the academic requirements for the degree of Masters of Science in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban

31 July 2013

As the candidate's supervisor I have approved this thesis for submission.

Signed: ________________ Name: ___________________________

Date: ________________

Abstract

The impossibility to reach an entire population, owing to time and budget constraints, results in the need for sampling to estimate population parameters. There are various methods of sampling and this thesis deals with a specific method of probability sampling, known as systematic sampling. Problems within the systematic sampling context include: (*i*) If the size of the population is not a multiple of the size of the sample, then conventional systematic sampling (also known as linear systematic sampling) will either result in variable sample sizes, or constant sample sizes that are greater than required; (*ii*) Linear systematic sampling is not the most preferred probability sampling design for populations that exhibit linear trend; (*iii*) An unbiased estimate of the sampling variance cannot be obtained from a single systematic sample. I will attempt to make an original contribution to the current body of knowledge, by introducing three new modified systematic sampling designs to address the problems mentioned in (*ii*) and (*iii*) above.

We will first discuss the measures to compare the various probability sampling designs, before providing a review of systematic sampling. Thereafter, the methodology of linear systematic sampling will be examined as well as two other methodologies to overcome the problem in (*i*). We will then obtain efficiency related formulas for the methodologies, after which we will demonstrate that the efficiency of systematic sampling depends on the correlation of the population units, which in turn depends on the arrangement and structure of the population. As a result, we will compare linear systematic sampling with other common probability sampling designs, under various population structures. Further designs of linear systematic sampling (including a new proposed design), which are considered to be optimal for populations that exhibit linear trend, will then be examined to resolve the problem mentioned in (*ii*). Thereafter, we will tackle the problem in (*iii*) by exploring various strategies, which include two new designs. Finally, we will obtain numerical comparisons for all the designs discussed in this thesis, on various population structures, before providing a comprehensive report on the thesis.

Keywords: linear systematic sampling; intra-class correlation coefficient; super-population model

Contents

List of acronyms

MSE	mean square error
CI	confidence interval
CL	confidence level
LSS	linear systematic sampling
SRS	simple random sampling
SRSWR	simple random sampling with replacement
SRSWOR	simple random sampling without replacement
STR	stratified random sampling
FIM	fractional interval method
CSS	circular systematic sampling
ICC	intra-class correlation coefficient
ANOVA	analysis of variance
YEC	Yates' end corrections
CESS	centered systematic sampling
BSS	balanced systematic sampling
MSS	modified systematic sampling
BMSS	balanced modified systematic sampling
MLSS	multiple-start linear systematic sampling
MBMSS	multiple-start balanced modified systematic sampling
PSS	partially systematic sampling
NSS	new systematic sampling
NPSS	new partially systematic sampling
BRS	balanced random sampling
BMRS	balanced modified random sampling
BMSSEC	balanced modified systematic sampling with end corrections

List of Figures

List of Tables

Preface

The experimental work described in this dissertation was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Durban, from August 2011 to July 2013, under the supervision of Professor Delia North.

These studies represent original work by the author and have not otherwise been submitted in any form for any degree or diploma to any tertiary institution. Where use has been made of the work of others it is duly acknowledged in the text.

Declaration - Plagiarism

I, ______________________________ declare that

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

   a. Their words have been rewritten but the general information attributed to them has been referenced.

   b. Where their exact words have been used, then their writing has been placed in italics and inside quotation marks, and referenced.

5. This thesis does not contain, text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed: ____________________

Acknowledgements

I would like to express my sincere appreciation to the NRF (National Research Foundation) and the College of Agriculture, Engineering and Science, for the financial support I received during the preparation of this thesis. I am also thankful for the moral support that I received from my parents. Furthermore, I would like to extend my gratitude to Prof. R. Arnab and Prof. T. Zewotir for their assistance and guidance. Finally, I'd like to thank my supervisor for her patience, support and particularly for advising me to pursue this research topic.

Chapter 1

INTRODUCTION

1.1 Overview of Sampling

Statistics involves the study of collecting, organizing, analysing, interpreting, explaining and presenting of data. Governments, clientele, medical agencies, institutions and organizations (both profit and non-profit) regularly use statistics to influence decision-making, such as deciding between different options, strategizing, implementing new policies, reviewing current policy situations, etc.

Once a problem has been identified and is presented to a statistician, he/she must then develop a workable goal before designing the research approach. Collection of appropriate data is then followed by the suitable analysis of the data and finally the findings are reported thereon. The above-mentioned phases of the research cycle are interdependent and are not unrelated. Failure to implement a good research design results in inefficient data collection methods, which in turn results in unfavourable data analysis and finally leads to an incorrect research report. It is clear that each phase of this research cycle is vital. We will now take a closer look at the data collection phase.

There are three fundamental types of statistical investigations, namely, surveys, experiments and observational studies. Each of these provides a different platform for data collection. Some examples of data collection methods are questionnaires, case studies, behaviour observation checklists, performance tests, etc. More often than not, one cannot collect data from an entire population (group of data that contains all the possible units that one is interested in investigating) owing to time constraints, money constraints and the common problem of it being virtually impossible to reach an entire population at a specific point or period in time. Consequently, we usually opt to draw a sample from a population, where a sample is defined as a selected group of units/subset from the population. Typically, we denote the population size by N and the sample size by n, with $N > n$. To make generalizations about a population, based on results from the sample, one needs to ensure that the sample is representative of the population. In other words, the characteristics of the sample should accurately reflect the characteristics of the population. Two conditions for drawing a representative sample are that (i) the sample must be of sufficient size so as to capture all aspects of the population and (ii) should be drawn in such a way so as to reduce bias, where bias is defined as the distortion of sample characteristics from the corresponding population characteristics.

A particular sampling design is generally employed to draw a sample that provides an estimate of a population parameter by means of a sample statistic, also known as an estimator. A point estimator, which is calculated from the sample data, is a single value that is used to estimate the population parameter.

An estimator is *unbiased* if its expected value is equal to the population parameter, otherwise it is referred to as being *biased*. For a biased estimator, the level of bias is the difference between the expected value of the point estimator and the true value of the population parameter. Accuracy is a term that is related to bias, such that unbiased estimators are on average equal to the population parameter and are thus considered to be perfectly accurate estimates of the population parameter.

Furthermore, a point estimator is a variable and therefore has a distribution, where the variance (or precision) of the point estimator, also known as the sampling variance, tells us by how much the point estimator varies from sample to sample. One may obtain more than one unbiased estimator of the population parameter (i.e. unbiased estimators of a population parameter are not unique) and the comparison of these corresponding sampling variances is then used to find the most precise estimator, i.e. the unbiased estimator with the lowest variance. The most precise unbiased estimator, which yields optimum results, will thus exhibit minimum variance as well as give the correct estimate of the corresponding population parameter on average and subsequently offer the most information about that population parameter, based on the sample. In addition, the relative efficiency between two point estimators is given as the ratio of their variances, i.e. the relative efficiency of point estimator a, when compared to point estimator b, is given as the variance of point estimator b divided by the variance of point estimator a. If this value is less than one, we can then deduce that point estimator b is preferred to point estimator a, while a value for the mentioned ratio which is greater than one, would imply that point estimator a is preferred to point estimator b. This measure is most effective when both point estimators are unbiased, or if both estimators exhibit the same degree of bias.

It should be noted that while unbiased (or accurate) estimators are desired, there may exist cases where biased estimators provide more information about the associated population parameter than unbiased ones, since certain biased estimators which exhibit a low level of bias, may offer a higher degree of precision than its counterparts. This resulting effect is thus a trade-off between accuracy and precision. A measure that captures this trade-off is known as the mean square error (MSE) of the point estimator and is usually used when comparing point estimators, where one or more of these point estimators are biased. The MSE of a point estimator is found by taking the sum of the variance of the associated estimator (measure of precision) and the squared bias of the corresponding estimator (measure of accuracy), such that preference is given for an estimator which exhibits a minimum MSE.

An interval estimate uses a range of values with two endpoints, such that the population parameter is likely to fall within the specified range. An example of an interval estimate is a confidence interval (CI), which is calculated from the point estimator and measures the precision of the point estimator in estimating the population parameter. The explanation of a 95% CI, where 95% is the confidence level (CL), is given as follows: If a sample was repeatedly drawn many times, according to a particular sampling design, where each time we calculated the point estimator and the corresponding CI for the associated population parameter, then we would expect that in 95% of the cases, the population parameter would lie within the CI. It should be noted that a narrower CI denotes a more precise point estimator, as the variance of the estimator has an inverse effect on the width of the CI. Three factors that affect the width of the CI are (*i*) the sampling variance, (*ii*) the size of the sample and (*iii*) the CL. A lower sampling variance, which results in a narrower CI, can be achieved by employing a sampling design which obtains samples that are relatively similar, i.e. smaller variation between samples. Alternatively, one can increase the size of the sample to achieve a narrower CI, but larger sample sizes incur greater costs, so that one needs to balance this trade-off. A higher CL will result in an increased probability that the population parameter will lie within the CI, such that a wider CI is needed, hence one also needs to balance this trade-off. For practical situations, the sampling variance is unknown and is thus estimated from the sample, so that we then construct the corresponding CIs according to this estimate.

It is thus of paramount importance to explore the effects of the various sampling designs on the level of bias, variance, MSE and CI of the associated point estimators, as well as the ability to produce an unbiased estimator of the corresponding sampling variance, when attempting to estimate the required population parameter. We will now take a closer look at the various probability sampling designs that can be employed for the estimation of population parameters.

There are various probability sampling designs to draw a representative sample, e.g. simple random sampling (SRS), stratified sampling, cluster sampling, systematic sampling etc. SRS involves randomly selecting a sample of size n from a population of size N, such that each population unit has an equal chance of selection for the sample, at each stage of the random selection. Simple random sampling with replacement (SRSWR) randomly selects a population unit for the sample, notes/records it, and then replaces it into the population to be eligible for the next random selection. There is thus a possibility that sampling units may be repeated when conducting SRSWR, since the population units that are selected for the sample are replaced, thus having a chance of being selected again in the random selections that follow. Simple random sampling without replacement (SRSWOR) is a slight adaption of SRSWR, which ensures a sample of distinct population units, because it randomly selects each population unit for the sample, as in the case of SRSWR, but now without replacing it before the next random selection. Stratified sampling involves dividing the entire population into subgroups (strata) according to some characteristic and then a particular form of random selection is independently carried out within each subgroup (stratum), such that the randomly selected population units for all the strata collectively form the stratified sample. It should be noted that if a simple random sample (with or without replacement) is drawn from each stratum, then this sampling design is termed as *stratified random sampling* (STR) and the randomly selected population units for all the strata collectively form a stratified random sample. Cluster sampling entails dividing the population into groups (clusters) and then randomly selecting entire clusters, so that the selected units within the clusters collectively form a cluster sample. Systematic sampling involves randomly selecting a population unit from the first k population units and selecting every kth population unit thereafter, such that the selected population units collectively form a systematic sample with sampling interval k, where the value of k is found by dividing the population size (N) by the sample size

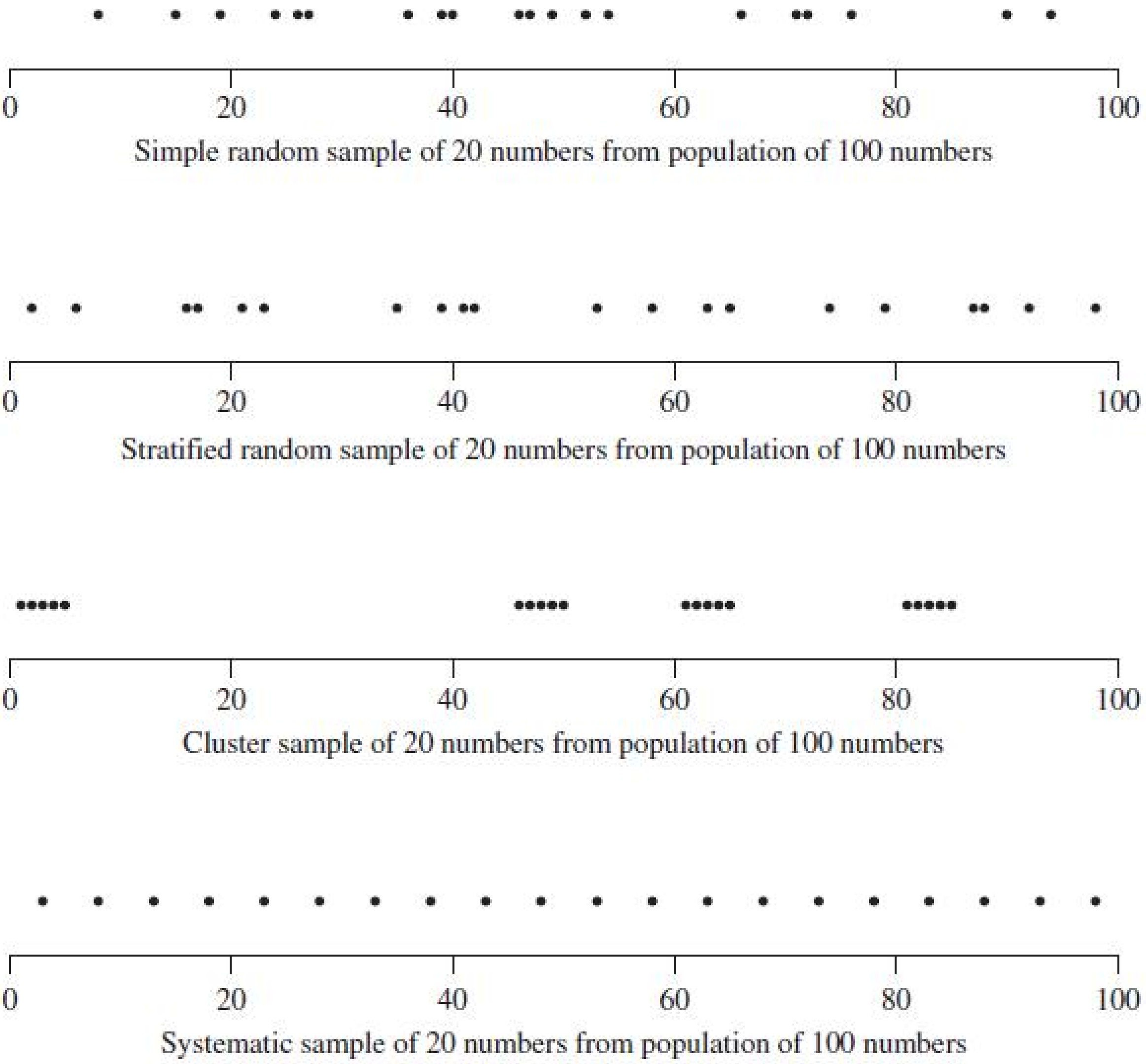
(n). It can be further noted that if we were to draw independent systematic samples from each stratum, then the corresponding sampling design is termed as *stratified systematic sampling* and the randomly selected population units for all the strata collectively form a stratified systematic sample (this will be discussed later in Section 4.2.5).

Figure 1.1 visually depicts the differences between SRSWOR, STR, cluster sampling and systematic sampling, where each design selects a sample of $n = 20$ integers from a population of $N = 100$ integers $\{1, ..., 100\}$ (Lohr 2010, p.27). We omit SRSWR, since we wish to make comparisons where the sample obtained for each design contains n distinct sampling units. However, we can expect the results to be similar to that of SRSWOR.

The stratified random sample is obtained by first dividing the population into 10 strata of equal size, such that the first strata contains the integers from 1 to 10, the second strata contains the integers from 2 to 20, and so forth. Two integers are then selected from each stratum, using SRSWOR, and these randomly selected integers collectively form the stratified random sample. The cluster sample is obtained by first dividing the population into 20 clusters of equal size, such that the first cluster contains the integers from 1 to 5, the second cluster contains the integers from 6 to 10, and so forth. Four clusters are then selected using SRSWOR and the integers within these clusters collectively form the cluster sample. A systematic sample is obtained by randomly selecting an integer from the first 5 integers (i.e. $k = N/n = 100/20 = 5$) and then selecting every fifth integer thereafter. In the example given in Figure 1.1, the integer 3 was randomly selected and thereafter every fifth value, i.e. 3, 8, 13, ..., 98. From Figure 1.1, it follows directly that a systematic sample ensures a more even spread of the sample over the entire population, as compared to the other probability sampling designs, for the population under consideration.

All the theory discussed thus far is a representation of the authors' understanding, which is interpreted from a variety of sources and can be broadly found in any standard introductory chapter on sample survey, e.g. Kalton (1983), Lehtonen & Pahkinen (2004), Lohr (2010), etc.

This thesis narrows the study to focus on the performance of systematic sampling for various population structures, while making comparisons to the other probability sampling designs mentioned above, as well as the designs which will be introduced later on.



Simple random sample of 20 numbers from population of 100 numbers

Stratified random sample of 20 numbers from population of 100 numbers

Cluster sample of 20 numbers from population of 100 numbers

Systematic sample of 20 numbers from population of 100 numbers

Figure 1.1: Graphic depiction of common probability sampling designs, when drawing a sample of 20 integers from a population of 100 integers $\{1, ..., 100\}$

1.2 Systematic Sampling

Comprehensive theoretical discussions on systematic sampling were initially provided by Madow & Madow (1944), Cochran (1946) and Yates (1948). Systematic sampling is commonly used in forestry, land use/cover area frame surveys, census, record sampling and for household and establishment surveys (Murthy & Rao 1988). Some applications of systematic sampling for forestry are provided by Hasel (1938), Finney (1948) and Zinger (1964), while applications for land use/cover area frames are given by Osborne (1942), Dunn & Harrison (1993) and D'Orazio (2003). Other examples of systematic sampling are provided in the fields of soil sampling (see Manson (1992) and Jacobsen (1998)) and environmental studies (see McArthur (1987) and NRC (2000)). Good summaries for the topic under consideration are given by Murthy (1967), Cochran (1977), Iachan (1982), Bellhouse (1988) and Murthy & Rao (1988).

Systematic sampling entails the following: Suppose that a sample of n units is to be selected from a population of size N using systematic sampling. The first step is always to present the population as a *"frame"* (set of all possible sampling units), with each population unit to be identified by a unique identifier, say unit number 1, unit number 2, ..., unit number N. A simple way to draw this sample would then be to determine a suitable sampling interval, say k, and to select units at equal intervals on the frame, where the value of k is found by dividing N by n. For example, if $k = 5$, then a unit is randomly selected from the frame, along with every fifth unit thereafter. More specifically, a 1-in-k systematic sample is obtained by randomly selecting a unit from the first k units in the frame and every kth unit thereafter. Another simple way to look at systematic sampling is that we are dividing the population into k possible samples and then selecting one of these samples at random. We next discuss some of the merits and shortcomings of using systematic sampling, as opposed to other probability sampling designs.

1.2.1 Advantages of systematic sampling

Some advantages of systematic sampling were noted by Daniel (2012) as follows:

(i) Systematic sampling is considered to be straightforward and inexpensive.

(ii) Systematic sampling is generally the preferred probability sampling design when there is a list of names or items available, in particular, for the case when records are numbered consecutively, or when population units can be ranked consecutively by attaching an integer to each of them.

(iii) Systematic sampling is often an economical design when the selection procedure is done manually, since only one randomization is required to select the first sampling unit and that particular sampling unit defines the sample, whereas in the case of SRSWR, SRSWOR and STR, we require n randomizations and this can be time-consuming for large sample and population sizes.

Moreover, systematic sampling provides a useful alternative to SRS, since (Scheaffer, Mendenhall & Ott 1995):

(i) Systematic sampling is much simpler to practise in the field and hence is less likely to have selection errors by field-workers, particularly if there is no good frame.

(i) Systematic sampling often supplies us with greater information per unit cost for populations with certain patterns in the arrangement of units.

Systematic sampling is usually more precise than SRS, since a systematic sample is more likely to contain units which are spread more evenly over the population as listed in the frame, when compared to the sampled units in a simple random sample, as shown in the earlier example.

1.2.2 Disadvantages of systematic sampling

The key shortcomings when conducting systematic sampling are that:

(i) If N is not a multiple of n, then systematic sampling will either result in variable sample sizes or constant sample sizes that are greater than n. Consequently, the former scenario results in biased estimates of the population parameters, while the latter scenario is undesired since sample sizes are usually fixed in advance. These scenarios will be extensively discussed in Chapters 2 and 3.

(ii) Systematic sampling is susceptible to periodic distortions, since the process of selection can interact with a population that exhibits periodic characteristics.

    Example 1.1: If there is a sampling frame containing adult residents in an area that consisted only of couples and a list is arranged as husband, wife, husband, wife etc. Now, if every tenth person is to be sampled, then the sample chosen will be only husbands or only wives.

    So, if a sampling design coincides with the periodicity of that characteristic, then the sampling design is considered to be non-random and the common property of systematic sampling being random is then compromised. A further discussion on this will be given in Chapter 4.

(iii) Systematic sampling is not the most preferred probability sampling design for populations that exhibit linear trend, as discussed in Chapter 4.

(iv) Certain pairs of population units will have a zero probability of being selected in the sample, which results in the estimation of the sampling variance being more complex (Daniel 2012). This disadvantage will be further explained in Chapter 6.

The fundamental aims of this thesis are to tackle these disadvantages, which will more often than not result in us providing modified systematic sampling designs. One may solve the disadvantage in (iv) by opting to maintain the conventional systematic sampling design, as shown in Chapter 6.

1.2.3 Scope of thesis

This thesis is divided into nine chapters. Chapter 2 discusses the methodology of the systematic sampling design, as well as discussing two common designs for dealing with the disadvantage in (i), i.e. *the fractional interval method* (FIM) and *circular systematic sampling* (CSS). In Chapter 3, we obtain an estimate for the population mean, the corresponding sampling variance and the intra-class correlation coefficient (ICC), when conducting either systematic sampling or CSS. The corresponding sampling variance obtained in Chapter 3 will then be used in Chapter 4, where we will compare the efficiency of systematic sampling to the other probability sampling designs, for various population structures. The population structures that will be discussed are populations in random order, populations that exhibit linear trend, periodic populations (solution to the disadvantage in (ii)), auto-correlated populations and stratified populations. Chapter 5 deals with various designs of systematic sampling, which are optimal for populations that exhibit linear trend, i.e. *Yates end corrections method* (YEC), *centered systematic sampling* (CESS), *balanced systematic sampling* (BSS), *modified systematic sampling* (MSS) and a new proposed design termed as *balanced modified systematic sampling* (BMSS). Error comparisons for each design to all the previously discussed designs are obtained, so as to solve the disadvantage in (iii). The problem of estimating the sampling variance will then be explored in Chapters 6 and 7, where various approaches (which include a new design for each chapter) are examined. Strategies to tackle the disadvantage in (iv) are to: (*i*) construct slightly biased variance estimators based on certain assumptions; (*ii*) supplement the systematic sample with independent sample(s); (*iii*) supplement the systematic sample with a dependent sample. We will examine strategies related to (*i*) and (*ii*) in Chapter 6, while strategies related to (*iii*) will be examined in Chapter 7. In Chapter 8, we provide numerical analysis for all the designs discussed, by considering various population structures. Finally, in Chapter 9 we pool all the theory and results from the previous chapters with a concluding comprehensive report of the different aspects and variations of systematic sampling.

Chapter 2

METHODOLOGY

The theory behind systematic sampling for the case where the sampling interval k=N/n is an integer (or if N is a multiple of n) is fairly straightforward, since all possible systematic samples are of size n. However, the theory behind systematic sampling for the case where k is not an integer (or if N is not a multiple of n) is a bit more complex, since systematic samples sizes may vary. The two common methods that alleviate the problem of variable sample sizes are the FIM and CSS. Both of these methods will be discussed in detail later on in this chapter.

Before discussing the methodologies, we first need to develop some notation which will be used throughout this thesis. For a population of size N, we respectively denote the population units and the corresponding variate values by U_i and y_i, for $i \in \{1, ..., N\}$. A variate value is defined as a single value that is usually a number (quantitative), which is a reading on our variable of interest, taken on the corresponding population unit. In this notation we use Y to denote our general variable of interest, which is associated with the variate values of each population unit, i.e. the variate value y_5, correspondingly represents the variate value of the fifth population unit (U_5), which denotes the particular value for this population unit that is associated with the variable of interest (Y).

Example 2.1: Suppose that we are interested in the average amount of household income for a given population. The population units will be the households and each household has a corresponding variate value (i.e. income) attached to it. These variate values are associated with our variable of interest, which is the average amount of household income for the population.

The theory presented in this chapter is a representation of the authors' understanding of the corresponding literature, interpreted from a variety of sample survey sources (see

Kish (1965), Murthy (1967) and Särndal et al. (2002)) and where minor contributions to the field are made, this will be clearly indicated in the text.

2.1 Case (A): If $k = N/n$ is an Integer

Suppose that we are to draw a sample of size n from a population of size N, using systematic sampling, where k is an integer. The corresponding methodology is given as follows:

(i) Randomly select an integer between 1 and k, say i, where $1 \leq i \leq k$.

(ii) The sample units chosen will be those elements with population unit numbers given by

$$i + (j-1)k, \qquad \text{for } j = 1, ..., n. \tag{2.1}$$

This process of selecting a systematic sample is known as *linear systematic sampling* (LSS). This particular method of selection is commonly known as the *restricted selection method*, since the selection of the first sampling unit is at random and restricted to the first k population units, i.e. the first sampling unit is chosen by a random selection from the first k population units.

Table 2.1 contains a list of possible values of random start i along with the corresponding sample outcomes, when selecting a sample of size n from a population of size N, using LSS when k is an integer. From Table 2.1, it is clear that the selection of the first population unit automatically determines the entire sample.

Table 2.1: Samples for possible values of i using LSS, where k is an integer

Possible values of i	Sample
$i = 1$	S_1=$\{U_1, U_{1+k}, U_{1+2k}, ..., U_{1+(n-1)k}\}$
$\vdots$	$\vdots$
$i = h$	S_h=$\{U_h, U_{h+k}, U_{h+2k}, ..., U_{h+(n-1)k}\}$
$\vdots$	$\vdots$
$i = k$	S_k=$\{U_k, U_{2k}, ..., U_{nk}\}$

Another way of looking at systematic sampling is to take into account that the entire population of size N is equally divided into k sampling units, each of size n. The process of selecting a systematic sample is thus equivalent to selecting one of these k sampling units at random. Systematic sampling is thus a form of cluster sampling, where each possible systematic sample can be viewed as a cluster.

For a given numbering of the population units, we are thus selecting one cluster of units from the k possible clusters, with a probability of $1/k$. Note that all the clusters collectively form the entire population, where we select one cluster at random. Each population unit (U_i) belongs to one and one cluster alone. Hence, the probability of selecting a cluster (i.e. P(Cluster is selected) $= 1/k$, for all $i \in \{1, ..., k\}$) is also the probability that a particular population unit is selected. The first-order inclusion probabilities are thus given by

$$\pi_i = \mathrm{P}(U_i \text{ is selected in the sample}) = 1/k, \qquad \text{for all } i \in \{1, ..., N\}.$$

We have therefore demonstrated that systematic sampling is a probability sampling design, as it is possible to determine the probability of selection of each population unit. Furthermore, if for some $i, j \in \{1, ..., N\}$

$$\pi_{ij} = \mathrm{P}(U_i \text{ and } U_j \text{ are both selected for the sample}),$$

then for all $i, j \in \{1, ..., N\}$, the second-order inclusion probabilities are given by

$$\pi_{ij} = \begin{cases} 0 & \text{if } U_i \text{ and } U_j \text{ are in different clusters,} \\ 1/k & \text{if } U_i \text{ and } U_j \text{ are in the same cluster.} \end{cases}$$

This demonstrates that it is impossible to have certain pairs of population units within the same systematic sample. The implications of this result will be shown in Chapter 6.

2.2 Case (B): If $k = N/n$ is not an Integer

If N is not a multiple of n, then we may represent this by $N = nk + c$, where $0 < c < n$. Now, if c/k is an integer, then by conducting LSS with a sampling interval of k, we obtain samples of fixed size given by $n + c/k$. As a result, this reduces to the LSS design given in (2.1), where the sample sizes are now $n + c/k$. This scenario is demonstrated by the following example.

Example 2.2: Suppose that we have a population of size $N = 27$ and we wish to draw a sample of size $n = 7$, using LSS. We thus note that $k = 3$ and $c = 6$, thus satisfying

$N = nk + c$ and $0 < c < n$. The possible systematic samples are then defined as:

(i) $S_1 = \{U_1, U_4, U_7, U_{10}, U_{13}, U_{16}, U_{19}, U_{22}, U_{25}\}$, for $i = 1$;

(ii) $S_2 = \{U_2, U_5, U_8, U_{11}, U_{14}, U_{17}, U_{20}, U_{23}, U_{26}\}$, for $i = 2$;

(iii) $S_3 = \{U_3, U_6, U_9, U_{12}, U_{15}, U_{18}, U_{21}, U_{24}, U_{27}\}$, for $i = 3$.

We thus obtain possible samples of constant size $n+c/k = 7+6/3 = 9$. The corresponding sample size is greater than the desired size and this is unadvisable since sample sizes are usually fixed in advance owing to budget constraints. Moreover, if c/k is not an integer, then by conducting LSS with a sampling interval of k, we obtain samples of variable size given by either $n+\text{INT}(c/k)$ or $n+\text{INT}(c/k)+1$, where $\text{INT}(a)$ is defined as the first integer before a. This is demonstrated by the following example.

Example 2.3: Suppose that we have a population of size $N = 19$ and we wish to draw a sample of size $n = 5$, using LSS. We thus note that $k = 3$ and $c = 4$, thus satisfying $N = nk + c$ and $0 < c < n$. The possible systematic samples are then defined as:

(i) $S_1 = \{U_1, U_4, U_7, U_{10}, U_{13}, U_{16}, U_{19}\}$, for $i = 1$;

(ii) $S_2 = \{U_2, U_5, U_8, U_{11}, U_{14}, U_{17}\}$, for $i = 2$;

(iii) $S_3 = \{U_3, U_6, U_9, U_{12}, U_{15}, U_{18}\}$, for $i = 3$.

We thus obtain variable sample sizes of either $n + \text{INT}(c/k) = 6$ or $n + \text{INT}(c/k) + 1 = 7$. As discussed in Chapter 1, a representative sample would require a fixed sample of sufficient size, such that samples with variable size may over-represent or under-represent the population. As a result, LSS with variable sample sizes results in biased estimates of population parameters (refer to Section 3.2) and is thus not an advisable approach. It should be noted that the probability of selecting a particular population unit is given as $1/k = 1/3$, irrespective of whether U_i belongs to a sample of size $n = 6$ or a sample of size $n = 7$, since only one sample is randomly selected from the $k = 3$ possible samples and each population unit falls into one of these $k = 3$ samples only.

In light of the above discussion, various strategies have been proposed over the years, for the case when k is not an integer, such that all samples are of fixed sample size n. Two common strategies are to (i) conduct systematic sampling such that the sampling interval has a fractional value (termed as the FIM) and (ii) to conduct systematic sampling in a circular fashion, which is termed as CSS. We will now discuss these two approaches.

2.2.1 Fractional interval method

This method was introduced by Kish (1965) and further investigated by Murthy (1967). It selects a linear systematic sample by giving k a fractional value, where $k = N/n$ and n is a fixed sample size. A random number, say i, is selected from the uniform distribution on the interval $(0, k]$. The sample units chosen will be those elements with population unit numbers given by α, where

$$\alpha - 1 < i + (j-1)k \leq \alpha, \qquad \text{for } j = 1, ..., n. \tag{2.2}$$

This is best demonstrated by the following example.

Example 2.4: Suppose that we have a population of size $N = 14$ and we wish to draw a sample of size $n = 3$, using the FIM. For this situation, the sampling interval is given by $k = 14/3$. Next, suppose that a random number $i = 1/5$ is selected from the uniform distribution, on the interval $(0, 14/3]$. By applying (2.2), we thus conclude that the population unit numbers 1, 5 and 10 are respectively chosen. The systematic sample, for the above situation, is subsequently given as $S_i = \{U_1, U_5, U_{10}\}$. This is demonstrated graphically in Figure 2.1.

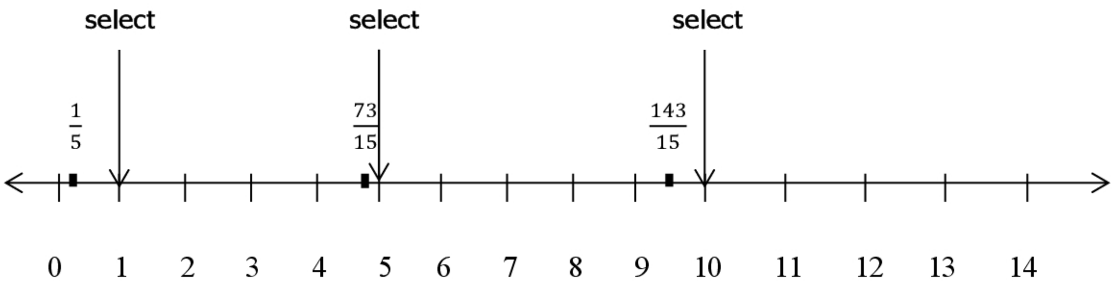


Figure 2.1: Selecting a sample of size 3 from a population of size 14, using the FIM with $i = 1/5$

The remainder of this subsection is solely due to the author. Table 2.2 contains a list of possible values of random start i along with the corresponding sample outcomes, when selecting a sample of size $n = 3$ from a population of size $N = 14$, using the FIM. From Table 2.2, it follows that each distinct sample is given by the interval

$$(t-1)/n < i \leq t/n, \qquad \text{for } t = 1, ..., N. \tag{2.3}$$

Table 2.2: Samples for possible values of i using the FIM, where $N = 14$ and $n = 3$

Possible values of i	Sample
$0 < i \leq 1/3$	S_1=$\{U_1, U_5, U_{10}\}$
$1/3 < i \leq 2/3$	S_2=$\{U_1, U_6, U_{10}\}$
$2/3 < i \leq 1$	S_3=$\{U_1, U_6, U_{11}\}$
$1 < i \leq 4/3$	S_4=$\{U_2, U_6, U_{11}\}$
$4/3 < i \leq 5/3$	S_5=$\{U_2, U_7, U_{11}\}$
$5/3 < i \leq 2$	S_6=$\{U_2, U_7, U_{12}\}$
$2 < i \leq 7/3$	S_7=$\{U_3, U_7, U_{12}\}$
$7/3 < i \leq 8/3$	S_8=$\{U_3, U_8, U_{12}\}$
$8/3 < i \leq 3$	S_9=$\{U_3, U_8, U_{13}\}$
$3 < i \leq 10/3$	S_{10}=$\{U_4, U_8, U_{13}\}$
$10/3 < i \leq 11/3$	S_{11}=$\{U_4, U_9, U_{13}\}$
$11/3 < i \leq 4$	S_{12}=$\{U_4, U_9, U_{14}\}$
$4 < i \leq 13/3$	S_{13}=$\{U_5, U_9, U_{14}\}$
$13/3 < i \leq 14/3$	S_{14}=$\{U_5, U_{10}, U_{14}\}$

The next example shows that in some cases we obtain samples that coincide, when applying the interval given by (2.3).

Example 2.5: Suppose that we have a population of size $N = 10$ and we wish to draw a sample of size $n = 4$, using the FIM. For this situation, the sampling interval is given by $k = 10/4$. Table 2.3 contains a list of possible values of random start i along with the corresponding sample outcomes, when selecting a sample of size 4 from a population of size 10, using the FIM, where each interval is as defined by (2.3).

It thus follows from Table 2.2 and Table 2.3 that there are N distinct samples in the former table, whereas we obtain samples that coincide in the latter table. We can thus show that each distinct sample for Table 2.3 (and more specifically for the case where $2N/n$ is an integer) is given by the interval

$$(t-1)/2 < i \leq t/2, \qquad \text{for } t = 1, ..., 2N/n. \tag{2.4}$$

Table 2.3: Samples for possible values of i using the FIM, where $N = 10$ and $n = 4$

Possible values of i	Sample
$0 < i \leq 1/4$	S_1={U_1, U_3, U_6, U_8}
$1/4 < i \leq 2/4$	S_2={U_1, U_3, U_6, U_8}
$2/4 < i \leq 3/4$	S_3={U_1, U_4, U_6, U_9}
$3/4 < i \leq 1$	S_4={U_1, U_4, U_6, U_9}
$1 < i \leq 5/4$	S_5={U_2, U_4, U_7, U_9}
$5/4 < i \leq 6/4$	S_6={U_2, U_4, U_7, U_9}
$6/4 < i \leq 7/4$	S_7={U_2, U_5, U_7, U_{10}}
$7/4 < i \leq 2$	S_8={U_2, U_5, U_7, U_{10}}
$2 < i \leq 9/4$	S_9={U_3, U_5, U_8, U_{10}}
$9/4 < i \leq 10/4$	S_{10}={U_3, U_5, U_8, U_{10}}

Using (2.3) and (2.4), we can thus define the probability of selecting a specific sample, using the FIM, to be

$$\mathrm{P}(S_i \text{ is selected}) = \begin{cases} n/2N & \text{if } 2N/n \text{ is an integer,} \\ 1/N & \text{otherwise.} \end{cases}$$

This result follows since each interval for the possible values of i is of equal length, thus each possible sample has an equal probability of being the selected sample. Furthermore, the possible samples obtained using the FIM are not mutually exclusive, since there are population units that occur more than once within the possible samples. In fact, there is an equal probability of inclusion in the sample for every population unit, regardless of whether $2N/n$ is an integer or not, i.e.

$$\pi_i = \mathrm{P}(U_i \text{ is in the sample}) = n/N, \qquad \text{for all } i \in \{1, ..., N\}.$$

2.2.2 Circular systematic sampling

CSS is another design that produces samples of fixed sample size n. We will first focus on the methodology of CSS, before explaining the choice of sampling interval. Thereafter, we will discuss the relationship between CSS and the other systematic sampling designs mentioned in this chapter.

Methodology

CSS was first introduced by Lahiri (1954) and involves the arrangement of sampling units in a circular fashion, such that the last population unit (U_N) is followed by the first unit (U_1). A random integer, say r, is selected from the interval $[1, N]$ with probability $1/N$, and k is taken to be the closest integer to N/n. The sample is then given by the population unit U_r and every kth population unit thereafter, until a sample of size n is obtained. By letting k be the nearest integer to N/n, we ensure a more even spread of the sample over the population (Murthy 1967). There are N possible samples in total for CSS, with each sample having a probability of $1/N$ of being selected. This method of selection is known as the *unrestricted selection method*, since the selection of the first sampling unit is at random and unrestricted to the entire frame, i.e. the first sampling unit is chosen by a random selection from the N population units. The above-mentioned methodology is best demonstrated by example.

Example 2.6: Suppose that we have a population of size $N = 14$ and we want to draw a sample of size $n = 3$, using CSS. For this situation, $k = 5$ since $N/n = 14/3 = 4.\overline{6}$ is nearest to integer 5. Also, suppose a random integer, say $r = 7$, is selected from the interval $[1, 14]$. For a sample of size $n = 3$, LSS will result in the selection of population units U_7, U_{12} and U_{17}, owing to the random start of $r = 7$ and sampling interval $k = 5$. This sample contains the population unit U_{17} which is non-existent, since there are only 14 population units in the frame. However, if we start counting the population units again, such that U_{15} corresponds to population unit U_1 (i.e. circular transversal), then U_{17} will correspond to population unit U_3 and the circular systematic sample will subsequently be found to be $S_7 = \{U_3, U_7, U_{12}\}$. This example is visually depicted in Figure 2.2.

Table 2.4 contains a list of all possible values of a random start r and the corresponding sample outcomes, when selecting a sample of size $n = 3$ from a population of size $N = 14$, using CSS. From Table 2.4, it thus follows that the possible samples are not mutually exclusive, since there are population units that occur more than once within the possible samples. In fact, just as in the case with the FIM, there is an equal probability of inclusion in the sample for every unit, i.e.

$$\pi_i = \mathrm{P}(U_i \text{ is in the sample}) = n/N, \qquad \text{for all } i \in \{1, ..., N\}.$$

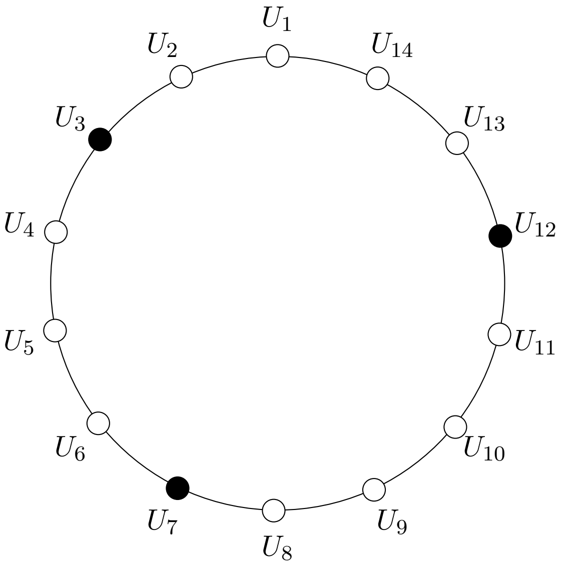


Figure 2.2: Selecting a sample of size 3 from a population of size 14, using CSS with $r = 7$

Table 2.4: Samples for possible values of i using CSS, where $N = 14$ and $n = 3$

Possible values of i	Sample
$r = 1$	S_1=$\{U_1, U_6, U_{11}\}$
$r = 2$	S_2=$\{U_2, U_7, U_{12}\}$
$r = 3$	S_3=$\{U_3, U_8, U_{13}\}$
$r = 4$	S_4=$\{U_4, U_9, U_{14}\}$
$r = 5$	S_5=$\{U_1, U_5, U_{10}\}$
$r = 6$	S_6=$\{U_2, U_6, U_{11}\}$
$r = 7$	S_7=$\{U_3, U_7, U_{12}\}$
$r = 8$	S_8=$\{U_4, U_8, U_{13}\}$
$r = 9$	S_9=$\{U_5, U_9, U_{14}\}$
$r = 10$	S_{10}=$\{U_1, U_6, U_{10}\}$
$r = 11$	S_{11}=$\{U_2, U_7, U_{11}\}$
$r = 12$	S_{12}=$\{U_3, U_8, U_{12}\}$
$r = 13$	S_{13}=$\{U_4, U_9, U_{13}\}$
$r = 14$	S_{14}=$\{U_5, U_{10}, U_{14}\}$

Choice of sampling interval

For the case where N is a multiple of k, Sudhakar (1978) showed that the sampling units will coincide when N/n is rounded up. This situation is demonstrated by the next example.

Example 2.7: Suppose that we draw a sample of size $n = 9$ from a population of size $N = 24$, using CSS. For this situation, $k = 3$ since $N/n = 24/9 = 2.\overline{6}$ is nearest to integer 3. Also, suppose that a random integer, say $r = 3$, is selected from the interval $[1, 14]$. This subsequently results in $S_3 = \{U_3, U_6, U_9, U_{12}, U_{15}, U_{18}, U_{21}, U_{24}\}$, i.e. the first sampling unit will coincide with the ninth sampling unit.

Sudhakar (1978) argued that one can achieve n distinct sampling units, if and only if N and k are co-prime (i.e. $N \neq (n-1)k$), by considering this result to be applicable when n is not fixed beforehand. A summary of this result suggests that k be chosen beforehand, with N and k being co-prime to obtain a sample of n distinct population units. The shortcoming of this approach is that sample sizes are usually fixed beforehand owing to budget constraints. Bellhouse (1984) suggests an alternative approach to overcome this shortcoming. He proposed that a new sampling interval, k', be defined as

$$k' = \begin{cases} \mathrm{INT}(N/n) & \text{if } N = (n-1)k, \\ \mathrm{INT}(N/n + 1/2) & \text{if } N \neq (n-1)k. \end{cases} \tag{2.5}$$

Hence, by using (2.5), where N/n is rounded down when N is a multiple of k, we obtain n distinct sampling units. On the contrary, Sengupta & Chattopadhyay (1987) argued that coincidence of sampling units is still possible when N/n is rounded up in (2.5), as shown in the next example.

Example 2.8: Suppose that we wish to draw a sample of size $n = 22$ from a population of size $N = 60$, using CSS. For this situation, $k = 3$ since $N/n = 60/22 = 2.\overline{72}$ is nearest to integer 3. By noting that $N \neq (n-1)k$, we then apply (2.5) to get $k' = \mathrm{INT}(60/22+1/2) = 3$. Now, for any random start, the first and the second sampling units coincide with the $(n-1)$th and nth unit, respectively.

Sengupta & Chattopadhyay (1987) provides a theorem which states that for any circular systematic sample, one can achieve n distinct sampling units, if and only if $\mathrm{lcm}(N, k) \geq nk$, or equivalently if and only if $\gcd(N, k) \leq N/n$, where $\mathrm{lcm}(a, b)$ and $\gcd(a, b)$ respectively denote the lowest common multiple and the greatest common divisor, for constants a and b. This theorem is not in disagreement with Sudhakar's (1978) results and can be used as an extension to Bellhouse's (1984) approach. Consequently, a new sampling interval, k^*, suggested by the author, can thus be defined by using the above-mentioned theorem along with (2.5), such that

$$k^* = \begin{cases} \text{INT}(N/n) & \text{if } \text{lcm}(N,k) \geq nk \text{ (or } \text{gcd}(N,k) \leq N/n), \\ \text{INT}(N/n + 1/2) & \text{otherwise.} \end{cases} \tag{2.6}$$

It should be noted that most authors suggest the use of a sampling interval given by INT(N/n). While this sampling interval will always result in samples of n distinct sampling units, it does not ensure an even spread of the sample over the population, as would a sampling interval of INT$(N/n + 1/2)$, i.e. INT$(N/n + 1/2) \geq$ INT(N/n), resulting in the selection of units that are further apart in the frame when applying the sampling interval INT$(N/n + 1/2)$, as opposed to the sampling interval INT(N/n).

Relationship to other systematic sampling designs

We have thus shown that for both the FIM and CSS, any population unit will have a chance of n/N of being in the sample and every possible sample will be of size n. With the assumptions that $2N/n$ is not an integer and lcm$(N,k) \geq nk$ (or gcd$(N,k) \leq N/n$), we thus conclude that the FIM and CSS are equivalent designs, since both designs have the same probability of selection for each possible sample and they both define the same set of possible samples (refer to Tables 2.2 and 2.4).

If N is a multiple of n, then CSS reduces to LSS. Moreover, if $N \gg n$, then the difference between LSS and CSS is negligible (Murthy 1967). It should be noted that for LSS, the probability of selecting a sample $(1/k)$ is also the probability that a particular population unit is selected, whereas for CSS these probabilities are not equal, i.e. $\text{P}(S_i \text{ is selected}) = 1/N \neq n/N = \text{P}(U_i \text{ is in the sample}),$ for all $i \in \{1, ..., N\}$. Moreover, the possible samples that are defined for CSS (and the FIM) are not mutually exclusive, unlike LSS which defines k mutually exclusive samples.

For the remainder of this thesis, we will use CSS as the preferred design over the FIM, since CSS is equivalent to the FIM in most cases (i.e. $2N/n$ is not an integer in most cases and lcm$(N,k) < nk$ seldom occurs) and it is easier to apply CSS, as opposed to the FIM.

In the next chapter, we focus our attention on the derivation of formulae which are associated with the variable of interest, for both LSS and CSS.

Chapter 3

ESTIMATION OF THE POPULATION MEAN

This chapter is divided into two parts. The first part consists of deriving formulae which are associated with the estimation of the population mean $(\overline{Y})$, for the case when k is an integer, i.e. we will conduct LSS. In the second part of this chapter, we will derive related formulae for the case when k is not an integer, conducting either LSS or CSS.

3.1 Case (A): If $k = N/n$ is an Integer

We will first derive formulae for an estimate of $\overline{Y}$ and the corresponding sampling variance, followed by obtaining an alternative formula for the sampling variance, which will be expressed in terms of the ICC. Thereafter, we will discuss the link between the ICC and the analysis of variance (ANOVA). Throughout this section we will assume that k is an integer, i.e. $N = nk$.

3.1.1 Population mean estimation

Theorem 3.1: An unbiased estimator of $\overline{Y}$ and the corresponding sampling variance, when conducting LSS, are respectively given by

$$\hat{\overline{Y}} = \overline{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij} = \frac{G_i}{n}, \tag{3.1}$$

$$\text{Var}(\overline{y}_i) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2, \tag{3.2}$$

where $\overline{y}_i = \sum_{j=1}^{n} y_{ij}/n$ denotes the ith systematic sample mean (i.e. the mean of the sample with random start i) and y_{ij} denotes the variate value of the population unit corresponding to the jth unit of the ith systematic sample, with $i = 1, ..., k$ and $j = 1, ..., n$. Note that $G_i = \sum_{j=1}^{n} y_{ij}$, for $i \in \{1, ..., k\}$, is the ith systematic sample total, where the probability that the systematic sample total is G_i, is the same as the probability of selecting that particular systematic sample, i.e. $\mathrm{P}(G_i) = 1/k$.

Proof: By using the fact that $\mathrm{P}(G_i) = 1/k$, we obtain

$$\begin{aligned}\mathrm{E}(G_i) &= \sum_{i=1}^{k} G_i \times \mathrm{P}(i\text{th systematic sample is selected}) \\ &= \frac{1}{k}\sum_{i=1}^{k} G_i = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij} = \frac{Y_{.}}{k}. \qquad (3.3)\end{aligned}$$

Thus, by using (3.3), it follows that

$$\mathrm{E}(\overline{y}_i) = \mathrm{E}\left(\frac{1}{n}\sum_{j=1}^{n} y_{ij}\right) = \mathrm{E}\left(\frac{G_i}{n}\right) = \frac{1}{n}\mathrm{E}(G_i) = \frac{Y_{.}}{nk} = \frac{Y_{.}}{N} = \overline{Y}.$$

We thus conclude that $\overline{y}_i = \sum_{j=1}^{n} y_{ij}/n$ is an unbiased estimator of $\overline{Y}$. Now, by applying (3.1), the sampling variance of $\overline{y}_i$ is then expressed as

$$\mathrm{Var}(\overline{y}_i) = \mathrm{Var}\left(\frac{G_i}{n}\right) = \frac{1}{n^2}\mathrm{Var}(G_i). \qquad (3.4)$$

Accordingly, by using (3.1), (3.3) and (3.4), we obtain

$$\mathrm{Var}(\overline{y}_i) = \frac{1}{n^2}\sum_{i=1}^{k}\left[G_i - \mathrm{E}(G_i)\right]^2 \times \mathrm{P}(G_i) = \frac{1}{k}\sum_{i=1}^{k}\left(\frac{G_i}{n} - \frac{Y_{.}}{nk}\right)^2 = \frac{1}{k}\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2.$$

Note that $\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2$ is a measure of the variation for the sample means, which is calculated as the sum of the deviations of each sample mean from the population mean. If the sample means are relatively similar, then we obtain a small sampling variance, which in turn improves the reliability of this estimation procedure.

Cochran (1977) provides us with three additional formulae for $\mathrm{Var}(\overline{y}_i)$, where two of these formulae are used to compare LSS with SRS (one of these formulae is derived in the next section). The third formula, which will be given in Section 4.1.3, expresses $\mathrm{Var}(\overline{y}_i)$ in terms of the corresponding sampling variance when conducting STR, and will be used to compare LSS with STR.

3.1.2 Intra-class correlation coefficient

The ICC between pairs of population units that lie within the same systematic sample is defined as

$$\rho = \mathrm{Cov}(y_{ij}, y_{il})/\sigma^2, \qquad \text{for } j, l = 1, ..., n, (j \neq l) \text{ and } i = 1, ..., k, \tag{3.5}$$

where

$$\mathrm{Cov}(y_{ij}, y_{il}) = \frac{1}{nk(n-1)} \sum_{i=1}^{k} \sum_{j=1}^{n} \sum_{\substack{l=1 \\ j \neq l}}^{n} (y_{ij} - \overline{Y})(y_{il} - \overline{Y}), \tag{3.6}$$

such that y_{ij} and y_{il} are random variables that represent two distinct units from the ith systematic sample and

$$\sigma^2 \triangleq \frac{1}{nk} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \overline{Y})^2 \tag{3.7}$$

is defined as the population variance. The divisor in (3.6) is obtained by noting that there are k systematic samples with n sampling units within each sample and $(n-1)$ comparisons for each particular sampling unit.

Using the above notation, we next express $\mathrm{Var}(\overline{y}_i)$ in terms of the ICC, before obtaining alternative formulae for the ICC which are related to the ANOVA. An alternative formula to (3.2), as shown below, will be used to obtain efficiency comparisons in Chapter 4.

Theorem 3.2: The sampling variance given in (3.2) can be written as

$$\mathrm{Var}(\overline{y}_i) = \frac{S_Y^2}{n} \left(\frac{N-1}{N} \right) \Big[1 + (n-1)\rho \Big], \tag{3.8}$$

where

$$S_Y^2 \triangleq \frac{1}{N-1} \sum_{j=1}^{N} \left(y_i - \overline{Y} \right)^2 = \frac{1}{N-1} \sum_{i=1}^{k} \sum_{j=1}^{n} \left(y_{ij} - \overline{Y} \right)^2 \tag{3.9}$$

is the adjusted population variance.

Proof: By applying (3.2), we obtain

$$
\begin{aligned}
\text{Var}(\overline{y}_i) &= \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 \\
&= \frac{1}{kn^2}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \overline{Y})\right]^2 \\
&= \frac{1}{Nn}\sum_{i=1}^{k}\left[\sum_{j=1}^{n}(y_{ij} - \overline{y})^2 + 2\sum_{j=1}^{n}\sum_{l>j}^{n}(y_{ij} - \overline{Y})(y_{il} - \overline{Y})\right] \\
&= \frac{1}{Nn}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2 + \frac{2}{Nn}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l>j}^{n}(y_{ij} - \overline{Y})(y_{il} - \overline{Y}) \\
&= \frac{1}{Nn}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2 + \frac{1}{Nn}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l\neq j}^{n}(y_{ij} - \overline{Y})(y_{il} - \overline{Y}). \qquad (3.10)
\end{aligned}
$$

Now, by using (3.5), (3.6) and (3.7), it follows that

$$
\begin{aligned}
\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l\neq j}^{n}(y_{ij} - \overline{Y})(y_{il} - \overline{Y}) &= nk(n-1)\text{Cov}(y_{ij}, y_{il}) \\
&= nk(n-1)\rho\sigma^2 \\
&= (n-1)\rho\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2. \qquad (3.11)
\end{aligned}
$$

Finally, substituting (3.11) into (3.10) and then applying (3.9), results in

$$
\begin{aligned}
\text{Var}(\overline{y}_i) &= \frac{1}{Nn}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2 + \frac{1}{Nn}(n-1)\rho\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2 \\
&= \frac{S_Y^2(N-1)}{Nn} + \frac{(n-1)\rho S_Y^2(N-1)}{Nn} = \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)\left[1 + (n-1)\rho\right].
\end{aligned}
$$

We thus conclude that positive correlation between population units that lie within the same systematic sample, increases $\text{Var}(\overline{y}_i)$ by a multiplier of $(n-1)$ (Cochran 1977). We further note from (3.8), that one cannot be certain that the sampling variance will decrease if the sample size is increased, or equivalently if the sampling interval is decreased, since $N = nk$ is fixed. This is in direct contrast to SRS and STR, where larger samples result in lower sampling variances. Empirical results given by Madow (1946) show us the erratic behaviour of the sampling variance as the sample size increases, when conducting LSS.

We will now discuss the ICC as a measure of homogeneity, by explaining the ANOVA. This approach is used in many standard sample survey textbooks, such as Särndal et al. (2002) and Lohr (2010). The ANOVA, which is used to explain the variance decomposition of N population units that are divided into k clusters of size n, is given as follows:

(i) The total sum of squares (SST), measures the total variation between all population units, and is given by

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2. \tag{3.12}$$

(ii) The sum of squares between clusters (SSB), measures the variation among the cluster means, and is given by

$$SSB = n\sum_{i=1}^{k}(\overline{y}_{ij} - \overline{Y})^2. \tag{3.13}$$

(iii) The sum of squares within clusters (SSW), measures the variation among population units that are within the same cluster, and is given by

$$SSW = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{y}_i)^2. \tag{3.14}$$

(iv) Standard sample survey texts (Särndal et al. 2002) bear evidence that equations (3.12), (3.13) and (3.14) are related by

$$SST = SSW + SSB. \tag{3.15}$$

(v) The adjusted population variance, given by (3.9), is commonly referred to as the total mean square (MST), and is expressed as

$$MST = S_Y^2 = \frac{SST}{N-1} = \frac{1}{N-1}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij} - \overline{Y})^2. \tag{3.16}$$

(vi) The variance between clusters, known as the mean square between clusters (MSB), is given by

$$MSB = \frac{SSB}{k-1} = \frac{n}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2. \tag{3.17}$$

(vii) The variance within clusters, known as the mean square within clusters (MSW), is given by

$$MSW = \frac{SSW}{N-k} = \frac{1}{N-k}\sum_{i=1}^{k}\sum_{j=1}^{n}(\overline{y}_i - \overline{Y})^2. \tag{3.18}$$

Since LSS is a form of cluster sampling (refer to Section 2.1), we can use (i) to (vii) for LSS. The formulae in (3.12) to (3.18) will remain the same and the only difference will be that we are now considering k systematic samples, instead of k clusters. The sum of squares between systematic samples (SSB) is thus a measure of the variation among systematic sample means, while the sum of squares within systematic samples (SSW) measures the variation among population units that are within the same systematic sample. Furthermore, the variance between systematic samples (S^2_{bsys}) is given by (3.17), while the variance among population units that are within the same systematic sample (S^2_{wsys}) is given by (3.18).

For any given discrete population, SST is clearly fixed, so that an increase in SSW results in a corresponding decrease in SSB (refer to (3.15)). Analytical work done by Stuart (1976) concludes that cluster sampling should be done in such a way that the clusters are made as internally heterogeneous (large variation between the population units that lie within the clusters) as possible and/or as externally homogeneous (small variation between the clusters) as possible to obtain maximum precision in estimation. We thus obtain maximum precision of estimates when conducting LSS, if the population units that lie within the same systematic sample vary as much as possible (i.e. maximize SSW), while attaining minimum difference between the k systematic sample means (i.e. minimize SSB). This will consequently result in a lower sampling variance. One can achieve the goal of lowering the sampling variance by rearranging the population units, so that different systematic samples are formed. This is in direct contrast to SRS, in which the arrangement of population units has no effect on the sampling variance. Different orderings/arrangements of population units and the corresponding effect on estimation, when conducting LSS, will be discussed in Chapters 4 and 5. We next obtain alternative formulae for the ICC, using the ANOVA given above.

By using equations (3.5), (3.6), (3.7), (3.10) and (3.12), we obtain

$$\begin{aligned}\rho &= \mathrm{Cov}(y_{ij}, y_{il})/\sigma^2 \\ &= \left[(n-1)\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\overline{Y})^2\right]^{-1}\sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{l\neq j}^{n}(y_{ij}-\overline{Y})(y_{il}-\overline{Y}) \\ &= \frac{1}{(n-1)SST}\big[Nn\mathrm{Var}(\overline{y}_i) - SST\big]. \qquad (3.19)\end{aligned}$$

Now, by applying (3.2) and (3.13), it follows that

$$Nn\text{Var}(\overline{y}_i) = \frac{Nn}{k}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 = n^2\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2 = nSSB. \tag{3.20}$$

Hence, substituting (3.20) into (3.19), results in

$$\rho = \frac{1}{(n-1)SST}[nSSB - SST]. \tag{3.21}$$

Another expression for ρ can be found by using (3.15) and (3.21), such that

$$\begin{aligned}\rho &= \frac{1}{(n-1)SST}[n(SST - SSW) - SST] \\ &= \frac{1}{(n-1)SST}[(n-1)SST - nSSW] = 1 - \frac{nSSW}{(n-1)SST}\ .\end{aligned} \tag{3.22}$$

A final measure of ρ is obtained by applying (3.16), (3.18) and (3.22), i.e.

$$\begin{aligned}\rho &= 1 - \frac{nSSW}{(n-1)SST} \\ &= 1 - \frac{n(N-k)MSW}{(n-1)(N-1)MST} \\ &= 1 - \frac{nk(n-1)MSW}{(n-1)(N-1)MST} = 1 - \left(\frac{N}{N-1}\right)\frac{MSW}{MST}.\end{aligned} \tag{3.23}$$

Now, if N is large, then $N/(N-1) \approxeq 1$ and by substituting this result into (3.23), we obtain

$$\rho \approxeq 1 - \frac{MSW}{MST}. \tag{3.24}$$

By referring to (3.24), we thus note that $\rho > 0$ when $MST > MSW$, i.e. when the adjusted population variance is greater than the variance among population units that lie within the same systematic sample ($S_Y^2 > S_{wsys}^2$). The population units that lie within the same systematic sample will thus tend to contain similar values and are labelled as homogeneous. In contrast, we can expect $\rho < 0$ when $MST < MSW$, i.e. when the adjusted population variance is less than the variance among population units that lie within the same systematic sample ($S_Y^2 < S_{wsys}^2$). In this case, the population units that lie within the same systematic sample will tend to contain dissimilar values and are labelled as heterogeneous. Complete homogeneity within the systematic samples indicates no variation among the population units within each systematic sample, so that $SSW = 0$ for this scenario, resulting in $\rho = 1$ or $\rho_{max} = 1$ (refer to (3.22)). By substituting this result into (3.8), we obtain $\text{Var}(\overline{y}_i) = S_Y^2(N-1)/N$, i.e. the sampling variance is at a maximum value. Conversely, complete heterogeneity within the systematic samples indicates that there is maximum variation among the population units within each systematic sample. From (3.15), we thus conclude that $SSW = SSW_{max} = SST$ for this scenario and since SST is fixed, it follows that $SSB = SSB_{min} = 0$. By substituting $SSW = SST$ into (3.22), we obtain $\rho = -1/(n-1)$ or $\rho_{min} = -1/(n-1)$. Stuart (1976) concludes that complete heterogeneity within clusters provides optimal results in terms of precision. It is thus desirable to obtain an ordering/arrangement of the units which results in ρ being as close to the value of $-1/(n-1)$ as possible. One can easily verify that $\text{Var}(\overline{y}_i) = 0$, by substituting $\rho = -1/(n-1)$ into (3.8).

3.2 Case (B): If $k = N/n$ is not an Integer

We next derive a formula for an estimate of $\overline{Y}$ and then proceed to find the associated level of bias for this estimator as well as the corresponding sampling variance, when conducting LSS for the case when k is not an integer. We follow this by obtaining formulae for an estimate of $\overline{Y}$ and the corresponding sampling variance, when conducting CSS. Finally, we will discuss the ICC for both LSS and CSS.

3.2.1 Population mean estimation

We now assume that k is not an integer, such that $N = nk + c$, where $0 < c < n$ and c/k is not an integer. In Chapter 2, we showed that if we apply LSS for this situation, then we either obtain samples of size $n + \text{INT}(c/k)$ or $n + \text{INT}(c/k) + 1$. Consequently, samples are either over-representative or under-representative of the population and thus one cannot obtain unbiased estimates of the population parameters. As a result, we obtain a biased estimate of $\overline{Y}$, as shown in the next theorem.

Theorem 3.3: Suppose that we draw a sample of size n from a population of size N, using LSS, where k is not an integer. If c denotes the remainder, where $N = nk + c$, $0 < c < n$ and c/k is not an integer, then a biased estimator of $\overline{Y}$, the associated level of bias for the estimator and the corresponding sampling variance, are respectively given by

$$\hat{\overline{Y}} = \overline{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij} = \frac{G_i}{n_i}, \tag{3.25}$$

$$\mathrm{B}\left(\hat{\overline{Y}}\right) = \sum_{i=1}^{k} \overline{y}_i \left(\frac{1}{k} - \frac{n_i}{N}\right), \tag{3.26}$$

$$\mathrm{Var}(\overline{y}_i) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \overline{\overline{y}})^2, \tag{3.27}$$

where n_i, $\overline{y}_i = \sum_{j=1}^{n_i} y_{ij}/n_i$, for $i \in \{1, ..., k\}$, and $\overline{\overline{y}} = \sum_{i=1}^{k} \overline{y}_i/k$ respectively denotes the size of the ith systematic sample, the ith systematic sample mean and the average of the k systematic samples. Note that n_i is either $n + \mathrm{INT}(c/k)$ or $n + \mathrm{INT}(c/k) + 1$ and G_i/n_i, for $i \in \{1, ..., k\}$, is the ith systematic sample mean, where the probability that the systematic sample mean is G_i/n_i, is the same as the probability of selecting that particular systematic sample, i.e. $\mathrm{P}(G_i/n_i) = 1/k$.

Proof: Now, since there are k possible systematic samples, we note that

$$\mathrm{E}\left(\hat{\overline{Y}}\right) = \mathrm{E}(\overline{y}_i) = \mathrm{E}\left(\frac{G_i}{n_i}\right) = \mathrm{E}\left[\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i}\right] = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{y_{ij}}{n_i k} = \sum_{i=1}^{k}\frac{\overline{y}_i}{k} = \overline{\overline{y}}. \tag{3.28}$$

In addition, $\overline{Y}$ is defined as

$$\overline{Y} \triangleq \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}. \tag{3.29}$$

Hence, by comparing (3.28) to (3.29) and noting that $N \neq n_i k$, for all $i \in \{1, ..., k\}$, we thus conclude that G_i/n_i is a biased estimator of $\overline{Y}$. To obtain the level of bias we use (3.25), (3.28) and (3.29), such that

$$\begin{aligned}\mathrm{B}\left(\hat{\overline{Y}}\right) &\triangleq \mathrm{E}\left(\hat{\overline{Y}}\right) - \overline{Y}\\ &= \frac{1}{k}\sum_{i=1}^{k}\overline{y}_i - \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij} = \frac{1}{k}\sum_{i=1}^{k}\overline{y}_i - \frac{1}{N}\sum_{i=1}^{k}\overline{y}_i n_i = \sum_{i=1}^{k}\overline{y}_i\left(\frac{1}{k} - \frac{n_i}{N}\right).\end{aligned}$$

Finally, we can find the sampling variance by applying (3.25) and (3.28), i.e.

$$\mathrm{Var}(\overline{y}_i) = \mathrm{Var}\left(\frac{G_i}{n_i}\right) = \sum_{i=1}^{k}\left[\frac{G_i}{n_i} - \mathrm{E}\left(\frac{G_i}{n_i}\right)\right]^2 \times \mathrm{P}\left(\frac{G_i}{n_i}\right) = \frac{1}{k}\sum_{i=1}^{k}(\overline{y}_i - \overline{\overline{y}})^2.$$

We next derive formulae for an estimate of $\overline{Y}$ and the corresponding sampling variance, when conducting CSS, such that the sample size is fixed.

Theorem 3.4: Suppose that we draw a sample of size n from a population of size N, using CSS, where the random starting point r is an integer drawn from the interval $[1, N]$ and k is not an integer. Let k^* and c denote respectively the sampling interval that is given by (2.6) and a non-zero integer, such that $N = nk^* + c$. An unbiased estimator of $\overline{Y}$ and the corresponding sampling variance are thus respectively given as

$$\hat{\overline{Y}} = \overline{y}_r = \frac{G'_r}{n}, \tag{3.30}$$

$$\mathrm{Var}(\overline{y}_r) = \frac{1}{N}\sum_{r=1}^{N}(\overline{y}_r - \overline{Y})^2, \tag{3.31}$$

where $\overline{y}_r = G'_r/n$ is the rth circular systematic sample mean and

$$G'_r = \begin{cases} \sum_{j=1}^{n} y_{(j-1)k^*+r} & \text{for } r = 1, ..., k^* + c; \\ \sum_{j=1}^{n-1} y_{(j-1)k^*+r} + y_{(n-1)k^*+r-N} & \text{for } r = k^* + c + 1, ..., 2k^* + c; \\ \sum_{j=1}^{n-2} y_{(j-1)k^*+r} + y_{(n-1)k^*+r-N} + y_{(n-2)k^*+r-N} & \text{for } r = 2k^* + c + 1, ..., 3k^* + c; \\ \vdots & \vdots \\ y_r + y_{(n-1)k^*+r-N} + y_{(n-2)k^*+r-N} + ... + y_{k^*+r-N} & \text{for } r = (n-1)k^* + c + 1, ..., N; \end{cases}$$

is the rth circular systematic sample total. It should be noted that the probability that the circular systematic sample total is G'_r, is the same as the probability of selecting that particular circular systematic sample, i.e. $\mathrm{P}(G'_r) = 1/N$.

Proof: Now, since there are N possible samples that could be selected, we note that

$$\mathrm{E}(\hat{\overline{Y}}) = \mathrm{E}(\overline{y}_r) = \mathrm{E}\left(\frac{G'_r}{n}\right) = \frac{1}{N}\sum_{r=1}^{N}\frac{G'_r}{n} = \frac{1}{Nn}\sum_{r=1}^{N}G'_r = \frac{nY.}{Nn} = \overline{Y}, \tag{3.32}$$

which follows since each population unit is repeated n times, when referring to all the possible samples (see Section 2.2.3). We thus conclude that $\overline{y}_r = G'_r/n$ is an unbiased estimator of $\overline{Y}$. Now, by using (3.30), we obtain

$$\mathrm{Var}(\overline{y}_r) = \mathrm{Var}\left(\frac{G'_r}{n}\right) = \frac{1}{n^2}\mathrm{Var}(G'_r) = \frac{1}{n^2}\sum_{r=1}^{N}\left[G'_r - \mathrm{E}(G'_r)\right]^2 \times \mathrm{P}(G'_r), \tag{3.33}$$

which follows since there are N possible circular systematic samples. Also, by using (3.32), we note that

$$\mathrm{E}(G'_r) = n\overline{Y}. \tag{3.34}$$

Thus, by applying (3.30), (3.33) and (3.34), it follows that

$$\mathrm{Var}(\overline{y}_r) = \frac{1}{Nn^2}\sum_{r=1}^{N}\left[G'_r - n\overline{Y}\right]^2 = \frac{1}{N}\sum_{r=1}^{N}\left[\frac{G'_r}{n} - \overline{Y}\right]^2 = \frac{1}{N}\sum_{r=1}^{N}\left[\overline{y}_r - \overline{Y}\right]^2.$$

Theorems 3.3 and 3.4 thus indicate another comparative advantage of CSS, since we are able to obtain an unbiased estimate of $\overline{Y}$ for CSS, whereas an unbiased estimate of $\overline{Y}$ is unobtainable when conducting LSS, for the case where $N = nk + c$, $0 < c < n$ and c/k is not an integer. Moreover, if $n_i = N/k$ for all i (i.e. all systematic samples are of equal size), then the level of bias in (3.26) reduces to zero, which then makes it possible to obtain an unbiased estimate for $\overline{Y}$, when conducting LSS. If $N = nk + c$, $0 < c < n$ and c/k is an integer, then $n_i = n + c/k$, for all i. Consequently, it is then possible to obtain an unbiased estimate of $\overline{Y}$, where the corresponding formulae are obtained by replacing n in the previous section, with $n + c/k$.

3.2.2 Intra-class correlation coefficient

The ICC requires that all clusters/systematic samples be of equal size. It is thus not applicable to use this measure of homogeneity for LSS when $N = nk + c$, $0 < c < n$ and c/k is not an integer. An alternative measure, δ, which is related to ρ, is obtained by assuming n and N to be large, such that $n \approxeq n + 1$ and $(N-1)/(N-k) \approxeq 1$ (Särndal et al. 2002). Thus, by using these values together with (3.16), (3.18) and (3.22), it follows that

$$\rho = 1 - \frac{nSSW}{(n-1)SST} \approxeq 1 - \frac{SSW}{SST} \approxeq 1 - \frac{(N-1)SSW}{(N-k)SST} = 1 - \frac{MSW}{MST} = \delta. \tag{3.35}$$

By substituting $SSW = 0$ into (3.35), we obtain $\delta = 1$ or $\delta_{max} = 1$, which is a result of complete homogeneity within the systematic samples. Conversely, by substituting $SSW = SST$ into (3.35), it follows that $\delta = -(k-1)/(N-k)$ or $\delta_{min} = -(k-1)/(N-k)$, which is a result of complete heterogeneity within the systematic samples. In Section 3.1.2, we concluded that complete heterogeneity within systematic samples provided optimal results for LSS, in terms of precision. It is thus desirable to obtain an ordering/arrangement of the units which results in δ being as close to the value of $-(k-1)/(N-k)$ as possible, if we were to conduct LSS, for the case where $N = nk + c$, $0 < c < n$ and c/k is not an integer.

There is an analogue to Theorem 3.2, which can be used to express $\text{Var}(\overline{y}_r)$ (the sampling variance when conducting CSS) in terms of the ICC. This result was noted by Murthy & Rao (1988) and given as

$$\text{Var}(\overline{y}_r) = \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)[1 - (n-1)\rho'],$$

where $\rho' = 2\sum_{r=1}^{N}\sum_{j=0}^{n-1}\sum_{j'>j}^{n-1}(y_{r+jk^*} - \overline{Y})(y_{r+j'k^*} - \overline{Y})/n(n-1)(N-1)S_Y^2$ denotes the ICC between pairs of population units that lie within the same circular systematic sample, with the deviations calculated from $\overline{Y}$.

Throughout this chapter we derived estimates for $\overline{Y}$ and the corresponding sampling variances. We can, however, obtain an estimate of the population total $Y.$ and the corresponding sampling variance, by noting that $Y. \triangleq N\overline{Y}$, such that

$$\hat{Y}. = N\hat{\overline{Y}}. \tag{3.36}$$

The corresponding sampling variance is thus given as

$$\text{Var}(\hat{Y}.) = \text{Var}(N\hat{\overline{Y}}) = N^2\text{Var}(\hat{\overline{Y}}). \tag{3.37}$$

We can now obtain the formulae for an estimate of $Y.$ and the corresponding sampling variance, by substituting the relative formulae into (3.36) and (3.7) respectively. However, for the purpose of this thesis, we will only consider estimating $\overline{Y}$ and refer to the variance of this estimate as the sampling variance.

In the next chapter, we will use (3.8) to compare the efficiency of LSS with the other probability sampling designs. We will also discuss various population structures, where we will obtain related efficiency comparisons within each population structure.

Chapter 4

EFFICIENCY AND POPULATION STRUCTURES

In the first part of this chapter we compare the efficiency of LSS to that of SRSWR, SRSWOR and STR. In Chapter 1, we noted that the variance of an unbiased estimate of the population parameter is a comparative measure, with the classic notion that the better estimate is the one which exhibits the lower variance. Consequently, we use the relative efficiency between two different estimators, produced by two different sampling designs, as a measure of efficiency (i.e. the ratio of the sampling variances), since we obtain unbiased estimates of $\overline{Y}$ when conducting either SRSWR, SRSWOR, STR (Cochran 1977) or LSS (see Theorem 3.1). It is easily deduced from (3.8) that the sampling variance, when conducting LSS, depends on n (or k, since $N = nk$ is fixed) and the ICC, as S_Y^2 and N are fixed. In Section 3.1.2, we mentioned that a larger value of n (or a smaller value of k) does not necessarily lead to a lower sampling variance, i.e. the sampling variance, when conducting LSS, does not vary consistently with n. Hence, the only factor that proportionately affects the sampling variance is the ICC, which depends on (i) the ordering of population units from which the systematic sample is to be drawn, (ii) the amount of correlation between successive population units and (iii) is also related to n (Murthy & Rao 1988). Consequently, for the second part of this chapter, we will compare the efficiency of LSS to the other probability sampling designs, by considering various population structures.

Throughout this chapter we assume that k is an integer, however we usually can apply the results obtained in this chapter to LSS when k is not an integer (Murthy & Rao 1988).

Theoretical efficiency comparisons of CSS to the other probability sampling designs, on various population structures, will be left for further studies. However, we will empirically test the efficiency of CSS against other probability sampling designs in Chapter 8. In this chapter and all subsequent chapters, we will use the notation $\overline{y}_{LSS}, \overline{y}_{SRSWR}, \overline{y}_{SRSWOR}$ and $\overline{y}_{STR}$ to denote the sample means, when conducting LSS, SRSWR, SRSWOR or STR, respectively.

4.1 Efficiency of Linear Systematic Sampling

We next compare the efficiency of LSS to each of the other probability sampling designs mentioned in Chapter 1, i.e. comparing their respective sampling variances. In this section, the formula for $\text{Var}(\overline{y}_{LSS})$ is given by (3.8), while other corresponding sampling variance formulae are given by Cochran (1977).

4.1.1 Comparison to SRSWR

Suppose that we draw a sample of size n from a population of size N, using SRSWR. An unbiased estimator of $\overline{Y}$ is given by $\overline{y}_{SRSWR}$, with the corresponding sampling variance expressed as

$$\text{Var}(\overline{y}_{SRSWR}) = \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right). \tag{4.1}$$

The relative efficiency of SRSWR, with respect to LSS, is thus given by

$$\frac{\text{Var}(\overline{y}_{LSS})}{\text{Var}(\overline{y}_{SRSWR})} = \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)\left[1+(n-1)\rho\right]\left[\frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)\right]^{-1} = 1+(n-1)\rho.$$

Clearly, if $\rho < 0$, then LSS is more efficient than SRSWR. By using (3.24), we thus conclude that if $MST < MSW$, then $\rho < 0$ and consequently LSS is then more efficient than SRSWR. This then translates to imply that the more heterogeneous the population units that lie within the same systematic sample, the greater the efficiency gains when choosing LSS over SRSWR. Conversely, we conclude that if $MST > MSW$, then $\rho > 0$ and LSS is then less efficient than SRSWR. This consequently means that the more homogeneous the population units that lie within the same systematic sample, the greater the efficiency loss when choosing LSS over SRSWR. If we substitute $\rho = 0$ (i.e. no correlation amongst population units that lie within the same systematic sample) into (3.8) and then compare this result with (4.1), we see that $\text{Var}(\overline{y}_{LSS}) = \text{Var}(\overline{y}_{SRSWR})$ and thus conclude that no efficiency is gained when choosing one design over the other for this scenario.

4.1.2 Comparison to SRSWOR

Suppose that we draw a sample of size n from a population of size N, using SRSWOR. An unbiased estimator of $\overline{Y}$ is given by $\overline{y}_{SRSWOR}$, with the corresponding sampling variance expressed as

$$\text{Var}(\overline{y}_{SRSWOR}) = \frac{S_Y^2}{n}\left(\frac{N-n}{N}\right). \tag{4.2}$$

The relative efficiency of SRSWOR, with respect to LSS, is thus given by

$$\begin{aligned}\frac{\text{Var}(\overline{y}_{LSS})}{\text{Var}(\overline{y}_{SRSWR})} &= \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)\left[1+(n-1)\rho\right]\left[\frac{S_Y^2}{n}\left(\frac{N-n}{N}\right)\right]^{-1} \\ &= \left(\frac{N-1}{N-n}\right)\left[1+(n-1)\rho\right] \approxeq 1+(n-1)\rho,\end{aligned}$$

which follows if we assume N to be relatively larger than n. The discussion of the effect of $\rho < 0$ and $\rho > 0$, given in the previous section, thus applies provided that we assume N to be relatively larger than n. By substituting $\rho = 0$ into (3.8) and comparing this result with (4.2), we see that $\text{Var}(\overline{y}_{LSS}) > \text{Var}(\overline{y}_{SRSWOR})$ and consequently there would be a gain in efficiency when choosing SRSWOR over LSS for this scenario. From the above result on the relative efficiency, we note that this gain in efficiency tends to zero as N becomes relatively larger than n. For convenience and simplicity, one may thus choose LSS as the preferred sampling design. By substituting $\rho = -1/(N-1)$ into (3.8) and then comparing this result with (4.2), we see that $\text{Var}(\overline{y}_{LSS}) = \text{Var}(\overline{y}_{SRSWOR})$, with no efficiency being gained when choosing one design over the other for this situation. More specifically, we thus conclude that LSS is more efficient than SRSWOR, if and only if $\rho < -1/(N-1)$. By applying (4.1) and (4.2), we obtain the relative efficiency of SRSWOR, with respect to SRSWR, given by

$$\frac{\text{Var}(\overline{y}_{SRSWR})}{\text{Var}(\overline{y}_{SRSWOR})} = \frac{S_Y^2}{n}\left(\frac{N-1}{N}\right)\left[\frac{S_Y^2}{n}\left(\frac{N-n}{N}\right)\right]^{-1} = \frac{N-1}{N-n}.$$

This result shows us that SRSWOR is always more efficient that SRSWR (except when $n = 1$) and both designs will be approximately equally efficient when N is large, i.e. larger population sizes result in a higher probability of obtaining distinct sampling units for SRSWR. Both designs are equivalent when $n = 1$ and this result is trivial, since there is only one unit sampled and replacement of the sampling unit thereafter does not affect the sample. In light of this result, we will focus on SRSWOR as being our preferred SRS design. We thus obtain efficiency comparisons between LSS and SRSWOR, before applying the results to make efficiency comparisons between LSS and SRSWR, e.g. if we find that LSS is more efficient than SRSWOR, then we conclude that LSS is more efficient than SRSWR (i.e. transitive law).

4.1.3 Comparison to STR

Suppose that we draw a sample of size n from a population of size N, using STR with the assumption of equally sized strata of size k, such that one unit is selected per stratum. An unbiased estimator of $\overline{Y}$ is given by $\overline{y}_{STR}$, with the corresponding sampling variance expressed as

$$\mathrm{Var}(\overline{y}_{STR}) = \frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right), \tag{4.3}$$

where $S^2_{wst} = \sum\limits_{s=1}^{n}\sum\limits_{i=1}^{k}(y_{is}-\overline{y}_{\cdot s})^2/n(k-1)$ denotes the variance amongst population units that are within the same stratum; y_{is} denotes the variate value of the population unit corresponding to the ith element of the sth stratum; and $\overline{y}_{\cdot s} = \sum\limits_{i=1}^{k} y_{is}/k$ is the sth stratum mean, for $s \in \{1, ..., n\}$. The degrees of freedom for S^2_{wst} is given as $n(k-1)$, since we have one parameter (i.e. stratum mean) within each stratum of k units, resulting in each of the n strata having $(k-1)$ degrees of freedom.

We next attempt to find an expression for comparing the efficiency of LSS in terms of STR, when estimating $\overline{Y}$. With the assumption that $N = nk$, we define LSS as being the process of dividing the N population units into n strata of k population units each and then selecting one population unit from each stratum, where the sampling unit selected is located in the same position for every stratum. The LSS design, which was depicted by Table 2.1, is thus transposed and can be compared to STR, with one unit drawn from each stratum. The arrangement is such that, the first k units belong to the first stratum, the second k units belong to the second stratum, and so forth. Cochran (1977) states that there is a similar theorem to that of Theorem 3.2, which gives an expression for $\mathrm{Var}(\overline{y}_{LSS})$ and can be used to compare LSS to STR. The corresponding result is given by

$$\mathrm{Var}\left(\overline{y}_{LSS}\right) = \frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right)\left[1+(n-1)\rho_{wst}\right], \tag{4.4}$$

where $\rho_{wst} = 2\sum\limits_{i=1}^{k}\sum\limits_{s=1}^{n}\sum\limits_{l>s}^{n}\left(y_{is}-\overline{y}_{\cdot s}\right)\left(y_{il}-\overline{y}_{\cdot l}\right)/\left[n\left(n-1\right)\left(k-1\right)S^2_{wst}\right]$ is the ICC between pairs of population units that lie within the same systematic sample, where the deviations are calculated from their respective stratum means. We next obtain the relative efficiency of STR, with respect to LSS, by using (4.3) and (4.4), i.e.

$$\frac{\mathrm{Var}\left(\overline{y}_{LSS}\right)}{\mathrm{Var}\left(\overline{y}_{STR}\right)} = \frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right)\left[1+(n-1)\,\rho_{wst}\right]\left\{\frac{S^2_{wst}}{n}\left(\frac{N-n}{N}\right)\right\}^{-1} = 1+(n-1)\,\rho_{wst}.$$

Clearly, if $\rho_{wst} < 0$, then LSS is more efficient than STR and if $\rho_{wst} > 0$, then LSS is less efficient than STR. By substituting $\rho_{wst} = 0$ into (4.4) and then comparing this result with (4.3), we see that $\mathrm{Var}(\overline{y}_{LSS}) = \mathrm{Var}(\overline{y}_{STR})$ and thus conclude that no efficiency is gained when choosing one design over the other for this scenario.

It consequently follows from the above discussion, that the only comparable factor which affects the efficiency of LSS is the ICC. The ICC depends on the ordering of the population units from which a systematic sample is to be drawn, the amount of correlation between successive elements in the population and is also related to n. We will thus consider different population structures to examine the efficiency of LSS.

4.2 Population Structures

We will now discuss the various population structures (i.e. random ordered, linear trend, periodic, auto-correlated and stratified populations) and obtain efficiency comparisons for each population structure.

4.2.1 Population in random order

The following theorem shows the relationship between SRSWOR and LSS, for a population that is in random order.

Theorem 4.1: For a randomly ordered population of size N, the probability of selecting any specific sample of size n using either, LSS or SRSWOR, is $1/(C^N_n) = n!(N-n)!/N!$, which results in both designs being equivalent.

Proof: This proof, which was first presented by Madow & Madow (1944), is given by Murthy (1967). The total number of possible samples of size n that can be drawn from a population of size N, using SRSWOR, is given as $\mathrm{N}(S) = C^N_n$, i.e. the order of the sampling units does not have any effect and thus do not matter. The probability of selecting any specific sample S_i (i.e. $\mathrm{N}(S_i) = 1$) of n sampling units is thus given by $\mathrm{P}(S_i) = \mathrm{N}(S_i)/\mathrm{N}(S) = 1/(C^N_n)$. By assuming $N = nk$, such that LSS is equivalent to CSS (see Section 2.2.3), we then use the unrestricted selection method, since the population units are random and the order doesn't matter. There are $N!$ permutations of orderings of the population units, which results in $\mathrm{N}(S) = N(N!)$ being the total number of possible systematic samples, which may be repeated. Furthermore, if we consider LSS of equally spaced intervals, of size k, we get $(N-n)!$ possible orderings of choosing a specific sample. Hence, the number of systematic samples which contains a specific set of n sampling units is given as $\mathrm{N}(S_i) = N(N-n)!n!$, since there are $n!$ orderings of the n sampling units and the first sampling unit may be any one of the N population units. The probability of selecting any specific sample of n sampling units, using LSS, is thus given as

$$\mathrm{P}\left(S_i\right) = \frac{\mathrm{N}\left(S_i\right)}{\mathrm{N}\left(S\right)} = \frac{N\left(N-n\right)!n!}{N\left(N!\right)} = \left[\frac{N!}{(N-n)!n!}\right]^{-1} = \frac{1}{C_n^N}.$$

The equivalence of $\mathrm{Var}(\overline{y}_{SRSWOR})$ and $\mathrm{Var}(\overline{y}_{LSS})$, for any single finite population in random order, is not exactly true, since $\mathrm{Var}(\overline{y}_{LSS})$ is not proportional to n (refer to Section 3.2.2). However, Madow & Madow (1944) proved that LSS is expected to be equally efficient to SRSWOR, by considering all $N!$ permutations of the finite randomly ordered population of size N, i.e. $\mathrm{E}[\mathrm{Var}(\overline{y}_{LSS})] = \mathrm{Var}(\overline{y}_{SRSWOR})$.

Example 4.1: Suppose that we are required to conduct a survey for a company on their employees' work related traveling expenses. In addition, suppose that we sample from a list of their employees and that this list is arranged in ascending order according to their surnames. Although the list is arranged in ascending order, the population is considered to be random, since there is no relation to work related traveling expenses (variable of interest) and surnames (ordered variable). Therefore, the order of the population units do not matter and LSS can be viewed as SRSWOR.

To compare LSS to STR, we compare (4.2) to (4.3), such that the relative efficiency of STR, with respect to LSS, is given by

$$\frac{\mathrm{Var}\left(\overline{y}_{LSS}\right)}{\mathrm{Var}\left(\overline{y}_{STR}\right)} = \frac{\mathrm{Var}\left(\overline{y}_{SRSWOR}\right)}{\mathrm{Var}\left(\overline{y}_{STR}\right)} = \frac{S_Y^2}{n}\left(\frac{N-n}{N}\right)\left[\frac{S_{wst}^2}{n}\left(\frac{N-n}{N}\right)\right]^{-1} = \frac{S_Y^2}{S_{wst}^2}.$$

We thus conclude that if the variance amongst population units which lie within the same stratum is greater than the adjusted population variance, then LSS is more efficient than STR. In contrast, if the variance amongst population units which lie within the same stratum is less than the adjusted population variance, then LSS is less efficient than STR.

4.2.2 Populations that exhibit linear trend

Trend is defined as a general path that is followed by the variate values of a population, as the population unit numbers ($i \in \{1, ..., N\}$) increase in a sequence and/or as time progresses, if the variate indices indicate points in time. If the variate values tend to increase as the population unit numbers increases, then the trend is said to be positive. Conversely, if the variate values tend to decrease as the population unit numbers increases, then the trend is said to be negative. Trend can either be linear or non-linear (parabolic trend, quadratic trend, exponential trend etc.). For the purpose of this thesis, we will only consider linear trends. We will first discuss a hypothetical linear trend model, in which the variate values of the population units exhibit arithmetic progression, i.e. a perfect linear trend model. We then obtain formulae for $\text{Var}(\overline{y}_{LSS}), \text{Var}(\overline{y}_{SRSWOR})$ and $\text{Var}(\overline{y}_{STR})$, for this model, before obtaining related efficiency comparisons. Finally, we will discuss the ICC, when comparing LSS to the other probability sampling designs for the population under consideration.

Perfect linear trend model

Mathematical evidence, originally given by Madow & Madow (1944) and later discussed by Murthy (1967) and Cochran (1977), is used to show the efficiency of LSS under the presence of linear trend in a population. A hypothetical population that exhibits linear trend may be represented by the model

$$y_i = a + bi, \qquad \text{for } i = 1, ..., N. \tag{4.5}$$

This model depends on constants a and b, where the variate values are increasing by a constant factor b, resulting in the population exhibiting perfect linear trend. By applying (4.5), we obtain

$$\begin{aligned}\overline{Y} &= \frac{1}{N}\sum_{i=1}^{N} y_i \\ &= \frac{1}{N}\left[(a+b) + ... + (a+Nb)\right] = \frac{1}{N}\left[Na + b\sum_{i=1}^{N} i\right] = a + \frac{b\,(N+1)}{2}.\end{aligned} \tag{4.6}$$

Furthermore, by using (4.5) and (4.6), we obtain

$$\begin{aligned}
S_y^2 &= \frac{1}{(N-1)} \sum_{i=1}^{N} \left[a + bi - \left\{ a + \frac{b\,(N+1)}{2} \right\} \right]^2 \\
&= \frac{b^2}{(N-1)} \sum_{i=1}^{N} \left[i - \frac{(N+1)}{2} \right]^2 \\
&= \frac{b^2}{(N-1)} \left[\sum_{i=1}^{N} i^2 - N \left(\frac{N+1}{2} \right)^2 \right] \\
&= \frac{b^2}{(N-1)} \left[\frac{N\,(N+1)\,(2N+1)}{6} - \frac{N(N+1)^2}{4} \right] \\
&= \frac{b^2 N\,(N+1)}{(N-1)} \left(\frac{4N+2-3N-3}{12} \right) = \frac{b^2 N\,(N+1)}{12}. \qquad (4.7)
\end{aligned}$$

Efficiency comparisons

By applying (2.1), for the model in (4.5), we note that

$$\begin{aligned}
\overline{y}_i &= \frac{1}{n} \left[(a+bi) + (a + b\{i+k\}) + ... + (a + b\{i + (n-1)k\}) \right] \\
&= a + bi + \frac{bk}{n} \sum_{i=1}^{(n-1)} i = a + bi + \frac{bk\,(n-1)}{2}. \qquad (4.8)
\end{aligned}$$

Thus, by using (3.2), (4.6) and (4.8), we obtain

$$\begin{aligned}
\operatorname{Var}(\overline{y}_{LSS}) &= \frac{1}{k} \sum_{i=1}^{k} \left[a + bi + \frac{bk\,(n-1)}{2} - \left\{ a + \frac{b\,(N+1)}{2} \right\} \right]^2 \\
&= \frac{b^2}{k} \sum_{i=1}^{k} \left[i + \frac{nk-k}{2} - \frac{(N+1)}{2} \right]^2 \\
&= \frac{b^2}{k} \sum_{i=1}^{k} \left[i - \frac{(k+1)}{2} \right]^2 \\
&= \frac{b^2}{k} \left[\sum_{i=1}^{k} i^2 - k \left(\frac{k+1}{2} \right)^2 \right] \\
&= \frac{b^2}{k} \left[\frac{k\,(k+1)\,(2k+1)}{6} - \frac{k(k+1)^2}{4} \right] \\
&= b^2\,(k+1) \left(\frac{4k+2-3k-3}{12} \right) = \frac{b^2\,(k+1)\,(k-1)}{12} = \frac{b^2\,(k^2-1)}{12}. \qquad (4.9)
\end{aligned}$$

Moreover, if we use (4.2) and (4.7), then

$$\begin{aligned}
\operatorname{Var}(\overline{y}_{SRSWOR}) &= \frac{b^2 N\,(N+1)}{12n} \left(\frac{N-n}{N} \right) \\
&= \frac{b^2 N\,(N+1)}{12n} \left[\frac{n\,(k-1)}{N} \right] = \frac{b^2\,(N+1)\,(k-1)}{12}. \qquad (4.10)
\end{aligned}$$

Finally, using (4.3) and (4.7), results in

$$\operatorname{Var}\left(\overline{y}_{STR}\right)=\frac{b^2 k\left(k+1\right)}{12n}\left[\frac{n\left(k-1\right)}{nk}\right]=\frac{b^2\left(k+1\right)\left(k-1\right)}{12n}=\frac{b^2\left(k^2-1\right)}{12n}, \tag{4.11}$$

which follows since S_{wst}^2 is obtained by substituting k for N in (4.3), i.e. under the perfect trend model, for the setup of population units into strata (see Section 4.1.3), the sum of squares between population units within a stratum is related to the total sum of squares, with the only difference being that we are now considering k population units in $\sum\limits_{i=1}^{k}\left(y_{is}-\overline{y}_{\cdot s}\right)^2$, instead of N population units in $\sum\limits_{i=1}^{N}\left(y_i-\overline{Y}\right)^2$.

The relative efficiency of SRSWOR, with respect to LSS, is obtained by using (4.9) and (4.10), such that

$$\frac{\operatorname{Var}\left(\overline{y}_{LSS}\right)}{\operatorname{Var}\left(\overline{y}_{SRSWOR}\right)}=\frac{b^2\left(k^2-1\right)}{12}\left[\frac{b^2\left(N+1\right)\left(k-1\right)}{12}\right]^{-1}=\frac{\left(k+1\right)\left(k-1\right)}{\left(N+1\right)\left(k-1\right)}=\frac{k+1}{N+1}<1,$$

which follows if $n\geq 2$. We thus conclude that LSS is more efficient than SRSWOR if $n\geq 2$. Similarly, by using (4.9) and (4.11), we obtain the relative efficiency of STR, with respect to LSS, i.e.

$$\frac{\operatorname{Var}\left(\overline{y}_{LSS}\right)}{\operatorname{Var}\left(\overline{y}_{STR}\right)}=\frac{b^2\left(k^2-1\right)}{12}\left[\frac{b^2\left(k^2-1\right)}{12n}\right]^{-1}=n,$$

which is greater than 1 if $n\geq 2$. We thus conclude that LSS is less efficient than STR by a factor of n. Now, by using the transitive law with the assumption that $n\geq 2$, we show that

$$\operatorname{Var}\left(\overline{y}_{STR}\right)<\operatorname{Var}\left(\overline{y}_{LSS}\right)<\operatorname{Var}\left(\overline{y}_{SRSWOR}\right)<\operatorname{Var}\left(\overline{y}_{SRSWR}\right).$$

Hence, we conclude that STR is the most efficient probability sampling design for populations that exhibit linear trend. Furthermore, substituting $n=1$ into (4.9), (4.10) and (4.11), results in

$$\operatorname{Var}\left(\overline{y}_{STR}\right)=\operatorname{Var}\left(\overline{y}_{LSS}\right)=\operatorname{Var}\left(\overline{y}_{SRSWOR}\right)=\operatorname{Var}\left(\overline{y}_{SRSWR}\right).$$

which follows since we have proven the equivalence between SRSWR and SRSWOR in Section 4.1.2, for the case when $n=1$. Also, by applying (4.9) and (4.10), while assuming N to be relatively larger than k, we obtain an approximation for the relative efficiency of SRSWOR, with respect to LSS, given by

$$\begin{aligned}\frac{\operatorname{Var}\left(\overline{y}_{LSS}\right)}{\operatorname{Var}\left(\overline{y}_{SRSWOR}\right)}&=\frac{b^2\left(k^2-1\right)}{12}\left[\frac{b^2\left(N+1\right)\left(k-1\right)}{12}\right]^{-1}\\&=\frac{\left(k+1\right)\left(k-1\right)}{\left(N+1\right)\left(k-1\right)}=\frac{\left(k+1\right)}{\left(N+1\right)}\approxeq\frac{k}{N}=\frac{1}{n}.\end{aligned}$$

The relation of the efficiencies for the sampling designs, when N is relatively larger than k, is thus given as

$$\operatorname{Var}(\overline{y}_{STR}) : \operatorname{Var}(\overline{y}_{LSS}) : \operatorname{Var}(\overline{y}_{SRSWOR}) : \operatorname{Var}(\overline{y}_{SRSWR}) \approxeq \frac{1}{n^2} : \frac{1}{n} : 1 : 1.$$

We thus conclude that STR is more efficient than LSS by a factor of n, which in turn is more efficient than both SRSWOR and SRSWR, by an approximate factor of n, when N is relatively larger than k. As a result, STR is more efficient than both SRSWOR and SRSWR, by an approximate factor of n^2 (by using the transitive law).

Intra-class correlation coefficient

For populations that exhibit linear trend, we can expect a high degree of variation between population units that lie within the same systematic sample. The cross products of the pairs of population units that lie within the same systematic sample, with deviations calculated from $\overline{Y}$, are thus predominantly negative. A negative ICC ($\rho < 0$) is thus achieved and this results in LSS being more efficient than both SRSWR and SRSWOR, for this scenario. It should be noted that the greater the degree of trend in a population, the greater the efficiency gains when choosing LSS over either SRSWR or SRSWOR.

Now, let us view LSS in terms of STR (as in Section 4.1.3) for populations that exhibit linear trend. Strata are thus likely to be internally homogenous and if the ith unit is selected for each strata, then the deviation between any sampling unit and its respective stratum mean, will likely have the same coefficient as the deviation of other sampling units from their respective stratum means. Both deviations from their respective stratum means are likely to be either positive or negative, resulting in their cross products being predominantly positive, i.e. $\rho_{wst} > 0$. We thus conclude that STR is more efficient than LSS for populations that exhibit linear trend. It should be noted that the greater the degree of trend in a population, the greater the efficiency loss when choosing LSS over STR.

4.2.3 Periodic populations

All periodic populations have a period, where the period is defined as an interval, in which the variate values of a population perform a complete cycle. The variate values follow regular oscillations that are repeated (cycle), i.e. the variate values of the population monotonically increase and then monotonically decrease at regular intervals. "*Periodic* *variations are likely to occur in certain natural populations such as land fertility, forest growth, events over time, etc. and in records such as payroll, census list of individuals arranged by households, etc.*" (Murthy & Rao 1988, p.157). Finney (1950) provides an example for a population that exhibits unexplainable periodic variation, but this claim was considered to be invalid by Milne (1959), since the calculations applied by Finney was owing to measurement errors, which created a false result of periodic variation. Other examples of periodic populations are given by Madow (1946) and Matérn (1960). Initial comparisons between LSS and STR for periodic populations, were obtained by Madow & Madow (1944).

To view the effect of LSS for periodic populations, we will consider a discrete hypothetical periodic population, given in Figure 4.1, where the variate values of the population are plot against the population unit numbers ($i = 1, ..., 24$). The period for this hypothetical population is given as AB=8, since we obtain a complete cycle between AB, before the cycle gets repeated. Moreover, the population mean can easily be interpreted from the graph, i.e. $\overline{Y} = 2.5$. Now, if we conduct LSS with $k = 8$ and a random start $i = 1$, then we obtain a sample which has no variation amongst the sampling units, i.e. we obtain complete homogeneity within the systematic samples, such that $\rho = 1$ and $\text{Var}(\overline{y}_{LSS}) = S_Y^2(N-1)/N$ (see equation (3.8)). This is no different to randomly selecting a sample of size one and hence results in SRSWR, SRSWOR and STR, providing more efficient results than LSS. Conducting LSS with $k = 8$ and random start $i = 3$, results in a sample that correctly estimates $\overline{Y}$. Nevertheless, we still consider this estimator to be inefficient and inaccurate, since all possible systematic sample means, with k being equal to the period, will not capture the variance explained by the population and will also, on average inaccurately estimate $\overline{Y}$. Finally, conducting LSS with $k = 4$ and random start $i = 4$, results in the variate values of each successive pair of sampling units being equidistant from $\overline{Y}$, i.e. the average of the variate values, for every successive pair of sampling units, is equivalent to $\overline{Y}$. Furthermore, the sampling variance is zero when k is half the period and n is even, since all the possible systematic sample means are equal to $\overline{Y}$, i.e. there is complete heterogeneity within the systematic samples and thus $\rho = -1/(n-1)$, which results in $\text{Var}(\overline{y}_{LSS}) = 0$ (see Section 3.1.2). Consequently, LSS is more efficient than SRSWR, SRSWOR, and STR for this scenario. By defining a similar hypothetical population, we can then show that $\text{Var}(\overline{y}_{LSS}) \neq 0$ if n is odd and k is half the period. This is owing to us obtaining an extra unit sampling after pairing the $(n-1)$ successive

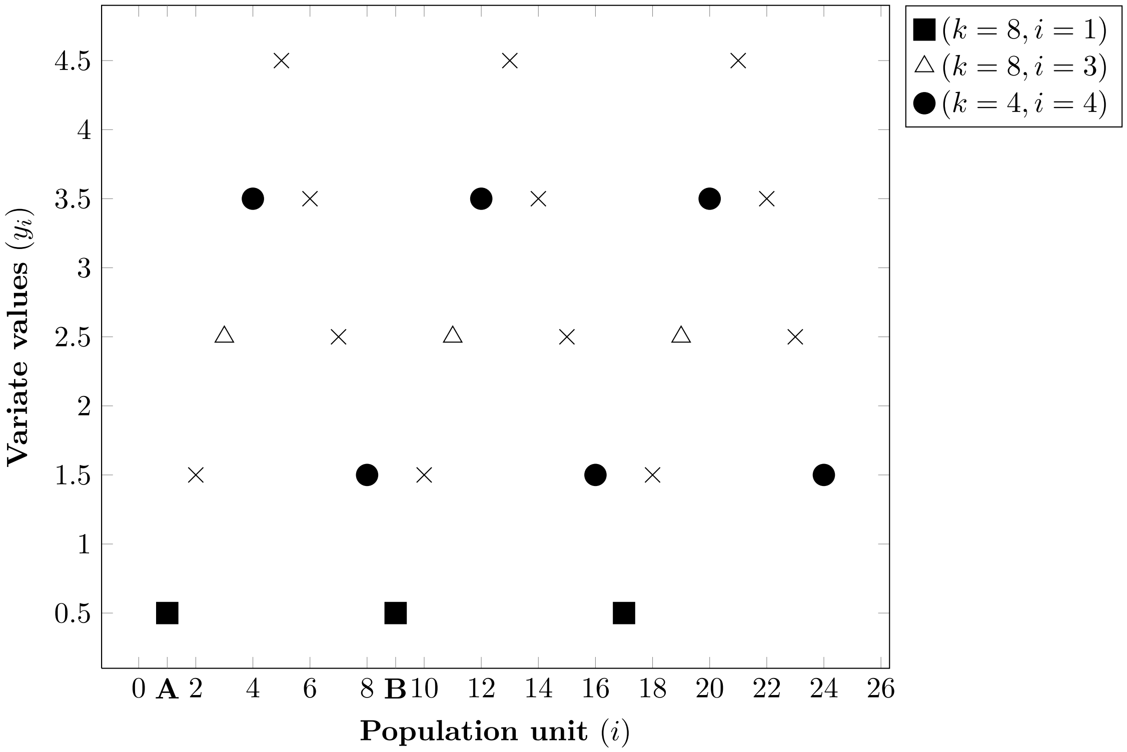


Figure 4.1: Selecting samples from a periodic population of size 24 with period AB=8, using LSS

sampling units, where the average of the variate values of each of these pairs is equal to $\overline{Y}$. Nevertheless, LSS is still much more efficient than SRSWR, SRSWOR and STR for this scenario.

A hypothetical population of this nature and those that exhibit an exact sine curve are rare in practice. However, we can generalize the results for the hypothetical population and apply it to all realistic periodic populations. We thus conclude that LSS is more efficient than SRSWR, SRSWOR and STR, for periodic populations, if k is equal an odd multiple of half the period (Cochran 1977). Conversely, LSS is less efficient than SRSWR, SRSWOR and STR, if k is an integral multiple of the period (Cochran 1977). Moreover, we obtain efficiency gains when conducting LSS when n is even, as opposed to the case when n is odd, if the sampling interval is equal to an odd multiple of half the period. It is thus of great importance that a sampler recognizes if a population exhibits periodic variation before sampling, so as to remove any periodicity bias, by selecting an appropriate sampling interval.

4.2.4 Auto-correlated populations

So far we have shown that the efficiency of LSS depends on the arrangement/ordering of population units and we provided some assumptions, i.e. random order, linear trend and periodic. These assumptions are based on a single finite population which results in inconsistencies, when comparing the efficiencies between sampling designs. We will therefore consider a more realistic method of comparison, as shown below, which was originally introduced by Cochran (1946).

Auto-correlated populations are described by the phenomenon where variate values of population units which occur closer, in a given population, are more alike (higher correlation), as compared to those that occur further apart. In this notation, we use ρ_u to denote the serial correlation for the pair of population units y_i and y_j $(i \neq j)$, such that $u = |i - j|$ represents the distance between these pairs. To test if a population exhibits autocorrelation, one can plot a correlogram, where ρ_u, for y_i and y_j, is plot against u (Cochran 1977). Cochran (1946) used this notation to introduce the super-population model, which assumes that the population units for a finite population are drawn at random from an infinite super-population. Now, we can obtain efficiency comparisons from an average of many finite populations and these results will converge with the finite population results, as the finite population increases, i.e. as N becomes larger. Accordingly, the super-population model is given by

$$\mathrm{E}_m\left(y_i\right) = \mu, \qquad \mathrm{E}_m(y_i - \mu)^2 = \sigma^2, \qquad \mathrm{E}_m\left(y_i - \mu\right)\left(y_{i+u} - \mu\right) = \rho_u \sigma^2, \tag{4.12}$$

where the function E_m denotes the average of all possible finite populations, which can be selected from this super-population. The model is based on the assumptions that:

$$\rho_u \geq 0 \text{ (i.e. } \rho_u \text{ is positive);} \tag{4.13}$$

$$\Delta\rho_u = \rho_{u+1} - \rho_u \leq 0 \text{ (i.e. } \rho_u \text{ is decreasing).} \tag{4.14}$$

We will first consider some preliminary results, before comparing LSS to the other probability sampling designs.

A common identity used in the ANOVA is given as

$$N \sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2 = \frac{1}{2} \sum_{i=1}^{N} \sum_{\substack{j=1 \\ j \neq i}}^{N} (y_i - y_j)^2,$$

which can be expressed as

$$\begin{aligned}\sum_{i=1}^{N}\left(y_i-\overline{Y}\right) &= \frac{1}{N}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(y_i-y_j\right)^2\\ &= \frac{(N-1)}{2}\times\frac{2}{N\,(N-1)}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(y_i-y_j\right)^2\\ &= \frac{(N-1)}{2}\mathrm{E}(y_i-y_j)^2\\ &= \frac{(N-1)}{2}\mathrm{E}\big[\left(y_i-\overline{Y}\right)-\left(y_j-\overline{Y}\right)\big]^2\\ &= \frac{(N-1)}{2}\mathrm{E}\left[\left(y_i-\overline{Y}\right)^2+\left(y_j-\overline{Y}\right)-2\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\right], \qquad (4.15)\end{aligned}$$

since there are $N(N-1)/2$ pairs of (y_i, y_j), where $j > i$. Now, in the $N(N-1)/2$ pairs of (y_i, y_j), there are $(N-1)$ pairs where $u = 1$, $(N-2)$ pairs where $u = 2$, and so forth. Thus, by averaging (4.15) over all possible finite populations, we obtain

$$\begin{aligned}\mathrm{E}_m\left[\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2\right] &= \frac{N-1}{2}\mathrm{E}\left[\mathrm{E}_m\left\{\left(y_i-\overline{Y}\right)^2+\left(y_i-\overline{Y}\right)^2-2\left(y_i-\overline{Y}\right)\left(y_j-\overline{Y}\right)\right\}\right]\\ &= \frac{N-1}{2}\mathrm{E}\left[2\sigma^2-2\sigma^2\rho_u\right]\\ &= (N-1)\,\sigma^2\left[1-\frac{2}{N\,(N-1)}\sum_{u=1}^{N-1}(N-u)\,\rho_u\right], \qquad (4.16)\end{aligned}$$

which follows from (4.12). The expected sum of squares for a single stratum is obtained by replacing N in (4.16) by k, since there are k population units within each stratum (refer to reasoning in Section 4.2.2). Moreover, the expected sum of squares for each of the n strata is equivalent, resulting in

$$\mathrm{E}_m\,(SS \text{ within strata}) = n\,(k-1)\,\sigma^2\left[1-\frac{2}{k\,(k-1)}\sum_{u=1}^{k-1}(k-u)\,\rho_u\right]. \qquad (4.17)$$

The expected sum of squares for a single systematic sample is obtained by replacing ρ_u and N in (4.16), by ρ_{ku} and n respectively, since correlations between consecutive population units are $\rho_k, \rho_{2k}, \rho_{3k}, ...$, rather than $\rho_1, \rho_2, \rho_3, ...$, and there are n units in the sample. Furthermore, the expected sum of squares for each of the k systematic samples is equivalent, which results in

$$\mathrm{E}_m\,(SSW) = k\,(n-1)\,\sigma^2\left[1-\frac{2}{n\,(n-1)}\sum_{u=1}^{n-1}(n-u)\,\rho_{ku}\right]. \qquad (4.18)$$

Now, if we apply (4.2) and (4.16), then we obtain the expected variance of $\overline{y}_{SRSWOR}$, given by

$$\begin{aligned}\sigma^2_{SRSWOR} &= \frac{(k-1)}{N(N-1)}\mathrm{E}_m\left[\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2\right]\\ &= \frac{(k-1)\,\sigma^2}{N}\left[1-\frac{2}{N(N-1)}\sum_{u=1}^{N-1}(N-u)\,\rho_u\right]. \qquad (4.19)\end{aligned}$$

Likewise, by applying (4.3) and (4.17), we obtain the expected variance of $\overline{y}_{STR}$, which is given as

$$\begin{aligned}\sigma^2_{STR} &= \mathrm{E}_m\left[\frac{(N-n)}{n^2(k-1)\,N}\sum_{s=1}^{n}\sum_{i=1}^{k}\left(y_{is}-\overline{y}_{\cdot s}\right)^2\right]\\ &= \frac{1}{Nn}\mathrm{E}_m(SS\text{ within strata}) = \frac{(k-1)\,\sigma^2}{N}\left[1-\frac{2}{k(k-1)}\sum_{u=1}^{k-1}(k-u)\,\rho_u\right]. \qquad (4.20)\end{aligned}$$

Finally, by respectively using (3.20), (3.15), (3.12), (4.16) and (4.18), we obtain the expected variance of $\overline{y}_{LSS}$, which is given by

$$\begin{aligned}\sigma^2_{LSS} &= \mathrm{E}_m\left[\frac{SSB}{N}\right]\\ &= \frac{1}{N}\left[\mathrm{E}_m\left\{\sum_{i=1}^{N}\left(y_i-\overline{Y}\right)^2\right\}-\mathrm{E}_m\left(SSW\right)\right]\\ &= \frac{\sigma^2}{N}\left[(N-1)-\frac{2}{N}\sum_{u=1}^{N-1}(N-u)\,\rho_u-\left\{k(n-1)-\frac{2k}{n}\sum_{u=1}^{n-1}(n-u)\rho_{ku}\right\}\right]\\ &= \frac{\sigma^2}{N}\left[(k-1)-\frac{2}{N}\sum_{u=1}^{N-1}(N-u)\,\rho_u+\frac{2k}{n}\sum_{u=1}^{n-1}(n-u)\,\rho_{ku}\right]\\ &= \frac{(k-1)\,\sigma^2}{N}\left[1-\frac{2}{N(k-1)}\sum_{u=1}^{N-1}(N-u)\,\rho_u+\frac{2k}{n(k-1)}\sum_{u=1}^{n-1}(n-u)\,\rho_{ku}\right]. \qquad (4.21)\end{aligned}$$

The expected relative efficiency of STR, with respect to SRSWOR, is then obtained by applying (4.19) and (4.20), such that

$$\begin{aligned}\frac{\sigma^2_{SRSWOR}}{\sigma^2_{STR}} &= \frac{(k-1)\,\sigma^2}{N}\left[1-\frac{2\sum_{u=1}^{N-1}(N-u)\,\rho_u}{N(N-1)}\right]\left[\frac{(k-1)\,\sigma^2}{N}\left\{1-\frac{2\sum_{u=1}^{k-1}(k-u)\,\rho_u}{k(k-1)}\right\}\right]^{-1}\\ &= 1-\frac{2}{N(N-1)}\sum_{u=1}^{N-1}(N-u)\,\rho_u\left[1-\frac{2}{k(k-1)}\sum_{u=1}^{k-1}(k-u)\,\rho_u\right]^{-1} > 1,\end{aligned}$$

which follows if we use the conditions in (4.13) and (4.14) (Cochran 1946). By using the relationship between SRSWR and SRSWOR, as in Section 4.1.2, it thus follows that

$$\sigma^2_{STR} \le \sigma^2_{SRSWOR} \le \sigma^2_{SRSWR}, \qquad (4.22)$$

for all $n \geq 1$, while equivalence occurs when $n = 1$. It should be noted that there is no general result for the expected relative efficiency of LSS, with respect to either SRSWOR or STR. However, we can obtain results by introducing an additional assumption, as shown in the next theorem, and then apply the transitive law to (4.22).

Theorem 3.1: If we provide an additional assumption to (4.13) and (4.14), such that the correlogram is convex, i.e.

$$\Delta^2 \rho_u = \delta_u^2 = \rho_{u+1} + \rho_{u-1} - 2\rho_u \geq 0, \qquad \text{for } u = 2, ..., (N-2), \tag{4.23}$$

then

$$\sigma_{LSS}^2 \leq \sigma_{STR}^2 \leq \sigma_{SRSWOR}^2 \leq \sigma_{SRSWR}^2,$$

for any n, where LSS and STR are equally efficient when $\delta_u^2 = 0$, for $u = 2, ..., (N-2)$.

Proof: The complete proof of this theorem is given by Cochran (1946); however, we will provide a simplified version for the special case, which considers the sampling of two units from two strata. The variance of $\overline{y}_{STR}$ is thus given as

$$\begin{aligned} \text{Var}\left(\overline{y}_{STR}\right) &= \text{Var}\left[\frac{y_i + y_j}{2}\right] \\ &= \frac{1}{4}\left[\text{Var}\left(y_i\right) + \text{Var}\left(y_j\right) + 2\text{Cov}\left(y_i, y_j\right)\right] \\ &= \frac{1}{4}\left[2\sigma^2 + 2\sigma^2 \rho_u\right] = \frac{\sigma^2}{2}\left[1 - \rho_u\right]. \end{aligned} \tag{4.24}$$

Table 4.1 shows the number of combinations for each pair of population units, with respect to their distance apart, for a population of size $2k$. By using Table 4.1, we obtain the total number of combinations of pairs of population units, which is given as

$$\begin{aligned} T_c &= 1 + 2 + ... + (k-1) + k + (k-1) + ... + 1 \\ &= k + 2\sum_{u=1}^{k-1} u = k + 2\left[\frac{k(k-1)}{2}\right] = k^2. \end{aligned} \tag{4.25}$$

Table 4.1: Number of combinations for each pair of population units, with their corresponding distance apart, when $N = 2k$

Distance Apart	1	2	...	$k-1$	k	$k+1$	...	$2k-1$
Number of Combinations	1	2	...	$k-1$	k	$k-1$	...	1

Thus, by applying (4.24) and (4.25), we can obtain the expected variance of $\overline{y}_{STR}$, over k^2 combinations, which is given by

$$\begin{aligned}\sigma^2_{STR} &= \mathrm{E}_m\left[\frac{\sigma^2}{2}\left(1+\rho_u\right)\right]\\ &= \frac{\sigma^2}{2}\mathrm{E}_m\left[1+\rho_u\right] = \frac{\sigma^2}{2k^2}\sum\left[1+\rho_u\right] = \frac{\sigma^2}{2k^2}\left[k^2+\sum\rho_u\right]. \qquad (4.26)\end{aligned}$$

Now, by using Table 4.1, it follows that

$$\begin{aligned}\sum\rho_u &= 1\rho_1 + ... + (k-1)\,\rho_{k-1} + k\rho_k + (k-1)\,\rho_{k+1} + ... + 1\rho_{2k-1}\\ &= \sum_{u=1}^{k-1} u\rho_u + k\rho_k + \sum_{u=1}^{k-1} u\rho_{2k-u}. \qquad (4.27)\end{aligned}$$

Hence, substituting (4.25) and (4.27) into (4.26), results in

$$\begin{aligned}\sigma^2_{STR} &= \frac{\sigma^2}{2k^2}\left[k + 2\sum_{u=1}^{k-1} u + \sum_{u=1}^{k-1} u\rho_u + k\rho_k + \sum_{u=1}^{k-1} u\rho_{2k-u}\right]\\ &= \frac{\sigma^2}{2k^2}\left[\sum_{u=1}^{k-1} u\left(2+\rho_u+\rho_{2k-u}\right) + k\left(1+\rho_k\right)\right]. \qquad (4.28)\end{aligned}$$

The variance of $\overline{y}_{LSS}$ is obtained by replacing ρ_u in (4.24) with ρ_k, since each pair of sampling units is k units apart if $n = 2$. Accordingly,

$$\mathrm{Var}(\overline{y}_{LSS}) = \frac{\sigma^2}{2}[1+\rho_k]. \qquad (4.29)$$

Similarly, as in the case of STR, we use Table 4.1 to obtain the total number of combinations, which is given as k^2 in (4.25). By applying (4.29), we can thus obtain the expected variance of $\overline{y}_{LSS}$, over k^2 combinations, which is given as

$$\begin{aligned}\sigma^2_{LSS} &= \mathrm{E}_m\left[\frac{\sigma^2}{2}\left(1+\rho_k\right)\right]\\ &= \frac{\sigma^2}{2}\mathrm{E}_m\left[1+\rho_k\right] = \frac{\sigma^2}{2k^2}\sum\left[1+\rho_k\right] = \frac{\sigma^2}{2k^2}\left[k^2+\sum\rho_k\right]. \qquad (4.30)\end{aligned}$$

By using Table 4.1, where every value in the distance apart row is replaced with k, we obtain

$$\begin{aligned}\sum\rho_u &= 1\rho_k + 2\rho_k + ... + (k-1)\,\rho_k + k\rho_k + (k-1)\,\rho_k + ... + 1\rho_k\\ &= 2\sum_{u=1}^{k-1} u\rho_k + k\rho_k. \qquad (4.31)\end{aligned}$$

Hence, substituting (4.25) and (4.31) into (4.30), results in

$$\begin{aligned}\sigma^2_{LSS} &= \frac{\sigma^2}{2k^2}\left[k + 2\sum_{u=1}^{k-1} u + 2\sum_{u=1}^{k-1} u\rho_k + k\rho_k\right] \\ &= \frac{\sigma^2}{2k^2}\left[\sum_{u=1}^{k-1} u\left(2+2\rho_k\right) + k\left(1+\rho_k\right)\right]. \qquad (4.32)\end{aligned}$$

By comparing (4.28) to (4.32), we note that $\sigma^2_{LSS} \leq \sigma^2_{STR}$ if and only if

$$\sum_{u=1}^{k-1} u\left(2+2\rho_k\right) \leq \sum_{u=1}^{k-1} u\left(2+\rho_u+\rho_{2k-u}\right),$$

or equivalently, if and only if

$$\sum_{u=1}^{k-1} u\left[\left(\rho_u-\rho_k\right)+\left(\rho_{2k-u}-\rho_k\right)\right] \geq 0. \qquad (4.33)$$

Now, by using the assumptions given in (4.14) and (4.23), we obtain

$$\delta^2_u = \rho_{u+1} + \rho_{u-1} - 2\rho_u \geq 0, \qquad \text{for } u = 2, ..., (N-2),$$

i.e.

$$\nabla_{u-1} = \rho_{u-1} - \rho_u \geq \nabla_u = \rho_u - \rho_{u+1} \geq 0.$$

Thus

$$\nabla_1 \geq \nabla_2 \geq ... \geq 0,$$

implies

$$\nabla_u + \nabla_{u-1} + ... + \nabla_{u+t-1} \geq \nabla_{u+t} + \nabla_{u+t+1} + ... + \nabla_{u+2t-1},$$

i.e.

$$\rho_u - \rho_{u+t} \geq \rho_{u+t} - \rho_{u+2t}, \qquad \text{for } t \geq 1.$$

For $t = k - u$, we obtain

$$\rho_u - \rho_k \geq \rho_k - \rho_{2k-u}.$$

We then conclude the proof by applying the above result to (4.33).

Cochran (1977) notes that the assumptions in (4.13), (4.14) and (4.23) are satisfied for the cases where the correlograms are linear, exponential and hyperbolic tangent. Examples of natural populations that exhibit positive convex decreasing correlograms are given by: (*i*) a linear autocorrelation function of $\rho_u = (l-u)/l$, for particular classes of economic time series, proposed by Wold (1938); (*ii*) an exponential autocorrelation function of $\rho_u = \exp(-\lambda u)$, for forestry and land use/cover area frame surveys, proposed by

Osborne (1942) and Matérn (1947); (*iii*) a hyperbolic tangent autocorrelation function of $\rho_u = \tanh(u^{-3/5})$, for investigating the weekly rainfall for two meteorological stations that are located at a distance of u units apart, proposed by Fisher & Mackenzie (1922). Furthermore, Bellhouse (1988) noted that any process which is autoregressive and has real roots, with respect to the characteristic equation, will exhibit a positive convex decreasing correlogram.

4.2.5 Stratified populations

If the entire population is divided into groups (or strata) that are internally homogenous and externally heterogeneous (i.e. population units within a stratum are alike according to some characteristic and strata differ amongst each other according to some characteristic), then this population is known as a stratified population. Stratified populations may be naturally defined or they may be defined by a sampler, e.g. a sampler may divide the population into strata according to some characteristic, which is related to the variable of interest. Examples of naturally defined stratified populations exist in certain multi-stage sampling designs, where the strata may be defined as provinces, municipalities, regions etc.

Multi-stage designs involve a nesting population structure of more than two categorical stages, such that the first stage involves dividing the population into primary sampling units (PSUs), the second stage involves dividing the primary sampling units into secondary sampling units (SSUs), and so forth, until a sampling design is employed to select the final stage sampling units, which collectively form the sample. If at one of the stages we divide the sampling units into strata and then at the next stage we apply an independent SRS design within each stratum, then we are simply employing a STR design for those two particular stages of sampling. Alternatively, we can use an independent systematic sampling design within each stratum for the corresponding stage of sampling, which is then termed as stratified systematic sampling. Stratified systematic sampling and STR are thus referred to as a two-stage sampling designs, where the first stage involves the constructing of strata and the second stage involves the independent random selection of units within the strata. An example of stratified systematic sampling for a large scale survey is given by Arnab & North (2012).

Madow & Madow (1944) provides us with two stratified systematic sampling designs, where the first design assumes that the sampling interval within each stratum is equal, while the second design assumes unequal sampling intervals, i.e. for a population that is divided into L strata, the sampling interval for the jth stratum is given by k_j, for $j \in \{1, ..., L\}$.

Stratified systematic sampling is more often than not, more efficient than STR, for the case where strata are considered to be large and more than one unit is drawn from each stratum. This is because stratified systematic sampling ensures a more even spread of the stratum sample over the corresponding stratum, for each of the strata, as compared to STR (Murthy & Rao 1988). This preference is used with the objective to reduce the variance within strata and results in efficiency gains, if and only if systematic sampling within strata is more precise than SRS within strata.

So far we have considered sampling from populations in its original state. It may thus be advantageous to order the population units before sampling, so as to make systematic sampling more efficient. A sampler may then opt to order the population in ascending/descending order according to some auxiliary variable (a readily available variable which is correlated to the variable of interest, such that the variate values of this variable are easier to obtain, than those of the variable of interest). The resulting effect is a population that approximately exhibits linear trend, where the stronger the degree of correlation between the auxiliary variable and the variable of interest, results in a stronger degree of linear trend. The results obtained in this chapter, for populations that exhibit linear trend, may then apply. In the next chapter, we will examine and compare various designs of LSS, for populations that exhibit linear trend.

Chapter 5

LINEAR SYSTEMATIC SAMPLING DESIGNS IN THE PRESENCE OF LINEAR TREND

In this chapter, we will discuss LSS designs which are considered to be optimal, when sampling from a population that exhibits linear trend. By assuming a linear trend model averaged over the super-population model, we will compare the various designs by comparing the expected MSEs of the corresponding sample means. We will first introduce some preliminary results, in which we will provide a generic formula for calculating the expected MSEs of the sample means. Thereafter, we will discuss some designs, which include YEC, CESS, BSS, MSS and a new proposed design termed as BMSS. It should be noted that the designs which are discussed in this chapter are not restricted to populations that exhibit linear trend only, such that they could also be shown to be useful for other population structures (numerical results for these designs on other population structures are given in Chapter 8). Note that throughout this chapter we will assume that k is an integer, so that we will be conducting sampling linearly.

5.0 Preliminary Results

The model for a hypothetical population that exhibits perfect linear trend, given by (4.5), is considered to be unrealistic. A more realistic model for linear trend is given by

$$y_i = a + bi + e_i, \qquad \text{for } i = 1, ..., N, \tag{5.1}$$

where e_i denotes the random error which follows the super-population model, given by Cochran (1946), such that

$$\mathrm{E}_m\left(e_i\right) = 0, \qquad \mathrm{E}_m\left(e_i^2\right) = \sigma^2, \qquad \mathrm{E}_m\left(e_ie_j\right) = 0\left(i \neq j\right). \tag{5.2}$$

By using (5.1), we obtain

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{1}{N}\sum_{i=1}^{N} a + \frac{b}{N}\sum_{i=1}^{N} i + \frac{1}{N}\sum_{i=1}^{N} e_i = a + \frac{b\left(N+1\right)}{2} + \overline{\overline{e}}, \tag{5.3}$$

which follows since $\overline{\overline{e}} = \sum\limits_{i=1}^{N} e_i/N$ is the mean random error of the population.

Theorem 5.1: With the assumption of equal weights $(1/n)$ being applied to all the sampling units, the expected MSE of any sample mean, for the model in (5.1), is given by

$$\begin{aligned} \mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{LT}\right) &\triangleq \mathrm{E}_m\left[\mathrm{E}\left(\left\{\overline{y}_{LT} - \overline{Y}\right\}^2\right)\right] \\ &= \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) + \mathrm{Var}\left(\overline{y}_{PLT}\right), \end{aligned} \tag{5.4}$$

where $\overline{y}_{PLT}$ denotes a linear unbiased estimator of (4.6), using the probability sampling design associated with $\overline{y}_{LT}$.

Proof: By using (5.1) and (4.5), we obtain

$$\overline{y}_{LT} = \overline{y}_{PLT} + \overline{e}_i,$$

where $\overline{e}_i = \sum e_i/n$ denotes the mean random error of the sample and $\sum$ denotes the sum over the sample. Now, by using the above expression along with (5.3) and (4.6), we obtain an expression for the expected MSE of $\overline{y}_{LT}$, given by

$$\begin{aligned} \mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{LT}\right) &\triangleq \mathrm{E}_m\left[\mathrm{E}\left(\left\{\overline{y}_{LT} - \overline{Y}\right\}^2\right)\right] \\ &= \mathrm{E}_m\left[\mathrm{E}\left(\left\{\overline{y}_{PLT} - a - \frac{b\left(N+1\right)}{2}\right\}^2 + \left\{\overline{e}_i - \overline{\overline{e}}\right\}^2\right)\right] \\ &= \mathrm{E}_m\mathrm{Var}\left(\overline{y}_{PLT}\right) + \mathrm{E}\left[\mathrm{E}_m\left(\overline{e}_i^2 - 2\overline{e}_i\overline{\overline{e}} + \overline{\overline{e}}^2\right)\right], \end{aligned} \tag{5.5}$$

which follows if we use the conditions in (5.2), i.e.

$$\mathrm{E}_m\mathrm{E}\left[2\left(\overline{y}_{PLT}-a-\frac{b\,(N+1)}{2}\right)\left(\overline{e}_i-\overline{\overline{e}}\right)\right]=0.$$

Furthermore, by applying the conditions in (5.2), we obtain

$$\begin{aligned}
\mathrm{E}_m\left(\overline{e}_i^2\right) &= \mathrm{E}_m\left[\left(\frac{1}{n}\sum e_i\right)^2\right] \\
&= \mathrm{E}_m\left[\frac{1}{n^2}\left(\sum e_i^2+\sum\sum_{j\neq i} e_ie_j\right)\right] \\
&= \frac{1}{n^2}\left[\sum \mathrm{E}_m\left(e_i^2\right)+\sum\sum_{j\neq i}\mathrm{E}_m\left(e_ie_j\right)\right] = \frac{1}{n^2}\sum\sigma^2=\frac{\sigma^2}{n},
\end{aligned} \tag{5.6}$$

$$\begin{aligned}
\mathrm{E}_m\left(\overline{\overline{e}}^2\right) &= \mathrm{E}_m\left[\left(\frac{1}{N}\sum_{j=1}^{N} e_j\right)^2\right] \\
&= \mathrm{E}_m\left[\frac{1}{N^2}\left(\sum_{j=1}^{N} e_j^2+\sum_{i=1}^{N}\sum_{j\neq i}^{N} e_ie_j\right)\right] \\
&= \frac{1}{N^2}\left[\sum_{j=1}^{N}\mathrm{E}_m\left(e_j^2\right)+\sum_{i=1}^{N}\sum_{j\neq i}^{N}\mathrm{E}_m\left(e_ie_j\right)\right] = \frac{1}{N^2}\sum_{j=1}^{N}\sigma^2=\frac{\sigma^2}{N},
\end{aligned} \tag{5.7}$$

and

$$\mathrm{E}_m\left(\overline{e}_i\overline{\overline{e}}\right)=\mathrm{E}_m\left(\frac{1}{nN}\sum\sum_{j=1}^{N} e_ie_j\right)=\frac{1}{nN}\sum\sum_{j=1}^{N}\mathrm{E}_m\left(e_ie_j\right)=\frac{n\sigma^2}{nN}=\frac{\sigma^2}{N}, \tag{5.8}$$

which follows since a sample of size n results in $e_i = e_j$ occurring n times. We then conclude the proof by substituting (5.6), (5.7) and (5.8) into (5.5), i.e.

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{LT}\right)=\mathrm{E}_m\mathrm{Var}\left(\overline{y}_{PLT}\right)+\mathrm{E}\left[\frac{\sigma^2}{n}-\frac{2\sigma^2}{N}+\frac{\sigma^2}{N}\right]=\mathrm{Var}\left(\overline{y}_{PLT}\right)+\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right).$$

It should be noted that Var($\overline{y}_{PLT}$) represents the linear trend component, while the assumption of equal weights being applied to all the sampling units results in a minimum expected error variance component, represented by $\sigma^2(1/n - 1/N)$. Hence, the most desirable sampling design(s) for populations that exhibit linear trend, are those that are associated with estimator(s) that completely remove the linear trend component (i.e. $\overline{y}_{PLT} = a + b(N + 1)/2$) and exhibit minimum expected error variance.

By substituting either (4.9), (4.10) or (4.11) into (5.4), we respectively obtain the expected MSEs for $\overline{y}_{LSS}$ $\overline{y}_{SRSWOR}$, and $\overline{y}_{STR}$, such that

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{LSS}\right) = \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{b^2\left(k^2-1\right)}{12}, \tag{5.9}$$

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{SRSWOR}\right) = \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{b^2\left(N+1\right)\left(k-1\right)}{12}, \tag{5.10}$$

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{STR}\right) = \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{b^2\left(k^2-1\right)}{12n}, \tag{5.11}$$

which follows since there are equal weights being applied to all the sampling units and $\overline{y}_{LSS}$, $\overline{y}_{SRSWOR}$, and $\overline{y}_{STR}$, are all design unbiased estimators of $\overline{Y}$. With the assumption of $n \geq 2$, we obtain error comparisons using (5.9) to (5.11), where the result is then given as

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{STR}\right) < \mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{LSS}\right) < \mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{SRSWOR}\right) < \mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{SRSWR}\right). \tag{5.12}$$

We will next discuss various LSS designs for populations that exhibit linear trend. Within each design we will (*i*) discuss the corresponding methodology, (*ii*) obtain expected MSE formulae, and (*iii*) compare the corresponding expected errors, to that of all the previously discussed designs.

5.1 Yates End Corrections

5.1.1 Methodology

Yates (1948) proposed an estimator that eliminates linear trend. The design is equivalent to LSS; however, an estimate of $\overline{Y}$ is obtained by applying weights to the first and the last sampling units, as shown in the next theorem.

Theorem 5.2: The YEC estimator of $\overline{Y}$ with random start i, for $i \in \{1, ..., k\}$, is given as

$$\overline{y}_{YEC} = \overline{y}_{LSS} + \frac{(2i-k-1)}{2\left(n-1\right)k}\left(y_i - y_{i+(n-1)k}\right). \tag{5.13}$$

Proof: An estimate of $\overline{Y}$ with random start i, for $i \in \{1, ..., k\}$, is given as

$$\overline{y}_{YEC} = \frac{1}{n}\left[\lambda_1 y_i + \left(\sum_{j=1}^{n-2} y_{i+jk}\right) + \lambda_2 y_{i+(n-1)k}\right], \tag{5.14}$$

where λ_1 and λ_2 are the weights applied to the first and the last sampling units, respectively. By substituting (4.5) into (5.14) and then equating this result to (4.6), we obtain

$$\begin{aligned}\overline{y}_{YEC} &= \frac{1}{n}\left[\lambda_1\left(a+bi\right)+\left(\sum_{j=1}^{n-2} a+b\left(i+jk\right)\right)+\lambda_2\left\{a+b\left[i+\left(n-1\right)k\right]\right\}\right] \\ &= a+\frac{b\left(N+1\right)}{2}. \end{aligned} \tag{5.15}$$

Now, by equating the coefficients of a in (5.15), it follows that

$$\lambda_1 = 2-\lambda_2. \tag{5.16}$$

Similarly, by equating the coefficients of b in (5.15), we obtain

$$\frac{1}{n}\left[\lambda_1 i+\left(n-2\right)i+\frac{\left(n-1\right)\left(n-2\right)k}{2}+\lambda_2 i+\lambda_2\left(n-1\right)k\right]=\frac{N+1}{2}. \tag{5.17}$$

Substituting (5.16) into (5.17), results in

$$2\left[2i-\lambda_2 i+\left(n-2\right)i+\frac{\left(n-1\right)\left(n-2\right)k}{2}+\lambda_2 i+\lambda_2\left(n-1\right)k\right]=n\left(N+1\right),$$

which simplifies to

$$\lambda_2 = 1-\frac{n\left(2i-k-1\right)}{2\left(n-1\right)k}. \tag{5.18}$$

The weight applied to the first sampling unit is thus obtained by substituting (5.18) into (5.16), such that

$$\lambda_1 = 1+\frac{n\left(2i-k-1\right)}{2\left(n-1\right)k}. \tag{5.19}$$

We then conclude the proof, by substituting (5.18) and (5.19) into (5.14), i.e.

$$\begin{aligned}\overline{y}_{YEC} &= \frac{1}{n}\left[y_i+\frac{n\left(2i-k-1\right)}{2\left(n-1\right)k}y_i+\left(\sum_{j=1}^{n-2} y_{i+jk}\right)+y_{i+(n-1)k}-\frac{n\left(2i-k-1\right)}{2\left(n-1\right)k}y_{i+(n-1)k}\right] \\ &= \overline{y}_{LSS}+\frac{\left(2i-k-1\right)}{2\left(n-1\right)k}\left(y_i-y_{i+(n-1)k}\right),\end{aligned}$$

where $y_i+\sum_{j=1}^{n-2} y_{i+jk}+y_{i+(n-1)k}=\sum_{j=1}^{n} y_{i+(j-1)k}=n\overline{y}_{LSS}$. This estimator is unbiased and $\mathrm{Var}(\overline{y}_{YEC})=0$, for the model in (4.5) (i.e. the linear trend component is completely removed), since we constructed $\overline{y}_{YEC}$ by equating it to (4.6). For the realistic linear trend model in (5.1), we can expect this estimator to be slightly biased; however, estimator $\overline{y}_{YEC}$ is usually more efficient than $\overline{y}_{LSS}$ (Murthy & Rao 1988).

5.1.2 Expected mean square error

The generic expected MSE formula, given in (5.4), assumes equal weights are applied to all the sampling units and is thus not applicable for the YEC estimator. The more appropriate method for obtaining the expected MSE of $\overline{y}_{YEC}$ is given as follows:

$$\begin{aligned}\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{YEC}\right) &\triangleq \mathrm{E}_m\left[\mathrm{E}\left(\left\{\overline{y}_{YEC}-\overline{Y}\right\}^2\right)\right]\\ &= \mathrm{E}\left\{\mathrm{E}_m\left[\left(\overline{y}_{YEC}-\overline{Y}\right)^2\right]\right\} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{E}_m\left[\overline{y}_{YEC}-\overline{Y}\right]^2, \qquad (5.20)\end{aligned}$$

which follows since there are k possible samples. Now, by using the model in (5.1), we obtain

$$\begin{aligned}\overline{y}_{LSS} &= \frac{1}{n}\sum_{j=1}^{n} y_{\{i+(j-1)k\}}\\ &= \frac{1}{n}\left[na+nbi+bk\sum_{j=1}^{n-1} j+\sum_{j=1}^{n} e_{\{i+(j-1)k\}}\right]\\ &= a+bi+\frac{bk\left(n-1\right)}{2}+\overline{e}_i = a+b\left[i+\frac{k\left(n-1\right)}{2}\right]+\overline{e}_i, \qquad (5.21)\end{aligned}$$

which follows since $\overline{e}_i = \sum\limits_{j=1}^{n} e_{\{i+(j-1)k\}}/n$. Moreover, by using (5.3) and (5.21), it follows that

$$\begin{aligned}\overline{y}_{LSS}-\overline{Y} &= a+b\left[i+\frac{k\left(n-1\right)}{2}\right]+\overline{e}_i-a-\frac{b\left(N+1\right)}{2}-\overline{\overline{e}}\\ &= b\left[i+\frac{\left(N-k\right)}{2}-\frac{\left(N+1\right)}{2}\right]+\overline{e}_i-\overline{\overline{e}} = b\left[\frac{2i-k-1}{2}\right]+\overline{e}_i-\overline{\overline{e}}. \qquad (5.22)\end{aligned}$$

In addition, by applying the model in (5.1), we obtain

$$\begin{aligned}y_i-y_{i+(n-1)k} &= a+bi+e_i-\left[a+b\left\{i+\left(n-1\right)k\right\}+e_{i+(n-1)k}\right]\\ &= e_i-b\left(n-1\right)k-e_{i+(n-1)k}. \qquad (5.23)\end{aligned}$$

If we use (5.13), (5.22) and (5.23), then it follows that

$$\begin{aligned}\mathrm{E}_m\left[\left(\overline{y}_{YEC}-\overline{Y}\right)^2\right] &= \mathrm{E}_m\left[\overline{y}_{LSS}+\frac{\left(2i-k-1\right)}{2\left(n-1\right)k}\left(y_i-y_{i+(n-1)k}\right)-\overline{Y}\right]^2\\ &= \mathrm{E}_m\left[b\left(\frac{2i-k-1}{2}\right)+\overline{e}_i-\overline{\overline{e}}\right.\\ &\qquad\left.+\frac{\left(2i-k-1\right)}{2\left(n-1\right)k}\left(e_i-b\left(n-1\right)k-e_{i+(n-1)k}\right)\right]^2\\ &= \mathrm{E}_m\left[\overline{e}_i-\overline{\overline{e}}+\frac{\left(2i-k-1\right)}{2\left(n-1\right)k}\left(e_i-e_{i+(n-1)k}\right)\right]^2. \qquad (5.24)\end{aligned}$$

By applying the conditions in (5.2), we obtain

$$\begin{aligned}
\mathrm{E}_m\left[\overline{e}_i\left(e_i-e_{i+(n-1)k}\right)\right] &= \mathrm{E}_m\left[\left(\frac{1}{n}\sum_{j=1}^{n}e_{\{i+(j-1)k\}}\right)\left(e_i-e_{i+(n-1)k}\right)\right]\\
&= \frac{1}{n}\left[\sum_{j=1}^{n}\mathrm{E}_m\left(e_{\{i+(j-1)k\}}e_i\right)-\sum_{j=1}^{n}\mathrm{E}_m\left(e_{\{i+(j-1)k\}}e_{i+(n-1)k}\right)\right]\\
&= \frac{\sigma^2-\sigma^2}{n}=0, \qquad (5.25)
\end{aligned}$$

where $e_{i+(j-1)k}=e_i$ and $e_{i+(j-1)k}=e_{i+(n-1)k}$ each occur once, for $j=1,...,n$. Likewise,

$$\begin{aligned}
\mathrm{E}_m\left[\overline{\overline{e}}\left(e_i-e_{i+(n-1)k}\right)\right] &= \mathrm{E}_m\left[\left(\frac{1}{N}\sum_{j=1}^{N}e_j\right)\left(e_i-e_{i+(n-1)k}\right)\right]\\
&= \frac{1}{N}\left[\sum_{j=1}^{N}\mathrm{E}_m\left(e_ie_j\right)-\sum_{j=1}^{N}\mathrm{E}_m\left(e_{i+(n-1)k}e_j\right)\right]\\
&= \frac{\sigma^2-\sigma^2}{N}=0, \qquad (5.26)
\end{aligned}$$

where $e_i=e_j$ and $e_{i+(n-1)k}=e_j$ each occur once, for $j=1,...,N$ and $i\in\{1,...,k\}$. Furthermore, by applying the conditions in (5.2) along with the assumption of $n\geq 2$, we obtain

$$\begin{aligned}
\mathrm{E}_m\left[\left(e_i-e_{i+(n-1)k}\right)^2\right] &= \mathrm{E}_m\left[e_i^2-2e_ie_{i+(n-1)k}+e_{i+(n-1)k}^2\right]\\
&= \mathrm{E}_m\left(e_i^2\right)-2\mathrm{E}_m\left(e_ie_{i+(n-1)k}\right)+\mathrm{E}_m\left(e_{i+(n-1)k}^2\right)\\
&= \sigma^2+\sigma^2=2\sigma^2. \qquad (5.27)
\end{aligned}$$

Expanding (5.24) and then substituting (5.6), (5.7), (5.8), (5.25), (5.26) and (5.27), results in

$$\mathrm{E}_m\left[\left(\overline{y}_{YEC}-\overline{Y}\right)^2\right]=\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{\sigma^2(2i-k-1)^2}{2(n-1)^2k^2}. \qquad (5.28)$$

Now, by substituting (5.28) into (5.20), it follows that

$$\begin{aligned}
\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{YEC}\right) &= \frac{1}{k}\sum_{i=1}^{k}\left[\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{\sigma^2(2i-k-1)^2}{2(n-1)^2k^2}\right]\\
&= \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{\sigma^2}{2(n-1)^2k^3}\sum_{i=1}^{k}(2i-k-1)^2. \qquad (5.29)
\end{aligned}$$

The summation term on the right hand side of (5.29) simplifies to

$$\sum_{i=1}^{k}(2i-k-1)^2=\frac{k\left(k^2-1\right)}{3}. \qquad (5.30)$$

Finally, substituting (5.30) into (5.29), results in

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{YEC}\right) = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) + \frac{\sigma^2\left(k^2-1\right)}{6(n-1)^2k^2}. \tag{5.31}$$

5.1.3 Error comparisons

If we compare (5.9) to (5.31), then we see that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{YEC}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$, if and only if $\sigma^2 < b^2(n-1)^2k^2/2$. Furthermore, by comparing (5.11) to (5.31), we note that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{YEC}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{STR})$, if and only if $\sigma^2 < b^2(n-1)^2k^2/2n$. Thus, by using (5.12), we conclude that $\overline{y}_{YEC}$ is expected to be subject to the least amount of error, when compared to $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$, if and only if $\sigma^2 < b^2(n-1)^2k^2/2n$.

5.2 Centered Systematic Sampling

5.2.1 Methodology

CESS was first discussed by Madow (1953) and involves selecting the centrally located sample from all the possible linear systematic samples, thus resulting in no randomization. If k is odd, then the sample units chosen will be those elements with population unit numbers given by

$$\left[(2j-1)\,k+1\right]/2, \qquad \text{for } j = 1, ..., n. \tag{5.32}$$

The sample is thus selected by applying LSS with a predetermined start of $i = (k+1)/2$. If k is even, then then the population unit numbers of the sampling units are given by either

$$(2j-1)\,k/2, \qquad \text{for } j = 1, ..., n, \tag{5.33}$$

or

$$\left[(2j-1)\,k+2\right]/2, \qquad \text{for } j = 1, ..., n, \tag{5.34}$$

with probability 1/2. The sample is thus selected by applying LSS, where the predetermined start is either $i = k/2$ or $i = (k+2)/2$, with probability 1/2.

The corresponding estimator of $\overline{Y}$(i.e. $\overline{y}_{CESS}$) is obtained by using (5.32), (5.33) and (5.34), such that

$$\overline{y}_{CESS} = \begin{cases} n^{-1}\sum\limits_{j=1}^{n} y_{[(2j-1)k+1]/2} & \text{if } k \text{ is odd,} \\ n^{-1}\sum\limits_{j=1}^{n} y_{(2j-1)k/2} \text{ or } n^{-1}\sum\limits_{j=1}^{n} y_{[(2j-1)k+2]/2} & \text{if } k \text{ is even.} \end{cases} \tag{5.35}$$

It should be noted that for CESS, some population units have no chance of being selected for the sample and thus $\overline{y}_{CESS}$ is subject to bias (Murthy 1967). However, $\overline{y}_{CESS}$ is unbiased for the model in (4.5), as shown in the next section.

5.2.2 Expected mean square error

If k is odd, then an estimator of (4.6) is obtained by using (4.5) and (5.35), such that

$$\overline{y}_{CSS(k\ \mathrm{odd})} = a + \frac{b}{2n}\left[k\sum_{j=1}^{n}(2j-1)+n\right] = a + \frac{b}{2}\left[k\left(n+1\right)-k+1\right] = a + \frac{b\left(N+1\right)}{2}.$$

Similarly, if k is even, then $\overline{y}_{k/2} = a + bN/2$ and $\overline{y}_{(k+2)/2} = a + b(N+2)/2$. By selecting $\overline{y}_{k/2}$ or $\overline{y}_{(k+2)/2}$ with probability 1/2, we can then show that $\overline{y}_{CSS(k\ \mathrm{even})}$ is an unbiased estimator of (4.6), i.e. $\mathrm{E}\left[\overline{y}_{CESS(k\ \mathrm{even})}\right] = \left[\overline{y}_{k/2} + \overline{y}_{(k+2)/2}\right]/2 = a + b(N+1)/2$. We may thus use (5.4) to obtain the expected MSE of $\overline{y}_{CESS}$, since there are equal weights being applied to all the sampling units and $\overline{y}_{CESS}$ is unbiased for the perfect linear trend model. The variance of $\overline{y}_{(k+1)/2}$ is thus given as

$$\mathrm{Var}(\overline{y}_{(k+1)/2}) = 0. \tag{5.36}$$

Furthermore, the variances of $\overline{y}_{k/2}$ and $\overline{y}_{(k+2)/2}$ are respectively given as

$$\begin{aligned}\mathrm{Var}(\overline{y}_{k/2}) &= \mathrm{E}\left[\left(a+\frac{bN}{2}-\left\{a+\frac{b\left(N+1\right)}{2}\right\}\right)^2\right]\\ &= \mathrm{E}\left[\left(-\frac{b}{2}\right)^2\right] = \frac{b^2}{4}\end{aligned} \tag{5.37}$$

and

$$\begin{aligned}\mathrm{Var}(\overline{y}_{(k+2)/2}) &= \mathrm{E}\left[\left(a+\frac{b\left(N+2\right)}{2}-\left\{a+\frac{b\left(N+1\right)}{2}\right\}\right)^2\right]\\ &= \mathrm{E}\left[\left(\frac{b}{2}\right)^2\right] = \frac{b^2}{4}.\end{aligned} \tag{5.38}$$

If we assume the model in (5.1), then by substituting (5.36) into (5.4) when k is odd and either (5.37) or (5.38) into (5.4) when k is even, we obtain

$$\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) = \begin{cases}\sigma^2\left(1/n-1/N\right) & \text{if } k \text{ is odd,}\\ \sigma^2\left(1/n-1/N\right)+b^2/4 & \text{if } k \text{ is even.}\end{cases} \tag{5.39}$$

We thus obtain a complete removal of the trend component in (5.39), when k is odd.

5.2.3 Error comparisons

By comparing (5.9) to (5.39), we conclude that:

(i) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$, when k is odd;

(ii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$, when $k \in \{4, 6, ...\}$;

(iii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$, when $k = 2$.

Furthermore, by comparing (5.11) to (5.39), we get $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{STR})$ when k is odd and for the case when k is even, provided that $k^2 - 1 > 3n$. Hence, by using (5.12), we thus conclude that $\overline{y}_{CESS}$ is expected to be subject to the least amount of error, when compared to $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$ when k is odd and for the case when k is even, if and only if $k^2 - 1 - 3n$ is non-zero and positive. Finally, if we compare (5.31) to (5.39), it follows that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{YEC})$ when k is odd and for the case when k is even, provided that $\sigma^2 > 3b^2(n-1)^2k^2/2(k^2-1)$.

5.3 Balanced Systematic Sampling

5.3.1 Methodology

The methodology of this design was first introduced by Sethi (1965) and later termed as BSS by Murthy (1967, p.165). One way of viewing a balanced arrangement is that it reverses the order, with respect to the population unit numbers, of every alternative set of k population units. LSS is then applied to this balanced arrangement to select the balanced systematic sample.

Instead of applying LSS to a balanced arrangement, one can alternatively use the following equivalent methodology, given by Murthy (1967, p.165). If n is even, then a balanced systematic sample is chosen by selecting those elements with population unit numbers given by

$$i + 2jk, \qquad 2(j+1)k - i + 1, \qquad \text{for } j = 0, ..., (n-2)/2. \tag{5.40}$$

The population is thus divided into $n/2$ groups, each consisting of $2k$ population units. A random start is then selected from the interval $[1, k]$, before selecting a pair of sampling units from each group according to the random start, such that each unit is paired with a unit that occurs at an equivalent position at the other end of the respective group, i.e.

the first sampling unit in the group is paired with the last unit, the second sampling unit in the group is paired with the second last unit, and so forth. This results in optimum sampling within each of the $n/2$ groups, when n is even (Murthy 1967). Alternatively, if n is odd, then a balanced systematic sample is chosen by selecting those elements with population unit numbers given by

$$i + 2jk, \qquad 2(j+1)k - i + 1, \qquad i + (n-1)k, \qquad \text{for } j = 0, ..., (n-3)/2. \tag{5.41}$$

For this situation, the first $N - k$ population units are divided into $(n-1)/2$ groups, each consisting of $2k$ population units. A random start is then selected from the interval $[1, k]$, before selecting two units from each group using the pairing technique above. The nth sampling unit, which corresponds to the random start, is then selected from a group which consists of the last k population units. Conversely, this does not result in optimum sampling, since we obtain an extra sampling unit after selecting the $(n-1)$ paired units using optimum sampling, where the variate value of this extra sampling unit will on average give the sample an uneven weighting, for the population under consideration.

The corresponding estimator of $\overline{Y}$ (i.e. $\overline{y}_{BSS}$) with random start i, for $i \in \{1, ..., k\}$, is thus obtained by using (5.40) and (5.41), such that

$$\overline{y}_{BSS} = \begin{cases} n^{-1} \sum_{j=0}^{(n-2)/2} (y_{i+2jk} + y_{2(j+1)k-i+1}) & \text{if } n \text{ is even,} \\ n^{-1} \left[\sum_{j=0}^{(n-3)/2} (y_{i+2jk} + y_{2(j+1)k-i+1}) + y_{i+(n-1)k} \right] & \text{if } n \text{ is odd.} \end{cases} \tag{5.42}$$

This estimator is design unbiased since each population unit has an equal chance, $1/k$, of being selected.

5.3.2 Expected mean square error

If n is even, then an estimator of (4.6) is obtained by using (4.5) and (5.42), i.e.

$$\begin{aligned} \overline{y}_{BSS(n \text{ even})} &= \frac{1}{n} \left[na + b \left(\frac{ni}{2} + 2k \sum_{i=1}^{(n/2-1)} i + 2k \sum_{i=1}^{n/2} i - \frac{ni}{2} + \frac{n}{2} \right) \right] \\ &= a + \frac{b}{n} \left[\frac{nk}{2} \left(\frac{n}{2} - 1 \right) + \frac{nk}{2} \left(\frac{n}{2} + 1 \right) + \frac{n}{2} \right] \\ &= a + b \left[\frac{nk}{4} - \frac{k}{2} + \frac{nk}{4} + \frac{k}{2} + \frac{1}{2} \right] \\ &= a + b \left[\frac{N}{2} + \frac{1}{2} \right] = a + \frac{b\,(N+1)}{2}. \end{aligned} \tag{5.43}$$

Similarly, when n is odd, we find that

$$\overline{y}_{BSS(n \text{ odd})} = a + \frac{b}{2} \left[N + 1 + \frac{(2i - k - 1)}{n} \right]. \tag{5.44}$$

The variance of $\overline{y}_{BSS(n\text{ even})}$ is obtained by using (4.6) and (5.43), i.e.

$$\mathrm{Var}(\overline{y}_{BSS(n\text{ even})}) = 0. \tag{5.45}$$

Furthermore, we obtain the variance of $\overline{y}_{BSS(n\text{ odd})}$ by using (4.6), (5.44) and (5.30), such that

$$\begin{aligned}
\mathrm{Var}(\overline{y}_{BSS(n\text{ odd})}) &= \mathrm{E}\left[\left(a+\frac{b}{2}\left\{N+1+\frac{(2i-k-1)}{n}\right\}-\left\{a+\frac{b}{2}(N+1)\right\}\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{b}{2}\left\{\frac{2i-k-1}{n}\right\}\right)^2\right] \\
&= \frac{b^2}{4n^2k}\sum_{i=1}^{k}(2i-k-1)^2 = \frac{b^2\left(k^2-1\right)}{12n^2}.
\end{aligned} \tag{5.46}$$

We next use (5.4) to obtain the expected MSE of $\overline{y}_{BSS}$, since there are equal weights being applied to all the sampling units and $\overline{y}_{BSS}$ is design unbiased. Consequently, if we assume the model in (5.1), then by substituting (5.45) into (5.4) when n is even and (5.46) when n is odd, we obtain

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{BSS}\right) = \begin{cases} \sigma^2\left(1/n-1/N\right) & \text{if } n \text{ is even,} \\ \sigma^2\left(1/n-1/N\right)+b^2(k^2-1)/12n^2 & \text{if } n \text{ is odd.} \end{cases} \tag{5.47}$$

We thus obtain a complete removal of the trend component in (5.47), when n is even.

5.3.3 Error comparisons

By comparing (5.9) to (5.47), we see that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$ when $n \geq 2$. Moreover, by comparing (5.11) to (5.47), we note that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{STR})$ when $n \geq 2$. By using (5.12), we thus conclude that $\overline{y}_{BSS}$ is expected to be subject to the least amount of error, when compared to $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$, if $n \geq 2$ (equality occurs when $n = 1$). Furthermore, if we compare (5.31) to (5.47), then we see that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{YEC})$ when n is even and for the case when n is odd, if and only if $\sigma^2 > b^2(n-1)^2k^2/2n^2$. Finally, by comparing (5.39) to (5.47), we note that

(i) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is even and k is odd;

(ii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) > \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is odd and k is odd;

(iii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is even and k is even;

(iv) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is odd and k is even, if and only if $(k^2-1)/n^2 < 3$.

5.4 Modified Systematic Sampling

5.4.1 Methodology

The method of MSS was first considered by Singh, Jindal & Garg (1968). It creates a sample by selecting pairs of sampling units that are equidistant from each end of the population. A modified arrangement is one that reverses the order of the second half of population units, i.e. the first $n/2$ groups of k population units are monotonically increasing, with respect to their population unit numbers, and the last $n/2$ groups of k population units are monotonically decreasing, or vice versa. If n is odd, then the number of groups which are monotonically increasing is not equal the number of groups which are monotonically decreasing. To obtain a modified arrangement, we thus leave the middle group of k units in the order of the initial arrangement, e.g. if we have five groups of k units each (i.e. $n = 5$ is odd), then a modified arrangement is applied by only reversing the order of the population units in the last two groups, such that the last population unit in the third group (U_{3k}) is followed by the last population unit in the fifth group (U_{5k}) and so forth, until the final population unit is the first unit in the fourth group (U_{3k+1}). LSS is then applied to this modified arrangement to select the modified systematic sample. It should be noted that the modified systematic sample is spread evenly over the population, except in the middle.

Instead of applying LSS to a modified arrangement, one can alternatively use the following equivalent methodology, given by Singh et al. (1968). If n is even, then a modified systematic sample is chosen by selecting those elements with population unit numbers given by

$$i + jk, \qquad N - jk - i + 1, \qquad \text{for } j = 0, ..., (n-2)/2. \tag{5.48}$$

The population is thus divided into two groups, such that the first $N/2$ population units belong to the first group and the last $N/2$ population units belong to the second group. We then select $n/2$ sampling units from the first group using LSS and pair each of these units with a corresponding unit from the second group, such that the units paired occur at opposite ends of each group, i.e. the first unit in the first group is paired with the last unit in the second group, the second unit in the first group is paired with the second last unit in the second group, and so forth. This results in optimum sampling for the case when n is even (Sethi 1965). Alternatively, if n is odd, then a modified systematic sample is chosen by selecting those elements with population unit numbers given by

$$i + jk, \qquad N - jk - i + 1, \qquad i + (n-1)k/2, \qquad \text{for } j = 0, ..., (n-3)/2. \tag{5.49}$$

For this situation, the first $(N-k)/2$ population units belong to the first group, the last $(N-k)/2$ population units belong to the second group and the middle set of k population units belong to the third group. We then select $(n-1)/2$ sampling units from the first group using LSS and pair each of these units with a corresponding unit from the second group, as defined for the case where n is even. The nth sampling unit, which corresponds to the random start, is then selected from the third group. Just as in the case of n being odd for BSS, we do not achieve optimum sampling, since we obtain an extra sampling unit after selecting the $(n-1)$ sampling units using optimum sampling.

The corresponding estimator of $\overline{Y}$ (i.e. $\overline{y}_{MSS}$) with random start i, for $i \in \{1, ..., k\}$, is thus obtained by using (5.48) and (5.49), such that

$$\overline{y}_{MSS} = \begin{cases} n^{-1} \sum_{j=0}^{(n-2)/2} (y_{i+jk} + y_{N-jk-i+1}) & \text{if } n \text{ is even,} \\ n^{-1} [\sum_{j=0}^{(n-3)/2} (y_{i+jk} + y_{N-jk-i+1}) + y_{i+(n-1)k/2}] & \text{if } n \text{ is odd.} \end{cases} \tag{5.50}$$

This estimator is design unbiased since each population unit has an equal chance, $1/k$, of being selected.

5.4.2 Expected mean square error

If n is even, then an estimator of (4.6) is obtained by using (4.5) and (5.50), such that

$$\begin{aligned} \overline{y}_{MSS(n\text{even})} &= \frac{1}{n} \left[na + b \left\{ \frac{ni}{2} + k \sum_{i=1}^{(n-2)/2} i + \frac{Nn}{2} - \frac{ni}{2} - k \sum_{i=1}^{(n-2)/2} i + \frac{n}{2} \right\} \right] \\ &= a + \frac{b}{n} \left[\frac{n(N+1)}{2} \right] = a + \frac{b(N+1)}{2}. \end{aligned} \tag{5.51}$$

Likewise, when n is odd, we find that $\overline{y}_{MSS(n \text{ odd})} = a + b(N+1+(2i-k-1)/n)/2$. Now, in MSS, we obtain the same estimator as that for BSS, when considering the model in (4.5), i.e. $\overline{y}_{MSS(n \text{ even})} = \overline{y}_{BSS(n \text{ even})}$ and $\overline{y}_{MSS(n \text{ odd})} = \overline{y}_{BSS(n \text{ odd})}$. By noting that equal weights are being applied to all the sampling units for MSS, we then use (5.47), such that

$$\mathrm{E}_m\mathrm{MSE}\left(\overline{y}_{MSS}\right) = \begin{cases} \sigma^2 \left(1/n - 1/N\right) & \text{if } n \text{ is even,} \\ \sigma^2 \left(1/n - 1/N\right) + b^2(k^2-1)/12n^2 & \text{if } n \text{ is odd.} \end{cases} \tag{5.52}$$

There is thus a complete removal of the trend component in (5.52), when n is even. We can now apply the error comparisons, given in Section 5.3.3, for MSS.

5.5 Balanced Modified Systematic Sampling

5.5.1 Methodology

The author next proposes a new design, which uses MSS in conjunction with a balanced arrangement and is thus termed as BMSS. It creates a sample by applying MSS on a balanced arrangement. A balanced modified arrangement is achieved by reversing the order, with respect to the population unit numbers, of every alternative set of k population units, before reversing the order of a group of population units that occur at the end of the population, i.e. we are using a modified arrangement on the balanced arrangement. For the case where n is even, we apply a balanced arrangement, before reversing the order of the last $n/2$ sets of k population units. For the case where n is odd, we apply a balanced arrangement, before reversing the order of the last $(n-1)/2$ sets of k population units. LSS is then applied to this balanced modified arrangement to obtain a balanced modified systematic sample. It should be noted that BMSS reduces to LSS when $n = 2$. We will thus assume $n > 2$ for the remainder of this section.

Instead of applying LSS to a balanced modified arrangement, one can alternatively use the following equivalent methodology. The population unit numbers of the sampling units, when selecting a balanced modified systematic sample, is given as follows:

(A) if $n/2$ is an even integer, then

$$i + 2jk, \qquad 2(j+1)k - i + 1, \qquad \text{for } j = 0, ..., (n-4)/4, \tag{5.53}$$

and

$$N + i - k - 2jk, \qquad N - i - k - 2jk + 1, \qquad \text{for } j = 0, ..., (n-4)/4; \tag{5.54}$$

(B) if $n/2$ is an odd integer, then

$$i + 2jk, \qquad N + i - k - 2jk, \qquad \text{for } j = 0, ..., (n-2)/4, \tag{5.55}$$

and

$$2(j+1)k - i + 1, \qquad N - i - k - 2jk + 1, \qquad \text{for } j = 0, ..., (n-6)/4; \tag{5.56}$$

(C) if $n = 3$, then

$$i + 2jk, \quad 2(j+1)k - i + 1, \quad N - i - 2jk + 1, \quad \text{for } j = 0, ..., (n-3)/4; \tag{5.57}$$

(D) if $n \neq 3$ and $(n+1)/2$ is an even integer, then

$$i + 2jk, \quad 2(j+1)k - i + 1, \quad N - i - 2jk + 1, \quad \text{for } j = 0, ..., (n-3)/4, \tag{5.58}$$

and

$$N + i - 2(j+1)k, \qquad \text{for } j = 0, ..., (n-7)/4; \tag{5.59}$$

(E) if $(n+1)/2$ is an odd integer, then

$$i + 2jk, \quad 2(j+1)k - i + 1, \quad i + (n-1)k/2, \quad \text{for } j = 0, ..., (n-5)/4, \tag{5.60}$$

and

$$N - i - 2jk + 1, \qquad N + i - 2(j+1)k, \qquad \text{for } j = 0, ..., (n-5)/4. \tag{5.61}$$

The corresponding estimator of $\overline{Y}$ (i.e. $\overline{y}_{BMSS}$) with random start i, for $i \in \{1, ..., k\}$, is thus the average of the sample variate values chosen, using the respective population unit numbers for the various cases given above. The resulting estimator is design unbiased since each population unit has an equal chance, $1/k$, of being selected.

5.5.2 Expected mean square error

If $n/2$ is an even integer, then an estimator of (4.6) is obtained by using (4.5), (5.53) and (5.54), such that

$$\overline{y}_{BMSS(A)} = a + b(N+1)/2 = \overline{Y},$$

which results in

$$\text{Var}\left[\overline{y}_{BMSS(A)}\right] = 0. \tag{5.62}$$

Similarly, if $n/2$ is an odd integer, then by using (4.5), (5.55) and (5.56), we obtain

$$\overline{y}_{BMSS(B)} = a + b[N + 1 + 2(2i - k - 1)/n]/2.$$

Thus, by applying (5.46), we obtain

$$\begin{aligned}\text{Var}\left[\overline{y}_{BMSS(B)}\right] &= \text{E}\left[\left\{a + \frac{b}{2}\left(N + 1 + \frac{2\,(2i-k-1)}{n}\right) - \left(a + \frac{b\,(N+1)}{2}\right)\right\}^2\right] \\ &= 4\text{E}\left[\left\{\frac{b}{2}\left(\frac{2i-k-1}{n}\right)\right\}^2\right] = \frac{b^2\left(k^2-1\right)}{3n^2}. \end{aligned} \tag{5.63}$$

For the case when $n = 3$, we find that

$$\overline{y}_{BMSS(C)} = a + \frac{b}{2}\left[\frac{10k - 2i + 4}{3}\right],$$

which is obtained by using (4.5) and (5.57). The corresponding variance is thus given as

$$\begin{aligned}\text{Var}\left[\overline{y}_{BMSS(C)}\right] &= \text{E}\left[\left\{a + \frac{b}{2}\left(\frac{10k - 2i + 4}{3}\right) - \left(a + \frac{b\,[3k+1]}{2}\right)\right\}^2\right] \\ &= \text{E}\left[\left\{\frac{b}{2}\left(\frac{10k - 2i + 4 - 9k - 3}{3}\right)\right\}^2\right] \\ &= \text{E}\left[\left\{\frac{b}{2}\left(\frac{2i - k - 1}{n}\right)\right\}^2\right] = \frac{b^2\left(k^2 - 1\right)}{12n^2}, \qquad (5.64)\end{aligned}$$

which follows if we apply (5.46). By using (4.5), (5.58) and (5.59), for the case when $n \neq 3$ and $(n+1)/2$ is an even integer, we obtain

$$\overline{y}_{BMSS(D)} = a + \frac{b}{2}\left[N + 1 - \frac{(2i - k - 1)}{n}\right]$$

such that

$$\begin{aligned}\text{Var}\left[\overline{y}_{BMSS(D)}\right] &= \text{E}\left[\left\{a + \frac{b}{2}\left(N + 1 - \frac{(2i - k - 1)}{n}\right) - \left(a + \frac{b\,(N+1)}{2}\right)\right\}^2\right] \\ &= \text{E}\left[\left\{\frac{b}{2}\left(\frac{2i - k - 1}{n}\right)\right\}^2\right] = \frac{b^2\left(k^2 - 1\right)}{12n^2}, \qquad (5.65)\end{aligned}$$

which follows if we apply (5.46). Finally, if $(n+1)/2$ is an odd integer, then

$$\overline{y}_{BMSS(E)} = a + \frac{b}{2}\left[N + 1 + \frac{(2i - k - 1)}{n}\right],$$

which is obtained by using (4.5), (5.60) and (5.61). The corresponding variance is found by noting the equivalence of $\overline{y}_{BMSS(E)}$ to $\overline{y}_{BSS(n\text{ odd})}$ in (5.44), such that we then use (5.46), i.e.

$$\text{Var}\left[\overline{y}_{BMSS(E)}\right] = \frac{b^2\left(k^2 - 1\right)}{12n^2}. \qquad (5.66)$$

We next use (5.4) to obtain the expected MSE of $\overline{y}_{BMSS}$, since there are equal weights being applied to all the sampling units and $\overline{y}_{BMSS}$ is design unbiased. Consequently, if we assume the model in (5.1), then by substituting (5.62) to (5.66) into (5.4), for the respective cases, we obtain

$$\text{E}_m\text{MSE}\left(\overline{y}_{BMSS}\right) = \begin{cases} \sigma^2(1/n - 1/N) & \text{if } n/2 \text{ is an even integer;} \\ \sigma^2(1/n - 1/N) + b^2(k^2 - 1)/3n^2 & \text{if } n/2 \text{ is an odd integer;} \\ \sigma^2(1/n - 1/N) + b^2(k^2 - 1)/12n^2 & \text{if } n \text{ is odd.} \end{cases} \qquad (5.67)$$

We thus obtain a complete removal of the trend component when $n/2$ is an even integer.

5.5.3 Error comparisons

We will now compare $\overline{y}_{BMSS}$ to all the other estimators, by assuming that $n > 2$. By comparing (5.9) to (5.67), we see that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS})$ for all the cases. Likewise, we obtain the same result when we compare (5.11) to (5.67), i.e. $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{STR})$. Thus, by using (5.12), we conclude that $\overline{y}_{BMSS}$ is expected to be subject to the least amount of error, when compared to $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$, except when $n = 2$. If $n = 2$, then

$$\mathrm{E}_m\mathrm{MSE}(\overline{y}_{STR}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{LSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{SRSWOR}).$$

Moreover, by comparing (5.31) to (5.67), we see that $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{YEC})$, when:

(i) $n/2$ is an even integer;

(ii) $n/2$ is an odd integer, if and only if $\sigma^2 > 2b^2(n-1)^2k^2/n^2$;

(iii) n is odd, if and only if $\sigma^2 > b^2(n-1)^2k^2/2n^2$.

Also, by comparing (5.39) to (5.67), we note that:

(i) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when $n/2$ is an even integer and k is odd;

(ii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when $n/2$ is an even integer and k is even;

(iii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) > \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when $n/2$ is an odd integer and k is odd;

(iv) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when $n/2$ is an odd integer and k is even, if and only if $(k^2-1)/n^2 < 3/4$;

(v) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) > \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is odd and k is odd;

(vi) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) < \mathrm{E}_m\mathrm{MSE}(\overline{y}_{CESS})$, when n is odd and k is even, if and only if $(k^2-1)/n^2 < 3$.

Finally, the comparison of (5.47) to (5.67), results in:

(i) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS})$, when $n/2$ is an even integer;

(ii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) > \mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS})$, when $n/2$ is an odd integer;

(iii) $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{BMSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS})$, when n is odd.

The above comparison can then be used to compare $\overline{y}_{BMSS}$ to $\overline{y}_{MSS}$, since $\mathrm{E}_m\mathrm{MSE}(\overline{y}_{MSS}) = \mathrm{E}_m\mathrm{MSE}(\overline{y}_{BSS})$. We thus conclude that $\overline{y}_{BMSS}$ is expected to be subject to the same amount of error, when compared to both $\overline{y}_{BSS}$ and $\overline{y}_{MSS}$, except for the case when $n/2$ is an odd integer, which results in the former estimator being expected to be subject to more error than the latter estimators.

Table 5.1 provides a summary of all the estimators mentioned in this chapter, when sampling from a population that exhibits a rough linear trend, given by the model in (5.1). The second column of Table 5.1 shows us what restrictions are placed on each estimator. The third column is a measure of the linear trend component. All estimators, except for $\overline{y}_{YEC}$, which have a zero in the third column, have completely removed linear trend. Estimator $\overline{y}_{YEC}$ completely removes the linear trend component, but this consequently results in a larger corresponding error variance component, since different weights are being applied to the first and the last sampling units. The last column in Table 5.1 shows us which estimators are unbiased, when estimating (5.3). From Table 5.1, we conclude that the estimators which are subject to the least amount of error are $\overline{y}_{CSS(k \text{ odd})}$, $\overline{y}_{BSS(n \text{ even})}$, $\overline{y}_{MSS(n \text{ even})}$ and $\overline{y}_{BMSS(n/2 \text{ an even integer})}$, for populations that exhibit a rough linear trend.

Throughout this thesis, we have estimated $\overline{Y}$ for various populations and we have provided expressions for the variance of each of these estimates. All of these variance expressions require us to have full knowledge of the population. Since we cannot study an entire population (Chapter 1), we thus need to estimate these variance expressions. In the next chapter, we will discuss the problem that arises when estimating $\mathrm{Var}(\overline{y}_{LSS})$ and then provide some solutions to the problem.

Table 5.1: The removal of the trend component for populations that exhibit a rough linear trend

Estimator	Conditions	$\mathrm{E}_m\mathrm{MSE}(\hat{\overline{Y}}) - \sigma^2(1/n - 1/N)$	Trend Component	Accuracy when estimating $\overline{Y}$
$\overline{y}_{SRSWOR}$	none	$b^2(k-1)(N+1)/12$	Not Removed	Unbiased
$\overline{y}_{STR}$	none	$b^2(k^2-1)/12n$	Not Removed	Unbiased
$\overline{y}_{LSS}$	none	$b^2(k^2-1)/12$	Not Removed	Unbiased
$\overline{y}_{YEC}$	none	$\sigma^2(k^2-1)/[6(n-1)^2k^2]$	Removed	Biased
$\overline{y}_{CESS}$	k is odd	0	Removed	Biased
$\overline{y}_{CESS}$	k is even	$b^2/4$	Not Removed	Biased
$\overline{y}_{BSS}$	n is even	0	Removed	Unbiased
$\overline{y}_{BSS}$	n is odd	$b^2(k^2-1)/12n^2$	Not Removed	Unbiased
$\overline{y}_{MSS}$	n is even	0	Removed	Unbiased
$\overline{y}_{MSS}$	n is odd	$b^2(k^2-1)/12n^2$	Not Removed	Unbiased
$\overline{y}_{BMSS}$	$n/2$ is an even integer	0	Removed	Unbiased
$\overline{y}_{BMSS}$	$n/2$ is odd integer	$b^2(k^2-1)/3n^2$	Not Removed	Unbiased
$\overline{y}_{BMSS}$	n is odd	$b^2(k^2-1)/12n^2$	Not Removed	Unbiased

Chapter 6

ESTIMATION OF THE SAMPLING VARIANCE

If $\overline{y}$ is an unbiased estimator of $\overline{Y}$, then $\text{Var}(\overline{y}) = \text{E}(\overline{y}^2) - \overline{Y}^2$ can be unbiasedly estimated by $\overline{y}^2 - Est(\overline{Y}^2)$, where $Est(\overline{Y}^2)$ denotes an unbiased estimate of $\overline{Y}^2 = \sum_{i=1}^{N} y_i^2/N^2 + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j/N^2$ (Murthy 1967). Now, if we conduct LSS, then an unbiased estimate of $\sum_{i=1}^{N} y_i^2/N^2$ is given by $\sum_{j=1}^{n} y_{i+(j-1)k}^2/n^2k$, since there is an equal probability of each population unit being included in the sample. Unfortunately, an unbiased estimate of $\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j/N^2$ is unobtainable when conducting LSS with a single start, since certain pairs of population units have a zero probability of being included in the sample, i.e. $\pi_{ij} = 0$ for certain combination values of i and j. This results in it being impossible to obtain an unbiased estimate of $\text{Var}(\overline{y}_{LSS})$ from a single start.

In light of the above result, we will first construct estimators of $\text{Var}(\overline{y}_{LSS})$ and find the least biased estimator, for various population structures. Thereafter, we will examine some designs which result in an unbiased estimate of the sampling variance. The designs that will be discussed are *multiple-start linear systematic sampling* (MLSS), *partially systematic sampling* (PSS) and a new proposed design termed as *multiple-start balanced modified systematic sampling* (MBMSS). For simplicity reasons we will assume that $k = N/n$ is an integer; however, the usual variance estimation problem is also applicable for CSS.

6.1 Variance Estimation from a Single Systematic Sample

In this section we will study eight estimators of $\text{Var}(\overline{y}_{LSS})$ on various population structures. Each estimator is based on the linear systematic sample with random start i, where

$i \in \{1, ..., k\}$. We will first define the estimators, before comparing them amongst each other on various population structures, so as to find the most accurate estimator for each underlying population model. We will thus use the expected value of each estimator, the corresponding expected bias and the expected relative bias, as comparative measures. It should be noted that there are many variance estimators which can be constructed and utilized for LSS. However, we will only define eight estimators, which widely represent the different types of variance estimators that can be used for LSS. The theory presented for this section is given by Wolter (2007).

6.1.1 Eight estimators of the variance

The first estimator is defined by assuming that the population is in random order. From Section 4.2.1, we note that LSS is equivalent to SRSWOR for this situation and hence this estimator is given as

$$v_1 = s^2 \left(\frac{1}{n} - \frac{1}{N} \right) ,$$

where $s^2 = \sum_{j=1}^{n} (y_{ij} - \overline{y}_{LSS})^2/(n-1)$ is the sample variance. If a randomly ordered population is expected (i.e. $\text{Var}(\overline{y}_{LSS}) \approxeq \text{Var}(\overline{y}_{SRSWOR})$), then it is trivial that v_1 will approximately be an unbiased estimator of $\text{Var}(\overline{y}_{LSS})$. If LSS is more efficient than SRSWOR, such that $\text{Var}(\overline{y}_{LSS}) < \text{Var}(\overline{y}_{SRSWOR})$, then v_1 provides an overestimate of $\text{Var}(\overline{y}_{LSS})$. Conversely, if LSS is less efficient than SRSWOR, such that $\text{Var}(\overline{y}_{LSS}) > \text{Var}(\overline{y}_{SRSWOR})$, then v_1 provides an underestimate of $\text{Var}(\overline{y}_{LSS})$. It is common practice for a survey statistician to use v_1 as an estimate of $\text{Var}(\overline{y}_{LSS})$ and this can result in a badly biased estimate if the population exhibits some structure, other than random.

A second estimator is constructed by assuming that the systematic sample is a stratified random sample, where two population units are selected from each successive stratum, of size $2k$. This estimator, which is based on non-overlapping differences, is then given as

$$v_2 = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{n} \sum_{j=1}^{n/2} a_{i,2j}^2 ,$$

where $a_{ij} = \Delta y_{ij} = y_{ij} - y_{i,j-1}$, such that Δ denotes the first difference operator. The objective of the third estimator is to increase the degrees of freedom in v_2. This estimator, which is based on overlapping differences, is then given as

$$v_3 = \left(\frac{1}{n} - \frac{1}{N} \right) \frac{1}{2(n-1)} \sum_{j=2}^{n} a_{ij}^2 .$$

Other estimators, which are based on higher-order contrasts than a_{ij}, have been suggested in literature. Examples of this are given by the next three estimators:

$$v_4 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{6(n-2)} \sum_{j=3}^{n} b_{ij}^2 ,$$

$$v_5 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{3.5(n-4)} \sum_{j=5}^{n} c_{ij}^2 ,$$

and

$$v_6 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{1}{7.5(n-8)} \sum_{j=9}^{n} d_{ij}^2 ,$$

where

$$b_{ij} = \Delta a_{ij} = \Delta^2 y_{ij} = y_{ij} - 2y_{i,j-1} + y_{i,j-2},$$

$$c_{ij} = \frac{1}{2}\Delta^4 y_{ij} + \Delta^2 y_{i,j-1} = \frac{y_{ij}}{2} - y_{i,j-1} + y_{i,j-2} - y_{i,j-3} + \frac{y_{i,j-4}}{2}$$

and

$$d_{ij} = \frac{1}{2}\Delta^8 y_{ij} + 3\Delta^6 y_{i,j-1} + 5\Delta^4 y_{i,j-2} + 2\Delta^2 y_{i,j-3} = \frac{y_{ij}}{2} - y_{i,j-1} + - ... + \frac{y_{i,j-8}}{2},$$

respectively denotes the second difference, a linear combination of the second and fourth differences and a linear combination of the second, fourth, sixth and eighth differences of the sample data. The corresponding degrees of freedom are given by $6(n-2)$, $3.5(n-4)$ and $7.5(n-8)$, which respectively represents the product of the sum of squares of the coefficients in b_{ij}, c_{ij}, and d_{ij} and the number of contrasts for the summations in v_4, v_5 and v_6.

Another estimator, which was initially studied by Koop (1971), can be constructed by splitting the original linear systematic sample into sub-samples of equal size. By letting p and n/p be integers and by splitting the linear systematic sample into p sub-samples of size n/p, we then show that the αth systematic sub-sample mean is given by

$$\overline{y}_\alpha = \frac{p}{n} \sum_{j=1}^{n/p} y_{i,p(j-1)+\alpha}, \qquad \text{for } \alpha = 1, ..., p.$$

An estimator of $\text{Var}(\overline{y}_{LSS})$ is thus given by

$$v_7 = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{n}{p(p-1)} \sum_{\alpha=1}^{p} (\overline{y}_\alpha - \overline{y}_{LSS})^2.$$

Koop examined this estimator when $p = 2$, i.e. splitting the original linear systematic sample into half. He obtained expressions for the bias of v_7, relative to $\text{Var}(\overline{y}_{LSS})$, in terms of the ICC.

The final type of estimator is constructed by making assumptions on the correlation between the population units. Examples of such assumptions were initially examined by Cochran (1946) and later by Osborne (1942) and Matérn (1947) for forestry and land use/cover area frame surveys. These authors assumed a super-population model (see Section 4.2.4) from which correlation arises and a further assumption that the correlation between two population units, which are k units apart, is given as $\rho_k = \exp(-\lambda k)$, where λ is a constant. Accordingly, an estimator of $\text{Var}(\overline{y}_{LSS})$ is given as

$$v_8 = \begin{cases} \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \left[1 + 2/\ln(\hat{\rho}_k) + 2\hat{\rho}_k/(1-\hat{\rho}_k)\right] & \text{if } \hat{\rho}_k > 0, \\ \left(\frac{1}{n} - \frac{1}{N}\right) s^2 & \text{if } \hat{\rho}_k \leq 0. \end{cases}$$

where

$$\hat{\rho}_k = \frac{1}{(n-1)s^2} \sum_{j=2}^{n} \left(y_{ij} - \overline{y}_{LSS}\right)\left(y_{i,j-1} - \overline{y}_{LSS}\right)$$

is an estimate of ρ_k.

6.1.2 Theoretical properties of the eight estimators

A comparative measure that will used in this section is the expected relative bias of estimator v_α, which is defined as

$$\text{R}_m(v_\alpha) = \text{B}_m(v_\alpha)/\text{E}_m\left[\text{Var}(\overline{y}_{LSS})\right], \qquad \text{for } \alpha = 1, ..., 8,$$

where $\text{B}_m(v_\alpha) = \text{E}_m\left[\text{E}(v_\alpha)\right] - \text{E}_m\left[\text{Var}(\overline{y}_{LSS})\right]$ is the expected bias of estimator v_α.

A simple mathematical model, in which the variate values of a population consists of a trend and a random error component, is given as

$$y_{ij} = \mu_{ij} + e_{ij}, \qquad \text{for } i = 1, ..., k, \text{ and } j = 1, ..., n, \qquad (6.1)$$

where the μ_{ij}'s denotes fixed constants and the e_{ij}'s denotes the random errors. We now will use the model above to represent the various population structures.

Random model

A randomly ordered population model can be represented by

$$\mu_{ij} = \mu, \qquad (6.2)$$

where the e_{ij}'s are given as iid $N(0, \sigma^2)$ random variables. From the definition of the model, we assume that there is no correlation present, so that

$$\mathrm{E}_m\left[\mathrm{Var}(\overline{y}_{LSS})\right] = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

(refer to Cochran (1946)). By using (6.1) and (6.2), we obtain

$$\begin{aligned}
\mathrm{E}\left(s^2\right) &= \mathrm{E}\left[\frac{1}{n-1}\sum_{j=1}^{n}\left(y_{ij} - \overline{y}_{LSS}\right)^2\right] \\
&= \frac{1}{n-1}\sum_{j=1}^{n}\mathrm{E}\left(e_{ij}^2 - 2e_{ij}\overline{e}_i + \overline{e}_i^2\right) = \frac{1}{n-1}\sum_{j=1}^{n}\left(\sigma^2 - \frac{\sigma^2}{n}\right) = \sigma^2, \qquad (6.3)
\end{aligned}$$

which follows since

$$\sum_{j=1}^{n}\mathrm{E}\left(e_{ij}\overline{e}_i\right) = \sum_{j=1}^{n}\mathrm{E}\left[e_{ij}\left(e_{i1} + \ldots + e_{ij} + \ldots + e_{in}\right)/n\right] = \sigma^2/n$$

and $\mathrm{E}\left(\overline{e}_i^2\right) = \sigma^2/n$. Thus, by applying (6.3), we conclude that v_1 is expected to be an unbiased estimator of $\mathrm{Var}(\overline{y}_{LSS})$, i.e.

$$\mathrm{E}_m\left[\mathrm{E}\left(v_1\right)\right] = \mathrm{E}_m\left[\mathrm{E}\left(\frac{s^2}{n}\left\{\frac{N-n}{N}\right\}\right)\right] = \mathrm{E}_m\left[\frac{\sigma^2}{n}\left(\frac{N-n}{N}\right)\right] = \mathrm{E}_m\left[\mathrm{Var}\left(\overline{y}_{LSS}\right)\right].$$

Furthermore, we can also deduce that v_2 is expected to be an unbiased estimator of $\mathrm{Var}(\overline{y}_{LSS})$, since

$$\begin{aligned}
\mathrm{E}_m\left[\mathrm{E}\left(v_2\right)\right] &= \mathrm{E}_m\left[\mathrm{E}\left\{\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n}\sum_{j=1}^{n/2}a_{i,2j}^2\right\}\right] \\
&= \mathrm{E}_m\left[\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n}\sum_{j=1}^{n/2}\mathrm{E}(y_{i,2j} - y_{i,2j-1})^2\right] \\
&= \mathrm{E}_m\left[\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n}\sum_{j=1}^{n/2}\mathrm{E}\left(e_{i,2j}^2 - 2e_{i,2j}e_{i,2j-1} + e_{i,2j-1}^2\right)\right] \\
&= \mathrm{E}_m\left[\left(\frac{1}{n} - \frac{1}{N}\right)\frac{1}{n}\sum_{j=1}^{n/2}2\sigma^2\right] = \mathrm{E}_m\left[\left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2\right] = \mathrm{E}_m\left[\mathrm{Var}\left(\overline{y}_{LSS}\right)\right].
\end{aligned}$$

Likewise, we can then show that estimators v_3 to v_7 are expected to be unbiased estimates of $\mathrm{Var}(\overline{y}_{LSS})$. Wolter (2007) states that an exact expression for $\mathrm{B}_m(v_8)$ is unobtainable without making some assumptions on the distribution; however, we can expect $\mathrm{B}_m(v_8) \approxeq 0$.

Linear trend model

A population that exhibits linear trend is given by the model

$$\mu_{ij} = a + b\left[i + (j-1)k\right], \qquad (6.4)$$

where a and b denote fixed constants and the e_{ij}'s are given as iid $N(0,\sigma^2)$ random variables. The expected variance of $\overline{y}_{LSS}$, under this model, is given as

$$\mathrm{E}_m\left[\mathrm{Var}\left(\overline{y}_{LSS}\right)\right] = \sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{b^2\left(k^2-1\right)}{12}$$

(refer to (5.9)). An expression for the expected value of v_1 is given by

$$\begin{aligned}\mathrm{E}_m\left[\mathrm{E}\left(v_1\right)\right] &= \mathrm{E}_m\left[\mathrm{E}\left\{\left(\frac{1}{n}-\frac{1}{N}\right)s^2\right\}\right]\\ &= \mathrm{E}_m\left[\left(\frac{1}{n}-\frac{1}{N}\right)\frac{1}{n-1}\sum_{j=1}^{n}\mathrm{E}(y_{ij}-\overline{y}_{LSS})^2\right].\end{aligned} \tag{6.5}$$

Now, by using (5.21) and (6.4), it follows that

$$\begin{aligned}\sum_{j=1}^{n}\mathrm{E}(y_{ij}-\overline{y}_{LSS})^2 &= \sum_{j=1}^{n}\mathrm{E}\left[a+b\left\{i+(j-1)k\right\}+e_{ij}-\left\{a+b\left[i+\frac{k(n-1)}{2}\right]+\overline{e}_i\right\}\right]^2\\ &= \sum_{j=1}^{n}\mathrm{E}\left[bk\left\{\left(j-1-\frac{(n-1)}{2}\right)+\left(e_{ij}-\overline{e}_i\right)\right\}\right]^2\\ &= \sum_{j=1}^{n}\left[b^2k^2\left\{(j-1)-\frac{(n-1)}{2}\right\}^2+\mathrm{E}\left(e_{i,j}^2-2e_{ij}\overline{e}_i+\overline{e}_i^2\right)\right]\\ &= b^2k^2\left[\frac{n\left(n-1\right)\left(2n-1\right)}{6}-\frac{n(n-1)^2}{2}+\frac{n(n-1)^2}{4}\right]+\sum_{j=1}^{n}\left(\sigma^2-\frac{\sigma^2}{n}\right)\\ &= b^2k^2n\left(n-1\right)\left[\frac{4n-2-6n+6+3n-3}{12}\right]+n\sigma^2\left(1-\frac{1}{n}\right)\\ &= \frac{b^2k^2n\left(n-1\right)\left(n+1\right)}{12}+\sigma^2\left(n-1\right).\end{aligned} \tag{6.6}$$

On substituting (6.6) into (6.5), we obtain

$$\mathrm{E}_m\left[\mathrm{E}\left(v_1\right)\right] = \mathrm{E}_m\left[\left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{b^2k^2n\left(n+1\right)}{12}+\sigma^2\right)\right] = \left(\frac{1}{n}-\frac{1}{N}\right)\left(\frac{b^2k^2n\left(n+1\right)}{12}+\sigma^2\right).$$

Similarly, we can obtain the expected values of estimators v_2 to v_7, for the model under consideration. Table 6.1 gives us the expected values of v_1 through to v_7 along with the corresponding approximate expected relative bias, under the assumptions that k is large and b is not very close to zero (see Wolter (2007)). The expected value of v_8 is obtained by assuming $\hat{\rho}_k > 0$ (i.e. $\ln(\hat{\rho}_k)$ is defined) and then approximating v_8 by the functions $\mathrm{E}_m\left[\mathrm{E}(s^2)\right]$ and $\mathrm{E}_m\left[\mathrm{E}(\hat{\rho}_k s^2)\right]$, such that

$$\mathrm{E}_m\left[\mathrm{E}\left(v_8\right)\right] = \left(\frac{1}{n}-\frac{1}{N}\right)\gamma_1\left[1+\frac{2}{\ln\left(\gamma_2/\gamma_1\right)}+\frac{2\gamma_2/\gamma_1}{\left(1-\gamma_2/\gamma_1\right)}\right],$$

where $\gamma_1 = \mathrm{E}_m\left[\mathrm{E}\left(s^2\right)\right] = b^2k^2n\left(n+1\right)/12+\sigma^2$ (which follows from equation (6.6)) and $\gamma_2 = \mathrm{E}_m\left[\mathrm{E}\left(\hat{\rho}_k s^2\right)\right] = b^2k^2\left(n-3\right)\left(n+1\right)/12-\sigma^2/n$.

Table 6.1: Expected values of v_1 to v_7 with their corresponding expected relative bias, under the linear trend model

Estimator	$\mathrm{E}_m\left[\mathrm{E}(v_\alpha)\right]$	$\mathrm{R}_m(v_\alpha)$
v_1	$(1/n-1/N)\left[b^2k^2n(n+1)/12+\sigma^2\right]$	n
v_2	$(1/n-1/N)\left[b^2k^2/2+\sigma^2\right]$	$-(n-6)/n$
v_3	$(1/n-1/N)\left[b^2k^2/2+\sigma^2\right]$	$-(n-6)/n$
v_4	$(1/n-1/N)\sigma^2$	-1
v_5	$(1/n-1/N)\sigma^2$	-1
v_6	$(1/n-1/N)\sigma^2$	-1
v_7	$(1/n-1/N)\left[b^2k^2n(p+1)/12+\sigma^2\right]$	k

From Table 6.1, it is clear that estimators v_2 and v_3 are the least biased and hence the preferred estimators of $\mathrm{Var}(\overline{y}_{LSS})$, for the linear trend model. It should be noted that Cochran (1977) promotes the use of estimator v_4 for linear trend populations. From Table 6.1, we see that estimators v_4, v_5 and v_6 eliminate the linear trend component and this is not desirable since $\mathrm{Var}(\overline{y}_{LSS})$ is a function of linear trend. However, estimators v_4, v_5 and v_6 should not be disregarded for linear trend populations, i.e. if we compare (5.47), (5.52) and (5.67) to the expected values of v_4, v_5 and v_6 in Table 6.1, with the notion that $\overline{y}_{BSS}$, $\overline{y}_{MSS}$ and $\overline{y}_{BMSS}$ are all unbiased estimators of $\overline{Y}$, then we note that estimators v_4, v_5 and v_6 are unbiased estimators of $\mathrm{Var}\left[\overline{y}_{BSS(n\text{ even})}\right]$, $\mathrm{Var}\left[\overline{y}_{MSS(n\text{ even})}\right]$ and $\mathrm{Var}\left[\overline{y}_{BMSS(n/2\text{ even integer})}\right]$. Moreover, these estimators are slightly biased for $\mathrm{Var}(\overline{y}_{YEC})$. It should be noted that it is impossible to estimate $\mathrm{Var}(\overline{y}_{CESS})$, since certain population units have no chance of being selected (Murthy 1967).

Periodic population model

Let us consider an example of an exact periodic population, given by the model

$$\mu_{ij} = a\sin\left(b[i+(j-1)k]\right),$$

where $e_{ij}=0$, for all values of i and j. Now, if we let $b=\pi/2$, then the period is given by $2\pi/b=4$. Furthermore, if we suppose that $k=4$, then it follows that:

(i) $\overline{y}_1 = \frac{1}{n}\sum\limits_{j=1}^{n} a\sin\left(\frac{\pi}{2}\left[1+(j-1)\,4\right]\right) = \frac{1}{n}\sum\limits_{j=1}^{n} a = a;$

(ii) $\overline{y}_2 = \frac{1}{n}\sum\limits_{j=1}^{n} a\sin\left(\frac{\pi}{2}\left[2+(j-1)\,4\right]\right) = \frac{1}{n}\sum\limits_{j=1}^{n} 0 = 0;$

(iii) $\overline{y}_3 = \frac{1}{n}\sum\limits_{j=1}^{n} a\sin\left(\frac{\pi}{2}\left[3+(j-1)\,4\right]\right) = \frac{1}{n}\sum\limits_{j=1}^{n} (-a) = -a;$

(iv) $\overline{y}_4 = \frac{1}{n}\sum\limits_{j=1}^{n} a\sin\left(\frac{\pi}{2}\left[4+(j-1)\,4\right]\right) = \frac{1}{n}\sum\limits_{j=1}^{n} 0 = 0.$

Note that we cannot provide an accurate estimate of $\mathrm{Var}(\overline{y}_{LSS})$ from a single sample, i.e. the actual value of $\mathrm{Var}(\overline{y}_{LSS})$ is $a^2/2$; however, all the variance estimators equate to zero, since the variate values of all the sampling units, from any one of the four samples, are equivalent. On the other hand, if k is equal to an odd multiple of half the period, then $\mathrm{Var}(\overline{y}_{LSS}) = 0$ and all estimators will be relatively large, since there is maximum variation in the variate values of the sampling units (see Section 4.2.3). This illustration again highlights the dangers of conducting LSS on periodic populations.

Auto-correlated population model

As shown in Section 4.2.4, an auto-correlated population assumes that the variate values of the population are correlated and hence the random errors are correlated, i.e. the e_{ij}s, given in (6.1), are not correlated random variables. An example of this is given by the underlying model used to construct v_8, which is represented by the first-order autoregressive process, given by

$$y_t - \mu = \phi\left(y_{t-1} - \mu\right) + \varepsilon_t, \qquad \text{for } t = 1, ..., N, \tag{6.7}$$

where ε_t denotes uncorrelated $(0, \sigma^2)$ random variables and ϕ denotes the first-order autocorrelation coefficient, such that $0 < \phi < 1$. Thus, by using (4.21), we obtain

$$\begin{aligned}\mathrm{E}_m\left[\mathrm{Var}\left(\overline{y}_{LSS}\right)\right] = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right) \times \Bigg[1 - \frac{2}{(k-1)}\frac{\left(\phi - \phi^N\right)}{(1-\phi)} + \frac{2}{N\,(k-1)} \\ \left\{\frac{\left(\phi - \phi^N\right)}{(1-\phi)^2} - \frac{(N-1)\,\phi^N}{(1-\phi)}\right\} + \frac{2k}{(k-1)}\frac{\left(\phi^k - \phi^N\right)}{(1-\phi^k)} \\ - \frac{2k}{n\,(k-1)}\left\{\frac{\left(\phi^k - \phi^N\right)}{(1-\phi^k)^2} - \frac{(n-1)\,\phi^N}{(1-\phi^k)}\right\}\Bigg].\end{aligned} \tag{6.8}$$

We can then approximate (6.8) to the order $0\left(n^{-2}\right)$, such that

$$\mathrm{E}_m\left[\mathrm{Var}\left(\overline{y}_{LSS}\right)\right] = \sigma^2\left(\frac{1}{n} - \frac{1}{N}\right)\left[1 - \frac{2}{(k-1)}\frac{\phi}{(1-\phi)} + \frac{2k}{(k-1)}\frac{\phi^k}{(1-\phi^k)}\right] + 0\left(n^{-2}\right), \tag{6.9}$$

where n is an index to a sequence and k is fixed. The corresponding expected values of all the estimators are given in Table 6.2. Large-n approximations, similar to that of (6.9), are provided for the expected values of v_1, v_7 and v_8. Estimator v_8 provides a good estimate of $\mathrm{Var}(\overline{y}_{LSS})$, since $-2\phi/k(1-\phi) \approxeq 2/\ln(\phi^k)$ (Cochran 1946) and thus $\mathrm{E}_m[\mathrm{E}(v_8)]$ in Table 6.2 is almost identical to (6.9).

Comparisons of the expected biases of the estimators are dependent on the values of ϕ and k, such that the differences between the expected values of the estimators are negligible when ϕ is small, and gets larger as ϕ increases, whereas these differences decrease as k increases (assuming ϕ is fixed). It is thus likely that estimator v_8 will provide an underestimate of $\mathrm{Var}(\overline{y}_{LSS})$, whilst the remaining estimators (especially v_1) are likely to provide an overestimate. Furthermore, estimator v_8 is likely to exhibit the smallest absolute bias, unless when ϕ is small, since the approximation of $2/\ln(\phi^k)$ is inadequate when $\phi \approxeq 0$.

Table 6.2: Expected values of v_1 to v_8 for the auto-correlated population model

Estimator	$\mathrm{E}_m\left[\mathrm{E}(v_\alpha)\right]$
v_1	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2+0\left(n^{-2}\right)$
v_2	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left(1-\phi^k\right)$
v_3	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left(1-\phi^k\right)$
v_4	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1-\frac{4\phi^3}{3}+\frac{\phi^{2k}}{3}\right]$
v_5	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1-\frac{12\phi^k}{7}+\frac{8\phi^{2k}}{7}-\frac{4\phi^{3k}}{7}+\frac{\phi^{4k}}{7}\right]$
v_6	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1-\frac{28\phi^k}{15}+\frac{24\phi^{2k}}{15}-\frac{20\phi^{3k}}{15}+\frac{16\phi^{4k}}{15}-\frac{12\phi^{5k}}{15}+\frac{8\phi^{6k}}{15}-\frac{4\phi^{7k}}{15}+\frac{\phi^{8k}}{15}\right]$
v_7	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1+\frac{2}{(p-1)}\right]\left[\frac{p\phi^{pk}}{(1-\phi^{pk})}-\frac{\phi^k}{(1-\phi^k)}\right]+0\left(n^{-2}\right)$
v_8	$\left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1+\frac{2}{\ln(\phi^k)}+\frac{2\phi^k}{(1-\phi^k)}\right]+0\left(n^{-2}\right)$

Stratification effects model

The classic notion of LSS being a process of selecting one population unit from each of the n strata, of size k (as in Section 4.1.3), is given by the model

$$\mu_{ij} = \mu_j,$$

such that μ_{ij} denotes a constant value within each stratum and the e_{ij}'s are given as iid $N(0, \sigma^2)$ random variables. Under this model, the expected variance of $\overline{y}_{LSS}$ is given as

$$\mathrm{E}_m \left[\mathrm{Var}(\overline{y}_{LSS})\right] = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma^2$$

and the corresponding expected values of all the estimators are as given in Table 6.3. With the highly probable assumptions of n being large and $\hat{\rho}_k > 0$, we provide an approximation for the expected value of v_8 in Table 6.3.

Table 6.3: Expected values of v_1 to v_8 for the stratification effects model

Estimator	$\mathrm{E}_m\left[\mathrm{E}(v_\alpha)\right]$
v_1	$(1/n - 1/N)\left[\sum_{j=1}^{n}\left(\mu_j - \overline{\mu}\right)^2/(n-1) + \sigma^2\right]$[a]
v_2	$(1/n - 1/N)\left[\sum_{j=1}^{n/2}\left(\mu_{2j-1} - \mu_{2j}\right)^2/n + \sigma^2\right]$
v_3	$(1/n - 1/N)\left[\sum_{j=1}^{(n-1)}\left(\mu_j - \mu_{j+1}\right)^2/2(n-1) + \sigma^2\right]$
v_4	$(1/n - 1/N)\left[\sum_{j=1}^{n-2}\left(\mu_j - 2\mu_{j+1} + \mu_{j+2}\right)^2/6(n-2) + \sigma^2\right]$
v_5	$(1/n - 1/N)\left[\sum_{j=1}^{n-4}\left(\mu_j/2 - \mu_{j+1} + \mu_{j+2} - \mu_{j+3} + \mu_{j+4}/2\right)^2/3.5(n-4) + \sigma^2\right]$
v_6	$(1/n - 1/N)\left[\sum_{j=1}^{n-8}\left(\mu_j/2 - \mu_{j+1} + -... + \mu_{j+8}/2\right)^2/7.5(n-8) + \sigma^2\right]$
v_7	$(1/n - 1/N)\left[n\sum_{\alpha=1}^{p}\left(\overline{\mu}_\alpha - \overline{\mu}\right)^2/p(p-1) + \sigma^2\right]$[b]
v_8	$(1/n - 1/N)(\omega_1 + \sigma^2)\left[1 + 2/\ln\kappa + 2\kappa/(1-\kappa)\right]$[c]

[a] $\overline{\mu} = \sum_{j=1}^{n}\mu_j/n$.

[b] $\overline{\mu}_\alpha = \alpha$-th systematic sub-sample mean of the μ_j.

[c] $\kappa = \omega_2/(\omega_1 + \sigma^2)$, $\omega_1 = \sum_{j=1}^{n}(\mu_j - \overline{\mu})^2/(n-1)$ and $\omega_2 = \sum_{j=1}^{n-1}(\mu_j - \overline{\mu})(\mu_{j+1} - \overline{\mu})/(n-1)$.

From Table 6.3, we note that the differences between v_1 to v_7 is negligible, if the mean of each stratum (μ_j, for $j \ \in \{1, ..., n\}$) is approximately equal. However, this is not common in practice, since stratification is such that strata are externally heterogeneous.

We will next evaluate the bias of each estimator, by assigning arbitrary values for μ_j, n, σ^2 and p. Accordingly, Table 6.4 contains the expected relative bias of each estimator, for the cases $\mu_j = j, j^{1/2}, j^{-1}$ and $\ln(j) + \sin(j)$, where $n = 20$, $\sigma^2 = 100$ and $p = 2$. Examples of linear trend between the strata are given by $\mu_j = j, j^{1/2}$ and j^{-1}, while an example of non-linear trend is given by $\mu_j = \ln(j) + \sin(j)$. From Table 6.4, it is clear that estimators v_4, v_5 and v_6 have the lowest level of bias and are thus preferred for the stratification effects model. The contrasts for these estimators are likely to eliminate the linear trend component in the stratum means, μ_j, which is appropriate since $\text{Var}(\overline{y}_{LSS})$ is not a function of this trend. Amongst these suitable estimators, we can further conclude that estimator v_6 is the most preferred estimator when the trend component is non-linear.

If the stratum means for neighbouring, non-overlapping pairs of strata are approximately equal (i.e. $\mu_{2j} \approxeq \mu_{2j-1}$, for $j = 1, ..., n/2$), then the expected bias of estimator v_2 will be the smallest. Also, if the average of the stratum means for neighbouring, non-overlapping groups of p strata are approximately equal (i.e. $\overline{\mu}_1 \approxeq \overline{\mu}_2 \approxeq ... \approxeq \overline{\mu}_\alpha \approxeq ...\overline{\mu}_p$, where $\mu_\alpha = p \sum_{j=1}^{n/p} \mu_{\alpha+(j-1)p}/n$), then the expected bias of estimator v_7 will be the smallest. It should be noted that equality for each of these two cases results in the expected bias for the corresponding estimators being zero.

Table 6.4: Expected relative bias multiplied by σ^2 for v_1 to v_8 under the stratification effects model

Estimator	$\mu_j = j$	$\mu_j = j^{1/2}$	$\mu_j = j^{-1}$	$\mu_j = \ln(j) + \sin(j)$
v_1	35.000	1.046	0.050	0.965
v_2	0.500	0.022	0.013	0.243
v_3	0.0500	0.020	0.008	0.243
v_4	0.000	0.000	0.001	0.073
v_5	0.000	0.000	0.001	0.034
v_6	0.000	0.000	0.000	0.013
v_7	5.000	0.177	0.022	0.206
v_8	−0.670	−0.396	−0.239	−0.373

We will next examine three designs which allow for an unbiased estimate of the sampling variance, two of which apply multiple random starts to the LSS design as well as the BMSS design (see Section 5.5), while the third design supplements the linear systematic sample with an independent simple random sample. Within each design we shall discuss the corresponding methodology, before obtaining expressions for the sample mean, the corresponding sampling variance and an unbiased estimate of the sampling variance. Thereafter, we will discuss the efficiency of the relative designs.

6.2 Multiple-Start Linear Systematic Sampling

6.2.1 Methodology

The method of inter-penetrating sub-sampling (or replicated sampling) was initially discussed by Mahalanobis (1946) and Tukey (1950) and later in the context of systematic sampling (i.e. MLSS) by Deming (1950), Gautschi (1957), Shiue (1960) and Tornqvist (1963). MLSS involves the selection of more than one linear systematic sample by applying the corresponding number of random starts.

The method of selecting a sample of size nm (where nm is now the required sample size with m being an integer) from a population of size N, using MLSS, is given as follows:

(i) Randomly select m integers $(i_1, ..., i_m)$ from the first k integers, such that $2 \leq m < k$.

(ii) The population unit numbers of the sampling units are then given by

$$i_h + (j-1)k, \qquad \text{for } h = 1, ..., m \text{ and } j = 1, ..., n.$$

This method can be viewed as selecting m clusters (each of size n) from k clusters, using SRS, where the clusters are defined as in Table 2.1.

6.2.2 Estimation formulae

Tornqvist (1963) suggested the use of SRSWR for step (i) above, such that an unbiased estimate of $\overline{Y}$ is given by

$$\overline{y}^{(m)}_{LSS(WR)} = \frac{1}{m}\sum_{h=1}^{m} \overline{y}_{i'_h}, \tag{6.10}$$

where (i) the superscript m denotes the number of random starts; (ii) $i'_h \in \{i'_1 ...., i'_m\}$ for step (i) denotes integers, where $\mathrm{P}(i'_h = i'_j) = 1/k$ for $j \in \{1, ..., h-1, h+1, ..., m\}$; and (iii) $\overline{y}_{i'_h}$ denotes the mean of the sample that is chosen with random starts i'_h in step (ii).

For this setup, the sample means can be viewed as population units, where we select a set of m sample means from the k possible sample means, using SRSWR. The corresponding adjusted population variance is thus obtained by replacing y_i and N in (3.9) with $\overline{y}_i$ and k respectively, such that

$$S_{\overline{y}}^2 = \frac{1}{k-1}\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2,$$

where the replacement of y_i and N in $\overline{Y} = \sum_{i=1}^{N} y_i/N$, by $\overline{y}_i$ and k respectively, results in $\sum_{i=1}^{k} \overline{y}_i/k = \overline{Y}$. Moreover, the variance of the estimator in (6.10) is obtained by replacing S_Y^2, N and n in (4.1) by $S_{\overline{y}}^2$, k and m respectively, such that

$$\begin{aligned}\text{Var}\left(\overline{y}_{LSS(WR)}^{(m)}\right) &= \left(\frac{k-1}{mk}\right)\frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2 \\ &= \frac{1}{mk}\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2 = \frac{\text{Var}\left(\overline{y}_{LSS}\right)}{m}, \qquad (6.11)\end{aligned}$$

which follows from (3.2). An unbiased estimate of (6.11) is then given as

$$v_9 = \frac{1}{m\,(m-1)}\sum_{h=1}^{m}\left(\overline{y}_{i'_h} - \overline{y}_{LSS(WR)}^{(m)}\right)^2,$$

since an unbiased estimate of Var($\overline{y}_{LSS}$) is given by

$$\frac{1}{(m-1)}\sum_{h=1}^{m}\left(\overline{y}_{i'_h} - \overline{y}_{LSS(WR)}^{(m)}\right)^2.$$

Note that the degrees of freedom (given by $(m-1)$) is adjusted for error, since we are sampling with replacement (Cochran 1977, pp.29-30).

Gautschi (1957) suggested the use of SRSWOR for step (i), such that an unbiased estimate of $\overline{Y}$ is given by

$$\overline{y}_{LSS(WOR)}^{(m)} = \frac{1}{m}\sum_{h=1}^{m}\overline{y}_{i_h} = \frac{1}{nm}\sum_{h=1}^{m}\sum_{j=1}^{n} y_{i_h+(j-1)k}, \qquad (6.12)$$

where *(i)* $i_h \in \{i_1, ..., i_m\}$ for step (i) denotes m distinct integers, i.e. P$(i_h = i_j) = 0$ for all $j \in \{1, ..., h-1, h+1, ..., m\}$; and *(ii)* $\overline{y}_{i_h}$ denotes the mean of the sample that is chosen with random starts i_h in step (ii). The variance of the estimator in (6.12) is then obtained by replacing S_Y^2, N and n in (4.2), by $S_{\overline{y}}^2$, k and m respectively, such that

$$\text{Var}\left(\overline{y}_{LSS(WOR)}^{(m)}\right) = \left(\frac{k-m}{mk}\right)\frac{1}{k-1}\sum_{i=1}^{k}\left(\overline{y}_i - \overline{Y}\right)^2 = \left(\frac{k-m}{k-1}\right)\frac{\text{Var}\left(\overline{y}_{LSS}\right)}{m}. \qquad (6.13)$$

Thus, if LSS is an efficient design for a sample of size n, then we can expect MLSS to be efficient at a sample of size nm. An unbiased estimate of (6.13) is then given by

$$v_{10} = \left(\frac{k-m}{mk}\right)\frac{1}{m-1}\sum_{h=1}^{m}\left(\overline{y}_{i_h} - \overline{y}^{(m)}_{LSS(WOR)}\right)^2,$$

since an unbiased estimate of $\sum_{i=1}^{k}(\overline{y}_i - \overline{Y})^2/(k-1)$ is given by the expression

$$\sum_{h=1}^{m}\left(\overline{y}_{i_h} - \overline{y}^{(m)}_{LSS(WOR)}\right)^2/(m-1),$$

i.e. we are now sampling without replacement and thus do not adjust the degrees of freedom, as in the case of SRSWR (Cochran 1977, pp.21-27).

6.2.3 Efficiency comparisons

By using (6.11) and (6.13), we obtain the relative efficiency of $\overline{y}^{(m)}_{LSS(WOR)}$, with respect to $\overline{y}^{(m)}_{LSS(WR)}$, given by

$$\frac{\text{Var}\left(\overline{y}^{(m)}_{LSS(WR)}\right)}{\text{Var}\left(\overline{y}^{(m)}_{LSS(WOR)}\right)} = \left[\frac{\text{Var}\left(\overline{y}_{LSS}\right)}{m}\right]\left[\left(\frac{k-m}{k-1}\right)\frac{\text{Var}\left(\overline{y}_{LSS}\right)}{m}\right]^{-1} = \frac{k-1}{k-m} > 1,$$

with the assumption of $2 \leq m < k$. For the remainder of this thesis, we will thus only consider selecting the m random starts using SRSWOR, where $\overline{y}^{(m)}_{MLSS} = \overline{y}^{(m)}_{LSS(WOR)}$ now denotes the sample mean, when conducting MLSS.

Gautschi (1957) examined the efficiency of LSS, when compared to MLSS, under the super-population model for various population structures, where appropriate adjustments to the sample size were applied for the former design (see Section 6.3.3). Under the super-population model in (4.12), if we denote the expected variance of $\overline{y}^{(m)}_{MLSS}$ by σ^2_{MLSS} , then a summary of Gautschi's results is given as follows:

(i) for randomly ordered populations, $\sigma^2_{MLSS} = \sigma^2_{LSS}$;

(ii) for populations that exhibit linear trend, $\sigma^2_{MLSS} > \sigma^2_{LSS}$;

(iii) for auto-correlated populations, where the correlogram is assumed to be linear, $\sigma^2_{MLSS} > \sigma^2_{LSS}$;

We are thus presented with a trade-off between accuracy (unbiased estimate of the sampling variance) and precision (efficiently estimating $\overline{Y}$), when comparing LSS to MLSS for populations that exhibit linear trend and auto-correlated populations.

6.3 Multiple-Start Balanced Modified Systematic Sampling

6.3.1 Methodology

The author next proposes a new design, which considers the application of BMSS (as discussed in Section 5.5) with m multiple random starts. With the assumption of $n > 2$, we note that there are five possible cases for the methodology of BMSS, i.e.

(A) $n/2$ is an even integer;

(B) $n/2$ is an odd integer;

(C) $n = 3$;

(D) $n \neq 3$ and $(n+1)/2$ is an even integer;

(E) $(n+1)/2$ is an odd integer.

Accordingly, the method of selecting a sample of size nm from a population of size N, using MBMSS, is given as follows:

(i) Randomly select m integers $(i_1, ..., i_m)$ from the first k integers, using SRSWOR, where $2 \leq m < k$.

(ii) For $h = 1, ..., m$, the sample units chosen for the respective cases will be those elements with population unit numbers given by

Case(A):

$$i_h + 2jk, \qquad 2(j+1)k - i_h + 1, \quad \text{for } j = 0, ..., (n-4)/4, \tag{6.14}$$

$$N + i_h - k - 2jk, \quad N - i_h - k - 2jk + 1, \quad \text{for } j = 0, ..., (n-4)/4, \tag{6.15}$$

Case(B):

$$i_h + 2jk, \qquad N + i_h - k - 2jk, \quad \text{for } j = 0, ..., (n-2)/4, \tag{6.16}$$

$$2(j+1)k - i_h + 1, \quad N - i_h - k - 2jk + 1, \quad \text{for } j = 0, ..., (n-6)/4, \tag{6.17}$$

Case(C):

$$i_h + 2jk, \qquad 2(j+1)k - i_h + 1, \qquad N - i_h - 2jk + 1,$$
$$\text{for } j = 0, ..., (n-3)/4, \tag{6.18}$$

Case(D): (6.18) as well as

$$N + i_h - 2(j+1)k, \qquad \text{for } j = 0, ..., (n-7)/4, \tag{6.19}$$

Case(E):

$$i_h + 2jk, \qquad 2(j+1)k - i_h + 1, \qquad i_h + (n-1)k/2,$$

$$\text{for } j = 0, ..., (n-5)/4, \tag{6.20}$$

$$N - i_h - 2jk + 1, \qquad N + i_h - 2(j+1)k, \quad \text{for } j = 0, ..., (n-5)/4. \tag{6.21}$$

6.3.2 Estimation formulae

By using (6.14) and (6.15), we obtain the sample mean for case (A), given as

$$\overline{y}^{(m)}_{MBMSS(A)} = \frac{1}{nm} \sum_{h=1}^{m} \sum_{j=0}^{(n/4-1)} \left[y_{i_h+2jk} + y_{2(j+1)k-i_h+1} + y_{N+i_h-k-2jk} + y_{N-i_h-k-2jk+1} \right]. \tag{6.22}$$

Theorem 6.1: The sample mean, given by (6.22), is an unbiased estimate of $\overline{Y}$.

Proof: For $i = 1, ..., k$, we denote the ith balanced modified systematic sample total by

$$T_i = \sum_{j=0}^{(n/4-1)} \left[y_{i+2jk} + y_{2(j+1)k-i+1} + y_{N+i-k-2jk} + y_{N-i-k-2jk+1} \right],$$

such that by using (6.22), we obtain

$$\mathrm{E}\left(\overline{y}^{(m)}_{MBMSS(A)}\right) = \frac{1}{nm} \sum_{h=1}^{m} \mathrm{E}\left(T_i\right) = \frac{m}{nm} \sum_{i=1}^{k} T_i \left(\frac{1}{k}\right) = \frac{1}{nk} \sum_{i=1}^{k} T_i = \frac{Y.}{nk} = \overline{Y}.$$

The sample means (T_i/n, for $i = 1, ..., k$) can now be viewed as population units, such that we are selecting m sample means from the k possible sample means, using SRSWOR. We thus replace y_i, N and n in (4.2) by T_i/n, k and m respectively, to obtain the variance of $\overline{y}^{(m)}_{MBMSS}$, written as

$$\mathrm{Var}\left(\overline{y}^{(m)}_{MBMSS(A)}\right) = \frac{S_T^2}{m} \left(\frac{k-m}{k}\right) = \left(\frac{k-m}{k-1}\right) \frac{\mathrm{Var}\left(\overline{y}_{BMSS(A)}\right)}{m}, \tag{6.23}$$

where

$$S_T^2 = \frac{1}{k-1} \sum_{i=1}^{k} \left(\frac{1}{n} T_i - \overline{Y}\right)^2 = \frac{k}{k-1} \mathrm{Var}\left(\overline{y}_{BMSS(A)}\right),$$

such that the replacement of y_i and N in $\overline{Y} = \sum_{i=1}^{N} y_i/N$, by T_i/n and k respectively, results in $\sum_{i=1}^{k} T_i/nk = Y./N = \overline{Y}$. Thus, we obtain efficient results for MBMSS when the sample size is nm, if and only if BMSS is an efficient design for a sample of size n. Furthermore, if we let

$$\overline{T} = \frac{1}{m}\sum_{h=1}^{m} T_{i_h} = n\overline{y}^{(m)}_{MBMSS(A)},$$

then it is not difficult to show that an unbiased estimate of (6.23), is given by

$$v_{11} = \left(\frac{k-m}{mk}\right)\frac{1}{n^2\,(m-1)}\sum_{h=1}^{m}\left(T_{i_h}-\overline{T}\right)^2,$$

i.e. $\sum\limits_{h=1}^{m}\left(T_{i_h}-\overline{T}\right)^2/n^2\,(m-1)$ is an unbiased estimate of S_T^2.

Similarly, we can obtain unbiased estimates of the population mean for cases (B) to (E) by using the corresponding population unit numbers in (6.16) to (6.21). We can then apply the above method, so as to obtain the corresponding sampling variances and the associated unbiased variance estimators.

6.3.3 Efficiency comparisons

We will now compare the efficiency of MBMSS, to that of SRSWOR, STR, LSS and MLSS, under various population structures. We will assume that $n/2$ is an even integer, such that the expressions obtained in the previous section may then be used. It should be noted that the results obtained in this section need not necessarily apply for the cases (B) to (E).

Before considering the efficiency comparisons on various population structures, we first need to make appropriate adjustments to the sample size. Accordingly, suppose there exists an integer l, such that $k = lm$. We then randomly select a population unit from the first l units and every lth unit thereafter, so as to obtain a linear systematic sample of size nm. We can now replace n and k by nm and l respectively, for the corresponding variance expressions obtained for LSS, SRSWOR and STR.

Population in random order

A population in random order can be represented by the model

$$y_i = \mu + e_i, \qquad \text{for } i = 1, ..., N, \tag{6.24}$$

where the random errors are drawn from a super-population (Cochran 1946), such that

$$\mathrm{E}_m(e_i) = 0, \qquad \mathrm{E}_m(e_i^2) = \sigma^2 \qquad \text{and} \qquad \mathrm{E}_m(e_ie_j) = 0\ (i \neq j).$$

By substituting (6.24) into T_i and noting that $\overline{Y} = \mu + \overline{\overline{e}}$, we then use (6.23) to obtain an expression for the expected variance of $\overline{y}^{(m)}_{MBMSS(A)}$, i.e.

$$\begin{aligned}
\sigma^2_{MBMSS(A)} &= \mathrm{E}_m \left[\frac{k-m}{mk\,(k-1)} \sum_{i=1}^{k} \left(\frac{1}{n} T_i - \overline{Y} \right)^2 \right] \\
&= \frac{k-m}{mk\,(k-1)} \sum_{i=1}^{k} \mathrm{E}_m \left[\frac{1}{n} \left(n\mu + e_T \right) - \left(\mu + \overline{\overline{e}} \right) \right]^2 \\
&= \frac{k-m}{mk\,(k-1)} \sum_{i=1}^{k} \mathrm{E}_m \left[\frac{1}{n} e_T - \overline{\overline{e}} \right]^2 \\
&= \frac{k-m}{mk\,(k-1)} \sum_{i=1}^{k} \mathrm{E}_m \left[\frac{e_T^2}{n^2} - \frac{2e_T\overline{\overline{e}}}{n} + \overline{\overline{e}}^2 \right], \qquad (6.25)
\end{aligned}$$

where

$$e_T = \sum_{j=0}^{(n/4-1)} \left(e_{i+2jk+}e_{2(j+1)k-i+1} + e_{N+i-k-2jk} + e_{N-i-k-2jk+1} \right).$$

Now, since there are n terms in e_T, it follows that

$$\mathrm{E}\left(e_T^2 \right) = n\sigma^2 \qquad (6.26)$$

and

$$\mathrm{E}_m \left(e_T \overline{\overline{e}} \right) = \frac{1}{N} \mathrm{E}_m \left[e_T \left(e_1 + ... + e_N \right) \right] = \frac{n\sigma^2}{N}. \qquad (6.27)$$

Thus, by substituting (5.7), (6.26) and (6.27) into (6.25), it follows that

$$\begin{aligned}
\sigma^2_{MBMSS(A)} &= \frac{k-m}{mk\,(k-1)} \sum_{i=1}^{k} \left(\frac{\sigma^2}{n} - \frac{2\sigma^2}{N} + \frac{\sigma^2}{N} \right) \\
&= \frac{(k-m)\,k}{mk\,(k-1)} \left[\frac{N-n}{nN} \right] \sigma^2 = \frac{(k-m)\,n\sigma^2}{mnN} = \frac{l-1}{N}\sigma^2, \qquad (6.28)
\end{aligned}$$

since $k = lm$. Similarly, the expected sampling variances for the comparative designs can be show to be equivalent to (6.28). We can thus conclude that MBMSS is as equally efficient as SRSWOR, STR, LSS and MLSS, for populations in random order.

Population with linear trend

If we substitute the model in (5.1) into T_i, we note that

$$\begin{aligned}
T_i &= \sum_{j=0}^{(n/4-1)} \left[4a + b\,(2N+2) \right] + e_T \\
&= na + bn \left(\frac{N+1}{2} \right) + e_T.
\end{aligned}$$

Taking the expectation of (6.23) and then substituting the above expression along with (5.3), results in

$$\begin{aligned}
\sigma^2_{MBMSS(A)} &= \frac{k-m}{mk\left(k-1\right)}\sum_{i=1}^{k}\mathrm{E}_m\left(\frac{1}{n}T_i-\overline{Y}\right)^2\\
&= \frac{k-m}{mk\left(k-1\right)}\sum_{i=1}^{k}\mathrm{E}_m\left[\frac{1}{n}\left\{na+bn\left(\frac{N+1}{2}\right)+e_T\right\}-\left\{a+\frac{b\left(N+1\right)}{2}+\overline{\overline{e}}\right\}\right]^2\\
&= \frac{k-m}{mk\left(k-1\right)}\sum_{i=1}^{k}\mathrm{E}\Big[\frac{e_T}{n}-\overline{\overline{e}}\Big]^2=\frac{l-1}{N}\sigma^2, \qquad (6.29)
\end{aligned}$$

which follows from (6.25) and (6.28). By replacing k and n in (5.9), (5.10) and (5.11), by l and nm respectively, we then obtain the corresponding expected variances of $\overline{y}_{LSS}$, $\overline{y}_{SRSWOR}$ and $\overline{y}_{STR}$ under the model in (5.1), i.e.

$$\sigma^2_{LSS}=\sigma^2\left(\frac{1}{nm}-\frac{1}{N}\right)+\frac{b^2\left(l^2-1\right)}{12}=\frac{l-1}{N}\sigma^2+\frac{b^2\left(l-1\right)\left(l+1\right)}{12}, \qquad (6.30)$$

$$\begin{aligned}
\sigma^2_{SRSWOR} &= \sigma^2\left(\frac{1}{nm}-\frac{1}{N}\right)+\frac{b^2\left(N+1\right)\left(l-1\right)}{12}\\
&= \frac{l-1}{N}\sigma^2+\frac{b^2\left(N+1\right)\left(l-1\right)}{12} \qquad (6.31)
\end{aligned}$$

and

$$\sigma^2_{STR}=\sigma^2\left(\frac{1}{nm}-\frac{1}{N}\right)+\frac{b^2\left(l^2-1\right)}{12nm}=\frac{l-1}{N}\sigma^2+\frac{b^2\left(l-1\right)\left(l+1\right)}{12nm}, \qquad (6.32)$$

where $\overline{y}_{LSS}$, $\overline{y}_{SRSWOR}$, $\overline{y}_{STR}$ are unbiased estimates of $\overline{Y}$, resulting in the expected MSEs of these estimates being equivalent to the corresponding expected sampling variances. Furthermore, by taking the expectation of (6.13) and substituting (5.9) into it, we obtain

$$\begin{aligned}
\sigma^2_{MLSS} &= \frac{k-m}{m\left(k-1\right)}\left[\sigma^2\left(\frac{1}{n}-\frac{1}{N}\right)+\frac{b^2\left(k^2-1\right)}{12}\right]\\
&= \frac{k-m}{m\left(k-1\right)}\left[\frac{k-1}{N}\sigma^2+\frac{b^2\left(k-1\right)\left(k+1\right)}{12}\right].
\end{aligned}$$

Remembering that $k=lm$, it follows that

$$\begin{aligned}
\sigma^2_{MLSS} &= \frac{lm-m}{m\left(lm-1\right)}\left[\frac{lm-1}{N}\sigma^2+\frac{b^2\left(lm-1\right)\left(lm+1\right)}{12}\right]\\
&= \frac{l-1}{N}\sigma^2+\frac{b^2\left(l-1\right)\left(lm+1\right)}{12}. \qquad (6.33)
\end{aligned}$$

By comparing (6.30), (6.31), (6.32) and (6.33) to (6.29), we conclude that MBMSS is more efficient than LSS, SRSWOR, STR and MLSS.

Periodic population

Let us assume that y_i exhibits an exact periodic function with period $2h$ and $N = 2Qh$, where h and Q are positive integers. Moreover, suppose that the worst case scenario for LSS is observed, i.e. $k = 2ah$, where $a \in \{1, 2, ...\}$, such that $Q = an$. Now, the variate values corresponding to this scenario are given by $y_j = y_{j+2h} = ... = y_{j+2h(Q-1)}$, for $j = 1, ..., 2h$, which results in LSS being equivalent to the random selection of a single population unit, i.e. the distance between each sampling unit is k, which is an integral multiple of the period. Thus, $N = 2ah = k$ and by replacing $\overline{y}_i$ and k in (3.2) with y_i and N respectively, we then obtain $\text{Var}(\overline{y}_{LSS}) = \sum_{i=1}^{N}(y_i - \overline{Y})^2/N = \sigma^2$.

Let us next consider BMSS for this scenario. If we assume that $n/2$ is an even integer, then by using (5.53) and (5.54), we obtain the variate values of the ith balanced modified systematic sample, for $i = 1, ..., 2ah$, i.e.

$$y_{i+4ahj}, \quad y_{4(j+1)ah-i+1}, \quad y_{2Qh+i-2ah-4ahj}, \quad y_{2Qh-i-2ah-4ahj+1}, \quad \text{for } j = 0, ..., n/4 - 1.$$

By comparing y_{i+4ahj} to $y_{4(j+1)ah-i+1}$, we note that the distance between these units is given by $4ah-2i+1$, which is not a function of $2h$, since $i \in \{1, ..., 2ah\}$ and h is a positive integer. We thus conclude that $y_{i+4ahj} \neq y_{4(j+1)ah-i+1}$. Also, by comparing y_{i+4ahj} to $y_{2Qh+i-2ah-4ahj}$, we note that the distance apart is $2Qh - 2ah - 8ahj = 2h(Q - a - 4aj)$, which results in $y_{i+4ahj} = y_{2Qh+i-2ah-4ahj}$. Moreover, the distance between $y_{4(j+1)ah-i+1}$ and $y_{2Qh-i-2ah-4ahj+1}$ is given by $2Qh - 6ah - 8ahj = 2h(Q - 3a - 4aj)$, resulting in $y_{4(j+1)ah-i+1} = y_{2Qh-i-2ah-4ahj+1}$. By using the transitive law, we thus conclude that a balanced modified systematic sample of size n (where $n/2$ is an even integer), is equivalent to the random selection of two population units with distinct variate values, resulting in BMSS being twice as efficient as LSS. Now, since MLSS and MBMSS are equivalent to the selection of m linear systematic samples and m balanced modified systematic samples respectively, we thus conclude that MBMSS is twice as efficient as MLSS, which in turn is much more efficient than LSS, i.e. MLSS is equivalent to SRSWOR of m sampling units with distinct variate values. By assuming $n/2$ to be an even integer, the best case scenario (i.e. $k = ch$ for $c \in \{1, 3, ..., \}$) results in

$$\text{Var}\left[\overline{y}^{(m)}_{MBMSS(A)}\right] = \text{Var}\left[\overline{y}^{(m)}_{MLSS}\right] = \text{Var}\left[\overline{y}_{BMSS}\right] = \text{Var}\left[\overline{y}_{LSS}\right] = 0.$$

Hence, MBMSS is equally efficient as MLSS and LSS, and more efficient than SRSWOR and STR, when k is an odd multiple of half the period. If we use a more realistic periodic population model under the assumptions of the super-population model in (4.12), then it will not be difficult to show that MBMSS is less efficient than STR, which in turn is less efficient than SRSWOR, for the worst case scenario.

Auto-correlated population

If we assume that the population units are correlated, as discussed in Section 4.2.4, then the sum of the serial correlations for the ith balanced modified systematic sample is given as

$$\sum_{j=0}^{n/4-1}\left[\rho_{2k-2i+1}+\rho_{N-k-4jk}+\rho_{N-2i-k-4jk+1}+\rho_{N+2i-3k-4jk-1}+\rho_{N-3k-4jk}+\rho_{2i-1}\right].$$

Therefore, the sum of the serial correlations for MBMSS can be written as

$$\sum_{h=1}^{m}\sum_{j=0}^{n/4-1}\left[\rho_{2k-2i_h+1}+\rho_{N-k-4jk}+\rho_{N-2i_h-k-4jk+1}+\rho_{N+2i_h-3k-4jk-1}+\rho_{N-3k-4jk}+\rho_{2i_h-1}\right].$$

Remembering that $k=lm$, we use the above expression, along with (4.16) and (4.18), to obtain the expected variance of $\overline{y}_{MBMSS(A)}$, written as

$$\begin{aligned}\sigma^2_{MBMSS(A)} = \frac{lm-1}{N}\sigma^2+\frac{2\sigma^2}{n^2}\sum_{h=1}^{m}\sum_{j=0}^{n/4-1}&\big[\rho_{2lm-2i_h+1}+\rho_{N-lm-4jlm}+\rho_{N-2i_h-lm-4jlm+1}\\ &+\rho_{N+2i_h-3lm-4jlm-1}+\rho_{N-3lm-4jlm}+\rho_{2i_h-1}\big]\\ &-\frac{2\sigma^2}{N^2}\sum_{u=1}^{N}(N-u)\rho_u. \qquad (6.34)\end{aligned}$$

By replacing k and n in (4.19) to (4.21) by l and nm respectively, we then obtain the corresponding expected variance of $\overline{y}_{SRSWOR}$, $\overline{y}_{STR}$ and $\overline{y}_{LSS}$, i.e.

$$\sigma^2_{SRSWOR} = \frac{l-1}{N}\sigma^2+\frac{2(l-1)\sigma^2}{N^2(N-1)}\sum_{u=1}^{N-1}(N-u)\rho_u, \qquad (6.35)$$

$$\sigma^2_{STR} = \frac{l-1}{N}\sigma^2+\frac{2\sigma^2}{Nl}\sum_{u=1}^{l-1}(l-u)\rho_u, \qquad (6.36)$$

$$\sigma^2_{LSS} = \frac{l-1}{N}\sigma^2-\frac{2\sigma^2}{N^2}\sum_{u=1}^{N-1}(N-u)\rho_u+\frac{2\sigma^2}{n^2m^2}\sum_{u=1}^{nm-1}(nm-u)\rho_{lu}. \qquad (6.37)$$

Furthermore, by taking the expectation of (6.13) and substituting (4.21) into it, we obtain

$$\sigma^2_{MLSS} = \frac{k-m}{m(k-1)}\frac{(k-1)\sigma^2}{N}\left[1-\frac{2}{N(k-1)}\sum_{u=1}^{N-1}(N-u)\rho_u+\frac{2k}{n(k-1)}\sum_{u=1}^{n-1}(n-u)\rho_{ku}\right].$$

Remembering that $k = lm$, reduces the above expression to

$$\sigma^2_{MLSS} = \frac{l-1}{N}\sigma^2 - \frac{2(l-1)\sigma^2}{N^2(lm-1)}\sum_{u=1}^{N-1}(N-u)\rho_u + \frac{2(l-1)^2}{n^2(lm-1)}\sum_{u=1}^{n-1}(n-u)\rho_{lmu}. \tag{6.38}$$

From (6.34) to (6.38), it is difficult to obtain simple theoretical comparisons and we will thus resort to some numerical comparisons in Chapter 8.

6.4 Partially Systematic Sampling

6.4.1 Methodology

Supplementing a linear systematic sample with an independent simple random sample, termed as PSS, was first noted by Zinger (1963, 1964) and later discussed in detail by Zinger (1980) and Wu (1984). The corresponding method for selecting a sample of size $n = n_1 + n_2$ from a population of size N, is given as follows:

(i) If we suppose that $k = N/n_1$ is an integer, then randomly select an integer between 1 and k, say i, where $1 \leq i \leq k$.

(ii) Select a sample of size n_1 using LSS, such that the sample units chosen will be those elements with population unit numbers given by

$$i + (j-1)k, \qquad \text{for } j = 1, ..., n_1. \tag{6.39}$$

(iii) Select n_2 sampling units from the remaining $N-n_1$ population units using SRSWOR.

(iv) The partially systematic sample is then given as the sample in step (ii), supplemented with the sample in step (iii).

6.4.2 Estimation formulae

If we let $\overline{y}_s$ and $\overline{y}_r$ denote the means resulting from steps (ii) and (iii) respectively, then an unbiased estimate of $\overline{Y}$, given by Zinger (1980), is the weighted average of the corresponding means, i.e.

$$\overline{y}_{PSS} = \overline{y}(\beta) = (1-\beta)\overline{y}_s + \beta\overline{y}_r, \qquad 0 \leq \beta \leq 1. \tag{6.40}$$

Zinger further provides the corresponding variance of (6.40), by taking the expectations with respect to the designs in (ii) and (iii), such that

$$\text{Var}(\overline{y}(\beta)) = \alpha_1(\beta)S^2_Y + \alpha_2(\beta)\text{Var}(\overline{y}_s), \tag{6.41}$$

where

$$\alpha_1(\beta) = \frac{\beta^2(N-1)(N-n_1-n_2)}{n_2 N(N-n_1-1)},$$
$$\alpha_2(\beta) = \left[1 - \frac{\beta k}{(k-1)}\right]^2 - \frac{\beta^2(N-n_1-n_2)}{n_2(k-1)^2(N-n_1-1)}$$

and

$$\mathrm{Var}(\overline{y}_s) = \frac{S_Y^2}{n_1}\left(\frac{N-1}{N}\right)[1+(n_1-1)\rho].$$

Suppose that $Q_s = \sum(y_i - \overline{y}_s)^2$, $Q_r = \sum(y_i - \overline{y}_r)^2$ and $Q_b = \sum(\overline{y}_s - \overline{y}_r)^2$ respectively denote the sum of squares within the linear systematic sample, the sum of squares within the simple random sample and the sum of squares between both samples. An unbiased estimate of (6.41) is provided by Wu (1984) and given as

$$v_{11} = v(\overline{y}(\beta)) = B(Q_s + \lambda Q_r) + DQ_b,$$

where

$$B = \frac{d_2\alpha_1(\beta) - d_1\alpha_2(\beta)}{d_2(n_1+\lambda c_1) + d_1(n_1+\lambda c_2)}, \qquad D = \frac{\alpha_1(\beta)[n_1+\lambda c_2] + \alpha_2(\beta)[n_1+\lambda c_1]}{d_2(n_1+\lambda c_1)+d_1(n_1+\lambda c_2)},$$

$$c_1 = \frac{(n_2-1)(N-n_1)}{(N-n_1-1)}, \qquad c_2 = \frac{n_1^2(n_2-1)}{(N-n_1)(N-n_1-1)},$$

$$d_1 = \frac{(N-n_1-n_2)}{n_2(N-n_1-1)}, \qquad d_2 = \frac{(n_2N^2 - n_2N - n_1^2 - n_1n_2)}{n_2(N-n_1)(N-n_1-1)}.$$

Wu further noted that $v(\overline{y}(\beta))$ will always be non-negative if

$$\text{(a) } \lambda \geq 0 \qquad \text{and} \qquad \text{(b) } \beta \geq (k-1)/2k. \qquad (6.42)$$

By letting $\lambda = 1$ and $\beta = (k-1)/2k$, we obtain an estimator which eliminates Q_s and Q_r, i.e.

$$v\left(\overline{y}\left(\frac{k-1}{2k}\right)\right) = \left(\frac{k-1}{2k}\right)^2 [\overline{y}_s - \overline{y}_r]^2.$$

The values of $\beta = n_2/(n_1+n_2)$ and $\beta = 1/2$ respectively correspond to a natural weighted average and an unweighted average of $\overline{y}_s$ and $\overline{y}_r$ (Zinger 1980). Zinger examined these two values along with the assumption that $\lambda = 1$, for the estimator $v(\overline{y}(\beta))$. He showed that $v(\overline{y}(n_2/(n_1+n_2)))$ is prone to producing negative values, while

$$v\left[\overline{y}\left(\frac{1}{2}\right)\right] = \frac{(N-n_1-n_2)(Q_s+Q_r)}{N[(n_2k-2)(n_1+n_2)+(n_1-n_2)k]}$$

$$+\left[\frac{1}{4}-\frac{n_2(\{N-n_1\}\{n_1+n_2\}-N)}{N(\{n_2k-2\}\{n_1+n_2\}+\{n_1-n_2\}k)}\right]Q_b$$

will always be non-negative and is thus the preferred weighting.

The assumption that k is large and $n_2 \leq n_1$ is usually met in most practical cases, since if $n_2 > n_1$ is true, then it is more beneficial to use MLSS, owing to its simplicity (Wu 1984). Accordingly, the optimum value of β, say β_{opt}, which minimizes (6.41) and results in LSS being more efficient than SRSWOR, is usually smaller than $(k-1)/2k \approxeq 1/2$ (Wolter 2007). By using (6.42), we thus conclude that there is a trade-off between efficiency and non-negative unbiased estimation of the sampling variance, when choosing an appropriate value of β. Wu suggests the following approach to overcome this trade-off:

(i) if $\beta_{opt} > (k-1)/2k$, then use $\overline{y}\left(\beta_{opt}\right)$ and $v\left(\overline{y}\left(\beta_{opt}\right)\right)$;

(ii) if $0.2 \leq \beta_{opt} \leq (k-1)/2k$, then use either $\overline{y}((k-1)/2k)$ and $v(\overline{y}((k-1)/2k))$, or $\overline{y}(1/2)$ and $v(\overline{y}(1/2))$;

(iii) if $\beta_{opt} < 0.2$, then use $\overline{y}(\beta_{opt})$ and the truncated variance estimator, given by $v_+(\overline{y}(\beta_{opt})) = \max\{\overline{y}(\beta_{opt}), 0\}$.

Wus approach is reasonably justified, except for case (iii), where a zero variance estimate is no more dangerous than a negative variance estimate (Wolter 2007).

6.4.3 Efficiency comparisons

Zinger (1980) studied the efficiency of PSS, when compared to SRSWOR, by assuming four artificial populations, two of which, exhibit stratification effects and linear trend (Cochran 1977, pp.211-212) as well as two that are in random order. He then concluded that PSS is more efficient than SRSWOR for Cochran's artificial populations, as well as one of the artificial populations in random order, provided that $n_2 > 2$. Zinger further compared PSS with MLSS by making the appropriate adjustments to the sample size. No simple conclusion was drawn for Cochran's artificial populations, i.e. PSS was more efficient than MLSS in some cases, while MLSS was more efficient than PSS in other cases.

It should be noted that Wu (1984) considered a modified approach to PSS, where the linear systematic sample of size n_1 is supplemented by another linear systematic sample, of size n_2. He examined two cases where (i) $n_1 = n_2$ and (ii) $n_1 = n_2 l$, such that l is an integer greater than one. It is trivial that case (i) reduces to MLSS with $m = 2$, while case (ii) does not seem to have any comparative advantage over MLSS (Wolter 2007). Moreover, this approach is not directly comparable to the usual PSS design, since we obtain values of β, which is not the same as that of PSS.

We next examine further designs which result in an unbiased estimate of the corresponding sampling variances, by supplementing a linear systematic sample with a dependent sample.

Chapter 7

SUPPLEMENTING A SYSTEMATIC SAMPLE WITH A DEPENDENT SAMPLE

In this chapter, we will examine some designs which supplement a systematic sample with a dependent sample. The designs that will be discussed are *new systematic sampling* (NSS), *new partially systematic sampling* (NPSS), *balanced random sampling* (BRS) and a new proposed design termed as *balanced modified random sampling* (BMRS). In NSS, we supplement a circular systematic sample with a sample of continuous units, while NPSS involves the supplementation of a circular systematic sample with a simple random sample without replacement. BRS is a slight adaption of MSS, whereby half the sample is selected using SRSWOR and the other half of the sample are the paired units, using the MSS pairing technique, as discussed in Section 5.4.1. Similarly, BMRS divides the population into groups, before conducting SRSWOR within each group and then pairs these sampling units using the MSS pairing technique. Before discussing each design, we will obtain some preliminary results to aid us in obtaining expressions for an estimate $\overline{Y}$, the corresponding sampling variance and an estimate of the sampling variance.

7.0 Preliminary Results

Theorem 7.1: Suppose that we draw a without-replacement sample of size n from a population of size N, such that the sample space is given as S. If $\pi_i > 0$, for all $i \in \{1, ..., N\}$, then the Horvitz & Thompson (1952) unbiased estimator of $\overline{Y}$ and the corresponding variance, are respectively given by

$$\hat{\overline{Y}}_{HT} = \frac{1}{N} \sum_{U_i \in S} \frac{y_i}{\pi_i}, \tag{7.1}$$

$$\text{Var}\left(\hat{\overline{Y}}_{HT}\right) = \frac{1}{N^2} \sum_{i=1}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{(\pi_{ij} - \pi_i \pi_j)}{\pi_i \pi_j} y_i y_j. \tag{7.2}$$

Proof: For $i = 1, ..., N$, let t_i be a random variable, such that

$$t_i = \begin{cases} 1 & \text{if } i\text{th unit is drawn,} \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, if we have a sample of size $n = 1$, then t_i follows a Bernoulli distribution with probability π_i, i.e. $t_i \sim \text{BIN}(1, \pi_i)$. Hence, $\text{E}(t_i) = \pi_i$ and $\text{Var}(t_i) = \pi_i(1 - \pi_i)$. Now, if we take y_i to be a fixed variable, then

$$\text{E}\left(\hat{\overline{Y}}_{HT}\right) = \text{E}\left(\frac{1}{N} \sum_{U_i \in S} \frac{y_i}{\pi_i}\right) = \frac{1}{N}\text{E}\left(\sum_{i=1}^{N} \frac{t_i y_i}{\pi_i}\right) = \frac{1}{N} \sum_{i=1}^{N} y_i = \overline{Y}.$$

Moreover, if we assume that $\pi_i = 0$, for some $i \in \{1, ..., N\}$, then $\text{E}(t_i) \neq \pi_i = 0$ and thus

$$\text{E}\left(\hat{\overline{Y}}_{HT}\right) \neq \overline{Y}.$$

The variance of the estimate in (7.1) is then given as

$$\begin{aligned} \text{Var}\left(\hat{\overline{Y}}_{HT}\right) &= \text{Var}\left(\frac{1}{N} \sum_{i=1}^{N} \frac{t_i y_i}{\pi_i}\right) \\ &= \sum_{i=1}^{N} \left(\frac{y_i}{N\pi_i}\right)^2 \text{Var}\,(t_i) + \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{y_i y_j}{N^2 \pi_i \pi_j} \text{Cov}\,(t_i t_j). \end{aligned} \tag{7.3}$$

Now,

$$\text{Cov}\,(t_i t_j) = \text{E}\,(t_i t_j) - \text{E}\,(t_i)\,\text{E}\,(t_j) = \pi_{ij} - \pi_i \pi_j,$$

since $t_i t_j = 1$ if and only if both units U_i and U_j are in the sample. By substituting this expression as well as $\text{Var}(t_i) = \pi_i(1 - \pi_i)$ into (7.3), we obtain the result in (7.2).

Theorem 7.2: With the assumption of a sampling design that exhibits fixed sample sizes, a form of the variance expression in (7.2), proposed by Sen (1953) and Yates & Grundy (1953), is given by

$$\text{Var}\left(\hat{\overline{Y}}_{HT}\right) = \frac{1}{2N^2} \sum_{i=1}^{N} \sum_{j \neq i}^{N} (\pi_i \pi_j - \pi_{ij}) \left[\left(\frac{y_i}{\pi_i}\right) - \left(\frac{y_j}{\pi_j}\right)\right]^2. \tag{7.4}$$

Proof: By using the given assumption, we obtain $\sum_{i=1}^{N} t_i = n$ for every possible sample, which results in

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \mathrm{E}\left(t_i\right) = \mathrm{E}\left(\sum_{i=1}^{N} t_i\right) = n. \tag{7.5}$$

Furthermore,

$$\sum_{j\neq i}^{N} \pi_{ij} = (n-1)\pi_i, \tag{7.6}$$

since there are $(n-1)$ values for j once the ith population unit is selected. By using (7.5) and (7.6), it then follows that

$$\begin{aligned}\frac{1}{\pi_i}\sum_{j\neq i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right) &= \sum_{j\neq i}^{N}\pi_i - (n-1)\\ &= \sum_{j=1}^{N}\pi_j - \pi_i - (n-1) = \left(1-\pi_i\right). \end{aligned}\tag{7.7}$$

Finally, by substituting (7.7) into (7.2), we obtain

$$\begin{aligned}\mathrm{Var}\left(\hat{\bar{Y}}_{HT}\right) &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\frac{y_i^2\left(\pi_i\pi_j - \pi_{ij}\right)}{\pi_i^2} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\frac{y_iy_j}{\pi_i\pi_j}\\ &= \frac{1}{N^2}\left[\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left\{\left(\frac{y_i}{\pi_i}\right)^2 + \left(\frac{y_j}{\pi_j}\right)^2\right\} - 2\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\frac{y_iy_j}{\pi_i\pi_j}\right]\\ &= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j>i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left[\left(\frac{y_i}{\pi_i}\right)^2 + \left(\frac{y_j}{\pi_j}\right)^2 - \frac{2y_iy_j}{\pi_i\pi_j}\right]\\ &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left[\left(\frac{y_i}{\pi_i}\right) - \left(\frac{y_j}{\pi_j}\right)\right]^2.\end{aligned}$$

Theorem 7.3: If we assume that $\pi_{ij} > 0$, for all $i, j \in \{1, ..., N\}$ $(i \neq j)$, then an unbiased estimator of (7.4) is given by

$$v_{YG} = \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N} t_it_j\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left[\left(\frac{y_i}{\pi_i}\right) - \left(\frac{y_j}{\pi_j}\right)\right]^2, \tag{7.8}$$

Proof: The expected value of v_{YG} is given as

$$\begin{aligned}\mathrm{E}(v_{YG}) &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\mathrm{E}\left(t_it_j\right)\left(\frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\right)\left[\left(\frac{y_i}{\pi_i}\right) - \left(\frac{y_j}{\pi_j}\right)\right]^2\\ &= \frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j\neq i}^{N}\left(\pi_i\pi_j - \pi_{ij}\right)\left[\left(\frac{y_i}{\pi_i}\right) - \left(\frac{y_j}{\pi_j}\right)\right]^2 = \mathrm{Var}\left(\hat{\bar{Y}}_{HT}\right),\end{aligned}$$

which follows since the expected value of the product of both indicator variables is given as $\mathrm{E}(t_i t_j) = \mathrm{P}(U_i \text{ and } U_j \in S) = \pi_{ij}$. Now, if we assume that $\pi_{ij} = 0$, for some i and $j \in \{1, ..., N\}$ $(i \neq j)$, then $\mathrm{E}(t_i t_j) \neq \pi_{ij} = 0$, so that

$$\mathrm{E}(v_{YG}) \neq \mathrm{Var}\left(\hat{Y}_{HT}\right).$$

Two basic properties of a good sampling design are that $\pi_i > 0$, for all $i \in \{1, ..., N\}$, and $\pi_{ij} > 0$, for all $i, j \in \{1, ..., N\}$ $(i \neq j)$. The condition on the first-order inclusion probabilities ensures that a sampling design produces samples of distinct sampling units, while the condition on the second-order inclusion probabilities ensures that it is possible to obtain an unbiased estimate of the sampling variance, provided that the corresponding sampling design exhibits fixed sample sizes. Thus, for each sampling design, we will aim to prove that the conditions of the inclusion probabilities hold true and/or obtain expressions for the inclusion probabilities. We can then substitute these expressions into (7.1), (7.4) and (7.8), so as to obtain the corresponding formulae. Furthermore, we will also discuss the methodology and efficiency of each design.

7.1 New Systematic Sampling

7.1.1 Methodology

NSS was first introduced by Singh & Singh (1977). The corresponding methodology to select a sample of size n from a population of size N is given as follows:

(i) Randomly select an integer r from the interval $[1, N]$.

(ii) Let $u \leq n$ be an integer, such that the sample units selected (of size u) for the sample S'_u, will be those elements with population unit numbers given by

$$r + i, \qquad \text{for } i = 0, 1, ..., u - 1. \tag{7.9}$$

(iii) With a sampling interval of $k'' = \mathrm{INT}[(N-u)/(n-u)]$, the remaining $n-u$ sampling units selected for the sample S''_u are given by the population unit numbers

$$r + u - 1 + jk'', \qquad \text{for } j = 1, ..., n - u. \tag{7.10}$$

(iv) The new systematic sample is then given as $S_u = S'_u \cup S''_u$.

The method of selection for the sample S''_u is the unrestricted method, which is applied in a circular fashion, such that $U_{N+r} = U_r$ (refer to Section 2.2.3). Furthermore, the probabilities of selecting the samples are given as

$$\mathrm{P}(S'_u) = \mathrm{P}(S''_u) = \mathrm{P}(S_u) = 1/N, \qquad\qquad \text{for } r \in \{1, ..., N\}.$$

7.1.2 Inclusion probabilities

Theorem 7.4: Under the NSS design, a necessary and sufficient condition for all the sampling units to be distinct is given as $(n-u)k'' \leq N-u$, where $N-u$ and $(n-u)$ represents the population size and the sample size respectively, when selecting sample S''_u.
Proof: The population unit numbers for the sample, corresponding with random start r, is obtained by using (7.9) and (7.10), such that

$$r, r+1, ..., r+u-1, r+u-1+k'', r+u-1+2k'', ..., r+u-1+(n-u)k''.$$

Now, if a sample consists of U_r as the first sampling unit and at most U_{r-1} as the last sampling unit, then it is trivial that all sampling units will be distinct. Thus, by using the rth sample above, we conclude that all sampling units are distinct, if the population unit number of the last sampling unit is less than the population unit number of the first sampling unit, i.e.

$$r+u-1+(n-u)k'' < N+r,$$

where $U_{N+r} = U_r$, as shown above. We can then conclude the proof, since for all values of i and j, it follows that $(i+1) \leq j$ if $i < j$.

Theorem 7.5: Under the NSS design, all second-order inclusion probabilities are non-zero if

(i) $k'' \leq u$;

(ii) $u + (n-u)k'' \geq N/2 + 1$.

Proof: Table 7.1 represents all the possible distances between pairs of population units for sample S_u, i.e. we use the sample in the previous theorem to calculate the distances apart. Now, if we assume $k'' > u$ such that $k'' = u+1$, then the distance u will not reflect in Table 7.1. Similarly, if we let $k'' = u+2$, then the distances u and $u+1$ will not reflect in Table 7.1. Finally, if we substitute any value of k'' such that $k'' \leq u$, then we note that all the possible distances are fully represented in Table 7.1. Moreover, if we conduct CSS

Table 7.1: Distances between pairs of population units, which are selected by conducting NSS with random start r

Population unit numbers	$r+1$	$r+2$	$\dots$	$r+u-1$	$r+u-1+k''$	$r+u-1+2k''$	$\dots$	$r+u-1+(n-u)k''$
r	1	2	$\dots$	$u-1$	$k''+u-1$	$2k''+u-1$	$\dots$	$(n-u)k''+u-1$
$r+1$	0	1	$\dots$	$u-2$	$k''+u-2$	$2k''+u-2$	$\dots$	$(n-u)k''+u-2$
$\vdots$				$\vdots$	$\vdots$	$\vdots$		$\vdots$
$r+u-2$				1	$k''+1$	$2k''+1$	$\dots$	$(n-u)k''+1$
$r+u-1$					k''	$2k''$	$\dots$	$(n-u)k''$
$r+u-1+k''$					0	k''	$\dots$	$(n-u-1)k''$
$r+u-1+2k''$						0	$\dots$	$(n-u-2)k''$
$\vdots$								$\vdots$
$r+u-1+(n-u-1)k''$								k''
$r+u-1+(n-u)k''$								0

for a population of size N, then the distance between two sampling units will never exceed $N/2$ (Singh & Singh 1977). Thus, just as in the previous explanation, we only obtain all the possible distances in Table 7.1, if $(n-u)k''+u-1 \geq N/2$, where $(n-u)k''+u-1$ represents the distance between the first sampling unit in (7.9) and the last sampling unit in (7.10), using the circular transversal concept in the previous theorem. We then conclude the proof, since all the possible distances occur at least once. Singh & Singh showed that by using the conditions in this theorem, we obtain a limitation to the sample size, given by

$$n \geq \sqrt{(2N+4)} - 1.$$

The first-order inclusion probability for the population unit U_i in the sample S''_u is given as

$$\pi_{i(S'_u)} = \mathrm{E}\left(I'\right) = \frac{1}{N}\sum_{r=1}^{N} I' = \frac{u}{N}, \tag{7.11}$$

where S'_u in the subscript indicates that we are considering the corresponding sample space and

$$I' = \begin{cases} 1 & \text{if } U_i \in S'_u, \\ 0 & \text{otherwise,} \end{cases}$$

denotes an indicator variable, such that each population unit occurs u times when applying (7.9), for $r = 1, ..., N$. Likewise, the first-order inclusion probability for the population unit U_i in the sample S''_u is given as

$$\pi_{i(S''_u)} = \mathrm{E}\left(I''\right) = \frac{1}{N}\sum_{r=1}^{N} I'' = \frac{n-u}{N}, \tag{7.12}$$

where

$$I'' = \begin{cases} 1 & \text{if } U_i \in S''_u, \\ 0 & \text{otherwise,} \end{cases}$$

denotes an indicator variable, such that each population unit occurs $(n-u)$ times when applying (7.10), for $r = 1, ..., N$, i.e. we use CSS where the sample size is $(n-u)$ (refer to Section 2.2.3 by replacing n with $(n-u)$ in π_i). The first-order inclusion probability for the population unit U_i in the sample S_u is then obtained by using (7.11) and (7.12), such that

$$\pi_i = \pi_{i(S'_u)} + \pi_{i(S''_u)} = \frac{n}{N},$$

since S'_u and S''_u are mutually exclusive. Exact values for the π_{ij}s are calculated for various cases by Singh & Singh (1977).

7.1.3 Efficiency comparisons

If $\overline{y}_{NSS}$ denotes the sample mean when conducting NSS, then the variance of $\overline{y}_{NSS}$ can be written as

$$\mathrm{Var}\,(\overline{y}_{NSS}) = \frac{N-1}{N}S_Y^2 - \frac{1}{Nn}\sum_{i=1}^{N}\sum_{j=1}^{n}(y_{ij}-\overline{y}_i)^2 \tag{7.13}$$

(Singh & Singh 1977). By using (7.13), Singh & Singh compared the efficiency of their design, to that of LSS and SRSWOR, under different population structures. A summary of the results is given as follows:

(i) For a population in random order, $\sigma^2_{NSS} = \sigma^2_{LSS} = \sigma^2_{SRSWOR}$, where σ^2_{NSS} denotes the expected sampling variance, when conducting NSS, under the super-population model in (4.12).

(ii) If k is an odd multiple of half the period, $n-u$ is even and $u = 2k''$, then the means of the y_i's associated with samples S'_u and S''_u are equivalent to $\overline{Y}$ and hence $\mathrm{Var}(\overline{y}_{NSS}) = 0$. NSS is thus expected to be more efficient than LSS for periodic populations, since the range of the extreme values for $\mathrm{Var}(\overline{y}_{NSS})$ is less than that of $\mathrm{Var}(\overline{y}_{LSS})$. Also, since (7.9) is defined as a continuous selection of u sampling units, we obtain greater variation between the sampling units, which in turn results in efficiency gains over SRSWOR, of size u.

(iii) For both auto-correlated populations and populations that exhibit linear trend, NSS is less efficient than LSS, but more efficient than SRSWOR, provided that u is small and k'' is large. We are thus presented with a trade-off between an unbiased estimate of the sampling variance and efficient estimation of the population mean $\overline{Y}$.

7.2 New Partially Systematic Sampling

7.2.1 Methodology

Leu & Tsui (1996) adopted a modified approach to NSS, termed as NPSS. The corresponding methodology to select a sample of size n from a population of size N is given as follows:

(i) Randomly select an integer t from the interval $[1, N]$.

(ii) If $k = N/n$ is an integer, then let $u = 2$, otherwise define k as the closest integer to $N/(n-1)$ and let u be an integer, where $2 \leq u \leq \text{INT}(n/2) + 1$. Furthermore, let $a = N - (n-u)k$.

(iii) Select a sample S'_t, of size u, from the sample space $S_T = \{U_t, U_{t+1}, ..., U_{t+a-1}\}$, using SRSWOR.

(iv) The remaining $n-u$ sampling units, selected for the sample S''_t, will be those elements with population unit numbers given by

$$t + a - 1 + ik, \qquad \text{for } i = 1, ..., n-u. \tag{7.14}$$

(v) The new partially systematic sample is then defined by $S_t{=}S'_t \cup S''_t$.

It should be noted that, just as in the previous design, we obtain $U_{N+t} = U_t$.

Leu & Tsui (1996) then used the theory of an optimal choice of the sampling interval, given by Bellhouse (1984), to ensure an even spread of the sample. If we let $k_1 = \text{INT}[N/(n-1)]$ and $k_2 = \text{INT}(N/n) + 1$, then the choices for parameters k and u are given as follows:

(i) if $k_1 = 1$, then select $u = \text{INT}(n/2)$ and $k = 1$;

(ii) if $k = N/n$ is an integer, then select $k = N/n$ and $u = 2$. Otherwise, select k as follows:

   (a) if $k_1 \geq 2$ and $k_1 \geq k_2$, then select $k = k_1$ and $u = 2$;

   (b) if $k_1 \geq 2$ and $k_1 < k_2$, then let u be a minimal integer which satisfies $(u-1)k_2 \geq n$, such that if $u \geq k_2$, then select $k = k_1$ and if $u < k_2$, then select k as either k_1 or k_2, subject to which one is closer to a/u.

These recommendations of the choices of u and k, along with the restrictions in the next theorem, will not result in any limitations on the sample size, which results in a comparative advantage of NPSS over NSS.

7.2.2 Inclusion probabilities

Under NPSS, we obtain n distinct sampling, since $N = a + (n-u)k$ results in only one circular transversal. As a result, $\pi_i > 0$, for all $i \in \{1, ..., N\}$.

Theorem 7.6: Under the NPSS design, all second-order inclusion probabilities are non-zero if

$$\text{(a)}\;\; u \geq 2 \qquad \text{and} \qquad \text{(b)}\;\; a \geq k.$$

The proof to this theorem is given by Leu & Tsui (1996).

Any fixed population unit U_i in the sample space $S_T = \{U_t, U_{t+1}, ..., U_{t+a-1}\}$, will occur a times, for $t = 1, ..., N$. Furthermore, since there are N possible values for t in the sample space S_T, we obtain

$$\pi_{i(S_t')} = \frac{1}{N} \times a\mathrm{P}\left(U_i \in S_T | i \in \{t, t+1, \ldots, t+a-1\}\right) = \left(\frac{a}{N}\right)\frac{u}{a} = \frac{u}{N}, \qquad (7.15)$$

as we are selecting u sampling units from the a units in S_T, using SRSWOR. The first-order inclusion probability for the population unit U_i in the sample S_t'' is given by

$$\pi_{i(S_t'')} = \mathrm{E}\left(I_t''\right) = \frac{1}{N}\sum_{i=1}^{N} I_t'' = \frac{n-u}{N}, \qquad (7.16)$$

where

$$I_t'' = \begin{cases} 1 & \text{if } U_i \in S_t'', \\ 0 & \text{otherwise,} \end{cases}$$

represents an indicator variable, such that each population unit occurs $(n-u)$ times when applying (7.13), for $t = 1, ..., N$, i.e. we use CSS where the sample size is $(n-u)$ (refer to Section 2.2.3 by replacing n with $(n-u)$ in π_i). The first-order inclusion probability for the population unit U_i in the sample S_t is then obtained by using (7.15) and (7.16), i.e.

$$\pi_i = \pi_{i(S_t')} + \pi_{i(S_t'')} = \frac{n}{N},$$

since S_t' and S_t'' are mutually exclusive. Exact values for the π_{ij}'s are calculated for various cases by Leu & Tsui (1996).

7.2.3 Efficiency comparisons

For $t = 1, ..., N$, Leu & Tsui (1996) used the following notation:

$$\overline{y}_{ut} = \text{the mean of the } y_i\text{'s for the sample } S_t',$$

$$\overline{y}_{st} = \text{the mean of the } y_i\text{'s for the sample } S_t'',$$

$$\overline{y}_{\cdot t} = \sum_{i=1}^{a-1} y_{it}/a, \text{where } y_{it} = y_{t+i},$$

$$\overline{y}_{NPSS} = [u\overline{y}_{at} + (n-u)\overline{y}_{st}]/n = \text{NPSS mean}.$$

Accordingly, the variance of $\overline{y}_{NPSS}$ can be written as

$$\text{Var}\left(\overline{y}_{NPSS}\right) = \frac{u}{n^2 N}\left(1-\frac{u}{a}\right)\sum_{t=1}^{N}\frac{\sum\limits_{i=0}^{a-1}\left(y_{it}-\overline{y}_{\cdot t}\right)^2}{a-1} + \frac{1}{N}\sum_{t=1}^{N}\left[\frac{u\overline{y}_{\cdot t}+(n-u)\,\overline{y}_{st}}{n}-\overline{Y}\right]^2 \tag{7.17}$$

(Leu & Tsui 1996). By using (7.17), Leu & Tsui compared the efficiency of their design to that of LSS and SRSWOR, under different population structures. Under the super-population model in (4.12), if we denote the expected variance of $\overline{y}_{NPSS}$ as σ^2_{NPSS}, then a summary of Leu & Tsui's results are given follows:

(i) For a population in random order, $\sigma^2_{NPSS} = \sigma^2_{LSS} = \sigma^2_{SRSWOR}$.

(ii) For the model in (4.5): $\text{Var}(\overline{y}_{NPSS}) > \text{Var}(\overline{y}_{LSS})$; $\text{Var}(\overline{y}_{NPSS}) < \text{Var}(\overline{y}_{SRSWOR})$ when $n > 2$; and $\text{Var}(\overline{y}_{NPSS}) = \text{Var}(\overline{y}_{SRSWOR})$ when $n = 2$.

(iii) If k is an integral multiple of the period, then NPSS is more efficient than LSS for periodic populations, since we are selecting $u \geq 2$ units from the from the sample space S_T, allowing for NPSS to offer more variation within the corresponding samples. If k is an odd multiple of half the period and if the y_i's of the u units selected are the same as that of the linear systematic sample, then $\text{Var}(\overline{y}_{NPSS}) = 0$. NPSS is thus expected to be on average more efficient than LSS for periodic populations, since the range of the extreme values for $\text{Var}(\overline{y}_{NPSS})$ is less than that of $\text{Var}(\overline{y}_{LSS})$.

(iv) For auto-correlated populations in the form of a linear, exponential or hyperbolic correlogram, we obtain $\sigma^2_{LSS} < \sigma^2_{NPSS} < \sigma^2_{SRSWOR}$.

We are thus presented with a trade-off between an unbiased estimate of the sampling variance and efficient estimation of $\overline{Y}$, when considering auto-correlated populations and populations that exhibit linear trend.

Leu & Tsui (1996) further compared the efficiency of their design to that of MLSS, for linear trend populations and auto-correlated populations. By denoting the number of random starts by l, such that $N = nk$ and $n = ml$, they concluded that:

(i) For the model in (4.5):

  (a) $\text{Var}(\overline{y}_{NPSS}) = \text{Var}(\overline{y}^{(l)}_{MLSS})$, when $l = 2$ and $m = 1$;

  (b) $\text{Var}(\overline{y}_{NPSS}) < \text{Var}(\overline{y}^{(l)}_{MLSS})$, for all other cases, except when $l = 2$ and $m < 5$, as well as the case when $l \geq 3$ and $m = 1$.

(ii) For auto-correlated populations:

(a) $\text{Var}(\overline{y}_{NPSS}) < \text{Var}(\overline{y}_{MLSS}^{(l)})$, under the exponential correlogram;

(b) $\text{Var}(\overline{y}_{NPSS}) < \text{Var}(\overline{y}_{MLSS}^{(l)})$, under the hyperbolic correlogram, except when $n = 4$, which results in $\text{Var}(\overline{y}_{NPSS}) > \text{Var}(\overline{y}_{MLSS}^{(l)})$;

(c) $\text{Var}(\overline{y}_{NPSS}) < \text{Var}(\overline{y}_{MLSS}^{(l)})$, under the linear correlogram, except when $l = 2$ and $n = 4$ and 6, which results in $\text{Var}(\overline{y}_{NPSS}) > \text{Var}(\overline{y}_{MLSS}^{(l)})$.

We thus conclude that NPSS is more often than not, more efficient than MLSS for auto-correlated populations and populations that exhibit linear trend.

7.3 Balanced Random Sampling

7.3.1 Methodology

A design which adopts the advantages of both SRSWOR and MSS is known as BRS and is originally attributed to Singh & Garg (1979). BRS divides the population into two groups, before selecting sampling units from the first group using SRSWOR and then pairing these units with units from the second group, using the MSS pairing technique, as discussed in Section 5.4.1. We can thus expect BMRS to perform particularly well for populations that exhibit linear trend.

If we assume that n and N are even, then the methodology of BRS is given as follows:

(i) Select a sample S'_r (of size $n/2$) from the sample space $S' = \{U_1, U_2, ..., U_{N/2}\}$, using SRSWOR, where the population unit numbers of the sampling units are given by

$$r_i, \qquad \text{for } i = 1, ..., n/2.$$

(ii) A dependent sample to S'_r, given by S''_r, is of size $n/2$ and belongs to the sample space $S'' = \{U_{N/2+1}, U_{N/2+2}, ..., U_N\}$, where the population unit numbers of the sampling units are given by

$$N + 1 - r_i, \qquad \text{for } i = 1, ..., n/2.$$

(iii) The balanced random sample is then given as $S_r = S'_r \cup S''_r$.

7.3.2 Inclusion probabilities

The first-order inclusion probability for the unit U_i in the sample S_r' is given as

$$\pi_{i(S_r')} = \begin{cases} 0 & \text{if } U_i \in S'', \\ (n/2)/(N/2) = n/N & \text{if } U_i \in S', \end{cases} \tag{7.18}$$

since we are selecting $n/2$ units from the sample space S', of size $N/2$. Likewise, the first-order inclusion for the unit U_i in the sample S_r'', is given as

$$\pi_{i(S_r'')} = \begin{cases} 0 & \text{if } U_i \in S', \\ (n/2)/(N/2) = n/N & \text{if } U_i \in S''. \end{cases} \tag{7.19}$$

By noting that S_r' and S_r'' are mutually exclusive, we then use (7.18) and (7.19) to obtain the first-order inclusion probability for the unit U_i, for $i \in \{1, ..., N\}$, in the sample S_r, i.e.

$$\pi_{i(S_r)} = \pi_{i(S_r')} + \pi_{i(S_r'')} = \frac{n}{N}.$$

To obtain the second-order inclusion probabilities, we first note that there are two possible cases for the π_{ij}'s in the sample S_r, i.e. (i) $i + j = N + 1$ and (ii) $i + j \neq N + 1$. For case (i), the selection of unit U_i from a sample space is paired with the unit U_j from the other sample space. Thus, the probability of selecting units U_i and U_j, given that U_i and U_j are paired, is equivalent to the probability of selecting U_i, i.e.

$$\pi_{ij(S_r)} = \pi_{i(S_r)} = \frac{n}{N}. \tag{7.20}$$

For case (ii), unit U_i is not paired with unit U_j from the other sample space, thus the probability of selecting the units U_i and U_j, given that U_i and another paired unit are already selected, is given as

$$\begin{aligned} \pi_{ij(S_r)} &= \pi_{i(S_r)} \times \mathrm{P}\left(U_j \text{ is selected} | U_i \text{ and another unit are selected}\right) \\ &= \left(\frac{n}{N}\right) \frac{n-2}{N-2}. \end{aligned} \tag{7.21}$$

The second-order inclusion probability for the pair of units (U_i, U_j) in S_r $(i \neq j)$, is thus obtained using (7.20) and (7.21), such that

$$\pi_{ij(S_r)} = \begin{cases} n/N & \text{if } i + j = N + 1, \\ n(n-2)/N(N-2) & \text{otherwise.} \end{cases}$$

7.3.3 Efficiency comparisons

If we denote $\overline{y}_{BRS}$ as the sample mean when conducting BRS, then the variance of $\overline{y}_{BRS}$ can be written as

$$\operatorname{Var}\left(\overline{y}_{BRS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right) \frac{2}{N-2} \left[\sum_{i=1}^{N} \left(y_i - \overline{Y}\right)^2 - \frac{1}{2} \sum_{i=1}^{N/2} \left(y_i - y_{N+1-i}\right)^2\right] \tag{7.22}$$

(Singh & Garg 1979). By using (7.22), Singh & Garg compared the efficiency of their design to that of LSS, SRSWOR and STR, under different population structures. A summary of the results is given as follows:

(i) For both the model in (4.5) and periodic populations, $\operatorname{Var}(\overline{y}_{BRS}) = 0$ and hence BRS is more efficient than all the designs.

(ii) If we assume a linear correlogram for auto-correlated populations, then BRS is twice as efficient as SRSWOR, but less efficient than both LSS and STR.

Singh & Garg (1979) further provided methodologies, along with the corresponding inclusion probabilities, for other cases of BRS, i.e. *(i)* N and n are both odd; *(ii)* N is odd and n is even; and *(iii)* N is even and n is odd. However, for the purposes of this thesis we shall only consider the case discussed in this section.

7.4 Balanced Modified Random Sampling

7.4.1 Methodology

The author next proposes a new design, which adopts the advantages of both SRSWOR and MSS, termed as BMRS. BMRS divides the population into $n/4$ groups, before using a MSS pairing technique within each group. We then select two pairs within each group, using SRSWOR. We can thus expect BMRS to perform particularly well for populations that exhibit linear trend. For this design, we will assume that k is an integer and $n/2$ is an even integer.

The methodology to select a sample of size n from a population of size N, using BMRS, is given as follows:

(i) Select $n/4$ pairs of integers $\left(\{i_1,\ i_2\}, \{i_3,\ i_4\}, ..., \{i_{n/2-1},\ i_{n/2}\}\right)$ from the first $2k$ integers, using an independent SRSWOR selection for each pair, i.e. $i_{2s-1} \neq i_{2s}$, for $s = 1, ..., n/4$.

(ii) The randomly selected sample, S_i, will contain those units with population unit numbers given by

$$i_{2s-1} + 4(s-1)k, \qquad i_{2s} + 4(s-1)k, \qquad \text{for } s = 1, ..., n/4.$$

(iii) The dependent sample containing the paired sampling units, S_j, will be those units with population unit numbers given by

$$4k - i_{2s-1} + 4(s-1)k + 1, \quad 4k - i_{2s} + 4(s-1)k + 1, \quad \text{for } s = 1, ..., n/4.$$

(iv) The balanced modified random sample is then given as $S_s = S_i \cup S_j$.

Table 7.2 depicts the above-mentioned design, whereby the population is divided into $n/4$ groups, each containing $2k$ pairs of population units and SRSWOR is applied within each group to select two pairs of units, which collectively represent a balanced modified random sample. From Table 7.2, we can easily verify that BMRS reduces to BRS if $n = 4$, i.e. we only consider group 1 and replace $4k$ with N.

Table 7.2: Pairs of population units for the BMRS design

Group 1	Group 2	...	Group $s = n/4$
$\{U_1,\ U_{4k}\}$	$\{U_{4k+1},\ U_{8k}\}$	...	$\{U_{(n-4)k+1},\ U_{nk}\}$
$\{U_2,\ U_{4k-1}\}$	$\{U_{4k+2},\ U_{8k-1}\}$	...	$\{U_{(n-4)k+2},\ U_{nk-1}\}$
$\vdots$	$\vdots$		$\vdots$
$\{U_{2k},\ U_{2k+1}\}$	$\{U_{6k},\ U_{6k+1}\}$	...	$\{U_{(n-2)k},\ U_{(n-2)k+1}\}$

7.4.2 Inclusion probabilities

Theorem 7.7: Under BMRS, the first-order inclusion probability of the population unit U_i, for all $i = 1, ..., N$, is given by $\pi_i = 1/k$.

Proof: Under BMRS, each group contains $2k$ pairs of units and thus the selection of two distinct pairs within each group is equivalent to applying SRSWOR to select two units from a population of size $2k$, i.e. $\pi_i = 2/2k = 1/k$.

Theorem 7.8: Under BMRS, the second-order inclusion probabilities for the pair of population units $\{U_i,\ U_j\}$, for all $i, j = 1, ..., N$ $(j \neq i)$ and $s = 1, ..., n/4$, are given by

$$\pi_{ij} = \begin{cases} 1/k & \text{if } i + j = 4k + 8k(s-1) + 1, \\ 1/k(2k-1) & \text{if } i, j \in \{1 + 4k(s-1), ..., 4ks\} \text{ and } i + j \neq 4k + 8k(s-1) + 1, \\ 1/k^2 & \text{otherwise.} \end{cases}$$

Proof: From Table 7.2, if we assume that U_i and U_j are paired, then the sum of their population unit numbers is given by $i + j = 4k + 8k(s-1) + 1$, for all $i, j = 1, ..., N$ $(j \neq i)$ and $s = 1, ..., n/4$. For this case, the probability of selecting units $\{U_i,\ U_j\}$ is equivalent to the probability of only selecting unit U_i, i.e. $\pi_{ij} = \pi_i = 1/k$. Furthermore, if we assume units U_i and U_j are not paired, but to belong to the same group (i.e. i and $j \in \{1 + 4k(s-1), ..., 4ks\}$ and $i + j \neq 4k + 8k(s-1) + 1$ for all $i, j = 1, ..., N$ $(j \neq i)$, where $s = 1, ..., n/4$), then

$$\pi_{ij} = \pi_i \times \mathrm{P}\left(U_j \text{ is selected}|U_i \text{ is selected}\right) = \frac{1}{k}\left(\frac{1}{2k-1}\right),$$

since we are selecting one pair from the remaining $2k-1$ pairs, after selecting the unit U_i. The final possibility of selecting $\{U_i,\ U_j\}$ for the sample is that U_i and U_j belong to different groups, such that

$$\pi_{ij} = \pi_i \times \mathrm{P}\left(U_j \text{ is selected}|U_i \text{ is selected}\right) = \frac{1}{k}\left(\frac{1}{k}\right) = \frac{1}{k^2},$$

since the selection of unit U_i is independent from the selection of unit U_j.

7.4.3 Efficiency comparisons

Theorem 7.9: Under BMRS, if we denote the sample mean by $\overline{y}_{BMRS}$, then the variance of $\overline{y}_{BMRS}$ can be written as

$$\mathrm{Var}\left(\overline{y}_{BMRS}\right) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{4}{n\,(2k-1)}\sum_{s=1}^{n/4}\left[S_t\,(s) - S_w\,(s)\right], \tag{7.23}$$

where

$S_w(s) = \sum_{i=1}^{4k}(y_{i+4k(s-1)} - y_{4k+1-i+4k(s-1)})^2/4 =$ the sum of squares within the pairs of units for group s,

$S_t(s) = \sum_{i=1}^{4k}\left(y_{i+4k(s-1)} - \overline{Y}_s\right)^2 =$ the total sum of squares within group s and

$\overline{Y}_s = \sum_{i=1}^{4k} y_{i+4k(s-1)}/4k =$ the mean of group s.

Proof: Under our proposed design, we can view each group from Table 7.2 as a stratum. For this situation, we denote $n_s = 4$ as the number of sampling units selected from stratum s and $N_s = 4k$ as the size of stratum s. Furthermore, if we denote the variance within stratum s by σ_s^2, then the sampling variance, when conducting STR (based on the selection of n_s sampling units from stratum s, of size N_s), is given by

$$\operatorname{Var}\left(\overline{y}_{STR}\right) = \frac{1}{N^2}\sum_{s=1}^{n/4}\frac{N_s\left(N_s - n_s\right)\sigma_s^2}{n_s}, \tag{7.24}$$

where $N = \sum_{s=1}^{n/4} N_s$ and $n = \sum_{s=1}^{n/4} n_s$. By applying the relevant substitutions, we obtain

$$\frac{N_s\left(N_s - n_s\right)}{n_s N^2} = \frac{4k\left(4k-4\right)}{4N^2} = 4\left(\frac{k^2}{N^2} - \frac{k}{N^2}\right) = \frac{4}{n}\left(\frac{1}{n} - \frac{1}{N}\right). \tag{7.25}$$

Let us now consider the selection of pairs of units within each stratum, of size $2k$ pairs, such that the variance within each stratum is equivalent to the mean sum of squares between the pairs, i.e.

$$\sigma_s^2 = \frac{1}{2k-1}S_b\left(s\right) = \frac{1}{2k-1}\left[S_t\left(s\right) - S_w\left(s\right)\right], \tag{7.26}$$

where $S_b(s)$ denotes the sum of squares between the pairs of units within stratum s. We then conclude the proof by substituting (7.25) and (7.26) into (7.24).

We will next use (7.23) to compare BMRS to SRSWOR, LSS and STR (based on the selection of one unit per stratum), under different population structures.

Population in random order

For the model in (6.24), if we let $\overline{e}_s = \sum_{i=1}^{4k} e_{i+4k(s-1)}/4k$, then we obtain

$$\begin{aligned}\mathrm{E}_m\left[S_t\left(s\right)\right] &= \sum_{i=1}^{4k}\mathrm{E}_m\left[\mu + e_{i+4k(s-1)} - \left(\mu + \overline{e}_s\right)\right]^2 \\ &= \sum_{i=1}^{4k}\mathrm{E}_m\left[e_{i+4k(s-1)} - \overline{e}_s\right]^2 = \sum_{i=1}^{4k}\left(\sigma^2 - \frac{\sigma^2}{4k}\right) = 4k\sigma^2 - \sigma^2,\end{aligned} \tag{7.27}$$

which follows since $e_{i+4k(s-1)}$ occurs once in e_s , for $i \in \{1, ..., 4k\}$ and $s \in \{1, ..., n/4\}$. Also,

$$\begin{aligned}\mathrm{E}_m\left[S_w\left(s\right)\right] &= \frac{1}{4}\sum_{i=1}^{4k}\mathrm{E}_m\left[\mu + e_{i+4k(s-1)} - \left(\mu + e_{4k+1-i+4k(s-1)}\right)\right]^2 \\ &= \frac{1}{4}\sum_{i=1}^{4k}\left[\mathrm{E}_m\left(e_{i+4k(s-1)}^2\right) + \mathrm{E}_m\left(e_{4k+1-i+4k(s-1)}^2\right)\right] = 2k\sigma^2,\end{aligned} \tag{7.28}$$

since $i+4k(s-1) \neq 4k+1-i+4k(s-1)$. We then obtain the expected variance of $\overline{y}_{BMRS}$, by using (7.23), (7.27) and (7.28), i.e.

$$\begin{aligned}\sigma^2_{BMRS} &= \left(\frac{1}{n}-\frac{1}{N}\right)\frac{4}{n\left(2k-1\right)}\sum_{s=1}^{n/4}\left(4k\sigma^2-\sigma^2-2k\sigma^2\right)\\ &= \left(\frac{1}{n}-\frac{1}{N}\right)\frac{4}{n\left(2k-1\right)}\sum_{s=1}^{n/4}\sigma^2\left(2k-1\right) = \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2. \qquad (7.29)\end{aligned}$$

The expected variances of $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$ are known to be equal to (7.29). We thus conclude that BMRS is as equally efficient to SRSWOR, LSS and STR, for populations in random order.

Population with linear trend

If we assume the model in (5.1), then

$$\overline{Y}_s = \frac{1}{4k}\sum_{i=1}^{4k}\left[a+b\left\{i+4k\left(s-1\right)\right\}+e_{i+4k(s-1)}\right] = a+\frac{b\left(4k+1\right)}{2}+4bk\left(s-1\right)+\overline{e}_s.$$

Consequently, we obtain

$$\begin{aligned}\mathrm{E}_m\left[S_t\left(s\right)\right] &= \sum_{i=1}^{4k}\mathrm{E}_m\left[b\left\{i+4k\left(s-1\right)\right\}+e_{i+4k(s-1)}-\left\{\frac{b\left(4k+1\right)}{2}+4bk\left(s-1\right)+\overline{e}_s\right\}\right]^2\\ &= \mathrm{E}_m\left[b^2\sum_{i=1}^{4k}\left\{i-\frac{\left(4k+1\right)}{2}\right\}^2\right]+\sum_{i=1}^{4k}\mathrm{E}_m\left[e_{i+4k(s-1)}-\overline{e}_s\right]^2\\ &= \mathrm{E}_m\left[b^2\sum_{i=1}^{4k}\left\{i-\frac{\left(4k+1\right)}{2}\right\}^2\right]+4k\sigma^2-\sigma^2, \qquad (7.30)\end{aligned}$$

which follows from (7.27). Moreover,

$$\begin{aligned}\mathrm{E}_m\left[S_w\left(s\right)\right] &= \frac{1}{4}\sum_{i=1}^{4k}\mathrm{E}_m\Bigg[b\left(i+4k\left\{s-1\right\}\right)+e_{i+4k(s-1)}\\ &\quad -\left(b\left\{4k+1-i+4k\left(s-1\right)\right\}+e_{4k+1-i+4k(s-1)}\right)\Bigg]^2\\ &= \mathrm{E}_m\left[\frac{b^2}{4}\sum_{i=1}^{4k}\left(2i-4k-1\right)^2\right]+\frac{1}{4}\sum_{i=1}^{4k}\left[\mathrm{E}_m\left(e^2_{i+4k(s-1)}\right)+\mathrm{E}_m\left(e^2_{4k+1-i+4k(s-1)}\right)\right]\\ &= \mathrm{E}_m\left[b^2\sum_{i=1}^{4k}\left\{i-\frac{\left(4k+1\right)}{2}\right\}^2\right]+2k\sigma^2 \qquad (7.31)\end{aligned}$$

By using (7.23), (7.29), (7.30) and (7.31), we obtain

$$\sigma^2_{BMRS} = \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2. \qquad (7.32)$$

Thus, by comparing (7.32) to (5.9), (5.10) and (5.11) (i.e. $\overline{y}_{SRSWOR}$, $\overline{y}_{LSS}$ and $\overline{y}_{STR}$ are all unbiased estimators of the population mean), we conclude that BMRS is more efficient than SRSWOR, LSS and STR, for populations that exhibit linear trend.

Periodic populations

Let us assume that y_i exhibits an exact periodic function with period $2h$ and $N = 2Qh$. Under this assumption, we then note that each group in Table 7.2 is equivalent for the two cases of $k = ah$, for $a \in \{1, 3, ...\}$, and $k = 2ah$, for $a \in \{1, 2, ...\}$. Accordingly, we obtain

$$S_t(1) = S_t(2) = ... = S_t(n/4), \quad S_w(1) = S_w(2) = ... = S_w(n/4) \quad \text{and} \quad \overline{Y}_s = \overline{Y},$$

Thus, if we only consider the first group in Table 7.2, then (7.23) reduces to

$$\begin{aligned} \text{Var}\left(\overline{y}_{BMRS}\right) &= \left(\frac{1}{4} - \frac{1}{4k}\right)\frac{1}{2k-1}\left[\sum_{i=1}^{4k}\left(y_i - \overline{Y}\right)^2 - \frac{1}{4}\sum_{i=1}^{4k}\left(y_i - y_{4k+1-i}\right)^2\right] \\ &= \frac{k-1}{2k-1}\left[\sigma^2 - \frac{1}{16k}\sum_{i=1}^{4k}\left(y_i - y_{4k+1-i}\right)^2\right]. \end{aligned} \tag{7.33}$$

If we suppose the best case scenario for LSS, i.e. $k = ah$, for $a \in \{1, 3, ...\}$, then it is known that $\text{Var}(\overline{y}_{LSS}) = 0$ (see Section 4.2.3) and by comparing this result to (7.33), we thus conclude that LSS is more efficient that BMRS for this situation. On the other hand, if we suppose the worst case scenario for LSS, i.e. $k = 2ah$, for $a \in \{1, 2, ...\}$, then $\text{Var}(\overline{y}_{LSS}) = \sigma^2$ (see Section 6.3.3) and by comparing this result to (7.33), we thus conclude that BMRS is more efficient than LSS for this situation. When comparing (7.33) to either (4.2) or (4.3), we note that simple theoretical comparisons are difficult to deduce and we will thus resort to some numerical comparisons in the next chapter.

It should be noted that the author is in disagreement with the corresponding results obtained in Singh and Garg's (1979) paper. Accordingly, the latter authors deduced that $\text{Var}(\overline{y}_{BRS}) = 0$ under the exact periodic function and this is incorrect, since BMRS reduces to BRS when $n = 4$ and by referring to (3.7), (4.15) and (7.33), we obtain

$$\sigma^2 = \frac{1}{8k}\sum_{i=1}^{4k}\sum_{j \neq i}^{4k}\left(y_i - y_j\right)^2 = \frac{2}{16k}\sum_{i=1}^{4k}\sum_{j \neq i}^{4k}\left(y_i - y_j\right)^2 > \frac{1}{16k}\sum_{i=1}^{4k}\left(y_i - y_{4k+1-i}\right)^2,$$

which follows since $(y_i - y_{4k+1-i})$ is the minimum difference between any pair of units, under the exact periodic function. Further evidence of this is provided in the next chapter. A possible explanation for this is that Singh & Garg may have assumed that a perfect linear trend pattern is repeated after every set of $2h$ units in the population, rather than variate values which are monotonically increasing and monotonically decreasing at every set of $2h$ units in the population. One can easily verify that this assumption results in $\text{Var}(\overline{y}_{BRS}) = \text{Var}(\overline{y}_{BMRS}) = 0$.

Auto-correlated population

Under the model in (4.12), we obtain

$$\text{E}_m\left[S_t\left(s\right)\right] = \left(4k-1\right)\sigma^2\left[1-\frac{2}{4k\left(4k-1\right)}\sum_{u=1}^{4k-1}\left(4k-u\right)\rho_u\right], \tag{7.34}$$

which follows if we replace N in (4.16) by $4k$, since there are $4k$ population units within group s. Moreover, by using (4.15) and (4.16), it follows that

$$\begin{aligned}\text{E}_m\left[S_w\left(s\right)\right] &= \frac{1}{4}\text{E}_m\left[\sum_{i=1}^{4k}\left(y_{i+4k(s-1)}-y_{4k+1-i+4k(s-1)}\right)^2\right]\\ &= \frac{1}{4}\sum_{i=1}^{4k}\left[2\sigma^2-2\sigma^2\rho_{|4k+1-2i|}\right]\\ &= \frac{\sigma^2}{2}\left[4k-\sum_{i=1}^{4k}\rho_{|4k+1-2i|}\right]\\ &= 2k\sigma^2\left[1-\frac{1}{4k}\sum_{i=1}^{4k}\rho_{|4k+1-2i|}\right].\end{aligned} \tag{7.35}$$

Now, on substituting (7.34) and (7.35) into (7.23), it follows that

$$\sigma^2_{BMRS} = \left(\frac{1}{n}-\frac{1}{N}\right)\sigma^2\left[1-\frac{1}{2k\left(2k-1\right)}\sum_{u=1}^{4k-1}\left(4k-u\right)\rho_u+\frac{1}{2\left(2k-1\right)}\sum_{i=1}^{4k}\rho_{|4k+1-2i|}\right].$$

By comparing this result to the related expected variance expressions in Section 4.2.4, we conclude that it is difficult to obtain simple theoretical comparisons and we will thus resort to some numerical comparisons in the next chapter.

Chapter 8

NUMERICAL ANALYSIS

This chapter focuses on numerically comparing the various designs presented in this thesis, under various population structures. The populations under consideration will be artificial populations, which exhibit the structures given in Section 4.2, as well as a natural population. To obtain the artificial populations, we apply Monte Carlo simulations by using the statistical software package R. For each artificial population model, we will simulate $G = 1000$ finite populations, each of size $N = 120$. By using the simple mathematical model in (6.1), we present the specifications for the artificial populations in Table 8.1. The final population, of size $N = 176$, is the strip-wise complete enumeration on length and timber volume for ten blocks of the blacks mountain experimental forest, studied by Hasel (1942). The variable of interest (Y) is the total amount of timber volume and P6A represents the population in its natural state (i.e. arranged as in the frame), while P6B is a result of the population arranged according to the length of the strips. These arrangements are depicted in Figures 8.1 and 8.2. From Figure 8.1, we note that P6A is approximately a stratified population, since the blocks in which the strips occur can be considered as strata. By arranging the population in ascending order according to an auxiliary variable, which in this case is the length of the strips, we obtain a population that approximately exhibits linear trend, P6B, represented by Figure 8.2.

We will next compare all the sampling designs discussed in this thesis, for the above-mentioned populations. The comparative measures that will be used are the MSEs of the corresponding sample means and the corresponding percentages of CIs which contain the true population mean. For the artificial populations, we will average these comparative measures over the 1000 populations, so as to obtain their respective expected quantities. The CL used for the intervals will be 95% (nominal rate), with the classic notion that the

Table 8.1: Specifications for the Artificial Populations

Population	Description	Trend Component (μ_{ij})	Error Component (e_{ij})
P1	Random	0	iid $N(0, 100)$
P2	Linear Trend	$10i$	iid $N(0, 100)$
P3	Periodic	$10\sin[\pi/2 \times (i + \{j-1\}k)]$	iid $N(0, 1)$
P4	Stratified	j	iid $N(0, 1)$
P5	Auto-correlated	0	$e_{ij} = \rho e_{i-1,j} + \epsilon_{ij}$ $e_{11} \sim N(0, 100/(1-\rho^2))$ ϵ_{ij} iid $N(0, 100)$ $\rho = 0.6$

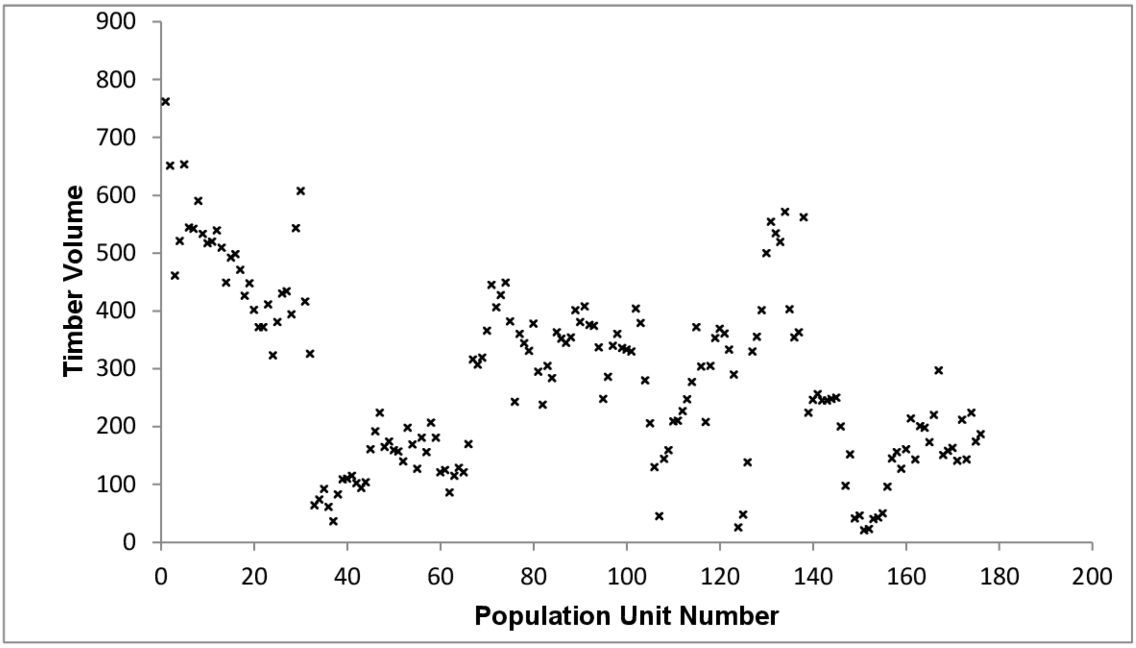


Figure 8.1: Timber volume for ten blocks of the blacks mountain experimental forest, arranged according to the frame

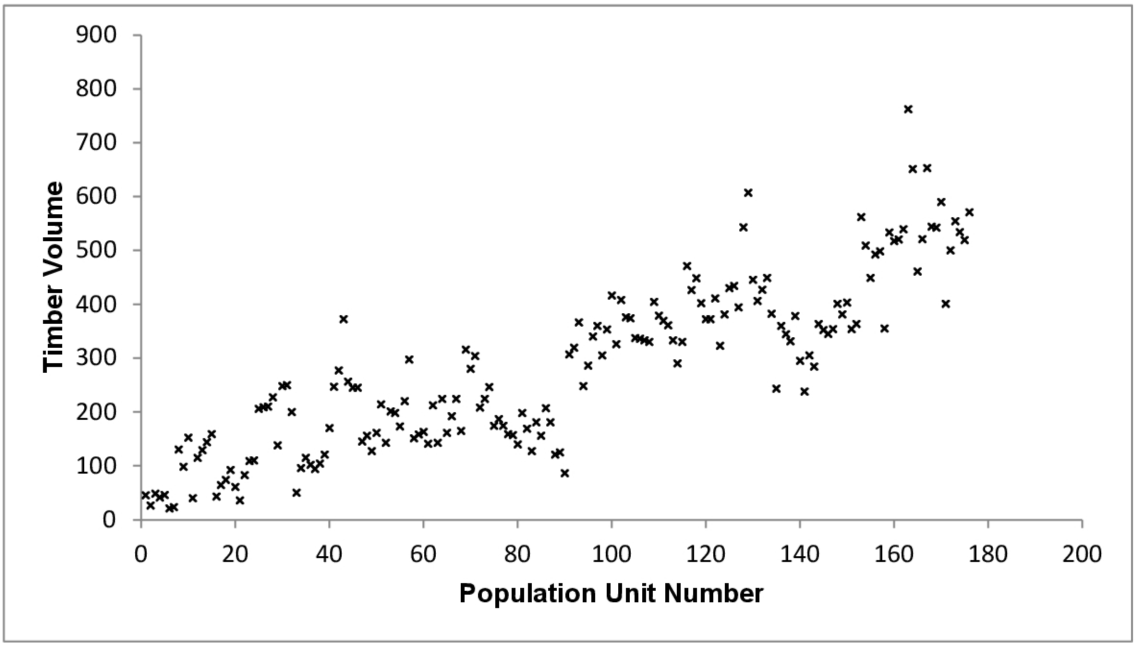


Figure 8.2: Timber volume for ten blocks of the blacks mountain experimental forest, arranged according to the length of strips

better estimate is the one which exhibits a higher expected percentage and estimates which exhibit an expected percentage lower than the nominal rate are considered undesirable. If the design is not applicable for the specific sample size, then we will denote this in the analysis by N/A, i.e. if $n = 7$, then $k = N/n$ is not an integer, which results in LSS, YEC, CESS, BSS, MSS, BMSS MLSS, MBMSS, BRS and BMRS being inapplicable.

The expected MSEs of the sample means, related to the various sampling designs discussed in this thesis, are presented in Tables 8.2 to 8.8. For the multiple-start designs, we consider the number of random starts as $m = 2$ and 3. For the PSS design, we let n_1 be a maximum value, such that the sampling interval, given by N/n_1, is an integer. Furthermore, we consider an unweighted average of the corresponding sample means, i.e. $\beta = 1/2$. For NSS and NPSS, we let $u = 2$, which results in a maximum number of sampling units obtained using the systematic selection procedure.

Table 8.2: Expected Mean Square Errors of the Sample Means, for all the Designs, under P1

Estimator	k is an integer					k is not an integer			
	$n = 4$	$n = 8$	$n = 12$	$n = 20$	$n = 40$	$n = 7$	$n = 14$	$n = 21$	$n = 35$
$\overline{y}_{LSS}$	24.1171	11.8282	7.3189	4.0130	1.5706	N/A	N/A	N/A	N/A
$\overline{y}_{CSS}$	24.1171	11.8282	7.3189	4.0130	1.5706	13.4561	6.2936	3.8407	2.0531
$\overline{y}_{SRSWR}$	24.7135	12.3689	8.2131	4.9590	2.4829	14.0995	7.1275	4.6956	2.8308
$\overline{y}_{SRSWOR}$	24.0905	11.6414	7.4539	4.1672	1.6692	13.3886	6.3489	3.9064	2.0220
$\overline{y}_{STR}$	24.0978	11.6333	7.4496	4.1533	1.6690	N/A	N/A	N/A	N/A
$\overline{y}_{YEC}$	25.8936	12.1568	7.4464	4.0525	1.5798	N/A	N/A	N/A	N/A
$\overline{y}_{CESS}$	23.1768	11.5978	7.0436	4.1361	1.7106	N/A	N/A	N/A	N/A
$\overline{y}_{BSS}$	24.1830	11.7578	7.3670	4.1693	1.6494	N/A	N/A	N/A	N/A
$\overline{y}_{MSS}$	24.1140	11.6340	7.4864	4.1096	1.6992	N/A	N/A	N/A	N/A
$\overline{y}_{BMSS}$	24.1664	11.7268	7.5624	4.0420	1.7294	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MLSS}$	24.0590	11.7560	7.4168	4.1747	1.6335	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	7.3795	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	24.0590	11.6287	7.5186	4.1996	1.6405	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	7.4569	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{PSS}$	32.3294	15.8065	14.0645	5.8319	2.5091	28.1821	13.8507	25.2847	4.9966
$\overline{y}_{NSS}$	24.2335	11.7237	7.5253	4.0288	1.5779	13.2634	6.2688	3.9762	1.9931
$\overline{y}_{NPSS}$	24.0926	11.7311	7.3644	4.0402	1.5804	13.3525	6.2798	3.9675	1.9960
$\overline{y}_{BRS}$	24.2161	11.6569	7.4258	4.1522	1.6736	N/A	6.3403	N/A	N/A
$\overline{y}_{BMRS}$	24.2161	11.6195	7.4445	4.1447	1.6642	N/A	N/A	N/A	N/A

Table 8.3: Expected Mean Square Errors of the Sample Means, for all the Designs, under P2

Estimator	k is an integer					k is not an integer			
	$n = 4$	$n = 5$	$n = 8$	$n = 15$	$n = 30$	$n = 7$	$n = 14$	$n = 21$	$n = 35$
$\overline{y}_{LSS}$	7516.81	4815.78	1880.63	530.52	126.96	N/A	N/A	N/A	N/A
$\overline{y}_{CSS}$	7516.81	4815.78	1880.63	530.52	126.96	2457.46	1141.20	2139.22	1963.77
$\overline{y}_{SRSWR}$	30032.96	24015.72	15010.74	8006.26	4003.48	17156.76	8576.90	5717.11	3430.60
$\overline{y}_{SRSWOR}$	29274.85	23208.47	14127.76	7064.35	3027.84	16291.71	7639.93	4756.25	2450.63
$\overline{y}_{STR}$	1897.33	978.44	245.26	40.77	6.65	N/A	N/A	N/A	N/A
$\overline{y}_{YEC}$	26.15	20.17	12.07	5.87	2.45	N/A	N/A	N/A	N/A
$\overline{y}_{CESS}$	47.55	43.96	11.57	30.69	26.56	N/A	N/A	N/A	N/A
$\overline{y}_{BSS}$	24.48	212.37	11.67	8.00	2.48	N/A	N/A	N/A	N/A
$\overline{y}_{MSS}$	24.30	211.70	11.72	8.16	2.38	N/A	N/A	N/A	N/A
$\overline{y}_{BMSS}$	24.08	212.58	11.47	8.03	3.05	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MLSS}$	14771.20	N/A	3629.72	N/A	227.41	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	N/A	1464.67	326.42	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	14771.20	N/A	11.71	N/A	3.56	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	N/A	63.915	15.32	N/A	N/A	N/A	N/A
$\overline{y}_{PSS}$	33198.56	31769.22	15626.34	9984.24	4769.03	30771.55	15032.67	30058.99	5747.58
$\overline{y}_{NSS}$	49581.34	27316.76	7523.39	2048.86	155.95	11062.42	1014.82	543.13	2059.47
$\overline{y}_{NPSS}$	20086.41	12079.56	3914.02	848.92	161.66	6668.22	1150.03	582.98	1815.98
$\overline{y}_{BRS}$	24.38	N/A	11.68	N/A	2.50	N/A	6.30	N/A	N/A
$\overline{y}_{BMRS}$	24.38	N/A	11.62	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.4: Expected Mean Square Errors of the Sample Means, for all the Designs, under P3

Estimator	k is an odd multiple of half the period $(=4)$			k is an integral multiple of the period $(=4)$		
	$k=6$	$k=10$	$k=30$	$k=8$	$k=12$	$k=20$
$\overline{y}_{LSS}$	0.042	0.074	0.243	50.043	50.096	50.168
$\overline{y}_{CSS}$	0.042	0.074	0.243	50.043	50.096	50.168
$\overline{y}_{SRSWR}$	2.555	4.247	12.736	3.398	5.099	8.500
$\overline{y}_{SRSWOR}$	2.147	3.854	12.415	2.999	4.714	8.143
$\overline{y}_{STR}$	2.408	4.155	12.702	3.391	5.192	8.492
$\overline{y}_{YEC}$	0.088	0.211	2.122	50.048	50.096	50.177
$\overline{y}_{CESS}$	0.042	0.072	0.253	0.055	0.093	0.154
$\overline{y}_{BSS}$	25.094	25.060	25.192	25.133	25.077	25.204
$\overline{y}_{MSS}$	0.041	0.076	0.242	25.132	25.076	25.205
$\overline{y}_{BMSS}$	0.041	0.076	0.243	25.131	26.076	27.874
$\overline{y}^{(2)}_{MLSS}$	22.816	23.743	24.793	N/A	24.007	24.522
$\overline{y}^{(3)}_{MLSS}$	N/A	0.075	N/A	15.271	N/A	16.263
$\overline{y}^{(2)}_{MBMSS}$	11.881	13.229	24.793	N/A	12.522	13.711
$\overline{y}^{(3)}_{MBMSS}$	N/A	0.075	N/A	7.956	N/A	16.263
$\overline{y}_{PSS}$	11.621	16.603	24.660	4.183	6.342	24.208
$\overline{y}_{NSS}$	0.292	1.463	12.731	0.279	1.092	2.935
$\overline{y}_{NPSS}$	0.270	0.732	6.380	38.017	33.051	25.092
$\overline{y}_{BRS}$	2.165	3.887	12.506	N/A	4.749	8.223
$\overline{y}_{BMRS}$	2.319	4.021	12.506	N/A	N/A	N/A

Table 8.5: Expected Mean Square Errors of the Sample Means, for all the Designs, under P4

Estimator	$n=4$	$n=8$	$n=12$	$n=15$	$n=20$	$n=24$	$n=40$
$\overline{y}_{LSS}$	0.2410	0.1166	0.0749	0.0595	0.0416	0.0335	0.0178
$\overline{y}_{CSS}$	0.2410	0.1166	0.0749	0.0595	0.0416	0.0335	0.0178
$\overline{y}_{SRSWR}$	0.5588	0.7809	1.0772	1.3116	1.7136	2.0380	3.3540
$\overline{y}_{SRSWOR}$	0.5447	0.7350	0.9776	1.1572	1.4400	1.6441	2.2548
$\overline{y}_{STR}$	0.2408	0.1170	0.0750	0.0583	0.0420	0.0332	0.0167
$\overline{y}_{YEC}$	0.3410	0.2014	0.1589	0.1418	0.1217	0.1150	0.0934
$\overline{y}_{CESS}$	0.2480	0.1185	0.0748	0.0546	0.0434	0.0335	0.0174
$\overline{y}_{BSS}$	0.2382	0.1149	0.0755	0.0593	0.0422	0.0347	0.0179
$\overline{y}_{MSS}$	0.2422	0.1153	0.0760	0.0594	0.0416	0.0342	0.0178
$\overline{y}_{BMSS}$	0.2420	0.1169	0.0759	0.0576	0.0420	0.0336	0.0170
$\overline{y}^{(2)}_{MLSS}$	0.3611	0.2387	0.1924	N/A	0.1553	0.1439	0.1168
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	0.2858	0.2596	N/A	0.2240	N/A
$\overline{y}^{(2)}_{MBMSS}$	0.3611	0.1151	0.0894	N/A	0.0461	0.0339	0.0171
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	0.0748	0.0672	N/A	0.0333	N/A
$\overline{y}_{PSS}$	0.6693	0.8416	1.6440	1.6475	1.6845	2.9900	3.0326
$\overline{y}_{NSS}$	0.7173	0.4489	0.2598	0.3775	0.0811	0.0655	0.0339
$\overline{y}_{NPSS}$	0.4332	0.2410	0.1613	0.1287	0.0928	0.0751	0.0390
$\overline{y}_{BRS}$	0.2403	0.1170	0.0749	N/A	0.0421	0.0332	0.0169
$\overline{y}_{BMRS}$	0.2403	0.1163	0.0746	N/A	0.0415	0.0333	0.0168

Table 8.6: Expected Mean Square Errors of the Sample Means, for all the Designs, under P5

Estimator	k is an integer					k is not an integer			
	$n = 4$	$n = 8$	$n = 12$	$n = 20$	$n = 40$	$n = 7$	$n = 14$	$n = 21$	$n = 35$
$\overline{y}_{LSS}$	33.7346	14.5548	8.1797	3.3629	0.9137	N/A	N/A	N/A	N/A
$\overline{y}_{CSS}$	33.7346	14.5548	8.1797	3.3629	0.9137	16.9654	6.4819	3.5319	1.6610
$\overline{y}_{SRSWR}$	37.6927	18.9476	12.6871	7.5341	3.7712	21.5407	10.8990	7.2581	4.3249
$\overline{y}_{SRSWOR}$	36.7424	17.8331	11.5144	6.3312	2.5353	20.4546	9.7084	6.0383	3.0892
$\overline{y}_{STR}$	34.2508	15.0272	8.8248	4.1640	1.2563	N/A	N/A	N/A	N/A
$\overline{y}_{YEC}$	35.2859	14.9049	8.2796	3.4130	0.9043	N/A	N/A	N/A	N/A
$\overline{y}_{CESS}$	32.0458	13.6244	8.0175	3.3941	0.8865	N/A	N/A	N/A	N/A
$\overline{y}_{BSS}$	35.9407	16.3556	9.7836	4.7931	1.5417	N/A	N/A	N/A	N/A
$\overline{y}_{MSS}$	34.3421	14.3914	8.2704	3.4935	0.8940	N/A	N/A	N/A	N/A
$\overline{y}_{BMSS}$	35.0916	16.1305	9.8832	5.1639	1.6202	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MLSS}$	35.8917	16.5382	9.6693	4.8015	1.3847	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	10.5461	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	35.8917	17.0015	10.4837	5.5156	1.9756	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	11.0116	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{PSS}$	48.8885	23.4229	21.0201	8.2683	3.5548	42.3953	20.4566	38.6164	7.4010
$\overline{y}_{NSS}$	64.0810	17.6734	9.3008	3.7816	0.9718	22.2468	7.1643	3.4683	1.7829
$\overline{y}_{NPSS}$	35.7295	15.4253	8.7452	3.6426	0.9683	19.0205	6.9333	3.5352	1.6162
$\overline{y}_{BRS}$	36.5276	17.5760	11.4888	6.2496	2.4953	N/A	9.6201	N/A	N/A
$\overline{y}_{BMRS}$	37.5276	17.1221	11.5665	5.4813	1.8571	N/A	N/A	N/A	N/A

Table 8.7: Expected Mean Square Errors of the Sample Means, for all the Designs, under P6A

Estimator	k is an integer					k is not an integer				
	$n = 4$	$n = 8$	$n = 11$	$n = 16$	$n = 22$	$n = 6$	$n = 12$	$n = 18$	$n = 24$	$n = 30$
$\overline{y}_{LSS}$	5976.01	3158.77	536.08	325.60	130.05	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{CSS}$	5976.01	3158.77	536.08	325.60	130.05	2288.39	836.37	331.98	238.15	373.51
$\overline{y}_{SRSWR}$	6028.67	3014.33	2192.24	1507.17	1096.12	4019.11	2009.56	1339.70	1004.78	803.82
$\overline{y}_{SRSWOR}$	5925.32	2893.76	2066.97	1377.98	964.59	3904.28	1883.24	1209.56	872.24	670.62
$\overline{y}_{STR}$	4986.96	1607.81	574.74	433.55	196.46	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{YEC}$	6083.59	1888.84	337.01	163.29	118.04	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{CESS}$	7075.10	87.68	8.60	20.92	477.03	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{BSS}$	5441.31	2352.94	494.80	95.19	154.12	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{MSS}$	2948.21	436.83	1021.60	163.28	192.87	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{BMSS}$	8783.34	2278.57	469.53	180.81	118.10	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MLSS}$	6401.00	2918.52	N/A	1504.18	250.17	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	6401.00	4289.54	N/A	1085.03	219.11	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{PSS}$	6160.41	2905.85	2701.18	1292.96	1038.98	4421.63	6144.82	3060.91	3018.19	742.91
$\overline{y}_{NSS}$	14081.52	3132.36	1071.56	565.25	186.57	6648.86	1026.50	407.22	296.44	157.25
$\overline{y}_{NPSS}$	6299.56	2718.36	896.58	447.03	185.45	2635.68	766.45	405.99	254.87	161.87
$\overline{y}_{BRS}$	3207.38	1566.40	N/A	745.90	522.13	2113.39	1019.40	654.74	472.40	363.01
$\overline{y}_{BMRS}$	3207.38	3555.48	N/A	739.38	N/A	N/A	N/A	N/A	N/A	N/A

Table 8.8: Expected Mean Square Errors of the Sample Means, for all the Designs, under P6B

Estimator	k is an integer					k is not an integer				
	$n=4$	$n=8$	$n=11$	$n=16$	$n=22$	$n=6$	$n=12$	$n=18$	$n=24$	$n=30$
$\overline{y}_{LSS}$	3302.58	674.22	381.01	320.28	38.65	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{CSS}$	3302.58	674.22	381.01	320.28	38.65	1430.37	547.70	368.18	149.75	453.60
$\overline{y}_{SRSWR}$	6028.67	3014.33	2192.24	1507.17	1096.12	4019.11	2009.56	1339.70	1004.78	803.82
$\overline{y}_{SRSWOR}$	5925.32	2893.76	2066.97	1377.98	964.59	3904.28	1883.24	1209.56	872.24	670.62
$\overline{y}_{STR}$	1842.24	561.14	424.11	240.45	130.71	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{YEC}$	1480.44	397.03	126.27	194.71	26.66	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{CESS}$	66.20	1957.05	154.55	540.03	23.43	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{BSS}$	1351.92	506.27	493.61	130.78	42.03	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{MSS}$	1446.73	481.13	489.71	148.21	37.93	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{BMSS}$	1285.59	603.53	187.93	171.68	301.60	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MLSS}$	2920.36	1612.89	N/A	321.06	177.81	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MLSS}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	2920.36	627.84	N/A	287.39	87.70	N/A	N/A	N/A	N/A	N/A
$\overline{y}^{(3)}_{MBMSS}$	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
$\overline{y}_{PSS}$	4440.21	2268.31	2138.25	1259.22	1037.91	3784.17	6111.22	3059.84	3001.63	726.18
$\overline{y}_{NSS}$	9014.36	2499.37	1036.68	371.87	76.49	4602.05	853.25	218.25	189.42	68.45
$\overline{y}_{NPSS}$	5044.98	1104.14	603.41	379.79	87.21	2692.81	586.11	228.26	169.56	73.90
$\overline{y}_{BRS}$	1364.48	666.37	N/A	317.32	222.12	899.08	433.67	278.54	200.97	154.43
$\overline{y}_{BMRS}$	1364.48	670.98	N/A	238.95	N/A	N/A	N/A	N/A	N/A	N/A

From Table 8.2, we note that most of the estimators are approximately subject to the same amount of error, as expected for P1. The expected error of estimation is relatively large for $\overline{y}_{PSS}$ which is owing to the weights (i.e. α_1 and α_2) being applied to the respective variance components. Moreover, $\overline{y}_{SRSWR}$ is slightly subject to more error than the other estimators, since we are sampling with replacement. It should be noted that if $G \to \infty$, then the corresponding expected MSEs for all the estimators, apart from $\overline{y}_{PSS}$ and $\overline{y}_{SRSWR}$, will tend to converge. Also, if $N \to \infty$, then the expected MSE of $\overline{y}_{SRSWR}$ will converge to that of the other estimators.

The conclusions from Table 8.3 are as follows:

(i) If $n = 4$, then the expected MSEs of $\overline{y}_{BSS}$, $\overline{y}_{MSS}$, $\overline{y}_{BMSS}$, $\overline{y}_{BRS}$ and $\overline{y}_{BMRS}$ are the smallest and are thus preferred. We further note that $\overline{y}_{YEC}$ is slightly subject to more error, while $\overline{y}_{CESS}$ is approximately subject to twice as much error, when compared to these estimators. This is expected, since $\overline{y}_{CESS}$ does not completely remove the linear trend component (i.e. k is even), unlike the favourable estimators, and $\overline{y}_{YEC}$ eliminates the linear trend component while increasing the expected error variance component.

(ii) If $n = 5$, then the expected MSE of $\overline{y}_{YEC}$ is the smallest. Just as in the case for $n = 4$, we do not obtain a removal of the linear trend component for estimator $\overline{y}_{CESS}$, resulting in $\overline{y}_{CESS}$ being roughly prone to twice as much error, when related to $\overline{y}_{YEC}$. Furthermore, since n is odd, the linear trend component is not entirely eliminated when conducting either BSS, MSS or BMSS, thus contributing to the corresponding estimators being susceptible to approximately 10 times as much error, when compared to $\overline{y}_{YEC}$.

(iii) The explanation for the case of $n = 8$ and $n = 4$ is similar, with the only difference being that k is now odd for the former case. As a result, $\overline{y}_{CESS}$ completely removes the linear trend component and is thus approximately prone to the same amount of minimum error as $\overline{y}_{BSS}$, $\overline{y}_{MSS}$, $\overline{y}_{BMSS}$, $\overline{y}_{BRS}$ and $\overline{y}_{BMRS}$. Furthermore, the expected MSE of $\overline{y}^{(2)}_{MBMSS}$ is approximately equivalent to the preferred estimators, since MBMSS involves the selection of two balanced modified systematic samples of size $n = 4$ and this is optimal, i.e. $n/2m$ is an even integer.

(iv) For the case where $n = 15$, we obtain related results to the situation when $n = 5$; however, the expected MSE of $\overline{y}_{YEC}$, relative to that of either $\overline{y}_{BSS}$, $\overline{y}_{MSS}$ or

$\overline{y}_{BMSS}$, is approximately 7.5 times smaller, when compared to the case when $n = 5$. Moreover, the expected MSE of $\overline{y}_{YEC}$, relative to that of $\overline{y}_{CESS}$, is approximately 2.5 times larger than is the case when n=5.

(v) If $n = 30$, then estimators $\overline{y}_{YEC}$, $\overline{y}_{BSS}$, $\overline{y}_{MSS}$ and $\overline{y}_{BRS}$ are advantageous over the other estimators. As expected, this is the only case where $\overline{y}_{BMSS}$ is not subject to the same amount of error as that of $\overline{y}_{BSS}$ and $\overline{y}_{MSS}$, i.e. $n/2 = 15$ is an odd integer. Moreover, the expected MSE of $\overline{y}_{CESS}$ is roughly 11 times larger, than that of the favourable estimators.

(vi) All the other estimators, which were not mentioned in (i) to (v), are heavily susceptible to error. The expected MSE of $\overline{y}_{STR}$ (which is the smallest of these least preferable estimators) gets closer to the expected MSEs of the favourable estimators as n increases, however, there is still a great difference when $n = 30$.

(vii) For the situation when $k = N/n$ is not an integer, there is no design which offers good results, apart from BRS, which is only applicable when $n = 14$ is even.

For P3, we note that $\sigma^2 \approxeq 50$, which follows since the corresponding sin curve is such that $\sin(\pi/2) = 1$, $\sin(\pi) = 0$, $\sin(3\pi/2) = -1$, $\sin(2\pi) = 0$, $\sin(5\pi/4) = 1$, ..., resulting in $y_1 \approxeq 10$, $y_2 \approxeq 0$, $y_3 \approxeq -10$, $y_4 \approxeq 0$, $y_5 \approxeq 10$, ... and $\overline{Y} \approxeq 0$, i.e. the population size is divisible by the period = 4. Accordingly, by examining Table 8.4, we conclude the following:

(i) If k is an odd multiple of half the period, then the most preferable designs are LSS, CSS, CESS, MSS and BMSS. The expected MSE of $\overline{y}_{BSS}$ is given by $\sigma^2/2 \approxeq 25$, since BSS reduces to the selection of two population units with different variate values. We further note the relationship between the multiple-start designs to that of the corresponding single-start designs. Consequently, if $k = 10$, then $\overline{y}^{(3)}_{MLSS}$ and $\overline{y}^{(3)}_{MBMSS}$ are comprised of three linear systematic samples and three balanced modified systematic samples, each of size $n = 4$, respectively. This sample size corresponds to optimal cases for LSS and BMSS and thus $\overline{y}^{(3)}_{MLSS}$ and $\overline{y}^{(3)}_{MBMSS}$ are both favourable for this situation.

(ii) If k is an integral multiple of the period, then the most advantageous design is CESS. For this situation, LSS reduces to a simple random sample of size one, resulting in the expected MSEs of $\overline{y}_{LSS}$, $\overline{y}_{CSS}$ and $\overline{y}_{YEC}$, all being approximately equivalent to

σ^2. Moreover, the explanation for the expected MSE of $\overline{y}_{BSS}$ in (i) also applies here, as well as for $\overline{y}_{MSS}$. The expected MSE of $\overline{y}_{BMSS}$ is equivalent to that of $\overline{y}_{BSS}$ and $\overline{y}_{MSS}$ when $k = 8$ and slightly greater than that of $\overline{y}_{BSS}$ and $\overline{y}_{MSS}$ when $k = 12$ and 20 (i.e. $n/2$ is an odd integer).

(iii) The expected MSEs of $\overline{y}_{BRS}$ and $\overline{y}_{BMRS}$ are not approximately equal to zero, i.e. by referring to (7.33), we note that $\sigma^2 \approxeq 50 \neq (y_i - y_{4k+1-i})^2/4 \approxeq 25$.

From Table 8.5, we conclude that estimators $\overline{y}_{LSS}$, $\overline{y}_{CSS}$, $\overline{y}_{STR}$, $\overline{y}_{CESS}$, $\overline{y}_{BSS}$, $\overline{y}_{MSS}$, $\overline{y}_{BMSS}$, $\overline{y}_{BRS}$ and $\overline{y}_{BMRS}$ are all approximately subject to the same amount of minimum error, when comparing all the estimators. The explanation of the optimality of MBMSS, given for point (i) above, also applies here. We further note that MBMSS is always favoured over MLSS, except for the case when $n = 4$, which results in both designs being equivalent, i.e. $n/m = 2$.

If k is an integer, then by referring to Table 8.6, we conclude that the most preferable sampling designs are LSS, CSS, CESS and MSS, under P5. Between these designs we can further deduce that $\overline{y}_{CESS}$ is susceptible to the least amount of error for most cases. For the case where $n = 20$, the most advantageous designs are LSS and CSS. Moreover, $\overline{y}_{YEC}$ is marginally subject to more error than that of $\overline{y}_{CESS}$, for the case where $n = 40$. If k is not an integer, then $\overline{y}_{CSS}$ is prone to the least amount of error for most cases. In addition, as n increases, there is a greater reduction in the expected MSEs of $\overline{y}_{NSS}$ and $\overline{y}_{NPSS}$, when compared to all the other applicable estimators. More notably, the case $n = 21$ results in a minimum expected MSE of estimator $\overline{y}_{NSS}$, while estimator $\overline{y}_{NPSS}$ is prone to the least amount of error if $n = 35$.

From Table 8.7, we note that $\overline{y}_{CESS}$ is on average susceptible to the least amount of error, for the case where k is an integer. When compared to all the other estimators, the expected MSE of $\overline{y}_{MSS}$ is a minimum when $n = 4$, while the expected MSEs of $\overline{y}_{YEC}$ and $\overline{y}_{BMSS}$ are minimum when $n = 22$. If k is not an integer, then CSS and NPSS are on average the most advantageous designs. For the rearranged population, given by Table 8.8, we see that the estimators related to CESS and YEC on average perform the best when k is an integer, while the most favourable designs, for the case where k is not an integer, are on average CSS and BRS. The performances of estimators $\overline{y}_{NSS}$ and $\overline{y}_{NPSS}$ are very poor for small sample sizes and improve drastically as the sample size increases. More notably, if $n = 18$ or $n = 30$, then the corresponding estimators are subject to minimum error, when compared to the other estimators. A major feature, when comparing Table 8.7 and Table 8.8, is the fact that in most instances, we obtain a reduction in the expected MSE of the estimators, when rearranging the population. In addition, the expected MSEs related to the SRS designs are constant, i.e. there is no efficiency gain by rearranging the population, when conducting SRS. Notable increases in the expected MSEs, when rearranging the population, are for $\overline{y}_{CESS}$ (n = 8, 11 and 16), $\overline{y}_{BSS}$ (n = 16), $\overline{y}_{MSS}$ (n = 8), $\overline{y}_{YEC}$ (n = 16), $\overline{y}_{BMSS}$ (n = 22) and $\overline{y}_{CSS}$ (n = 18 and 30). We may thus conclude that by rearranging the population according to a correlated auxiliary variable, we most likely will obtain a reduction in error, when estimating the population mean.

Further conclusions, which can be drawn from Tables 8.3 to 8.8, are noted as follows:

(i) If k is an integer, then $\overline{y}_{LSS}$ is subject to the same amount of error as is association with $\overline{y}_{CSS}$, for all populations.

(ii) The LSS design is usually preferred over the SRS designs, except when k is an integral multiple of the period for P3 and when $n = 8$ for P6A. Furthermore, if $n = 4$, then $\overline{y}_{SRSWOR}$ is prone to less error, when compared to $\overline{y}_{LSS}$, under P6A.

(iii) The LSS design offers a strict improvement over the STR design for P3 (when k is an odd multiple of half the period), P5, P6A (when $n = 11$, 16 and 22) and P6B (when $n = 11$ and 22). Approximate equivalence in the expected MSEs of $\overline{y}_{LSS}$ and $\overline{y}_{STR}$, occur for P4. The STR design is favoured over LSS for all other populations/cases.

(iv) Under P2, the sample means for the designs from Chapter 5 are susceptible to less error, when compared to that for LSS, SRSWR, SRSWOR and STR, except for the case when $n = 30$, which results in STR being advantageous over CESS. Moreover, we generally obtain the same result for P6B.

(v) A key feature for the MLSS design is that the expected MSE of $\overline{y}^{(m)}_{MLSS}$, for a sample of size n, is approximately less than m times smaller than the expected MSE of $\overline{y}_{LSS}$, for a sample of size n/m. Hence, if LSS is a favourable design for a sample of size n/m, then we note that MLSS is a preferable design for a sample of size n. A similar relation applies to MBMSS and BMSS.

(vi) Estimator $\overline{y}_{PSS}$ performs poorly for all populations and in most cases offers no advantage, in terms of a lower expected MSE, over the other estimators.

(vii) The NSS and NPSS designs offer improvement over SRS when n is not small. The rate of reduction in the expected MSEs of $\overline{y}_{NSS}$ and $\overline{y}_{NPSS}$ are greater than that of $\overline{y}_{SRSWR}$ and $\overline{y}_{SRSWOR}$, as n increases. For most populations, estimators $\overline{y}_{NSS}$ and $\overline{y}_{NPSS}$ are usually subject to more error than estimator $\overline{y}_{LSS}$, while the opposite holds true for P3, when k is an integral multiple of the period.

(viii) Estimators $\overline{y}_{BRS}$ and $\overline{y}_{BMRS}$ are approximately prone to the same amount of error in most cases. Furthermore, the associated designs are usually preferred over the SRS designs for most populations. If k is an integral multiple of the period, then estimator $\overline{y}_{BRS}$ is susceptible to less error, when compared to $\overline{y}_{LSS}$, $\overline{y}_{SRSWR}$ and $\overline{y}_{STR}$, for P3, while estimator $\overline{y}_{BRS}$ is marginally subject to more error than $\overline{y}_{SRSWR}$. Moreover, if k is an odd multiple of half the period, then the balanced random designs are preferred to STR and SRSWR, but are less favourable when compared to LSS and SRSWOR, under P3. For P5, the LSS and STR designs are advantageous over the balanced random sampling designs.

Table 8.9 represents the Monte Carlo simulations for the expected CLs associated with the sample means, using the corresponding sampling designs for the populations under consideration. The artificial populations are for a fixed sample of size $n = 12$, while the natural population is for a fixed sample of size $n = 16$. We omit the CESS, since there is no randomization involved with this design. By noting that the nominal rate is 95%, we refer to Table 8.9 and conclude the following:

(i) All estimators are able to produce expected CLs that are slightly above, or approximately equivalent to the nominal rate, under P1 and P5.

(ii) For P2, all estimators generate an expected CL that is above the nominal rate, where the associated CIs for estimators $\overline{y}_{LSS}$ and $\overline{y}_{CSS}$ will always contain $\overline{Y}$.

(iii) Under P3, estimators $\overline{y}^{(2)}_{MBMSS}$ and $\overline{y}_{NSS}$ will almost surely exhibit CIs that will contain $\overline{Y}$, whereas the CIs associated with $\overline{y}_{BSS}$ will always contain $\overline{Y}$. The expected CL for estimator $\overline{y}^{(2)}_{MLSS}$ is approximately 5 percentage points below the nominal rate. In addition, the expected CL for estimators $\overline{y}_{SRSWR}$, $\overline{y}_{STR}$, $\overline{y}_{YEC}$ and $\overline{y}_{NPSS}$ are all slightly below the nominal rate, while all other estimators generate expected CLs above the nominal rate.

Table 8.9: Expected Confidence Levels Constructed Using the Estimators Associated with The Sampling Designs

Estimator	Population						
	P1	P2	P3	P4	P5	P6A	P6B
$\overline{y}_{LSS}$	95.89	100	95.74	95.60	96.59	90.92	100
$\overline{y}_{CSS}$	95.89	100	95.71	95.62	96.58	90.89	100
$\overline{y}_{SRSWR}$	95.01	95.06	94.91	95.13	95.06	95.05	95.07
$\overline{y}_{SRSWOR}$	95.04	95.08	95.07	95.06	95.02	95.03	95.09
$\overline{y}_{STR}$	94.99	95.09	94.87	95.05	95.08	95.09	95.01
$\overline{y}_{YEC}$	96.04	95.96	94.36	97.00	96.53	100	100
$\overline{y}_{BSS}$	95.73	96.13	100	95.98	96.18	100	90.84
$\overline{y}_{MSS}$	95.94	95.95	95.95	95.87	96.48	90.90	100
$\overline{y}_{BMSS}$	95.86	96.00	96.20	96.17	96.54	100	100
$\overline{y}^{(2)}_{MLSS}$	95.34	96.00	89.81	95.86	95.48	98.69	95.67
$\overline{y}^{(3)}_{MLSS}$	95.21	95.33	95.21	95.32	95.28	N/A	N/A
$\overline{y}^{(2)}_{MBMSS}$	95.34	95.97	99.67	95.33	95.44	96.53	96.09
$\overline{y}^{(3)}_{MBMSS}$	95.23	95.21	95.22	95.21	95.30	N/A	N/A
$\overline{y}_{PSS}$	95.07	96.06	97.50	95.85	95.47	97.23	97.43
$\overline{y}_{NSS}$	95.43	96.54	99.78	96.25	95.93	96.60	96.02
$\overline{y}_{NPSS}$	95.36	95.79	94.76	94.03	95.54	38.45	34.84
$\overline{y}_{BRS}$	95.11	95.09	97.25	95.07	95.07	95.10	94.88
$\overline{y}_{BMRS}$	95.13	95.10	97.13	95.08	95.13	95.15	94.99

(iv) The expected CLs for estimators $\overline{y}_{YEC}$, $\overline{y}_{BMSS}$ and $\overline{y}_{NSS}$ are marginally advantageous over the other estimators for P4. Apart from estimator $\overline{y}_{NPSS}$, which produces an expected CL of approximately 1 percentage point below the nominal rate, all other estimators generate expected CLs which are slightly above the nominal rate, for P4.

(v) Under P6A, the corresponding CIs for estimators $\overline{y}_{YEC}$, $\overline{y}_{BSS}$, and $\overline{y}_{BMSS}$ will always contain $\overline{Y}$. In addition, the expected CL of estimator $\overline{y}^{(2)}_{MLSS}$ is well above the nominal rate, while the expected CLs associated with estimators $\overline{y}_{LSS}$, $\overline{y}_{CSS}$ and $\overline{y}_{MSS}$ are roughly 4 percentage points below the nominal rate. All other estimators, apart from $\overline{y}_{NPSS}$, are able to generate expected CLs above the nominal rate.

(vi) The CIs corresponding to estimators $\overline{y}_{LSS}$, $\overline{y}_{CSS}$, $\overline{y}_{YEC}$, $\overline{y}_{MSS}$ and $\overline{y}_{BMSS}$ will always contain $\overline{Y}$, under P6B. Furthermore, the expected CL of estimator $\overline{y}_{BSS}$ is roughly 4 percentage points below the nominal rate. All other estimators, apart from $\overline{y}_{NPSS}$, are able to generate expected CLs that are approximately close to the nominal rate.

(vii) For the natural population, given by P6A and P6B, estimator $\overline{y}_{NPSS}$ performs very poorly in providing a CI which contains $\overline{Y}$.

(viii) The expected CLs associated with estimators $\overline{y}_{YEC}$ and $\overline{y}_{BMSS}$, are on average better than that of all the other estimators, when considering all population structures.

A similar simulation study for the eight variance estimators in Section 6.1 can then be obtained by using the expected bias of the estimators as a further comparative measure. An example of such a study is given by Wolter (2007). Consequently, all the theory discussed in Section 6.1 is supported by the numerical results provided by Wolter.

We next provide a comprehensive report on all the work coved in this thesis as well as some remarks on future studies for the topic on hand.

Chapter 9

CONCLUSIONS

The conventional systematic sampling design, also known as LSS, is often used for large scale sampling (Arnab & North 2012), since it is simple to apply and further ensures a more even spread of the sample over the listed population elements, as opposed to the case when conducting SRSWR, SRSWOR or STR. However, there are shortcomings within the systematic sampling context and a recap of the key disadvantages discussed earlier is as follows:

(i) If N is not a multiple of n, then LSS will either result in variable sample sizes or constant sample sizes that are greater than n. Consequently, the former situation results in biased estimates of population parameters, while the latter situation is undesired since sample sizes are usually fixed in advance.

(ii) LSS is susceptible to periodic distortions, since the process of selection can negatively interact with the population structure where periodic characteristics are evident.

(iii) LSS is not the most preferred probability sampling design for populations that exhibit linear trend.

(iv) Certain pairs of population units will have a zero probability of being selected in the sample, which results in it being impossible to obtain an unbiased estimate of the sampling variance, from a single sample.

In this thesis, we provided solutions to the above-mentioned shortcomings and as a result, we generally introduced some modified systematic sampling designs. The designs that aided us in solving problem (i) are the FIM and CSS. For problem (iii), we presented four variations of systematic sampling (i.e. YEC, CESS, BSS, MSS) as well as a new proposed design known as BMSS. To tackle problem (iv), we first considered the application of LSS and subsequently proposed eight estimators of the sampling variance, after which we made assumptions on the population structure, so as to find the estimator with the least amount of bias, under each underlying population structure. We then provided some modified systematic sampling designs to solve problem (iv), given by MLSS, PSS, NSS, NPSS and BRS, as well as two new proposed designs given by MBMSS and BMRS. The notion of supplementing a systematic sample with independent sample(s) was suggested for MLSS, MBMSS and PSS, while the supplementation of a systematic sample with a dependent sample was suggested for NSS, NPSS, BRS and BMRS. We then numerically tested the designs against each other, under various population structures, so as to find the best design for each underlying population structure, i.e. we numerically provided a solution to problem (ii).

9.1 Conclusions

The significant results obtained in this thesis are summarized as follows:

(i) The FIM and CSS are advantageous over LSS when k is not an integer, since the former designs result in samples of fixed sample size n. We are thus able to obtain unbiased estimates of the population parameters for the required sample size, when applying the former designs, whereas this is not possible for LSS. In addition, the FIM is more often than not equivalent to CSS, while discrepancies between the two designs occur when $2N/n$ is an integer or when lcm(N, k)$<$ nk, which are rare in practice. Moreover, the FIM and CSS reduce to LSS when k is an integer. The usual CSS design, which uses a sampling interval of INT$(N/n + 1/2)$, may result in sampling units that coincide, so that we thus suggested using the sampling interval given in (2.6) to obtain distinct sampling units, while ensuring an even spread of the sample over the population.

(ii) When conducting LSS, maximum precision of estimates is obtained when the population units that lie within the same systematic sample vary as much as possible, i.e. the variance within the possible systematic samples is high, which consequently is equivalent to saying that the difference between the possible systematic samples that could be selected is as low as possible. Consequently, the sampling variance is dependent on the arrangement of the population units from which the systematic samples are to be drawn. Evidence of this is given in (3.8), where $\text{Var}(\overline{y}_{LSS})$ varies consistently with the ICC, which in turn largely depends on the ordering of population units and the amount of correlation between successive population units. This is in direct contrast to the SRS designs, where the arrangement of the population units has no effect on the corresponding sampling variances. A summary on the efficiency of LSS, when compared to SRSWR, SRSWOR and STR, is given as follows:

(a) For randomly ordered populations: LSS is expected to be equivalent to both SRSWOR and STR, and more efficient than SRSWR.

(b) For populations that exhibit linear trend: LSS is more efficient than both SRSWR and SRSWOR, but less efficient than STR.

(c) For periodic populations: If k is an odd multiple of half the period, then $\text{Var}(\overline{y}_{LSS}) \approxeq 0$ and thus LSS is more efficient than SRSWR, SRSWOR and STR. On the other hand, if k is an integral multiple of the period, then $\text{Var}(\overline{y}_{LSS}) = \sigma^2$ and thus LSS is less efficient than SRSWR, SRSWOR and STR.

(d) For auto-correlated populations: In the case of a positive convex decreasing correlogram, LSS is more efficient than SRSWR, SRSWOR and STR. This assumption applies for the cases where the correlograms are linear, exponential and hyperbolic tangent, as well as for any process which is autoregressive and has real roots, with respect to the characteristic equation.

(e) For stratified populations: stratified systematic sampling is more often than not, more efficient than STR, for the case where strata are considered to be large and more than one unit is to be drawn from each stratum for the sample.

(iii) In Chapter 5, we considered variations of the LSS design that are optimal for populations that exhibit linear trend. Under the assumption that k is an integer, the corresponding designs are summarized as follows:

(a) The YEC estimator inherits the LSS design, with the only difference being that an estimate of $\overline{Y}$ is obtained by applying appropriate weights to the first and last sampling units. As a result, we obtain a complete removal of the linear trend component; however, the unevenly weighted sampling units result in a larger error variance component. Consequently, this estimator is slightly biased in practice. Nevertheless, preference is given to this estimator over $\overline{y}_{LSS}$.

(b) CESS involves the selection of the centrally located linear systematic sample and there is thus no randomization required. Consequently, some population units have no chance of being selected for the sample when applying CESS, so that $\overline{y}_{CESS}$ is prone to exhibit an element of bias. A further disadvantage of this estimation procedure is that it is impossible to provide an estimate of the error associated with $\overline{y}_{CESS}$, when estimating $\overline{Y}$.

(c) BSS entails reversing the order, with respect to the population unit numbers, of every alternative set of k population units, before applying LSS on this balanced arrangement. Estimator $\overline{y}_{BSS}$ is consequently design unbiased.

(d) MSS divides the population into two groups and then reverses the population units in the second group, with respect to their population unit numbers, i.e. If n is even, then we reverse the order of the last $n/2$ sets of k populations and if n is odd, then we reverse the order of the last $(n-1)/2$ sets of k population units. LSS is then applied to this modified arrangement. Estimator $\overline{y}_{MSS}$ is consequently design unbiased.

(e) BMSS involves the application of MSS on a balanced arrangement, i.e. we apply a balanced arrangement and then apply a modified arrangement, before conducting LSS on this balanced modified arrangement. Estimator $\overline{y}_{BMSS}$ is consequently design unbiased. BMSS reduces to LSS if $n = 2$.

For populations that exhibit a rough linear trend, estimators $\overline{y}_{YEC}$ and $\overline{y}_{CESS}$ are generally expected to be subject to less error, when compared to estimators $\overline{y}_{LSS}$, $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$ and $\overline{y}_{STR}$. Moreover, estimators $\overline{y}_{BSS}$ and $\overline{y}_{MSS}$ are always expected to be subject to less error than estimators $\overline{y}_{LSS}$, $\overline{y}_{SRSWR}$, $\overline{y}_{SRSWOR}$ and $\overline{y}_{STR}$, for $n \geq 2$. On the other hand, estimator $\overline{y}_{BMSS}$ is prone to less error when compared to the latter estimators if $n > 2$. When comparing these designs amongst each other, we provide the following recommendations for the most appropriate design(s), under the assumption of linear trend:

(a) if k is even and $n/2$ is an even integer, then it is best to use BSS, MSS, or BMSS;

(b) if k is even and $n/2$ is an odd integer, then it is best to use BSS or MSS;

(c) if k is even and n is odd, then it is best to use YEC;

(d) if k is odd and $n/2$ is an even integer, then it is best to use CESS, BSS, MSS, or BMSS;

(e) if k is odd and $n/2$ is an odd integer, then it is best to use CESS, BSS or MSS;

(f) if k is odd and n is odd, then it is best to use CESS.

(iv) If we maintain the usual systematic sampling design, then we need to construct estimators, which are based on certain assumptions, to estimate $\text{Var}(\overline{y}_{LSS})$. These estimators will be biased if the population exhibits some structure, other than random. Accordingly, we constructed eight estimators (refer to estimators v_1 to v_8 in Section 6.1.1) and tested them on various population structures. The following recommendations are given for each population structure:

(a) For randomly ordered populations: estimators v_1 to v_7 are unbiased and are thus preferred, while estimator v_8 is expected to be slightly biased.

(b) For populations that exhibit linear trend: Estimators v_2 and v_3 are least biased and thus favourable. Estimators v_4, v_5 and v_6 remove the linear trend component and are thus not desirable since $\text{Var}(\overline{y}_{LSS})$ is a function of linear trend. However, these estimators are unbiased estimators of $\text{Var}\left[\overline{y}_{BSS(n\text{ even})}\right]$, $\text{Var}\left[\overline{y}_{MSS(n\text{ even})}\right]$ and $\text{Var}\left[\overline{y}_{BMSS(n/2\text{ even integer})}\right]$, and are slightly biased estimators of $\text{Var}(\overline{y}_{YEC})$. It is impossible to obtain an estimate for the error associated with $\overline{y}_{CESS}$, when estimating $\overline{Y}$.

(c) For periodic populations: all estimators are heavily biased and we thus cannot provide an adequate estimate of $\text{Var}(\overline{y}_{LSS})$.

(d) For auto-correlated populations: Estimator v_8 is most likely to exhibit the smallest absolute bias and thus provides a good estimate of $\text{Var}(\overline{y}_{LSS})$. Furthermore, estimator v_8 will likely provide an overestimate of $\text{Var}(\overline{y}_{LSS})$, while all other estimators tend to underestimate $\text{Var}(\overline{y}_{LSS})$.

(e) For stratified populations: Estimators v_4, v_5 and v_6 are the least biased and are thus most favoured. These estimators tend to eliminate the trend component in the stratum means and are thus ideal since $\text{Var}(\overline{y}_{LSS})$ is not a function of trend. If the trend component is non-linear, then estimator v_6 is preferred over estimators v_4 and v_5.

(v) MLSS and MBMSS involve the selection of m linear systematic samples and m balanced modified systematic samples, respectively. We obtain more efficient results if the m samples are selected using SRSWOR, as opposed to selecting them using SRSWR. If LSS is a preferable design for a sample of size n/m, then we note that MLSS is a favoured design for a sample of size n. The same relationship holds true for BMSS and MBMSS. Equivalence between MLSS and MBMSS occurs when $n/m = 2$, since LSS is equivalent to BMSS when $n = 2$.

(vi) PSS involves the supplementation of a linear systematic sample with an independent sample, using SRSWOR. A natural weighted average of the corresponding sample means provide efficient results, when estimating $\overline{Y}$; however, the associated estimator of the sampling variance (v_{11}) may assume negative values for this situation. An unweighted average of the corresponding sample means will always result in $v_{11} \geq 0$; however, we do not obtain efficient results when estimating $\overline{Y}$ for this scenario. We are thus presented with a trade-off, where we can either have v_{11} that may assume negative values and an estimate of $\overline{Y}$ that is subject to less error than that of $\overline{y}_{SRSWOR}$, or $v_{11} \geq 0$ and an estimate of $\overline{Y}$ that is subject to more error than $\overline{y}_{SRSWOR}$. In Chapter 9, we noted that an unweighted average of the corresponding sample means results in $\overline{y}_{PSS}$ being heavily subject to error, when estimating $\overline{Y}$. We thus conclude that PSS is not a desirable design.

(vii) NSS selects a continuous set of population units from the frame, chosen according to a random start, and supplements these units with a dependent circular systematic sample, so that the selected units collectively form the new systematic sample. Assumptions on the design to produce positive second-order inclusion probabilities result in a restriction on n. A modified design known as NPSS, randomly selects a sample space of size a from the frame, before selecting u units within this sample space, using SRSWOR. These sampling units are then supplemented with a circular systematic sample of size $n-u$, which is selected according to the randomly selected sample space, so that all the selected units collectively form the new partially systematic sample. Furthermore, NPSS involves a modified selection of the circular systematic sample, when compared to NSS, where appropriate values of k and u are chosen to ensure distinct sampling units, as well as an even spread of the sample over the population. The corresponding choices of u and k, along with the restrictions in

Theorem 7.6, will not result in any limitations on n and thus NPSS is advantageous over NSS on this basis.

(viii) BRS first selects $n/2$ units from the first $N/2$ population units, using SRSWOR, before selecting units that occur at an equivalent position at the other end of the population, i.e. we use a MSS pairing technique. BMRS divides the population into $n/4$ groups and then uses the MSS pairing technique within each group, before selecting two pairs of units for each group, using SRSWOR. BMRS thus reduces to BRS when $n = 4$.

(ix) Three important characteristics, when comparing estimators of a specific population parameter, are given as follows:

(A) the best estimator will exhibit minimum MSE;

(B) the best estimator will exhibit the highest percentage of CIs, which contains the true population mean;

(C) it must be possible to find an unbiased estimate of the corresponding sampling variance of the estimator.

Accordingly, by comparing all the designs, we provide the following recommendations:

(a) For randomly ordered populations: use any design, except SRSWR and PSS.

(b) For populations that exhibit linear trend: If k is an integer, then use either BRS (if N and n are even), BMRS (if $n/2$ is an even integer) or MBMSS (if $n/m = 4$). If these designs are inapplicable, then use the designs from Chapter 5, according to the recommendations in (iii)((a) to (f)), such that we are able to satisfy both (A) and (C) if n is even (refer to (iv)(b)). Furthermore, if k is not an integer, then use NPSS (if n is not small), otherwise, use CSS.

(c) For periodic populations: If mk is an odd multiple of half the period, then use MBMSS, i.e. MBMSS is preferred over MLSS on the basis of (B). If the assumptions that correspond to m are not true and k is an odd multiple of half the period, then use LSS, CSS, CESS, MSS or BMSS, if preference is given to (A) over (C), alternatively, use NPSS if preference is given to (C) over (A). If the assumptions that correspond to m are not true and k is an integral multiple of the period, then use CESS if preference is given to (A) over (C), otherwise use NSS if preference is given to (C) over (A).

(d) For auto-correlated populations: If k is an integer, then use CESS if preference is given to (A) over (C), alternatively, use NSS if preference is given to (C) over (A). Moreover, if we assume n to be large, then use NPSS if k is not an integer, otherwise use CSS if and only if preference is given to (A) over (C).

(e) For stratified populations: use STR, MBMSS (if $n/2m$ is an even integer), BRS (if N and n are even) or BMRS (if $n/2$ is an even integer).

(x) If we arrange a population in ascending/descending order according to an auxiliary variable, then we obtain a population that approximately exhibits linear trend, where the stronger the degree of correlation between the auxiliary variable and the variable of interest, results in the rearranged population exhibiting a stronger degree of linear trend. The designs in Chapter 5, as well as MBMSS, BRS and BMRS are then optimal for this situation.

Finally, we note in passing that all the systematic sampling designs and variance estimators require us to have knowledge of the population structure, so that we may apply the most suitable design and/or variance estimator for the corresponding population structure. In practice we are not given the population structure and thus, the onus is on the survey statistician to gather as much information about the population as possible, prior to sampling, so as to estimate the population structure. This may involve the building of appropriate models, where we can then apply the most suitable design and/or variance estimator, as presented in this thesis, according to the estimated population structure.

9.2 Future Studies

This thesis primarily focused on solving problems (ii) to (iv). For large scale sampling, we are often presented with the scenario where N is not a multiple of n, and thus problem (i) often occurs in practical situations. For our future studies, we will thus provide: (a) further designs to solve problem (i); (b) modifications to some of the designs presented in this thesis; and (c) an extension to the BMRS design. A brief overview of such studies is given as follows:

(i) *Balanced modified circular systematic sampling*:

This proposed design uses CSS on a balanced modified arrangement. We first divide the population into $n+1$ groups, such that the first n groups contain k population units each and the last group is of size c, i.e. $N = nk+c$, $0 < c < n$. We then reverse the order, with respect to the population unit numbers, of every alternative group. Thereafter, we reverse the order of the last $n/2$ groups of units if n is even, otherwise we reverse the order of the last $(n+1)/2$ groups of units when n is odd. We then apply CSS, with our proposed sampling interval in (2.6), to this balanced modified arrangement, so as to obtain the balanced modified circular systematic sample.

(ii) *Remainder balanced modified systematic sampling*:

If $k = N/n$ is not an integer, then the population size can be expressed as $N = nk + c = (n-c)k + c(k+1)$, where $0 < c < n$. We then divide the population into two groups, where the first group contains the first $(n-c)k$ population units and the second group contains the last $c(k+1)$ units. Now, select $(n-c)$ units with a sampling interval of k in the first group, using BMSS. Also, select c units with a sampling interval of $(k+1)$ in the second group, using BMSS. The selected units then collectively form the remainder balanced modified systematic sample.

(iii) *Multiple-start remainder balanced modified systematic sampling*:

If there exists integers p and q, such that $(n-c)/p$ and c/q are integers, then we select p balanced modified systematic samples of size $(n-c)/p$ from the first group and supplement this with q balanced modified systematic samples of size c/q in the second group, using the groups and sampling intervals in (ii). This design results in an unbiased estimate of the associated sampling variance, since every possible pair of population units will have a chance of being selected for the sample.

(iv) *New partially balanced modified systematic sampling*:

By referring to Section 7.2.1, we select the $n-u$ sampling units for the sample S_t'', using balanced modified circular systematic sampling, as discussed in (i).

(v) Further methodologies for BMRS are given as follows:

(a) If $n/2$ is an odd integer: We first group the first set of $2k$ population units with the last set of $2k$ population units, the second set of $2k$ population units with the second last set of $2k$ population units, and so forth, such that the first

$(n-2)/4$ groups each contain $4k$ population units and the last group contains $2k$ population units. Apply BMRS on the first $(n-2)/4$ groups to select $(n-2)$ sampling units and then supplement these sampling units with two units from the last group, which are selected using SRSWOR. It should be noted that if $n = 2$, then BMRS reduces to MSS.

(b) If $n = 3$: For this situation, we view the entire population as a group. We then select two units using MSS and supplement these units by randomly selecting a unit from the remaining $N - 2$ population units.

(c) If $n \neq 3$ and $(n+1)/2$ is an even integer: We use the grouping method defined in (a), such that the first $(n-3)/4$ groups each contain $4k$ population units and the last group contains $3k$ population units. We then apply BMRS on the first $(n-3)/4$ groups to select $(n-3)$ sampling units and supplement this with three units from the last group, according to (b).

(d) If $(n+1)/2$ is an odd integer: We use the grouping method defined in (a), such that the first $(n-5)/4$ groups each contain $4k$ population units and the last group contains $5k$ population units. Next, apply BMRS on the first $(n-5)/4$ groups to select $(n-5)$ sampling units. We then use the MSS pairing technique for the last group, before randomly selecting two pairs of units using SRSWOR and then supplement this with a randomly selected unit from the remaining $5k - 4$ population units.

(vii) *Balanced modified systematic sampling with end corrections* (BMSSEC):

By applying weights to the first and last sampling units of $\overline{y}_{BMSS}$ when $n/2$ is not an even integer, we obtain an estimator that completely removes the linear trend component. The corresponding estimator will exhibit a lower MSE, when compared to $\overline{y}_{BMSS}$ for this scenario. This estimator is given in the next theorem.

Theorem 9.1: The BMSSEC estimator of $\overline{Y}$ with random start i, for $i \in \{1, ..., k\}$, is given as

$$\overline{y}_{BMSSEC} = \overline{y}_{BMSS} + \frac{[(x_n + x_1) - K]}{n\,(x_n - x_1)}\,(y_{x_1} - y_{x_n})\,, \tag{9.1}$$

where $K = n(N+1)/2 - \sum_{j=2}^{n-1} x_j$ and $x_1, ..., x_n$ are the population unit numbers of the sampling units, when conducting BMSS, which are arranged in ascending order, e.g. if the balanced modified systematic sample is y_7, y_2 and y_{12} then $x_1 = 2$, $x_2 = 7$ and $x_3 = 12$.

Proof: An estimate of $\overline{Y}$ with random start i, for $i \in \{1, ..., k\}$, is given as

$$\overline{y}_{BMSSEC} = \frac{1}{n}\left[\psi_1 y_{x_1} + \sum_{j=2}^{(n-1)} y_{x_j} + \psi_2 y_{x_n}\right], \tag{9.2}$$

where ψ_1 and ψ_2 are the weights applied to the first and the last sampling units, respectively. By substituting (4.5) into (9.2) and then equating this result to (4.6), we obtain

$$\overline{y}_{BMSSEC} = \frac{1}{n}\left[\psi_1\left(a + bx_1\right) + \sum_{j=2}^{(n-1)}\left(a + bx_j\right) + \psi_2\left(a + bx_n\right)\right] = a + \frac{b(N+1)}{2}. \tag{9.3}$$

By equating the coefficients of a in (9.3), it follows that

$$\psi_1 = 2 - \psi_2. \tag{9.4}$$

Similarly, by equating the coefficients of b in (9.3), we obtain

$$\frac{1}{n}\left[\psi_1 x_1 + \sum_{j=2}^{(n-1)} x_j + \psi_2 x_n\right] = \frac{N+1}{2}. \tag{9.5}$$

Substituting (9.4) into (9.5) results in

$$2\left[2x_1 - \psi_2 x_1 + \sum_{j=2}^{(n-1)} x_j + \psi_2 x_n\right] = n\left(N+1\right),$$

which simplifies to

$$\psi_2 = \frac{K - 2x_1}{x_n - x_1}. \tag{9.6}$$

The weight applied to the first sampling unit is thus obtained by substituting (9.6) into (9.4), such that

$$\psi_1 = \frac{2x_n - K}{x_n - x_1}. \tag{9.7}$$

We thus conclude the proof by substituting (9.6) and (9.7) into (9.2), i.e.

$$\begin{aligned}
\overline{y}_{BMSSEC} &= \frac{1}{n}\left[\frac{(2x_n - K)}{(x_n - x_1)} y_{x_1} + \sum_{j=2}^{n-1} y_{x_j} + \frac{(K - 2x_1)}{(x_n - x_1)} y_{x_n}\right] \\
&= \overline{y}_{BMSS} + \frac{1}{n}\left[\frac{(2x_n - K)}{(x_n - x_1)} y_{x_1} + \frac{(K - 2x_1)}{(x_n - x_1)} y_{x_n} - y_{x_1} - y_{x_n}\right] \\
&= \overline{y}_{BMSS} + \frac{1}{n}\left[\frac{\{2x_n - K - (x_n - x_1)\}}{(x_n - x_1)} y_{x_1} + \frac{\{K - 2x_1 - (x_n - x_1)\}}{(x_n - x_1)} y_{x_n}\right] \\
&= \overline{y}_{BMSS} + \frac{\{(x_n + x_1) - K\}}{n\left(x_n - x_1\right)}\left(y_{x_1} - y_{x_n}\right),
\end{aligned}$$

where $\overline{y}_{x_1} + \sum_{j=1}^{n-2} y_{x_j} + y_{x_n} = \sum_{j=1}^{n} y_{x_j} = n\overline{y}_{BMSS}$.

In closing, we note that this thesis specifically dealt with systematic sampling with an equal probability of selection. However, it may be advantageous to conduct systematic sampling with unequal probabilities, i.e. to consider the situation where each population unit has an auxiliary variable of size attached to it. The systematic sample would then be selected in a way that ensures that the probability of selection is positively correlated with the size measures of the items. This is commonly referred to as pps (probability proportionate to size) systematic sampling. Moreover, we only considered sampling in one-dimension. Systematic sampling is often used for spatial sampling. The usual systematic sampling problems, which were presented in this thesis, also applies to these fields, so that we may translate the designs and theory discussed in this thesis to pps systematic sampling and spatial systematic sampling, so as to find the suitable solutions for the problems within these fields.

Bibliography

Arnab, R. & North, D. (2012), ‘An appraisal of household income and expenditure survey design’, *Pakistan Journal of Statistics* **28**(4), 423–436.

Bellhouse, D. R. (1984), ‘On the choice of the sampling interval in circular systematic sampling’, *Sankhyā: The Indian Journal of Statistics, Series B,* **46**(2), 247–248.

Bellhouse, D. R. (1988), Systematic sampling, *in* P. R. Krishnaiah & C. R. Rao, eds, ‘Handbook of Statistics’, Vol. 6, Elsevier, Amsterdam, chapter 6, pp. 125–145.

Cochran, W. (1946), ‘Relative accuracy of systematic and stratified random samples for a certain class of populations’, *The Annals of Mathematical Statistics* **17**, 164–177.

Cochran, W. (1977), *Sampling Techniques*, 3rd edn, John Wiley & Sons, New York.

Daniel, J. (2012), *Sampling Essentials: Practical Guidelines for Making Sampling Choices*, SAGE Publications, California.

Deming, W. E. (1950), *Some theory of sampling*, John Wiley and Sons, New York.

D’Orazio, M. (2003), ‘Estimating the variance of the sample mean in two-dimensional systematic sampling’, *Journal of Agricultural, Biological, and Environmental Statistics* **8**(3), 280–295.

Dunn, R. & Harrison, A. R. (1993), ‘Two-dimensional systematic sampling of land use’, *Journal of the Royal Statistical Society, Series C (Applied Statistics),* **42**(4), 585–601.

Finney, D. J. (1948), ‘Random and systematic sampling in timber surveys’, *Forestry* **22**, 64–99.

Finney, D. J. (1950), ‘An example of periodic variation in forest sampling’, *Forestry* **23**, 96–111.

Fisher, R. A. & Mackenzie, W. A. (1922), 'The correlation of weekly rainfall', *Quarterly Journal of the Royal Meteorological Society* **48**, 234–245.

Gautschi, W. (1957), 'Some remarks on systematic sampling', *The Annals of Mathematical Statistics* **28**(2), 385–394.

Hasel, A. A. (1938), 'Sampling error in timber surveys', *Journal of Agricultural Research* **57**(10), 713–736.

Hasel, A. A. (1942), 'Estimation of volume in timber stands by strip sampling', *The Annals of Mathematical Statistics* **13**(2), 179–206.

Horvitz, D. G. & Thompson, D. (1952), 'A generalization of sampling without replacement from a finite universe', *Journal of the American Statistical Association* **47**(260), 663–685.

Iachan, R. (1982), 'Systematic sampling: a critical review', *International Statistical Review* **50**(3), 293–303.

Jacobsen, J. S. (1998), Soil sampling, Technical report, Montana State University Extension Publications.

Kalton, G. (1983), *Introduction to Survey Sampling*, SAGE Publications, Beverly Hills.

Kish, L. (1965), *Survey sampling*, John Wiley & Sons, New York.

Koop, J. C. (1971), 'On splitting a systematic sample for variance estimation', *The Annals of Mathematical Statistics* **42**(3), 1084–1087.

Lahiri, D. B. (1954), On the question of bias of systematic sampling, *in* 'Proceedings of World Population Conference', Vol. 6, pp. 349–362.

Lehtonen, R. & Pahkinen, E. (2004), *Basic Sampling Techniques*, 2nd edn, John Wiley & Sons, Chichester, pp. 9–58.

Leu, C.-H. & Tsui, K.-W. (1996), 'New partially systematic sampling', *Statistica Sinica* **6**, 616–630.

Lohr, S. (2010), *Sampling: Design and Analysis*, 2nd edn, Cengage Learning, Boston: Brooks/Cole.

Madow, L. H. (1946), 'Systematic sampling and its relation to other sampling designs', *Journal of the American Statistical Association* **41**(234), 204–217.

Madow, W. G. (1953), 'On the theory of systematic sampling, III. comparison of centered and random start systematic sampling', *The Annals of Mathematical Statistics* **24**(1), 101–106.

Madow, W. G. & Madow, L. H. (1944), 'On the theory of systematic sampling, I', *The Annals of Mathematical Statistics* **15**, 1–24.

Mahalanobis, P. C. (1946), 'Recent experiments in statistical sampling in the indian statistical institute', *Journal of the Royal Statistical Society* **109**(4), 325–370.

Manson, B. (1992), Preparation of soil sampling protocols: sampling techniques and strategies, Technical report, Encironmental Monitoring Systems Laboratory, Office of Research and Development, U.S. Environmental Protection Agency, Las Vegas, Nevada.

Matérn, B. (1947), 'Methods of estimating the accuracy of line and sample plot surveys', *Meddelanden från Statens Skogsforskningsinstitut* **36**, 1–138.

Matérn, B. (1960), 'Spatial variation: stochastic models and their application to some problems in forest surveys and other sampling investigations', *Meddelanden från Statens Skogsforskningsinstitut* **49**(5), 1–144.

McArthur, R. D. (1987), 'An evaluation of sample designs for estimating a locally concentrated pollutant', *Communications in Statistics - Simulations and Computing* **16**, 735–759.

Milne, A. (1959), 'The centric systematic area-sample treated as a random sample', *Biometrics* **15**, 270–297.

Murthy, M. (1967), *Sampling: Theory and Methods*, Statistical Publishing Society, Calcutta.

Murthy, M. N. & Rao, T. J. (1988), Systematic sampling with illustrative examples, *in* P. R. Krishnaiah & C. R. Rao, eds, 'Handbook of Statistics', Vol. 6, Elsevier, Amsterdam, chapter 7, pp. 147–185.

NRC (2000), *Multi-Agency Radiation Survey and Site Investigation Manual (MARSSIM)*, NUREG-1575, Revision 1 edn, Washington, DC.

Osborne, J. G. (1942), 'Sampling errors of systematic and random surveys of cover-type areas', *Journal of the American Statistical Association* **37**(218), 256–264.

Särndal, C., Swensson, B. & Wretman, J. (2002), *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Scheaffer, R., Mendenhall, W. & Ott, L. (1995), *Elementary Survey Sampling*, 5th edn, Duxbury, Pacific Groove.

Sen, A. R. (1953), 'On the estimate of the variance in sampling with varying probabilities', *Indian Society of Agricultural Statistics* **5**, 119–127.

Sengupta, S. & Chattopadhyay, S. (1987), 'A note on circular systematic sampling', *Sankhyā: The Indian Journal of Statistics, Series B,* **49**(2), 186–187.

Sethi, V. K. (1965), 'On optimal pairing of units', *Sankhyā: The Indian Journal of Statistics, Series B,* **27**, 315–320.

Shiue, C. J. (1960), 'Systematic sampling with multiple random starts', *Forest Science* **6**, 42–50.

Singh, D., Jindal, K. K. & Garg, J. N. (1968), 'On modified systematic sampling', *Biometrika* **55**(3), 541–546.

Singh, D. & Singh, P. (1977), 'New systematic sampling', *Journal of Statistical Planning and Inference* **1**(2), 163–177.

Singh, P. & Garg, J. N. (1979), 'On balanced random sampling', *Sankhyā: The Indian Journal of Statistics, Series C,* **41**, 60–68.

Stuart, A. (1976), *Basic Ideas of Scientific Sampling*, 2nd edn, Charles Griffin and Company Ltd, London.

Sudhakar, K. (1978), 'A note on circular systematic sampling design', *Sankhyā: The Indian Journal of Statistics, Series C,* **40**(1), 72–73.

Tornqvist, L. (1963), 'The theory of replicated systematic cluster sampling with random start', *Review of the International Statistical Institute* **31**(1), 11–23.

Tukey, J. W. (1950), 'Some sampling simplified', *Journal of the American Statistical Association* **45**(252), 501–519.

Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*, Almqvist & Wiksell, Stockholm.

Wolter, K. (2007), Variance estimation for systematic sampling, *in* 'Introduction to Variance Estimation', Springer, New York, pp. 298–353.

Wu, C. F. J. (1984), 'Estimation in systematic sampling with supplementary observations', *Sankhyā: The Indian Journal of Statistics, Series B,* **46**(3), 306–315.

Yates, F. (1948), 'Systematic sampling', *Philosophical Transactions of the Royal Society of London, Series A* **241**(834), 345–377.

Yates, F. & Grundy, P. M. (1953), 'Selection without replacement from within strata with probability proportional to size', *Journal of the Royal Statistical Society, Series B (Methodological)* **15**(2), 253–261.

Zinger, A. (1963), 'Estimations de variances avec échantillonnage systématique', *Revue de Statistique Appliquée* **11**(2), 89–97.

Zinger, A. (1964), 'Systematic sampling in forestry', *Biometrics* **20**(3), 553–565.

Zinger, A. (1980), 'Variance estimation in partially systematic sampling', *Journal of the American Statistical Association* **75**(369), 206–211.