SOME ASPECTS OF RANKED SET SAMPLING

DISSERTATION

Submitted in partial fulfillment of the requirement for the award of the degree of

Master of Philosophy

IN

STATISTICS

BY

SABA RIYAZ

Under the Supervision of

DR. TARIQ RASHID JAN
(Assistant Professor)

Post-Graduate Department of Statistics
Faculty of Physical and Material Sciences
University of Kashmir, Srinagar

September (2012)

Dedicated to My Parents, who Have
Encouraged Me throughout My Life,
in All of My Educational Endeavors

Post-Graduate Department of Statistics

University of Kashmir, Srinagar- 190006

Certificate

This is to certify that the scholar **Ms. Saba Riyaz** has carried out the present dissertation entitled "**Some Aspects of Ranked Set Sampling**" under my supervision and the work is suitable for submission for the award of the Degree of Master of Philosophy in Statistics. It is further certified that the work has not been submitted in part or full for the award of this or any other degree elsewhere.

Dr. Tariq Rashid Jan

(Supervisor)

ACKNOWLEDGEMENT

It would not have been possible to write this thesis without support and help of people around me, to only some of whom it is possible to give particular mention here.

Above all, I would like to thank my husband Mirza Villayat Zaman and his parents for their personal support, affection and great patience at all times to carry my work forward. My parents, brother and sisters have given me immense support throughout, for which my mere expression of thanks does not suffice.

This dissertation would not have been possible without guidance, input, support and patience of my supervisor Dr. Tariq Rashid Jan, Asst. Professor in the Department of Statistics of Kashmir University, whose transformation of knowledge and wisdom remains unsurpassed. Advice and support of Prof. Aquil Ahmad, Head of the Department, has been invaluable, both at academic and personal level, for which I am extremely grateful.

I would like to acknowledge and thank senior faculty in the department, Prof. Anwar Hassan, Dr. M.A.K Baig and Dr. Sheikh Parvaiz Ahmad, my fellow research scholars and friends for their constant encouragement during the process.

I would like to acknowledge academic and technical facilities provided by University of Kashmir which I have benefitted from, since the days of my post graduation; and the team at Virus Computers who helped to give a final shape to this document.

Finally, my son Yawar Abass progressed from crawling to standing on his feet while I worked day and night on my dissertation. Enjoying his smile and historical moments of getting on his feet, I acknowledge his patience.

Saba Riyaz

PREFACE

Cost-effective sampling methods are of a major concern in statistics, especially when the measurement of the characteristic of interest is costly and / or time-consuming. In the early 1950's in seeking to effectively estimate the yield of pasture in Australia, McIntyre proposed a sampling method which later came to be known as ranked set sampling (RSS). The notion of RSS provides an effective way to achieve observational economy under certain particular conditions. Although the method remained dormant for a long time, its value was rediscovered in the last 25 years or so because of its cost-effective nature. There have been many new developments from the original idea of McIntyre, which made the method applicable in a much wider range of fields than originally intended. More and more applications of RSS have been cited in the literature.

The basic premise for RSS is an infinite population under study and the assumption that a set of sampling units drawn from the population can be ranked by certain means rather cheaply without the actual measurement of the variable of interest, which is costly and / or time-consuming. This assumption may look rather restrictive at first sight, but it turns out that there are plenty of situations in practice where this is satisfied.

The topic of this dissertation is ***'SOME ASPECTS of RANKED SET SAMPLING '****. This dissertation is divided into five chapters with a comprehensive bibliography given at the end.*

Chapter-I presents a brief review of various types of sampling methods and the structural differences between ranked set samples and simple random samples are discussed.

Chapter-II deals with the estimation of parameters of Generalized Geometric distribution using Ranked Set Sampling procedure. These estimates are

compared with the ordered least squares estimates and it is shown that the relative precisions of estimators using Ranked set sampling are higher than those of the ordered least squares estimation.

Chapter-III deals with estimation of the means of Bivariate Normal distribution using Moving Extreme Ranked Set Sampling with concomitant variable. The estimators obtained are compared to their counterparts based on simple random sampling thus showing that they are more efficient. The issue of robustness of the procedure is addressed and real trees data set has been used for illustration.

Chapter-IV deals with estimation of Simple Linear Regression Model using L Ranked Set Sampling. It is shown that estimated regression model based on LRSS is highly efficient compared to the estimators based on Simple Random Sampling, Extreme Ranked Set Sampling and Ranked Set Sampling.

Chapter-V In this chapter, a new RSS method i.e., Stratified Quartile Ranked Set Sampling (SQRSS) is compared with simple random sampling (SRS), stratified simple random sampling (SSRS) and stratified ranked set sampling (SRSS) methods. It is shown that the SQRSS estimators are unbiased of the population mean of symmetric distributions and that the SQRSS is more efficient than its counterparts using SRS, SSRS and SRSS based on the same number of measured units..

Chapter-VI In this chapter some novel applications of Ranked Set sampling have been discussed.

Table of Contents

Chapter 1

Sampling Designs

1.1 Introduction

"Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring reliability of useful statistical information through the theory of probability"

- Deming *(1950)*

Our knowledge, our attitudes, and our actions are based to a very large extend on samples. A person's opinion of an institution that conducts thousands of transactions every day is often determined by the one or two encounters he has had with the institution in the course of several years. Travelers, who spend 10 days in a foreign country and then proceed to write a book telling the inhabitants how to revive their industries, reform their political system, balance the budget, and improve the food in their hotels, are a familiar figure of fun. But in a real sense, they differ from the political scientist who devotes 20 years to living and studying in the country only in that they base their conclusions on a much smaller sample of experience and are less likely to be aware of the extent of their ignorance.

Sampling, or sample survey is a method of drawing an inference about the characteristic of a population or universe by observing only a part of population. Such methods are extensively used by government bodies throughout the world for assessing, among others, different characteristics of national economy as are required for taking decisions regarding the imposition of taxes, fixation of prices and minimum wages, etc and for the planning and projection of future economic structure. Thus, surveys are conducted for estimation of yield rates and acreages under different crops, estimation of value added by manufacture in the industries sector, estimation of number of unemployed persons in the labor forces, construction of cost of living indices for persons in different professions, and so on.

The enumeration of population by sampling methods, proposed by Laplace in 1783, came into widespread use only by the mid-thirties of this century. During the last few decades there has been tremendous development in the methods of analysis of and drawing inference from the data obtained through survey sampling.

1.2 Classical Types of Sampling Designs

For achieving desired correct results from a sample survey, the execution of sample design is of utmost importance and hence proper selection of the sampling methods becomes imperative. The sampling techniques can be broadly classified into following categories; viz Probability and Non-Probability sampling, which are enumerated as follows:-

1 - Probability Sampling

a) Simple Random Sampling

b) Stratified Sampling

c) Systematic Sampling

d) Cluster Sampling

e) Multi – Stage Sampling

f) Multi – phase Sampling

g) Area Sampling

2 - Non- Probability Sampling

a) Convenience Sampling

b) Quota Sampling

c) Judgmental Sampling.

Probability Sampling

The Probability sampling is the scientific technique which draws sample from the population based on the application of probability methods, wherein each unit of the probability has some predefined probability of inclusion of an event into the drawn sample.

The samples will therefore be selected in the following manner

- Each unit is drawn on the basis of randomness
- Each unit has the same chance of being selected.
- Probability of selection of a unit is proportional to the sample size.

Thus the samples are drawn based on random procedure and not on any judgmental method.

These sampling techniques are described below:

a) Simple Random Sampling

Simple random sampling (SRS) is the most basic form of probability sampling and the most commonly used sampling approach for collecting data from a population with the goal of making inferences about unknown features of the population. The observations in an SRS are mutually independent if the sampling is from an infinite population or with replacement from a finite population and they are dependent if sampling from a finite population without replacement. In either situation, however, there is a probabilistic guarantee that each measured observation in an SRS can be considered representative of the population.

In a simple random sampling, the elements of the sample are drawn at random and such that each and every unit of the population has an equal chance of being selected. If we have a population of N elements we can select n sets of elements out of such a population (where n is fairly large), and the possible sets of n elements will be ${}^{N}C_{n}$, following the same probability of selection for every such set of elements. The basic aim is to achieve randomness in drawing the elements of a sample to ensure all possible samples to have the same chance of being selected. We can use either lottery system or the Random Number table system, both either with replacement of the drawn number or without replacement.

In lottery system, all the elements of the population are allotted identical identification; say some type and size of paper with elements numbers written on each. After proper folding the papers in the same manner and thorough mixing of these papers, we can choose any paper at random without any bias either through the container system or taking out each paper blindly. When N is very large, this method becomes cumbersome and difficult to manage. In that case, we use the method of Random Number Tables. From the Random Number Tables, the numbers can be selected from the list, where numbers have already been arranged in Random order. We can select Numbers either through the rows or through columns. Various Random Number Tables in use are:

(i) Tippett's (1927) random number tables of 41,600 digits,

(ii) Fisher and Yates' (1938) random number tables of 15,000 digits,

(iii) Kendall and Smiths (1939) table of random number of 100,000 digits,

(iv) Random number tables of 1 million digits prepared by Rand Corporation (1955),

(v) C.R. Rao, S.K. Mitra, A. Matthai, and K.G. Ramamurthy (1966), table of random numbers of 20,000 digits.

Simple random sampling is most useful when the population of interest is relatively homogeneous. The main advantages of this design are:

- It provides statistically unbiased estimates of the mean, proportions, and variability.
- It is easy to understand and easy to implement.
- Sample size calculations and data analysis are very straightforward

b) Stratified Sampling:

This design is useful for estimating a parameter when the target population is heterogeneous. In stratified sampling, the target population is separated into non-overlapping strata, or subpopulations that are known or thought to be more homogeneous, so that there tends to be less variation among sampling units in the same stratum than among sampling units in different strata. Strata may be chosen on the basis of pre-existing information or professional judgment about the units of the population.

In a sample survey the necessity of stratification is often dictated by administrative requirements or convenience. For a state wise survey, for instance, it is often convenient to draw samples independently from each district and carry out survey operations for each district to take care of the survey operations under its jurisdiction. Thus for administrative convenience, each district may be treated as a stratum. Since a stratified sample consists of units selected separately from each stratum, such a sample is expected to be a better representation for the population than a simple random sample selected from the whole universe. In practice, the population

often consists of heterogeneous units (with respect to the character under study). For a socio – economic survey, for instance, people may live in rural areas, urban localities, ordinary domestic houses, hostels, hospitals, jails, etc. It is evident that the sampling problem will be different for these different sectors of the population and each such sector should be treated as a separate stratum. Again, administrators may require estimates for different strata separately along with the estimate for the population as a whole. This can be achieved through stratified sampling.

Advantages of this sampling design are:

- It has potential for achieving greater precision in estimates of the mean and variance.
- It allows computation of reliable estimates for population subgroups of special interest.
- Greater precision can be obtained if the measurement of interest is strongly correlated with the variable used to make the strata.

c) Systematic Sampling:

A very simple form of sampling for its design and execution is used when the numbers of population are arranged in an order, the order corresponding to consecutive numbers. In this type of sampling, the first sample unit is selected at random and the remaining units are automatically selected on a definite sequence at equal spacing from one another. This design provides a practical, easy and convenient way often used in field surveys for designating sample locations and ensures uniform coverage of the population.

d) Cluster Sampling

Sometimes it is not possible to have a list of all the units of study in the population so that drawing a simple random sample is not feasible. However, a list of some bigger units each consisting of several smaller units (study units) may be available from which a sample may be drawn.

In cluster sampling the population is first divided into a number of non-overlapping clusters. A cluster is a collection of a number of smaller units which are

the ultimate objects of study and in respect of which survey results are to be computed. We shall often refer to these smaller units as elementary units. A simple random sample of clusters is selected and all the elementary units belonging to the selected clusters are surveyed.

Cluster sampling is often used in agricultural surveys for determination of area under crops where a randomly selected point on a cadastral map determines a cluster of plots of a specified total size. In a survey on the industrial products, batches of products coming out from a production process within specified lengths of time may from clusters.

e) Multi-stage Sampling:

Use of cluster sampling technique under certain circumstances is cheaper, but it is less efficient than the individual sampling. Thus as a combination, we can use Multistage Sampling, in which we can select cluster samples and then studying only a sample of units in each cluster. This is called Two-stage Sampling. Similar concept can be extended to bring in Multistage sampling, where sampling units at each stage being done from each of the sampling units.

f) Multi-phase Sampling:

This type of sampling is adopted when sampling units of the same type are the objects of different phases of observation. In this case all the units of a phase in a sample are studied with respect to the same characteristics. This concept can be extended to Multiphase sampling. In this case information collected during one phase is then used in the second or subsequent phases.

g) Area Sampling:

When we use cluster sampling concept for the elementary units of population in a particular geographical area, it is called Area Sampling. In this case, we can study the community behavior index of a particular community living in a particular locality or part of the country, but selection of sample in each area should be random for enumerated elements. Thus the enumeration of elements is necessary only in the limited number of selected areas.

Non-Probability Sampling

As against the Probability Sampling, the non-Probability Sampling is a procedure of selection of a sample without the use of randomization. It is based on convenience or judgment and hence is likely to be biased. The sampling variation in such a case is very uncertain and cannot be estimated.

In this category we can have samplings done either on the convenience basis such as picking up names from the telephone directory or on the basis of quota such as quota of candidates under one category fixed for the interview. We can first sample out the total population based on categories as per quota list and then selection of these lists without any fixed procedure. We can also follow a judgmental method of non-probability sampling, when the sample elements are either picked up on previous experience basis or with no set rule procedure, but based on hunch. It is also called as opinion sampling. This is used only when there is better evidence or selection procedure in vogue.

1.3 Ranked Set Sampling

1.3.1 Introduction

Despite the assurance that there is a probabilistic guarantee that each measured observation in a simple random sample can be considered representative of the population, there remains a distinct possibility that a specific SRS might not provide a truly representative picture of the population. With this issue in mind, statisticians have developed a variety of ways to guard against obtaining such unrepresentative samples. Sampling designs such as stratified sampling, cluster sampling, etc., all provide additional structure on the sampling process to improve the likelihood that the collected sample data provide a good representation of the underlying population. A secondary goal in most data collection settings is to minimize the costs associated with obtaining the data, including both the cost of initially selecting the population units for measurement and in making the actual measurements. Ranked set sampling (RSS) is a relatively recent development that addresses both of these issues. It uses additional information from the population to provide more structure to the data collection process and increases the likelihood that the collected sample data will, in fact, provide a representative picture of the population. In addition, it is designed to

minimize the number of measured observations (i.e., the sample size) required to achieve the desired precision in making inferences about the population. Ranked set sampling has a potential to be used in environmental, ecological, biological, medical, social, agricultural sciences as well as business applications.

Ranked Set Sampling is an innovative sampling design originally developed by McIntyre (1952) for situations where taking the actual measurements for sample units is difficult (e.g., costly, destructive, time-consuming) but there are mechanisms readily available for either informally or formally ranking a set of sample units. The unique feature of ranked set sampling is that it combines simple random sampling with the field investigator's professional knowledge and judgment to pick places to collect samples. The use of ranked set sampling increases the chance that the collected samples will yield representative measurements; that is, measurements that span the range of low, medium, and high values in the population. This results in better estimates of the mean as well as improved performance of many statistical procedures. Moreover, ranked set sampling can be more cost-efficient than simple random sampling because fewer samples need to be collected and measured.

The use of professional judgment in the process of selecting sampling locations is a powerful incentive to use ranked set sampling. Professional judgment is typically applied by visually assessing some characteristic or feature of various potential sampling locations in the field, where the characteristic or feature is a good indicator of the relative amount of the variable of interest that is present.

In particular, McIntyre was interested specifically in improving the precision in estimation of average yield from larger plots of arable crops without a substantial increase in the number of fields from which detailed expensive and tedious measurements needed to be collected. The RSS approach, however, is applicable in any situation where minimizing sample size while retaining precision of our statistical inferences is important. For lots of cases such as the one McIntyre had, RSS can replace the use of SRS in these designs, to the benefit of sample estimates.

1.3.2 Significance of Ranked Set Sampling

Typically the most expensive and time consuming part of this process is laboratory analysis. For example, suppose we wish to estimate the mean bone density

of students at a university. The lab work is so costly that the budget is only enough to analyze samples from four students. Furthermore, in order to acknowledge the inherent uncertainty, we need to present this estimate with a confidence interval within which we expect the true population mean to lie with desired confidence. The simplest way to obtain our sample is to randomly select four students from the university's population, then take bone samples from them and measure their bone densities. While the arithmetic average of the four bone densities is an unbiased point estimate of the population mean, the associated confidence interval can be very large, reflecting the high degree of uncertainty with estimating a population mean from only four measurements. This is because we have no control over which individuals of the population enter the sample. The only way to overcome such a problem with simple random sampling is to increase the sample size which sometimes is not realistic and applicable. RSS was proposed to help improve efficiency in such cases without increasing sample size. In the last few years there has been an explosion of interest in and a tremendous amount of methodological development of ranked set sampling procedures. One reason for this increase is the recognition by statisticians of the need for more cost-effective sampling procedures, such as those that use a priori knowledge or can otherwise provide the needed information with a significant reduction in cost over the more traditional simple random sampling approaches. Nowhere is this need more evident than in the field of environmental monitoring and assessment. In the past, the assessment of most environmental problems was relatively straight forward. Many of the major environmental problems could be detected using the human senses (a river was burning because of chemical wastes being discharged directly into the river; the air in large cities could be seen and smelled etc). In the last forty years, our knowledge of anthropogenic pollution and our ability to measure minute quantities (parts per billion or trillion) of toxic chemicals in our environment has dramatically improved. Now we have identified hundreds of man-made toxic chemicals in our environment. The cost of measuring and monitoring these chemicals and assessing their environmental impact is extremely high. It requires sophisticated measurement and careful sampling of large potentially impacted areas. These measurements can range in cost from a few dollars to several thousand per sample. The large range and diversity of media from which samples

must be drawn create additional costs. Thus any sampling method which allows fewer observations to provide the same information (currently known as `Observational Economy') is particularly valuable in environmental applications. Ranked set sampling can provide observational economy under very special circumstances - namely, when sample units can be easily and inexpensively gathered and ranked among themselves, but are expensive to measure accurately. This situation arises very naturally in agriculture and forestry, where the earliest applications occurred. It is easy and cheap to judge fairly accurately by observation, for example, which of several trees contains the largest volume of wood, which the next largest, and so on down to the least. It is much more expensive to actually measure the amount of wood in each. The same type of circumstance arises in some environmental applications. For example, consider the problem of assessing the status of a hazardous waste site, (i.e., determining if a site has toxic chemicals in excess of a set standard). We often know a great deal about the sites from records, photos and physical characteristics. This knowledge will allow us to rank the areas from which we will sample in terms of high to low levels of toxic pollution. This would limit the number of expensive samples necessary to assess the status of the hazardous waste site (i.e., does it require clean up or not).

The core idea of Ranked-set sampling is to create hypothetical stratified samples based on ranks. For example, we could randomly select two trees and judge by expert opinion which tree contains more volume of wood. The smaller of the two is selected for accurate measurement later. Next, select another two trees, this time select the larger one. Continue this procedure until 20 trees being selected and ranked; 10 of them are selected for accurate measurements. As we can see from the above example, one main disadvantage of simple random sampling is when a sample of units is drawn at random from a population the units may not constitute as a representative sample of the population. For example, when we draw a random sample of 5 students from the population of all students of a university in order to estimate the average weight, it is possible with positive probability that all students in the sample are obese.

1.3.3 A historical note

Ranked set sampling was basically the idea first proposed by McIntyre in his effort to find a more efficient method to estimate the yield of pastures. Measuring yield of pasture plots requires mowing and weighing the hay which is a very time-consuming process. But an experienced person can rank by eye inspection fairly accurately the yields of a small number of plots without actual measurement. McIntyre adopted the following process. A random sample of m pasture lots is ranked by visual inspection with respect to the amount of yield. From this sample, the lot with rank 1 is taken for cutting and weighing. Then again a random sample is taken and ranked. From the second sample, the lot with rank 2 is taken, and so on. When all the selected lots for ranks from 1 to m have been taken and measured, the cycle is repeated over again and again until a total of r cycles are completed. McIntyre illustrated the gain in efficiency by a computation involving five distributions. He observed that the relative efficiency, in not much less than $(m+1)/2$ for symmetric or moderately asymmetric distributions, and that the relative efficiency decreases with increase in the asymmetry of the underlying distribution but is always greater than 1. He also mentioned the problems of optimal allocation of measurements among the ranks and the problems of ranking errors and possible correlation among the units within a set, etc. Since only a fraction of the sampled units are quantified, the method presumes that the physical acquisition of units is cheap as compared with their quantification.

McIntyre's proposal remained buried in literature for over a decade until Halls and Dell (1966) conducted a field trial evaluating the applicability of RSS to the estimation of forage yields in a pine-hardwood forest. The term 'Ranked Set Sampling' which is in current use, was coined by them. The first theoretical result for ranked set sampling was given by Takahasi and Wakimoto (1968). They proved that when ranking is perfect, the ranked set sample mean is an unbiased estimator of the population mean, and the variance of the ranked set sample mean is always smaller than the variance of the mean of a simple random sample of the same size. Dell and Clutter (1972) also obtained similar results but without restricting to the case of perfect ranking. They demonstrated that, for comparable sample sizes, the RSS

procedure results in more accurate parameter estimators than simple random sampling. Equivalently, RSS requires fewer measured observations than SRS to attain the same level of precision. The improvement in precision comes about because RSS adds structure to the data, in the form of the sampler's ranking, that is absent in SRS. This added structure is similar to stratifying the population prior to taking a SRS. Whereas stratified SRS uses auxiliary information from the entire population, however, RSS uses auxiliary information from only the units in the initial sample; it does not require the availability of auxiliary information for all units in the population.

Dell and Clutter (1972) and David and Levine (1972) were the first to give some theoretical results on imperfect ranking. Stokes (1976) (1977) considered the use of concomitant variables in RSS. Till then the attention had been focused mainly on the non-parametric estimation of population mean. Stokes (1980a) considered the estimation of population variance and the estimation of correlation coefficient of a Bivariate Normal population based on RSS. Many procedures were yet to be investigated and developed.

The middle of 1980's was a turning point in the development of the theory and methodology of RSS. Since then, many variations of the ranked set sampling have been proposed and various statistical procedures for non-parametric and parametric estimation have been investigated and a sound theoretical foundation has been laid.

Several researchers have studied ranked set sampling, but a complete review of applications and theoretical framework on RSS is available in Patil et al. (1994a), Kaur et al. (1995), and Johnson et al. (1996). Patil et al. (1993a) studied the RSS method when sampling is from a finite population. They gave explicit expressions for the variance and relative precision of RSS estimators for several set sizes when the population follows a linear or quadratic trend. Patil et al. (1993b) studied the relative precision of ranked set sampling estimators with the regression estimator when the ranking is done on the basis of an auxiliary variable. The same authors (1994a) classified various papers on RSS into three groups: (i) theory, (ii) methods, and (iii) applications. They reviewed various aspects of RSS in a single unified notation. For additional applications of RSS and its multivariate considerations see Johnson et al.

(1993), Patil et al. (1994c), Patil et al. (1994b), and Gore et al. (1993). Some of these references also discuss the problems in implementing RSS. A comprehensive bibliography on RSS up to 1999 was provided by Patil et al. (1999). The field of RSS has been in its florescence period in the past few years and many new developments in RSS have been made. A few are mentioned as follows. The Fisher information theory of RSS has been established; Chen (2000a) and Bai and Chen (2003). A host of new RSS schemes such as the adaptive RSS and multi-layer RSS using either variables of interest or concomitant variables have been devised, Al-Saleh and Zheng (2002), Chen (2002) and Chen and Shen (2003). Optimal RSS designs have been developed for various problems such as estimation of parameters in parametric families, estimation of quantiles and distribution-free tests, see Chen and Bai (2000), Chen (2001a) and Ozturk and Wolfe (2000a, b, c, 2001). The issues with cost in RSS have been reasonably handled; see Nahhas et al. (2002) and Wang et al. (2004). The RSS has been considered for many more statistical procedures such as density estimation, quantile estimation, U-statistics, M-statistics, variance estimation and rank regression etc., Chen (1999, 2000b), Presnell and Bohn (1999), Zhao and Chen (2002), MacEachern et al. (2002) and Ozturk (2002). Special RSS schemes have been applied to the designs for treatment comparisons, Ozturk and MacEachern (2004) and Chen et al. (2006a). A comprehensive coverage of RSS which includes the most recent developments of RSS was given in a monograph by Chen et al. (2004).

Several variations of ranked set sampling method have been proposed and developed by researchers to come up with more efficient estimators of a population mean. A few references are given as follows. Samawi et al. (1996) introduced Extreme Ranked Set Sampling and obtained an unbiased estimator of the mean which outperforms the usual mean of a simple random sample of the same size for symmetric distributions. Muttlak (1997) suggested Median Ranked Set Sampling to increase the efficiency and to reduce ranking errors over ranked set sampling method and proved its better performance in estimating the mean of a variable of interest for some symmetric distributions. Hossain and Muttlak (1999) introduced Paired Ranked Set Sampling, Al-Saleh and Al-Kadiri (2000) introduced Double Ranked Set Sampling, Hossain and Muttlak (2001) introduced Selected Ranked Set Sampling and Al-Saleh and Al-Omari (2002) introduced Multistage Ranked Set Sampling. Muttlak

(2003a) proposed Percentile Ranked Set Sampling and Muttlak (2003b) proposed the use of Quartile Ranked Set Sampling (RSS) for estimating the population mean. Jemain and Al-Omari (2006) suggested Double Quartile Ranked Set Sampling (DQRSS) for estimating the population mean. Two-stage Median Ranked Set Sampling was developed by Jemain et al. (2007a). Al-Nasser (2007) introduced L-Ranked Set Sampling Design as a generalization of some of the above mentioned ranked set type sampling methods and proved the optimal property of his proposed estimators for symmetric family of distributions. Al-Nasser and Radaideh (2008) used L Ranked-Set Sampling (LRSS) to estimate a simple linear regression model. They showed that the estimated regression model based on LRSS is highly efficient compared to the estimators based on simple random sampling, Extreme ranked set sampling and ranked set sampling. Balanced Group Ranked Set Sampling was developed by Jemain et al. (2009). In addition, various modifications of RSS have been suggested for the estimation of population ratio. Samawi and Muttlak (2001), for example, used Median Ranked Set Sampling to estimate the population ratio. Samawi and Tawalbeh (2002) introduced Double Median Ranked Set Sampling (DMRSS) method for estimating the population mean and ratio. More recently, Stratified Percentile Ranked Set Sampling (SPRSS) method has been suggested for estimating the population mean by Al-Omari et al. (2011). They compared the SPRSS method with the Simple Random Sampling (SRS), Stratified Simple Random Sampling (SSRS) and Stratified Ranked Set Sampling (SRSS). It was shown that SPRSS estimator is an unbiased estimator of the population mean of symmetric distributions and is more efficient than its counterparts using SRS, SSRS and SRSS based on the same number of measured units. Stratified Quartile Ranked Set Sampling (SQRSS) has been given by Syam et al. (2012) and it has been shown that the SQRSS estimators are unbiased of the population mean of symmetric distributions.

1.3.4 Description of Ranked Set Sampling

The original method of getting a Ranked set sample as obtained by McIntyre is described as follows. First, a simple random sample of size *m* is drawn from the population and the *m* sampling units are ranked with respect to the variable of interest, say height (X), by judgment *without* actual measurement. Then the unit with rank 1 is

identified and measured for X. The remaining units of the sample are discarded. Next, another simple random sample of size *m* is drawn and the units of the sample are ranked by judgment, the unit with rank 2 is measured for X and the remaining units are discarded. The process is continued until a sample of size *m* is obtained and ranked and the unit with rank *m* (highest rank) is taken for measurement of X. This whole process is called a cycle. The cycle is then repeated *r* times and it yields a ranked set sample of size *n=rm.*

The essence of RSS is conceptually similar to the classical stratified sampling. RSS can be considered as post-stratifying the sampling units according to their ranks in the sample.

For the General RSS scheme we select *m* random sets each of size *m* from the target population. In practice, *m* usually takes values such as 2, 3, or 4. Each set is then ranked by convenient (cheap) method in context of the variable of interest.

In Matrix notation, we have

$$\begin{matrix} X_{(1)1} & X_{(1)2} & \cdots & X_{(1)m} \\ X_{(2)1} & X_{(2)2} & \cdots & X_{(2)m} \\ \cdots & \cdots & \cdots & \cdots \\ X_{(m)1} & X_{(m)2} & \cdots & X_{(m)m} \end{matrix}$$

After ranking, only the diagonal units are selected and actually measured. This constitutes the ranked set sample. That is, we have only measures $X_{(1)1}$, $X_{(2)2}$, …., $X_{(m)m}$, by obtaining the unit with the smallest rank from the first row, the second smallest rank from the second row and so on until the largest unit from the m^{th} row. This represents one cycle of RSS. We can repeat the whole procedure *r* times to get a RSS of size *n = rm*. It is to be noted here that RSS requires m^2 units to be taken, but only *m* of them are actually measured.

The ranks which the units in a set receive may not necessarily tally with the numerical orders of their latent X values. If the ranks do tally with the numerical orders, the ranking is said to be perfect, otherwise it is said to be imperfect. When ranking is perfect, the ranks are put in parentheses, else they are put in brackets. Thus

$X_{(m)}$ and $X_{[m]}$ is the generic notation for measurements with rank *m* when ranking in perfect and imperfect, respectively.

1.3.5 Important Mathematical Results

For a simple random sample of size n, i.e., $x_1, x_2, \ldots x_n$, from a population with mean μ and variance σ^2 the traditional non-parametric estimator of μ is given by

$$\bar{X}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

With the variance

$$V(\bar{X}_{SRS}) = \frac{\sigma^2}{n}$$

Now with RSS, for $n=rm$ we have the following,

$$\bar{X}_{RSS} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} X_{(i)ij}$$

And

$$V(\bar{X}_{RSS}) = \frac{\sigma^2}{n} - \frac{1}{rm^2}\sum_{i=1}^{n}(\mu_{(i)} - \mu)^2$$

Where $\mu_{(i)}$ is the mean of the *i*th ranked set, and is given by $\mu_{(i)} = \frac{1}{r}\sum_{j=1}^{r} x_{i(i)j}$.

We can see a variance reduction factor of $\frac{1}{rm^2}\sum_{i=1}^{n}(\mu_{(i)} - \mu)^2$ in the expression for $V(\bar{X}_{RSS})$ above, associated with $\bar{X}_{RSS}$. As the rankings become more accurate, the term $\sum_{i=1}^{n}(\mu_{(i)} - \mu)^2$ becomes larger, and the overall variance of $\bar{X}_{RSS}$ decreases.

Takahasi and Wakimoto (1968) were the first to study the mathematical theory of RSS in detail and they also defined Relative Precision (RP) as

$$RP = \frac{V(\bar{X}_{SRS})}{V(\bar{X}_{RSS})}$$

They also showed that

$$0 \leq RP \leq \frac{m+1}{2}$$

Under the equal allocation of each order statistic, RSS will always result in as precise an estimate as SRS, if not better.

1.3.6 Balanced Ranked Set Sampling

In a balanced RSS, the number of measurements made on each ranked statistic is the same for all the ranks. A balanced ranked set sampling produces a data set as follows:

$$\begin{matrix} X_{[1]1} & X_{[1]2} & \dots & X_{[1]m} \\ X_{[2]1} & X_{[2]2} & \dots & X_{[2]m} \\ \dots & \dots & \dots & \dots \\ X_{[m]1} & X_{[m]2} & \dots & X_{[m]m} \end{matrix}$$

Here all the $X_{[k]i}$'s are mutually independent and the $X_{[k]i}$'s in the same row are identically distributed. The measured observations $X_{[1]1}$, $X_{[2]2}$ …, $X_{[m]m}$ constitute a balanced ranked set sample of size m, where the descriptor 'balanced' refers to the fact that we have collected one judgment order statistic for each of the ranks 1,2,…, *m*.

1.3.7 Unbalanced Ranked Set Sampling

An alternative to balanced RSS is unbalanced RSS. Instead of having all of the ranks represented equally in the subsample, one could measure the variable of interest on certain ranks more frequently than on others.

An unbalanced RSS is one in which the ranked order statistics are not quantified the same number of times. An unbalanced ranked set sample is given as follows:

$$\begin{matrix} X_{[1]1} & X_{[1]2} & \dots & X_{[1]m_1} \\ X_{[2]1} & X_{[2]2} & \dots & X_{[2]m_2} \\ \dots & \dots & \dots & \dots \\ X_{[m]1} & X_{[m]2} & \dots & X_{[m]m_k} \end{matrix}$$

Here all the $X_{[k]i}$'s are independent and the $X_{[k]i}$'s with the same j are also identically distributed.

There are situations where measuring differing numbers of the various judgment order statistics (unbalanced RSS) can lead to improved RSS procedures. The choice of set size remains important for this unbalanced RSS setting but the concept of a cycle is no longer necessary, since we do not need to have the same measurement counts for every judgment order statistic.

Just as with balanced RSS, the measured units in an unbalanced RSS are mutually independent, but now the numbers of measured units in each of the ranks are not necessarily equal. Balanced RSS corresponds to the special case where $m_1 = m_2 = \ldots = m_k$.

There are a number of factors to consider when deciding whether to use balanced or unbalanced RSS, mostly related to the type of inferences of interest and what is known about the shape of the underlying distribution. There has been a substantial amount of research on the best way to allocate observations to the judgment order statistics in unbalanced RSS. The optimal allocation depends on the parameter being estimated and the statistical inference being performed. Stokes (1995) and Bhoj (1997) were the first to demonstrate the optimality of unbalanced RSS for estimation of a location parameter within the context of a parametric family and Kaur et al. (1997) obtained corresponding results for positively skewed distributions. Ozturk and Wolfe (2000a) provided the optimal allocation for a variety of nonparametric test procedures. Wolfe (2004) described an RSS procedure where the measured subsample consisted of ranked units judged to be the medians of their respective sets. Such a method is ideal for estimating the population median (although other parameters might be difficult to estimate once the data are collected this way). Chen et al. (2006b) described how Neyman allocation is optimal for obtaining an unbalanced RSS when one is interested in estimation of a population proportion.

Early RSS research assumed that the ordering assigned by the researcher corresponded perfectly to how the items would have been ordered if the researcher had used the actual value of the variable of interest to rank them. In this situation, the judgment order statistics are the true order statistics. Dell and Clutter (1972) considered the more realistic scenario in which the rankings are not perfect. It is easy to imagine that visual judgment can lead to imperfect rankings, particularly among

units with ranks in the middle of the ordered set. These items may be so similar in attributes that the researcher has difficulty ordering them. MacEachern et al. (2004) developed a method that allows the researcher to assign probabilities to the ranks instead of forcing him or her to assign a single distinct rank to each item. When using an auxiliary variable to estimate a ranking based on the variable of interest, imperfect rankings can occur when the two variables are not perfectly correlated.

Regardless of whether or not the rankings are perfect, $\hat{\mu}_{rss}$ is an unbiased estimator for the population mean so long as the errors are not related to the ranking procedure (Dell and Clutter, 1972). The more accurate the rankings, the more precise this estimator will be. The performance of the RSS estimator vis-a-vis SRS can be evaluated by examining the relative precision of the estimators. Nahhas et al. (2002) showed that the relative precision of RSS (i.e., the ratio of the variance of the mean estimator under RSS to the variance of the mean estimator under SRS) improves as the rankings become more accurate.

For a given number of quantified observations, the precision of the RSS estimator for the mean, proportion, or total is at least as good as that of the SRS estimator. A SRS is equivalent to a RSS when the ranks are assigned randomly. Thus, as long as the ranking of the initial sample is better than a random ranking, a RSS provides an estimator with less variability than the estimator from a SRS of the same size.

The advantages of RSS will be maximized, therefore, when the researcher chooses a ranking method where

(1) the rankings are perfect or close to perfect, and

(2) the cost of collecting and ranking the initial observations is significantly lower than that of quantifying the selected observations.

Amarjot et al. (1996) developed a cost model for comparing RSS to stratified simple random sampling. Nahhas et al. (2002) provided a method for determining the optimal set size taking into account the various costs associated with RSS. Wang et al. (2004) evaluated the cost-effectiveness of quantifying multiple units from the same set.

1.4 Some Variations of Ranked Set Sampling

There are a number of variations of ranked set sampling method proposed and developed by researchers. Some of them are described below.

1.4.1 Extreme Ranked set Sampling (ERSS)

The extreme ranked set sampling (ERSS) procedure was introduced by Samawi et al (1996a). In this procedure, we select *m* random samples of size *m* units from the population and rank the units within each sample with respect to a variable of interest by visual inspection. If the sample size *m* is even (ERSS-Even), select from *m*/2 samples the smallest unit and from the remaining *m*/2 samples the largest unit for actual measurement. If the sample size is odd, select from (*m*-1) samples the smallest unit, from the other (*m*-1) samples the largest unit and for the remaining sample, we have two options. Either select the median of the sample for actual measurement (ERSS-odd-Median) or take the average of the measures of the smallest and the largest units (ERSS-odd-both). The cycle may be repeated *r* times to get *rm* units. These *rm* units form the ERSS data.

We can see that the ERSS in practical applications can be performed with fewer errors in ranking the units since all we have to do is find the largest or the smallest of the sample and measure it. The ERSS method is very easy to apply in the field and will save time in performing the ranking of the units with respect to the variable of interest. In addition, this method will reduce the errors in ranking and hence increase the efficiency of the ERSS when compared to RSS.

1.4.2 Median Ranked Set Sampling (MRSS)

Muttlak (1997) proposed median ranked set sampling (MRSS) method which consists of selecting *m* random samples each of size *m* units from the population and rank the units within each sample with respect to the variable of interest. If the sample size *m* is odd, then from each sample select for measurement the ((*m*+1)/2)th smallest rank (the median of the sample). If the sample size *m* is even, then select for measurement the (*m*/2)th smallest rank from the first *m*/2 samples, and the ((*m*+2)/2)th smallest rank from the second *m*/2 samples. The cycle can be repeated *r* times if needed to obtain a sample of size *rm*.

1.4.3 Paired Ranked Set Sampling (PRSS)

Hossain and Muttlak (1999) gave the Paired ranked set sampling (PRSS). In this method two sets of *m* random elements are required to obtain a sample of size two. At first *m* elements are selected randomly and ordered, the *m*th smallest element of the set is considered for measurement, where $1 \leq k \leq m$ is pre-determined, Similarly, second set of size *m* elements is again selected randomly and ordered, and the $(m - k+1)^{th}$ smallest of the set is measured. The procedure can be repeated *r* times to obtain a sample of size 2*r*. Note that in the usual RSS method the sample size is required to be a multiple of *m* and in the PRSS method it is required to be a multiple of 2 and does not depend on the choice of the set size *m*.

1.4.4 Double Ranked Set Sampling (DRSS)

The double ranked set sampling (DRSS) procedure was given by Al-Saleh and Al -Kadiri (2000). It can be described as the following: Identify m^3 units from the target population and divide these units randomly into *m* sets each of size m^2. The procedure of ranked set sampling is applied on each m^2 units to obtain *m* ranked set sampling each of size *m*, then again apply the ranked set sampling procedure on the *m* ranked set sampling sets obtained in the first stage to obtain a DRSS of size *m* .

1.4.5 Moving Extremes Ranked Set Sampling (MERSS)

Al-Odat and Al-Saleh (2001) introduced the concept of varied set size RSS, which is coined here as Moving Extremes Ranked Set Sampling (MERSS). They investigated this modification non-parametrically and found that the procedure can be more efficient and applicable than the simple random sampling technique (SRS).

The procedure of MERSS is described as follows:

Step 1: Select *m* random samples of size *m* from the population.

Step 2: Identify the maximum of each set by eye or by some other relatively inexpensive method, without actual measurement of the characteristic of interest. Measure accurately the selected unit.

Step 3: Again select *m* random samples of size *m* from the population and identify the minimum in these. Measure it accurately.

These 2*m* units constitute a MERSS of size n=2*m*

Repeat the above steps *r* time until the desired sample size, k = 2*rm* is obtained.

This sample is called Moving Extremes Ranked Set Sample (MERSS).

1.4.6 Selected Ranked Set Sampling (SRSS)

Hossain and Muttlak (2001) considered the situation where, instead of selecting *m* random sets of size *m* elements each as in the RSS, only *k* sets of *m*>*k* elements are selected, and instead of measuring the *i*th smallest order statistic of the *i*th set, m_i^{th} smallest order statistic of the m_i^{th} set is considered for measurement the values of $m_1, m_2, \dots\dots, m_k (1 \leq m_1 < m_2 < \cdots < m_k \leq m)$ are required to be determined beforehand.

1.4.7 Percentile Ranked Set Sampling (PRSS)

Percentile ranked set sampling (PRSS) was given by Muttlak (2003a). In this procedure, select *m* random samples of size *m* units from the population and rank the units within each sample with respect to a variable of interest. If the sample size is even, select for measurement from the first *m*/2 samples the ($p(m+1)$)th smallest ranked unit and from the second *m*/2 samples the ($q(m+1)$)th smallest ranked unit, where $0 \leq p \leq 1$ and $q = 1 - p$. If the sample size is odd, select from the first (*m*-1)/2 samples the ($p\ (m+1)$) th smallest ranked unit and from the other (*m*-1)/2 samples the ($q\ (m+1)$) th smallest ranked unit and select from the remaining sample the median for that sample for actual measurement. The cycle may be repeated *r* times if needed to get *rm* units. These *rm* units form the PRSS data. Note that we will always take the nearest integer of $p\ (m+1)$th and $q(m+1)$th.

1.4.8 Quartile Ranked Set Sampling (QRSS)

Quartile Ranked Set sampling method was given by Muttlak (2003b). In this method, *m* units are selected from the population and we rank the units within each sample with respect to a variable of interest. If the sample size is even, select for measurement from the first *m*/2 samples the q_1 (*m*+1)th smallest ranked unit and from the second *m*/2 samples, the q_3(*m*+1)th smallest ranked unit. If the sample size is odd, select from the first $(m-1)/2$ samples the $q_1(m+1)$th smallest rank and for the last

(m-1)/2 samples the q_3(m+1)th smallest rank, and from the remaining sample the median for that sample for actual measurement. The cycle can be repeated r times if needed to get a sample of size rm units. Note that we always take the nearest integer of q_1(m+1)th and q_3(m+1)th where q_1=0.25 and q_3=0.75.

1.4.9 Double Quartile Ranked Set Samples (DQRSS)

The Double Quartile Ranked Set Sampling (DQRSS) procedure was given by Jemain & Al-Omari (2006) can be described as follows. Select m^3 units from the population and divide them into m^2 samples each of size m. If the sample size is even, select from the first $m^2/2$ sample the [q_1(m+1)]th smallest rank, from the second $m^2/2$ samples the [q_3 (m+1)] th smallest rank. If the sample size is odd, select from the first m(m–1)/2 samples the [q_1(m+1)) th smallest rank, the median from the next m samples and the [q_3(m+1)] the smallest rank from the second m(m–1)/2 samples. This step yield m sets each of size m. Apply the QRSS procedure on the m sets obtained earlier to get a DQRSS sample of size m. The whole cycle may be repeated r times to obtain a sample of size rm from DQRSS.

1.4.10 Two-stage ranked set sampling (TSRSS)

The TSRSS procedure was given by Jemain et al. (2007) and it can be summarized as the followings. First, randomly select $m^3 = 27k^3$ (k = 1, 2,...) units from the target population and divide these units randomly into $m^2 = 9k^2$ sets each of size m. Then, allocate these $9k^2$ sets into three groups, each of $3k^2$ sets. From each set in the first group select the smallest rank unit, from each set in the second group select the median rank unit, and from each set in the third group select the largest rank unit. This step yields k sets in each group. Finally, without doing any actual quantification, from the k sets in the first group select the smallest rank unit, from the k sets in the second group select the median rank unit, and from the k sets in the third group select the largest rank unit. This step yields one set of size $m = 3k$. If the procedure is repeated r times, a sample of size rm is obtained.

1.4.11 Multistage ranked set sampling

This method was given by Al-Saleh & Al-Omari (2002). The MSRSS procedure is described as follows:

Step 1: Randomly selected m^{r+1} sample units from the target population, where r is the number of stages and m is the set size.

Step 2: Allocate the m^{r+1} selected units randomly into m^{r-1} sets, each of size m^2.

Step3: For each set in Step (2), ranked set sampling procedure is applied; to obtain a (judgment) ranked set of size m. This step yields m^{r-1} (judgment) ranked sets, of size m each.

Step 4: Without doing any actual quantification on these ranked sets, repeat Step (3) on the m^{r-1} ranked set to obtain m^{r-2} second stage (judgment) ranked sets, each of size m.

Step 5: The process is continued using Step (3), without doing any actual quantification, until we end up with one rth stage (judgment) ranked set of size m.

Step 6: Finally, the m identified elements in Step (5) are now quantified for the variable of interest

1.4.12 L Ranked Set Sampling (LRSS)

Al-Nasser (2007) suggested a robust RSS procedure, based on the idea of L statistic, which will is referred as L ranked set sampling (LRSS). The main idea of this procedure is to discard the data in the tails of a data set (trimming), or replace data in the tails of a data set with the next most extreme data value (winsorizing). In order to plan LRSS design, m random samples should be selected each of size m, where m is typically small to reduce ranking error. For the sake of convenience it is assumed that the judgment ranking is as good as actual ranking. LRSS has the following steps:

Step 1: Select m random samples each of size m units

Step 2: Rank the units within each sample with respect to a variable of interest by a visual inspection or any other cost-effective method.

Step 3: Select the LRSS coefficient, $k = [mp]$, such that $0 \leq p < 0.5$, where [x] is the largest integer value less than or equal to x.

Step 4: For each of the first k ranked samples, select the unit with rank $k + 1$ for actual measurement.

Step 5: For each of the last k ranked samples, select the unit with rank $m - k$ for actual measurement.

Step 6: For $j = k+1, k + 2, \ldots\ldots, m - k - 1$, the unit with rank j in the jth ranked sample is selected for actual measurement.

Step 7: The cycle may be repeated r times to obtain the desired sample size $n=rm$.

1.4.13 Balanced Group Ranked Set Sampling (BGRSS)

Jemain et al. (2009) proposed the Balanced groups ranked set samples method (BGRSS) for estimating the population mean with samples of size $m=3k$ where ($k=1,2,\ldots$). It was found that the BGRSS produced unbiased estimators with smaller variance than commonly used simple random sampling for symmetric distribution. The balanced groups of ranked set sampling can be described as follows:

Step 1: Randomly select $m=3k$, ($k=1, 2,..$) sets each of size m from the target population, and rank the units within each set with respect to the variable of interest.

Step 2: Allocate the $3k$ selected sets randomly into three groups, each of size k sets.

Step 3: For each group in step (2), select for measurement the lowest ranked unit from each set in the first group, and the median unit from each set in the second group, and the largest ranked unit from each set in the third group.

By this way we have a measured sample of size $m=3k$ units in one cycle. The Steps 1-3 can be repeated r times to increase the sample size to $3rk$ out of $9\ r\ k^2$ units.

Indeed, the BGRSS method is easy to be applied since we only need to identify and measure the lowest rank units of the first k sets and the medians of the second k sets and the largest rank units from the last k sets. Here, k is any positive integer. However, for practical purposes, k should be small in order to have a small sample size, so that the ranking is easy and errors in ranking are reduced.

1.4.14 Percentile double ranked set sampling (PDRSS)

This method was proposed and explored by Al-Omari & Jaber (2008). To obtain a sample of size m based on PDRSS method, the following steps are required to be carried out:

Step 1: Randomly select m^3 units from the target population and divide them into m samples each of size m^2.

Step 2: Apply the RSS method on the m sets; this step yields m ranked set samples each of size m.

Step 3: Without doing any actual quantifications on the m sets obtained in Step 2, apply the PRSS method described above. Repeat the process r times to obtain a set of size rm from initial m^3r units.

Note that if the sample size $m \leq 3$, the percentile double ranked set sampling will be reduced to the usual ranked set sampling procedure. However, we will always take the nearest integer of the $(p(m+1))$th and $(q(m+1))$th, where $q = 1- p$ and $0 \leq p \leq 1$.

1.4.15 Stratified Percentile Ranked Set Sampling (SPRSS)

This procedure is given by Al-Omari et al. (2011). In the classical stratified sampling method, the population of N units is divided into L non overlapping subpopulations each of $N_1, N_2, \ldots, N_L$ units, respectively, such that $N_1 + N_2 + \ldots + N_L = N$. These subpopulations are called strata. Then the samples are drawn independently from each strata, producing samples sizes denoted by $n_1, n_2, \ldots, n_L$, such that the total sample size is $n = \sum_{h=1}^{L} n_h$.

If a simple random sample is taken from each stratum, the whole procedure is known as stratified simple random sampling (SSRS). If the percentile ranked set sampling is used to select the sample units from each stratum, then the whole procedure is called a stratified percentile ranked set sampling (SPRSS).

1.4.16 Stratified Quartile Ranked Set Samples (SQRSS)

Syam et al. (2012) introduced and explored the method of Stratified Quartile Ranked Set Sampling (SQRSS). In stratified sampling method, the population of N units is divided into L non overlapping subpopulations each of $N_1, N_2, \ldots, N_L$ units,

respectively, such that $N_1 + N_2 + ... + N_L = N$. These subpopulations are called strata. Then the samples are drawn independently from each strata, producing samples sizes denoted by $n_1, n_2, ..., n_L$, such that the total sample size is $n = \sum_{h=1}^{L} n_h$. If a simple random sample is taken from each stratum, the whole procedure is known as stratified simple random sampling (SSRS). If the quartile ranked set sampling method is used to select the sample units from each stratum then the whole procedure is called a stratified quartile ranked set sampling (SQRSS).

CHAPTER 2

ESTIMATION OF THE PARAMETERS OF THE GENERALIZED GEOMETRIC DISTRIBUTION USING RANKED SET SAMPLING

2.1 Introduction

In statistics as we quantify observations, we use a mathematical approach to account for and explain the observations generated by a phenomenon. Distributions are fitted to observed data to find a pattern which may lead the investigator to see whether some generating model can be set up for the process.

The discrete probability distributions form a basic and promising field of study in the domain of statistics and have many important applications in a wide variety of disciplines, such as biological and medical, social, physical sciences quality control, engineering and so-on. The field of discrete distributions has been found to have a huge potential for wider exploration. Since last twenty years or so, a vast amount of literature has appeared in this field. A large number of discrete distributions have been evolved, many authors obtained different generalizations of some classical distributions, either by compounding two or more discrete/continuous distributions or by dropping some assumptions in classical distributions. At present there exists large number of generalizations of basic distributions in statistical literature. A good account of these distributions is available in Patil and Joshi (1968), Johnson Kotz (1969) and Johnson Kotz and Kemp (1992). The usefulness of a distribution to a greater extent rests on its structural properties and the basic assumptions inherent in the very derivation of the distribution. These properties considerably help very much in recognizing the empirical situations where the distributions may be applied successfully.

Some of the distributions and their properties discussed below.

2.1.1 Binomial Distribution (BD)

The Binomial distribution is one of the oldest distributions derived by James Bernoulli in 1713.

A random variable X is said to have Binomial Distribution with parameters n and p if its probability mass function (p.m.f) is given by

$$P(X = x) = b(x; n, p) = \begin{cases} \binom{n}{x} p^x q^{n-x} & x = 0,1,\cdots n \\ 0 & otherwise \end{cases}$$

n ranges over the set of positive integers and p satisfies $0 \le p \le 1$, $q = 1 - p$. The probabilities are terms of binomial expansion of $(q + p)^n$ hence the name binomial distribution. When $n = 1$, binomial distribution reduces to Bernoulli distribution, whose p.m.f is given as:

$$P(X = x) = p^x q^{1-x} \qquad x = 0,1$$
$$= 0, \qquad otherwise$$

2.1.2 Geometric Series Distribution (GSD)

In the binomial distribution we consider a fixed number of Bernoullian trials and the probability of a number of successes with probability of success at a trial being p and that of failure being q, $p+q=1$. The concept is extended by considering an infinite sequence of such trials and getting interested in the probability "when does the first success occur?"

Let X be a random variable representing the number of trials after which the first success occurs. Now for any positive integer $x \ge 0$,

$P(X = x) = P$ [first x trials are failures & $(x+1)$th trial is a success].

$$P(X = x) = g(x; p) = \begin{cases} pq^x & x = 0,1,2,\ldots. \\ 0 & otherwise \end{cases}$$

and $0 < p < 1$

The random variable X with p. m. f given above is said to follow a geometric distribution with parameter p. The reason is, for x=0, 1, 2,…, $p(x)$ gives different terms of a geometric series. This may also be called the distribution of discrete waiting time (in terms of no. of failures) till the first success.

On the other hand if X is taken as the no. of trials required for first success, then

$$P[X = x] = q^{x-1} p, \quad x=0,1,2,\ldots.$$

All these forms are used frequently in literature.

2.1.3 The 'Memoryless' property of Geometric Distribution

Given that there is no success up to the first r number of trials, the conditional probability of having a success at the (r+1)th trial is independent of r (the no. of trials resulting in failures). This property is called 'Memoryless' property. It implies that the system forgets its previous history regarding the number of past failures.

The conditional probability of a success at the $(r+1)^{th}$ trial given that there are no successes up to the rth trial, (i.e., all r trials are failures), is given by:

$$P[(X = r+1)/X > r] = \frac{P[X=r+1]}{P[X>r]} = \frac{pq^r}{q^r} = p$$

which is not only independent of r but is the probability of success at any trial.

Significance of this property

Suppose a machine works (or fails) according to the geometric distribution. Suppose each trial corresponds to a period of one month. The lack of memory property implies that the chance of machine failure during the 1st or 2nd or 3rd… month of operation is same as that of failing in hundredth, thousandth or any month of operation. In other words, this property implies that the machine (with geometric distribution) forgets its age while failing and the chance of failing at any age remains the same. This is a unique property of geometric distribution among discrete distributions.

2.2 Generalized Geometric Series Distribution (GGSD)

The Generalized Geometric Series Distribution (GGSD) was obtained by Mishra (1982) using the lattice path analysis. This distribution has two parameters θ and β and its pmf is given by:

$$P(X = x) = \frac{1}{1+\beta x}\binom{1+\beta x}{x}\theta^x(1-\theta)^{1+\beta x-x} \qquad x = 0,1,2,3,\ldots.$$

$$= 0 \quad for\ x \geq t,\ \ if\ 1+\beta t < 0 \qquad (2.2.1)$$

$$0 < \theta < 1, \qquad \beta = 0 \text{ or } 1 \leq \beta < \theta^{-1}$$

This distribution reduces to the geometric series distribution with parameter θ, if $\beta = 1$ and to Bernoulli distribution at $\beta = 0$.

Also, the GGSD is a member of Lagrangian distribution, i.e.

$$f(t) = f(0) + \sum_{s=1}^{\infty} \frac{u^s}{s!} \left[\frac{d^{s-1}}{dt^{s-1}}(g(t))^s f'(t)t = 0\right]$$

and can be obtained by taking $g(t) = (1 - \theta + \theta t)^{\beta}$ and $f(t) = (1 - \theta + \theta t)$.

Sometimes we find discrete distribution for the values of random variable x=1,2,….Such cases are called zero truncated distributions. The zero truncated GGSD is given by

$$P(X = x) = \frac{1}{1+\beta x}\begin{bmatrix}1+\beta x \\ x\end{bmatrix}\theta^{x-1}(1-\theta)^{1+\beta x-x} \qquad x = 1,2,3, \ldots.$$

$$= 0 \quad for\ x \geq t,\ \ if\ 1 + \beta t < 0 \qquad (2.2.2)$$

$$0 < \theta < 1, \qquad \beta = 0\ or\ 1 \leq \beta < \theta^{-1}$$

2.2.1 Size Biased Generalized Geometric Series Distribution (SBGGSD)

The p.m.f of size-biased generalized geometric series distribution (SBGGSD) is given by

$$P(X = x) = (1 - \theta\beta)\begin{bmatrix}\beta x \\ x-1\end{bmatrix}\theta^{x-1}(1-\theta)^{1+\beta x-x} \qquad x = 1,2,3, \ldots.$$

$$= 0 \quad for\ x \geq t,\ \ if\ 1 + \beta t < 0$$

$$0 < \theta < 1, \qquad |\theta\beta| < 1$$

2.2.2 Moments of Generalized Geometric Series Distribution

The first four moments about origin of GGSD are as follows:

$$\mu_1' = \frac{\theta}{1-\theta\beta} \qquad (2.2.2.1)$$

$$\mu_2' = \frac{\theta(1-\theta)}{(1-\theta\beta)^3} + \frac{\theta^2}{(1-\theta\beta)^2}$$

$$\mu_3' = \frac{\theta^3}{(1-\theta\beta)^3} + \frac{3\theta^2(1-\theta)}{(1-\theta\beta)^4} + \frac{\theta(1-\theta)}{(1-\theta\beta)^5}\left[1-2\theta+\theta\beta(2-\theta)\right]$$

$$\mu_4' = \frac{\theta^4}{(1-\theta\beta)^4} + \frac{6\theta^3(1-\theta)}{(1-\theta\beta)^5} + \frac{\theta^2(1-\theta)\left[7-11\theta+4\theta\beta(2-\theta)\right]}{(1-\theta\beta)^6}$$

$$+\frac{\theta(1-\theta)\left[1-60+6\theta^2+2\theta\beta(4-9\theta+4\theta^2)+\theta^2\beta^2(6-6\theta+\theta^2)\right]}{(1-\theta\beta)^7}$$

The central moments of the GGSD are:

$$\mu_2 = \frac{\theta(1-\theta)}{(1-\theta\beta)^3} \qquad (2.2.2.2)$$

$$\mu_3 = \frac{\theta(1-\theta)}{(1-\theta\beta)^5}\left[1-2\theta+\theta\beta(2-\theta)\right] \qquad (2.2.2.3)$$

The moments about origin of the zero truncated GGSD (2.2.2) may be obtained by just dividing the corresponding moments of GGSD (2.2.1) by θ, we get:

$$\mu_1' = \frac{1}{(1-\theta\beta)}$$

$$\mu_2' = \frac{\theta}{(1-\theta\beta)^2} + \frac{(1-\theta)}{(1-\theta\beta)^3} \qquad (2.2.2.4)$$

2.3 Some Other Generalizations of Geometric Distribution

2.3.1 Generalized Geometric Series Distribution I (GGSD-I)

To find GGSD-I Singh (1989) used the second form of the Lagrange's expansion, i.e.

$$\frac{f(z)}{1-\frac{zg'(z)}{g(z)}} = \sum_{x=0}^{\infty}\frac{1}{x!}\cdot\frac{\partial^x}{\partial_z^x}\left[f(z)\{g(z)\}^x\right]_{z=0}\left[\frac{z}{g(z)}\right]^x$$

where *f(z)* and *g(z)* are positive continuous functions

The probability mass function of GGSD-I is given by

$$P(X=x) = (1-\theta\beta)\binom{x\beta}{x}\theta^x(1-\theta)^{x\beta-x}$$

where $0 < \theta < 1$, $|\theta\beta| < 1$ and $x = 0,1,2\ldots$

2.3.2 Size-Biased Generalized Geometric Series Distribution-I (SBGGSD-I)

The probability mass function of GGSD-I is given by

$$P(X = x) = (1-\theta\beta)\binom{x\beta}{x}\theta^x(1-\theta)^{x\beta-x}$$

where $0 < \theta < 1, |\theta\beta| < 1$ and $x = 0,1,2 \ldots$

The size-biased version of GGSD-I is given by the p.m.f:

$$P_1(X = x) = \frac{(1-\theta\beta)^3}{\theta\beta}(x\beta - x + 1)\binom{x\beta}{x-1}\theta^x(1-\theta)^{x\beta x-x-1}$$

$$0 < \theta < 1, |\theta\beta| < 1 \text{ and } x = 0,1,2 \ldots$$

The SBGGSD-I reduces to Size Biased Geometric Series Distribution (SBGSD) when $\beta = 1$.

2.3.3 Generalized Geometric Series Distribution – II (GGSD-II)

The probability function of a two parameter generalized geometric distribution with parameters *b* and *q*

$$p_r^* = \frac{r\ \mathrm{p_r}}{\sum_{r=1}^{\infty} r\ p_r} = \frac{\alpha^*}{(r-1)!}\int_0^{\infty}(1+bx)^{-q}x^{r-1}e^{-x}dx$$

$$r = 1, 2, \ldots; b > 0, q > 0$$

where $\left(\alpha^*\right)^{-1} = \sum_{r=1}^{\infty}\frac{1}{(r-1)!}\int_0^{\infty}(1+bx)^{-q}x^{r-1}e^{-x}dx$

$$= \frac{1}{b(q-1)}, \qquad q > 1$$

2.3.4 Generalized Geometric Series Distribution-III (GGSD-III)

The probability function of GGSD-III defined by Tripathi and Gupta (1987) is given by

$$p_r = C\cdot\frac{(r-1)!\theta^{r-1}}{(\lambda+1)_{r-1}} \quad , \ \lambda > -1,\ 0 < \theta \le 1, \quad \mathrm{r} = 1,2\cdots$$

where $C = p_1$ with

$$p_1 = \left[\sum_{r=1}^{\infty} \frac{(r-1)!\theta^{r-1}}{(\lambda+1)_{r-1}} \right]^{-1}$$

2.4 Estimation of Generalized Geometric Series Distribution by Method of Moments

Suppose a random sample of size n is taken from GGSD model (2.2.1). Let the observed frequencies be $n_0, n_1, \ldots n_k$ where k is the largest value of x in sample such that $n = \sum_{i=1}^{k} n_x$ .Let the first two sample moments for GGSD model be denoted as

$$m_1 = \frac{1}{n} \sum_{x=0}^{k} x n_x \qquad and \qquad m_2 = \frac{1}{n} \sum_{x=0}^{k} x^2 n_x$$

2.4.1 First Two Moment Method (TMM)

By using elimination between the expression (2.2.2.1) and (2.2.2.2) and replacing μ_1' and μ_2' by respective sample moments, we get from (2.2.2.1)

$$\hat{\beta} = \frac{(m_1 - \theta)}{m_1 \hat{\theta}} \tag{2.4.1.1}$$

Also

$$\frac{(m_2 - m_1^2)}{m_1^3} = \frac{1-\theta}{\theta^2}$$

which gives quadratic equation in θ as

$$m_1^{-3}\left(m_2 - m_1^2\right)\theta^2 + \theta - 1 = 0$$

The admissible roots of θ is given by

$$\hat{\theta} = \frac{-1 + \left(1 + 4k\right)^{1/2}}{2k} \tag{2.4.1.2}$$

where $k = \left(m_2 - m_1^2\right) m_1^{-3}$

2.4.2 Zero frequency and first moment method (ZFFM)

Let P_0 be the probability of the zero class in GGSD (2.2.1)

$$P_0 = f_0 n^{-1}$$

We equate P_0 to the corresponding sample proportion of zeros to get

$$f_0 n^{-1} = (1-\theta)$$

which gives $\hat{\theta} = \left(1 - f_0 n^{-1}\right)$

In addition we have

$$\mu_1' = \frac{\theta}{1-\theta\beta}$$

replacing μ_1' by corresponding sample moments m_1 and after simplification we get the estimate of β as

$$\hat{\beta} = \frac{m_1 - \hat{\theta}}{m_1 \hat{\theta}} \qquad (2.4.2.1)$$

2.4.3 First two moments and Ratio of first two frequencies (MORA)

Let P_1 be the probability of the "one" class and P_0 be the probability of "zero" class in GGSD (2.2.1). The ration of "one" class to the "zero" class is given by

$$\frac{P_1}{P_0} - \theta(1-\theta)^{\beta-1} = f_r$$

Squaring this term, we get

$$\theta^2 (1-\theta)^{2\beta-2} = f_r^2$$

which gives

$$\theta^2 = \frac{f_r^2}{(1-\theta)^{2\beta-2}}$$

Also we have

$$\mu_1' = \frac{\theta}{(1-\theta\beta)}$$

Substituting the value of $(1-\theta\beta)^3$ in (2.2.2.2) we have

$$\mu_2' = \frac{(1-\theta)\mu_1^3}{\theta^2}$$

which gives $\theta^2 = (1-\theta)\mu_1'^3 \cdot \mu_2^{-1}$

on combining (2.2.2.3) and (2.2.2.4)we have

$$\frac{f_r^2}{(1-\theta)^{2\beta-2}} = (1-\theta)\mu_1^3\mu_2^{-1}$$

$$= \mu_2 f_r^2 \mu_1^{-3} = (1-\theta)^{2\beta-1}$$

Applying log, we get

$$2\beta-1\ \log(1-\theta) - \log(\mu_2 f_r^2 \mu_1^{-3}) \qquad (2.4.3.1)$$

Also ration of first two moments gives

$$\frac{\mu_1}{\mu_2} = \frac{\theta(1-\theta\beta)^2}{(1-\theta)}$$

which on simplification gives

$$\hat{\beta} = \frac{1}{\theta} - \frac{1}{\theta}\left[\frac{\mu_1(1-\theta)}{\mu_2}\right]^{1/2} \qquad (2.4.3.2)$$

using relation (2.4.3.2)in (2.4.3.1)on simplification, we have

$$f(\theta) = \left\{\frac{2}{\theta} - \frac{2}{\theta}\left[\frac{\mu_1(1-\theta)}{\mu_2}\right]^{1/2} - 1\right\}\log(1-\theta) - \log\left(\mu_2 f_r^2 \mu_1^{-3}\right) = 0$$

replacing the first two sample moments to their corresponding population moments, we have

$$f(\theta) = \left\{\frac{2}{\theta} - \frac{2}{\theta}\left[\frac{m_1(1-\theta)}{m_2 - m_1^2}\right]^{1/2}\right\}\log(1-\theta) - \log\left(\left(m_2 - m_1^2\right) f_r^2 m_1^{-3}\right) = 0$$

We solve $f(\theta)$ iteratively to obtain, $\hat{\theta}$ the MORA estimator of parameter θ. The estimate of β can be obtained by using (2.4.2.1).

2.5 Maximum Likelihood Estimation

The likelihood function of GGSD based on the random sample $x_1, \ldots, x_n$ is given by

$$L = \frac{\theta^{n\bar{x}}(1-\theta)^{n+n(\beta-1)\bar{x}}\prod_{x=1}^{k}\prod_{j=1}^{x-1}(1+\beta x-j)^{nx}}{\prod_{x=0}^{k}(x_i!)^{fx}}$$

The two likelihood equation are obtained as

$$\frac{\partial}{\partial\theta}\log L = \frac{n\bar{x}}{\theta} - \frac{n[1+(\beta-1)\bar{x}]}{1-\theta} = 0 \qquad (2.5.1)$$

$$\frac{\partial}{\partial\beta}\log L = n\bar{x}\log(1-\theta) + \sum_{x=2}^{k}\sum_{j=1}^{x-1}\frac{xn_x}{(1+\beta x-j)} = 0 \qquad (2.5.2)$$

from (2.5.1)we have

$$\hat{\theta} = \frac{\bar{x}}{1+\beta\bar{x}} \qquad (2.5.3)$$

when substituted in (2.5.2)gives

$$= n\bar{x}\log\left[\frac{1+(\beta\text{-}1)\bar{x}}{1+\beta\bar{x}}\right] + \sum_{x=2}^{k}\sum_{j=1}^{x-1}\frac{xn_x}{(1+\beta x-j)} = 0$$

The equation can be solved for β applying some iteration technique. The estimate of β when substituted in (2.5.3) gives an estimate of θ.

2.6 Bayesian Estimation of Parameters

Mishra (1982) defined GGSD as

$$P(X=x) = \frac{1}{1+\beta x}\binom{1+\beta x}{x}\theta^x(1-\theta)^{1+\beta x-x} \qquad x = 0,1,2,\cdots \qquad 0<\theta<1$$

The likelihood function is obtained as

$$L(x/\theta,\beta)=\prod_{i=1}^{n}\left\{\frac{1}{1+\beta xi}\binom{1+\beta xi}{\text{xi}}\right\}\theta^{\Sigma x_i}(1-\theta)^{n+\beta\Sigma x_i-\Sigma x_i}$$

$$\text{K}\theta^{\text{y}}(1-\theta)^{\text{n}+\beta\text{y}-\text{y}}$$

where $K=\prod_{i=1}^{n}\frac{1}{1+\beta x_i}\binom{1+\beta x_i}{\text{x}_\text{i}}$

and $y=\sum_{i=1}^{n}x_i$

Since $0<\theta<1$, it is assumed that prior information of θ is given by a beta distribution with density function

$$g(\theta;a,b)=\frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}$$

$$a,b>0, \qquad 0<\theta<1$$

The posterior distribution of θ is defined as $\pi(\theta/y)=\frac{L(\underline{x}/\theta,\beta)g(\theta;a,b)}{\int_{\Omega}L(\underline{x}/\theta,\beta)g(\theta;a,b)d\theta}$

$$\pi(\theta/y)=\frac{\theta^{y}(1-\theta)^{n+(\beta-1)y}\cdot\theta^{a-1}(1-\theta)^{b-1}}{\int_{0}^{1}\theta^{y}(1-\theta)^{n+(\beta-1)y}\theta^{a-1}(1-\theta)^{b-1}d\theta}$$

$$\pi(\theta/y)=\frac{\theta^{a+y-1}(1-\theta)^{n+(\beta-1)y+b-1}}{\int_{0}^{1}\theta^{y+a-1}(1-\theta)^{n+(\beta-1)y+b-1}d\theta}$$

$$\pi(\theta/y)=\frac{\theta^{a+y-1}(1-\theta)^{n+(\beta-1)y+b-1}}{B(y+a,n+(\beta-1)y+b)}$$

The Bayes estimator for parametric function $\phi(\theta)$

$$\overset{*}{\phi}(\theta)=\int_{0}^{1}\phi(\theta)\pi(\theta/y)d\theta$$

$$\overset{*}{\phi}(\theta)=\frac{\int_{0}^{1}\phi(\theta)\theta^{a+y-1}(1-\theta)^{n+(\beta-1)y+b-1}d\theta}{B(y+a,n+(\beta-1)y+b)}$$

If we take $\phi(\theta) = \theta$ then Bayes estimator θ is

$$\overset{*}{\theta} = \frac{\int_0^1 \theta^{a+y}(1-\theta)^{n+(\beta-1)y+b-1} d\theta}{B\left(y+a, n+(\beta-1)y+b\right)}$$

$$\overset{*}{\theta} = \frac{B\left(y+a+1, n+(\beta-1)y+b\right)}{B\left(y+a, n+(\beta-1)y+b\right)}$$

$$\overset{*}{\theta} = \frac{a+y}{n+a+b+\beta y}$$

Which is also identical to MLE of GGSD if $a = b = 0$

2.7 A Quick Method for Estimating Generalized Geometric Series Distribution

A quick method for estimating the parameters of generalized geometric series distribution (GGSD) was given by Hassan, Mishra and Jan (2002) for the case when non-zero frequencies are found only up to a finite number of values of the variable. In such cases only one parameter θ is estimated which is based on the mean of the observed distribution, the parameter β being obtain just by counting the number of non-zero frequency classes. The estimator is simple and quick in practice.

Let $t-1$ be the highest observed value having non-zero frequency. From the condition of GGSD (2.2.1)

$$P(X = x) = 0 \quad for \quad x \le t \quad \text{if } 1+\beta t - t \le 0$$

we may have $1 + \beta t - t = 0$, which gives minimum value of β, say β_0 as

$$\beta_0 = \frac{t-1}{t}$$

Substituting this value of β in the expression for the mean of GGSD (2.2.2.1) and replacing $\boldsymbol{\mu_1'}$ by sample mean $\bar{\boldsymbol{x}}$ we get the estimate of θ, $\hat{\boldsymbol{\theta}}$ as

$$\hat{\theta} = \frac{\bar{x}}{1+\bar{x}\beta_0}$$

which is same if β in replaced by β_0 in maximum likelihood estimator of θ (2.5.3)

The value of β_0 is obtained directly from the non-zero frequency classes and may be treated as predetermined as n in case of binomial distribution.

2.8 Estimation of Parameters μ and σ Based on Ranked Set Sampling

The estimation of location and scale parameters, based on an ordered sample, has been discussed by Lloyd (1952). Downton (1954) obtained least squares estimates explicitly for a class of two-parameter distributions having the form $f\{(x-\mu)/\sigma\}/\sigma$. More specifically, consider the generalized geometric random variable X, with pdf as follows:

$$f(x) = p\sigma^{-1}b^{-p}\left(\frac{x-\mu}{\sigma}+a\right)^{p-1}, \quad \mu - a\sigma \le x \le \mu + (b-a)\sigma \qquad (2.8.1)$$

$$= 0, \text{ otherwise}$$

Where $p \ge 1, a = \sqrt{p(p+2)}, b = (p+1)\sqrt{(p+2)/p}$.

The rectangular distribution ($p = 1$) and the right triangular distribution $(p = 2)$ are special cases of the above distribution.

It can be shown that $E(X) = \mu$ and $Var(X) = \sigma^2$.

Downton derived the least squares estimates of μ and σ based on the ordered observations $x_{(1)} < x_{(2)} < \ldots < x_{(m)}$. He gave all the intermediate computations but did not write the explicit formulae for the estimates. Instead, the explicit formulae for the estimates and their variance-covariance matrices are given for special cases of p = 2 and p=1.

Bhoj and Ahsanullah (1996) derived the estimates of μ and σ for the random variable X whose pdf is given in (2.8.1) for any p based on ranked set observations. The variances and covariance of the estimates are also given. They are compared with those of ordered least squares estimates given by Downton for special values of p.

Ranked set sampling procedure is used for the joint estimation of population mean μ and standard deviation σ of the two-parameter distribution given in (2.8.1) by using least squares methods. For this, we first measure accurately a ranked set sample of m observations, i.e., $x_{(11)}, x_{(22)}, \ldots, x_{(mm)}$. These m observations are then

used to estimate μ and σ of the generalized geometric distribution whose pdf is given in (2.8.1). Since each set is an independent sample and only one element in each set is quantified, all quantified elements are independent. Further, $x_{(ii)}$ is an i^{th} ordered observation in the i^{th} set.

We define $y_{(ii)} = (x_{(ii)} - \mu)/\sigma$.

Let $E(y_{(ii)}) = \alpha_{(i)}$ and $var(y_{(ii)}) = d_{ii}$. Considering the pdf of the i^{th} order statistic from the generalized geometric distribution, it can be shown that

$$\alpha_i = (p+1)\sqrt{\frac{p+2}{p}}\left\{\frac{m^{m-i+1)}p^{m-i+1}}{(mp+1)^{(m-i+1)}} - \frac{p}{p+1}\right\} \tag{2.8.2}$$

And

$$d_{ii} = \frac{(p+1)^2(p+2)}{p}\left[\frac{m^{(m-i+1)}p^{m-i+1}}{(mp+2)^{(m-i+1)}} - \left\{\frac{m^{(m-i+1)}p^{m-i+1}}{(mp+1)^{(m-i+1)}}\right\}^2\right] \tag{2.8.3}$$

where

$$m^{(s)} = m(m-1)\dots(m-s+1)$$

and

$$(mp+k)^{(s)} = (mp+k)\{(m-1)p+k\}\dots\{(m-s+1)p+k\}.$$

In terms of original *x*'s we have

$$E(x_{(ii)}) = \mu + \sigma\alpha_i, \qquad var(x_{(ii)}) = d_{ii}\sigma^2$$

Let

$$X' = (x_{(11)}, x_{(22)}, \dots, x_{(mm)})$$

$$1' = (1,1,\dots,1)$$

$$\alpha' = (\alpha_1, \alpha_2, \dots, \alpha_m)$$

and $V(X) = \sum \sigma^2$

where $\sum$ is a $m \times m$ diagonal matrix with d_{ii} as the (i,i)th element, $i = 1, 2, \ldots, m$.

Then

$$E(X) = \mu_1 + \sigma\alpha$$

$$= A\theta, say$$

where

$$A' = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \alpha_1 & \alpha_2 & \ldots & \alpha_m \end{pmatrix} and\ \theta' = (\mu, \sigma)$$

The minimum variance linear unbiased estimators (MVLUE) of θ can be obtained by using least squares theorem of Gauss and Markov. Let $\tilde{\theta}$ denote the MVLUE of θ, then

$$\tilde{\theta} = (A'\Sigma^{-1}A)^{-1}A'\Sigma^{-1}X$$

and the variance-covariance matrix of $\tilde{\theta}$ is given by $(A'\Sigma^{-1}A)^{-1}\sigma^2$ σ^2.

On simplification, we have

$$\tilde{\mu} = \sum w_i x_{(ii)} \qquad (2.8.4)$$

$$\tilde{\sigma} = \sum v_i x_{(ii)} \qquad (2.8.5)$$

Where

$$w_i = \frac{1}{D}\left(\frac{T_1}{d_{ii}} + \frac{\alpha_i T_3}{d_{ii}}\right) \qquad (2.8.6)$$

$$v_i = \frac{1}{D}\left(\frac{\alpha_i T_2}{d_{ii}} + \frac{T_3}{d_{ii}}\right) \qquad (2.8.7)$$

$$T_1 = \sum_{i=1}^{n} \frac{\alpha_i^2}{d_{ii}}, \qquad T_2 = \sum_{i=1}^{n} \frac{1}{d_{ii}}, \qquad T_3 = -\sum_{i=1}^{n} \frac{\alpha_i}{d_{ii}}$$

and

$$D = T_1 T_2 - T_3^2$$

The variances and covariances of $\tilde{\mu}$ and $\tilde{\sigma}$ are given by

$$var(\tilde{\mu}) = \frac{\sigma^2 T_1}{D} \tag{2.8.8}$$

$$var(\tilde{\sigma}) = \frac{\sigma^2 T_2}{D} \tag{2.8.9}$$

And

$$cov(\tilde{\mu}, \tilde{\sigma}) = \frac{\sigma^2 T_3}{D} \tag{2.8.10}$$

The variances and covariance of these estimators are compared with those of ordered least squares estimators given by Downton for two values of p, p= 1 and p = 2.

2.8.1. Right Triangular Distribution

We get the following pdf of the right triangular distribution by substituting p = 2 in (2.8.1).

$$f(x) = \begin{cases} \frac{1}{9\sigma}\left[\frac{x-\mu}{\sigma} + 2\sqrt{2}\right], & \mu - 2\sqrt{2}\sigma \le x \le \mu + \sqrt{2}\sigma \\ 0, & otherwise \end{cases}$$

Downton gave the expressions for $\hat{\mu}$, $\hat{\sigma}$, var($\hat{\mu}$), var($\hat{\sigma}$), and cov($\hat{\mu}$,$\hat{\sigma}$), where $\hat{\mu}$ and $\hat{\sigma}$ are the MVLUE of μ and σ based on *m* ordered statistics. The MVLUE estimators of μ, and σ based on ranked set sampling are obtained by substituting $p = 2$ in (2.8.2) and (2.8.3) to compute α_i and d_{ii}, and then using these to calculate w_i and v_i in (2.8.6) and (2.8.7).

To facilitate computations of the estimators $\tilde{\mu}$ and $\bar{\sigma}$, the coefficients w_i and v_i are given in Tables 2.1 and 2.2 for $2 \le m \le 15$. Table 2.3 gives var($\tilde{\mu}$)/σ^2, var($\tilde{\sigma}$)/σ^2, and cov($\tilde{\mu}$,$\tilde{\sigma}$)/σ^2 for m = 2,3,.., 15, for comparing the precision of our estimators with those of Downton. It also gives the generalized variance of $\hat{\mu}$ and $\hat{\sigma}$ and $\tilde{\mu}$ and $\tilde{\sigma}$, where

$$Gvar\ (\hat{\mu},\hat{\sigma}) = var\ (\hat{\mu})\ var\ (\hat{\sigma}) — (cov\ (\hat{\mu},\hat{\sigma}))^2$$

and

$$Gvar\ (\tilde{\mu},\tilde{\sigma}) = var\ (\tilde{\mu})\ var\ (\tilde{\sigma}) — (cov\ (\tilde{\mu},\tilde{\sigma}))^2.$$

Following are the three relative precisions for comparison purposes:

$$RP_1 = \frac{var(\hat{\mu})}{var(\tilde{\mu})}, \qquad RP_2 = \frac{var(\hat{\sigma})}{var(\tilde{\sigma})}, \qquad \text{and} \qquad RP_3 = \frac{Gvar(\hat{\mu}, \hat{\sigma})}{Gvar(\tilde{\mu}, \tilde{\sigma})}$$

These are given in Table 2.3 for m = 2, 3… 15. It can be seen that the ranked set sampling estimator $\tilde{\mu}$ is uniformly better than the ordered least square estimator$\hat{\mu}$and the gain in precision is quite substantial. The gain in precision in terms of the generalized variance is even more dramatic for $m \geq 2$. The ranked set estimator of σ, $\tilde{\sigma}$ is more efficient than $\hat{\sigma}$ for $m \geq 5$. In this case, the gain in precision is not as great as the one attained in estimating μ.

2.8.2. Rectangular Distribution

The following pdf of the rectangular distribution centered at μ with variance σ^2 is obtained by substituting p = 1 in (2.8.1).

$$f(x) = \begin{cases} \dfrac{1}{2\sqrt{3}\sigma}, & \mu - 2\sqrt{3}\sigma \leq x \leq \mu + \sqrt{3}\sigma \\ 0, & otherwise \end{cases}$$

The MVLUEs for μ and σ given by Downton are based only on the largest and smallest observations with variances and covariance given by

$$var(\hat{\mu}) = \frac{6\sigma^2}{(m+1)(m+2)}, \qquad var(\hat{\sigma}) = \frac{2\sigma^2}{(m-1)(m+2)}$$

And

$$cov(\hat{\mu}, \hat{\sigma}) = 0$$

To drive the estimators for μ and σ based on ranked set sampling, the expressions for α_i and d_{ii} are obtained from (2.8.2) and (2.8.3), which are given by

$$\alpha_i = \sqrt{3}\left\{\frac{2i}{m+1} - 1\right\}, \qquad \text{and} \qquad d_{ii} = \frac{12i(m-i+1)}{(m-1)^2(m-2)},$$

Substituting these values in T_1, T_2 and T_3 and simplifying, we obtain from (2.8.6), (2.8.7), (2.8.8), (2.8.9) and (2.8.10)

$$\tilde{\mu} = \sum_{i=1}^{m} w_i^* x_{(ii)}, \qquad \tilde{\sigma} = \sum_{i=1}^{m} v_i^* x_{(ii)}$$

Where

$$w_i^* = \frac{m+1}{2i(m-i+1)S_m}$$

$$v_i^* = \frac{(m+1)(2i-m-1)}{2\sqrt{3}i(m-i+1)\{(m+1)S_m - 2m\}}$$

$$\text{var}(\tilde{\mu}) = \frac{6\sigma^2}{2\sqrt{3}i(m+1)\{(m+1)S_m - 2m\}},$$

$$var(\tilde{\sigma}) = \frac{2\sigma^2}{(m+2)\{(m+1)S_m - 2m\}},$$

$$cov(\tilde{\mu}, \tilde{\sigma}) = 0 \qquad and \qquad S_m = \sum_{i=1}^{m} \frac{1}{i}$$

It is clear that $var(\tilde{\mu}) < var(\hat{\mu})$ for all $m \geq 2$ since $S_m > 1$. However $var(\tilde{\sigma}) < var(\hat{\sigma})$ when $m > 5$. S_m can be read from Table 2.4.

The variances of RSS estimators are compared with those based on ordered least squares methods to assess the effectiveness of the ranked set sampling procedure. Table 2.4 gives the variances for both sets of estimators and the following two relative precisions:

$$RP_4 = \frac{var(\hat{\mu})}{var(\tilde{\mu})} = S_m, \qquad RP_5 = \frac{var(\hat{\sigma})}{var(\tilde{\sigma})} = \frac{(m+1)S_m - 2m}{m-1}$$

Since both sets of estimators have covariance zero, the generalized variances of these estimators are not given. Note that $\tilde{\mu}$ is uniformly better than $\hat{\mu}$, and $\tilde{\sigma}$ is better than $\hat{\sigma}$ for $m>5$. The fact that ranked set sampling does not result in more efficient estimators of variance in small samples for both values of p is consistent with Stokes (1980b) results.

2.8.3. Comparison with Usual Ranked Set Estimator of μ

The usual ranked set estimator of the population mean is compared with $\tilde{\mu}$. Stokes proposed the estimator for population variance by using ranked set sample

data. However, that estimator is biased and therefore cannot be compared directly with the RSS estimators discussed in this article.

The usual estimator of the population mean based on ranked set sampling is $\bar{\mu} = \sum_{i=1}^{m} x_{(ii)}/m$, with variance $var\ (\bar{\mu}) = (\sigma^2/m^2)\sum_{i=1}^{m} d_{(ii)}$.When p = 1, $Var(\bar{\mu}) = 2\sigma^2/(m\ (m\ +\ 1))$.

The maximum relative efficiency of $\bar{\mu}$ with respect to the sample mean of a simple random sample is maximum when the underlying distribution is rectangular. The relative precision of $\bar{\mu}$ and $\tilde{\mu}$ is given by:

$$RP_6 = \frac{var(\bar{\mu})}{var(\tilde{\mu})} = \frac{(m+2)S_m}{3m}$$

Values of RP_6 are displayed in Table 2.5 for various values of m and two values of p, p= 1 and p = 2. It is clear that $\tilde{\mu}$ is better than $\bar{\mu}$ for m > 2. The gain in precision of $\tilde{\mu}$ over $\bar{\mu}$ is greater for p = 1 than for p = 2.

Table 2.1: Coefficients for estimating $\tilde{\mu}$ for right triangular distribution

m/w_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	0.5000	0.5000													
3	0.3450	0.3060	0.3489												
4	0.2674	0.2297	0.2283	0.2745											
5	0.2201	0.1869	0.1784	0.1850	0.2296										
6	0.1878	0.1588	0.1492	0.1479	0.1570	0.1993									
7	0.1643	0.1388	0.1293	0.1258	0.1272	0.1373	0.1773								
8	0.1464	0.1236	0.1147	0.1105	0.1095	0.1122	0.1225	0.1606							
9	0.1321	0.1117	0.1034	0.0991	0.0971	0.0973	0.1008	0.1111	0.1474						
10	0.1206	0.1020	0.0944	0.0901	0.0878	0.0869	0.0879	0.0917	0.1019	0.1367					
11	0.1110	0.0940	0.0870	0.0829	0.0804	0.0791	0.0789	0.0803	0.0843	0.0943	0.1278				
12	0.1028	0.0873	0.0807	0.0768	0.0743	0.0728	0.0721	0.0724	0.0741	0.0782	0.0880	0.1203			
13	0.0959	0.0815	0.0754	0.0717	0.0693	0.0677	0.0667	0.0665	0.0671	0.0689	0.0730	0.0827	0.1138		
14	0.0898	0.0765	0.0708	0.0673	0.0649	0.0633	0.0622	0.0617	0.0617	0.0625	0.0645	0.0686	0.0780	0.1083	
15	0.0845	0.0721	0.0667	0.0634	0.0612	0.0596	0.0584	0.0577	0.0574	0.0577	0.0586	0.0606	0.0647	0.0740	0.1034

Table 2.2: Coefficients for estimating $\tilde{\sigma}$ for right triangular distribution

m/v_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
2	-0.8839	0.8839													
3	-0.5366	-0.1230	0.6595												
4	-0.3840	-0.1918	0.0249	0.5509											
5	-0.2989	-0.1823	-0.0799	0.0753	0.4857										
6	-0.2448	-0.1637	-0.1022	-0.0280	0.0970	0.4417									
7	-0.2074	-0.1463	-0.1042	-0.0598	0.0004	0.1075	0.4097								
8	-0.1799	-0.1314	-0.1001	-0.0701	-0.0342	0.0177	0.1129	0.3851							
9	-0.1590	-0.1190	-0.0945	-0.0723	-0.0481	-0.0172	0.0290	0.1155	0.3655						
10	-0.1424	-0.1086	-0.0886	-0.0714	-0.0537	-0.0329	-0.0054	0.0368	0.1167	0.3495					
11	-0.1290	-0.0998	-0.0830	-0.0691	-0.0554	-0.0403	-0.0219	0.0032	0.0423	0.1171	0.3360				
12	-0.1179	-0.0923	-0.0779	-0.0663	-0.0553	-0.0437	-0.0304	-0.0136	0.0096	0.0463	0.1170	0.3245			
13	-0.1086	-0.0858	-0.0733	-0.0634	-0.0543	-0.0450	-0.0348	-0.0227	-0.0072	0.0146	0.0493	0.1166	0.3145		
14	-0.1007	-0.0802	-0.0691	-0.0606	-0.0528	-0.0452	-0.0370	-0.0278	-0.0166	-0.0020	0.0186	0.0516	0.1159	0.3058	
15	-0.0938	-0.0753	-0.0654	-0.0578	-0.0511	-0.0447	-0.0380	0.0306	-0.0221	-0.0117	0.0021	0.0217	0.0534	0.1152	0.2980

Table 2.3: Variances, covariance, & relative precisions for right triangular distribution

m	$\frac{var(\hat{\mu})}{\sigma^2}$	$\frac{var(\tilde{\mu})}{\sigma^2}$	$\frac{var(\hat{\sigma})}{\sigma^2}$	$\frac{var(\tilde{\sigma})}{\sigma^2}$	$\frac{cov(\hat{\mu},\hat{\sigma})}{\sigma^2}$	$\frac{cov(\tilde{\mu},\tilde{\sigma})}{\sigma^2}$	$\frac{Gvar(\hat{\mu},\hat{\sigma})}{\sigma^4}$	$\frac{Gcov(\tilde{\mu},\tilde{\sigma})}{\sigma^4}$	RP_1	RP_2	RP_3
2	0.50000	0.34000	0.56250	1.06250	0.17678	-0.17677	0.25000	0.33000	1.47058	0.52941	0.75757
3	0.31481	0.17168	0.24769	0.34052	0.12440	-0.09347	0.06250	0.04972	1.83376	0.72737	1.25698
4	0.22273	0.10366	0.15057	0.16559	0.09482	-0.05818	0.02455	0.01378	2.14869	0.90927	1.78129
5	0.16907	0.06943	0.10513	0.09727	0.07599	-0.03983	0.01200	0.00517	2.43521	1.08080	2.32236
6	0.13452	0.04977	0.07938	0.06376	0.06304	-0.02903	0.00670	0.00233	2.70265	1.24501	2.87607
7	0.11068	0.03744	0.06303	0.04491	0.05364	-0.02212	0.00410	0.00119	2.95615	1.40368	3.43963
8	0.09340	0.02920	0.05185	0.03328	0.04652	-0.01743	0.00268	0.00067	3.19893	1.55801	4.01144
9	0.08037	0.02341	0.04377	0.02562	0.04097	-0.01410	0.00184	0.00040	343307	1.70872	4.58977
10	0.07024	0.01919	0.03770	0.02031	0.03652	-0.01165	0.00131	0.00025	3.66017	1.85651	5.17438
11	0.06218	0.01602	0.03299	0.01648	0.03288	-0.00979	0.00097	0.00017	3.88124	2.00167	5.76357
12	0.05563	0.01358	0.02924	0.01363	0.02986	-0.00834	0.00073	0.00012	4.09722	2.14470	6.35806
13	0.05021	0.01165	0.02620	0.01146	0.02732	-0.00719	0.00057	0.00008	4.30866	2.28560	6.95515
14	0.04568	0.01011	0.02368	0.00976	0.02515	-0.00627	0.00045	0.00006	4.51621	2.42498	7.55780
15	0.04182	0.00886	0.02157	0.00842	0.02327	-0.00551	0.00036	0.00004	4.72021	2.56270	8.16273

Table 2.4: ***Variances and relative precisions for rectangular distribution***

n	$\frac{var(\hat{\mu})}{\sigma^2}$	$\frac{var(\tilde{\mu})}{\sigma^2}$	$\frac{var(\hat{\sigma})}{\sigma^2}$	$\frac{var(\tilde{\sigma})}{\sigma^2}$	RP_4	RP_5
2	0.50000	0.33333	0.50000	1.00000	1.50000	0.50000
3	0.30000	0.16364	0.20000	0.30000	1.83333	0.66667
4	0.20000	0.09600	0.11111	0.13793	2.08333	0.80555
5	0.14286	0.06257	0.07143	0.07722	2.28333	0.92500
6	0.10714	0.04373	0.05000	0.04854	2.45000	1.03000
7	0.08333	0.03214	0.03704	0.03296	2.59285	1.12381
8	0.06667	0.02453	0.02857	0.02364	2.71786	1.20867
9	0.05455	0.01928	0.02273	0.01767	2.82897	1.28621
10	0.04545	0.01552	0.01852	0.01364	2.92896	1.35762
11	0.03846	0.01274	0.01538	0.01080	3.01988	1.42385
12	0.03297	0.01062	0.01299	0.00874	3.10321	1.48561
13	0.02857	0.00898	0.01111	0.00720	3.18013	1.54349
14	0.02500	0.00769	0.00962	0.00602	3.25156	1.59795
15	0.02206	0.00665	0.00840	0.00509	3.31823	1.64940

Table 2.5: ***Comparison of relative efficiencies of $\tilde{\mu}$ and $\bar{\mu}$***

m	***RP_6***	
	Rectangular distribution	*Right triangular distribution*
2	1.0000	1.0000
3	1.0185	1.0033
4	1.0417	1.0065
5	1.0656	1.0091
6	1.0089	1.0113
7	1.1112	1.0129
8	1.1324	1.0143
9	1.1525	1.0155
10	1.1716	1.0164
11	1.1897	1.0172
12	1.2068	1.0178
13	1.2231	1.0184
14	1.2387	1.0189
15	1.2536	1.0193

CHAPTER 3

ESTIMATION OF THE MEANS OF THE BIVARIATE NORMAL DISTRIBUTION USING MOVING EXTREME RANKED SET SAMPLING WITH CONCOMITANT VARIABLE

3.1 Introduction

Ranking using a concomitant variable first proposed by Stokes (1977) greatly broadened the range of the application of RSS. There are abundant practical situations where a concomitant variable correlated with the variable of interest is available and the measurement of the concomitant variable is cheap and easy. Stokes studied RSS with concomitant variables; she assumed that the variable of interest X has a linear relation with another variable Y. There are situations, when several attributes are to be studied simultaneously using a single combined study rather than separate studies, one for each characteristics. For example, in situations where quantifications entail destruction of units as in uprooting of plants. In this chapter, Moving Extreme Ranked Set Sampling (MERSS) with concomitant variable for the estimation of the means of the bivariate normal distribution, given by Al-Saleh and Al-Ananbeh (2007) has been studied.

3.2 Moving Extreme Ranked Set Sampling with Concomitant Variable

Assume that (X, Y) is a bivariate random vector such that variable Y is difficult to measure or to order by judgment, but the concomitant variable X, which is correlated with Y, is easier to measure or to order by judgment.

The variable X may be used to acquire the rank of Y as follows:

a. Select m units from the bivariate normal distribution using m SRS of sizes $1,2,\ldots,m$, respectively. Identify by judgment the maximum of each set with respect to the variable X.

b. Repeat step a, but for the minimum.

c. Repeat the above two steps r times, if necessary, until the desired sample size, $n = 2rm$, is obtained.

d. Measure accurately the selected n judgment identified units for both variables.

The set of the n pairs obtained using the above procedure, is called a Moving Extreme ranked set sample (MERSS) with concomitant variable.

Assume that a random vector (X, Y) follows a bivariate normal distribution denoted by $BN(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ having joint density $f_{X,Y}(x,y)$ where, $-\infty < x, y, \mu_x, \mu_y < \infty$, $\sigma_x^2, \sigma_y^2 > 0, -1 < \rho < 1$. Let $\{(X_{(1:k)}, Y_{[1:k]}), (X_{(k:k)}, Y_{[k:k]}); k = 1,2,\ldots,m\}$ be a MERSS from $f_{X,Y}(x,y)$, based on the concomitant variable X. If judgment ranking is perfect then, $X_{(i:k)}$ and $Y_{[i:k]}$ are, respectively, the i^{th} smallest value of X from the k^{th} sample and the corresponding value of Y , where i =1 or k.

Then following Stokes (1977), we have

$$Y = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(X - \mu_x) + \varepsilon$$

Where X and ε are independent and ε has mean 0 and variance $\sigma_y^2(1-\rho^2)$, ρ is the correlation between X and Y and μ_x, μ_y, σ_x, σ_y are the means and standard deviations of the variable X and Y.

Note that the pairs of this sample are independent but not identically distributed. Joint density of$(X_{(k:k)}, Y_{[k:k]})$ and $(X_{(1:k)}, Y_{[1:k]})$ is denoted by $f_{k:k}(x,y)$ and $f_{1:k}(x,y)$ respectively:

$$f_{k:k}(x,y) = f_{X(k:k)}(x) f_{Y|X}(y|x)$$

and

$$f_{1:k}(x,y) = f_{X(1:k)}(x) f_{Y|X}(y|x),$$

Where $f_{Y|X}(y|x)$ is the conditional density of Y given X, (see Yang, 1977 and Stokes, 1980a).

Consider the following two estimators of μ_x and μ_y, respectively:

$$\hat{\mu}^*_{xMERSS} = \frac{1}{2m} \sum_{k=1}^{m} \left(X_{(1:k)} + X_{(k:k)}\right)$$

$$\hat{\mu}^*_{yMERSS} = \frac{1}{2m} \sum_{k=1}^{m} \left(Y_{(1:k)} + Y_{(k:k)}\right)$$

Let $\hat{\mu}^*_{xSRS}$ and $\hat{\mu}^*_{ySRS}$be the two corresponding estimators based on a bivariate SRS.

Theorem 3.1:

$\hat{\mu}^*_{xMERSS}$ and $\hat{\mu}^*_{yMERSS}$ are unbiased estimators of ,μ_x and μ_y, respectively

Proof:

Since

$$\left(\frac{X_{(1:k)}-\mu_x}{\sigma_x}\right)^d = -\left(\frac{X_{(k:k)}-\mu_x}{\sigma_x}\right)$$

it follows that

$$E\left(\frac{X_{(1:k)}+X_{(k:k)}}{2}\right)=\mu_x$$

Hence $\hat{\mu}^*_{xMERSS}$ is an unbiased estimator of μ_x.

Also, for i =1or k we have

$$E\,(Y_{[i:k]}) = E(E(Y_{[i:k]}|X_{\,(i:k)})) = E(\mu_y+\rho\frac{\sigma_y}{\sigma_x}(X_{(i:k)}-\mu_x)$$

Thus,

$$E\left(\sum_{k=1}^{m}\frac{Y_{[1:k]}+Y_{[k:k]}}{2m}\right)=\mu_{y.}$$

Let $\mu_{(i:k)}$ and $\sigma^2_{(i:k)}$be respectively, the mean and variance of the i^{th} standard normal order statistic of a SRS of size k . Let $\mu_{x(i:k)}$ and $\sigma^2_{x(i;k)}$ be, respectively, the mean and the variance of the i^{th}order statistic of a SRS of size k from the distribution of X.

Theorem 3. 2:

$$a. \quad Var(\hat{\mu}^*_{xMERSS}) = \frac{\sigma_x^2}{2m^2}\sum_{k=1}^{m}\sigma^2_{(k:K)}$$

$$b. \quad Var\left(\hat{\mu}^*_{yMERSS}\right) = \tfrac{1}{2m^2}\textstyle\sum_{k=1}^{m}\left[\sigma_y^2(1-\rho^2)+\rho^2\sigma_y^2\sigma^2_{(k:k)}\right]$$

Proof:

$$Var\left(\frac{X_{(i:k)}-\mu_x}{\sigma_x}\right)=\frac{1}{\sigma_x^2}\sigma^2_{x(i:k)}, \sigma^2_{x(i:k)}=\sigma^2_{(i:k)}\times\sigma_x^2$$

and

$$\left(\frac{X_{(i:k)}-\mu_x}{\sigma_x}\right)=\mu_{(i:k)}$$

i.e.

$$E\left(X_{(i:k)}\right) = \mu_{x(i:k)} = \mu_{(i:k)}\sigma_x + \mu_x$$

Therefore,

$$Var(\hat{\mu}^*_{xMERSS}) = \frac{\sigma_x^2}{2m^2}\sum_{k=1}^{m}\sigma^2_{(k:k)}$$

Also,

$$Var\left(Y_{[i:k]}\right) = E[Var\left(Y_{[i:k]}\middle|X_{[i:k]}\right)] + Var[E\left(Y_{[i:k]}\middle|X_{[i:k]}\right)]$$
$$= \sigma_y^2(1-\rho^2) + \rho^2\sigma_y^2\sigma^2_{(i:k)}$$

Therefore ,

$$Var\left(\hat{\mu}^*_{yMERSS}\right) = \frac{1}{2m^2}\sum_{k=1}^{m}\left[\sigma_y^2(1-\rho^2) + \rho^2\sigma_y^2\sigma^2_{(k:k)}\right]$$

The efficiency of $\hat{\mu}^*_{xMERSS}$w.r.t. $\hat{\mu}^*_{xSRS}$ is given by:

$$eff(\hat{\mu}^*_{xMERSS}; \hat{\mu}^*_{xSRS}) = \frac{Var(\hat{\mu}^*_{xSRS})}{Var(\hat{\mu}^*_{xMERSS})} = m\left(\sum_{k=1}^{m}\sigma^2_{(i:k)}\right)^{-1}$$

Theorem 3.3: $eff\ (of\ \hat{\mu}^*_{xMERSS}, \hat{\mu}^*_{xSRS}) \geq 1.$

Proof:

For the order statistics of a SRS of size m from N(0,1), $\sum_{j=1}^{m}\sigma^2_{(i,j:m)} = 1$ for $i = 1,\ldots,m$, where $\sigma^2_{(i,j:m)} = cov\left(X_{(i:m)}, X_{(j:m)}\right)$;in other words, the sum of the elements in a row or a column of the covariance matrix of the standard normal order statistics is 1 for any sample of size m (See Arnold et al.,1992, p. 91). Since $\sigma^2_{(i,j:m)} > 0$ (Lehmann, 1966), it follows that $\sigma^2_{(i:m)} \leq 1$ and $\sum_{k=1}^{m}\sigma^2_{(1:k)} \leq m$.

Hence,

$$eff(\hat{\mu}^*_{xMERSS}, \hat{\mu}^*_{xSRS}) \geq 1$$

Similarly,

$$eff\left(\hat{\mu}^*_{yMERSS}, \hat{\mu}^*_{ySRS}\right) = \left\{1 - \rho^2\left[1 - \frac{\sum_{k=1}^{m}\sigma^2_{(i:k)}}{m}\right]\right\}^{-1}$$

$$= \left\{1 - \rho^2\left\{1 - \left(eff(\hat{\mu}^*_{xMERSS}, \hat{\mu}^*_{xSRS})\right)^{-1}\right\}\right\}^{-1}$$

Tables 3.1 and 3.2, give *eff* $(\hat{\mu}^*_{xMERSS}, \hat{\mu}^*_{xSRS})$ and *eff* $(\hat{\mu}^*_{yMERSS}, \hat{\mu}^*_{ySRS})$, respectively, for various values of *m.* The two efficiencies are also presented graphically in Figures 3.1 and 3.2, with *r* standing for ρ. Based on Table 3.1, *eff* $(\hat{\mu}^*_{xMERSS}, \hat{\mu}^*_{xSRS})$ is always larger than 1 and is increasing in *m* . Based on Table 3.2, we conclude the *eff* $(\hat{\mu}^*_{yMERSS}, \hat{\mu}^*_{ySRS})$, is always larger than 1 and is increasing in *m* for fixed $|\rho|$; it is increasing in $|\rho|$ for fixed *m.* Note that the efficiency is very close to 1 when $|\rho|$ is small; thus for this method to be beneficial, it is necessary that the r elation between the two variables is fairly strong.

3.2.1 Comparing MERSS and RSS

For the purpose of comparing the MERSS procedure with the usual RSS procedure, if the bivariate sample is obtained using the balanced RSS with concomitant variable, then the efficiency of the procedure can be obtained using a result reported by Stokes (1977). The efficiency of the RSS estimator of μ_x w.r.t the SRS estimator is

$$E_1 = m\left(\sum_{k=1}^{m} \sigma^2_{(i:m)}\right)^{-1}$$

In practice, when using RSS, *m* should not be large. For example, for *m = 1, 2,3,4,5,* the values of E_1 are 1, 1.46, 1.91, 2.35, and 2.77, respectively. The efficiency of the estimator of, μ_y with respect to the corresponding estimator based on a bivariate SRS is given by

$$\left\{1 - \rho^2\{1 - E^{-1}\}\right\}^{-1}$$

Numerical values of the efficiency are given in Table 3.2 for *m = 1,...,5*. From a theoretical perspective, usual RSS with concomitant variable is significantly more efficient than MERSS. In choosing between the two procedures, the efficiency as well as the applicability should be taken into account. In practice, MERSS is easier to apply than RSS. Also the total number of sample points needed to be available to obtain a MERSS is much less than that needed to obtain a RSS of the same size. For example in order to obtain a MERSS of size 2*m*, we need to identify *m* (*m* –1) sample points; the number is $2m^2$ in the case of RSS.

3.3. Robustness of the MERSS procedure

It has been seen from previous sections, that if the underlying distribution is the bivariate normal, the MERSS is always preferable to bivariate SRS. One may ask about the suitability of this procedure if the bivariate normality assumption is not valid, i.e. is the favorable properties of the procedure robust against departure from normality. First of all, the unbiasedness of the estimators is established because the marginal distributions in the bivariate normal case are symmetric; thus, departure from symmetry may lead to biased estimators. For the study of robustness of the procedure, one possibility is to consider the model;

$$Y_i = \mu_y + \rho \frac{\sigma_y}{\sigma_x}(X_i - \mu_x) + Z_i$$

where Z_i are independent of X_i . Assume that the marginal distributions of X and Y are symmetric about their means, see Stokes (1977). $E(Z_i) = 0$ and $Var(Z_i) = {\sigma_y}^2(1-\rho^2)$. Furthermore, since Z_i are independent of X_i, we have

$$Y_{[i:m]} = \mu_y + \rho \frac{\sigma_y}{\sigma_x}\left(X_{(h:m)} - \mu_x\right) + Z_i$$

All properties of the estimators derived in section 3.2 are valid under this model. Note that the bivariate normal random variables satisfy the assumption of the above model.

3.4. Application

MERSS procedure has been illustrated using a real data set, which consists of the height X and the diameter Y of 1083 trees (Prodan1968). For this data set, regarded as a population, we have ρ = 0.715, μ_x = 21.6, μ_y = 22.6, σ_x = 2.96, σ_y = 5.62.

Al-Hadhrami (2001) investigated the bivariate normality of the data and suggested removing the lowest 20 values of each variable to achieve marginal normality. Table 3.3 gives the efficiency of the MERSS estimators of μ_x and μ_y. Based on Table 3.3, the efficiency is always larger than 1 and is increasing in m.

3.5. Conclusions

MERSS is a very useful modification of RSS, which allows for an increase in the set size *m* without introducing ranking errors. MERSS uses only the two extremes values, maximum or minimum of sets of varied size, but RSS needs the ranking of all the elements of each set. MERSS has been used with concomitant variable to estimate the two means of the bivariate normal distribution. It appears that the use of MERSS with concomitant variable is highly beneficial when compared to SRS for estimating the population means. The estimators obtained are unbiased and more efficient than those obtained using SRS; in addition, their asymptotic efficiency is always greater than one. For the procedure to be used in practice, it is essential that the maximum and the minimum can be identified for one of the variables by judgment, whereas the other variable should be highly correlated with the first variable.

Table 3.1: Efficiency of $\widehat{\mu}^{*}_{xMERSS}$ w.r.t. $\widehat{\mu}^{*}_{xSRS}$

m	1	2	3	4	5	6	7	8	9	10
eff	1.00	1.19	1.34	1.46	1.57	1.67	1.76	1.83	1.91	1.98

Table 3.2: Efficiency of $\widehat{\mu}^{*}_{yMERSS}(\widehat{\mu}^{*}_{yRSS})$ w.r.t $\widehat{\mu}^{*}_{ySRS}$

$m\downarrow \mid \rho \mid \rightarrow$	**0.2**	**0.4**	**0.6**	**0.8**
1	1.00(1.00)	1.00(1.00)	1.00(1.00)	1.00(1.00)
2	1.01(1.01)	1.03(1.05)	1.06(1.13)	1.11(1.26)
3	1.01(1.02)	1.04(1.08)	1.10(1.21)	1.19(1.44)
4	1.01(1.02)	1.05(1.10)	1.13(1.26)	1.25(1.58)
5	1.02(1.03)	1.06(1.11)	1.15(1.30)	1.30(1.69)
6	1.02	1.07	1.17	1.34
7	1.02	1.07	1.18	1.38
8	1.02	1.08	1.20	1.41
9	1.02	1.08	1.21	1.44
10	1.02	1.09	1.22	1.46

Table 3.3: Efficiency for trees data $\hat{\mu}^*_{xMERSS}$ ($\hat{\mu}^*_{yMERSS}$) w.r.t.$\hat{\mu}^*_{xSRS}$($\hat{\mu}^*_{ySRS}$)

m	$eff(\hat{\mu}^*_{xMERSS};\ \hat{\mu}^*_{xSRS})$	$eff(\hat{\mu}^*_{yMERSS};\ \hat{\mu}^*_{ySRS})$
1	1.00	1.00
2	1.19	1.11
3	1.35	1.20
4	1.45	1.25
5	1.56	1.28
6	1.64	1.33
7	1.74	1.35
8	1.76	1.37
9	1.82	1.35
10	1.81	1.32

Figure 3.1: The efficiency of Table 3.1

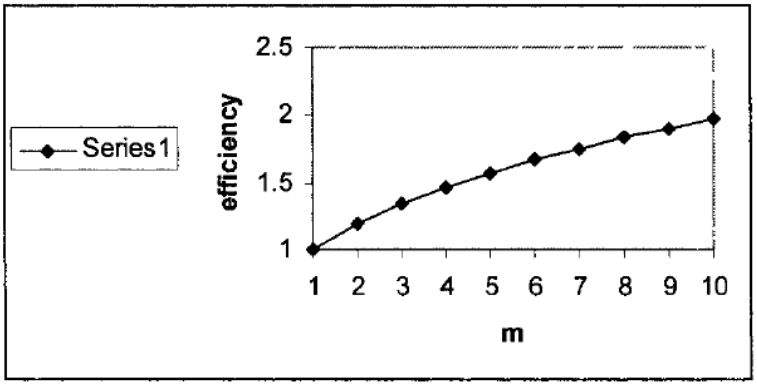
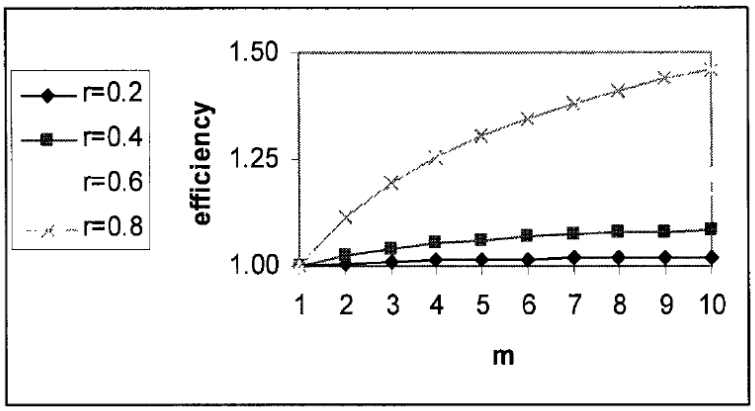


Figure 3.2: The efficiency of Table 3.2



Chapter 4

Estimation of Simple Linear Regression Model Using L Ranked Set Sampling

4.1 Introduction

Regression analysis is a conceptually simple method for investigating functional relationships among variables. The relationship is expressed in the form of an equation or model connecting the response variable *(Y)* and one (*X*) or more explanatory variables. The simple true relationship can be approximated by the regression model

$$Y = \alpha + \beta X + \varepsilon$$

Where ε is assumed to be random error, α, β are unknown regression parameters to be estimated from the data.

In areas such as medical studies, quantitative genetics, and ecological and environmental studies, there are abundant situations where, in the context of regression, the measurement of the response variable is costly or time consuming but the measurement of the predictor variable can be obtained easily with relatively negligible cost. For example, in the assessment of the association between certain biomarkers and exposure level in cancer studies, the measurement of biomarkers involves expensive and time-consuming laboratory investigation but the measurement of exposure level can be easily obtained. Other examples can be found in animal growth studies where the ages of animals need to be determined but aging an animal is usually time consuming and costly, and sometimes there is even need to sacrifice the animal. However, variables on the physical size of an animal, which are costly related to age, can be collected easily and cheaply. Sampling strategies that can reduce cost and increase efficiency are highly desirable in these cases.

Many authors have used RSS technique in regression analysis. Patil et al. (1993b) compared the RSS sample and SRS sample in relation to the concomitant variable and the regression estimate. Yu and Lam (1997) proposed a regression-type estimator based on RSS. They demonstrated that this estimator is always more efficient than the regression estimator using SRS and is also more efficient than the estimator proposed by Patil et al. (1993b) unless the correlation coefficient is low ($|\rho| < 0.4$). Muttlak (1995) used RSS to estimate the parameters of the simple linear regression model treating the regressor *X* as a constant. Chen (2001b) did an extensive study on the

properties of regression type estimates. Chen and Wang (2004) studied the optimal RSS for the regression analysis. Samawi and Ababneh (2001) and earlier Samawi et al (1996a), showed that the extreme ranked set sampling (ERSS) performed better than RSS at estimating model parameters.

The current study uses generalized ranked data procedure (LRSS) (Al-Nasser (2007)).

4.2. Estimation of Mean using L Ranked Set Sampling (LRSS)

Based on the LRSS scheme, explained in the estimator of the population mean when r=1 is defined as:

$$\hat{\mu}_{LRSS} = \frac{1}{M}\left(\sum_{i=1}^{K} X_{i(k+1)} + \sum_{i=k+1}^{m-k} X_{i(i)} + \sum_{i=m-k+1}^{m} X_{i(m-k)}\right)$$

and its variance is given by:

$$var(\hat{\mu}_{LRSS}) = \frac{1}{m^2}\left(\sum_{i=1}^{K} var\left(X_{i(k+1)}\right) + \sum_{i=k+1}^{m-k} var(X_{i(i)}) + \sum_{i=m-k+1}^{m} var\left(X_{i(m-k)}\right)\right)$$

Al-Nasser proved that $\hat{\mu}_{LRSS}$ is unbiased estimator of the population mean μ, and has smaller variance than $\hat{\mu}_{SRS}$ if the underlying distribution is symmetric.

4.3 Bivariate L Ranked Set Sampling (LRSS)

Al-Nasser and Radaideh (2008) used a modified bivariate LRSS to estimate parameter in the simple linear regression model.

In order to have a Bivariate L ranked set sample, the following steps are performed:

Step1: Randomly draw *m* independent sets each containing *m* bivariate sample units.

Step2: Rank the units within each sample with respect to the *X's* by visual inspection or any other cost effective method.

Step3: Select LRSS coefficient, $K = [mp]$ such that $0 \leq p < 0.5$, and $[X]$ the largest integer value less than or equal to X.

Step4: For each of the first $(k + 1)$ ranked s amples; select the unit with rank $k + 1$ and measure the Y value that corresponding to $x_{(k+1)i}$ and denote it by $y_{[k+1]i}$.

Step5: For $j=k+2,\ldots\ldots, m-k-1$, the unit with rank j in the j^{th} ranked sample is selected and measures the y value that corresponds.

Step6: The procedure continued until (m-k)th unit selected from the each of the last $(m - k)^{th}$ ranked samples, with respect to the first characteristic and measure the correspond y value.

For example, if $k = 1$ and $m = 5$ then the selected ranked sample will be as given in Table. 4.1.

Table 4.1: Selected Bivariate LRSS when m = 5 and k = 1.

$x_{(1)1},y_{[1]1}$	$x_{(1)2},y_{[1]2}$	$x_{(1)3},y_{[1]3}$	$x_{(1)4},y_{[1]4}$	$x_{(1)5},y_{[1]5}$
$x_{(2)1},y_{[2]1}$	$x_{(2)2},y_{[2]2}$	$x_{(2)3},y_{[2]3}$	$x_{(2)4},y_{[2]4}$	$x_{(2)5},y_{[2]5}$
$x_{(3)1},y_{[3]1}$	$x_{(3)2},y_{[3]2}$	$x_{(3)3},y_{[3]3}$	$x_{(3)4},y_{[3]4}$	$x_{(3)5},y_{[3]5}$
$x_{(4)1},y_{[4]1}$	$x_{(4)2},y_{[4]2}$	$x_{(4)3},y_{[4]3}$	$x_{(4)4},y_{[4]4}$	$x_{(4)5},y_{[4]5}$
$x_{(5)1},y_{[5]1}$	$x_{(5)2},y_{[5]2}$	$x_{(5)3},y_{[5]3}$	$x_{(5)4},y_{[5]4}$	$x_{(5)5},y_{[5]5}$

4.4 Estimating Simple Linear Regression Parameters

In completion of the sampling, let $d^{x}_{(i)j}$and $d^{y}_{(i)j}$be, respectively, X with rank k and the corresponding value of Y obtained from the i^{th} set in the j^{th} cycle.

Then, the regression equation based on bivariate LRSS can be modeled as:

$$d^{y}_{[i]j} = \alpha + \beta d^{x}_{(i)j} + d^{\varepsilon}_{[i]j} \qquad i = 1,\ldots,m \;\; j = 1,\ldots,r$$

$$d^{y}_{[i]j} = \begin{cases} Y_{[k+1]j} & i \le k \\ Y_{[i]j} & k+1 \le i \le m-k \;\; ; j = 1,2,\ldots,r \\ Y_{[m-k]j} & m-k+1 \le i \le m \end{cases} \qquad (4.4.1)$$

$$d^{x}_{(i)j} = \begin{cases} X_{(k+1)j} & i \le k \\ X_{(i)j} & k+1 \le i \le m-k \;\; ; j = 1,2,\ldots,r \\ X_{(m-k)j} & m-k+1 \le i \le m \end{cases}$$

where $d^{\varepsilon}_{[i]j}$ is the random error. Under the regular assumptions of simple linear regression model Draper and Smith (1981), the least square estimates of the regression parameters mentioned in (4.4.1) are given by:

$$\hat{\beta}_{LRSS} = \frac{\sum_{j=1}^{r}\left[\sum_{i=1}^{m}\left(d^{x}_{(i)j} - d^{\bar{x}}\right)\left(d^{y}_{[i]j} - d^{\bar{y}}\right)\right]}{\sum_{j=1}^{r}\left[\sum_{i=1}^{m}\left(d^{x}_{(i)j} - d^{\bar{x}}\right)^{2}\right]} \qquad (4.4.2)$$

And

$$\hat{\alpha}_{LRSS} = d^{\bar{y}} - \hat{\beta}_{LRSS} d^{\bar{x}}$$

Where

$$d^{\bar{y}} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} d^{y}_{[i]j} \qquad and \qquad d^{\bar{x}} = \frac{1}{rm}\sum_{j=1}^{r}\sum_{i=1}^{m} d^{x}_{(i)j}$$

Hereafter, the fitted model will be:

$$d^{\hat{y}}_{[i]j} = \hat{\alpha}_{LRSS} + \hat{\beta}_{LRSS}\, d^{x}_{(i)j}$$

note that, the estimated residuals are given by

$$d^{\varepsilon}_{[i]j} = d^{y}_{[i]j} - d^{\hat{y}}_{[i]j}$$

Theorem1: Assume that (4.4.1) is satisfied then:

1. $E\left(\hat{\beta}_{LRSS}\right) = \beta$
2. $E(\hat{\alpha}_{LRSS}) = \alpha$
3. $var\left(\hat{\beta}_{LRSS}\right) = \sum_{j=1}^{r}\left[\sum_{i}^{m}\left(\sigma^{2}_{[i]j} \times \frac{\left(d^{x}_{(i)j} - d^{\bar{x}}\right)^{2}}{Sxx^{2}}\right)\right]$
4. $var(\hat{\alpha}_{LRSS}) = \frac{1}{m^{2}}\sum_{j=1}^{r}\sum_{i=1}^{m}\sigma_{d}^{2} + \sum_{j=1}^{r}\left[\sum_{i}^{m}\left(\sigma^{2}_{[i]j} \times \frac{\left(d^{x}_{(i)j} - d^{\bar{x}}\right)^{2}}{Sxx^{2}}\right)\right]$

Proof: Without loss of generality suppose that r= 1 then $\hat{\beta}$ given (4.4.2) can be rewritten as

$$\hat{\beta}_{LRSS} = \frac{S_{d^x d^y}}{S_{d^x d^x}} = \sum_{i=1}^{m} c_i \times d_{[i]}^{y}$$

Where $c_i = \frac{\left(d_{(i)}^{x} - d^{\bar{x}}\right)^2}{S_{d^x d^x}}$, $S_{d^x d^x} = \sum_{i=1}^{m}\left(d_{(i)}^{x} - d^{\bar{x}}\right)^2$ and

$$S_{d^x d^y} = \sum_{i=1}^{m}\left(d_{(i)}^{x} - d^{\bar{x}}\right)\left(d_{(i)}^{y} - d^{\bar{y}}\right)$$

Now $E\left(\hat{\beta}_{LRSS}\right) = E\left(\sum_{i=1}^{m} c_i \times d_{[i]}^{y}\right) = \sum_{i=1}^{m}\left(c_i \times E(d_{[i]}^{y})\right)$

$$= \sum_{i=1}^{m}\left(c_i E\left(\alpha + \beta d_{(i)}^{x} + d_{[i]}^{\varepsilon}\right)\right) = \alpha \sum_{i=1}^{m}(c_i) + \beta \sum_{i=1}^{m}(c_i)\, d_{(i)}^{x}$$

Note that : $\sum_{i=1}^{m}(c_i) = \sum_{i=1}^{m}\left(\frac{\left(d_{(i)}^{x} - d^{\bar{x}}\right)}{S_{d^x d^x}}\right) = \frac{1}{S_{d^x d^x}}\sum_{i=1}^{m}\left(d_{(i)}^{x} - d^{\bar{x}}\right) = 0$

and

$$\sum_{i=1}^{m}\left(c_i \times d_{(i)}^{x}\right) = \frac{1}{S_{d^x d^x}}\sum_{i=1}^{m}\left(d_{(i)}^{x} - d^{\bar{x}}\right) d_{(i)}^{x}$$

$$= \frac{1}{S_{d^x d^x}}\left(\sum_{i=1}^{m}\left(\left(d_{(i)}^{x}\right)^2 - d_{(i)}^{x} d^{\bar{x}}\right)\right)$$

$$= \frac{1}{S_{d^x d^x}}\left(\sum_{i=1}^{m}\left(d_{(i)}^{x}\right)^2\right) - m(d^{\bar{x}})^2 = \frac{S_{d^x d^x}}{S_{d^x d^x}} = 1$$

Therefore; $E\left(\hat{\beta}_{LRSS}\right) = \beta$

2- Now for the intercept estimator we have

$$E(\hat{\alpha}_{LRSS}) = E(d^{\bar{y}}) - d^{\bar{x}} E\left(\hat{\beta}_{LRSS}\right) = (\alpha + \beta d^{\bar{x}}) - \beta d^{\bar{x}} = \alpha$$

3- $var\left(\hat{\beta}_{LRSS}\right) = var\left(\sum_{i=1}^{m} c_i \times d_{[i]}^{y}\right)$

$$= \sum_{i=1}^{m} c_i^2 \, var\left(d_{[i]}^{y}\right) = \sum_{i=1}^{m} \sigma_{[i]}^2 c_i^2 = \sum_{i=1}^{m} \left(\sigma_{[i]}^2 \frac{\left(d_{(i)}^{x} - d^{\bar{x}}\right)^2}{S_{d^x d^x}{}^2} \right)$$

4- $var(\hat{\alpha}_{LRSS}) = var(d^{\bar{y}}) + (d^{\bar{x}})^2 \, var(\hat{\beta}_{LRSS})$

$$= \frac{\sigma_{d^{\varepsilon}{}_{[i]}}^2}{m} + (d^{\bar{x}})^2 \left(\sum_{i=1}^{m} \left(\sigma_{[i]}^2 \frac{\left(d_{(i)}^{x} - d^{\bar{x}}\right)^2}{S_{d^x d^x}{}^2} \right) \right)$$

$$= \frac{1}{m^2} \sum_{i=1}^{m} \sigma_{d^{\varepsilon}{}_{[i]}}^2 + (d^{\bar{x}})^2 \left(\sum_{i=1}^{m} \left(\sigma_{[i]}^2 \frac{\left(d_{(i)}^{x} - d^{\bar{x}}\right)^2}{S_{d^x d^x}{}^2} \right) \right)$$

Following Yu and Lam (1997) the LRSS regression estimator is given by

$$\hat{\mu}_{LRSS.Reg} = d^{\bar{y}} + \hat{\beta}(\bar{X} - d^{\bar{x}})$$

Moreover, under model (4.4.1) and the above assumptions, then for fixed value of r we have

$$\frac{\left(\hat{\beta} - \beta\right)}{\sqrt{Var(\hat{\beta})}} \xrightarrow{L} N(0,1), as\ m \to \infty$$

and

$$\frac{(\hat{\alpha} - \alpha)}{\sqrt{Var(\hat{\alpha})}} \xrightarrow{L} N(0,1), as\ m \to \infty$$

The proof of these results are concluded directly using the ideas of RSS (Chen et al (2004).

4.5. Simulation Study

To illustrate the performance of the LRSS estimator's Monte Carlo simulation studies were conducted considering two cases inliers and outlier cases. The simulation plan has the following assumptions:

- Generate 10000 random samples using SRS, RSS, ERSS and LRSS (with k=1, 2).
- Set the number of cycles r = 5, 10, 20, and set size m = 5, 6, 7, 8.

- Initiate the strength of the association between the two variables by ρ = 0.1, 0.5 and 0.9.
- The intercept and the slope are initialed as α =0 and $\beta = \rho$.
- The error term is generated from (0, 1- ρ^2) and the regressor from N(0,1).
- Also, an outlier case is considered, by generating an outlier (one observation). For this observation we generate the error term from N $(0,5^2)$.
- The relative efficiency (RE) for the estimated model based on LRSS is computed according to the following expression:

$$RE = \frac{MSE\left(\hat{\mu}_{SRS.Reg}\right)}{MSE\left(\hat{\mu}_{LRSS.Reg}\right)}$$

The results of the MSE for the SLR model for inliers case is given in Table.4.2 – Table.4.4; and the results for outlier cases are given in Table 4.5 – Table 4.7.

The simulation results indicate that estimation of the simple linear regression model using LRSS is more efficient than using the traditional sampling techniques; SRS, ERSS or RSS. Moreover, when the data contains outliers the LRSS is shown to be a robust technique, and as the value of K increases the RE increases. Moreover, the RE of regression estimators decreases as the set size or the cycle size increases. Also, for fixed *r* and *m*, the RE decreases whenever ρ increases. It seems that, for a moderate or large sample size, the RE is slightly different when using either RSS or ERSS. However, using LRSS is generally more efficient than using SRS, ERSS or RSS for regression analysis.

4.6. Illustration Using Real Data

An illustration of the LRSS procedure in estimation using simple linear regression is discussed based on a real data set from Platt et al (1988).

4.6.1 Real Data Set

The original data were collected on seven variables about tree characteristics of which only two have been used here: *X*, the diameter in

centimeters at breast height and *Y*, the entire height in feet. The regression model is analyzed assuming that the population is consists of 375 trees. The summary statistics of the data are reported in Table.4.8.

Based on the entire measurements a random sample of size 75 is drawn by using different sampling schemes, SRS, RSS, ERSS, and LRSS (k= 1, 2). In RSS, ERSS and LRSS procedure we use m sets each of size m, where $m=5$, and repeat this cycle fifteen times "i.e., r = 15" to achieve a sample of size *75*. The summary statistics of the selected random samples is presented in Table.4.9.

It can be noted that, the average of regressor varied from 17.3-28.5 and response from 42.7-79.3 depends on which sampling scheme is used.

4.6.2 Data Analysis

In order to form the regression model based on different sampling scheme, a visual inspection using scatter plot is used (Figure. 4.1).

The scatter plots in Figure. 4.1 suggested that the relationship between both variables is not linear. Therefore, both variables are re-expressed by a natural logarithmic transformation. After here, the least square method is used for model fitting; the estimates of the regression parameters are given in Table. 4.10.

The results suggest that the RSS, $LRSS_1$ and $LRSS_2$ perform well compared to the SRS and ERSS in regards MSE point of view. Also, it can be noted that using RSS the intercept and slope have the minimum standard error and the highest fitting measure (i.e., 93.2%). Moreover, the residual plot and the normality p-p plot Figure.4.2 suggest that the model reasonably fits the data using these methods. In conclusion, from the data analysis and simulation results; the LRSS produced a satisfactory estimation for simple linear regression compared to the SRS and the other ranked data sampling schemes.

Table 4.2: RE for Regression model with ρ= 0.1					
r	***m***	**ERSS**	**RSS**	**$LRSS_1$**	**$LRSS_2$**
5	5	.562	.984	1.964	3.467
	6	.486	.978	1.759	3.446
	7	.452	.974	1.612	3.001
	8	.413	.982	1.526	2.651
10	5	.567	.981	1.959	3.438
	6	.493	.988	1.772	3.441
	7	.458	.991	1.638	3.042
	8	.418	.990	1.538	2.659
20	5	.573	.997	1.972	3.462
	6	.497	.995	1.783	3.463
	7	.460	.994	1.642	3.042
	8	.421	.998	1.550	2.680

Table 4.3: RE for Regression model with ρ= 0.5					
r	***m***	**ERSS**	**RSS**	**$LRSS_1$**	**$LRSS_2$**
5	5	.659	.978	1.723	2.869
	6	.601	.973	1.560	2.831
	7	.580	.983	1.470	2.536
	8	.550	.984	1.397	2.246
10	5	.672	.989	1.732	2.874
	6	.612	.988	1.581	2.854
	7	.588	.993	1.483	2.549
	8	.557	.992	1.409	2.266
20	5	.675	.996	1.738	2.861
	6	.618	.994	1.591	2.860
	7	.591	.995	1.487	2.547
	8	.559	.993	1.409	2.262

Table 4.4: RE for Regression model with ρ= 0.9					
r	***m***	**ERSS**	**RSS**	**$LRSS_1$**	**$LRSS_2$**
5	5	.907	.990	1.179	1.465
	6	.898	.989	1.137	1.464
	7	.893	.996	1.116	1.385
	8	.887	.995	1.097	1.317
10	5	.913	.991	1.182	1.476
	6	.903	.996	1.148	1.471
	7	.890	.995	1.117	1.389
	8	.885	.996	1.101	1.320
20	5	.918	.998	1.188	1.471
	6	.902	.999	1.149	1.470
	7	.895	.996	1.120	1.390
	8	.886	.997	1.102	1.319

Table 4.5: RE for Regression model with ρ= 0.1: outlier case					
r	***m***	**ERSS**	**RSS**	**$LRSS_1$**	**$LRSS_2$**
5	5	0.474	0.980	3.537	6.668
	6	0.404	0.987	2.992	6.205
	7	0.368	0.988	2.635	5.184
	8	0.330	0.979	2.361	4.286
10	5	0.477	0.986	3.576	6.681
	6	0.406	0.989	3.019	6.238
	7	0.372	0.996	2.663	5.210
	8	0.335	0.990	2.404	4.361
20	5	0.483	0.996	3.594	6.709
	6	0.411	0.996	3.056	6.297
	7	0.372	0.998	2.669	5.219
	8	0.335	0.996	2.404	4.359

Table 4.6: RE for Regression model with ρ= 0.5: outlier case

r	*m*	ERSS	RSS	$LRSS_1$	$LRSS_2$
5	5	0.491	0.976	3.240	5.689
	6	0.423	0.975	2.758	5.276
	7	0.389	0.977	2.43 1	4.396
	8	0.363	0.989	2.226	3.753
10	5	0.498	0.989	3.263	5.694
	6	0.434	0.995	2.794	5.300
	7	0.397	0.993	2.484	4.468
	8	0.365	0.996	2.250	3.760
20	5	0.503	1.000	3.286	5.723
	6	0.435	0.995	2.803	5.297
	7	0.398	0.992	2.475	4.440
	8	0.366	0.998	2.259	3.784

Table 4.7: RE for Regression model with ρ= 0.9: outlier case

r	*m*	ERSS	RSS	$LRSS_1$	$LRSS_2$
5	5	0.612	0.984	1.922	2.727
	6	0.554	0.975	1.718	2.544
	7	0.542	0.993	1.616	2.270
	8	0.506	0.983	1.508	2.014
10	5	0.618	0.998	1.928	2.724
	6	0.565	0.993	1.748	2.567
	7	0.536	0.991	1.620	2.268
	8	0.513	0.997	1.537	2.050
20	5	0.621	1.000	1.941	2.724
	6	0.571	0.998	1.745	2.554
	7	0.538	0.995	1.614	2.267
	8	0.513	1.002	1.538	2.039

Table 4.8: Summary Statistics for the Tree Data

	Diameter(x) in cm	Entire Height (y) in feet
N	375	375
Mean	21.8971	54.83
Std. Deviation	17.63671	57.656
Range	73.2	242

Table 4.9: Summary Statistics for the selected samples of size 75

		Range	Minimum	Maximum	Mean	Std. Deviation
SRS	*x*	66.90	2.30	69.20	20.1227	17.79634
	y	219.00	4.00	223.00	48.8933	58.30896
RSS	*x*	66.90	2.30	69.20	21.4427	18.96384
	y	219.00	4.00	223.00	55.8400	64.10023
$LRSS_1$	*x*	48.70	4.20	52.90	18.5333	13.57199
	y	205.00	6.00	211.00	42.7600	44.77718
$LRSS_2$	*x*	41.40	5.10	46.50	17.3213	11.50245
	y	203.00	8.00	211.00	43.3467	48.53516
ERSS	*x*	66.90	2.30	69.20	28.4773	23.89840
	y	219.00	4.00	223.00	79.3200	79.59614

Table 4.10: Regression Analysis of Tree data

Method	Constant	Log(Diameter)	Adj(R^2)	MSE
SRS	0.556* (0.130)	1.066* (0.047)	0.875	0.134
RSS	0.468* (0.099)	1.120* (0.035)	0.932	0.080
ERSS	0.525* (0.116)	1.112* (0.38)	0.920	0.138
$LRSS_1$	0.531* (0.124)	1.073* (0.045)	0.885	0.080
$LRSS_2$	0.614* (0.154)	1.063* (0.057)	0.826	0.090

Note: Standard Errors in parentheses; * Statistically Significant at 1%.

Figure 4.1: Scatter Plot of Tree Data by Using Different Sampling Schemes

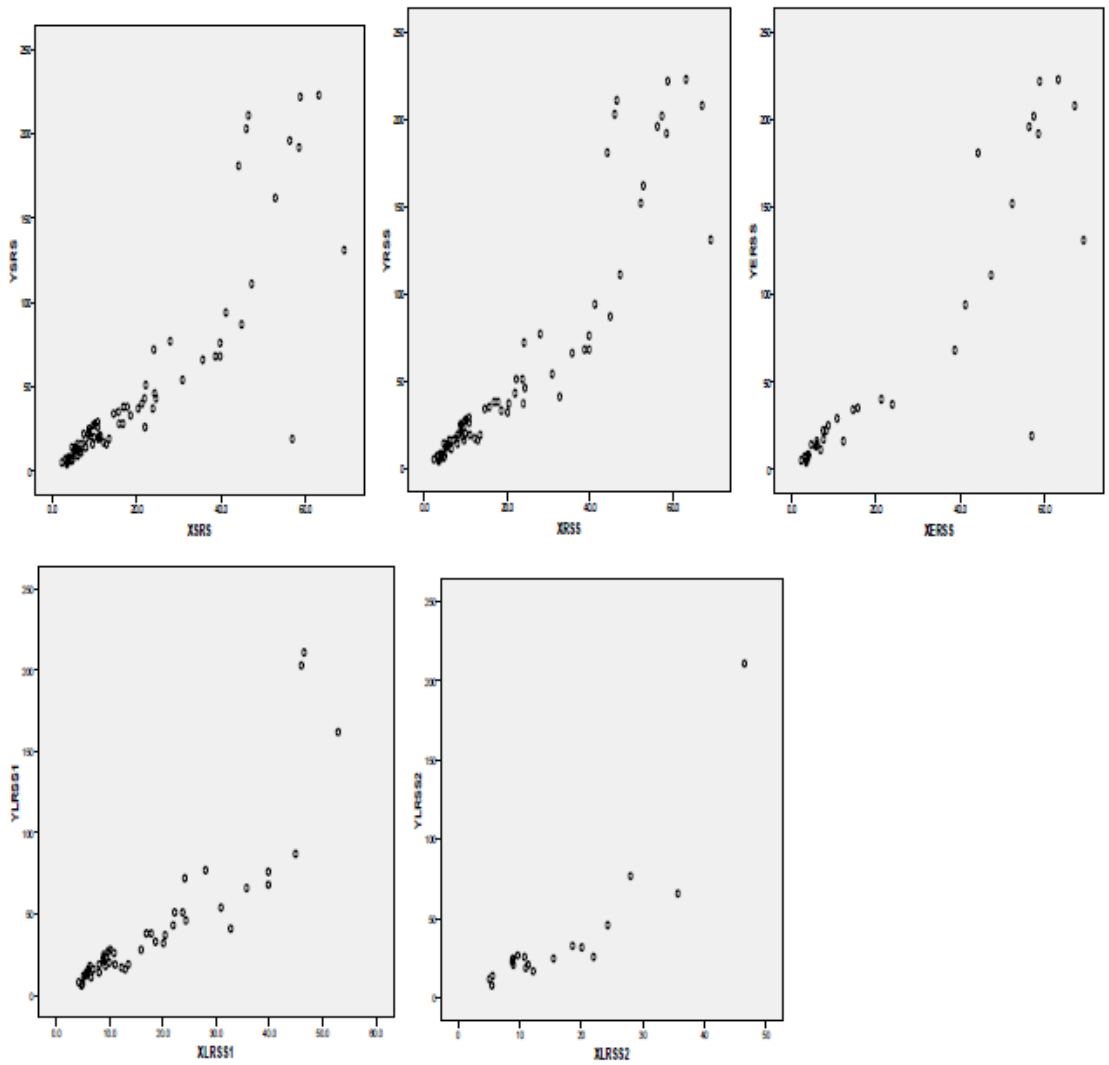
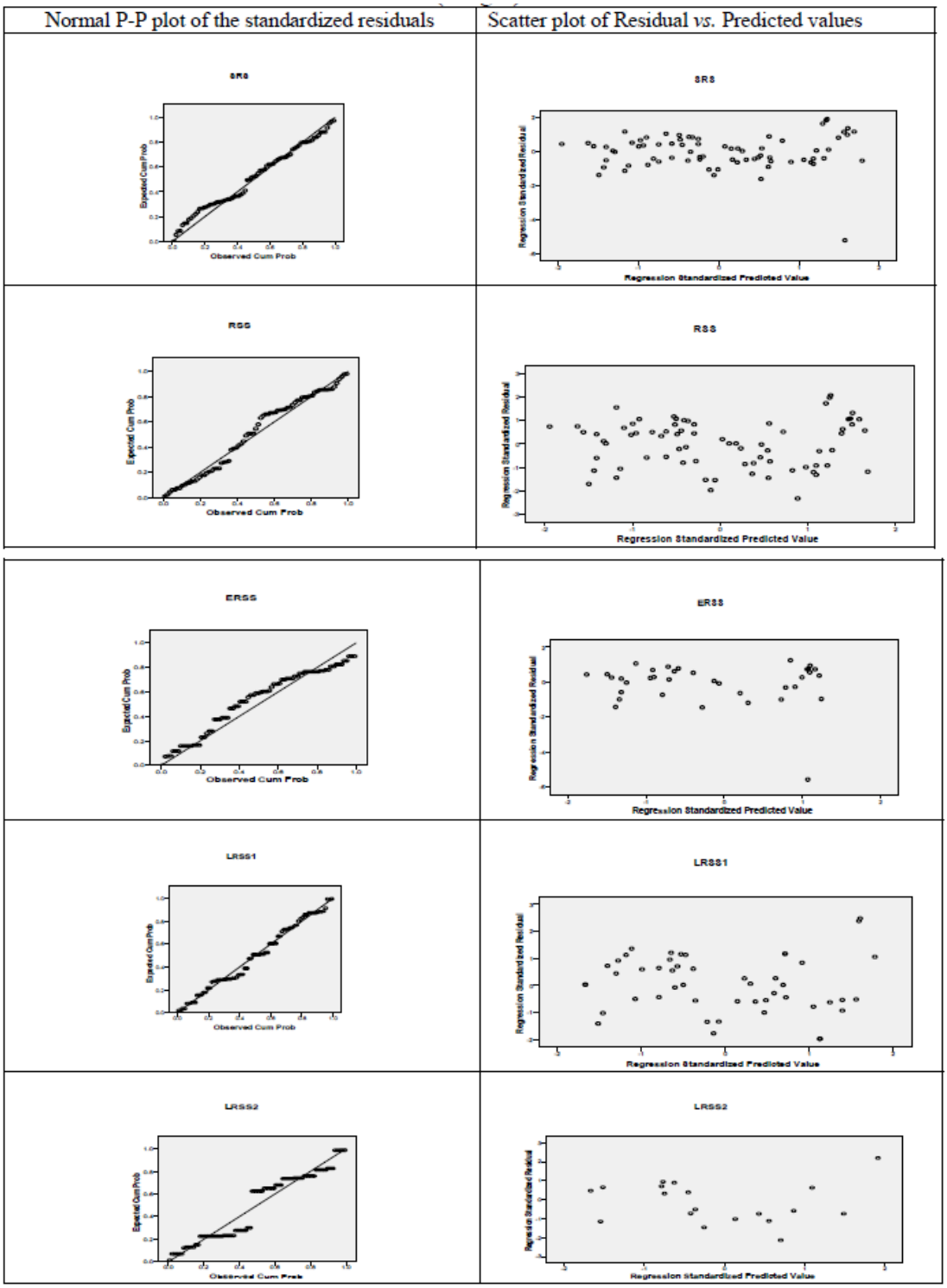


Figure 4.2: Residual Analysis using Different Sampling Scheme: "Response is Ln (Height)"



CHAPTER 5

ESTIMATION OF THE POPULATION MEAN USING STRATIFIED QUARTILE RANKED SET SAMPLING

1 Introduction

In stratified sampling method, the population of N units is divided into L non overlapping subpopulations each of $N_1, N_2, \ldots N_L$ units, respectively, and $N_1+N_2+,\ldots+N_L=N$. These subpopulations are called strata. For full benefit from stratification, the size of the hth subpopulation, denoted by N_h for $h=1,2,\ldots,L$, must be known. Then the samples are drawn independently from each stratum, producing samples sizes denoted by $n_1, n_2,\ldots,n_L$, such that the total sample size is $n = \sum_{h=1}^{L} n_h$.

If a simple random sample is taken from each stratum, the whole procedure is known as Stratified Simple Random Sampling (SSRS). On the other hand, in Stratified Ranked Set Sampling (SRSS) procedure, rather than selecting a simple random sample within each stratum, as is done in stratified simple random sampling , a ranked set sample is taken within each stratum. This sampling design combines the variance reduction that arises from stratifying the population with the increased precision RSS holds over SRS.

The estimator for the population mean under stratified ranked set sampling is as follows:

$$\bar{X}_{SSRS} = \sum_{h=1}^{L} \frac{N_h}{N} \left(\frac{1}{n_h} \sum_{i=1}^{n_h} X_{ih} \right)$$

where X_{ih} is the measurement for the ith unit sampled from stratum h, N_h is the population size in stratum h, N is the total population size, and L is the total number of strata. This quantity is the weighted average of the ranked set sampling estimators for the mean of each stratum. Under this sampling design, one needs to stratify the elements of the population a priori. A source of information must be available that permits classification of each element of the population into a stratum (e.g., a previous Census).

Muttlak (2003b) suggested Quartile Ranked Set Sampling (*1.4.8*) to estimate the population mean and showed that QRSS reduces the errors in ranking when compared to RSS. The quartile ranked ret sampling method is carried out by selecting n random samples each of size n units from the population of interest and ranking the units in

each sample with respect to a variable of interest. If the sample size n is even, select for measurement from the first $n/2$ samples the $q_1(n+1)$th smallest ranked unit and from the second $n/2$ samples the $q_3(n+1)$th smallest ranked unit. If the sample size n is odd, select for measurement from the first $(n-1)/2$ samples the $q_1(n+1)$th smallest ranked unit, from the last $(n-1)/2$ samples the $q_3(n+1)$th smallest ranked unit and from the remaining sample the median ranked unit. The cycle can be repeated r times if needed to get a sample of size nr units. Note that we always take the nearest integer of $q_1(n+1)$th and $q_3(n+1)$th where $q_1=0.25$ and $q_3=0.75$. If the quartile ranked set sampling method is used to select the sample units from each stratum then the whole procedure is called a stratified quartile ranked set sampling (SQRSS), (Syam and Ibrahim (2012)). The following is an example of SQRSS method for even sample size.

Suppose that we have two strata, i.e. $L=2$ and $h=1,2$. Let $X_{ih(q_1(n_h+1))}$ and $X_{ih(q_3(n_h+1))}$ be the $(q_1(n_h+1))$th and $(q_3(n_h+1))$th order statistics, respectively, of the ith sample in the hth stratum. Assume that from the first stratum we select a sample of size 6 and from the second stratum we want a sample of size 8. Then the process as shown below:

Stratum 1: Select 6 samples each of size 6 as follows:

$$X_{11(1)}, X_{11(2)}, X_{11(3)}, X_{11(4)}, X_{11(5)}, X_{11(6)}$$

$$X_{21(1)}, X_{21(2)}, X_{21(3)}, X_{21(4)}, X_{21(5)}, X_{21(6)}$$

$$X_{31(1)}, X_{31(2)}, X_{31(3)}, X_{31(4)}, X_{31(5)}, X_{31(6)}$$

$$X_{41(1)}, X_{41(2)}, X_{41(3)}, X_{41(4)}, X_{41(5)}, X_{41(6)}$$

$$X_{51(1)}, X_{51(2)}, X_{51(3)}, X_{51(4)}, X_{51(5)}, X_{51(6)}$$

$$X_{61(1)}, X_{61(2)}, X_{61(3)}, X_{61(4)}, X_{61(5)}, X_{61(6)}$$

For h=1, select the second order statistics, $X_{i1(q_1(n_1+1))} = X_{i1(2)}$ for i=1,2,3, and the 5th order statistics $X_{i1(q_3(n_h+1))} = X_{i1(5)}$ for i =4,5,6.

Thus, from the first stratum we have: $X_{11(2)}, X_{21(2)}, X_{31(2)}, X_{41(5)}, X_{51(5)}, X_{61(5)}$

Stratum 2: In the second stratum select 8 samples each of size 8 as follows:

$X_{12(1)}, X_{12(2)}, X_{12(3)}, X_{12(4)}, X_{12(5)}, X_{12(6)}, X_{12(7)}, X_{12(8)}$

$X_{22(1)}, X_{22(2)}, X_{22(3)}, X_{22(4)}, X_{22(5)}, X_{22(6)}, X_{22(7)}, X_{22(8)}$

$X_{32(1)}, X_{32(2)}, X_{32(3)}, X_{32(4)}, X_{32(5)}, X_{32(6)}, X_{32(7)}, X_{32(8)}$

$X_{42(1)}, X_{42(2)}, X_{42(3)}, X_{42(4)}, X_{42(5)}, X_{42(6)}, X_{42(7)}, X_{42(8)}$

$X_{52(1)}, X_{52(2)}, X_{52(3)}, X_{52(4)}, X_{52(5)}, X_{52(6)}, X_{52(7)}, X_{52(8)}$

$X_{62(1)}, X_{62(2)}, X_{62(3)}, X_{62(4)}, X_{62(5)}, X_{62(6)}, X_{62(7)}, X_{62(8)}$

$X_{72(1)}, X_{72(2)}, X_{72(3)}, X_{72(4)}, X_{72(5)}, X_{72(6)}, X_{72(7)}, X_{72(8)}$

$X_{82(1)}, X_{82(2)}, X_{82(3)}, X_{82(4)}, X_{82(5)}, X_{82(6)}, X_{82(7)}, X_{82(8)}$

For h=2, select $X_{i2(q_1(n_2+1))} = X_{i2(2)}$ for i=1,2,3,4 and $X_{i2(q_3(n_2+1))} = X_{i2(7)}$ for i =5,6,7,8.

Then we have $X_{12(2)}, X_{22(2)}, X_{32(2)}, X_{42(2)}, X_{52(7)}, X_{62(7)}, X_{72(7)}, X_{82(7)}$.

Therefore, the SQRSS units are $X_{11(2)}, X_{21(2)}, X_{31(2)}, X_{41(5)}, X_{51(5)}, X_{61(5)}, X_{12(2)}, X_{22(2)}, X_{32(2)}, X_{42(2)}, X_{52(7)}, X_{62(7)}, X_{72(7)}, X_{82(7)}$.

The mean of these units is used as an estimator of the population mean.

5.2 Estimation of Population Mean

Let $X_1, X_2, \ldots, X_n$ be n independent random variables from a probability density function $f(x)$, with mean μ and variance σ^2. The SRS estimator of the population mean based on a sample of size n is given by

$$\bar{X}_{SRS} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

With the variance

$$V(\bar{X}_{SRS}) = \frac{\sigma^2}{n}$$

The RSS estimator of population mean is given by

$$\bar{X}_{RSS} = \frac{1}{n}\sum_{i=1}^{n} X_{i(i)}$$

And

$$V(\bar{X}_{RSS}) = \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_{i=1}^{n}(\mu_{(i)} - \mu)^2$$

Where $\mu_{(i)}$ is the mean of the *i*th order statistic $X_{(i)}$ for a sample of size *n*.

The stratified quartile ranked set sampling estimator of the population mean when n_h is even, is defined as

$$\bar{X}_{SQRSS1} = \sum_{h=1}^{L}\frac{W_h}{n_h}\left[\sum_{i=1}^{\frac{n_h}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+2}{2}}^{n_h} X_{ih(q_3(n_h+1))}\right]$$

Where $W_h = \frac{N_h}{N}$, N_h is the stratum size and N is the total population size. The variance of SQRSS1 is given by

$$Var(\bar{X}_{SQRSS1}) = Var\left[\sum_{h=1}^{L}\frac{W_h}{n_h}\left[\sum_{i=1}^{\frac{n_h}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+2}{2}}^{n_h} X_{ih(q_3(n_h+1))}\right]\right]$$

$$= \sum_{h=1}^{L}\frac{{W_h}^2}{{n_h}^2}\left[\sum_{i=1}^{\frac{n_h}{2}} Var(X_{ih(q_1(n_h+1))}) + \sum_{i=\frac{n_h+2}{2}}^{n_h} Var(X_{ih(q_3(n_h+1))})\right]$$

$$= \sum_{h=1}^{L}\frac{{W_h}^2}{{n_h}^2}\left[\sum_{i=1}^{\frac{n_h}{2}} \sigma^2_{ih(q_1)} + \sum_{i=\frac{n_h+2}{2}}^{n_h} \sigma^2_{ih(q_3)}\right]$$

When the sample size n_h is odd, the SQRSS estimator is defined as

$$\bar{X}_{SQRSS2} = \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h-1}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+3}{2}}^{n_h} X_{ih(q_3(n_h+1))} + X_{\frac{n_h+1}{2}h\left(\frac{n_h+1}{2}\right)} \right]$$

With variance

$$Var(\bar{X}_{SQRSS2}) = Var\left[\sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h-1}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+3}{2}}^{n_h} X_{ih(q_3(n_h+1))} + X_{\frac{n_h+1}{2}h\left(\frac{n_h+1}{2}\right)} \right] \right]$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{{n_h}^2} \left[\sum_{i=1}^{\frac{n_h-1}{2}} Var(X_{ih(q_1(n_h+1))}) + \sum_{i=\frac{n_h+3}{2}}^{n_h} Var(X_{ih(q_3(n_h+1))}) + Var(X_{\frac{n_h+1}{2}h\left(\frac{n_h+1}{2}\right)}) \right]$$

$$= \sum_{h=1}^{L} \frac{W_h^2}{{n_h}^2} \left[\sum_{i=1}^{\frac{n_h-1}{2}} \sigma^2_{ih(q_1)} + \sum_{i=\frac{n_h+3}{2}}^{n_h} \sigma^2_{ih(q_3)}) + \sigma^2_{h\left(\frac{n_h+1}{2}\right)} \right]$$

Lemma: $\bar{X}_{SQRSS1}$ and $\bar{X}_{SQRSS2}$ are unbiased estimators of the mean of symmetric distributions

Proof:

If n_h is even, we have

$$E\left(\bar{X}_{SQRSS1}\right) = E\left[\sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+2}{2}}^{n_h} X_{ih(q_3(n_h+1))} \right] \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h}{2}} E(X_{ih(q_1(n_h+1))}) + \sum_{i=\frac{n_h+2}{2}}^{n_h} E(X_{ih(q_3(n_h+1))}) \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h}{2}} \mu_{h(q_1)} + \sum_{i=\frac{n_h+2}{2}}^{n_h} \mu_{h(q_3)} \right]$$

where $\mu_{h(q_1)}$ and $\mu_{h(q_3)}$ are the means of the order statistics corresponding to the first and third quartiles, respectively. Since the distribution is symmetric about μ, then $\mu_{h(q_1)+}\mu_{h(q_3)} = 2\mu_h$. Therefore, we have

$$E(\bar{X}_{SQRSS1}) = \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\frac{n_h}{2} \mu_{h(q_1)} + \frac{n_h}{2} \mu_{h(q_3)} \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\frac{n_h}{2} (\mu_{h(q_1)} + \mu_{h(q_3)}) \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\frac{n_h}{2} (2\mu_h) \right]$$

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

If n_h is odd, then

$$E(\bar{X}_{SQRSS2}) = E\left[\sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h-1}{2}} X_{ih(q_1(n_h+1))} + \sum_{i=\frac{n_h+3}{2}}^{n_h} X_{ih(q_3(n_h+1))} + X_{\frac{n_h+1}{2}h\left(\frac{n_h+1}{2}\right)} \right] \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h-1}{2}} E(X_{ih(q_1(n_h+1))}) + \sum_{i=\frac{n_h+3}{2}}^{n_h} E(X_{ih(q_3(n_h+1))}) + E(X_{\frac{n_h+1}{2}h\left(\frac{n_h+1}{2}\right)}) \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\sum_{i=1}^{\frac{n_h-1}{2}} \mu_{h(q_1)} + \sum_{i=\frac{n_h+3}{2}}^{n_h} \mu_{h(q_3)} + \mu_{h\left(\frac{n_h+1}{2}\right)} \right]$$

where $\mu_{h(q_1)}$ is the mean of the first quartile for the first $(\frac{n_h-1}{2})$samples in the *h*th stratum, $\mu_{h(q_3)}$ is the mean of the third quartile for the last $\left(\frac{n_h-1}{2}\right)$samples in the hth stratum, and μ_h is the mean for the stratum *h*. Since the distribution is symmetric about μ, then $\mu_{h(q_1)} + \mu_{h(q_3)} = 2\mu_h$. Therefore,

$$E\left(\bar{X}_{SQRSS2}\right) = \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\left(\frac{n_h-1}{2}\right) \mu_{h(q_1)} + \left(\frac{n_h-1}{2}\right) \mu_{h(q_3)} + \mu_{h\left(\frac{n_h+1}{2}\right)} \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\left(\frac{n_h-1}{2}\right) \left(\mu_{h(q_1)} + \mu_{h(q_3)}\right) + \mu_{h\left(\frac{n_h+1}{2}\right)} \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} \left[\left(\frac{n_h-1}{2} (2\mu_h) + \mu_h\right) \right]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} [(n_h-1)\mu_h + \mu_h]$$

$$= \sum_{h=1}^{L} \frac{W_h}{n_h} (n_h \mu_h)$$

$$= \sum_{h=1}^{L} W_h \mu_h = \mu$$

5.3 Simulation Study

A simulation study is conducted to investigate the performance of SQRSS in estimating the population mean. Symmetric and asymmetric distributions are considered for *n* = 7, 12, 14, 15, 18 by assuming that the population is partitioned into two or three strata. Using 100000 replications, estimates of the means, variances and mean square errors are computed. For each distribution it is assumed that the

distribution of each stratum follows that distribution. When the underlying distribution is symmetric, the efficiency of SQRSS relative to SRS, SSRS, SRSS, is given by:

$$eff\left(\bar{X}_{SQRSS}, \bar{X}_{SSRS}\right) = \frac{Var(\bar{X}_{SSRS})}{Var(\bar{X}_{SQRSS})}, eff\left(\bar{X}_{SQRSS}, \bar{X}_{SRSS}\right) = \frac{Var(\bar{X}_{SRSS})}{Var(\bar{X}_{SQRSS})},$$

$$eff\left(\bar{X}_{SQRSS}, \bar{X}_{SRS}\right) = \frac{Var(\bar{X}_{SRS})}{Var(\bar{X}_{SQRSS})},$$

Respectively, and if the distribution is asymmetric the efficiency is defines as

$$eff\left(\bar{X}_{SQRSS}, \bar{X}_{SSRS}\right) = \frac{MSE(\bar{X}_{SSRS})}{MSE(\bar{X}_{SQRSS})}, eff\left(\bar{X}_{SQRSS}, \bar{X}_{SRSS}\right) = \frac{MSE(\bar{X}_{SRSS})}{MSE(\bar{X}_{SQRSS})},$$

$$eff\left(\bar{X}_{SQRSS}, \bar{X}_{SRS}\right) = \frac{MSE(\bar{X}_{SRS})}{MSE(\bar{X}_{SQRSS})}$$

where MSE is the mean square error (MSE) which is defined as

$$MSE(\bar{X}) = Var(\bar{X}) + [Bias(\bar{X})]^2$$

Based on Tables 5.1-5.7, it is conclude that:

1. A gain in efficiency is attained using SQRSS method for estimating the population mean of the variable of interest. For example, for n=18 with n_1=4, n_2=6, and n_3=8, the efficiency of SQRSS1 with respect to SRSS is 1.9037 for estimating the mean of the uniform distribution.

2. SQRSS is more efficient than SRSS, SSRS and SRS based on the same number of measured units. For example, when n=12, the efficiency value of SQRSS1 with respect to SRSS, SSRS and SRS are 3.0702, 4.4249 and 4.3212, respectively, for estimating the mean of the normal distribution.

3. The suggested estimators are more efficient when the underlying distribution is symmetric as compared to some asymmetric distributions.

4. As the number of strata increases, the bias values decreases. For example, when n=18, for three strata the bias of SQRSS is 0.0008 while the bias is 0.0048 for two strata for estimating the mean of B (1, 2).

Table 5.1: The efficiency of SQRSS1 relative to SRSS, SSRS and SRS for n = 14 with n_1= 8 and n_2 = 6.

	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRSS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SSRS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRS})$
Uniform (0,1)	1.2951	1.1961	1.1765
Normal (0,1)	1.5421	1.7392	1.7081
Student T (3)	2.5941	3.0907	2.9233
Geometric (0.5)	2.3171	1.8348	1.8045
Exponential (1)	1.4827	2.9393	2.8866
Gamma (1,2)	2.8220	2.8187	2.7522
Beta (1,2)	2.0800	1.6000	1.4815
Beta (5,2)	1.5714	1.4615	1.3846
LogNormal(0,l)	2.4629	2.8177	2.7685
Weibull (1,2)	2.4512	2.4762	2.4286

Table 5.2: The efficiency of SQRSS2 relative to SRSS,SSRS and SRS for n=7 with n_1 = 4 and n_2 = 3

	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SRSS})$	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SSRS})$	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SRS})$
Uniform (0,1)	1.4044	1.9680	1.9520
Normal (0,1)	2.2923	1.3206	1.2979
Student T (3)	3.2733	4.1918	4.0163
Geometric (0.5)	3.1237	3.0990	3.0437
Exponential (1)	4.6853	4.5361	4.4577
Gamma (1,2)	4.5464	4.9583	4.8654
Beta (1,2)	2.6986	1.1096	1.0959
Beta (5,2)	1.2593	1.3704	1.3704
LogNormal(0,l)	1.0519	4.1557	4.0867
Weibull (1,2)	1.5090	1.2724	1.2480

Table 5.3: The efficiency of SQRSS1 relative to SRSS, SSRS and SRS for n = 12 with n_1 = 5 and n_2 = 7

	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRSS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SSRS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRS})$
Uniform (0,1)	2.0526	1.9726	1.9452
Normal (0,1)	3.0702	4.4249	4.3212
Student T (3)	3.7740	2.3829	2.3589
Geometric (0.5)	5.9230	5.5405	5.3883
Exponential (1)	5.1000	7.2101	6.9916
Gamma (1,2)	5.9486	5.5614	5.4599
Beta (1,2)	1.5625	1.4688	1.4375
Beta (5,2)	2.0000	1.6923	1.6154
LogNormal(0,l)	6.2195	8.3230	8.1325
Weibull (1,2)	1.8829	2.0674	2.0112

Table 5.4: The efficiency of SQRSS1 relative to SRSS, SSRS and SRS for n = 18 with n_1 = 4, n_2 = 6 and n_3 = 8

	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRSS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SSRS})$	$eff(\bar{X}_{SQRSS1}, \bar{X}_{SRS})$
Uniform (0,1)	1.9037	3.0625	2.8750
Normal (0,1)	2.4148	4.5581	4.3023
Student T (3)	3.0018	3.1649	2.9785
Geometric (0.5)	2.6504	3.0281	2.8414
Exponential (1)	4.4286	6.3913	6.0217
Gamma (1,2)	2.1230	3.4107	3.2293
Beta (1,2)	2.0000	3.4178	3.5317
Beta (5,2)	1.5346	3.8884	3.6292
LogNormal(0,l)	1.8972	3.1877	3.0023
Weibull (1,2)	1.9744	3.0625	3.0513

Table 5.5: The efficiency of SQRSS2 relative to SRSS, SSRS and SRS for n = 15 with n_1 = 3, n_2 = 5 and n_3 = 7

	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SRSS})$	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SSRS})$	$eff(\bar{X}_{SQRSS2}, \bar{X}_{SRS})$
Uniform (0,1)	1.1053	3.1579	2.9474
Normal (0,1)	1.2982	4.9792	4.6250
Student T (3)	2.5238	3.1970	3.0256
Geometric (0.5)	2.7542	6.2294	5.7835
Exponential (1)	3.3857	4.9247	4.5479
Gamma (1,2)	1.8189	3.6919	3.4249
Beta (1,2)	2.6345	5.7726	5.3396
Beta (5,2)	1.3050	3.8128	3.6009
LogNormal(0,l)	4.1032	5.4317	4.8524
Weibull (1,2)	2.8636	4.5294	4.2059

Table 5.6 : The bias values of SQRSSl for n =12, 14, 18

	n = 14	**n = 12**	**n = 18**	**n = 18**
	n_1 = 8 , n_2 =6	n_1 = 5 , n_2 = 7	n_1 =10 , n_2 = 8	n_1 = 4 , n_2 = 6, n_3 = 8
Geometric (0.5)	0.0168	0.0168	0.0086	0.0061
Exponential (1)	0.0308	0.0308	0.0297	0.0059
Gamma (1,2)	0.0599	0.0761	0.1124	0.0312
Beta (1,2)	0.0052	0.0052	0.0048	0.0008
Beta (5,2)	0.0347	0.0372	0.0253	0.0092
LogNormal(0,l)	0.0799	0.0852	0.0644	0.0158
Weibull (1,2)	0.0385	0.0458	0.0349	0.0118

Table 5.7: The bias values of SQRSS2 for n =7, 15		
	***n* = 7**	***n* = 18**
	$n_1 = 4$, $n_2 = 3$	$n_1 = 3$, $n_2 = 5$, $n_3 = 7$
Geometric (0.5)	0.0170	0.0085
Exponential (1)	0.0309	0.0062
Gamma (1,2)	0.0899	0.0369
Beta (1,2)	0.0054	0.0009
Beta (5,2)	0.0705	0.0094
LogNormal(0,l)	0.0974	0.0118
Weibull (1,2)	0.0770	0.0122

CHAPTER 6

APPLICATIONS

6.1 Introduction

The variety of RSS variants has tremendously enlarged the territory of the application of RSS from its original colony of agriculture and ecological studies to a vast and much diversified continent including the areas of clinical trials and genetic studies. This will be illustrated by the examples of application presented in the next section.

All the RSS variants share the same basic features and properties. These RSS schemes bear the similarity to stratified sampling. Samples ascertained through the RSS schemes contain more information than simple random samples of the same size, which explains why RSS is more efficient than simple random sampling as has been demonstrated by many particular statistical problems.

6.2 Case Studies

6.2.1 Forage Yields

Although McIntyre's original proposal of estimating pasture yields by "unbiased selective sampling using ranked sets" was made in 1952, no applications were apparently reported until fourteen years later. Halls and Dell in 1966 applied McIntyre's method, coining it "ranked set sampling" for estimating the weights of browse and herbage in a pine-hardwood forest of east Texas. These authors discovered RSS to be considerably more efficient than SRS.

Sets of three closely grouped quadrats were formed on a 300-acre tract. At select locations, metal frames of 3.1 square feet were placed at three randomly selected points within a circle of 13 foot radius as seen in figure 6.2.1. Quadrats were then ranked as lowest, intermediate and highest according to the perceived weight of browse and, separately, of herbage. Then, after clipping and drying, the separate weights of browse and herbage were determined for each quadrat. This was repeated for 126 sets for estimating browse and 124 sets for estimating herbage.

In order to simulate the SRS estimator for the mean weight of browse, one quadrat was randomly selected from each set without considering its rank. Since actual values were known for each quadrat, the RSS estimator was obtained by randomly choosing the ranks to be quantified for each set, resulting in 37 lowest ranks, 46 intermediate

ranks and 43 highest ranks. Halls and Dell also examined McIntyre's suggestion that unequal allocation might further improve the efficiency of estimation. Since the standard deviations for the order statistics were 7, 13 and 27.7 for the low, intermediate and high yield, respectively (ratio of 1:2:4), they selected 14 quadrats in the low group, 40 in the intermediate group and 72 in the high group. Note that perfect ranking was obtained for both RSS protocols because the actual values already known for each quadrat.

Results of these three sampling protocols are reported in Table 6.2.1. As expected under perfect ranking, precision due to RSS with approximately equal allocation increased, more than doubling for browse estimates. Furthermore, when allocation was proportional to the order statistic standard deviation, the precision increased still further, thus supporting McIntyre's contention.

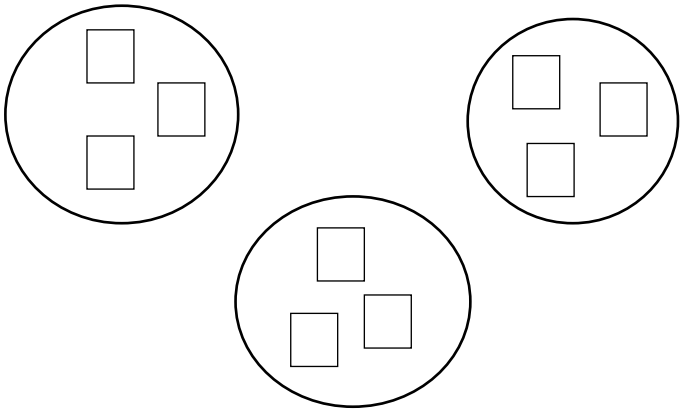


Figure 6.2.1: Within each circle, quadrats are randomly placed, followed by ranking and analysis of one appropriate quadrat. (not to scale)

Another very valuable aspect of this study was that two observers independently ranked the quadrats, one a professional range man and the other a woods worker. There was practically no difference in the ranking results between the two observers.

6.2.2 Seedling counts

The effectiveness of RSS for improving the sampling precision of seedling counts was studied by Evans (1967) in an area in central Louisiana that was seeded to Longleaf Pine (Pinuspalustris mill). After dividing the target area into 24 blocks, each block was then subdivided into 25 one-milacre plot. All 600 plots were initially measured to characterize the population, which is summarized in Table 6.2.2a. The population mean and standard deviation were calculated to be 1.675 and 1.36, respectively.

For the RSS protocol, three plots were randomly selected from each of the 24 blocks (sets), resulting in 72 identified plots. The three plots within each set were then visually ranked. One cycle consisted of selecting the lowest ranked plot from the first set. The second lowest from the second set and the highest ranked plot from the third set. Repeating the cycle eight times yielded 24 selected plots in the ranked set sample (m =3, r = 8). This whole procedure was repeated twice so that three separate field trials were performed, as summarized in Table 6.2.2b. Evans also computed the means and standard deviations of each rank using all 72 identified plots for each of the three field trials. These results are reproduced in the Table 6.2.2c for comparison to the RSS results in Table 6.2.2b.

In order to compare RSS to SRS, Evans resampled the 24 blocks (sets) 80 times to obtain two empirical distributions of the mean, one based on the RSS estimator and the other based on the SRS estimator, which is actually a stratified random sample estimator. The results of this "bootstrapping" exercise are reproduced in Table 6.2.2d where we see a significant reduction in the variance due to RSS.

6.2.3. Shrub Phytomass in Forest Stands

The performance of RSS for estimating shrub Phytomass (all vegetation between one and five meters high) was evaluated by Martin et al. (1980) at a forested site in Virginia. They investigated four major vegetation types along a decreasing moisture gradient: mixed hardwood, mixed oak, mixed oak and pine, and mixed pine. For each vegetation type, a 20m by 20m area was subjectively located which was further divided into 16 plots of equal size (5m by 5m).

For the RSS procedure, four sets of four plots were randomly selected from the 16 plots in each vegetation type. The plots in each set were then ranked by visual inspection, followed by quantifying the smallest ranked plot from the first set, the second smallest ranked plot from the second set and so on in the usual manner for RSS. This was repeated for each of the four vegetation types. For the SRS procedure, four out of the 16 plots in each vegetation type was randomly selected without replacement, followed by quantification of each selected plot. Again, this is actually a stratified random sample since each vegetation type is a separate stratum. Shrub Phytomass was also determined for all 64 plots to obtain a grand mean and variance for comparison. Their results are reproduced in Table 6.2.3 where we see a substantial increase in precision of the mean estimator associated with RSS.

6.2.4. Herbage Mass

In order to compare RSS with SRS for estimating herbage mass in pure grass swards and both herbage mass and clover content in mixed grass-clover swards, Cobby et al. (1985) conducted four experiments at Hurley (UK). Besides comparison of RSS to SRS, their objective was to assess the effects of the following factors on RSS: (i) imperfect ranking within sets, (ii) greater variation between sets than within sets, and (iii) asymmetric distribution of the quantified values.

The first two experiments were conducted by randomly selecting 15 locations, followed by randomly selecting three quadrats at each location and have several observers rank the quadrats within each set. For the last two experiments, 45 quadrats were drawn at random from the entire target area. This allowed an assessment of the effects of both spatial variation and ranking errors within sets.

Their results are reproduced in Table 6.2.4, where RP of both the worst and best observers are compared to the RP under perfect ranking, and the between and within set variances are presented for assessing spatial variation. These authors determined the main adverse factor to be within set clustering, and they recommend spacing quadrats within sets as far apart as possible when local spatial autocorrelation exists. With this in mind, they recommend RSS over SRS for sampling grass and grass-clover swards.

6.2.5. PCB Contamination Levels

Before being lead to believe that RSS is only for vegetation studies, let us consider estimating PCB concentrations in soil. Patil et al. (1994a) used measurements of this contaminant collected at a Pennsylvania site along the gas pipeline of the Texas Eastern Company. Table 6.2.5.1 provides the summary statistics of PCB values in two sampling grids (A and C) within this site. Since the distribution of these data was highly skewed, they examined the effects of unequal as well as equal allocation of samples. More specifically, they examined the following schemes:

a) Equal allocation of samples using all possible choices of sample units of each set size,
b) Equal allocation of samples for a particular sample, and
c) Unequal allocation of samples.

Considering set sizes 2, 3, and 4, the relative savings (RS) were computed as $\left[\frac{VAR(SRS)-Var(RSS)}{Var(SRS)}\right]$ taking into consideration all possible choices of sample units for each set size for both the grids under the equal allocation scheme. The results are given in Table 6.2.5.2, where it is evident that RS increases with set size but that the magnitude of RS is higher for grid C than for grid A. Note that the data for grid C is much less skewed than grid A, as seen in Table 6.2.5.1.

For comparing the performance of the RSS protocol relative to that of SRS with unequal allocation of samples, these authors considered two different proportional allocations for each set-size in order to decide the sample size for each rank. This has been done to show the impact of proportional allocation on the magnitude of relative savings accrued due to RSS over SRS. The results are given in Table 6.2.5.3, where the magnitudes of relative savings are seen to be quite substantial for each set size for both the grids.

While unequal allocation of samples into ranks can substantially increase RS when the underlying population follows a skewed distribution, this procedure does require some prior knowledge of the underlying distribution. For this purpose one may either take advantage of prior surveys of similar nature or conduct a pilot study.

This same problem also arises in determining the optimum sample size under Neyman's allocation scheme for stratified random sampling. Recent work by Kaur et al. (1994) has addressed the issue of optimum allocation when some knowledge about the underlying distribution is available, and they have devised a rule-of-thumb for allocating sample units based on skewness.

6.2.6 Application in population genetics.

In the second stage of RSS, if only the units with the smallest rank or the largest rank are chosen for full measurement, the RSS scheme is referred to as the extreme RSS. The extreme RSS has recently found important applications in genetics for quantitative trait loci (QTL) mapping.

A QTL is a gene which affects a quantitative trait of concern such as obesity, cholesterol level, etc. Suppose that a candidate QTL has two alleles, say Q and q, which form three possible genotypes QQ, Qq and qq. Let Q be the allele which causes larger values of the quantitative trait, if the candidate QTL is indeed a QTL. It is usually the case that the frequency of the Q allele is small. As a consequence of this fact, even a large random sample from the population will include only a few of individuals whose genotype at the QTL contains the Q allele. This makes the usual t-test, which compares the mean trait values between different genotypes of the QTL, infeasible. One approach adopted for detecting QTL using population data is to truncate the population at a certain quantile of the distribution of Y and take a random sample from the truncated portion and a random sample from the whole population. Then the two samples are genotyped and compared on the number of Q-alleles. If a significant difference exists, the candidate QTL is claimed as a true QTL, see Slatkin (1999) and Xu et al. (1999). In the implementation of the truncation approach, a large number of individuals have to be screened before a sample can be taken from the truncated portion. This causes tremendous practical difficulties, which hinders the application of the truncation approach in most of practical situations.

The extreme RSS provides an alternative to the truncation approach. In the extreme RSS, individuals are taken in sets. The individuals within each set are ranked according to their trait values, and the one with the largest trait value is put into an upper sample and the one with the smallest trait value is put into a lower sample. The two samples obtained this way are then genotyped and compared. This extreme RSS

approach has been applied for linkage disequilibrium mapping of QTL recently by Chen et al. (2005). It turns out that the extreme RSS approach can achieve comparable powers to that of the truncation approach but avoids all practical difficulties of the truncation approach. The extreme RSS has also been applied to a sib-pair regression model where extremely concordant and/or discordant sib-pairs are selected by the extreme RSS; see Zheng et al. (2006). The extreme RSS approach can be applied to many other genetic problems such as the TDT test (Spielman et al. 1993) and the gamete competition model (Sinsheimer et al. 2000), etc. The properties of the extreme RSS in those problems are yet to be investigated.

6.2.7 Application in regression analysis.

A new application of RSS discussed in this section concerns with the following linear regression model:

$$Y = \beta_0 + \beta_1 X + \cdots + \beta_P X^P + \epsilon$$

Here it is assumed that the response variable *Y* is expensive to measure but the predictor variable *X* can be measured cheaply. The RSS can be carried out by ranking with respect to the predictor variable. The interest is now on the estimation of the regression coefficients. By considering optimality criteria such as *D*-optimality or *A*-optimality based on the asymptotic variances and covariances of the estimated regression coefficients obtained from RSS samples, optimal unbalanced RSS schemes can be obtained to improve the efficiency of the estimation of the regression coefficients.

An application of the above approach to a lung cancer study is given in Chen and Wang (2004). In that study, the effect of smoking on lung cancer is investigated through three bio-markers: the polyphenol DNA adducts in blood mononuclear cells, the micro nuclei (MI)which are chromosomal fragments or whole chromosomes excluded from the nucleus at mitosis, and the sister chromatid exchanges (SCE) which involve the reciprocal exchange of genetic material during cell replication. The purpose of the study is to determine the relationship between the three bio-markers and smoking level through three separate quadratic regression models. In this problem, the measurement of the three markers is very expensive but the smoking level of people can be easily obtained and is available from a large number of people.

By using an *A*-optimal RSS scheme with set size 10, the relative efficiencies, measured by the ratios of the sums of variances of the estimated regression coefficients, are at least 2.

Another application of the optimal regression approach to a fishery study is given in Chen et al. (2004, Chapter 6). In the fishery study, one is interested in describing the growth of a special fish species *Tenualosa ilisha* in Bangladesh through a regression relationship between the age of a fish and its weight. Determining the age of a fish is an extremely time-consuming process. First, one of its otoliths is removed, cleaned and sent in a plastic bag to a lab. Then the otolith is embedded onto a microscope slide with thermoplastic cement and polished with wet and dry sandpaper until its mid-plane is reached. Finally, the polished otolith is viewed under immersion oil on a video screen attached to a microscope and the daily rings are counted along the longitudinal axis towards the posterior of the otolith. On the other hand, the weight of a fish can be easily obtained without any cost. It was demonstrated that, by using an optimal RSS scheme with set size 10, a relative efficiency 1.4 compared with simple random sampling in terms of the integrated mean square error of the regression function can be achieved.

In practical regression problems, the situation where the response variable is expensive to measure but the predictor variable can be easily and cheaply measured is abundant. The approach developed in Chen and Wang (2004) has a great potential in applications.

6.2.7 Application in treatment comparisons.

Another novel application of RSS is in treatment comparison experiments including many clinical trials. In RSS many more sampling units are sampled and discarded than those eventually fully measured. This might not be desirable in the situation where sampling units are not easy to obtain, which is especially the case in clinical trials. Ozturk and MacEachern (2004) and Chen et al. (2006a) separately considered an RSS approach which generates ranked set samples for each treatment but without discarding any sampling units. Chen (2007) elaborated on this approach in the following setting of Chen et al. (2006a). Assume that the responses of the experimental units to treatments are correlated with a common concomitant variable.

Let Y and Z be the responses to treatments 1 and 2 respectively and X the concomitant variable. The assumption formulates that

$$Y_{1i} = \alpha_1 + \beta_1 X_{1i} + \epsilon_{1i},$$

$$Y_{2i} = \alpha_2 + \beta_2 X_{2i} + \epsilon_{2i},$$

$$i = 1, \dots, n$$

where ϵ_{li} s are i.i.d. with mean zero and variance $\sigma_l^2, l = 1,2,$ and are independent from the X_{li}s. Let the set size k in RSS be even. A special case of the RSS schemes considered by Chen et al. (2006a) is as follows. The RSS is carried out two sets at a time. That is, each time two random sets of experimental units are taken and ranked separately according to the values of X. For the first ranked set, units with odd ranks are assigned to treatment 1 and units with even ranks are assigned to treatment 2. For the second ranked set, units with odd ranks are assigned to treatment 2 and units with even ranks are assigned to treatment 1. This process produces two correlated general RSS samples, each for each treatment. It does not discard any experimental units. It is shown in Chen et al. (2006a) that this method of treatment assignment is much more efficient than a simple random assignment.

Chen (2007) applied the above method to a retrospective study of a well known clinical trial called ACTG 320. The ACTG 320 clinical was a randomized double-blind multi center clinical trial comparing the effects of the three-drug combination of IDV+ZDV+3TC and the two-drug combination of ZDV+3TC on an AIDS-defining event. The background and more details on ACTG 320 can be found in Hammer et al. (1997) and Marschner et al. (1999).The effect of the drug combinations on a patient was measured by the HIV-1 RNA changes from the baseline HIV-1 RNA level of the patient. In the retrospective study, the response variable is taken as the measured HIV-1 RNA change at week 24. In the original study, a total of 1,080 patients were initially involved but only 639 patients remained on their initial treatments at week 24. The data of these 639 patients is used in the retrospective study. The data shows that the change in HIV-1 RNA level at week 24 is correlated with the pre-entry HIV-1 RNA level in both treatments. We take Y_1 and Y_2 as the RNA changes in $\log_{10}$ scale with the treatment of two-drug combination and

with the treatment of three-drug combination respectively, and take the concomitant variable *X* as the pre-entry HIV-1 RNA level.

In the retrospective application, The RSS protocol with $k = 4$ is applied. The details of the protocol are as follows. The patients are considered in groups of size 4. For each group, the pre-entry HIV-1 RNA levels of the four patients are ranked. For one group, the two patients with their pre-entry HIV-1 RNA levels ranked 1 and 3 are assigned to the treatment of two-drug combination, and the other two patients are assigned to the treatment of three-drug combination. For another group, the two patients with their pre-entry HIV-1 RNA levels ranked 2 and 4 are assigned to the treatment of two-drug combination, and the other two patients are assigned to the treatment of three-drug combination. This protocol does not incur any additional cost other than those needed by the simple random assignment.

From the original data, the following parameter values are computed:

$$\mu_{Y1} = 0.4195, \quad \sigma_{Y1}^2 = 0.0282, \ \mu_{Y2} = 2.4658, \sigma_{Y2}^2 = 0.1320,$$

$$\mu_X = 4.9448, \ \sigma_X^2 = 0.3685, \ \rho_{XY1} = 0.61, \ \rho_{XY2} = 0.43.$$

From these values, we obtain

$$\alpha_1 = -0.4147, \beta_1 = 0.1687, \sigma_1^2 = 0.0177,$$

$$\alpha_2 = 1.1930, \beta_2 = 0.2574, \sigma_2^2 = 0.1076.$$

These values are taken as if they are the true parameter values in the retrospective study. The pre-entry HIV-1 RNA level is assumed to be normally distributed. The relative efficiency of the general RSS protocol relative to the simple random assignment is computed theoretically. It is also simulated by a simulation study with 5,000 repetitions. The theoretical value of the relative efficiency is 1.21. The simulated approximation is 1.22 which is quite in line with the theoretical value. This relative efficiency implies that the precision achievable by including 639 patients in the trial with the RSS assignment could only be achieved by a simple random assignment with 781 patients.

The theoretical value of the relative efficiency is given by $\sigma_{SRS}^2 / \sigma_{4,4,1}^2$, where

$$\sigma_{SRS}^2 = \sigma_1^2 + \sigma_2^2 + (\beta_1^2 + \beta_2^2)\sigma_X^2,$$

$$\sigma_{4,4,1}^2 = \sigma_{SRS}^2 - \frac{(\beta_1+\beta_2)^2\sigma_X^2}{8}\sum_{r=1}^{4}(3\sigma_{(r,\overline{r+1}:4)} + \sigma_{(r,\overline{r+3}:4)}$$
$$-\frac{(\beta_1^2+\beta_2^2)\sigma_X^2}{8}\left[4\sum_{r<s}\sigma_{(r,s:4)} - \sum_{r=1}^{4}(3\sigma_{(r,\overline{r+1}:4)} + 2\sigma_{(r,\overline{r+2}:4)} + \sigma_{(r,\overline{r+3}:4)})\right]$$

Here, $\sigma_{(r,s:4)}$denotes the covariance of the rth and the sth order statistics of a simple random sample of size 4 from the standard normal distribution. The numerical values of $\sigma_{(r,s:4)}$ can be found in Krishnaiah and Sen (1984). The meaning of $\bar{s}$ is that $\bar{s} = s$, if $s \leq 4, \bar{s} = s - 4$, otherwise. It should be noted that the relative efficiency is affected by the correlation of the concomitant variable X with the response variables in the two treatments through $\beta_1$and β_2. In fact, $\beta_l = \rho_{XYl}(\sigma_{Yl}/\sigma_X), l = 1,2$. It is clear from the expression of $\sigma_{4,4,1}^2$that both the magnitude and the signs of ρ_{XY1} and ρ_{XY2} affect the relative efficiency. The relative efficiency is larger when the two correlation coefficients have the same sign than when they have opposite signs.

6.3 Discussion

The variety of the variants of RSS developed has broadened the range of application to a large extent than its earlier forms. RSS is still an active area of research. Below are discussed some further directions for the research of RSS.

(i) *The cost issue of RSS*. Without taking into account the cost involved in taking sampling units and ranking, which is assumed negligible, the larger the set size, the more efficient the RSS is compared to simple random sampling. However, in many practical problems where RSS has a potential application, the cost of taking sampling units and ranking, though much less than the full measurement, is not negligible, or the availability of sampling units is limited. In such cases, the cost issue arises. One needs to devise a sampling scheme such that the scheme is as efficient as possible, say, in terms of the accuracy of estimation for certain parameters, subject to a fixed cost, or such that the scheme is as less costly as possible subject to a required accuracy of estimation. There are multiple questions to be asked. Is RSS still more beneficial than SRS by a proper choice of the set size? If RSS is still beneficial, what is the optimal set size when the costs of taking sampling units, ranking and making the full measurement are given? If the original RSS which takes only one full measurement in a ranked set is not beneficial, is a general RSS scheme which takes

more than one full measurement in a ranked set beneficial? If yes, how many and what ranks? Nahhas et al. (2002) and Wang et al. (2004) addressed some aspects of the cost issue. But more research is needed on this issue.

(ii) *Design with observational data.* RSS can be used as a tool for design with observational data. The special case of polynomial regression has been addressed by Chen and Wang (2004). The more general case with multiple covariates is yet to be investigated.

(iii) *RSS as data reduction tools.* In the context of data reduction, one is faced with the problems caused by huge data sets. A data set could be so huge that it is even infeasible to compute the quantiles of the data set by the modern computers. In data reduction, one tries to discard the part of the data with less information, or equivalently retain the part of the data with more information. The part of the data retained can be viewed as a sample from the original data. RSS can play a role here. More research is needed in this regard especially when data involves many variables.

Table 6.2.1 Summary statistics for browse and herbage estimates

	Browse		**Herbage**	
	Mean	Variance of mean	Mean	Variance of mean
Unranked: random perfect ranking	14.9	4.55	7.3	1.00
Perfect ranking: Near equal allocation	13.2	2.18	7.0	0.73
Perfect ranking : Proportional allocation	12.9	1.91	7.2	0.58
(Source: Halls and Dell. 1966)				

Table 6.2.2: Data from Longleaf Pine Seedling Counts

(a) The frequency distribution of seedling counts in the 600 milacre plots.

Seedling counts	0	1	2	3	4	5	6	7	8	9
Frequency	110	201	157	75	33	17	3	3	0	1

(b) Means and variances of three ranked set sample trails. (*mr* = 24)

Trail	**Mean**	**Variance**
1	1.49	0.043
2	1.62	0.056
3	1.71	0.024

(c) Means and standard deviations of all seedlings for all ranks of three field trails and ranked set sampling

Trial	**Means**				**Mean Standard Deviations**		
	L	**M**	**H**		**L**	**M**	**H**
1	0.750	1.500	2.625	1.625	0.532	0.750	1.173
2	0.917	1.625	2.833	1.792	0.881	1.013	1.880
3	0.750	1.708	3.125	1.861	0.520	0.955	0.927

(d) Test of significance of ranked-set versus random sampling.

Method of sampling	**Number applications**	**Degree of freedom**	**Mean**	**Sum of squares**	**Variance**	**F**
Random	80	79	1.709	7.572	0.0958	3.91**
Ranked-set	80	79	1.647	1.939	0.0245	

****** Significant at the 0.01 level of probability (Source: Evans, 1967)

Table 6.2.3: RSS and SRS results for 16 measured plots across all vegetation types**.**

Sampling Method	Mean Phytomass (kg/ha)	Variance of the Mean (X 10^6)	Coefficient of Variation of the Mean (%)
All 64 Plots	2536	0.15	15
SRS	1976	4.54	108
RSS	2356	2.73	70
(Source: Martin et al. 1980)			

Table 6.2.4: Relative precisions (RP) ± s.e. of the worst and the best observers, and under perfect ranking; and the between and the within set variances while estimating herbage mass (grass and mixture) and clover contents.

Experiments	Relative Precisions (R P)			Variances	
	Worst	Best	Perfect	Between	Within
1 (Grass)	1.11 ± 0.09	1.23 ± 0.14	1.31 ± 0.17	0.24	0.31
2 (Mixture)	1.11 ± 0.09	1.27 ± 0.10	1.40 ± 0.16	0.07	0.09
3 (Grass)			1.66 ± 0.17	0.00	1.58
4 (Mixture)	1.36 ± 0.14	1.51 ± 0.15	1.55 ± 0.16	0.11	0.66
2 (Clover)	1.15 ± 0.12	1.34 ± 0.15	1.44 ± 0.16	16.3	34.4
4 (Clover)	1.36 ± 0.19	1.62 ± 0.18	1.72 ± 0.20	16.2	71.6
(Source: Cobby et al. 1985)					

Table 6.2.5.1: Descriptive statistics of PCB values in grids A and C

Characteristics	**Grid**	
	A	**C**
Number of observations	184	68
Mean	200.9	600.2
Standard Deviation	902.9	1585
Coefficient of Variation	4.49	2.64
Coefficient of skewness	9.27	4.64
Coefficient of Kurtosis	99.69	20.88

Table 6.2.5.2: Relative savings (RS) considering all possible combinations of each set size under perfect ranking situation with equal allocation**.**

Set size (m)	**Grid**	
	A **RS**	**C** **RS**
2	4	9
3	7	16
4	10	22

Table 6.2.5.3: Values of the sample mean, $X_{(m)u}$, relative precision, and relative savings under the perfect ranking protocol with unequal allocation of samples.

Set Size ***m***	**Grid**							
	A				**C**			
	Proportion of samples (exact No)	$\bar{X}_{(m)u}$	RP	RS	Proportion of samples (exact No)	$\bar{X}_{(m)u}$	RP	RS
2	1:10 (8 , 84)	205.9	1.724	42	1:10 (3 , 31)	535.2	2.041	51
2	1:15 (6 , 86)	203.1	1.818	45	1:15 (2 , 32)	520.4	2.174	54
3	1:4:20 (2 , 10 ,48)	203.6	2.174	54	1:1.7:1.5 (5 , 8 , 8)	560.1	1.471	32
3	1:4:25 (2 , 8 ,50)	201.1	2.326	57	1:2:7 (2 , 4 , 15)	615.2	1.923	48
4	1:3:5:16 (2,5,9,28)	247.1	1.695	41	1:2:3:4 (2, 3, 5, 6)	576.6	2.083	52
4	1:3:9:27 (2,2,10,30)	226.1	1.316	24	1:1:3:5 (2, 2, 4, 8)	802.4	1.449	31

Al-Hadhrami, S. (2001). Parametric estimation using moving extreme ranked set sampling. Master thesis, Sultan Qaboos University, Oman.

Al-Nasser, D. A. (2007). L Ranked set sampling: A generalization procedure for robust visual sampling. *Communications in Statistics*: *Simulation and Computation*, **36**(1), 33 – 43.

Al-Nasser, D. A., & Radaideh, A. (2008). Estimation of Simple Linear Regression Model Using L Ranked Set Sampling. *Int. J. Open Problems Compt. Math.*, **1**, (1), 18-33.

Al-Odat, M. T., & Al-Saleh, M. F. (2001). A variation of ranked set sampling. *Journal of Applied Statistical Science*, **10**, pp. 137-246.

Al-Omari, A. I., & Jaber, K. (2008). Percentile double ranked set sampling. *Journal of Mathematics and Statistics*, **4**(1), 60-64.

Al-Omari, A. I., Ibrahim, K., & Syam, M. I. (2011). Investigating the use of Stratified Percentile Ranked Set Sampling method for estimating the population mean. *Proyecciones Journal of Mathematics*, **30** (3), 351-368.

Al-Saleh, M. F., & Al- Ananbeh, A. M. (2007). Estimation of the means of the bivariate normal using moving extreme ranked set sampling with concomitant variable. *Statistical Papers*, **48**, pp. 179-195.

Al-Saleh, M. F., & Al- Kadiri, M. (2000). Double ranked set sampling. *Statist. Probab. Lett.*, **48**, pp. 205-212.

Al-Saleh, M. F., & Al- Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Interference*, **102**, pp. 273-286.

Al-Saleh, M. F., & Zheng, G. (2002). Estimation of bivariate characteristics using ranked set sampling. *The Australian and New Zealand Journal of Statistics*, **44**, pp. 221- 232.

Amarjot, Patil, G. P., Sharik, S. J., & Taillie, C. (1996). Environmental sampling with a concomitant variable: A comparison between ranked set sampling and stratified simple random sampling. *Journal of Applied Statistics*, **23**, pp. 231-255.

Arnold, B., Balakrishnan, N., & Nagaraja, H. (1992). A first course in order statistics. John Wiley and Sons, New York.

Bai, Z. D., Chen, Z. (2003). On the theory of ranked set sampling and its ramifications. *J Stat Plan Infer,* **109**, pp. 81-99.

Bhoj, D. S. (1997). New parametric ranked set sampling. *J. Appl. Statist. Sci*., **6**, pp. 275-289.

Bhoj, D. S., & Ahsanullah, M. (1996). Estimation of Parameters of the Generalized Geometric Distribution Using Ranked Set Sampling. *Biometrics,* **52**(2), 685-694.

Chen, H., Stasny, E. A., & Wolfe, D. A. (2006b). Unbalanced ranked set sampling for estimating a population proportion. *Biometrics*, **62**, pp. 150-158.

Chen, Z. (1999). Density estimation using ranked-set sampling data. *Environ EcolStat*, **6**, pp. 135-146.

Chen, Z. (2000a). The efficiency of ranked-set sampling relative to simple random sampling under Multi-parameter Families. *Stat Sin*, **10**, pp. 247-263.

Chen, Z. (2000b). On ranked-set sample quantiles and their applications. *J Stat Plan Infer.,* **83**, pp. 125-135.

Chen, Z. (2001a). The optimal ranked-set sampling scheme for inference on population quantiles. *Stat Sin* **11**, p.2337.

Chen, Z. (2001b). Ranked-set sampling with regression-type estimators. *Journal of Statistical Planning and Inference,* **92**, pp. 181-192.

Chen, Z. (2002). Adaptive ranked set sampling with multiple concomitant variables: an effective way to observational economy. *Bernoulli,* 8, pp. 313-322.

Chen, Z. (2007). Ranked set sampling: its essence and some new applications. *Environ Ecol Stat* **14**, pp. 355–363

Chen, Z., & Bai Z. D. (2000). The optimal ranked-set sampling scheme for parametric families. *Sankhya,* Ser A, **62**, pp. 178-192.

Chen, Z., & Shen, L. (2003). Two-layer ranked set sampling with concomitant variables. *J Stat Plan Infer,* **115**, pp. 45-57.

Chen, Z., & Wang, Y. (2004). Efficient Regression Analysis with Ranked-Set Sampling. *Biometrics* **60**, pp. 997–1004.

Chen, Z., Bai, Z., & Sinha, B. K. (2004) Ranked Set Sampling: Theory and Applications. Springer: New York.

Chen, Z., Liu, J., Shen, L., & Wang, Y. (2006a). General ranked set sampling for efficient treatment comparisons. *Stat Sin*

Chen, Z., Zheng, G., Ghosh, K., & Li. Z. (2005). Linkage disequilibrium mapping of quantitative trait loci by selective genotyping. *Am J Hum Genet*, **77**, pp. 661-669

Cobby, J. M., Ridout, M. S., Bassett, P. J., & Large, R. V. (1985). An investigation into the use of ranked set sampling on grass and grass-clover swards. *Grass and Forage Science,* **40**, pp. 257-263.

David, H. A., & Levine, D. N. (1972). Ranked set sampling in the presence of judgment error. *Biometrics*, **28**, pp. 553-555.

Dell, T. R., & Clutter, J. L. (1972). Ranked-set sampling theory with order statistics background. *Biometrics* **28**, pp. 545-555.

Downton, F. (1954). Least-squares estimates using ordered observations. *Annals of Mathematical Statistics,* **25**, pp. 303-316.

Draper N., & Smith, H. (1981). Applied Regression Analysis. 2nd edition. USA: John Wiley & sons, Inc..

Evans, M. J. (1967). Application of ranked set sampling to regeneration surveys in areas direct-seeded to longleaf pine. Masters Thesis, School of Forestry and Wildlife Management, Louisiana State University, Paton Rouge.

Gore, S. D., Patil, G. P., Sinha, A. K., & Taillie, C. (1993). Certain multivariate considerations in ranked set sampling and composite sampling designs. *In Multivariate Environmental Statistics*, G. P. Patil and C. R. Rao (eds). New York, New York: Elsevier Science Publishers, B.V.

Halls, L. K., & Dell, T. R. (1966). Trial of ranked set sampling for forage yields. *Forest Science* **12**, pp. 22-26.

Hammer, S. M., Squires, K. E., Hughes, M. D., Grimes, J. M, Demeter, L. M., Currier, J. S., Eron, J. J., Jr, Feinberg, J. E., Balfour, H. H., Jr, Deyton, L. R., Chodakewitz, J. A., Fischl, M. A. (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. *N Engl J Med*, **337**, pp. 725-733.

Hassan, A., Mishra, A., & Jan, T. R. (2002). A quick method for estimating generalized geometric series distribution, *Studia Scientiarum, Mathematicarium, Hungaria,* **39**, pp. 291- 295.

Hossain, S. S. & Muttlak, H. A. (1999). Paired ranked set sampling: A more efficient procedure. *Environmetrics*, **10**, pp. 195-212.

Hossain, S. S., & Muttlak , H. A. (2001). Selected ranked set sampling. *Aust. N. Z. J. Stat*., **43**, pp. 311-325.

Jemain, A. A., Al-Omari, A. I., & Ibrahim, K. (2007a). Two-stage ranked set sampling for estimating the population median. *Sains Malaysiana,* **37**(1), 95-99.

Jemain, A. A., Al-Omari, A.I. & Ibrahim, K. (2009). Balanced groups ranked set samples for estimating the population median. *Journal of Applied Statistical Science,* **17**, pp. 39-46.

Jermain, A. A. & Al-Omari A. I., (2006). Double quartile ranked set samples. *Pakistan Journal of Statistics*, **22**(3), 217-228.

Johnson, G. D., Nussbaum, B. D., Patil, G. P & Ross, P. N. (1996). Designing cost effective environmental sampling using concomitant information. *Chance*, **9**(1), 4-11.

Johnson, G. D., Patil, G. P., & Sinha, A. K. (1993). Ranked set sampling for vegetation research. *Abstracta Botanica*, **17**, pp. 87-102.

Johnson, N. L., & Kotz, S. (1969). Discrete distributions (First Edition) Bosten: Houghton Miffin.

Johnson, N. L., Kotz, S. & Kemp, A. W., (1992). Univariate discrete distributions (Second Edition) John Wiley and Sons. Inc.

Kaur, A., Patil, G. P., & Taillie, C. (1997). Unequal allocation models for ranked set sampling with skew distributions, *Biometrics*, **53**, pp. 123–130.

Kaur, A., Patil, G. P., & Tallie, C. (1994). Unequal allocation model for ranked set sampling with skew distributions. Technical report 94-0930, Centre for Statistical Ecology and Environmental Statistics, Department of Statistics, Pennsylvania State University, University Park, PA.

Kaur, A., Patil, G. P., Sinha, A.K. & Taillie, C. (1995). Ranked set sampling: An annotated bibliography. *Environmental and Ecological Statistics,* **2**, pp. 25-54.

Krishnaiah, P. R, & Sen, P. K. (1984). Tables for order statistics. In: Krishnaiah PR, Sen PK (eds), *Handbook of statistics*,. Elsevier Science Publishers **4**, pp. 892–897.

Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* **37**, pp. 1137-1153.

Lindsey, J. (1999). Multivariate elliptically contoured distributions for repeated measurements. *Biometrics* **55**, pp. 1277-1280.

Lloyd, E. H. (1952). Least-squares estimation of location and scale parameters using order statistics. *Biometrika* **39**, pp. 88-95.

MacEachern, S. N., Stasny, E. A., & Wolfe, D. A. (2004). Judgment poststratification with imprecise rankings. *Biometrics*, **60**, pp. 207-215.

MacEachern, S.N., Ozturk, O., Stark, G., Wolfe, D.A. (2002). A new ranked set sample estimator of variance. *J R StatSocSerB*, **64**, pp. 177-188

Marschner, I. C, Betensky, R. A., DeGruttola, V., Hammer, S. M., Kuritzkes, D. R. (1999). Clinical trials using HIV-1RNA-based primary endpoints:

statistical analysis and potential biases. *J Acquir Immune DeficSyndr HumRetrovirol,* **20**, pp. 220-227.

Martin, W. L., Sharik, T. L., Oderwald, R. G., & Smith, D. W. (1980). Evaluation of ranked set sampling for estimating shrub phytomass in Appalachian oak forests. Publication Number FWS-4-80, School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

McIntyre, G. A. (1952). A method of unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research,* **3**, pp. 385-390.

Mishra, A. (1982). A generalization of geometric series distribution. *J. Bihar Math. Soc.,* **6**, pp. 18-22.

Mukhopadhyay, P. (2009). Theory & Methods of Survey Sampling, Second Edition, PHI Learning Pvt. Ltd.

Muttlak, H. A. (1995). Parameters Estimation in a simple linear regression using rank set sampling. *Biometrical. J.*, **37**(7), 799–810.

Muttlak, H. A. (2003a). Modified ranked set sampling methods. *Pakistan Journal of Statistics*, **19**(3), 315-323.

Muttlak, H. A. (2003b). Investigating the use of quartile ranked set sampling for estimating the population mean. *Applied Mathematics and Computation,* **146**, pp. 437-443.

Muttlak, H.A. (1997). Median ranked set sampling. *Journal of Applied Statistical Science,* **6**, pp. 245-255.

Nahhas, R. W., Wolfe, D. A., & Chen, H. (2002). Ranked set sampling: cost and optimal set size. *Biometrics,* **58**, pp. 964-971.

Ozturk, O. (2002). Ranked set sample rank regression estimator. *J Am Stat Assoc.,* **97**, pp. 1180-1191.

Ozturk, O., & MacEachern, S. N. (2004). Order restricted randomized designs for control versus treatment comparison. *Ann Inst Stat Math,* **56**, pp. 701-720.

Ozturk, O., & Wolfe, D. A. (2000a). Optimal allocation procedure in ranked set sampling for unimodal and multi-modal distributions. *Environmental and Ecological Statistics,* **7**, pp. 343-356.

Ozturk, O., & Wolfe, D. A. (2000b). An improved ranked set two-sample Mann-Whitney-Wilcoxon test. *Can J Statist,* **28**, pp. 123-135.

Ozturk, O., & Wolfe, D. A. (2000c). Optimal allocation procedures in ranked set two-sample median test. *J Nonparamet Stat,* **13**, pp. 57-76.

Ozturk, O., & Wolfe, D. A. (2001). A new ranked set sampling protocol for the signed rank test. *J Stat Plan Infer*, **96**, pp. 351–370.

Patil, G. P., & Joshi, S. W. (1968). A dictionary and bibliography of discrete distribution. Olmir and Boyd. Edinburgh.

Patil, G. P., Gore, S. D., & Sinha, A. K. (1994c). Environmental chemistry, statistical modeling, and observational economy. *In Environmental Statistics, Assessment and Forecasting,* C. R.Cothern and N. P. Rose (eds),. Ann Arbor, Michigan 57-97 Lewis Publishers.

Patil, G. P., Sinha, A. K., & Taillie, C. (1993a). Ranked set sampling from a finite population in the presence of a trend on a site. *Journal of Applied Statistical Science,* **1**, pp. 51-65.

Patil, G. P., Sinha, A. K., & Taillie, C. (1993b). Relative precision of ranked set sampling: comparison with the regression estimator. *Environmetrics,* **4**, pp. 399-412.

Patil, G. P., Sinha, A. K., & Taillie, C. (1994a). Ranked set sampling. *In Handbook of Statistics,* G. P. Patil and C. R. Rao (eds), 167-199. New York, New York: Elsevier Science Publishers, B.V.

Patil, G. P., Sinha, A. K., & Taillie, C. (1994b). Ranked set sampling for multiple characteristics. *International Journal of Ecology and Environmental Sciences,* **20**, pp. 357-373.

Patil, G., Sinha, A., & Taillie, C. (1999). Ranked set sampling: Bibliography. *Environmental and Ecological Statistics,* **6**, pp. 91-98.

Platt , W. J., Evans, G. W., & Rathbun, S. L. (1988). The population dynamics of a long-lived conifer. *The Amer. Naturalist*., **131**, pp. 391–525.

Presnell, B., Bohn, L. L., (1999). U-statistics and imperfect ranking in ranked set sampling. *J Nonparamet Stat.*, **10**, pp. 111–126.

Prodan, M. (1968). Forest Biometrics. Program Press, London.

Samawi, H. M. & Tawalbeh, E. M. (2002). Double median ranked set sample: Comparison to other double ranked samples for mean and ratio estimators. *Journal of Modern Applied Statistical Methods*, **1**(2), 428-442.

Samawi, H. M., & Ababneh, F. (2001). On regression analysis using ranked set sample. *Journal of Statistical Research* (*JSR*), **35** (2), 93–105.

Samawi, H. M., & Muttlak, H. A. (2001). On ratio estimation using median ranked set sampling. *Journal of Applied Statistical Science*, **10**(2), 89-98.

Samawi, H. M., Ahmed, M. S., & Abu-Dayyeh, W. (1996). Estimating the population mean using extreme ranked set sampling, *Biometrical. J.*, **38** (5), 577– 586.

Singh, S. K. (1989). Some contributions to quasi binomial & discrete lagrangian prob. Distributions. PhD Theses, Patna University, Patna.

Sinsheimer, J. S., Blangero, J., Lange, K. (2000). Gamete competition models. *Am J Hum Genet*, **66**, pp. 1168–1172.

Slatkin, M. (1999). Disequilibrium mapping of a quantitative-trait locus in an expanding population. *Am J Hum Genet*, **64**, pp. 1765–1773.

Spielman, R. S., McGinnis, R.E., Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet*, **52**, pp. 506–516

Stokes, S. L. (1976). An investigation of the consequences of ranked set sampling. PhD Thesis, University of North Carolina, Chapel Hill NC.

Stokes, S. L. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics- Theory and Methods,* **6**, pp. 1207-1211.

Stokes, S. L. (1980a). Inferences on the correlation coefficient in bivariate normal populations from ranked set samples. *Journal of the American Statistical Association*, **75**, pp. 989- 995.

Stokes, S. L. (1980b). Estimation of variance using judgment ordered ranked-set samples. *Biometric,s* **36**, pp. 35-42.

Stokes, S. L. (1995). Parametric ranked set sampling. *Annals of the Institute of Statistical Mathematics*, **47**, pp. 465–482.

Syam, M. I., Ibrahim, K., & Al-Omari, A. I. (2012). The Efficiency of Stratified Quartile Ranked Set Sampling in Estimating the Population Mean, *Tamsui Oxford Journal of Information and Mathematical Sciences*, **28**(2), 175-190.

Takahasi, K., & Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, **20**, pp. 1-31.

Tripathi, R. C., & Gupta, R. C. (1987). Some generalizations of geometric series distribution, *Sankhya*, B, **49**(3), 218-223.

Wang, Y., Chen, Z., Liu, J. (2004). General ranked set sampling with cost considerations. *Biometrics*, **60**, pp. 556–561.

Wolfe, D. A. (2004). Ranked set sampling: An approach to more efficient data collection. *Statistical Science*, **19**, pp. 636-643.

Xu, X. P., Rogus, J. J., Terwedom, H. A. et al (1999). An extreme-sib-pair genome scan for genes regulating blood pressure. *Am J Hum Genet*, **64**, pp. 1694-1701.

Yang, S. (1977). General distribution theory of the concomitants of order statistics. *Annals of Statistics*, **5**, pp. 996-1002.

Yu, P. L. H., & Lam, K. (1997). Regression estimator in ranked set sampling. *Biometrics*, **53**, pp. 1070-1080.

Zhao, X., & Chen, Z. (2002). On the ranked-set sampling M-estimates for symmetric location families. *Ann Inst Stat Math*, **54**, pp. 626–640.

Zheng, G., Ghosh, K., Chen, Z., & Li, Z. (2006). Extreme rank selection for linkage analysis of quantitative traits using selected sib pairs. *Ann Hum Genet*, **70**, pp. 857–866.