On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares

M Iqbal Jeelani Bhat
(2011-380-D)

Division of Agricultural Statistics

Faculty of Postgraduate Studies

Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir

2014

On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares

M Iqbal Jeelani Bhat
(2011-380-D)

Thesis

Submitted to

The Faculty of Postgraduate Studies
Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir
in partial fulfilment of requirement for the award of the degree of

Doctor of Philosophy in Statistics

2014

Dedicated to

My Parents

With love and affectioin to
"My Nephews"

Shayan and Rehan

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Division of Agricultural Statistics, Shalimar Campus, Srinagar – 190 025
-::o::-

Certificate – I

This is to certify that the thesis entitled, **"On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares"** submitted in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy in Statistics**, to the **Faculty of Postgraduate Studies, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir** is a record of bonafide research work carried out by **Mr. M Iqbal Jeelani Bhat (Regd. No. 2011-380-D)** under my supervision and guidance. No part of the thesis has been submitted for any other degree or diploma.

It is further certified that information received during the course of investigation has duly been acknowledged.

(Dr. S.A. Mir)
Chairman
Advisory Committee

Endorsed

Head,
Division of Agricultural Statistics,
SKUAST-Kashmir, Shalimar

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Division of Agricultural Statistics, Shalimar Campus, Srinagar– 190 025

Certificate – II

We, the members of the Advisory Committee of **Mr. M Iqbal Jeelani Bhat (Regd. No. 2011-380-D)**, a candidate for the degree of **Doctor of Philosophy in Statistics**, have gone through the manuscript of the thesis entitled, **"On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares"** and recommend that it may be submitted by the student in partial fulfilment of the requirements for the award of the degree.

Advisory Committee

Chairman

Dr. S.A. Mir
Head,
Division of Agricultural Statistics,
SKUAST-Kashmir

Members

Dr. S. Maqbool
Assistant Professor, Division of Agricultural Statistics, SKUAST-Kashmir

Dr. Imran Khan
Assistant Professor, Division of Agricultural Statistics, SKUAST-Kashmir

Dr. Gul Zaffer
Professor, Division of Genetics and Plant Breeding, SKUAST-Kashmir

Dean PG Nominee

Dr. K.N. Singh,
Professor & Head, Division of Agronomy,
SKUAST-Kashmir

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Shalimar Campus, Srinagar – 190 025
-::o::-

Certificate – III

This is to certify that the thesis entitled, **"On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares"** submitted by **Mr. M Iqbal Jeelani Bhat (Regd. No. 2011-380-D)**, to the **Faculty of Postgraduate Studies, Sher-e-Kashmir University of Agricultural Sciences & Technology of Kashmir** in partial fulfilment of the requirements for the award of the degree of **Doctor of Philosophy in Statistics** was examined and approved by the Advisory Committee and External Examiner on …….

Chairman
Advisory Committee

External Examiner

Head,
Division of Agricultural Statistics

Director Resident Instruction-cum-Dean
Postgraduate Studies, SKUAST-Kashmir

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Division of Agricultural Statistics – 190 025
-::o::-

Name of the student	:	**M Iqbal Jeelani Bhat**
Registration No.	:	2011-380-D
Major subject	:	Statistics
Minor subjects	:	Genetics and Plant Breeding/Agronomy
Major advisor	:	**Dr. S.A. Mir,** Head, Division of Agricultural Statistics, SKUAST-Kashmir
Title of the Thesis	:	**"On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares"**

ABSTRACT

The present study was carried out on Rank set sampling with a view of increasing the efficiency of estimate of population mean. The basic premise for ranked set sampling (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 might be costly and/or time-consuming. The essence of RSS is similar to the classical stratified sampling. RSS can be considered as post-stratifying the sampling units according to their ranks in a sample. In present study simple linear regression models were considered with respect to samples taken from the identified sampling techniques like simple random sampling (SRS), systematic sampling (SYS) and rank set sampling (RSS). It was found that the coefficient of determination obtained from regression model based on rank set sample was higher than rest of two sampling schemes. Root mean square error, p values, coefficient of variation were much lower in rank set based regression model than others. Kernel density curves were more symmetric in case of rank set sample as compared to SRS and SYS. Using validation technique (Jacknifing) there was consistency in the measure of R^2, Adj R^2 and RMSE in case of RSS as compared to SRS and SYS. Ranked set sampling is introduced within the frame

work of stratified sampling. Rather than selecting a simple random sample within each stratum as is done in stratified simple random sampling (SSRS), a ranked set sample within each stratum is taken. From the simulation results it is concluded that RSS, when used in place of SRS in the final stage of stratified sampling, can provide considerably more accurate estimates of population means. New ratio estimators for RSS are proposed based on various combinations of known values of deciles, Median, Quartile deviation, coefficient of Skewness, Kurtosis, and Correlation coefficient of auxiliary variable. The proposed ratio estimators were more efficient than classical ratio estimators, and from various simulation results it was found that the efficiency of RSS estimators decreases as the correlation coefficient decreases, the efficiency increases as the set size *m* increases. Population mean under non responses is also studied under rank set sampling. Some new allocation schemes were proposed under RSS in order to study their effect on sampling variance. In most of the situations under different combinations of non-response rate and inverse ratio of sub-sampled non-response class, allocation schemes depending solely upon the knowledge of stratum size, non-response rate, mean squares of non-response group produces more precise estimates as compared to proportional allocation and other allocations based on knowledge of response and non-response rate only. From the results it is concluded that in addition to the knowledge of strata sizes, the knowledge of non-response rates and mean squares among non-response groups while allocating sample to different strata, improves the precision of the estimate. Different computer programmes were prepared using R-software and the analysis as per the objectives were carried out. In the preliminary study regression analysis and regression diagnostics was carriedout in SAS, while the simulation was carried out using the function `library (mvtnorm)` in R software. With the help of R -software new functions like `drss(m,r), varwts(n,h), makeAlloc (n,m)`and `ratio.est(n,N(x,y))`were developed. All these functions were run on real data set generated from forestry and horticultural crops.

Key words: Rank set sampling, Ratio estimators, Proposed ratio estimators, Non-Response, Allocations, SAS, R-software.

Signature of the Student

Dated_____

Signature of Major Advisor

Dated_____

ACKNOWLEDGEMENT

"Success belongs to those who are dare to act"

(Albert Einstein)

In the name of Almighty, "Allah", the most Benificient and Merciful, Billions of peace and Blessings be upon Holy Prophet (SAW). I bow in reverence to Almighty for giving me enough courage, patience and success in this venture.

I take this opportunity to express my sincere and profound gratitude to Dr. S.A. Mir, Head, Division of Agricultural Statistics, SKUAST-Kashmir, Shalimar Chairman of my advisory committee for his valuable guidance and constant encouragement throughout the course of study. His dynamic attitude, inspiring guidance and wholehearted encouragement led this task to its success and shall remain a life-long gifted memory for me.

I am highly greatful to Dr. Showkat Maqbool, Assistant Professor Division of Agricultural Statistics and Dr. Imran Khan, Assistant Professor Division of Agricultural Statistics for their constant encouragement, valuable suggestions and generous help during research and in the preparation of this manuscript. I extend my sincere thanks to members of my advisory committee Dr. Gul Zaffar, Professor, Division of Genetics and Plant Breeding; and Dr. K.N. Singh, Professor & Head, Division of Agronomy (Dean P.G Nominee) for their valuable guidance and suggestions during the study and for helping in finalization of the manuscript. I place on record my respect and thanks to Dr. Tariq Ahmad Raja and Dr. M.S. Pukhta for their whole hearted cooperation during the entire period of study. I am highly greatful to Miss Nageena Nazir, Assistant Professor, Division of Agricultural Statistics, SKUAST-Kashmir for her constant encouragement, valuable suggestions and generous help during preparation of this manuscript.

I am greatly thankful to Prof. Walid-bin-Abu, Professor, University of Yarmouk, Jordan for providing the valuble suggestions during the preparation of this manuscript.

I place on record my thanks to worthy Director Resident Instruction-cum-Dean Post Graduate Studies Prof. Badrul Hassan for his help and cooperation in this endeavour.

I place on record my gratitude to Dr. A.H. Mir, Professor (Ex-Head), Division of Agricultural Statistics, SKUAST-Kashmir, Shalimar.

I am very thankful to Hon'ble Vice Chancellor Dr. Tej Pratap for providing necessary facilities during the course of the study.

I place on record my gratitude to the University Grants Commission for providing the MANF national fellowship for perusing my doctoral programme.

I am highly thankfull to ARIS and Library staff members of SKUAST-Kashmir for their constant encouragement, valuable suggestions and generous help during preparation of this manuscript.

Special thanks to my friends for their cooperation, appreciation and nice company during my studies.

I am also thankful to all the non-teaching staff members of Division of Agricultural Statistics SKUAST-Kashmir for their co-operation.

Words fail to express my gratitude to my Father and my beautiful Mother for their good wishes, moral support, sustained help and constant encouragement which enabled me to complete this uphill task.

Last but not the least special thanks to Mr. Younus Ahmad Bhat and Mr. Rafiq Ahmad of M/s Universal Computers, Shalimar for composing this manuscript beautifully and giving it a final shape in shortest possible time.

M Iqbal Jeelani Bhat

Place : Shalimar, Srinagar

Dated :

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter – 1

INTRODUCTION

1.1 Introduction

The basic premise for Rank set sampling (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 original form of RSS conceived by McIntyre (1952) can be described as follows. First, a simple random sample of size *k* is drawn from the population and the *k* sampling units are ranked with respect to the variable of interest, say X, by judgment without actual measurement. Then the unit with rank 1 is identified and taken for the measurement of *X*. The remaining units of the sample are discarded. Next, another simple random sample of size *k* is drawn and the units of the sample are ranked by judgment, the unit with rank 2 is taken for the measurement of *X* and the remaining units are discarded. This process is continued until a simple random sample of size *k* is taken and ranked and the unit with rank *k* is taken for the measurement of *X*. This whole process is referred to as a cycle. The cycle then repeats *m* times and yields a ranked set sample of size $N = mk$. For k = 3, the sampling procedure is illustrated in Fig. 1.

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 a sample. Although the mechanism is different from the stratified sampling, the effect is the same in that the population is divided into sub-populations such that the units within each sub-population are as homogeneous as possible. In fact, we can consider any mechanism, not necessarily ranking the units according to their *X* values, which can post-stratify the sampling units in

Cycle 1

$$X_{(1)11} \leq X_{(2)11} \leq X_{(3)11} \rightarrow X_{(1)1}$$

$$X_{(1)21} \leq X_{(2)21} \leq X_{(3)21} \rightarrow X_{(2)1}$$

$$X_{(1)31} \leq X_{(2)31} \leq X_{(3)31} \rightarrow X_{(3)1}$$

Cycle 2

$$X_{(1)12} \leq X_{(2)12} \leq X_{(3)12} \rightarrow X_{(1)2}$$

$$X_{(1)22} \leq X_{(2)22} \leq X_{(3)22} \rightarrow X_{(2)2}$$

$$X_{(1)32} \leq X_{(2)32} \leq X_{(3)32} \rightarrow X_{(3)2}$$

$$\cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots \quad \cdots$$

Cycle m

$$X_{(1)1} \leq X_{(2)1} \leq X_{(3)1} \rightarrow X_{(1)}$$

$$X_{(1)2} \leq X_{(2)2} \leq X_{(3)2} \rightarrow X_{(2)}$$

$$X_{(1)3} \leq X_{(2)3} \leq X_{(3)3} \rightarrow X_{(3)}$$

Fig. 1: Demonstration of the ranked set sampling (RSS) procedure

such a way that it does not result in a random permutation of the units. The mechanism will then have similar effect to the ranking mechanism considered above.

RSS is applicable whenever a ranking mechanism can be found such that the ranking of sampling units is carried out easily and sampling is much cheaper than the measurement of the variable of interest. In particular, it is applicable in the following situations: (i) the ranking of a set of sampling units can be done easily by judgment relating to their latent values of the variable of interest through visual inspection or with the help of certain auxiliary means; (ii) there are certain easily obtainable concomitant variables available. By concomitant variables we mean variables that are not of major concern but are correlated with the variable of interest. These situations are abundant in practice.

RSS has been used in the assessments of animal carrying capacity of pastures and ranges. The sampling units in such assessments are well defined quadrats. The measurement of a quadrat involves mowing herbage and/or clipping browse within the quadrat and then drying and weighing the forage, which is quite time-consuming and destructive. However, nearby quadrats can be ranked rather precisely by visual inspection of an experienced person. If the variation between closely spaced quadrats is the same as that between widely spaced quadrats, quadrats taken nearby can be considered as random samples. RSS can increase greatly the efficiency in such situation. Hall and Dell (1966) reported an experiment by using ranked set sampling for estimating forage yield in shortleaf-loblolly pine-hardwood forests carried out on a 300-acre tract of the Stephen F. Austin Experimental Forest, near Nacogdoches, Texas, USA. In this investigation, the quadrats were determined by metal frames of 3.1 feet square. The measurement of a quadrat involves clipping the browse and herbage on the quadrat and then dry the forage to constant weight at 70^{o}C. Browse included the current season's leaf and twig growth of shrubs, woody vines, and trees available to deer or cattle. Herbage included all non-woody plants. Due to particular

features of the plantation in this area, the variation of the forage plants between closely spaced quadrats was essentially the same as that between widely spaced quadrats. The sampling procedure is as follows. First, set locations were established on a 2-chain grid. Then at each location, three metal frames were placed on the ground at randomly selected points within a circle 13 feet in radius. Two observers-one a professional range man, the other a woods worker-independently ranked the three quadrats. Browse and herbage were ranked separately. To investigate the efficiency of the ranked set sampling, all the forage on all the quadrats were clipped and dried. For browse, 126 sets of three quadrats were ranked and clipped and, for herbage, 124 sets were ranked and clipped. It turned out that, to achieve the same accuracy of the estimation with 100 randomly selected quadrants, only 48 quadrats need to be measured for browse and only 73 quadrats need to be measured for herbage by RSS.

Martin *et al.* (1980) evaluated RSS for estimating shrub phytomass in Appalachian Oak forest. In this application, sampling involves counting the number of each vegetation type in randomly selected blocks of forest stands, which is rather time-consuming. But the ranking of a small number of blocks by visual inspection can be done quite easily.

In the above two examples, ranking of a small number of sampling units is done by visual inspection. For the visual inspection to be possible, the sampling units must be nearby. As mentioned in Dell (1966), if the variation between nearby units is as same as that between far apart units, a set of sampling units taken nearby can be considered as a random sample. However, if the variation between nearby units is much smaller than that between far apart units, a set taken nearby is no longer a random sample from the population under study. To overcome this difficulty, either the entire area under study may be divided into sections such that within each section local variations are essentially the same as the overall variation and then apply RSS in each section, or some auxiliary tools

such as photos and video pictures may be used for the ranking of a randomly selected set of units.

The assessment of the status of hazard waste sites is usually costly. But, often, a great deal of knowledge about hazard waste sites can be obtained from records, photos and certain physical characteristics, etc., and then be used to rank the hazard waste sites. In certain cases, the contamination levels of hazardous waste sites can be indicated either by visual cues such as defoliation or soil discoloration, or by inexpensive indicators such as special chemically-responsive papers, or electromagnetic readings. A concrete example follows. In the early 1970s, an investigation was carried out to estimate the total amount of plutonium 239,240 Pu in surface soil within a fenced area adjacent to the Nevada Test Site of Nevada, U.S.A. Soil samples at different locations were collected and analyzed for 239,240 Pu using radiochemical techniques to determine the plutonium concentration per square meter of surface soil. The radiochemical analysis of the soil samples is costly. However, a concomitant variable, the Field Instrument for the Determination of Low Energy Radiation (FIDLER) counts per minute taken at soil sample location, can be obtained rather cheaply. The ranking of soil sample locations based on the FIDLER is rather cheap and simple. Gilbert and Eberhardt (1976). Yu and Lam (1997) applied retrospectively RSS with concomitant variables to the data collected from the above investigation and found that the statistical procedures based on RSS improves significantly those based on simple random samples.

In population census in certain countries, values of some variables are kept on a “short form” for all individuals. These records are available for a survey designer before a survey is carried out. In a survey, the values of survey variables are collected on a “long form”. The variables on the “short form”, which are easily obtainable, can be considered as concomitant variables. Therefore the RSS with concomitant variables can well be applied in surveys of this kind. The following case, which concerns the 1988 Test Population Census of Limeira, Sao

Paulo, Brazil, provides an example. The test census was carried out in two stages. In the first stage, a population of about 44,000 households was censused with a "short form" questionnaire. In the second stage, a systematic sample of about 10% of the population size was surveyed with a "long form" questionnaire. The "short form" contains variables such as sex, age and education of the head of household, ownership of house, car and color TV, the number of rooms and bathrooms, a proxy to the monthly income of the head of household, etc. The "long form" contains, besides the variables in the "short form", the actual monthly income of the head of household and other variables. The data obtained from the survey consists of the sample records of the "long form" for 426 heads of households. Details of the data were described in Silva and Skinner (1997). Chen (2002) used the data to illustrate the efficiency of RSS procedure and found that, for the estimation of the mean monthly income of the head of household, RSS is more efficient than simple random sampling in all statistical procedures considered.

Many quantitative traits of human such as hypertension and obesity can be attributed to genetic factors. In genetic linkage analysis, sibpair models are used for mapping quantitative trait loci in humans. To test whether or not a marker locus is associated with the quantitative trait under consideration, sib pairs are selected, the values of the quantitative trait of the pairs are measured, the genotypes at the locus of the pair are determined and the number of alleles that the pair have derived from a common parent [identical by descent (ibd)] at the locus is found. The data is then used to test whether the number of shared alleles ibd of the pair is correlated with the squared difference between the values of the quantitative trait of the pair. However, the power of the test by using a simple random sample is very low. To detect the association in existence, thousands of sib pairs are usually required. Risch and Zhang (1995) found that the power of the test can be significantly increased by selecting sib pairs with extremely

concordant or disconcordant trait values. In implementation, this requires to screen a large number of sib pairs before genotyping can be started, which will certainly be subject to practical limitations. To overcome the difficulty caused by practical limitations, RSS scheme can be employed for the selection of the sib pairs. Sib pairs can be screened in sets of an appropriate size. In each set, only the pair with either the smallest or the largest absolute difference in trait values is selected for genetyping. This procedure, while increasing the power of the test, is more practical.

RSS can be used in certain medical studies. For instance, it can be used in the determination of normal ranges of certain medical measures, which usually involves expensive laboratory tests. Samawi (1999) considered using RSS for the determination of normal ranges of bilirubin level in blood for new born babies. To establish such ranges, blood sample must be taken from the sampled babies and tested in a laboratory. But, on the other hand, the ranking of the bilirubin levels of a small number of babies can be done by observing whether their face, chest, lower parts of the body and the terminal parts of the whole body are yellowish, since, as the yellowish colour goes from face to the terminal parts of the whole body, the level of bilirubin in blood goes higher. RSS also has potential applications in clinical trials. Usually, the cost for a patient to go through a clinical trial is relatively high. However, the patients to be involved in the trial can be selected using the technique of RSS based on their information such as age, weight, height, blood pressure and health history etc., which can be obtained with a relatively negligible cost. RSS is expected to be more efficient than simple random sampling (SRS), when applicable, regardless how ranking is done. This is because, intuitively, a ranked set sample contains more information than a simple random sample of the same size, since a ranked set sample contains not only the

information carried by the measurements on the variable of interest but also the information carried by the ranks.

1.2 Historical background

The idea of ranked set sampling was first proposed by McIntyre (1952), 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 time consuming. 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 sampling scheme. Each time, a random sample of k pasture lots is taken and the lots are ranked by eye inspection with respect to the amount of yield. From the first sample, the lot with rank 1 is taken for cutting and weighing. From the second sample, the lot with rank 2 is taken, and so on. When each of the ranks from 1 to k has an associated lot being taken for cutting and weighing, the cycle repeats over again and again until a total of m cycles are completed. McIntyre illustrated the gain in efficiency by a computation involving various distributions. He observed that the relative efficiency, defined as the ratio of the variance of the mean of a simple random sample and the variance of the mean of a ranked set sample of the same size, is not much less than (k +1)/2 for symmetric or moderately asymmetric distributions, and that the relative efficiency diminishes with increasing asymmetry of the underlying distribution but is always greater than 1. McIntyre also illustrated the estimation of higher moments. In addition, McIntyre mentioned the problem of optimal allocation of the measurements among the ranks and the problems of ranking error and possible correlation among the units within a set, etc. Though there is no theoretical rigor, the work of McIntyre is pioneering and fundamental and farm the base line of modern RSS methodology.

The idea of RSS seemed buried in the literature for a long time until Halls and Dell (1959) conducted a field trial evaluating its applicability to the estimation of forage yields in a pine hardwood forest. The terminology ranked set

sampling was, in fact, coined by Halls and Dell (1966). The first theoretical result about RSS was obtained 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 (1952) later obtained similar results, however, without restricting to the case of perfect ranking. Dell and Clutter (1952) and David and Levine (1972) were the first to give some theoretical treatments on imperfect ranking. Stokes (1976, 1977) considered the use of concomitant variables in RSS. Up to this point, the attention had been focused mainly on the nonparametric estimation of population mean. A few years later, Stokes (1980) considered the estimation of population variance and the estimation of correlation coefficient of a bivariate normal population based on an RSS. However, other statistical procedures and new methodologies in the context of RSS had 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, various statistical procedures with RSS, non-parametric or parametric, have been investigated, variations of the original notion of RSS have been proposed and developed, and sound general theoretical foundations of RSS have been laid. A few references of these developments are given below. The estimation of cumulative distribution function with various settings of RSS was considered by Stokes and Sager (1988), Kvam and Samaniego (1993) and Chen (2000). The RSS version of distribution-free test procedures such as sign test, signed rank test and Mann-Whitney-Wilconxon test were investigated by Bohn and Wolfe (1992) and Hettmansperger (1995). The estimation of density function and population quantiles using RSS data were studied by Chen (1999, 2000). The RSS counterpart of ratio estimate was considered by Samawi and Muttlak (1996). The U-statistic and M-estimation based on RSS were considered, respectively, by Presnell and Bohn (1999) and

Zhao and Chen (2000). The RSS regression estimate was tackled by Patil *et al.* (1993), Yu and Lam (1997) and Chen (2001).

The parametric RSS assuming the knowledge of the family of the underlying distribution was studied by many authors, e.g., Abu-Dayyeh and Muttlak (1996), Bhoj (1997).

The optimal design in the context of unbalanced RSS was considered by Kaur *et al.* (1997), Ozturk and Wolfe (1998), Chen and Bai (2000) and Chen (2001). A general theory on parametric and non-parametric RSS was developed by Bai and Chen (2003). Ranking mechanisms based on the use of multiple concomitant variables were developed by Chen and Shen (2003). Rank set sampling procedures are used for obtaining the sub-sample from the set of non-respondents. Gaajendra and Bouza (2012) have applied double sampling to study the non-response situations, when rank set sampling is used in the sub-sample. They suggested that the first visit may serve for ranking accurately the sub-sampled non-respondents and applied two variations of rank set sampling (RSS), i.e. extreme-RSS and median-RSS for developing estimators of the population mean. Their expected variances and biases were obtained using Monte Carlo experiments. Various sampling strategies have been applied to study the behavior of non-responses.

1.3 Introduction to various sampling techniques

One of the vital issues in a sample survey is the choice of proper sampling techniques. In the choice of a sampling method, there are some methods of selection while others are control measures which help in grouping the population before the selection process. The basic sampling techniques which are employed are simple random sampling, systematic sampling and sampling with unequal probabilities of selection of units particularly with probability proportional to size. Among the control measures are procedures such as stratified sampling, cluster

sampling and multi-stage sampling, etc. Brief description of some of the important sampling techniques is given below:

1.3.1 Simple Random Sampling

Simple random sampling is a method of selecting *n* units out of the *N* such that every one of the ${}_NC_n$ distinct samples has an equal chance of being drawn. In practice a simple random sample is drawn unit by unit. The units in the population are numbered from *1* to *N*. A series of random numbers between *1* and *N* is then drawn, either by means of a table of random numbers or by means of a computer program that produces such table. At any draw the process used must give an equal chance of selection to any number in the population not already drawn. The units that bear these *n* numbers constitute the sample. It is easily verified that all ${}_NC_n$ distinct samples have an equal chance of being selected by this method. If a unit is selected and noted and then returned back to the population before the next drawing is made and this procedure repeated *n* times, it gives rise to a simple random sample of *n* units. This procedure is generally known as simple random sampling *with replacement (srswr).* If this procedure is repeated till *n* distinct units are selected and all repetitions are ignored. It is called a simple random sampling *without replacement (srswor).*Simple random sampling serves as a baseline for comparing the relative efficiency of other sampling method.

Some important results in simple random sampling and vital properties are as under:

Property 1 : The probability that a specified unit of the population being selected at any given draw is equal to the probability of its being selected at the first draw.

Property 2 : The probability of a specified unit being included in the sample is equal to *n/N*.

Corollary 1: The probability of a specified sample of n units, ignoring order, is $1/\binom{N}{n}$.

Corollary 2: In simple random sampling without replacement, the standard error of $\hat{Y}$ is given by;

$$\hat{\sigma}_{\hat{Y}} = NS\left[\frac{N-n}{nN}\right]^{1/2} = NS\left[\frac{1-f}{n}\right]^{1/2}$$

Corollary 3: In simple random sampling with replacement, the standard error of $\hat{y}$ can be written as;

$$V(\hat{y}) = \sigma^2/n \text{ and } \sigma_{\bar{y}} = \sigma/\sqrt{n}$$

Corollary 4: In simple random sampling with replacement, the variance and standard error of $\hat{Y} = N_{\bar{y}}$ can be written as ;

$$V(\hat{Y}) = N^2\sigma^2/n \text{ and } \sigma_{\hat{y}} = N\sigma/\sqrt{n}$$

Property 3: In simple random sampling $s^2 = \sum_i^n (y_i - \overline{y})^2/(n-1)$ is an unbiased estimator of $S^2 = N\sigma^2/(N-1)$

Corollary 1: An unbiased estimator of variance of $\bar{y}$ in random sampling without replacement, is given by;

$$v(\bar{y}) = (N-n)s^2/Nn = (1-f)\,s^2/n$$

Corollary 2: An unbiased estimator of the variance of $\hat{Y} = N\bar{y}$ in random sampling without replacement, is given by;
$v(\hat{Y}) = N(N-n)s^2/n = (1-f)N^2\,s^2/n$

Corollary 3: An unbiased estimator of variance of $\bar{y}$ in random sampling with replacement, is given by;

$$v(\bar{y}) = s^2/n$$

Corollary 4: An unbiased estimator of variance of $\hat{Y} = N^2s^2/n$ in random sampling with replacement, is given by

Estimators are generally calculated to estimate the population parameters. These estimators vary from sample to sample. A number of estimators are used to

draw inference about the population parameters. There are two general categories of estimators; (1) *ordered estimators* (2) *unordered estimators*.

The estimators that take into account the order in which the units are selected in the sample are called *ordered estimators*. An important ordered estimator is *Des Raj's ordered estimator*. Des Raj (1956) has proposed an estimator that makes the use of conditional probabilities without calculating inclusion probabilities which are difficult to calculate for many sampling schemes.

An estimator which does not depend on the order in which the units are drawn within the sample is called *unordered estimator*. Some of the important unordered estimators are;

Murthy's unordered estimator: Murthy (1957) suggested that an unordered estimator can be obtained by weighting all possible ordered estimators with their respective probabilities. In sampling *n* units without replacement from a finite population of *N* units, here will be $\binom{N}{n}$ unordered sample *(s).* Each unordered sample of size *n* can be ordered in *M (= n!)* ways, i.e., an unordered sample corresponds to *M* ordered samples.

Horvitz-Thompson estimator: Horvitz-Thompson (1952) suggested an estimator which is a biased estimator of population total. They proposed a general estimator of population total which possesses several desirable characteristics. This estimator with revised probabilities is always more efficient than usual estimators. The variance estimator under this scheme with revised probabilities is always non-negative. It may be noted that Horvitz-Thompson estimator can be used for any sampling design when the estimator has only distinct units in the sample. A typical survey sampling set up consists of a population *U* on *N* labeled units with a value y_i attached to the unit u_i for $i = 1,2,3,\ldots\ldots,N$. The finite population *U* a sample of size *n* is drawn without replacement. One of the problems of interest is to estimate $\bar{Y} = \sum_{i=1}^{N} yi\ /N$, the population means, by observing the *y*-values on a subset of units in the population. A popular choice is to select a sampling plan

that has both its first and its second order inclusion probabilities equal. It is noted that a simple random sampling plan is the most famous plan to achieve this. It is known that under simple random sampling, the first and second order inclusion probabilities are respectively given by $\pi_i = n/N$, for every individual unit, and $\pi_{ij} = n(n-1)/N(N-1)$ for all pairs of units, $i \neq j$. Having obtained these probabilities, the Horvitz-Thompson (1952) estimator reduces to the usual sample mean. The sample mean is unbiased with variance given by $V(\bar{y}) = \frac{N-n}{Nn}S^2$, which can be unbiasedly estimated by $\hat{V}(\bar{y}) = \frac{N-n}{Nn}s^2$, where $S^2 = \sum_{i=1}^{N}(y_i - \bar{Y})^2/N-1$ and $s^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/n-1$ have their usual meanings. It is clear that, for a fixed-size sampling design, the sampling error remains unchanged if the values π_i *and* π_{ij} are the same as that of simple random sampling. Horvitz and Thompson (1952), Murthy (1957) have shown that unordered estimators are more efficient that ordered estimators, because their variances are always less than or equal to that of ordered estimators.

1.3.2 Probability proportional to size sampling

A sampling procedure in which the units are selected with varying probabilities in proportion to some measure of the size of the sampling units is known as sampling with probability proportional to size (PPS). Under PPS sampling larger units has more probability of inclusion in the sample as compared to the unit of smaller size. Sampling with probability proportional to size (PPS) provides a practical technique when sampling from populations with large variation in the values of the study variable and often gives considerable gain in efficiency. Under PPS sampling an auxiliary size measure must be available and for efficient estimation the size measure should be strongly related to the study variable. More precisely, a size measure is sought for which ratio to the value of the study variable remains nearly a constant for all the population elements. These three basic sampling techniques along with control measures can be used to construct a manageable sampling design for a

complex sample surveys. In all the techniques, excluding simple random sampling, auxiliary information on the structure of the population is required. As a rule sampling error can be decreased by the proper use of auxiliary information.

The procedure of selecting a PPS sample consists in associating with each unit a number or a set of numbers equal to its size. The selection of units is done corresponding to a number chosen at random from totality of numbers. The commonly used selection procedures for a sample are; (1) Cumulative total method (2) Lahiri's method

Cumulative total method: let the size of the ith unit be denoted by X_i, the total size of N population units being $X = \sum X_i, i = 1,2,3, \dots, N$. Then, the selection procedure consists of following steps ;(1) Write down cumulative totals for sizes $X_i, i = 1,2,3, \dots, N$. (2) Choose a random number r, such that, $1 \leq r \leq X$. (3) Select ith population unit if $T_{i-1} = X_1 + X_2 + X_3 + \cdots \dots . . + X_{i-1}$ and $T_i = T_{i-1} + X_i$. The probability of selecting the ith unit, using this procedure, is given by $P_i = \frac{X_i}{X}$. Main difficulty in this procedure is the compulsion to compute successive cumulative totals, which becomes time consuming and costly when the population size is large.

Lahiri's method: Cumulative method involves cummulation of sizes and then writing down these cumulative totals, which is a tedious one. A procedure which avoids the need for calculating cumulative totals for each unit has been given by Lahiri (1951). It involves the following steps for selecting a sample;(1) select a random number say *i,* from *1 to N*.(2) select another random number *j,* such that *i* $\leq$ *j* $\leq$ *M,* where *M* is either equal to maximum of the sizes $\{X_i\}, i = 1,2,3, \dots \dots N,$ or is more than the maximum size in the population.

1.3.3 Systematic sampling

Systematic sampling is one of the most frequently used sample selection techniques. A list of population elements or a register serves as the selection frame from which every q^{th} element can be systematically selected. For

example, many population registers are alphabetically ordered by family name. The first member is selected at random among the first q element. The rest of the sample is selected by taking every q^{th} element thereafter down to the end of the list. Systematic sampling may in some cases be more effective than simple random sampling. This will occur, for example, if there is a certain relationship between the ordering of the frame population and the values of the study variable. The most common cases are those where the population is already stratified or a trend exists that follows the population ordering or there is a periodic trend; all these situations can also be reached by appropriate sorting procedures. Periodicity may be harmful in some cases especially if harmonic variation coincides with the sampling interval

1.3.4 Stratified sampling

In stratified sampling, the population consisting of N units is first divided into K sub-populations of $N_1, N_2, N_3 \ldots\ldots\ldots, N_k$ units respectively. These sub-populations are non overlapping and together they compromise the whole of the population *i.e* $\sum_{i=1}^{k} Ni = N$. These sub-populations are called strata. Stratification, the values of N_i's must be known. When the strata have been determined, a sample is drawn from each stratum, the drawings being made independently in different strata. If a simple random sample is taken in each stratum then the procedure is termed as stratified random sampling. As the sampling variance of the estimate of mean or total depends on within strata variation, the stratification of population is done in such a way that strata are homogeneous within themselves with respect to the variable under study. However, in many practical situations, it is usually difficult to stratify with respect to the variable under consideration especially because of physical and cost consideration. Generally the stratification is done according to administrative groupings, geographical regions and on the basis of auxiliary characters correlated with the character under study.

1.3.5 Multi-stage sampling

Generally, elements belong to the same cluster are more homogeneous as compared to those elements which belong to different clusters. Therefore, a comparatively representative sample can be obtained by enumerating each cluster partially and distributing the entire sample over more clusters. This will increase the cost of the survey but the proportionate increase in cost vis-à-vis cluster sampling will be less as compared to increase in the precision. This process of first selecting cluster and then further sampling units within a cluster is called as two-stage sampling. The clusters in a two-stage sample are called as primary-stage units (psu) and elements within a cluster are called as second-stage units (ssu). A two-stage sample has the advantage that after psu's are selected the frame of the ssu's is required only for the sampled psu's. The procedure allows the flexibility of using different sampling design at the different stages of selection of sampling units. A two-stage sampling procedure can be easily generalized to multi-stage sampling designs. Such a sampling design is commonly used in large scale surveys. It is operationally convenient, provides reasonable degree of precision and is cost-wise efficient.

1.3.6 Ratio and regression methods of estimation

In sampling theory if the auxiliary information, related to the character under study, is available on all the population units, then it may be advantageous to make use of this additional information in survey sampling. One way of using this additional information is in the sample selection with unequal probabilities of selection of units. The knowledge of auxiliary information may also be exploited at the estimation stage. The estimator can be developed in such a way that it makes use of this additional information. Ratio estimator and Regression estimators are the examples of such estimators. Obviously, it is assumed that the auxiliary information is available on all the sampling units. In case the auxiliary information is not available then it can be obtained easily without much burden on the cost. Another way the auxiliary information can be

used is at the stage of planning of survey. An example of this is the stratification of the population units by making use of the auxiliary information.

1.3.7 Cluster sampling

A cluster may be defined as a group of units. When the sampling units are clusters the method of sampling is known as cluster sampling. Cluster sampling is used when the frame of units is not available or it is expensive to construct such a frame. Thus, a list of all the farms in the districts may not be available but information on the list of villages is easily available. For carrying out any district level survey aimed at estimating the yield of a crop, it is practically feasible to select villages first and then enumerating the elements (in this case farms) in the selected village. The method is operationally convenient, less time consuming and more importantly such a method is cost-wise efficient. The efficiency of cluster sampling procedure increases as the heterogeneity between units belonging to same cluster increases. Cluster sampling becomes more efficient than element sampling if the units pertaining to same cluster are negatively, correlated.

1.4 Preliminaries

Some of the basic sampling concepts and notations to be used in this study hereafter are given below:

Population: The collection of all units of a specified type in a given region at a particular point or period of time is termed as a population or universe. Thus, we may consider a population of persons, families, farms, cattle in a region or a population of trees or birds in a forest or a population of fish in a tank etc. depending on the nature of data required.

Sampling Unit: Elementary units or group of units which besides being clearly defined, identifiable and observable are convenient for purpose of sampling are called sampling units. For instance, in a crop survey, a farm or a group of farms owned or operated by a household may be considered as the sampling unit.

Sampling Frame: A list of all the sampling units belonging to the population to be studied with their identification particulars or a map showing the boundaries of the sampling units is known as sampling frame. Examples of a frame are a list of farms and a list of suitable area segments like villages, blocks in India. The frame should be up to date and free from errors of omission and duplication of sampling units.

Population Parameter: Any function of all the population values or observations is termed as population parameter or simply parameter. e.g. population mean μ, population variance σ^2 etc.

Sample statistic or Estimator: Any function of the sample values which is free from population parameters is called sample statistic or estimator.eg$(\bar{x},\ s^2)$ Generally it is calculated to estimate the population parameters. The estimator varies from sample to sample. For a particular sample, the value of the estimator in called estimate.

Sampling distribution: For all possible samples from the population, the distribution of the sample statistic is called sampling distribution. The standard deviation of the sampling distribution is called standard error of the distribution.

Expected Value or average value of Estimator: Suppose in a probability sampling scheme π_i is the probability of selecting the i^{th} samples. Let t_i is the estimate of parameter θ from the i^{th} sample and let M_o be the total of all possible samples from the population. Then expected value or average value of the estimator 't' is defined by

$$\sum_{i=1}^{M_o} t_i\pi_i \text{ and denoted by E(t)} \qquad \text{i.e. E(t)} = \sum_{i=1}^{M_o} t_i\pi_i$$

Unbiased Estimator: The estimator 't' is said to be unbiased estimator of parameter θ if E(t) is equal to θ

i.e. $E(t) = \sum_{i=1}^{M_o} t_i \pi_i = \theta$

Biased Estimator : The estimator *'t'* is said to be biased estimator of parameter θ, if $E(t) \neq \theta$

Bias in the estimator *'t'* is given by

$B(t) = E(t) - \theta$

Consistent Estimator: For finite populations and estimator, *'t'* is said to be consistent estimator of θ if

$t \rightarrow \theta$ when $n \rightarrow N$

'n' is sample size and *'N'* is population size for infinite populations, *'t'* is said to be consistent if

$$\lim_{n \longrightarrow \infty} P\left[\left|t - \theta\right| > \in\right] = 0 \quad \in > 0$$

Mean Square Error (MSE) of the Estimator**:** Suppose in a probability sampling scheme t_i is the estimate of the parameter θ from the i^{th} sample, then $(t_i - \theta)$ is called sample error. It varies from sample to sample. The expected value of the square of sampling error (or expected value of the square of deviation) of the estimator *'t'* from its true value θ is termed as *MSE* of the estimator t and denoted by *MSE*

$$MSE\ (t) = E\ (t- \theta)^2 = \sum_{i=1}^{M_o} \left(t_i - \theta\right)^2 \pi_i$$

The square root of *MSE (t)* is called root mean square

Sampling Variance: The expected value of the square of deviation of the estimator *'t'* from its expected value is termed as variance of the estimator t and denoted by *V(t)* or σ^2

$V(t) = \sigma^2 = E\ [t - E(t)\ ]^2$

It is measure of divergence of the estimator from its expected value.

Coefficient of Variance: It is defined by standard error of *'t'* divided by expected value of *t.* i.e. $CV(t)= s.e\ (t)/E(t)$

Efficiency: Given two estimators t_1 and t_2 of parameter θ then t_1 is said to be more efficient than t_2, if $MSE\ (t_1) < MSE\ (t_2)$

On the other hand t_1 is said to be more precise estimator than t_2 if

$V(t_1) < V(t_2)$

Relative efficiency: Relative efficiency *(RE)* of t_1 as compared to t_2 is defined as $= MSE\ (t_2)/MSE\ (t_1)$

Relative precision: Relative precision t_1 as compared to t_2 is= $V\ (t_2) / V\ (t_1)$

Sampling and Non-Sampling Error: The error arising due to drawing inferences about the population on the basis of observations on a part (sample) of it is termed sampling error. The sampling error is non-existent in a complete enumeration survey since the whole population is surveyed. The errors other than sampling errors such as those arising through non-response, in-completeness and inaccuracy of response are termed non-sampling errors and are likely to be more wide-spread and important in a complete enumeration survey than in a sample survey. Non-sampling errors arise due to various causes right from the beginning stage when the survey is planned and designed to the final stage when the data are processed and analyzed. The sampling error usually decreases with increase in sample size (number of units selected in the sample) while the non-sampling error is likely to increase with increase in sample size. As regards the non-sampling error, it is likely to be more in the case of a complete enumeration survey than in the case of a sample survey since it is possible to reduce the non-sampling error to a great extent by using better organization and suitably trained personnel at the field and tabulation stages in the latter than in the former.

1.5 Ranking mechanism

The procedure is a two-stage scheme. At the first stage, simple random samples are drawn and a certain ranking mechanism is employed to rank the units in each simple random sample. At the second stage, actual measurements of the variable of interest are made on the units selected based on the ranking information obtained at the first stage. The judgment ranking relating to the latent values of the variable of interest, as originally considered by McIntyre (1952), provides one ranking mechanism.

Let us start with McIntyre's (1952) original ranking mechanism, i.e., ranking with respect to the latent values of the variable of interest. If the ranking is perfect, that is, the ranks of the units tally with the numerical orders of their latent values of the variable of interest, the measured values of the variable of interest are indeed order statistics. In this case, $f_{[r]} = f_{(r)}$, the density function of the *r*th order statistic of a simple random sample of size k from distribution *F*. We have

$$f_{(r)}(x) = \frac{k!}{(r-1)!(k-r)!} F^{r-1}(x)[1-F(x)]^{k-r} f(x)$$

It is then easy to verify that

$$f(x) = \frac{1}{k}\sum_{r=1}^{k} f_{(r)}(x)$$

for all *x*. This equality plays a very important role in RSS. It is this equality that gives rise to the merits of RSS. The equalities of this kind are regarded as the fundamental equalities. A ranking mechanism is said to be consistent if the following fundamental equality holds:

$$F(x) = \frac{1}{k}\sum_{r=1}^{k} F_{(r)}(x), \text{ for all } x.$$

Obviously, perfect ranking with respect to the latent values of X is consistent. Other consistent ranking mechanisms are as follows :

1.5.1 Imperfect ranking with respect to the variable of interest

When there are ranking errors, the density function of the ranked statistic with rank r is no longer $f_{(r)}$. However, we can express the corresponding cumulative distribution function $F_{[r]}$ in the form:

$$F_{[r]}(x) = \sum_{s=1}^{k} p_{sr} F_{(s)}(x),$$

Where p_{sr} denotes the probability with which the sth (numerical) order statistic is judged as having rank r. If these error probabilities are the same within each cycle of a balanced RSS, we have $\sum_{s=1}^{k} p_{sr} = \sum_{r=1}^{k} p_{sr} = 1$. Hence,

$$\frac{1}{k}\sum_{r=1}^{k} F_{[r]}(x) = \frac{1}{k}\sum_{r=1}^{k} \sum_{s=1}^{k} p_{sr} F_{(s)}(x)$$

$$= \frac{1}{k}\sum_{s=1}^{k} \sum_{r=1}^{k} p_{sr} F_{(s)}(x) = F(x).$$

1.5.2 Ranking with respect to a concomitant variable

There are cases in practical problems where the variable of interest, X, is hard to measure and difficult to rank as well but a concomitant variable, Y, can be easily measured. Then the concomitant variable can be used for the ranking of the sampling units. The RSS scheme is adapted in this situation as follows. At the first stage of RSS, the concomitant variable is measured on each unit in the simple random samples, and the units are ranked according to the numerical order of their values of the concomitant variable. Then the measured X values at the second stage are induced order statistics by the order of the Y values. Let $Y_{(r)}$ denote the rth order statistic of the Y 's and $X_{[r]}$ denote its corresponding X. Let $f[X|Y_{(r)}(x|y)]$ denote the conditional density function of X given $Y_{(r)} = y$ and $g_{(r)}(y)$ the marginal density function of $Y_{(r)}$. Then we have

$$f_{[r]}(x) = \int fx|Y_{(r)}(x|y)g_{(r)}(y)dy.$$

Then

$$f(x) = \int \sum_{r=1}^{k} \frac{1}{k} fx|Y_{(r)}(x|y)g_{(r)}(y)dy.$$

$$= \frac{1}{k}\sum_{r=1}^{k} f_{(r)}(x).$$

1.5.3 Multivariate samples obtained by ranking one of the variables

With-out loss of generality, let us consider the bivariate case. Suppose that inferences are to be made on the joint distribution of *X* and *Y*. The RSS scheme can be similarly adapted in this case. The scheme goes the same as the standard RSS. The sampling units are ranked according to one of the variables, say Y. However, for each item to be quantified, both variables are measured. Let *f*(*x*, *y*) denote the joint density function of *X* and *Y* and *f*[*r*](*x*, *y*) the joint density function of *X*[*r*] and *Y*[*r*]. Then

$$f_{[r]}(x, y) = fX\ |Y_{[r]}(x|y)g_{[r]}(y)$$

and

$$f(x, y) = \frac{1}{k}\sum_{r=1}^{k} f_{[r]}(x, y).$$

1.6 Estimation of population mean using ranked set sampling

Let $h_{(x)}$ be any function of *x*. Denote by μ_h the expectation of *h*(*X*), i.e., $\mu h = Eh(X)$. We consider in this section the estimation of μ_h by using a ranked set sample. Examples of *h*(*x*) include: (a) $h(x) = x^l$, l =1, 2, ⋯, corresponding to the estimation of population moments, (b) $h(x)= I\{x \leq c\}$ where $I\{\cdot\}$ is the usual indicator function, corresponding to the estimation of distribution function, (c) $h(x) = \frac{1}{\lambda}K\left(\frac{t-x}{\lambda}\right)$, where *K* is a given function and λ is a given constant,

corresponding to the estimation of density function. We assume that the variance of $h(X)$ exists, then

$$\hat{\mu}_{h.RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} h(X_{[r]i}).$$

We consider first the statistical properties of $\hat{\mu}_{h.RSS}$ and then the relative efficiency of RSS with respect to SRS in the estimation of population mean.

First, we have the following result.

Theorem 1.6.1. *Suppose that the ranking mechanism in RSS is consistent. Then,*

i) The estimator $\hat{\mu}_{h.RSS}$ is unbiased, i.e., $E\hat{\mu}_{h.RSS} = \mu_h$

ii) $\text{Var}(\hat{\mu}_{h.RSS}) \le \frac{\sigma_h^2}{mk}$, where σ_h^2 denotes the variance of $h(X)$, and the inequality is strict unless the ranking mechanism is purely random.

iii) As $m \to \infty$,

$$\sqrt{mk}\,(\hat{\mu}_{h.RSS} - \mu_h) \to N(0, \sigma^2_{h.RSS})$$

in distribution, where,

$$\sigma^2_{h.RSS} = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r].}$$

Here $\sigma^2_{h[r]}$ denotes the variance of $h(X_{[r]i})$

Proof : i) It follows from the fundamental equality that

$$E\hat{\mu}_{h.RSS} = \tfrac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} Eh(X_{[r]i}) = \frac{1}{k}\sum_{r=1}^{k} Eh(X_{[r]i})$$

$$= \frac{1}{k}\sum_{r=1}^{k}\int h(x)\,dF_{[r]}(x) = \int h(x)\,d\,\frac{1}{k}\sum_{r=1}^{k} F_{(r)}(x)$$

$$= \int h(x)\,dF(x)\mu_h$$

ii) $$\text{Var}\ (\hat{\mu}_{h.RSS}) = \frac{1}{(mk)^2}\sum_{r=1}^{k}\ \sum_{i=1}^{m}\text{Var}\ (h(X_{[r]}i)) = \frac{1}{mk^{\ 2}}\sum_{r=1}^{k}\text{Var}\ \ (h(X_{[r]}))$$

$$= \frac{1}{mk}\left(\frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - [Eh(X_{[r]})]^2\right)$$

$$= \frac{1}{mk}\left(m_{h2} - \frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2\right),$$

where m_{h2} denotes the second moment of $h(X)$. It follows from the Caushy-Schwarx inequality that

$$\frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2 \geq \left(\frac{1}{k}\sum_{r=1}^{k}Eh(X_{[r]})\right)^2 = \mu_h^2,$$

where the equality holds only when $Eh(X_{[1]}) = \cdots = Eh(X_{[r]})$ in which case the ranking mechanism is purely random.

iii) By the fundamental equality, $\mu_h = \frac{1}{k}\sum_{r=1}^{k}\mu_{h[r]}$, where $\mu_{h[r]}$ is the expectation of $h(X_{[r]i})$. Then, we can write

$$\sqrt{mk}\ (\hat{\mu}_{h.RSS} - \mu_h) = \frac{1}{\sqrt{k}}\sum_{r=1}^{k}\sqrt{m}\,[\frac{1}{m}\sum_{i=1}^{m}h(X_{[r]}i) - \ \mu_{h[r]}]$$

$$= \frac{1}{\sqrt{k}}\sum_{r=1}^{k}Z_{mr}, \text{say.}$$

By the multivariate central limit theorem, (Z_{ml}, $\cdots$, Z_{mk}) converges to a multivariate normal distribution with mean vector zero and covariance matrix given by $Diag\ (\sigma^2_{h[1],\ldots,}\sigma^2_{h[k]})$.

We know that $\sigma_h^2/(mk)$ is the variance of the moment estimator of μ_h based on a simple random sample of size mk. Theorem 1.6.1 implies that the moment estimator of μh based on an RSS sample always has a smaller variance than its counterpart based on an SRS sample of the same size. In the context of

RSS, the cost or effort for drawing sampling units from the population and then ranking them is negligible. When we compare the efficiency of a statistical procedure based on an RSS sample with that based on an SRS sample, we assume that the two samples have the same size. Let $\hat{\mu}_{h.SRS}$ denote the sample mean of a simple random sample of size *mk*. We define the relative efficiency of RSS with respect to SRS in the estimation of μ_h as follows:

$$RE\,(\hat{\mu}_{h.RSS}, \hat{\mu}_{h.SRS})\frac{Var\,(\hat{\mu}_{h.SRS})}{Var\,(\hat{\mu}_{h.RSS})}$$

Theorem 1.6.1 implies that RE $(\hat{\mu}_{h.RSS}, \hat{\mu}_{h.SRS}) \geq 1$. In order to investigate the relative efficiency in more detail, we derive the following :

$$\sigma^2_{h.RSS} = \frac{1}{k}\sum_{r=1}^{k}\sigma^2_{h[r]}$$

$$= \frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - [Eh(X_{[r]})]^2)$$

$$= \frac{1}{k}\sum_{r=1}^{k}(E[h(X_{[r]})]^2 - \mu_h^2 + \mu_h^2 - \frac{1}{k}\sum_{r=1}^{k}[Eh(X_{[r]})]^2$$

$$= \sigma_h^2 - \frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2.$$

Thus, we can express the relative efficiency as :

$$RE\,(\hat{\mu}h_{.RSS}, \hat{\mu}h_{.SRS}) = \frac{\sigma_h^2}{\sigma^2_{h.RSS}} = \left[1 - \frac{\frac{1}{k}\sum_{r=1}^{k}(\mu_{h[r]} - \mu_h)^2}{\sigma_h^2}\right]^{-1}$$

It is clear from the above expression that, as long as there is at least one *r* such that $\mu_{h[r]} \neq \mu h$, the relative efficiency is greater than 1. For a given underlying distribution and a given function h, the relative efficiency can be computed, at least, in principle.

Based on the computations on a number of underlying distributions, following conjectures are made : the relative efficiency of RSS with respect to SRS, in the estimation of population mean, is between 1 and (k+1)/2 where k is the set size; For symmetric underlying distributions, the relative efficiency is not much less than (k+1)/2, however, as the underlying distribution becomes asymmetric, the relative efficiency drops down but never diminishes to less than1. Takahasi and Wakimoto (1968) showed that, when ranking is perfect, $\frac{1}{k}\sum_{r=1}^{k}\sigma_{h[r]}^2$, as a function of *k*, decreases as *k* increases, which implies that the relative efficiency increases as *k* increases. A practical implication of this result is that, in the case of judgment ranking relating to the latent values of the variable of interest, when ranking accuracy can still be assured or, in other cases, when the cost of drawing sampling units and ranking by the given mechanism can still be kept at a negligible level, the set size *k* should be taken as large as possible. The relative efficiency for a number of underlying distributions are computed. The relative efficiency is affected by the underlying distribution, especially by the skewness and kurtosis. The notations μ, σ^2, γ and κ in the Table-1 stands, respectively, for the mean, variance, skewness and kurtosis.

Table-1 : *μ*, σ^2, γ, κ and the relative efficiency of RSS with *k* = 2, 3, 4 for some distributions

Distribution	μ	σ^2	γ	κ	2	3	4
Uniform	0.500	0.083	0.000	1.80	1.500	2.000	2.500
Exponential	1.000	1.000	2.000	9.00	1.333	1.636	1.920
Gamma	0.500	0.500	2.828	15.0	1.254	1.483	1.696
Normal	0.000	1.000	0.000	3.00	1.467	1.914	2.347
Beta	0.500	0.028	0.000	2,45	1.484	1.958	2.425
Weibull	2.000	20.00	6.619	87.7	1.127	1.236	1.334
χ^2	0.789	0.363	0.995	3.87	1.430	1.841	2.239
Triangular	0.500	0.042	0.000	2.40	1.485	1.961	2.430

1.7 Estimation of smooth-function-of-means using ranked set sampling

In this section we deal with the properties of RSS for a particular model, the smooth-function-of-means model, which refers to the situation where we are interested in the inference on a smooth function of population moments. Typical examples of smooth-function-of-means are (i) the variance, (ii) the co-efficient of variation, and (iii) the correlation coefficient, etc. Let $m_1,..., m_p$ denote p moments of F and g a p-variate smooth function with first derivatives. We consider the method-of-moment estimation of $g(m_1,..., mp)$.

The following notation will be used in this section. Let Z_l, l =1,..., p, be functions of X (~F) such that E $[Z_l]= m_l$. Let $n = km$. A simple random sample of size n is represented by $\{(Z_{1j}, ..., Z_{pj}): j =1,..., n\}$. A general RSS sample of size n is represented by $\{(Z_{1(r)i},..., Z_{p(r)i}): r =1,..., k; i = 1,..., m\}$. The simple random and ranked set sample moments are denoted, respectively, by

$$\tilde{Z}_l = \frac{1}{n}\sum_{j=1}^{n} Z_{lj}, l=1,..., p$$

and

$$\tilde{Z}_l = \frac{1}{km}\sum_{r=1}^{n}\sum_{i=1}^{m} Z_{l(r)i}, l=1,..., p.$$

Let $\tilde{Z}_{SRS} = (\tilde{Z}_1,..., \tilde{Z}_p)^T$ and $\tilde{Z}_{RSS} = (\tilde{Z}_1,..., \tilde{Z}_p)^T$. Denote by $\sum_{SRS}$ and $\sum_{SRS}$ the variance-covariance matrices of $\sqrt{n}\tilde{Z}_{SRS}$ and $\sqrt{n}\tilde{Z}_{RSS}$, respectively. Let ag denote the vector of the first partial derivatives of g evaluated at (m_1, …, m_p). Define

$$\eta = g(m_1,..., m_p)$$

$$\hat{\eta}_{SRS} = g(\tilde{Z}_1,..., \tilde{Z}_p)$$

$$\hat{\eta}_{RSS} = g(\tilde{Z}_1,..., \tilde{Z}_p)$$

We first state the asymptotic normality of $\hat{\eta}_{SRS}$ and $\hat{\eta}_{RSS}$, and then consider the asymptotic relative efficiency (ARE) of $\hat{\eta}_{RSS}$ with respect to $\hat{\eta}_{SRS}$

Theorem 1.7.1. *As* $m \to \infty$ (*hence* $n \to \infty$), *we have*

$$\sqrt{n}(\hat{\eta}_{SRS} - \eta) \to N(0, ag^T \sum_{SRS} \text{ag})$$

in distribution and

$$\sqrt{n}(\hat{\eta}_{SRS} - \eta) \to N(0, ag^T \sum_{RSS} ag)$$

in distribution

The above result follows from the multivariate central limit theorem. The proof is omitted. The ARE of $\hat{\eta}_{RSS}$ with respect to $\hat{\eta}_{SRS}$ is defined as

$$ARE(\hat{\eta}_{SRS}, \hat{\eta}_{SRS}) = \frac{ag^T \sum_{SRS} ag}{ag^T \sum_{RSS} ag}$$

The next theorem implies that the ARE of $\hat{\eta}_{RSS}$ with respect to $\hat{\eta}_{SRS}$ is always greater than 1.

Theorem 1.7.2. *Suppose that the ranking mechanism in RSS is consistent. Then we have that*

$$\sum_{SRS} \geq \sum_{RSS}$$

where $\sum_{SRS} \geq \sum_{RSS}$ means that $\sum_{SRS} \geq \sum_{RSS}$ is non-negative definite.

Proof : It suffices to prove that, for any vector of constants *a,*

$$\alpha^T \sum_{SRS} \alpha^T \sum_{RSS} \alpha \geq 0.$$

Define

$$Y = \alpha^T Z = \sum_{j=1}^{p} {}_{aj} Z_j \text{ and } \mu \text{Y} = \alpha^{\text{T}} \text{m} = \sum_{j=1}^{p} ajmj.$$

Then we have

$$\hat{\mu}_{Y.SRS} = \alpha^T \tilde{Z}_{SRS} \text{ and } \hat{\mu}_{\text{Y.RSS}} = \alpha^T \tilde{Z}_{RSS.}$$

It follows from Theorem 1.5.1 that

$$Var(\hat{\mu}_{Y.RSS}) \leq Var(\hat{\mu}_{\text{YSRS}}),$$

i.e.,

$$\alpha^T \sum_{SRS} a \geq a^T \sum_{RSS} a$$

The theorem is proved

In fact, it can be proved that, as long as there are at least two ranks, say *r* and *s*, such that $F_{[r]} \neq F_{[s]}$, then $\Sigma_{\text{SRS}} > \Sigma_{\text{RSS}}$.

It should be noted that, unlike in the estimation of means, the estimator of a smooth-function-of-means is no longer necessarily unbiased. It is only asymptotically unbiased. In this case, the relative efficiency of RSS with respect to SRS should be defined as the ratio of the mean square errors of the two estimators. The ARE, which is the limit of the relative efficiency as the samples size goes to infinity, does not take into account the bias for finite sample sizes. In general, the ARE can not be achieved when sample size is small.

1.8 Estimation of variance using rank set sampling

The natural estimates of σ^2 using an SRS sample and an RSS sample are given, respectively, by

$$S^2_{SRS} = \frac{1}{mk-1} \sum_{r=1}^{k} \sum_{i=1}^{m} (X_{ri} - \overline{X}_{SRS})^2,$$

where $\overline{X}_{SRS} = \frac{1}{mk} \sum_{r=1}^{k} \sum_{i=1}^{m} X_{ri}$, and

$$S^2_{RSS} = \frac{1}{mk-1} \sum_{r=1}^{k} \sum_{i=1}^{m} (X_{[r]i} - \overline{X}_{SRS})^2,$$

where $\overline{X}_{RSS} = \frac{1}{mk}\sum_{r=1}^{k}\sum_{i=1}^{m} X_{[r]i.}$

The properties of S^2_{RSS} were studied by Stokes (1980) Unlike the SRS version S^2_{SRS}, the RSS version S^2_{RSS} is biased. It can be derived, that :

$$E_{S^2_{RSS}} = \sigma^2 + \frac{1}{k(mk-1)}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2$$

An appropriate measure of relative efficiency of S^2_{RSS} with respect to S^2_{SRS} is then given by

$$RE(S^2_{RSS}, S^2_{SRS}) = \frac{Var(S^2_{SRS})}{MSE(S^2_{RSS})}$$

$$= \frac{Var(S^2_{SRS})}{Var(S^2_{RSS}) + [\frac{1}{k(mk-1)}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2]^2}$$

It can be easily seen that

$$RE(S^2_{RSS}, S^2_{SRS}) < ARE(S^2_{RSS}, S^2_{RSS})$$

Since

$$\frac{1}{k}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2 < \sigma^2,$$

it is clear that $\frac{1}{k(mk-1)}\sum_{r=1}^{k}(\mu_{[r]} - \mu)^2$ will decrease as either *k* or *m* increases.

i.e., the relative efficiency will converge increasingly to ARS as either *k* or *m* increases.

1.9 An overview of the data sets

For the present study two data sets were utilized, one was taken from apple and another one from Pinus. For apple data the block Ganderbal was selected for

the present study in District Ganderbal. District Ganderbal being inseparable part of the state naturally inherits the same characteristics which predominately exist in the economy of the state. Agriculture is the main source of income and employment in the district. More than half of the population, directly and indirectly derive their livelihood from it. Paddy, maize and horticulture are the principle crops grown in the district. There is good network of agricultural infrastructure available throughout the length and breadth of the district. Total area sown under different food and non-food crops is about 27735 hectares, out of which 15828 hectares constituting 57 per cent was under cereal food crops. At present 8738 hectares are under major horticulture crops with 3866 hectares constituting 44 per cent are under apple cultivation and out of 47916 MT of production of horticulture crops, apple production is 34873 MT which is 72 per cent of the total production. A survey was conducted for estimation of average yield of apple in the district Ganderbal at block level. Since at present 8738 hectares are under major horticulture crops with 3866 hectares constituting 44 per cent of the area is under apple cultivation in district Ganderbal. A total of 420 orchards were reported in the block Ganderbal covering an area of 772.8 hectares with 73,496 total number of trees. Total production of apple in the block was found out to be 6758.52 metric tons (Mt) with the productivity of 8.74 Mt/ha. American, Delicious and Maharaji were the main varieties of apple cultivated in the block.

Data on *Pinus wallichiana* was also considered in the present study. *Pinus wallichiana* is a coniferous evergreen tree native to the Himalaya, Karakoram and Hindu Kush mountains, from eastern Afghanistan east across northern Pakistan and India to Yunnan in southwest China. It grows in mountain valleys at altitudes of 1800–4300 m (rarely as low as 1200 m), between 30 m and 50 m in height. It favours a temperate climate with dry winters and wet summers. This tree is often known as 'Bhutan pine', (not to be confused with the recently described Bhutan white pine, *Pinus bhutanica*, a closely related species). Other

names include 'blue pine', 'Himalayan white pine' and 'Himalayan blue pine'. In the past, it was also known by the invalid botanic names *Pinus griffithii* McClelland or "*Pinus excelsa*" Wall., *Pinus chylla* Lodd. when the tree became available through the European nursery trade in 1836, nine years after Dr Wallich first introduced seeds to England. The leaves ("needles") are in fascicles (bundles) of five and are 12–18 cm long. They are noted for being flexible along their length, and often droop gracefully. The cones are long and slender, 16–32 cm, yellow-buff when mature, with thin scales; the seeds are 5–6 mm long with a 20–30 mm wing. Typical habitats are mountain screes and glacier forelands, but it will also form old growth forests as the primary species or in mixed forests withdeodar, birch, spruce, and fir. In some places it reaches the tree line. The wood is moderately hard, durable and highly resinous. It is a good firewood but gives off a pungent resinous smoke. It is a commercial source of turpentine which is superior quality than that of *P. roxburghii* but is not produced so freely. It is also a popular tree for planting in parks and large gardens, grown for its attractive foliage and large, decorative cones. It is also valued for its relatively high resistance to air pollution, tolerating this better than some other conifers. The data on *Pinus wallichiana* was taken from block Langate of District Baramaula from Forest department J&K.

While scanning the review of literature it has been found that no such work with regards to agricultural/Forestry data has so far been conducted. Therefore, the present investigation entitled “On some aspects of Rank set sampling and Non-Response situations utilizing R / SAS Softwares” was undertaken with following objectives:

1) To compare Rank set sampling scheme with other sampling techniques (Simple Random sampling, Systematic Sampling) and study the impact of these sampling techniques on estimates of simple regression models.
2) To study the impact of Rank set sampling on the estimators of population mean in Stratified sampling.

3) To develop class of Ratio estimators using Rank set sampling and their comparison with the classical Ratio estimators.

4) To study Rank set sampling in situations of Non-Response, while considering the problems of allocation.

5) To utilize R/ SAS softwares in the above methodology and to develop functions based on these softwares to study the allied data sets.

1.10 Review of literature

Ranked set sampling was first described as a method to increase the precision of estimated yield without the bias of researcher-chosen 'representative' samples (McIntyre 1952). He contended that the effectiveness of the method was dependent upon the information gained by ranking. RSS has been used to estimate pasture yield (McIntyre 1978), shrub phytomass (Martin *et al.,* 1980), mass herbage in a paddock (Cobby *et al.,* 1985) in order to achieve observational economy and increased precision over simple random sampling (SRS). In practice, ranking is bound to be performed with some error. The statistical theory of ranking error was developed by Takahasi & Wakimoto (1968). Dell & Clutter (1972) showed that the RSS estimator remains unbiased in the presence of unbiased ranking error and that when ranking is completely random the RSS estimator provides better precision than simple random sample estimator. The correlation between the concomitant variable and the variable of interest is proportional to ranking error. RSS was extended to ranking on a concomitant variable by Stokes (1977).

Rank set sampling has wide applications in regression modelling as it is used to improve parameter estimation in simple linear regression model. Many authors used RSS technique in regression analysis. Patil *et al.* (1993) compared the RSS sample and SRS sample in relation to the concomitant variable and the regression estimate. Muttlak (1995) used RSS to estimate the parameters of simple linear regression model treating the regressor as a constant. Gilbert (1995) recommended it for environmental research questions such as estimating

plutonium soil concentrations. Nussbaum and Sinha (1997) discussed the problem of quality testing reformulated gasoline with reference to RSS. Samawi (1997) proposed a regression-type estimator based on RSS. They demonstrated that this estimator is always more efficient than the regression estimator using SRS. Chen (2001) did an extensive study on the properties of regression type estimates. Ranked samples are proven to be more efficient than random samples. Hani (2002) suggested that using ranked samples increases the precision of regression analysis and argued that all residual analysis methods for the model diagnostics remain valid when ranked samples are used. Chen and Wang (2004) studied the optimal RSS for the regression analysis.

The effects of using RSS in multiple linear regression analysis were considered by Yaprak (2007) in terms of estimation of regression model parameters. Yaprak suggested that the estimators obtained based on RSS are more efficient than the estimators based on SRS when the sample size is small. Many forms of ranked set samples have been introduced recently for estimating the population mean and other parameters. For multiple characteristics estimation, Walid *et al.* (2011) suggested extreme ranked set sample and argued that this sampling scheme will prove to be more efficient than bivariate simple random sample and the usual univariate rank set sample for estimating population means. He suggested a simulation study to compare the efficiency of the estimators and showed that extreme ranked set sample gives unbiased and more efficient estimators than those obtained by using bivariate simple random sample and univariate ranked set sample, using the same number of quantified observations.

Murray (2000) applied RSS for estimating the performance of spray deposits on leaves of apple trees and contended that despite errors in ranking, RSS improved the precision of estimating the mean of percentage of upper leaf surface covered with deposit and the total deposit on the upper surface of the leaf as compared to SRS. Zehua (2000) suggested that maximum likelihood estimates

(MLE) based on RSS are always more efficient than their counterparts based on SRS. For comparable sample sizes, the RSS procedure results in more accurate parameter estimators than simple random sampling (SRS). Husby *et al.* (2005) used survey data on crop production in Ohio, and considered reduction in standard error for a variety of sample sizes while using several different auxiliary variables for ranking.

Zehua (2008) proposed schemes for the assignment of experimental units that may greatly improve the efficiency of the comparison in such situations. The proposed schemes based on general ranked set sampling were considered by him and relative efficiency and cost-effectiveness of the proposed schemes were studied concluding that the proposed scheme based on ranking provides results which are always more efficient than the traditional SRS when the total cost is same. Estimating a monitoring survey abundance index would be more efficient if the sampling sites were selected based on the information from previous surveys or catch rates of the fishery research. You (2009) suggested two practical examples from fishery research that RSS incorporates information on concomitant variables that are correlated with the variable of interest in the selection of samples, to demonstrate the approach: site selection for a fishery-independent monitoring survey in the Australian northern prawn fishery (NPF) and fish age prediction by simple linear regression modeling a short-lived tropical clupeoid. Both the strategies were based on RSS. The relative efficiencies of the new designs were derived analytically and sampling strategies were developed based on the idea of ranked set sampling (RSS) to increase efficiency and reduce the cost of sampling in fishery research.

Minzhu (2005) provided the estimation of quantiles from data sets generated with RSS. Based on this data he proposed new estimators and argued that such estimators have smaller asymptotic variances for all distributions Bouza (2009) proposed a modified ratio estimators of the population mean of the variable of interest involving the first or third quartiles of an auxiliary variable

that is correlated with the variable of interest. These estimators were investigated under simple random sampling (SRS) and ranked set sampling (RSS) method and these estimators were found to be approximately unbiased and the RSS estimators were more efficient than those based on SRS method for the same quartile and sample size. Many sampling methods have been suggested for estimating population median or second quartile in a situation when sampling units in a study can be ranked easily than quantified. Kamarulzaman (2011) illustrated the superiority of RSS over SRS through simulation studies.

Rank set sampling procedures are used for obtaining the sub-sample from the set of non-respondents. Gaajendra and Bouza (2012) suggested that the first visit may serve for ranking accurately the sub-sampled non-respondents and applied two variations of rank set sampling (RSS), i.e. extreme-RSS and median-RSS for developing estimators of the population mean. Their expected variances and biases were obtained using Monte Carlo experiments. Various sampling strategies have been applied to study the behavior of non-responses.

Chapter – 2

UTILIZATION OF R AND SAS SOFTWARES IN SAMPLE SURVEY DATA

2.1 Introduction to R and SAS softwares

The proposed investigation envisages utilization of R software and SAS packages for statistical and graphical studies, in view of the following important and distinguishing features of these softwares.

R-software is an integrated suit of software for data manipulation, calculation, and graphical display. It has an effective handling and storage facility. It has a large number of functions for data analysis. It contains numerous graphical facilities to display either directly at computer screen or as hard copy. It has its own programming language, which is very effective and simple. It has the feature of extendibility, thus a programme written in C, C++ or FORTRAN can easily be called in this software. R- Software is similar to S-PLUS software and both are implementations of S language, developed at Bell laboratories USA, which is birth place of C language and unix operating system. Three fundamental books written by Becker *et al.* (1988), Venables and Repley (2004) are of immense use for understanding this software's. Khan and Mir (2005) discuss in detail the application of R-software in agricultural data analyses. One of the important feature of R- software is that it is an Open Source and freely available on website http://cran-project.org. R language is essentially a functional language for all practical purposes of data analysis and graphics. Preliminary data analysis in case of Pinus data was carried out in R software.

Some of the important features of R and S-plus softwares are:

- These are an integrated suit of software for data manipulation, calculation, and graphical display.
- They have an effective handling and storage facility.

- They have suit of operators for calculations on arrays, in particular matrices and data frames.
- They have a large number of functions for data analysis.
- They contain numerous graphical facilities to display either directly at computer screen or as hard copy.
- They have their own programming language, which is very effective and simple.
- They have the feature of extendibility, thus a programme written in C, C++ or Fortran can easily be called in these softwares.

SAS is a sophisticated computer package containing many components. Originally the letters SAS stand for Statistical Analysis System. SAS software provides comprehensive statistical tools for a wide range of statistical analyses, including analysis of variance, categorical data analysis, cluster analysis, multiple imputation, multivariate analysis, nonparametric analysis, power and sample size computations, psychometric analysis, regression, survey data analysis, and survival analysis. SAS has also been utilized in studies, like, nonlinear mixed models, generalized linear models, correspondence analysis, and robust regression. For over three decades, SAS software has been used by programmers, analysts and scientists to manipulate and analyze data. SAS (Statistical Analysis System) software is comprehensive software which deals with many problems related to Statistical analysis, Spreadsheet, Data Creation, Graphics, etc. It is a layered, multivendor architecture. Regardless of the difference in hardware, operating systems, etc., the SAS applications look the same and produce the same results. The three components of the SAS System are Host, Portable Applications and Data. Host provides all the required interfaces between the SAS system and the operating environment. Functionalities and applications reside in Portable component and the user supplies the Data.

Windows of SAS

1. Program Editor : All the instructions are given here.
2. Log : Displays SAS statements submitted for execution and messages
3. Output : Gives the output generated

Rules for SAS Statements

1. SAS program communicates with computer by the SAS statements.
2. Each statement of SAS program must end with semicolon (**;**).
3. Each program must end with run statement.
4. Statements can be started from any column.
5. One can use upper case letters, lower case letters or the combination of the two.

Basic Sections of SAS Program

1. DATA section
2. CARDS section
3. PROCEDURE section

Data Section

Details of data section are given below :

Data value

A single unit of information, such as name of the specie to which the tree belongs, height of one tree, etc.

Variable

A set of values that describe a specific data characteristic e.g. diameters of all trees in a group. The variable can have a name upto a maximum of 8

characters and must begin with a letter or underscore. Variables are of two types:

Character Variable

It is a combination of letters of alphabet, numbers and special characters or symbols.

Numeric Variable

It consists of numbers with or without decimal points and with + or - ve signs.

Observation

A set of data values for the same item i.e. all measurement on a tree. Data section starts with Data statements as DATA NAME (it has to be supplied by the user);

Input Statements

Input statements are part of data section. This statement provides the **SAS** system the name of the variables with the format, if it is formatted.

List Directed Input

- Data are read in the order of variables given in input statement.
- Data values are separated by one or more spaces.
- Missing values are represented by period (.).
- Character values are followed by $ (dollar sign).

Univariate study in case of Apple data was carried out in SAS.

2.2 Preliminary study of data structure

Two data sets were considered for the present study To have a comprehensive look into the data its graphic as well as numerical summary is required. This job can be accomplished using the functions of R which are

specially meant for meeting this requirement. These functions will be discussed in what follows. Summary features of the data consist of the two main aspects; (i) Numerical summary (ii) Graphical summary.

2.2.1 Numerical summary

Numerical summary consists of summary features of the characters (variables) of breeding data. The main features are minimum, maximum, first quartile, second quartile (median), mean and third quartile. These summary features are reported for each variable of apple data. All summary features can be obtained by making use of the function `summary ()` of R software packages. A general format of the function is `summary (data)`

Where data stands for data object of which summary is required. This data may be a vector representing a single character. A better alternative is data frame which is a two dimensional array consisting of rows and columns, where rows represent number of observations and columns represent number of variables (characters). The data used for the present study is pinus data with 275 rows and 2 columns. Names of these characters can be obtained using the function `names ()` of R software packages.

```
> names(pinus)
[1] "dbh"    "height"
```

Now the data set pinus introduced in the above is analyzed to see its summary features.

```
> summary(pinus)
      dbh              height
 Min.   :  2.20   Min.   : 0.90
 1st Qu.:  6.85   1st Qu.: 4.80
 Median : 14.60   Median : 9.31
 Mean   : 21.45   Mean   :15.67
 3rd Qu.: 33.25   3rd Qu.:20.12
 Max.   :219.00   Max.   :71.77
```

2.2.2 Graphical summary

A graphical summary of the data set provides its visible picture which can be easily understood by a common man.

`Boxplot()`

`Boxplot()` is meant for comprehensive presentations of data. It shows centre as well as spread of a distribution. Thus, variability can be depicted along with point of centrality. A line in the box is placed at the median value. The width of the box is equal to inter quartile range IQR, which is the difference between the third and first quartile. The width of the box shows the variability present within the character. Wiskers are lines on both sides of the box that extend the edge of the box to either sides of the extreme value or to a distance of $1.5 \times IQR$ from the median whichever is less, (e.g., Khan and Mir (2005)). Box plot of data is obtained by using `boxplot()` function available in and R. Argument to this function is data object whose box plot is required. The general format of the function is `Boxplot(data).`

To get the box plot of `"dbh"` `"height"` of `pinus` data the commands are given Fig. 2.

```
op=par(mfrow=c(1,2))
# par can be used to set or query graphical parameters.
Parameters can be set by specifying them as arguments
to par in tag = value form, or by passing them as a
list of tagged values.
> boxplot(pinus$dbh,ylab="cm",main="dbh")
> boxplot(pinus$height,ylab="mt",main="height")
```

Quantile–Quantile plot (`QQ-Plot`)

The quantile quantile plot is a plot of one set of quantiles against another set of quantiles. There are two main forms of `QQ-Plot` these are `qqnorm` and `qqline`. The most frequently used form i.e., `qqnorm` checks whether data set comes from a particular normal distribution . In this type of plot one set of

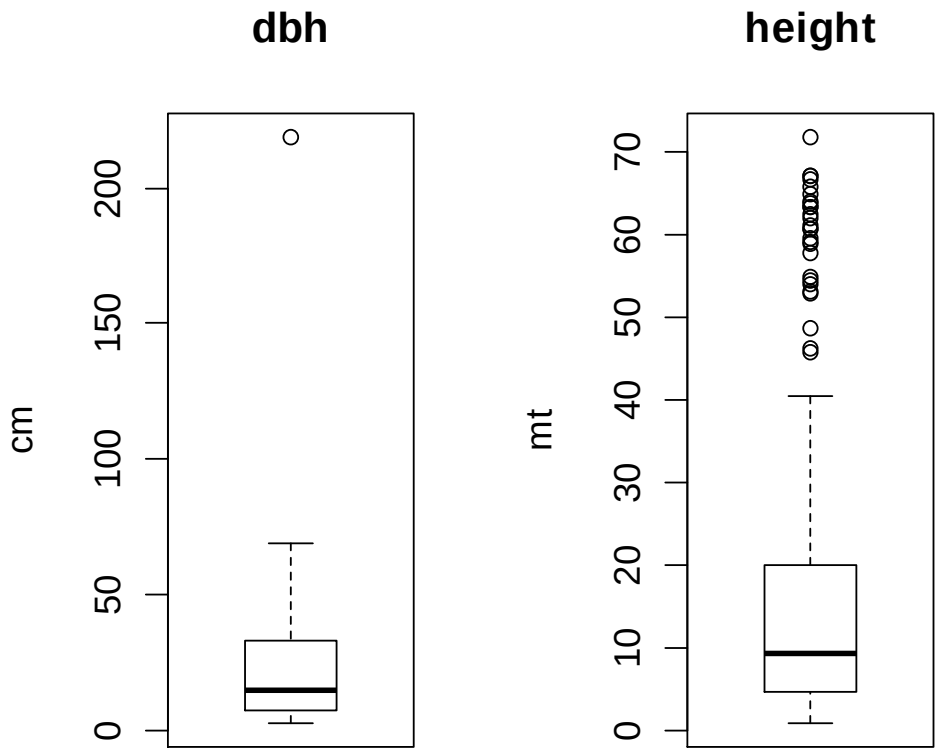

Fig. 2 : Box plot of `dbh,height`

quantiles consists of the ordered set of data values and the other set of quantiles are from normal distribution. If the points in the plot cluster along the straight line the data set probably has the normal distribution. The second form i.e., `qqline` fits a line through a normal `qqplot` to check the distribution shape. The general format of these functions is

```
qqnorm(data)

qqline(data)
```

To get `qqnorm` and `qqline` of the each character of pinus data the commands are given in Fig. 3.

```
> op=par(mfrow=c(1,2))
> qqnorm(pinus$dbh,main="dbh")
> qqline(pinus$dbh,main="dbh")
> qqnorm(pinus$height,main="height")
> qqline(pinus$height,main="height")
```

2.2.4 Regression analysis of the data

A linear multiple regression models was fitted for the Pinus data set. Multiple regression models is an extension of simple regression model. In such a case more than one regressor variables are involved. The model is specified by the linear equation,

$$y_i = \beta_0 + \beta_1 X_i + ... + \beta_k X_k + e_i, \quad i = 1,...,n$$

Where $\beta's$ are regression coefficients and $e_i's$ are distributed normally with mean zero and variance σ^2. This model can be extended to a general linear model defined as

$$y = X\beta + e$$

Which is equivalent to

$$y_i \sim N(\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}, \sigma^2), \quad i = 1,...,n$$

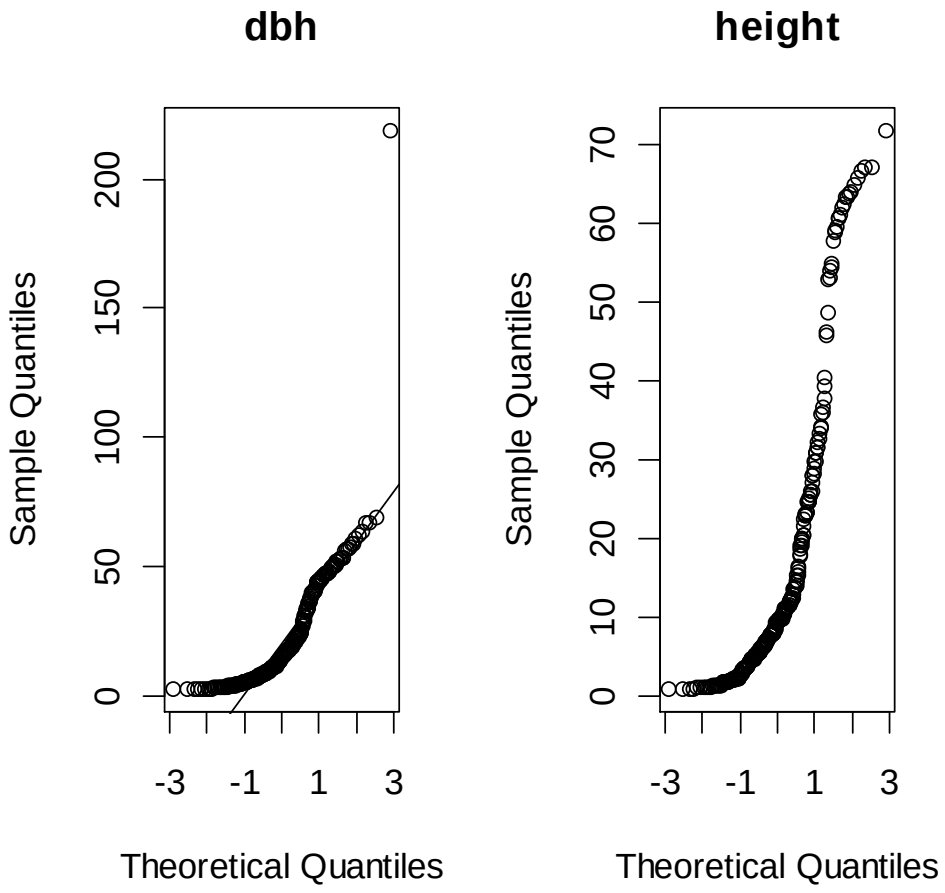

Fig. 3 : QQplot of the each character of Pinus data

For regression analysis, `lm` function of R software is used. Its main arguments are `formula` and data. Formula provides relationship between response and regressor variables while data provides data frame in which data on response and regressor variables are made available. Thus, general form of *lm* function is as follows:

```
reg<–lm(formula, data=data.frame)
```

For *pinus* data in which if 'height' is a response variable and 'dbh'is regressor variable then this model will be fitted as

```
reg<-lm(height~dbh,data=pinus)
# lm is used to fit linear models. It can be used to
carry out regression, single stratum analysis of
variance and analysis of covariance (although aov may
provide a more convenient interface for these).
```

> summary(reg)

```
# summary is a generic function used to produce result
summaries of the results of various model fitting
functions.

lm(formula = height ~ dbh, data = pinus)

Residuals:
       Min         1Q     Median         3Q        Max
-120.91078   -3.56472   -2.14159    0.01044   35.58641

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.83774    1.00261    2.83  0.00499 **
dbh          0.59810    0.03347   17.87  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Residual standard error: 11.61 on 273 degrees of
freedom
```

```
Multiple R-squared: 0.5391,     Adjusted R-squared:
0.5374
F-statistic: 319.3 on 1 and 273 DF,  p-value: < 2.2e-16
```

> anova(reg)

```
# Compute analysis of variance (or deviance) tables for
one or more fitted model objects.

Analysis of Variance Table

Response: height
           Df Sum Sq Mean Sq F value    Pr(>F)
dbh         1  43023   43023  319.33 < 2.2e-16 ***
Residuals 273  36781     135
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1
```

2.2.8 Validity of assumptions

The validity of the fitted multiple regression model for the pinus data is checked graphically. The assumption on the errors being i.i.d. normal random variables translates into the residuals being normally distributed. They are not independent as they add to and their variance is not uniform, but they should show no serial correlations. We can test for normality with histograms, boxplots and normal plots. We can test for correlations by looking if there are trends in the data. This can be done with plots of the residuals vs. time and order. We can test the assumption that the errors have the same variance with plots of residuals vs. time order and fitted values. Following functions of R are used for checking the validity of assumptions (Fig. 4).

```
> op=par(mfrow=c(2,2))

> plot(reg)

(# par can be used to set or query graphical
parameters. Parameters can be set by specifying them as
```

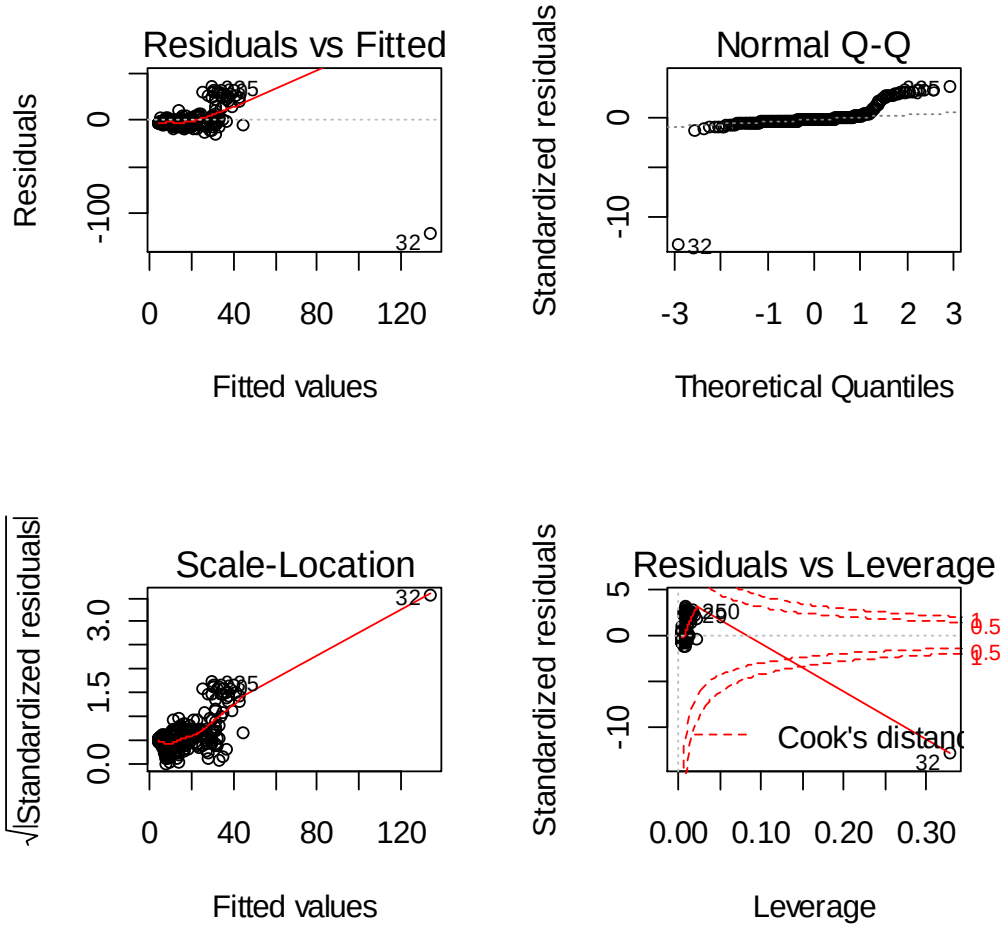

Fig. 4 : Graphs of fitted regression model of Pinus data

```
arguments to par in tag = value form, or by passing
them as a list of tagged values)
```

Residuals vs. fitted : This plots the fitted ($\hat{y}$) values against the residuals. Look for spread around the line y = 0 and no obvious trend.

Normal qqplot : The residuals are normal if this graph falls close to a straight line.

Scale-Location : This plot shows the square root of the standardized residuals. The tallest points are the largest residuals.

Cook's distance : This plot identifies points which have a lot of influence in the regression line.

2.2.10 Confidence interval

```
> confint(reg)

# Computes confidence intervals for one or more
parameters in a fitted model.

                2.5 %    97.5 %
(Intercept) 0.8639181 4.8115701
dbh         0.5322046 0.6639875
```

2.2.11 Correlation

```
> Cor (pinus)

#var, cov and cor compute the variance of x and the
covariance or correlation of x and y if these are
vectors. If x and y are matrices then the covariances
(or correlations) between the columns of x and the
columns of y are computed.

          dbh    height
dbh    1.0000000 0.7342398
height 0.7342398 1.0000000
```

2.2.11 Univariate study of Apple data using SAS

The UNIVARIATE Procedure under SAS takes the following arguments;

```
data apple"
input  yield Area   Trees  Bearing;
cards;
23.52  2.55   255    196
19.32  2.1    210    161
16.52  1.8    180    138
20.76  2.25   225    173
18     1.95   196    150
24.84  2.7    270    207
11     1.2    120    92
9.6    1.05   105    80
13.8   1.5    150    115
16.5   1.8    182    138
.
.
.
.
.
.
.
.
.
.
.
.
.
9.96   1.2    109    83
13.44  1.5    146    112
21     2.25   228    175
20.04  2.25   217    167
6.24   0.75   68     52
;
proc univaraite data = apple;
run;
```

Variable: yield

Moments

N	420	Sum Weights	420
Mean	16.0917143	Sum Observations	6758.52
Std Deviation	6.22854468	Variance	38.7947689
Skewness	0.16762684	Kurtosis	-1.1074731
Uncorrected SS	125011.181	Corrected SS	16255.0082
Coeff Variation	38.7065329	Std Error Mean	0.30392186

Basic Statistical Measures

Location		Variability	
Mean	16.09171	Std Deviation	6.22854
Median	15.28000	Variance	38.79477
Mode	14.04000	Range	25.49000
		Interquartile Range	10.08000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	52.94688	Pr > |t|	<.0001
Sign	M	210	Pr >= |M|	<.0001
Signed Rank	S	44205	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	30.480
99%	28.080
95%	26.400
90%	24.720
75% Q3	20.880
50% Median	15.280
25% Q1	10.800
10%	8.105
5%	6.720
1%	6.000
0% Min	4.990

Variable: yield

Extreme Observations

----Lowest----		----Highest----	
Value	Obs	Value	Obs
4.99	257	28.08	107
5.91	253	28.20	300
6.00	395	28.32	86
6.00	381	28.44	222
6.00	366	30.48	408

Variable: Area

Moments

N	420	Sum Weights	420
Mean	1.84	Sum Observations	772.8
Std Deviation	0.67870437	Variance	0.46063962
Skewness	0.10850838	Kurtosis	-1.0517595
Uncorrected SS	1614.96	Corrected SS	193.008
Coeff Variation	36.8861069	Std Error Mean	0.03311738

Basic Statistical Measures

Location		Variability	
Mean	1.840000	Std Deviation	0.67870
Median	1.800000	Variance	0.46064
Mode	1.500000	Range	2.55000
		Interquartile Range	1.20000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	55.55995	Pr > |t|	<.0001
Sign	M	210	Pr >= |M|	<.0001
Signed Rank	S	44205	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	3.30
99%	3.15
95%	3.00
90%	2.70
75% Q3	2.40
50% Median	1.80
25% Q1	1.20
10%	1.05
5%	0.75
1%	0.75
0% Min	0.75

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
0.75	420	3.15	300
0.75	395	3.30	92
0.75	394	3.30	199
0.75	393	3.30	222
0.75	382	3.30	408

Variable: Trees

Moments

N	420	Sum Weights	420
Mean	174.990476	Sum Observations	73496
Std Deviation	67.1602998	Variance	4510.50588
Skewness	0.1676306	Kurtosis	-1.1033162
Uncorrected SS	14751002	Corrected SS	1889901.96
Coeff Variation	38.3794029	Std Error Mean	3.27708708

Basic Statistical Measures

Location		Variability	
Mean	174.9905	Std Deviation	67.16030
Median	166.5000	Variance	4511
Mode	120.0000	Range	265.00000
		Interquartile Range	108.00000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------	
Student's t	t	53.39818	Pr > |t|	<.0001
Sign	M	210	Pr >= |M|	<.0001
Signed Rank	S	44205	Pr >= |S|	<.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	330.0
99%	304.0
95%	287.0
90%	268.0
75% Q3	226.0
50% Median	166.5
25% Q1	118.0
10%	90.0
5%	74.0
1%	68.0
0% Min	65.0

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
65	395	304	107
65	381	306	300
65	366	307	86
65	314	308	222
68	420	330	408

Variable: Bearing

Moments

N	420	Sum Weights	420
Mean	134.171429	Sum Observations	56352
Std Deviation	51.7450067	Variance	2677.54572
Skewness	0.1714324	Kurtosis	-1.1005816
Uncorrected SS	8682720	Corrected SS	1121891.66
Coeff Variation	38.5663381	Std Error Mean	2.52489779

Basic Statistical Measures

Location		Variability	
Mean	134.1714	Std Deviation	51.74501
Median	128.0000	Variance	2678
Mode	92.0000	Range	204.00000
		Interquartile Range	83.50000

Tests for Location: Mu0=0

Test	-Statistic-		-----p Value------
Student's t	t	53.13935	Pr > |t| <.0001
Sign	M	210	Pr >= |M| <.0001
Signed Rank	S	44205	Pr >= |S| <.0001

Quantiles (Definition 5)

Quantile	Estimate
100% Max	254.0
99%	234.0
95%	220.0
90%	206.0
75% Q3	173.5
50% Median	128.0
25% Q1	90.0
10%	67.5
5%	56.0
1%	50.0
0% Min	50.0

Extreme Observations

----Lowest----		----Highest---	
Value	Obs	Value	Obs
50	395	234	107
50	381	235	300
50	366	236	86
50	314	237	222
50	257	254	408

2.3 Development of computer programmes for the present study

One of the important feature of R- software is that it is an Open Source and freely available on website http://cran-project.org. R language is essentially a functional language for all practical purposes of data analysis and graphics. However, in case some specific situations data analyst is forced to develop his own functions according to his requirements. Consequently, few functions have

been developed according to the requirements of this study. Detailed codes are available inside text; however, a brief summary of these functions is reported here. A convention adopted throughout the thesis is that all the commands and output of the software are typed in Courier New font. Computer programming in R-software is essentially development of functions which are to be executed in the data analysis. Four functions have been developed.

These are `drss(m,k)`,`varwts(n,h)`,`makeAlloc(n,m)` and `ratio.est(n,N(x,y))`

```
drss <- function(m,k)
n<-55
id<-1:(11*5^2)
s1<-sample(id,n^2)
s2<-id[-s1]

block<-rep(1:(5*n),each=n)
d<-data.frame(block)
for(i in 1:n){
d$rss<-ifelse(d$block==i,sample(s1,n),0)
#s1<-s1[-which(s1==d$rss)]
}
for(i in 1:n){
d$rss<-ifelse(d$block==i,sample(s2,n),0)
#s2<-s2[-which(s2==d$rss)]
}
d <- transform(d,rss = ifelse(d[,"block"]<=n,
sample(s1), sample(s2)))
d <- transform(d,block.id = rep(1:n))
d <- transform(d,rss = ifelse(d[,"block"]<=n,
sample(s1), sample(s2)))
d <- dcast(d,block ~ block.id, value = rss)
```

```
R commands for ratio estimation
ratio.est <- function(x, y, mux = NA, N = NA) {
# estimate of a ratio and ratio estimate of population
mean and total.
# x is auxiliary variable, y is response, mux is
population mean
# of x (xbar is used if no value is given),
# N is population size (assumed infinite if no value
given),
```

```
if(length(x) != length(y)) stop("x and y must be same
length")
n <- length(x)
fpc <- 1
if(!is.na(N))
fpc <- (N - n)/N
r <- sum(y)/sum(x)
sr2 <- (1/(n - 1)) * sum((y - r * x)^2)
if(is.na(mux)) mx <- mean(x) else mx <- mux
cat("r=", r, " SE=", sqrt((fpc * sr2)/(mx^2 * n)),
"\n")
if(!is.na(mux))
cat("mu-hat=", r * mux, " SE=", sqrt((fpc * sr2)/n),
"\n")
if(!is.na(N) & !is.na(mux))
cat("tau-hat=", N*r * mux, " SE=", N*sqrt((fpc *
sr2)/n), "\n")
```

R commands for rank set sampling under stratification

```
# Variance of RSS under stratification mean estimator
with 3 strata
varwts <- rep((strata.pops/setsizes), setsizes)
vars <- c()
for(i in 1:3)
{
m <- setsizes[i]
v <- varOS[[m]]
vars1 <- strata.vars[i] * v
vars <- c(vars, vars1)
}
if(total) out <- sum(varwts^2 * vars/alloc) else
{
z <- varwts*sqrt(vars)
optkrh <- round(n*z/sum(z))
out <- sum(varwts^2 * vars/optkrh)
}
return(out)
}
```

R commands for allocation under non response in RSS

```
# This function makes allocations of sets for strata
and of
# observations of sets
makeAlloc <- function(n, M)
```

```
{
A <- combn(n-1, M-1)
out <- matrix(nrow=M, ncol=ncol(A))
out[1, ] <- A[1, ]
for(i in 2:nrow(A))
{
out[i, ] <- A[i, ] - A[(i-1), ]
}
out[M, ] <- n - A[M-1, ]
dimnames(out) <- NULL
return(out)
}
selectNeighbor <- function(setsizes, alloc)
{
M <- sum(setsizes) # total of set sizes
H <- length(setsizes) # number of strata
m <- rep(setsizes, setsizes) # lists set size for each
k value
s <- rep(1:H, setsizes) # lists stratum for each k
value
d <- 1:M
allocNew <- alloc
setsizeNew <- setsizes
# Do not remove observation from stratum w/ 1 set and 1
obs
d_exclude <- d[alloc == 1 & m == 1]
if(length(d_exclude) == 0) {sub_x <- d} else {sub_x <-
d[-d_exclude]}
# Sample stratum-rank from which to remove 1 obs
subone <- sample(x=sub_x, size=1)
allocNew[subone] <- alloc[subone] - 1
# Sample space of places to add; do not consider
stratum-rank
# from where obs removed; values in the 100's mean add
a new set
add_x <- c(d[-subone], 101:(100 + H))
# This happens if there was only 1 obs in a stratum-
rank, and this
# obs was removed; equivalent to removing a set
# We do not want to add obs back to this set, wo remove
from sample space
if(allocNew[subone] == 0)
{
s_exclude <- s[allocNew == 0]
add_x <- c(d[-subone], (101:(100 + H))[-s_exclude])
```

```
setsizeNew[s_exclude] <- setsizes[s_exclude] - 1
}
addone <- sample(x=add_x, size=1)
if(addone %/% 100 != 1) {allocNew[addone] <-
alloc[addone] + 1} else
{
c <- addone %% 100
setsizeNew[c] <- setsizes[c] + 1
allocNew <- c(allocNew[s <= c], 1, allocNew[s > c])
}
allocOut <- allocNew[allocNew != 0]
out <- list(kt = allocOut, mt = setsizeNew)
return(out)
}
# Parameters for simulated annealing
strata.means <- c(22.11, 12.09, 7.62)
strata.vars <- c(6.89, 7.63, 4.14)
H <- 3
n <- 50
load("varOS.Rdata")
source("importantFuncs.r")
runs <- list(NA, NA, NA, NA, NA, NA, NA, NA, NA, NA)
for(b in 1:10)
{
# Randomly select initial allocation
M0 <- sample(size=1, 1:n)
totN <- enum(n, H)
e <- seq(1, totN, by=totN/16)
e[17] <- totN + 1
samp <- rep(NA,)
for(i in 1:16)
{
samp[i] <- sample(x=e[i]:(e[i+1]-1), size=1)
}
s <- sample(x=samp, size=1)
x <- 0:(n-H)
N <- choose(n-1, n-H-x)*choose(H+x-1, x)
lowerN <- c(0, cumsum(N))
x0 <- max(x[s > lowerN])
rem <- s - lowerN[x0 + 1]
i0 <- rem%/%choose(n-1, n-H-x0)
j0 <- rem%%choose(n-1, n-H-x0)
if(j0 == 0) i0 <- i0 - 1
k0 <- makeAlloc(n, x0+H)[,j0]
m0 <- makeAlloc(x0+H, H)[,i0]
```

```
var0 <- objec(m0, k0)
out <- matrix(c(m0, NA, var0, 1), nrow=1)
colnames(out) <- c("m1","m2","m3","p","var","keep")
runs[[b]] <- list(OUT=out, K=list(k0))
t0 <- 0.1
y <- 2
stageLength <- rep(c(22,33,44), rep(3,2))
for(j in 1:length(stageLength))
{
for(i in 1:stageLength[j])
{
if(i == 1 && j==1)
{
t <- t0
mOld <- m0
kOld <- k0
varOld <- var0
}
newAlloc <- selectNeighbor(mOld, kOld)
mNew <- newAlloc$mt
kNew <- newAlloc$kt
varNew <- objec(mNew, kNew)
p <- min(1, exp((varOld-varNew)/t))
keep <- rbinom(1, 1, prob=p)
if(keep)
{
mOld <- mNew
kOld <- kNew
varOld <- varNew
}
runs[[b]]$OUT <- rbind(runs[[b]]$OUT, c(mNew, p,
varNew, keep))
runs[[b]]$K[[y]] <- kNew
y <- y + 1
if(y%%1000 == 0) print(y)
}
t <- 0.9*t
}
}
```

Chapter – 3

COMPARISON OF RANK SET SAMPLING SCHEME WITH OTHER SAMPLING TECHNIQUES BASED ON SIMPLE REGRESSION MODELS

In this chapter estimation of regression model and regression estimates of the mean of the response variable in context to rank set sampling are discussed, also it is shown mathematically that for estimation of mean of the response variable, regression estimates based on rank set sampling are better than simple random sampling sample mean as long as response variable and auxiliary variable are correlated. Also in this chapter simple linear regression models have been considered with respect to samples taken from the identified sampling techniques including RSS. In case the study pertains to establish the relationship between one dependent variable and one independent variable, it is simple regression, where as in the functional relationship between dependent and independent variable if the power of independent variable is one then it is the case of simple linear regression, otherwise the case is non-linear.

Simple regression model proceeds as follows:

$$Y = \beta_0 + \beta_1 X + \varepsilon ,$$

Where Y is dependent variable, X is independent variable or ranking or auxiliary variable, ε is a random disturbance/error term. The parameters β_1 is called regression coefficient. The parameter β_0 is the intercept of the regression line. β_0 gives the mean of Y when X = 0. The performance of RSS will be based on the efficiency of estimators relative to the identified sampling schemes as mentioned above. The relative efficiency (RE) for the estimated model based on RSS will be computed according to the following expression :

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

Where $MSE(\hat{\mu}_{d.\mathrm{Re}g})$ is the MSE (mean square error) of estimated regression model based on the sampling schemes (simple random sampling and systematic sampling) and $MSE(\hat{\mu}_{RSS.\mathrm{Re}g})$ is the MSE of the estimated regression model based on RSS. The method of estimation used in this chapter is the ordinary least squares method.

3.1 Regression analysis based on RSS with concomitant variables

Suppose that the variable Y and the concomitant variables follow a linear regression model, i.e.,

$$\mathrm{Y} = \alpha + \beta' X + e, \tag{3.1.1}$$

where β is a vector of unknown constant coefficients, and e is a random variable with mean zero and variance σ_e^2 and is independent of X.

Suppose that an RSS with certain ranking mechanism is implemented in m cycles. In a typical cycle i, for r =1,..., k, a simple random sample of k units with latent values (Y_{1ri}, X_{1ri}),..., (Y_{kri}, X_{kri}) is drawn from the population. The values of the X's are all measured. The k sampled units are ranked according to the ranking mechanism. Then the Y value of the unit with rank r is measured. The X and Y values of the unit with rank r are denoted, respectively, by $X_{[r]i}$ and $Y_{[r]i}$. In the completion of the sampling, we have a data set as follows :

$$\begin{gathered} (Y_{[1]i}, X_{[1]i}, X_{11i},..., X_{1ki}), \\ ... \\ (Y_{[k]i}, X_{[k]i}, X_{k1i},..., X_{kki}), \\ i = 1,..., m. \end{gathered} \tag{3.1.2}$$

Note that the setting above includes both the multi-layer RSS and the adaptive RSS. In fact, we can treat the ranks as indices of strata or categories. Thus, in the two-layer RSS, we can represent the double ranks by a single index in an appropriate manner. Based on this, we consider in this section two problems:

(i) the estimation of the regression coefficient, (ii) the estimation of μ_Y , (mean of *Y)*.

3.2 Estimation of regression coefficient with RSS

Denote by *G(y/x)* the conditional distribution of *Y* given *X*. Let {(Y_1, X_1), ..., (Y_n, X_n)} be a simple random sample from the joint distribution of *Y* and *X*. Suppose (R_1,...,R_n) is a random permutation of (1,...,*n*) determined only by (X_1,..., X_n).

Lemma 3.2.1 The random-permutation-induced statistics $Y_{R1},...,Y_{Rn}$ are conditionally independent given $(X_1,..., X_n)$ with conditional distribution functions $G(\cdot|X_{R1}),...,G(\cdot|X_{Rn})$ respectively

Then under simple regression model, we have

$$Y_{[r]i} = \alpha + \beta^T X_{[r]i} + \varepsilon_{ri}, \qquad (3.2.2)$$

$r = 1 ,..., k;\ i = 1,...,m,$

where ε_{ri} are independent identically distributed and are independent of $X_{[r]i}$. Thus, we can estimate α and β by least squares method as given below :

Let

$$\overline{X}_{RSS} = \frac{1}{mk} \sum_{r=1}^{k} \sum_{i=1}^{m} X_{[r]i}, \overline{Y}_{RSS} = \frac{1}{mk} \sum_{r=1}^{k} \sum_{i=1}^{m} Y_{[r]i},$$

$X_{RSS} = (X_{[1]1},..., X_{[1]m},..., X_{[k]1},..., X_{[k]m})$

$Y_{RSS} = (y_{[1]1},...,y_{[1]m},...,y_{[k]1},...,y_{[k]m})$

The least squares estimates of α and β based on (3.2.2) are then given, respectively, by

$$\hat{\sigma}_{RSS} = \overline{Y}_{RSS} - \hat{\beta}'_{RSS} \overline{X}_{RSS}, \qquad (3.2.3)$$

$$\hat{\beta}_{RSS} = [X'_{RSS} (I - \frac{11'}{mk}) X_{RSS}]^{-1} X'_{RSS} (I - \frac{11'}{mk}) Y_{RSS} \qquad (3.2.4)$$

It is obvious that $\hat{\sigma}_{RSS}$ and $\hat{\beta}_{RSS}$ are unbiased. Since both $\hat{\sigma}_{RSS}$ and $\hat{\beta}_{RSS}$ are of the form of smooth-function-of-means, as estimates of α and β, they are, at least asymptotically, as good as their counterparts based on an SRS, that is, their asymptotic variances at the order $O(\frac{1}{mk})$ are smaller than or equal to those based on an SRS. More specific conclusions can be made from the explicit expressions of the variances of $\hat{\sigma}_{RSS}$ and $\hat{\beta}_{RSS}$. These variances can be derived as

$$Var(\hat{\alpha}_{RSS}) = \alpha_e^2 E\left[\frac{1}{mk} + [\overline{X}_{RSS}^T (X'_{RSS} (I - \frac{11'}{mk}) X_{RSS}]^{-1} \overline{X}_{RSS}\right]$$

$$Var(\hat{\beta}_{RSS}) = \alpha_e^2 E\left[[X'_{RSS} (I - \frac{11'}{mk}) X_{RSS}]^{-1}\right]$$

where the expectations are taken with respect to the distribution of the *X*'s. Let Σ denote the variance-covariance matrix of *X*. We have

$$\frac{1}{mk} X'_{RSS} (I - \frac{11'}{mk}) X_{RSS} \to \sum,$$

In order to get the asymptotic expressions of the variances above at the order $O(\frac{1}{mk})$, we can replace $[X'_{RSS} (I - \frac{11'}{mk}) X_{RSS}]^{-1}$ *by* $\frac{1}{mk}\sum^{-1}$. Then, it is easy to see that the asymptotic variances at the order $O(\frac{1}{mk})$ are the same as those of their counterparts based on an SRS, which implies that, in the estimation of the regression coefficient, RSS and SRS are asymptotically equivalent.

3.3 Regression estimate of the mean of response variable with RSS

Let

$$\overline{X}_T = \frac{1}{mk^2} \sum_{r=1}^{k} \sum_{j=1}^{k} \sum_{i=1}^{m} X_{rji}.$$

We can define another estimate of μ_Y rather than $\overline{Y}_{RSS}$. The estimate is called the RSS regression estimate and is defined as

$$\overline{\mu}_{RSS.REG} = \hat{Y}_{RSS} + \hat{\beta}'_{RSS}\,(\overline{X}_T - \overline{X}_{RSS}) \qquad (3.3.1)$$

The RSS regression estimate of μ_Y is unbiased and its variance can be obtained as

$$Var\,(\hat{\mu}_{RSS.REG}) = \frac{\alpha_e^2}{mk}\{1 + \nabla_{RSS}\} + \frac{1}{mk^2}\beta'\Sigma\beta, \qquad (3.3.2)$$

where

$$\Delta_{RSS} = E[mk\,(\overline{X}_T - \overline{X}_{RSS})'[X'_{RSS}\,(I - \frac{11'}{mk})\,X_{RSS}] - 1(\overline{X}_T - \overline{X}_{RSS})]$$

If the ranked set sample is replaced by a simple random sample, we get an SRS regression estimate of μ_Y. The variance of the SRS regression estimate is also of the form (3.3.2) but with Δ_{RSS} replaced by the corresponding quantity Δ_{SRS} defined on the simple random sample. It follows from the asymptotic approximations of Δ_{RSS} and Δ_{SRS}, as long as the ranking mechanism in the RSS is consistent, we always have $\Delta_{RSS} < \Delta_{SRS}$ asymptotically. In other words, the RSS regression estimate is asymptotically more efficient than the SRS regression estimate.

3.4 Bivariate Rank Set Sampling

In this section bivariate ranked set sample (BVRSS), is introduced. A bivariate rank set sampling (BVRSS) can be obtained as follows :

Suppose *(X, Y)* is a bivariate random vector with the *(jpdf)* joint probability density function *f (x, y).*

1. A random sample of size k^4 is identified from the population and randomly allocated into k^2 pools of size k^2 each, where each pool is a square matrix with *k* rows and *k* columns.

2. In the first pool, identify by judgment the minimum value w.r.t. the first characteristic, for each of the *k* rows.

3. For the *k minima* obtained in Step 2, choose the pair that corresponds to the minimum value of the second characteristic, identified by judgment, for actual quantification. This pair, which resembles the label (1, 1), is the first element of the BVRSS sample.

4. Repeat Steps 2 and 3 for the second pool, but in step 3, the pair that corresponds to the second minimum value w.r.t the second characteristic is chosen for actual quantification. This pair resembled the label (1, 2).

5. The process continues until the label (*k*, *k*) is resembled from the $(k^2)^{th}$ (last) pool. The above procedure produces a BVRSS of size k^2 . Thus we have k^2 observations denote by : $(X[i](j), Y(i)[j])$, $i=1,2\ldots k$ and $j=1,2,\ldots k$.

6. The procedure can be repeated *m* times to obtain a sample of size $n = k^2m$ which will be denoted by $(X[i](j)k, Y(i)[j]k)$, $i=1,2\ldots k$ and $j=1,2,\ldots k$, $k=1,2,\ldots r$.

3.4.1 Some notations

The following notations will be used :

$E(X) = \mu_x$,

$E(Y) = \mu_Y$, $\text{var}(X) = \sigma_x^2$,

$\text{var}(Y) = \sigma_Y^2$, $\rho = \dfrac{\text{cov}(X,Y)}{\sigma_X \sigma_Y}$,

$E(X_{[i](j)}) = \mu_{X[i](j)}$, $E(Y_{(i)[j]}) = \mu_{Y(i)[j]}$,

$E(X^2_{[i](j)}) = \mu^{(2)}_{X[i](j)}$,

$E(Y^2_{[i](j)}) = \mu^{(2)}_{Y[i](j)}$,

$Var\ (X_{[i](j)}) = \sigma^2_{X[i](j)}$,

$Var\ (Y_{[i](j)}) = \sigma^2_{Y[i](j)}$,

$\text{cov}\ (X_{[i](j)}, Y_{(i)[j]}) = \sigma_{(X[i](j), y(i)[j])}$,

3.5 Simple linear regression using bivariate rank set sampling

The simple regression model of the two variables Y and X is defined by : $y_{(i)[j]k} = a + \beta X_{[i](j)k} + E_{ijk}$ where a is the model intercept, β is the model slope and E_{ijk} is the random error. The assumptions needed here for the purpose of parameters estimation are; the mean of the error is zero, its variance is finite. Also X_i and E_i are independent. Then the least squares estimates of a and β are respectively given by :

$$\hat{\alpha}_{bvrss} = \overline{Y}_{bvrss} - \hat{\beta}_{bvrss}\,\overline{X}_{bvrss}$$

$$\hat{\beta}_{bvrss} = \frac{\sum_{k.j.i}(X_{[i](j)k} - \overline{X}_{bvrss})(y_{(i)[j]k} - \overline{Y}_{bvrss.}}{\sum_{k.j.i}(X_{[i](j)k} - \overline{X}_{bvrss})^2})\overline{X}_{bvrss}$$

where

$$\overline{X}_{bvrss} = \frac{\sum_{k.j.i} X_{[i](j)k}}{n} \qquad \text{and}$$

$$\overline{Y}_{bvrss} = \frac{\sum_{k.j.i} Y_{[i](j)k}}{n}$$

$$\text{var}\,(\hat{Y}_{(i)[j]k}) = \frac{\sigma_e^2}{n}[1 + E(\frac{(X_{(i)[j]k} - \overline{X}_{bvrss})^2}{S^2_{X\,bvRSS}})] + \beta^2\sigma^2_{X[i](j)k}$$

Then the fitted model is

$$\overline{Y}_{(i)[j]ki} = \hat{a}_{bvrss} + \hat{\beta}_{bvrss}\,\overline{X}_{[i](j)k}$$

$$e_{(i)[j]k} = \hat{\epsilon}_{(i)[j]k} = Y_{(i)[j]k} - \hat{Y}_{[i](j)k}$$

Also, a consistent unbiased estimate for σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{\sum_{k.j.i} e_{(i)[j]k}^2}{n-k}$$

Where, *k* is the number of parameters to be estimated in simple linear regression model.

Assuming the conditions of the regression model above, then

1) $E(\hat{a}_{bvrss}) = a\, E(\hat{\beta}_{bvrss}) = \beta$

2) $Var(\hat{a}_{bvrss}) = \frac{a^2}{n}[1 + E(\frac{\overline{X}^2{}_{bvrss}}{S^2_{X,\,bvrss}})]$

where

$$S^2_{X,\,bvrss} = \frac{\sum_{k,j,i} (X_{(i)[j]k} - \overline{X}_{bvrss})^2}{n}$$

3) $Var(\hat{\beta}_{bvrss}) = \frac{a_e^2}{n}[1 + E(\frac{1}{S^2_{X,\,bvrss}})]$

4) $E(\hat{Y}_{(i)[j]k} / X_i = x_i) = a\,\beta x_{(i)[j]k}$

5) $Var(\hat{Y}_{(i)[j]k}) = \frac{a_e^2}{n}[1 + E(\frac{(x_{(i)[j]k} - X_{bvrss})^2}{S^2_{X,\,bvrss}})] + \beta^2 \sigma^2_{X(i)[j]k}$

6) $E(e_{(i)[j]k}) = 0$

7) $Var(e_{(i)[j]k}) = a_e^2[1 - [\frac{1}{n} + E(\frac{(x_{(i)[j]k} - \overline{X}_{bvrss})^2}{nS^2_{X,\,bvrss}})]$

8) $\mathrm{cov}(\hat{a}_{bvrss}, \hat{\beta}_{bvrss}) = -\frac{a_e^2}{n} E(\frac{(\overline{X}_{bvrss})^2}{\overline{Y}_{bvrss}})]$

9) $E(\hat{\sigma}_e^2) = \sigma_e^2$

10) $Var(\hat{\sigma}_e^2) = \frac{2\sigma_e^4}{n-k}$ (Assuming normality)

11) $\hat{\sigma}_e^2 \to \sigma_e^2$

From the above conditions, we can derive the efficiencies of the estimators of α and β using BVRSS relative to the estimators using BVSRS (Bivariate Simple random sampling) and BVSYS (Bivariate Systematic sampling) as follows:

$$\text{eff}\ (\hat{a}_{bvrss}, \hat{\beta}_{bvh}) = \left(\frac{[1 + E\frac{(\overline{X}^2{}_{bvh})}{S^2_{X,bvh}}}{[1 + E\frac{(\overline{X}^2{}_{bvrss})}{S^2_{X,bvrss}}}\right)]$$

where

bvh = bvsrs, bvsys

and

$$\text{Efficiency}\ (\hat{\beta}_{bvrss}, \hat{\beta}_{bvh}) = \left(\frac{E\frac{1}{S^2_{bvh}}}{E\frac{1}{S^2_{X,bvrss}}}\right)]$$

3.6 Numerical illustration

Assuming that the (X, Y) follows the bivariate normal distribution, the performance of simple regression model using BVSRS, BVSYS and BVRSS is judged with the help of a data set. The original data were collected on two variables of *Pinus wllichiana* : *X*, the diameter in centimetres at breast height and *Y*, the entire height in meters. The regression model is analysed assuming that the population is consists of *275* trees. The summary statistics of the data is reported as in Table-2.

Table-2 : Summary statistics of the Pinus data

	DBH (cm)	*Height (m)*
Mean	21.44	15.66
Standard Deviation	20.95	17.06
Range	216.80	70.87
Minimum	2.20	0.90
Maximum	219	71.77
Count	275	275

A sample size of 55 was fixed in all the sampling designs. Regression analysis and regression diagnostics in case of all the three sampling designs was carried out in SAS software using the function `poc reg,` which is given below:

```
data pinus;
input dbh height;
cards;
;
run;
ods rtf;
proc reg data=pinus;
model height=dbh;
run;
plot=diagnostics(unpack);
ods graphics off;
ods rtf close;
```

The layout of RSS is given in Table-3. Where row number 1,3,5,7,9 in each cycle represents the tree number and rows 2,4,6,8,10 represents the “dbh” in each cycle.

Table-3 : Layout of RSS

Cycles	Set size = k= 5 (N= 275, n = 55, means we have to repeat the process of ranking 11 times i.e. 11×5 = 55				
Cycle 1	1	2	3	4	5
	15.9	22	56.9	9.6	24.6
	6	7	8	9	10
	3.3	11.4	4.7	21.3	16.8
	11	12	13	14	15
	5.1	7.5	3.1	4.9	6.1
	16	17	18	19	20
	5.5	6.5	5.6	6.9	3.8
	21	22	23	24	25
	9.7	6.9	4.1	58.5	46
Cycle 2	26	27	28	29	30
	22.2	3.7	52.9	63.2	46.5
	31	32	33	34	35
	56.3	219	11	4.7	11
	36	37	38	39	40
	58.8	3.5	10.1	16.9	10.8
	41	42	43	44	45
	9	8	17.8	23.9	2.3
	46	47	48	49	50
	5.8	6	8.8	9.9	14.6
Cycle 3	51	52	53	54	55
	10.8	44.2	12.9	28	39.8
	56	57	58	59	60
	20.4	47.3	35.7	44.9	8.7
	61	62	63	64	65
	24.3	15.7	30.9	69.2	24.1
	66	67	68	69	70
	4.2	3.8	41.2	39.8	18.6
	71	72	73	74	75
	38.7	12.2	6	8	13.5

Contd…

Table-3 contd…

Cycle 4	76	77	78	79	80
	20.1	57.4	8.2	32.7	9.4
	81	82	83	84	85
	8.9	9.2	6.1	7.5	52.3
	86	87	88	89	90
	15.5	23.7	67.1	12.3	14
	91	92	93	94	95
	4.9	5.5	7.6	3.5	6.3
	96	97	98	99	100
	19	2.7	8.2	7.6	9.2
Cycle 5	101	102	103	104	105
	5.9	6.2	13.3	13.4	33.9
	106	107	108	109	110
	33.7	8.3	48	40.4	8.6
	111	112	113	114	115
	16	29.1	18.4	26.8	6.2
	116	117	118	119	120
	2.9	3	14.6	18.4	15
	121	122	123	124	125
	18.4	44.5	4.5	10.4	24
Cycle 6	126	127	128	129	130
	5.1	5.3	2.5	2.2	3.1
	131	132	133	134	135
	2.6	8.1	12.4	15.1	12.7
	136	137	138	139	140
	49	20.8	11.9	47.6	10.6
	141	142	143	144	145
	22.9	10.6	49.7	50.6	19.1
	146	147	148	149	150
	53	18	44.4	10.8	51.7

Contd…

Table-3 contd…

Cycle 7	151	152	153	154	155
	22.6	7.7	43.5	3.1	5
	156	157	158	159	160
	4.4	3.3	2.6	53.5	48.9
	161	162	163	164	165
	47.8	17.2	28.6	10.8	50.1
	166	167	168	169	170
	4.7	5.3	10.6	3.7	3.9
	171	172	173	174	175
	5.3	2.5	13.2	17.1	13.9
Cycle 8	176	177	178	179	180
	8	8.5	50.1	6.8	19.9
	181	182	183	184	185
	17.5	6.8	10.9	11.2	20.2
	186	187	188	189	190
	19.6	18.4	50.9	17.6	44.1
	191	192	193	194	195
	17	46.9	2.8	25.5	14.5
	196	197	198	199	200
	14.1	47.1	42.2	40.2	66.8
Cycle 9	201	202	203	204	205
	4.1	60.6	8	17.2	22
	206	207	208	209	210
	15.9	3.1	4.5	32	46.9
	211	212	213	214	215
	36.4	25.4	40	40.4	19.8
	216	217	218	219	220
	30.5	37.7	22.1	5.5	28.4
	221	222	223	224	225
	46.4	15.8	45.9	33.5	36.7

Contd…

Table-3 contd…

Cycle 10	226	227	228	229	230
	44	51.6	45	34	53.1
	231	232	233	234	235
	30.8	17.2	57	6.3	44.2
	236	237	238	239	240
	3	36.4	2.7	4.4	41.4
	241	242	243	244	245
	3.4	8.4	4.8	4.2	6.3
	246	247	248	249	250
	32.6	15.3	38.6	5.2	61.8
Cycle 11	251	252	253	254	255
	10.9	3.5	2.5	10.9	8.9
	256	257	258	259	260
	21	44.1	7	9.4	8
	261	262	263	264	265
	23	11.6	33	7.5	17.5
	266	267	268	269	270
	8.9	47.4	22	6.8	7.5
	271	272	273	274	275
	22.2	19.3	14.5	3.5	10.9

The relative efficiency of RSS with SRS and SYS along with R^2, Adj R^2 and other parameters of comparison are given the Table-4.

Regression diagnostics and kernel density plots in each sampling design are produced in Figs. 5 to 10.

The performance of RSS with SRS and SYS is also judged with the help of validation technique i.e. Jack-knifing carried out in SAS using function PROC JACKREG in Tables-5 to 7.

In present study simple linear regression models were considered with respect to samples taken from the sampling techniques like simple random sampling (SRS), systematic sampling (SYS) including rank set sampling (RSS). It was found that the R^2, Adj R^2 obtained from regression model based on rank set sample were higher than rest of two sampling schemes considered, also the parameters of comparison like root mean square error, p value and coefficient of variation were much lower in rank set based regression model than the other considered schemes. From density curves again the curves were more symmetric in case of rank set sample as compared to SRS and SYS. Also from validation technique (Jacknifing) there was consistency in the measure of R^2, Adj R^2 and RMSE (Root mean square error) in case of RSS as compared to SRS and SYS. The above results occurred because rank set samples are more regularly spaced than those obtained from SRS and SYS and therefore more representative of the population. As ranking of sampling units in the RSS procedure induces stratification at sample level which results the gain in precession.

Table-4 : Relative efficiency of RSS with SRS and SYS along with R^2, Adj R^2 and others measures of comparison

	RSS	SRS	SYS
RMSE	8.221	9.521	9.851
R^2	0.8029	0.771	0.7458
Adj R^2	0.7991	0.763	0.7268
ESS	2942.73	3600.26	4804.16
F value	414.78	215.85	209.21
CV	35.41	55.61	48.42
p value	0.0007	0.0034	0.0048
Dbh (β_1)	1.175	0.935	0.916

	Lower 95%	Upper 95%	Lower 95%	Upper 95%	Lower 95%	Upper 95%
Confidence Interval	1.05	1.29	0.64	1.06	0.60	1.08
	0.24		0.42		0.48	

	$RMSE_{srs}/RMSE_{rss}$	$RMSE_{sys}/RMSE_{rss}$
Relative Efficiency	1.15	1.19

Table-5 : Comparison of regression models using Jack-knifing technique

No. of Models	RSS (Rank set sampling)			SRS (Simple random sampling)			Systematic sampling		
	R^2	Adj R^2	RMSE	R^2	Adj R^2	RMSE	R^2	Adj R^2	RMSE
1	0.8021	0.7984	8.231	0.7972	0.7940	9.525	0.7455	0.7267	9.653
2	0.8019	0.7987	8.225	0.7901	0.7632	9.632	0.7135	0.6995	10.432
3	0.8011	0.7982	8.223	0.7992	0.7165	9.743	0.7194	0.6941	10.932
4	0.8015	0.7993	8.228	0.7832	0.7132	10.146	0.7065	0.6932	11.401
5	0.8011	0.7984	8.235	0.7814	0.7115	10.324	0.7034	0.6911	11.567
6	0.8018	0.7984	8.238	0.7801	0.7135	10.365	0.7010	0.6843	11.537
7	0.8022	0.7983	8.242	0.7832	0.7128	10.378	0.7001	0.6872	11.612
8	0.8024	0.7979	8.239	0.7733	0.7132	9.432	0.6942	0.6845	11.105
9	0.8020	0.7973	8.245	0.7632	0.7109	8.475	0.6837	0.6741	11.653
10	0.8023	0.7971	8.228	0.7774	0.7113	11.372	0.6135	0.6735	11.724
11	0.8029	0.7974	8.223	0.7701	0.6995	11.248	0.6347	0.6611	12.001
12	0.8015	0.7992	8.236	0.7732	0.6943	11.371	0.6451	0.6601	12.345
13	0.8013	0.7993	8.230	0.7632	0.6735	9.506	0.6458	0.6538	9.377
14	0.8016	0.7995	8.227	0.7448	0.6525	9.135	0.6743	0.6527	11.433
15	0.8012	0.7998	8.240	0.7452	0.7103	9.346	0.6748	0.6439	12.165
16	0.8015	0.7988	8.245	0.7456	0.6772	9.732	0.6551	0.6425	12.437
17	0.8021	0.7981	8.237	0.7480	0.6785	9.441	0.6449	0.6417	11.523
18	0.8022	0.7987	8.230	0.7495	0.6742	10.532	0.6442	0.6391	12.453
19	0.8023	0.7983	8.228	0.7501	0.6832	9.632	0.6767	0.6382	11.06
20	0.8025	0.7985	8.229	0.7832	0.6945	9.575	0.6743	0.6311	10.501
21	0.8015	0.7982	8.237	0.7827	0.7343	9.321	0.6859	0.6235	10.425
22	0.8014	0.7983	8.231	0.7773	0.7135	10.242	0.6866	0.6211	11.501
23	0.8016	0.7985	8.234	0.7721	0.7247	11.438	0.6977	0.6171	10.501
24	0.8013	0.7987	8.236	0.7732	0.7135	11.586	0.6542	0.6123	10.425
25	0.8015	0.7988	8.228	0.7560	0.6991	10.788	0.6321	0.6118	11.167
26	0.8030	0.7990	8.240	0.7470	0.6432	9.656	0.6354	0.6112	12.432
27	0.8028	0.7991	8.239	0.7321	0.6135	9.842	0.6366	0.6011	11.937
28	0.8029	0.7993	8.244	0.7215	0.7201	9.747	0.6501	0.5991	10.666
29	0.8023	0.7992	8.233	0.7721	0.7115	9.735	0.6554	0.5932	11.732
30	0.8024	0.7993	8.237	0.7232	0.6432	9.645	0.6932	0.5932	12.406
31	0.8012	0.7995	8.243	0.7245	0.6940	9.748	0.6142	0.5995	9.735
32	0.8017	0.7991	8.240	0.7243	0.6532	10.432	0.5995	0.6013	9.532
33	0.8015	0.7992	8.246	0.7165	0.6348	10.458	0.5812	0.6115	9.567
34	0.8030	0.7984	8.231	0.7237	0.6474	9.638	0.5994	0.5942	9.471
35	0.8017	0.7981	8.237	0.7354	0.6903	9.546	0.6011	0.5711	9.450
36	0.8016	0.7980	8.228	0.7380	0.7103	9.532	0.6045	0.5432	9.448
37	0.8027	0.7976	8.243	0.7410	0.6532	9.437	0.6741	0.5511	9.478
38	0.8023	0.7972	8.245	0.7580	0.6671	9.648	0.6849	0.5432	9.501
39	0.8024	0.7973	8.235	0.7595	0.6643	9.632	0.6978	0.5617	9.548
40	0.8020	0.7975	8.238	0.7610	0.6854	9.644	0.7015	0.6348	9.539
41	0.8023	0.7979	8.242	0.7623	0.6711	9.651	0.7135	0.6556	9.511
42	0.8025	0.7983	8.246	0.7651	0.6535	9.628	0.7143	0.6417	10.235
43	0.8021	0.7981	8.248	0.7659	0.7113	9.645	0.7211	0.6521	11.247
44	0.8026	0.7989	8.229	0.6906	0.7211	9.137	0.7312	0.6695	9.373
45	0.8028	0.7986	8.233	0.6972	0.7013	9.635	0.7154	0.6743	9.381
46	0.8019	0.7988	8.238	0.7643	0.6143	9.875	0.7312	0.6528	9.456
47	0.8020	0.7990	8.240	0.7511	0.6051	9.445	0.7221	0.6818	9.556
48	0.8017	0.7995	8.245	0.7451	0.6543	9.436	0.7145	0.6743	9.247
49	0.8015	0.7993	8.240	0.7559	0.6543	9.635	0.7116	0.6713	10.498
50	0.8013	0.7992	8.239	0.6972	0.6136	10.432	0.7234	0.6855	12.071
51	0.8023	0.7990	8.238	0.6907	0.6142	10.548	0.7112	0.6711	11.247
52	0.8024	0.7993	8.240	0.7643	0.6547	10.456	0.7135	0.6943	11.478
53	0.8025	0.7990	8.243	0.7511	0.6567	11.230	0.7149	0.6843	12.562
54	0.8021	0.7982	8.238	0.7432	0.6549	10.629	0.6235	0.6816	12.164
55	0.8029	0.7991	8.242	0.7979	0.7941	9.521	0.7458	0.7268	9.651

Table-6: Summary statistics in each sampling design based on Jacknifing

	RSS (Rank set sampling)			SRS (Simple random sampling)			Systematic sampling		
	R^2	Adj R^2	RSE	R^2	Adj R^2	RSE	R^2	Adj R^2	RSE
Mean	0.8020	0.7986	8.2264	0.7034	0.6832	9.9631	0.6759	0.6418	10.8037
Standard Deviation	0.0005	0.0007	0.0065	0.0262	0.0388	0.6675	0.0417	0.0434	1.0948
Range	0.0019	0.0027	0.0250	0.1086	0.1889	3.1110	0.1643	0.1835	3.3150
Largest	0.8030	0.7998	8.2480	0.7992	0.7940	11.5860	0.7455	0.7267	12.5620
Smallest	0.8011	0.7971	8.2230	0.6906	0.6051	8.4750	0.5812	0.5432	9.2470

Table-7: Design fitted based on 55 samples

RSS (Rank set sampling)			SRS (Simple random sampling)			Systematic sampling		
R^2	Adj R^2	RSE	R^2	Adj R^2	RSE	R^2	Adj R^2	RSE
0.8029	0.7991	8.221	0.771	0.763	9.521	0.7458	0.7268	9.851

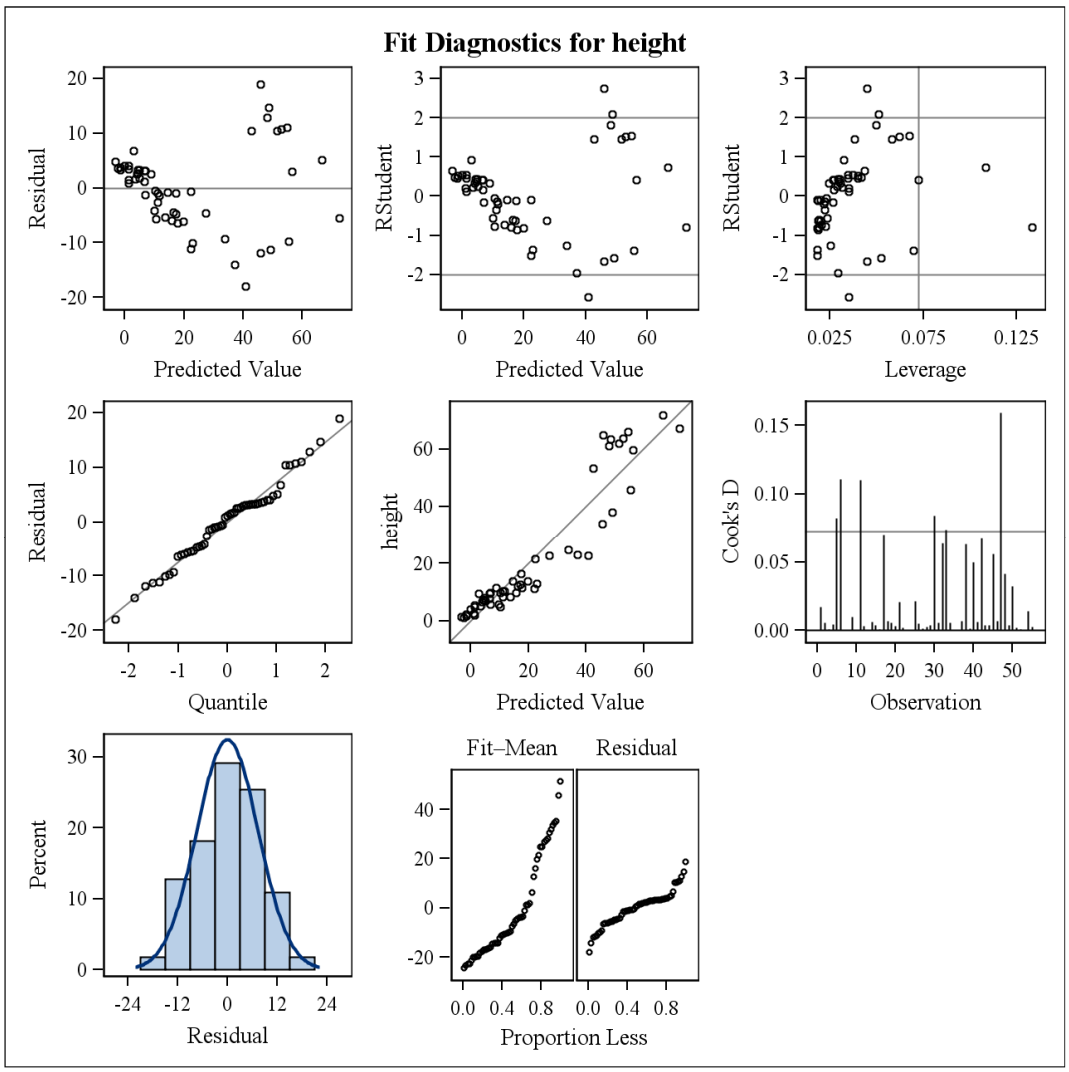

Fig. 5 : Rank Set Sampling

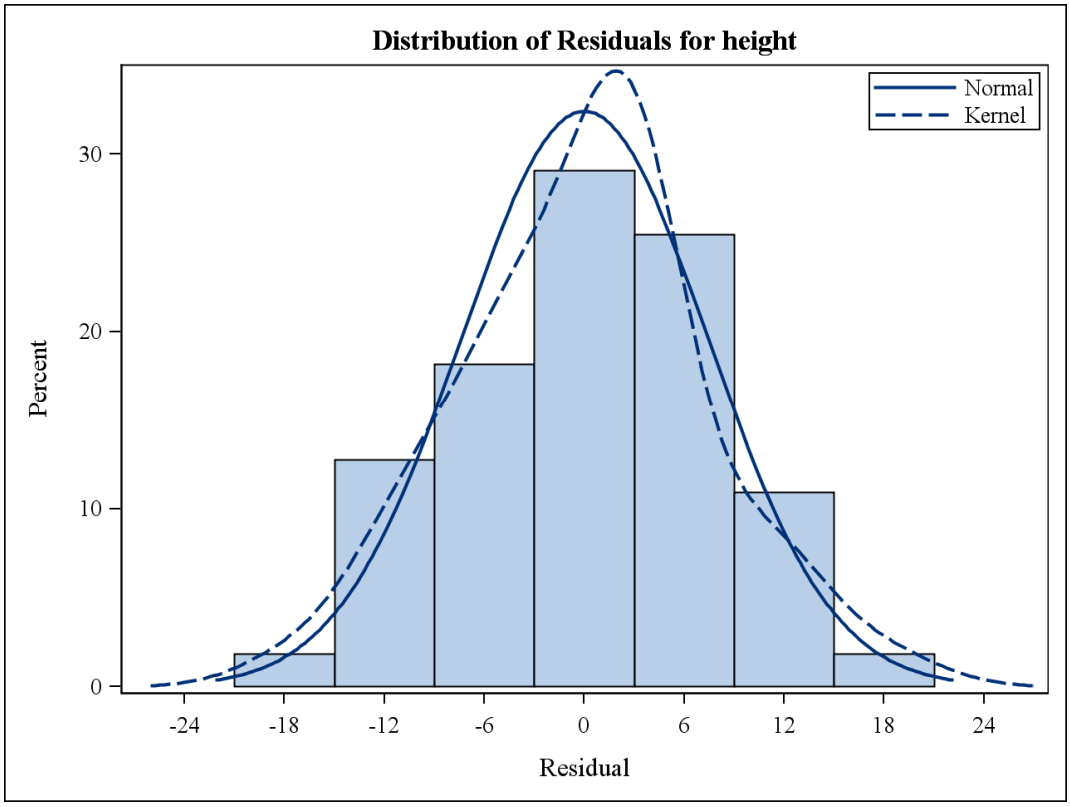

Fig. 6 : Kernel density plot in RSS

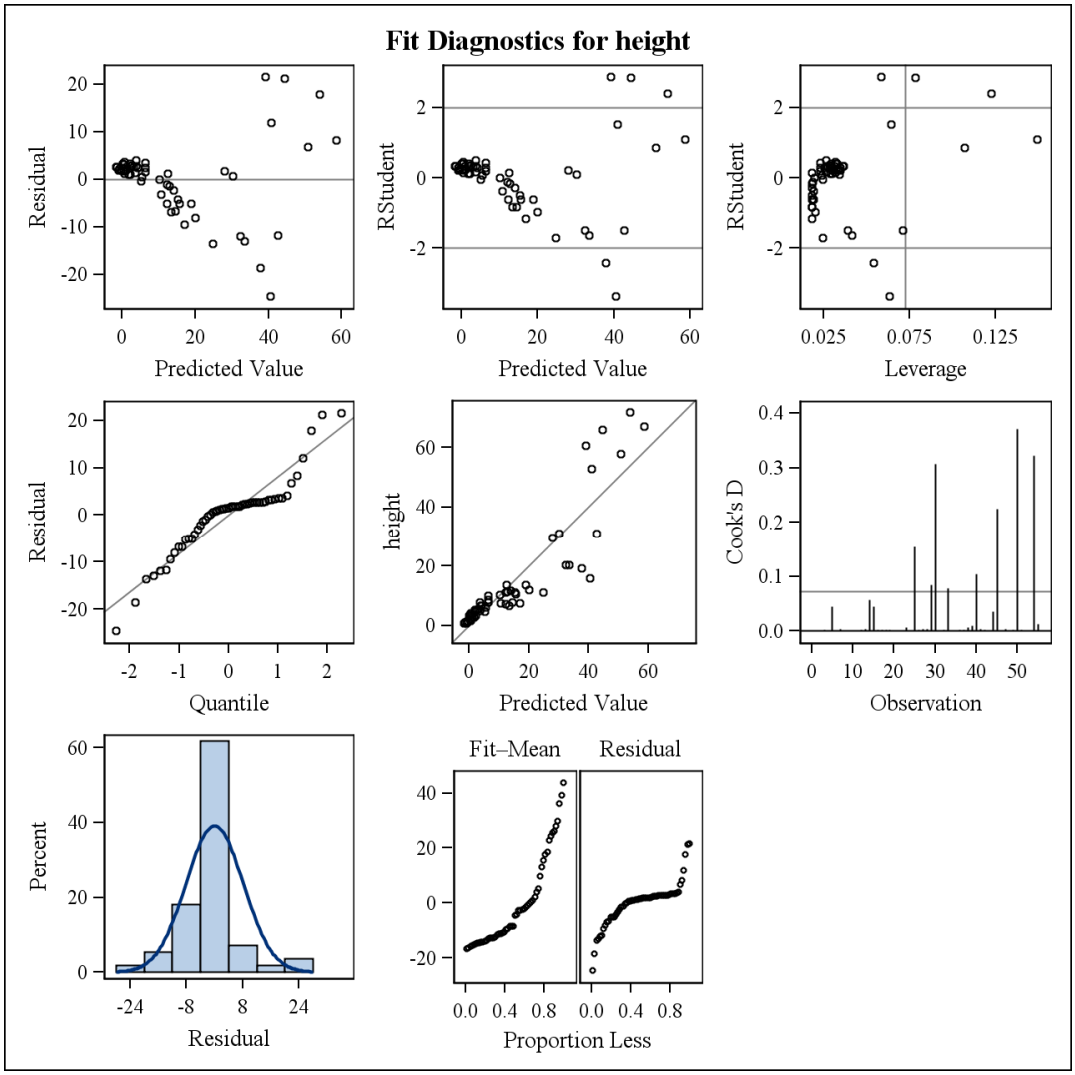

Fig. 7 : Simple Random Sampling

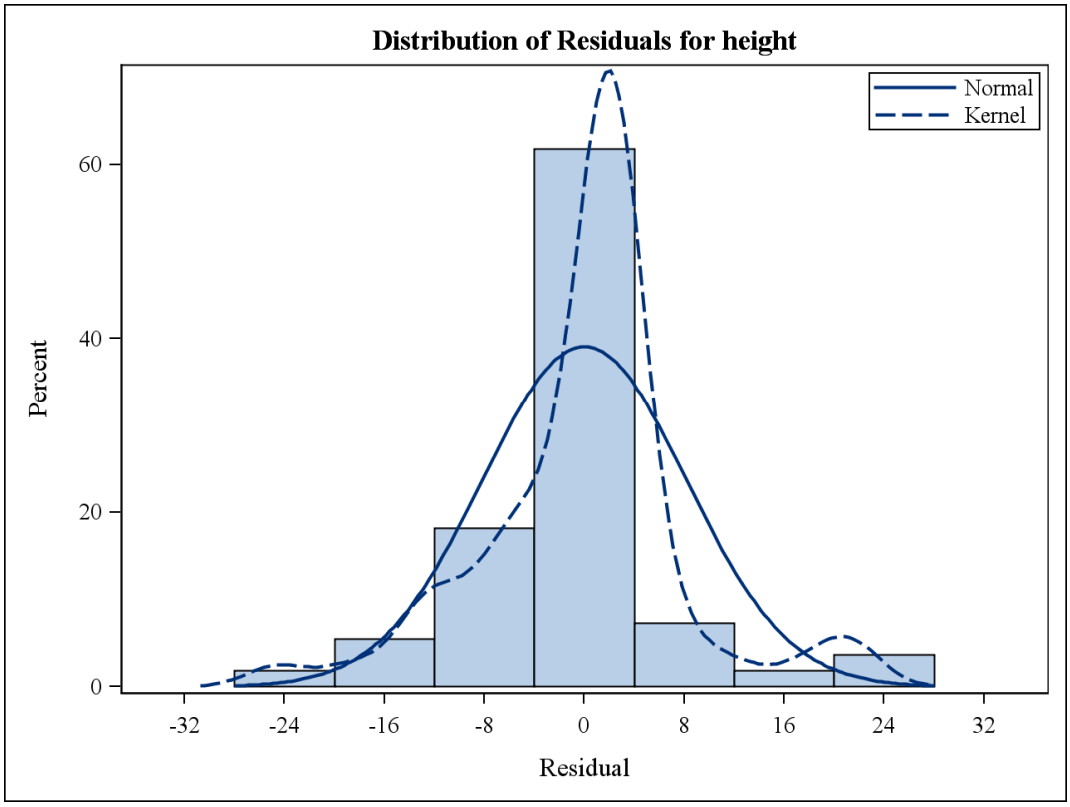

Fig. 8 : Kernel density plot in Simple Random Sampling

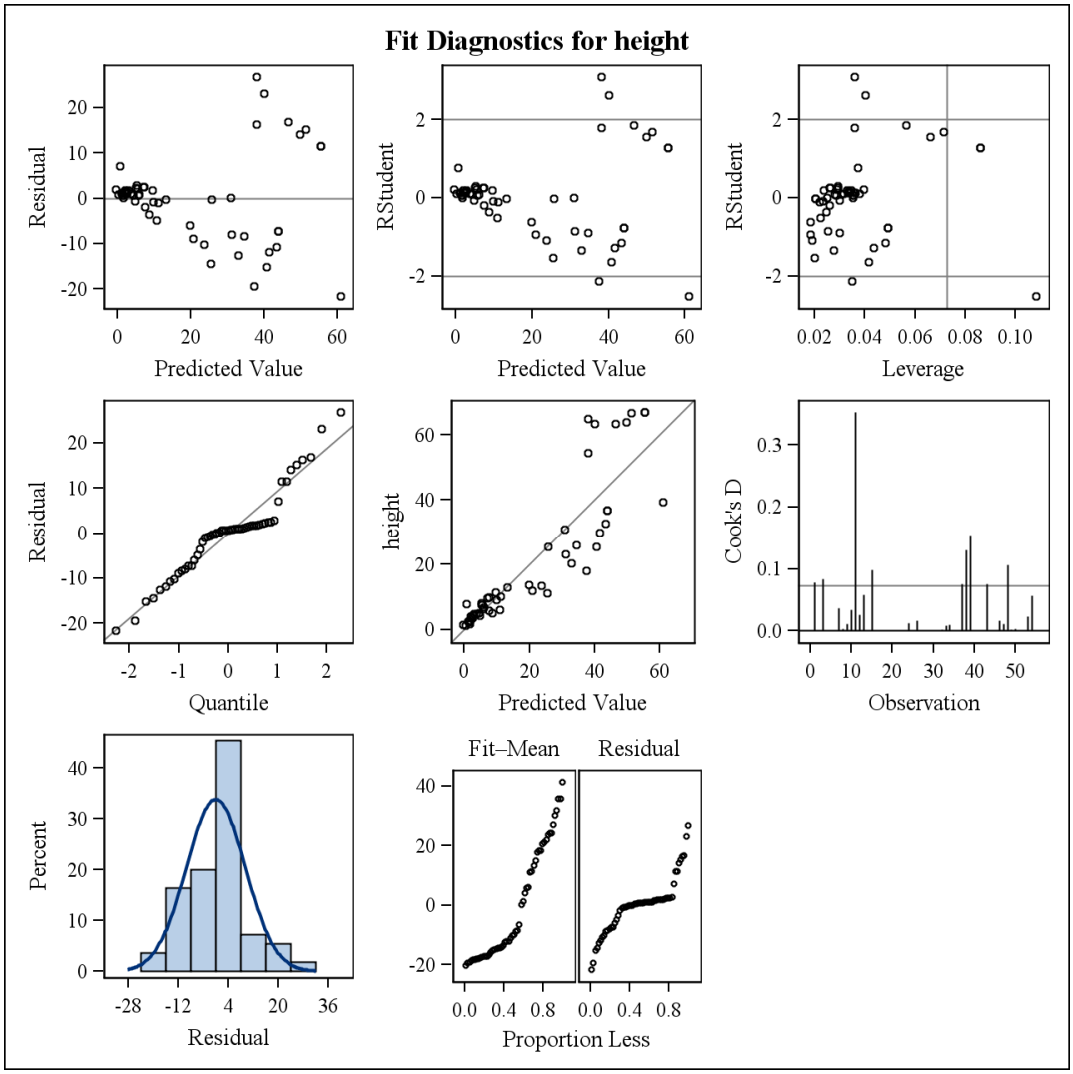

Fig. 9 : Systematic Sampling

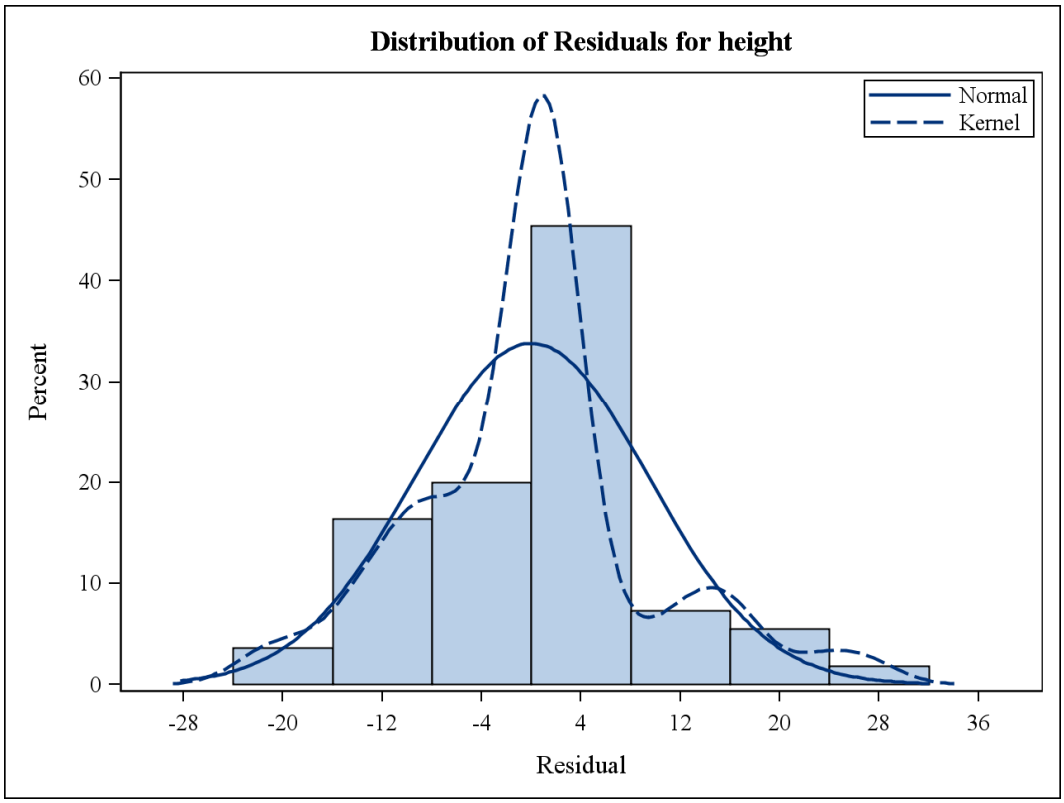

Fig. 10 : Kernel density plot in Systematic Sampling

Chapter – 4

IMPACT OF RANK SET SAMPLING ON THE ESTIMATORS OF POPULATION MEAN IN STRATIFIED SAMPLING

Ranked set sampling (RSS) is a method of collecting data that improves estimation by utilizing the sampler's judgment or auxiliary information about the relative sizes of the sampling units. Prior to quantifying the data, the researcher samples from the population and then ranks the sampled units based on his or her judgment about their relative sizes on the variable of interest. In survey sampling settings, a logical method of ranking the units is to order them based on the values of an auxiliary variable correlated with the variable of interest. Using the ranks of the units, the researcher creates a subset of the sample and quantifies the variable of interest for units in the subset. The measurements from this subset are used to estimate population parameters. The method was first proposed by McIntyre (1952) to increase the accuracy of crop yield estimates without increasing the number of observations that need to be quantified. The middle of 1980's was a turning point in the development of the theory and methodology of RSS. Since then, various statistical procedures with RSS, non-parametric or parametric, have been investigated, variations of the original notion of RSS have been proposed and developed, and sound general theoretical foundations of RSS have been laid.

4.1 Estimation procedure

In this chapter we have introduced the concept of Ranked set stratified sampling (RSSS) procedure to estimate the population mean. As we know whenever in surveys we are dealing with a heterogeneous population, we divide the whole population into homogenous groups under certain criterion. These groups are termed as starta. Then a sample is drawn randomly from each stratum independently. Such a sampling procedure is known as Stratified simple random

sampling (SSRS). A SSRS is a sampling plan in which a population is divided into L mutually exclusive strata and a simple random sample (SRS) of n_K elements is taken within each stratum k. Such a sampling procedure provides an increased precision in the estimates of population mean. As in case of SSRS where we take a random sample from each stratum, a ranked set sample of n_K elements is quantified within each stratum, $k = 1,2,3 \ldots . L$. Such a procedure will be called as ranked set stratified sampling (RSSS), which is nothing but a collection of L separate ranked set samples.

For all $i, i = 1,2,3 \ldots ., n_k$ and $(k = 1,2,3 \ldots . L)$, let

$\mu_k = E(Y'_{kij}),\ \sigma_k^2 = var(Y'_{kij}),$
$\mu_{k(i)} = E(Y'_{ki(i)}), \sigma_{k(i)}^2 = var(Y'_{ki(i)})\ for\ all\ j = 1,2,3, \ldots, n_k$

The mean μ of the variable Y is given by

$$\mu = \frac{1}{N}\sum_{k=1}^{L} N_k \mu_k = \sum_{K=1}^{L} Z_k \mu_k , \qquad (4.1.1)$$

Where, $Z_k = \frac{N_k}{N}$ (4.1.2)

Also, estimate of mean is given by $\bar{Y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki1}$ (4.1.3)

The mean and variance of $\bar{Y}_k$ are known to be $E(\bar{Y}_k) = \mu_k$ and $var(\bar{Y}_k) = \frac{\sigma_k^2}{n_k}$ respectively, assuming $N_k's$ are large enough. The estimate of the population mean μ using SSRS of size n is defined by

$$\bar{Y}_{SSRS} = \frac{1}{N}\sum_{k=1}^{L} N_k \bar{Y}_k = \sum_{k=1}^{L} Z_k \bar{Y}_k \qquad (4.1.4)$$

The mean and variances are known to be $E(\bar{Y}_{SSRS}) = \mu$ and

$var\ (\bar{Y}_{SSRS}) = \sum_{k=1}^{L} Z_k^2 \left(\frac{\sigma_k^2}{n_k}\right)$ respectively.

4.2 Rank set sampling under Stratification

In this section, rank set sampling is introduced under stratification, for this purpose a population is divided in L mutually exclusive and exhaustive starta. Let $Y_{k11}, Y_{k12}, \ldots, Y_{k1n_k}; Y_{k21}, Y_{k22}, \ldots, Y_{k2n_k}; \ldots\ldots\ldots; Y_{kn_k1}, Y_{kn_k2}, \ldots, Y_{kn_kn_k}$ be n_k independent random samples of size n_k each one is taken from each stratum ($k = 1,2,3 \ldots L.$). Assuming that each element in the sample has same density function and distribution function, then according to our description $Y'_{k11}, Y'_{k21}, \ldots\ldots, Y'_{kn_k1}$ could be considered as SRS from the $k-th$ stratum. Let $Y_{ki(1)}, Y_{ki(2)}, \ldots, Y_{ki(n_k)}$ be the ordered statistics of $i-th$ sample $Y_{ki1}, Y_{ki2}, \ldots, Y_{kin_k}$ $(i = 1,2,3 \ldots n_k)$ taken from $k-th$ stratum. Then, $Y_{k1(1)}, Y_{k1(2)}, \ldots, Y_{kn_k(n_k)}$ denotes the ranked set sample for the $k-th$ stratum. If $N_1, N_2, \ldots\ldots, N_L$ represent the number of sampling units within respective strata, and $n_1, n_2, \ldots\ldots, n_L$ represent the number of sampling units measured within each stratum, then $N = \sum_{k=1}^{L} N_k$ will be the total population size, and $n = \sum_{k=1}^{L} n_k$ will be the total sample size.

4.3 Notations of rank set sampling under stratification

If we select a RSS of n_k elements from N_k elements in the stratum and each sample element is measured, then estimate of mean using RSS will be

$$\bar{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki(i)} \qquad (4.3.1)$$

It can be shown that mean and variance of $\bar{Y}_{k(n_k)}$ is

$$E\left(\bar{Y}_{k(n_k)}\right) = \mu_k \text{ and} \qquad (4.3.2)$$

$$var\left(\bar{Y}_{k(n_k)}\right) = \frac{\sigma_k^2}{n_k} - \frac{1}{n_k^2} \sum_{i=1}^{n_k} T_{k(i)}^2, \text{ respectively} \qquad (4.3.3)$$

where, $T_{k(i)}^2 = (\mu_{k(i)} - \mu_k)$

Which means the estimate of population mean under Ranked set stratified sampling is

$$\bar{Y}_{RsSS} = \frac{1}{N}\sum_{k=1}^{L} N_k \bar{Y}_{k(n_k)} = \sum_{k=1}^{L} Z_k \bar{Y}_{k(n_k)} \qquad (4.3.4)$$

Estimate of variance is $var(\bar{Y}_{RSSS}) = \sum_{k=1}^{L} Z_k^2 \left(\frac{\sigma_k^2}{n_R} - \frac{1}{n_k^2}\sum_{i=1}^{n_k} T_{k(i)}^2\right)$

Therefore the relative efficiency of the estimator of population mean using SSRS with respect to ranked set stratified sampling is given as;

$$RE = \frac{var(\bar{Y}_{SSRS})}{var(\bar{Y}_{RSSS})} = \left\{1/1 - \frac{1}{var(SSRS)}\left(\sum_{k=1}^{L} \frac{Z_k^2}{n_k^2}\sum_{i=1}^{n_k} T_{k(i)}^2\right)\right\}$$

which is always greater than 1. Also the standard errors in case of ranked set sampling are much lower as compared to SSRS, which is proved with the help of the simulation on a real data .

4.4 Numerical illustration

The data on Apple production from district Ganderbal of Kashmir valley from 420 orchards in 30 villages was taken in this chapter. The variables choosen for the study where Yield (MT), Bearing trees, Total number of trees, Area (ha). We take samples of equal size for each sampling design and estimate the standard error in each sampling design. The sample sizes considered were 10, 20, 40, 60 and the set sizes considered were 5,10,15 shown in Table-8 using same number of strata's in case of SSRS and SRSS along with correlation coefficients ρ ranging from 0.90 to 0.40 . Three distinct simulations based on three combinations of sample sizes and set sizes for each sampling design; each simulation uses a combination of variables for stratification and/or ranking and quantification.

From the results of Table-8, it is observed that ranked set sampling, when used in place of simple random sampling in stratified sampling provides more accurate estimates of population means. The gain in precision of ranked set

estimator over the estimator based on simple random sampling under stratification occurs, when ranking is inexpensive relative to the cost of quantifying the variable of interest, also the results of simulations reveal that the ranked set stratified sampling procedure combines the variance reduction that arises from stratifying the population with the increased precision that ranked set sampling holds over simple random sampling.

Table-8: Results of simulation study

Correlation coefficient (ρ)	Sampling procedure	Variable combinations	No of sets or strata	STANDARD ERRORS			
				Sample sizes			
				10	20	40	60
90	Stratified random sampling	Yield *vs* Area	5	180.13	175.16	168.52	161.71
			10	177.43	170.27	163.04	158.38
			15	165.27	160.04	155.43	147.63
	Ranked set stratified sampling		5	177.43	171.32	162.63	155.43
			10	170.78	165.42	158.43	150.27
			15	162.43	156.32	150.01	147.63
70	Stratified random sampling	Bearing trees *vs* Area	5	1756.43	1730.39	1709.58	1683.54
			10	1743.52	1725.32	1700.04	1677.41
			15	1728.65	1718.42	1692.58	1657.32
	Ranked set stratified sampling		5	1715.38	1700.43	1688.43	1671.52
			10	1709.12	1692.18	1672.53	1654.37
			15	1690.35	1681.53	1669.43	1638.52
40	Stratified random sampling	Total trees *vs* Area	5	2271.52	2261.25	2242.63	2232.13
			10	2263.48	2258.1	2238.57	2221.08
			15	2259.33	2240.09	2230.01	2215.54
	Ranked set stratified sampling		5	2252.11	2241.51	2232.54	2221.09
			10	2248.27	2230.62	2225.63	2211.52
			15	2230.52	2221.11	2218.57	2207.54

Chapter – 5

DEVELOPMENT OF A NEW CLASS OF RATIO ESTIMATORS USING RANK SET SAMPLING AND THEIR COMPARISON WITH THE CLASSICAL RATIO ESTIMATORS

Sampling is not mere substitution of a partial coverage for a total coverage. Sampling is the science and art of controlling and measuring the reliability of useful statistical information through the theory of probability. The simplest and the most common method of sampling is simple random sampling in which a sample is drawn unit by unit, with equal probability of selection for each unit at each draw, where there is no additional information available. Most of the times in sample surveys, along with the variable of interest Y, information on auxiliary variable X, which is highly correlated with Y is also collected. This information on auxiliary variable may well be utilized to obtain a more efficient estimator of population mean. Ratio method of estimation is one such example which utilizes the information on auxiliary variable X, which is positively correlated with the variable of interest Y, in order to improve the precision of the estimate of population mean. An alternative method to SRS called Ranked Set Sampling (RSS) was introduced to increase the efficiency of the estimation of population mean (1952). The method is useful when the variable of interest is very expensive or difficult to measure but it can be easily ranked at a negligible cost. There are cases in practical situation where the variable of interest Y is difficult to measure and to rank but a concomitant variable X, which is highly correlated with Y, can be easily ranked and be used for the ranking of the sampling units. Some extension is done by Swami (1996) utilized both the rank and the measure of the concomitant variable and considered ratio estimation using RSS. The ratio estimation based on RSS is more efficient compared with the SRS ratio estimate.

5.1 Ratio estimation under simple random sampling

Let the variable of interest Y and the concomitant variable X is correlated with the coefficient of correlation ρ. The population ratio of these two variable is $R = {\mu_y}/{\mu_x}$ or $R = {\bar{Y}}/{\bar{X}}$ and its estimator is $\hat{R} = \frac{\bar{y}}{\bar{x}}$, where μ_y and μ_x are the populations means of the variables Y and X respectively, $\bar{y}$ and $\bar{x}$ are the sample means for μ_x and μ_y respectively. The ratio estimator is biased but the bias is negligible when the estimator is approximated using Taylor series expansion to the first degree (Cochran, 1977). The approximated Variance of $\hat{R}$ is

$$\mathbf{Var}\left(\hat{R}\right) \cong \frac{R^2}{n}\left(V_x^2 + V_y^2 - 2\rho_x V_x V_y\right)$$

where, $V_x = \frac{\sigma_x}{\mu_x}$, $V_y = \frac{\sigma_y}{\mu_y}$ and

$$\rho_{xy} = \Sigma_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)/N\sigma_x\sigma_y$$

σ_x and σ_y are the standard deviations of the population of the variables X and Y, respectively.

There are many existing ratio type estimators based on SRS. Improvements in the precision of the estimator through the use of Ratio method of estimation are achieved by introducing a large number of modified Ratio estimators which utilizes the information on known values of Co-efficient of variation, Co-efficient of kurtosis, Co-efficient of Skewness etc. Some of the modified Ratio estimators in case of SRS given by Murthy (1967), Cochran (1977), Prasad (1989), Sen (1993), Upadhyaya and Singh (1999), Singh and Tailor (2003, 2005), Singh *et al.* (2004), Kadilar and Cingi (2004, 2006), Koyuncu and Kadilar (2009), Yan and Tian (2010), Iqbal *et al.* (2013) are available in literature. These estimators are in the form

$$\hat{\mu}_{\text{SRS}} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(\alpha\bar{x}+\gamma)}(\alpha\mu_x + \gamma)$$

where $\beta = {\sigma_{xy}}/{\sigma_x^2}$

They suggested utilizing some known parameters of the concomitant variable *X*. where the α *and* γ are the main parameters of interest.

Kadilar and Chingi (2004), Koyuncu and Kadilar (2009) proposed some modified ratio estimators based on known values of coefficient of variation and coefficient of Kurtosis of the auxiliary variable under SRS, which are given below:

i) If $\alpha = 1$ and $\gamma = V_x$, the estimator be :

$$\hat{\mu}_{SRS1} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(\bar{x}+V_x)}(\mu_x + V_x)$$

where V_x is the coefficient of variation defined as $V_x = \sigma_x / \mu_x$

ii) If $\alpha = 1$ and $\gamma = K_x$, then the estimator be :

$$\hat{\mu}_{SRS2} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(\bar{x}+K_x)}(\mu_x + K_x)$$

where K_x is the coefficient of Kurtosis defined

as $K_x = \mu_{x4} / \mu_{x2}^2$, *where* $\mu_{xr} = E(X-\mu_x)^r$

iii) If $\alpha = K_x$ and $\gamma = V_x$, then the estimator be :

$$\hat{\mu}_{SRS3} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(K_x\bar{x}+V_x)}(K_x\mu_x + V_x)$$

iv) If $\alpha = V_x$ and $\gamma = K_x$, then the estimator be :

$$\hat{\mu}_{SRS4} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(V_x\bar{x}+K_x)}(V_x\mu_x + K_x)$$

which attain a general form given as below :

$$\bar{y}_{Ri} = \frac{\bar{y}+\beta(\mu_x-\bar{x})}{(\alpha\bar{x}+\gamma)}(\alpha\mu_x+\gamma)$$

for i = 1,2,3,4

The mean square error (MSE) of the above estimators are approximately

$$MSE(\bar{y}_R) \cong \frac{1-f}{n}\left[R_i^2\sigma_x^2 + \sigma_y^2(1-\rho^2)\right]$$

$$\cong MSE(\bar{y}_{Ri}) = \theta\bar{Y}^2\left[C_y^2 + D_i^2C_x^2 - 2D_i\rho C_yC_x\right]$$

for i= 1,2,3,4

where

$$D_i = \frac{\mu_y}{\alpha\mu_x+\gamma}$$

$D_1 = \frac{\mu_y}{\mu_x + V_x}$, $D_2 = \frac{\mu_y}{\mu_x + K_x}$, $D_3 = \frac{\mu_y K_x}{\mu_x K_x + V_x}$, and $D_4 = \frac{\mu_y V_x}{V_x \mu_x + K_x}$

$C_y = \frac{S_y}{\bar{Y}}$, $C_x = \frac{S_x}{\bar{X}}$, $\rho = \frac{S_{yx}}{S_y S_x}$, $\theta = \frac{1}{n}$ $(on\ ignoring, f = {}^{n}/_{N})$

$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{Y})^2}{N-1}$,

$S_x^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{X})^2}{N-1}$,

$S_{yx} = \frac{\sum(y_i - \bar{Y})(x_i - \bar{X})}{N-1}$

5.2 Ratio estimation under rank set sampling

Samawi and Muttlak (1996) suggested modified ratio estimators in case of rank set sampling. The procedure of ratio estimation under rank set sampling given by Samawi and Muttlak (1996) are as under:

Let *Y* be the variable if interest and *X* be a suitable concomitant variable which is correlated to Y and easy to rank. The summary of the RSS procedure is then as follows:

1. Select randomly m^2 bivariate units *(X,Y)* from the population.
2. Allocate the chosen units into *m* sets each if size *m*.
3. From the first set, the smallest X and the associated Y are measured. From the second set, the second smallest of *X* and the associated *Y* are measured, We continue in this way until the last set where the largest X and the associated Y are measured.
4. Repeat the steps above *r* times until getting the required number of elements.

The associated variable *Y* is then with error unless the relation between *X* and *Y* is perfect. Let us denote $(X_{j(i)}, Y_{j[i]})$ as the pair of the i^{th} order statistics of *X* and the associated element Y in the J^{th} cycle. Then the ranked set sample is $(X_{1(1)}, Y_{1[1]}),\ldots, (X_{1(m)}, Y_{1[m]})$,

$(X_{2(1)}, Y_{2[1]}),\ldots, (X_{2(m)}, Y_{2[m]}),$

.
.
.

$(X_{r(1)}, Y_{r[1]}),\ldots, (X_{r(m)}, Y_{r[m]})$

Then we define the sample means based on RSS by

$\bar{X}^* = (1/n)\sum_{i=1}^{n} X_{(i)}$ and $\bar{Y}^* = (1/n)\sum_{i=1}^{n} Y_{[i]}$ with variances are

$Var(\bar{X}^*) = \frac{\sigma_x^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{x(i)} - \mu_x)^2,$

$Var(\bar{Y}^*) = \frac{\sigma_y^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{y[i]} - \mu_y)^2$ and

$Cov(\bar{X}^*, \bar{Y}^*) = (1/m)\sigma_{xy} - (1/m^2)\sum_{i=1}^{m} k_{xy[i]}$ with

$$k_{xy[i]} = (\mu_{x(i)} - \mu_x)(\mu_{y[i]} - \mu_y)$$

$\hat{R}=\bar{y}^*/\bar{x}^*$ where $\bar{x}^* = \frac{1}{mr}\sum_{k=1}^{r}\sum_{i=1}^{m} X_{k(i)}$ and

$$\bar{y}^* = \frac{1}{mr}\sum_{k=1}^{r}\sum_{i=1}^{m} Y_{k(i)}$$

and the ratio estimator of the population mean of Y is

$$\bar{y}_{Rss} = \bar{y}^*\left(\frac{\bar{X}}{\bar{x}^*}\right)$$

and MSE is given below :

$$MSE(\bar{y}_{Rss}) = \theta\bar{Y}^2\left[C_y^2 + C_x^2 - 2\rho C_y C_x\right] - \{Z_{y(i)} - Z_{x(i)}\}^2$$

$$\theta = \frac{1}{mr},$$

$$C_y^2 = \frac{S_y^2}{\bar{Y}^2},$$

$$C_x^2 = \frac{S_x^2}{\bar{X}^2},$$

$$C_{yx} = \frac{S_{yx}}{\bar{X}\bar{Y}} = \rho C_y C_x,$$

$$Z_{x(i)}^2 = \frac{1}{m^2 r}\frac{1}{\bar{X}^2}\sum_{i=1}^{m} k_{x(i)}^2,$$

$$Z^2_{y(i)} = \frac{1}{m^2 r}\frac{1}{\bar{X}^2}\sum_{i=1}^{m} k^2_{y(i)},$$

$$Z_{yx(i)} = \frac{1}{m^2 r}\frac{1}{\bar{Y}\bar{X}}\sum_{i=1}^{m} k_{yx(i)}$$

Also $k_{x(i)} = (\mu_{x(i)} - \bar{X}), k_{y(i)} = (\mu_{y(i)} - \bar{Y})$

$$k_{yx(i)} = (\mu_{x(i)} - \bar{X})(\mu_{y(i)} - \bar{Y})$$

Using one degree Taylor series expansion, they showed that this estimator is more efficient than that from SRS with similar form.

5.3 New/proposed ratio estimators under Rank set sampling based on linear combination of known values of Median, Quartile deviation, coefficient of Skewness, Kurtosis, Correlation Coefficient of auxiliary variable

In this chapter we suggest to use similar form of estimators under RSS as given by Kadilar and Chingi (2004), Koyuncu and Kadilar (2009) under SRS. We assume that the population mean of the auxiliary variable is known beforehand, we also assume that the relation between X and Y is positive and approximately linear.

Based on RSS, we suggest ratio-type estimators for the mean in the form

$\hat{\mu}_{RSS} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\alpha\bar{x}^* + \gamma)}(\alpha\mu_x + \gamma)$ where α and γ are positive constants,

and $\beta = \sigma_{xy} / \sigma_x^2$

Let us take some special cases of α and γ involving various combinations of known values of Median, Quartile deviation, coefficient of Skewness, Kurtosis, Correlation of auxiliary variable in case of RSS.

CASE-1: If the Quartile Deviation and Median of the concomitant variable are available, we may choose this parameter to be values for σ and γ in the estimator above. For examples:

If α=1 and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Qd)}(\mu_x + Qd)$$

If α=1 and $\gamma = Md$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Md)}(\mu_x + Md)$$

If α=*Md* and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Md\bar{x}^* + Qd)}(Md\mu_x + Qd)$$

If α=*Qd* and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Qd\bar{x}^* + Md)}(Qd\mu_x + Md)$$

CASE-2: If the Quartile Deviation and Coefficient of Skewness (*Sk*) of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Sk$, then we have the estimator

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Sk)$$

If α=1 and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Qd)}(\mu_x + Qd)$$

If α= *Sk* and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Sk\bar{x}^* + Qd)}(Sk\mu_x + Qd)$$

If α=*Qd* and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Qd\bar{x}^* + Sk)}(Qd\mu_x + Sk)$$

CASE-3: If the Quartile Deviation and Coefficient of Variation ($V_x = \sigma_x / \mu_x$) of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If $\alpha = V_x$ and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(V_x \bar{x}^* + Qd)}(V_x \mu_x + Qd)$$

If α=1 and $\gamma = Vx$, then we have the estimator :

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Vx)}(\mu_x + Vx)$$

If $\alpha = Vx$ and $\gamma = Qd$, then we have the estimator

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Vx\, \bar{x}^* + Qd)}(Vx\, \mu_x + Qd)$$

If $\alpha = Qd$ and $\gamma = V_x$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Qd\bar{x}^* + V_x)}(Qd\mu_x + V_x)$$

CASE-4: If the Coefficient of Variation ($V_x = \sigma_x / \mu_x$) and Median of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Md)}(\mu_x + Md)$$

If α=1 and $\gamma = V_x$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + V_x)}(\mu_x + V_x)$$

If $\alpha = V_x$ and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(V_x \bar{x}^* + Md)}(V_x \mu_x + Md)$$

If $\alpha = Md$ and $\gamma = V_x$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Md\bar{x}^* + V_x)}(Md\mu_x + V_x)$$

CASE-5: If the Coefficient of Variation ($V_x = \sigma_x / \mu_x$) and Coefficient of Skewness (*Sk*) of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Sk)$$

If α=1 and $\gamma = V_x$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + V_x)}(\mu_x + V_x)$$

If $\alpha = V_x$ and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(V_x\,\bar{x}^* + Sk)}(V_x\mu_x + Sk)$$

If α=*Sk* and $\gamma = V_x$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Sk\bar{x}^* + V_x)}(Sk\mu_x + V_x)$$

CASE-6: If the Coefficient of Kurtosis K_x and Median of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Kx)}(\mu_x + Kx)$$

If α=1 and $\gamma = Md$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Md)}(\mu_x + Md)$$

If $\alpha = Md$ and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Md\,\bar{x}^* + Kx)}(Md\mu_x + Kx)$$

If $\alpha = K_x$ and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Kx\bar{x}^* + Md)}(Kx\mu_x + Md)$$

CASE-7: If the Coefficient of correlation ρ and Coefficient of Skewness (*Sk*) of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = \rho$, then we have the estimator

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + \rho)}(\mu_x + \rho)$$

If α=1 and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Sk)$$

If $\alpha = \rho$ and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\rho\,\bar{x}^* + Sk)}(\rho\mu_x + Sk)$$

If α=*Sk* and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Sk\bar{x}^* + \rho)}(Sk\mu_x + \rho)$$

CASE-8: If the Coefficient of correlation ρ and Quartile deviation *Qd* of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Qd)$$

If α=1 and $\gamma = \rho$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + \rho)}(\mu_x + \rho)$$

If $\alpha = \rho$ and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\rho\,\bar{x}^* + Qd)}(\rho\mu_x + Qd)$$

If α=*Qd* and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Qd\bar{x}^* + \rho)}(Qd\mu_x + \rho)$$

CASE-9: If the Coefficient of correlation ρ and Coefficient of Kurtosis K_x of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Kx)}(\mu_x + Kx)$$

If α=1 and $\gamma = \rho$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + \rho)}(\mu_x + \rho)$$

If $\alpha = \rho$ and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\rho\bar{x}^* + Kx)}(\rho\mu_x + Kx)$$

If $\alpha = K_x$ and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Kx\bar{x}^* + \rho)}(Kx\mu_x + \rho)$$

CASE-10: If the Coefficient of correlation ρ and Median *Md* of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Md)}(\mu_x + Md)$$

If α=1 and $\gamma = \rho$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + \rho)}(\mu_x + \rho)$$

If $\alpha = \rho$ and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\rho\bar{x}^* + Md)}(\rho\mu_x + Md)$$

If α= Md and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Md\,\bar{x}^* + \rho)}(Md\mu_x + \rho)$$

CASE-11: If the Quartile deviation Qd and Coefficient of Kurtosis K_x of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Kx)}(\mu_x + Kx)$$

If α=1 and $\gamma = Qd$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Qd)}(\mu_x + Qd)$$

If α= Qd and $\gamma = Kx$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Qd\,\bar{x}^* + Kx)}(Qd\mu_x + Kx)$$

If α= K_x and $\gamma = Qd$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Kx\bar{x}^* + Qd)}(Kx\mu_x + Qd)$$

CASE-12: If the Median and Coefficient of Skewness S_k of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Sk)$$

If α=1 and $\gamma = Md$, then we have the estimator

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Md)}(\mu_x + Md)$$

If α=Md and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Md\,\bar{x}^* + Sk)}(Md\mu_x + Sk)$$

If $\alpha = Sk$ and $\gamma = Md$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Sk\bar{x}^* + Md)}(Sk\mu_x + Md)$$

CASE-13: If the Coefficient of Kurtosis K_x and Coefficient of Skewness S_k of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = Md$, then we have the estimator

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Kx)}(\mu_x + Kx)$$

If α=1 and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + Sk)}(\mu_x + Sk)$$

If $\alpha = K_x$ and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Kx\ \bar{x}^* + Sk)}(Kx\ \mu_x + Sk)$$

If $\alpha = K_x$ and $\gamma = Sk$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(Sk\bar{x}^* + Kx)}(Sk\mu_x + Kx)$$

CASE-14: If the Coefficient of correlation ρ and Correlation variation V_x of the concomitant variable are available, we may choose this parameter to be values for α and γ in the estimator above. For examples:

If α=1 and $\gamma = V_x$, then we have the estimator :

$$\hat{\mu}_{RSS1} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + V_x)}(\mu_x + V_x)$$

If α=1 and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS2} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\bar{x}^* + \rho)}(\mu_x + \rho)$$

If $\alpha = \rho$ and $\gamma = V_x$, then we have the estimator :

$$\hat{\mu}_{RSS3} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\rho\,\bar{x}^* + V_x)}(\rho\mu_x + V_x)$$

If $\alpha = V_x$ and $\gamma = \rho$, then we have the estimator :

$$\hat{\mu}_{RSS4} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(V_x\,\bar{x}^* + \rho)}(V_x\mu_x + \rho)$$

Similarly replacing the $\bar{y}^*$ and $\bar{x}^*$ by $\bar{y}$ and $\bar{x}$, the above proposed ratio estimators of RSS based on linear combinations of known values of Median, Quartile deviation, coefficient of Correlation, Skewness and Kurtosis of the ranking or auxiliary variable can be used in case of SRS also

As we know, $\bar{y}_{Rss} = \bar{y}^*\left(\frac{\bar{x}}{\bar{x}^*}\right)$ also we can denote $\bar{y}_{Rss}$ as $\hat{\mu}_{Rss}$

Where

$$\bar{x}^* = \frac{1}{mr}\sum_{k=1}^{r}\sum_{i=1}^{m} X_{k(i)}$$

$$\bar{y}^* = \frac{1}{mr}\sum_{k=1}^{r}\sum_{i=1}^{m} Y_{k(i)}$$

Then

$$\bar{y}_{Rssi} = \frac{\bar{y}^* + \beta(\mu_x - \bar{x}^*)}{(\alpha\bar{x}^* + \gamma)}(\alpha\mu_x + \gamma)$$ for i= 1,2,3,4,……,56 (from case 1 to case 14)

The bias and mean square error of the above estimator based on RSS will be :

$$Bias(\bar{y}_{Rssi}) = \bar{Y}\left[\theta\left(G_i^2 C_x^2 - G_i\rho C_y C_x\right) - \left\{G_i^2 Z_{x(i)}^2 - G_i Z_{yx(i)}\right\}\right]$$

$$MSE(\bar{y}_{Rssi}) = \bar{Y}^2\left[\theta\left\{C_y^2 + G_i^2 C_x^2 - 2G_i\rho C_y C_x\right\} - \left\{Z_{y(i)} - G_i Z_{x(i)}\right\}^2\right]$$

Where $G_i = \frac{\gamma\alpha}{\alpha\bar{X}+\gamma}$

Where α and γ can be replaced by various linear combinations of known values of Median, quartile deviation, Coefficient of Skewness, kurtosis, Correlation and variation of the auxiliary variable which are mention in the above 14 cases

5.4 Bias and Mean square estimation of proposed estimators

Since the bias and mean square error of the above proposed estimators based on RSS will be in the form of :

$Bias(\bar{y}_{Rssi}) = \bar{Y}\left[\theta\left(G_i^2 C_x^2 - G_i \rho C_y C_x\right) - \left\{G_i^2 Z_{x(i)}^2 - G_i Z_{yx(i)}\right\}\right]$ and

$MSE(\bar{y}_{Rssi}) = \bar{Y}^2\left[\theta\left\{C_y^2 + G_i^2 C_x^2 - 2G_i \rho C_y C_x\right\} - \left\{Z_{y(i)} - G_i Z_{x(i)}\right\}^2\right]$ respectively

where $G_i = \frac{\bar{Y}\alpha}{\alpha\bar{X}+\gamma}$

To obtain the bias and MSE of $\bar{y}_{Rssi}$

We put $\bar{y}^* = \bar{Y}(1+\varepsilon_0)$ and $\bar{x}^* = \bar{X}(1+\varepsilon_1)$, so that the expectation of ε_0 and ε_1 is equal to zero

$E(\varepsilon_0) = E(\varepsilon_1) = 0$

$V(\varepsilon_0) = E(\varepsilon_0^2) = \frac{V(\bar{y}^*)}{\bar{Y}^2}$

$= \frac{1}{mr}\frac{1}{\bar{Y}^2}\left[\frac{1}{m}\sum_{i=1}^{m} k_{y(i)}^2\right] = [\theta C_y^2 - Z_{y(i)}^2]$

Similarly

$V(\varepsilon_1) = E(\varepsilon_1^2) = [\theta C_x^2 - Z_{x(i)}^2]$

and

$Cov(\varepsilon_0, \varepsilon_1) = E(\varepsilon_0, \varepsilon_1) = \frac{Cov(\bar{y}^*\bar{x}^*)}{\bar{Y}\bar{X}}$

$= \frac{1}{\bar{Y}\bar{X}}\frac{1}{mr}\left[S_{yx} - \frac{1}{m}\sum_{i=1}^{m} k_{yx(i)}\right] = [\theta\rho C_y C_x - Z_{yx(i)}]$

where

$\theta = \frac{1}{mr}$,

$$C_y^2 = \frac{S_y^2}{\bar{Y}^2},$$

$$C_x^2 = \frac{S_x^2}{\bar{X}^2},$$

$$C_{yx} = \frac{S_{yx}}{\bar{X}\bar{Y}} = \rho C_y C_x,$$

$$Z_{x(i)}^2 = \frac{1}{m^2 r}\frac{1}{\bar{X}^2}\sum_{i=1}^{m} k_{x(i)}^2,$$

$$Z_{y(i)}^2 = \frac{1}{m^2 r}\frac{1}{\bar{Y}^2}\sum_{i=1}^{m} k_{y(i)}^2,$$

$$Z_{yx(i)} = \frac{1}{m^2 r}\frac{1}{\bar{Y}\bar{X}}\sum_{i=1}^{m} k_{yx(i)}$$

Also $k_{x(i)} = (\mu_{x(i)} - \bar{X}), k_{y(i)} = (\mu_{y(i)} - \bar{Y})$
and
$k_{yx(i)} = (\mu_{x(i)} - \bar{X})(\mu_{y(i)} - \bar{Y})$

Proof of Bias :

$Bias(\bar{y}_{Rssi}) = E(\bar{y}_{Rssi} - \bar{Y})$

Here $\bar{y}_{Rssi} = \bar{Y}(1+\varepsilon_0)(1+G_i\varepsilon_1)^{-1}$

Suppose $(G_i\varepsilon_1) < 1$ so that $(1+G_i\varepsilon_1)^{-1}$ is expandable

$\bar{y}_{Rssi} = \bar{Y}(1+\varepsilon_0)\{1 - G_i\varepsilon_1 + G_i^2\varepsilon_1^2 + O(G_1\varepsilon_1)\}$

(Using Taylor series expansion, where $O(\varepsilon_1)$ with power more than 2 are neglected for large power of ε_1.

$Bias(\bar{y}_{Rssi}) = \bar{Y}[G_i^2 E(\varepsilon_1^2) - G_i E(\varepsilon_0\varepsilon_1)]$

$= \bar{Y}[G_i^2(\theta C_x^2 - Z_{x(i)}^2) - G_i(\theta\rho C_y C_x - Z_{yx(i)})]$

$\Rightarrow\gg Bias(\bar{y}_{Rssi}) = \bar{Y}[\theta(G_i^2 C_x^2 - G_i\rho C_y C_x) - \{G_i^2 Z_{x(i)}^2 - G_i Z_{yx(i)}\})]$

Proof of Mean Square Error:

$MSE(\bar{y}_{Rssi}) = E(\bar{y}_{Rssi} - \bar{Y})^2$

$$= \bar{Y}^2 E[\varepsilon_0 - G_i\varepsilon_1 + G_i^2\varepsilon_1^2 - 2G_i\varepsilon_0\varepsilon_1]^2$$

$$= \bar{Y}^2 E[\varepsilon_0^2 + G_i^2\varepsilon_1^2 - 2G_i\varepsilon_0\varepsilon_1]$$

$$= \bar{Y}^2[\theta C_y^2 - Z_{y(i)}^2 + G_i^2(\theta C_x^2 - Z_{x(i)}^2) - 2G_i(\theta\rho C_y C_x - Z_{yx(i)})]$$

$$= \bar{Y}^2[\theta\{C_y^2 + G_i^2 C_x^2 - 2G_i\rho C_y C_x\} - \{Z_{y(i)}^2 + G_i^2 Z_{x(i)}^2 - 2G_i Z_{yx(i)}\}]$$

$$\Rightarrow\gg MSE(\bar{y}_{Rssl}) = \bar{Y}^2[\theta\{C_y^2 + G_i^2 C_x^2 - 2G_i\rho C_y C_x\} - \{Z_{y(i)} - G_i Z_{x(i)}\}^2]$$

5.5 Efficiency comparison and numerical illustration

The behaviour of the above new/proposed estimators is studied and compared with the corresponding estimators from SRS. Let us assume that the variable of interest *Y* and a concomitant variable *X* are correlated with a correlation coefficient ρ. Assume also that *X* and *Y* have a bivariate normal distribution with parameter. Using the Pinus data set, different sample generated using SRS and RSS from a bivariate normal distribution using the above combinations of α *and* γ. From this distribution we generated 440 samples based on RSS with 10 cycles and another 440 samples using SRS. For each sample, the mean square errors are computed respectively. The above bivariate normal distribution was generated in R-Software using the function `mvtnorm from library (mvtnorm).` Different values of ρ and *m* are used and the results are shown in Tables-9 to 22.

The efficiency of RSS with respect to corresponding estimators of SRS

$$Ef_i = \frac{MSE\,(\hat{\mu}SRS)_i}{MSE\,(\hat{\mu}RSS)_i} = \frac{\theta\bar{Y}^2[C_y^2 + D_i^2 C_X^2 - 2D_i\rho C_y C_x]}{\bar{Y}^2[\theta\{C_y^2 + G_i^2 C_x^2 - 2G_i\rho C_y C_x\} - \{Z_{y(i)} - G_i Z_{x(i)}\}^2]}, \text{ where } (i = 1,2,3,4)$$

$MSE\,(\hat{\mu}_{SRS})_i - MSE\,(\hat{\mu}_{RSS})_i = A \geq 0$, where A is a non negative value

From the results of Tables-9 to 22 it is concluded the proposed ratio estimators under rank set sampling are more efficient than the ratio estimators based on simple random sampling. Also efficiency of RSS estimators decreases as the correlation coefficient decreases and efficiency increases as the set size *m* is increases.

Table-9 : Case-1: Linear combination of Quartile deviation and Median

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.89	1.86	1.74	1.42
	4	2.55	2.34	1.79	1.49
	6	2.82	2.64	1.88	1.52
	12	3.92	3.13	2.1	1.71
0.70	2	1.43	1.39	1.44	1.12
	4	1.84	1.74	1.85	1.14
	6	1.91	1.76	1.91	1.23
	12	2.24	2.01	2.24	1.35
0.40	2	1.18	1.16	1.22	1.04
	4	1.41	1.32	1.41	1.13
	6	1.49	1.38	1.49	1.26
	12	1.58	1.39	1.58	1.38

Table-10 : Case-2 : Linear combination of Quartile deviation and Coefficient of Skewness

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.09	1.97	1.65	2.12
	4	2.57	2.02	1.72	2.78
	6	2.87	2.11	1.75	3.05
	12	3.36	2.33	1.94	4.15
0.70	2	1.62	1.67	1.35	1.66
	4	1.97	2.08	1.37	2.07
	6	1.99	2.14	1.46	2.14
	12	2.24	2.47	1.58	2.47
0.40	2	1.39	1.45	1.27	1.41
	4	1.55	1.64	1.36	1.64
	6	1.61	1.72	1.49	1.72
	12	1.62	1.81	1.61	1.81

Table-11 : Case-3 :Linear combination of Quartile deviation and Coefficient of Variation

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.95	1.83	1.51	1.98
	4	2.43	1.88	1.58	2.64
	6	2.73	1.97	1.61	2.91
	12	3.22	2.19	1.8	4.01
0.70	2	1.48	1.53	1.21	1.52
	4	1.83	1.94	1.23	1.93
	6	1.85	2	1.32	2
	12	2.1	2.33	1.44	2.33
0.40	2	1.25	1.31	1.13	1.27
	4	1.41	1.5	1.22	1.5
	6	1.47	1.58	1.35	1.58
	12	1.48	1.67	1.47	1.67

Table-12 : Case-4 :Linear combination of Coefficient of Variation and Median

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.02	1.9	1.58	2.05
	4	2.5	1.95	1.65	2.71
	6	2.8	2.04	1.68	2.98
	12	3.29	2.26	1.87	4.08
0.70	2	1.55	1.6	1.28	1.59
	4	1.9	2.01	1.3	2
	6	1.92	2.07	1.39	2.07
	12	2.17	2.4	1.51	2.4
0.40	2	1.32	1.38	1.2	1.34
	4	1.48	1.57	1.29	1.57
	6	1.54	1.65	1.42	1.65
	12	1.55	1.74	1.54	1.74

Table-13 : Case-5 :Linear combination of Coefficient of Variation and coefficient of Skewness

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.99	1.87	1.55	2.02
	4	2.47	1.92	1.62	2.68
	6	2.77	2.01	1.65	2.95
	12	3.26	2.23	1.84	4.05
0.70	2	1.52	1.57	1.25	1.56
	4	1.87	1.98	1.27	1.97
	6	1.89	2.04	1.36	2.04
	12	2.14	2.37	1.48	2.37
0.40	2	1.29	1.35	1.17	1.31
	4	1.45	1.54	1.26	1.54
	6	1.51	1.62	1.39	1.62
	12	1.52	1.71	1.51	1.71

Table-14 : Case-6 :Linear combination of Median and Coefficient of Kurtosis

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.91	1.79	1.47	1.94
	4	2.39	1.84	1.54	2.6
	6	2.69	1.93	1.57	2.87
	12	3.18	2.15	1.76	3.97
0.70	2	1.44	1.49	1.17	1.48
	4	1.79	1.9	1.19	1.89
	6	1.81	1.96	1.28	1.96
	12	2.06	2.29	1.4	2.29
0.40	2	1.21	1.27	1.09	1.23
	4	1.37	1.46	1.18	1.46
	6	1.43	1.54	1.31	1.54
	12	1.44	1.63	1.43	1.63

Table-15 : Case-7 :Linear combination of Coefficient of Correlation and Coefficient of Skewness

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.12	2	1.68	2.15
	4	2.6	2.05	1.75	2.81
	6	2.9	2.14	1.78	3.08
	12	3.39	2.36	1.97	4.18
0.70	2	1.65	1.7	1.38	1.69
	4	2	2.11	1.4	2.1
	6	2.02	2.17	1.49	2.17
	12	2.27	2.5	1.61	2.5
0.40	2	1.42	1.48	1.3	1.44
	4	1.58	1.67	1.39	1.67
	6	1.64	1.75	1.52	1.75
	12	1.65	1.84	1.64	1.84

Table-16 : Case-8 :Linear combination of Coefficient of Correlation and Quartile deviation

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.00	1.88	1.56	2.03
	4	2.48	1.93	1.63	2.69
	6	2.78	2.02	1.66	2.96
	12	3.27	2.24	1.85	4.06
0.70	2	1.53	1.58	1.26	1.57
	4	1.88	1.99	1.28	1.98
	6	1.9	2.05	1.37	2.05
	12	2.15	2.38	1.49	2.38
0.40	2	1.3	1.36	1.18	1.32
	4	1.46	1.55	1.27	1.55
	6	1.52	1.63	1.4	1.63
	12	1.53	1.72	1.52	1.72

Table-17 : Case-9 : Linear combination of Coefficient of Correlation and Coefficient of Kurtosis

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.916	1.886	1.766	1.446
	4	2.576	2.366	1.816	1.516
	6	2.846	2.666	1.906	1.546
	12	3.946	3.156	2.126	1.736
0.70	2	1.456	1.416	1.466	1.146
	4	1.866	1.766	1.876	1.166
	6	1.936	1.786	1.936	1.256
	12	2.266	2.036	2.266	1.376
0.40	2	1.206	1.186	1.246	1.066
	4	1.436	1.346	1.436	1.156
	6	1.516	1.406	1.516	1.286
	12	1.606	1.416	1.606	1.406

Table-18 : Case-10 :Linear combination of Coefficient of Correlation and Median

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.86	1.83	1.71	1.39
	4	2.52	2.31	1.76	1.46
	6	2.79	2.61	1.85	1.49
	12	3.89	3.1	2.07	1.68
0.70	2	1.4	1.36	1.41	1.09
	4	1.81	1.71	1.82	1.11
	6	1.88	1.73	1.88	1.2
	12	2.21	1.98	2.21	1.32
0.40	2	1.15	1.13	1.19	1.01
	4	1.38	1.29	1.38	1.1
	6	1.46	1.35	1.46	1.23
	12	1.55	1.36	1.55	1.35

Table-19 : Case-11 : Linear combination of Quartile deviation and Coefficient of Kurtosis

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	1.97	1.94	1.82	1.5
	4	2.63	2.42	1.87	1.57
	6	2.9	2.72	1.96	1.6
	12	4	3.21	2.18	1.79
0.70	2	1.51	1.47	1.52	1.2
	4	1.92	1.82	1.93	1.22
	6	1.99	1.84	1.99	1.31
	12	2.32	2.09	2.32	1.43
0.40	2	1.26	1.24	1.3	1.12
	4	1.49	1.4	1.49	1.21
	6	1.57	1.46	1.57	1.34
	12	1.66	1.47	1.66	1.46

Table-20 : Case-12 :Linear combination of Median and Coefficient of Skewness

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.09	2.06	1.94	1.62
	4	2.75	2.54	1.99	1.69
	6	3.02	2.84	2.08	1.72
	12	4.12	3.33	2.3	1.91
0.70	2	1.63	1.59	1.64	1.32
	4	2.04	1.94	2.05	1.34
	6	2.11	1.96	2.11	1.43
	12	2.44	2.21	2.44	1.55
0.40	2	1.38	1.36	1.42	1.24
	4	1.61	1.52	1.61	1.33
	6	1.69	1.58	1.69	1.46
	12	1.78	1.59	1.78	1.58

Table-21 : Case-13 :Linear combination of Coefficient of Skewness and Coefficient of Kurtosis

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.17	2.14	2.02	1.7
	4	2.83	2.62	2.07	1.77
	6	3.1	2.92	2.16	1.8
	12	4.2	3.41	2.38	1.99
0.70	2	1.71	1.67	1.72	1.4
	4	2.12	2.02	2.13	1.42
	6	2.19	2.04	2.19	1.51
	12	2.52	2.29	2.52	1.63
0.40	2	1.46	1.44	1.5	1.32
	4	1.69	1.6	1.69	1.41
	6	1.77	1.66	1.77	1.54
	12	1.86	1.67	1.86	1.66

Table-22 : Case-14 :Linear combination of Coefficient of variation and Correlation

ρ	m	Ef_1	Ef_2	Ef_3	Ef_4
0.99	2	2.04	2.01	1.89	1.57
	4	2.7	2.49	1.94	1.64
	6	2.97	2.79	2.03	1.67
	12	4.07	3.28	2.25	1.86
0.70	2	1.58	1.54	1.59	1.27
	4	1.99	1.89	2	1.29
	6	2.06	1.91	2.06	1.38
	12	2.39	2.16	2.39	1.5
0.40	2	1.33	1.31	1.37	1.19
	4	1.56	1.47	1.56	1.28
	6	1.64	1.53	1.64	1.41
	12	1.73	1.54	1.73	1.53

5.6 New/proposed estimators using Deciles of Auxiliary variable under SRS and RSS schemes

In this section, we suggest new/proposed ratio estimators of population mean of the study variable using the linear combination of known values of Deciles of the auxiliary variable under rank set sampling and simple random sampling schemes. Mean square error of the proposed estimators under rank set sampling is calculated and compared. By this comparison, we demonstrate theoretically and numerically that the proposed estimators under rank set sampling are more efficient than estimators based on simple random sampling.

5.6.1 New/proposed estimators

Under simple random sampling let (X_1,Y_1), (X_2,Y_2), (X_3,Y_3),……..., (X_m,Y_m) be a bivariate random sample with probability density function (pdf) $f(x,y)$, cumulative distribution function *(cdf)* $F(x,y)$, means μ_x, μ_y, variances σ_x^2, σ_y^2 and correlation coefficient ρ.

Let $(X_{1(1)}, Y_{1[1]})$,…, $(X_{1(m)}, Y_{1[m]})$, $(X_{2(1)}, Y_{2[1]})$,…, $(X_{2(m)}, Y_{2[m]})$,…….. $(X_{r(1)}, Y_{r[1]})$,…, $(X_{r\,(m)}, Y_{r\,[m]})$ be m independent bivariate samples each of size m, then the proposed ratio estimators under SRS using deciles of auxiliary variable will take the form as given below :

$$\hat{\mu}_{YSRS1} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_1}{\bar{X}_{SRS} + D_1}\right),$$

$$\hat{\mu}_{YSRS2} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_2}{\bar{X}_{SRS} + D_2}\right),$$

$$\hat{\mu}_{YSRS3} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_3}{\bar{X}_{SRS} + D_3}\right),$$

$$\hat{\mu}_{YSRS4} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_4}{\bar{X}_{SRS} + D_4}\right),$$

$$\hat{\mu}_{YSRS5} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_5}{\bar{X}_{SRS} + D_5}\right),$$

$$\hat{\mu}_{YSRS6} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_6}{\bar{X}_{SRS} + D_6}\right),$$

$$\hat{\mu}_{YSRS7} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_7}{\bar{X}_{SRS} + D_7}\right),$$

$$\hat{\mu}_{YSRS8} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_8}{\bar{X}_{SRS} + D_8}\right),$$

$$\hat{\mu}_{YSRS9} = \bar{Y}_{SRS}\left(\frac{\mu_X + D_9}{\bar{X}_{SRS} + D_9}\right),$$

respectively, where $\bar{X}_{SRS} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\bar{Y}_{SRS} = = \frac{1}{m}\sum_{i=1}^{m} Y_i$ are sample means of auxiliary variable *X* and *Y*, D_h *are deciles, for* $h = 1,2,\ldots,9$

Using Taylor series, $\hat{\mu}_{YSRSh}\,(h = 1,2,3\ldots,9)$ can be approximated as :

$$\bar{\mu}_{YSRSh} \cong \bar{Y}_{SRS} - L_h(\bar{X}_{SRS} - \mu_X) + L_h G_h(\bar{X}_{SRS} - \mu_X)^2 - G_h(\bar{X}_{SRS} - \mu_X)(\bar{Y}_{SRS} - \mu_Y)$$

where $L_h = \frac{\mu_Y}{\mu_X + D_h}$ and $G_h = \frac{1}{\mu_X + D_h}$ for $h = 1,2,\ldots,9$. Using the first degree approximation to the above equation, the estimator in above equation is given by :

$$\bar{\mu}_{YSRSh} \cong \bar{Y}_{SRS} - L_h(\bar{X}_{SRS} - \mu_X),$$

with bias and MSE respectively are given as

$$Bias\,(\bar{\mu}_{YSRSh}) \cong 0,$$

and

$MSE(\bar{\mu}_{YSRSh}) \cong Var(\bar{Y}_{SRS}) + L_h^2 Var(\bar{X}_{SRS} - 2L_h Cov(\bar{X}_{SRS}, \bar{Y}_{SRS})$, respectively, where, $Cov(\bar{X}_{SRS}, \bar{Y}_{SRS}) = E\big((\bar{X}_{SRS} - \mu_X)(\bar{Y}_{SRS} - \mu_Y)\big)$. It can be noted for the first order approximation the estimators are unbiased :

Using

$$Cov(\bar{X}_{SRS}, \bar{Y}_{SRS}) = \beta\, var\,(\bar{X}_{SRS})$$

$$var(\bar{Y}_{SRS}) \cong \beta^2 var\,(\bar{X}_{SRS}) + \frac{1}{m}\rho_Y^2(1 - \rho^2),$$

where $\beta = \rho\frac{\sigma_Y}{\sigma_X}$.

The MSE of $\bar{\mu}_{YSRSh}$ therefore can be written as under

$$MSE(\bar{\mu}_{YSRSh}) \cong (L_h - \beta)^2 Var(\bar{X}_{SRS}) + \frac{1}{m}\sigma_Y^2(1-\rho^2)$$

$$\bar{\mu}_{YSRSh} = \frac{1}{m}(L_h^2\sigma_x^2 + \sigma_y^2 - 2L_h\sigma_x\sigma_y\rho)$$

Under rank set sampling, we assume that the ranking is performed on the variable *X* for estimating the population mean of the variable *Y*. The RSS method can be described as follows: Select *m* random samples each of size *m* bivariate units from the target population. From the first set of *m* units, the smallest ranked unit *X* is selected together with the associated *Y*, and from the second set of *m* units the second smallest ranked unit *X* is selected together with the associated *Y*. The procedure is continued until from the m_{th} set of *m* units the largest ranked unit *X* is selected with the associated *Y*. The procedure can be repeated *n* times to increase the sample size to *nm* RSS bivariate units.

Let $(X_{i(1)},Y_{i(1)})$, $(X_{i(2)},Y_{i(2)})\ldots\ldots\ldots,(X_{i(m)},Y_{i(m)})$ be the order statistics of $X_{i1}, X_{i2},\ldots\ldots, X_{im}$ and the judgment order of $Y_{i1}, Y_{i2},\ldots..,Y_{im}$ *for* $i= 1,2,\ldots,m$. Then the RSS units are $(X_{1(1)},Y_{1(1)})$, $(X_{2(2)},Y_{2(2)})$,……, $(X_{m(m)},Y_{m(m)})$. The proposed estimators of population mean μ_Y involving the deciles of the auxiliary variable X are defined below :

$$\hat{\mu}_{YRSS1} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_1}{\bar{X}_{RSS} + D_1}\right),$$

$$\hat{\mu}_{YRSS2} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_2}{\bar{X}_{RSS} + D_2}\right),$$

$$\hat{\mu}_{YRSS3} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_3}{\bar{X}_{RSS} + D_3}\right),$$

$$\hat{\mu}_{YRSS4} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_4}{\bar{X}_{RSS} + D_4}\right),$$

$$\hat{\mu}_{YRSS5} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_5}{\bar{X}_{RSS} + D_5}\right),$$

$$\hat{\mu}_{YRSS6} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_6}{\bar{X}_{RSS} + D_6}\right),$$

$$\hat{\mu}_{YRSS7} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_7}{\bar{X}_{RSS} + D_7}\right),$$

$$\hat{\mu}_{YRSS8} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_8}{\bar{X}_{RSS} + D_8}\right),$$

$$\hat{\mu}_{YRSS9} = \bar{Y}_{RSS}\left(\frac{\mu_X + D_9}{\bar{X}_{RSS} + D_9}\right),$$

respectively, where $\bar{X}_{RSSS} = \frac{1}{m}\sum_{i=1}^{m} X_i$ and $\bar{Y}_{RSS} = = \frac{1}{m}\sum_{i=1}^{m} Y_i$ with variances $var(\bar{X}_{RSS}) = \frac{\sigma_X^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{X[i]} - \mu_X)^2$

and $var(\bar{Y}_{RSS}) = \frac{\sigma_Y^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{Y[i]} - \mu_Y)^2$,

D_h *are deciles, for* $h = 1,2,\ldots,9$

$var(\bar{X}_{RSS}) = \frac{\sigma_X^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{X[i]} - \mu_X)^2$, and $\bar{Y}_{RSS} = \frac{1}{m}\sum_{i=1}^{m} Y_{i(i)}$ with variance $var(\bar{Y}_{RSS}) = \frac{\sigma_Y^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{Y[i]} - \mu_Y)^2$.

Using Taylor series, $\hat{\mu}_{YRSSh}\,(h = 1,2,3\ldots,9)$ can be approximated as :

$$\bar{\mu}_{YRSSh} \cong \bar{Y}_{RSS} - L_h(\bar{X}_{RSS} - \mu_X) + L_h G_h(\bar{X}_{RSS} - \mu_X)^2 - G_h(\bar{X}_{RSS} - \mu_X)(\bar{Y}_{RSS} - \mu_Y)$$

For the first order approximation, the above estimator can be written as

with bias and MSE respectively are given as

$$Bias\,(\bar{\mu}_{YRSSh}) \cong 0,$$

and

$$MSE(\bar{\mu}_{YRSSh}) \cong Var(\bar{Y}_{RSS}) + L_h^2 Var(\bar{X}_{RSS} - 2L_h Cov(\bar{X}_{RSS}, \bar{Y}_{RSS}),$$

respectively,

where $Cov(\bar{X}_{RSS}, \bar{Y}_{RSS}) = E\big((\bar{X}_{RSS} - \mu_X)(\bar{Y}_{RSS} - \mu_Y)\big)$. As in case of SRS, it is clear that to the first order of approximation the RSS estimators are unbiased.

using

$$Cov(\bar{X}_{RSS}, \bar{Y}_{RSS}) = \beta\, var\,(\bar{X}_{RSS})$$

$$var(\bar{Y}_{RSS}) \cong \beta^2 var\,(\bar{X}_{RSS}) + \frac{1}{m}\rho_Y^2(1-\rho^2),$$

The MSE of $\bar{\mu}_{YRSSh}$ therefore can be written as under :

$$MSE(\bar{\mu}_{YRSSh}) \cong (L_h - \beta)^2 Var(\bar{X}_{RSS}) + \frac{1}{m}\sigma_Y^2(1-\rho^2)$$

$$= (L_h - \beta)^2 \left(\frac{\sigma_X^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{X(i)} - \mu_X)^2\right) + \frac{1}{m}\sigma_Y^2(1-\rho^2)$$

where $L_h = \frac{\mu_Y}{\mu_X + D_h}$ and $G_h = \frac{1}{\mu_X + D_h}$ for *h=1,2,3….,9*

5.6.2 Relative efficiency of new/proposed estimators

The efficiency of $\hat{\mu}_{YSRSh}$ with respect to $\hat{\mu}_{YRSSh}$ for estimating the population mean μ_Y is defined as :

$$Relative\ efficiency\ (\hat{\mu}_{YRSSh}, \hat{\mu}_{YSRSh}) = \frac{MSE(\hat{\mu}_{YSRSh})}{MSE(\hat{\mu}_{YRSSh})}$$

$$\cong \frac{(L_h - \beta)^2 Var(\bar{X}_{SRS}) + \frac{1}{m}\sigma_Y^2(1-\rho^2)}{(L_h - \beta)^2 Var(\bar{X}_{RSS}) + \frac{1}{m}\sigma_Y^2(1-\rho^2)}$$

$$= \frac{(L_h - \beta)^2 Var(\bar{X}_{SRS}) + \frac{1}{m}\sigma_Y^2(1-\rho^2)}{(L_h - \beta)^2 \left(\frac{\sigma_X^2}{m} - \frac{1}{m^2}\sum_{i=1}^{m}(\mu_{X(i)} - \mu_X)^2\right) + \frac{1}{m}\sigma_Y^2(1-\rho^2)}$$

It is clear that, $Relative\ efficiency\ (\hat{\mu}_{YRSSh}, \hat{\mu}_{YSRSh}) > 1$. This implies that $\hat{\mu}_{YRSSh}\ for\ h = (1,2,3,..9)$ is more efficient than $\hat{\mu}_{YSRSh}$ based on the same number of measured units.

5.7 Numerical illustration

For empirical study `library (mvtnorm)` of R software was used to generate 500 replicates from a population of Pinus data with $\mu_x = 21.44, \mu_y = 15.66, \sigma_x = 20.95, \sigma_y = 17.06$ with ρ (0.99,0.70, 0.40). Based on 500 replications, the results for *m= 4, 8, 12, 16* using deciles (60.60, 83.00, 102.70, 111.20, 142.50, 210.20, 264.50, 304.40, 373.20, 643.00. The results are given in Tables-23.

It is concluded form the Table-23 that estimators under RSS performs better than estimators based on SRS. Utilizing the knowledge of the deciles of the auxiliary variable *X*, a gain in efficiency is obtained using RSS with respect to SRS for estimating the population mean of the variable of interest *Y*. Also the bias of the RSS estimators is small and efficiency of RSS estimators decreases as the correlation coefficient decreases. Finally the efficiency in case of RSS estimators increases as the set size *m* increases.

Table-23 : Efficiency comparison of modified ratio estimators

ρ		D_1				D_2				D_3			
		4	**8**	**12**	**16**	**4**	**8**	**12**	**16**	**4**	**8**	**12**	**16**
0.99	Efficiency	1.632	1.779	1.959	2.079	1.369	1.387	1.421	1.446	1.459	1.534	1.557	1.628
	Bias of RSS	0.073	0.062	0.058	0.055	0.106	0.078	0.066	0.063	0.156	0.1	0.073	0.073
	Bias of SRS	0.089	0.074	0.067	0.063	0.138	0.106	0.092	0.084	0.212	1.013	0.126	0.111
0.70	Efficiency	1.344	1.357	1.396	1.429	1.353	1.435	1.483	1.516	1.438	1.541	1.664	1.727
	Bias of RSS	0.084	0.066	0.059	0.058	0.132	0.091	0.073	0.065	0.168	0.107	0.081	0.073
	Bias of SRS	0.104	0.083	0.075	0.07	0.174	0.131	0.104	0.09	0.231	0.161	0.133	0.115
0.40	Efficiency	1.339	1.341	1.392	1.415	1.346	1.453	1.534	1.574	1.411	1.6	1.686	1.783
	Bias of RSS	0.097	0.073	0.063	0.059	0.142	0.096	0.073	0.206	0.176	0.106	0.091	0.073
	Bias of SRS	0.126	0.093	0.083	0.074	0.194	0.142	0.131	0.112	0.248	0.167	0.141	0.123

Contd…

Table-23 Contd….

ρ		D_4				D_5				D_6			
		4	**8**	**12**	**16**	**4**	**8**	**12**	**16**	**4**	**8**	**12**	**16**
0.99	Efficiency	1.652	1.799	1.979	2.099	1.389	1.407	1.441	1.466	1.479	1.554	1.577	1.648
	Bias of RSS	0.093	0.082	0.078	0.075	0.126	0.098	0.086	0.083	0.176	0.12	0.093	0.093
	Bias of SRS	0.109	0.094	0.087	0.083	0.158	0.126	0.112	0.104	0.232	1.033	0.146	0.131
0.70	Efficiency	1.364	1.377	1.416	1.449	1.373	1.455	1.503	1.536	1.458	1.561	1.684	1.747
	Bias of RSS	0.104	0.086	0.079	0.078	0.152	0.111	0.093	0.085	0.188	0.127	0.101	0.093
	Bias of SRS	0.124	0.103	0.095	0.09	0.194	0.151	0.124	0.11	0.251	0.181	0.153	0.135
0.40	Efficiency	1.359	1.361	1.412	1.435	1.366	1.473	1.554	1.594	1.431	1.62	1.706	1.803
	Bias of RSS	0.117	0.093	0.083	0.079	0.162	0.116	0.093	0.226	0.196	0.126	0.111	0.093
	Bias of SRS	0.146	0.113	0.103	0.094	0.214	0.162	0.132	0.12	0.268	0.187	0.161	0.143

Table-23 Contd….

ρ		D_7				D_8				D_9			
		4	**8**	**12**	**16**	**4**	**8**	**12**	**16**	**4**	**8**	**12**	**16**
0.99	Efficiency	1.803	2.303	2.792	3.243	1.707	1.988	2.303	2.55	1.542	1.686	1.883	1.95
	Bias of RSS	0.365	0.301	0.273	0.258	0.347	0.288	0.265	0.253	0.316	0.277	0.259	0.247
	Bias of SRS	0.444	0.374	0.33	0.31	0.413	0.348	0.317	0.299	0.365	0.315	0.293	0.281
0.70	Efficiency	1.763	2.2	2.665	3.02	1.626	1.847	2.073	2.211	1.503	1.65	1.765	0.848
	Bias of RSS	0.369	0.3	0.265	0.257	0.333	0.281	0.262	0.252	0.308	0.273	0.252	0.247
	Bias of SRS	0.427	0.363	0.33	0.306	0.384	0.331	0.306	0.294	0.352	0.307	0.289	0.275
0.40	Efficiency	1.709	2.077	2.421	2.778	1.573	1.756	1.945	2.047	1.481	1.584	1.711	1.752
	Bias of RSS	0.359	0.289	0.267	0.256	0.328	0.276	0.261	0.248	0.297	0.271	0.253	0.246
	Bias of SRS	0.421	0.351	0.319	0.303	0.37	0.319	0.298	0.755	0.344	0.301	0.281	0.271

Chapter – 6

RANK SET SAMPLING IN SITUATIONS OF NON-RESPONSE WHILE CONSIDERING THE PROBLEMS OF ALLOCATION

In this chapter, a detailed discussion on the effect of non-response on the estimator of population mean under ranked set stratified sampling is given, our main objective is to suggest some new allocation schemes utilizing the knowledge of strata sizes and non-response rates of different strata. The effects of proposed schemes on the sampling variance of the estimator are discussed and compared with the usual allocation scheme, namely, proportional allocation in presence of non-response. Basic method of Non-Response estimation given by Hansen and Hurtwiz (1946) is used in the present study.

6.1 Introduction

In Probability sampling when observation y_i on the i^{th} unit is the correct value for that unit, the error of estimate arises purely from the random sampling variation i.e. when fraction of units is measured instead of the complete population. This deviation of the sample statistics from the population parameters is usually called sampling error. It is well known that if all units of the population are measured, the estimate will be free from sampling error. But in practice it may not be always possible to get the true observation y_i on the i^{th} unit. Consequently the estimate based on sample will also involve errors different from sampling errors. All the errors in estimation, which are not because of sampling, are called non-sampling errors, i.e. these are the residual categories. The sampling errors arise because of sampling on a 'Part' from the 'Whole' population while non-sampling errors mainly arise because of some departure from the prescribed rules of the survey, such as survey design, field work, tabulation, analysis of data etc. This is the reason that the census results though free from sampling errors are subject to various types of non-sampling errors and sometimes these non-sampling errors may be more important than the sampling errors and thus may affect the results substantially. Non sampling errors arise due to numerous factors

and almost at every stage of survey from planning of the survey to report writing. Non-response is one of the important types of non-sampling errors. The first attempt to deal with the problem of non-response was perhaps made by Hansen and Hurwitz (1946) through Call back method. They assumed the population as divided into two classes, (i) response class where respondents respond in the first attempt and (ii) a non-response class where respondents do not respond in the first attempt. Another method to obtain unbiased estimators from the information collected from the respondent in the first attempt only was proposed by Politz and Simmons (1949). Kish and Hess (1954) proposed the adding of a sample of non-responding units from previous surveys for obtaining information about the non-respondents. Non-Response is becoming a concern in survey research, when people are not able or willing to answer questions asked by the interviewer at the time of data collection work. The extent and the effect of the non-response can vary greatly from one type of survey to another. It affects the quality of survey in two ways. Firstly, due to the reduction in available amount of data, the estimates of population parameters will be less precise. Secondly, if a relationship exists between the variable under investigation and response behaviours, statements made on the basis of the response are not valid for the total population. It is obvious that the extent of non-response must be kept as small as possible. In spite of many efforts, there still remains a considerable amount of non-response. In sample surveys, the population may be assumed to be composed of two parts. (i) Response group and (ii) Non- response group. In case when the units of the non-response group are such that after some additional efforts it is possible to get the information we refer such non-responding group of units as "Soft Core". In some cases, a part of non-response units are such that it is impossible to get any information, the set of these units are refereed as "Hard Core". While estimating the population mean of the character under study, the representation of responding and non-responding units is required to give the representative value of the population mean. In case of "Soft Core", the problem is to minimize the effect of non-response and make some adjustment which may provide the efficient

estimate. In case of mail surveys, Hansen and Hurwitz (1946) have suggested the method of sub-sampling from the non-responding units of the sample and proposed the estimator for population mean with its standard errors.

6.2 Hansen and Hurtwiz Technique of Non-Response

The usual theory of sampling is developed to assume that a finite population $U = \{u_1, \ldots \ldots u_N\}$ is composed by individuals that can be perfectly identified. A sample s of size $n \leq N$ is selected. The variable of interest Y is measured in each selected unit. Real life surveys should deal with problems that invalidate some initial assumptions and affect the properties of the statistical models. One of them appears when some of the units in the sample (responding units) do not give a response. The existence of non-response does not permit us to computer sample mean

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Which gives the estimate of μ population mean, because we obtain response only from the units in $s_1 = \{i \in s | i \ give \ a \ response \ at \ first \ visit\}$. This fact suggests that the population U is divided into two strata: U_1 units which give a response at first visit and U_2 which contains rest of the individuals. This is called response starta model and was first proposed by Hansen-Hurvitz (1946). They proposed to select a sub-sample s_2' of size n'_2 among the n_2 non-respondents grouped in the sample s_2 . Then we obtain information on the non-respondent's strata U_2 through s_2'.

6.3 Non-response under Stratified sampling

Let us consider a sample of size *n* is drawn from a finite population of size N. Let n_1 units in the sample responded and n_2 units did not respond, so that

$$n_1 + n_2 = n$$

The n_1 units may be regarded as a sample from the response class and n_2 units as a sample from the non-response class belonging to the population. Let us assume that N_1 and N_2 be the number of units in the response stratum and non-response stratum respectively in the population. Obviously, N_1 and N_2 are not known but their unbiased estimates can be obtained from the sample as

$$\hat{N}_1 = n_1 \frac{N}{n};\ \hat{N}_2 = n_2 \frac{N}{n};$$

Let m be the size of the sub-sample from n_2 non-respondents to be interviewed. Hansen and Hurwitz (1946) proposed an estimator to estimate the population means $\bar{Y}_{hh}$ of the study variable Y as

$$\bar{Y}_{hh} = \frac{n_1 \bar{y}_{01} + n_2 \bar{y}_{0m}}{n}$$

which is unbiased for $\bar{Y}_{hh}$, where $\bar{y}_{01}$ and $\bar{y}_{0m}$ are sample means based on samples of sizes n_1 and m respectively for the study variable.

The variance of $\bar{Y}_{hh}$ is given by

$$V(\bar{Y}_{hh}) = \left[\frac{1}{n} + \frac{1}{N}\right] S_0^2 + \frac{K-1}{n} W_2 S_{02}^2,$$

where $K = \frac{n_2}{m}$, $W_2 = \frac{N_2}{N}$, S_0^2 and S_{02}^2 are the mean squares of entire group and non-response group respectively in the population.

Let us consider a population consisting of N units divided into k strata. Let the size of i^{th} stratum is $N_i\ (1, 2, \ldots, k)$ and we decide to select a sample of size n from the entire population in such a way that n_i units are selected from the i^{th} stratum. Thus, we have

$$\sum_{i=1}^{k} n_i = n$$

Let the non-response occurs in each stratum. Then using Hansen and Hurwitz procedure, we select a sample of size m_i units out of n_{i2} non-respondent units in

the i^{th} stratum such that $n_{i2} = K_i m_i$, $K_i \geq 1$ and the information are observed on all the m_i units by interview method.

The Hansen-Hurwitz estimator of population mean $\bar{Y}_{hhi}$ for the i^{th} stratum will be

$$\bar{Y}_{ssrsi} = \frac{n_{i1}\bar{y}_{0i1} + n_{i2}\bar{y}_{0mi}}{n_i}, (1,2,\ldots,k)$$

where $\bar{y}_{0i1}$ and $\bar{y}_{0mi}$ are the sample means based on n_{i1} respondent units and m_i non-respondent units in the i^{th} stratum.

Combining the estimators over all strata, we get the estimator of population mean Y_{ssrsi}, given by

$$Y_{ssrsi} = \sum_{i=1}^{k} p_i Y_{ssrsi}$$

where $p_i = \frac{N_i}{N}$

Obviously, we have $E[\bar{Y}_{ssrsi}] = Y_{ssrs}$

The variance of $\bar{Y}_{ssrsi}$ is given by

$$V[\bar{Y}_{ssrsi}] = \sum_{i=1}^{k}\left(\frac{1}{n_i} - \frac{1}{N_i}\right)p_i^2 S_{0i}^2 + \sum_{i=1}^{k}\frac{(K_i-1)}{n_i} W_{i2} p_i^2 S_{0i2}^2$$

where $W_{i2} = \frac{N_{i2}}{N_i}$, S_{0i}^2 and S_{0i2}^2 are the mean squares of entire group and non-response group respectively in the i^{th} stratum.

It is easy to see that under 'proportional allocation' (PA), that is, when $n_i = np_i \; for \; i = 1,2,3,\ldots,k$, $V[\bar{Y}_{ssrsi}]$ is obtained as

$$V[\bar{Y}_{ssrsi}]_{PA} = \sum_{i=1}^{k}\left(\frac{1}{n} - \frac{1}{N}\right)p_i S_{0i}^2 + \frac{1}{n}\sum_{i=1}^{k}(K_i - 1)W_{i2} p_i S_{0i2}^2$$

6.4 Non response under ranked set stratified sampling

In this section, we have incorporated RSS for selecting the sub-sample among the non-respondents. RSS procedure is used for sub-sampling s_2. Take a sub-sample $s'_{2(rss)}$ from s using RSS procedure. That is we select n'_2 independent

samples of size n_2/K. The units are ranked accordingly with the variable closely related with the variable of interest Y.

Let there be n_2' independent samples

$Y_{11}, Y_{12} \ldots\ldots\ldots, Y_{1n_2'};\ Y_{21}, Y_{22} \ldots\ldots\ldots, Y_{2n_2'};\ \ldots\ldots\ldots;\ Y_{n_2'1}, Y_{n_2'2} \ldots\ldots\ldots, Y_{n_2'n_2'}$
they are ranked and we obtain

$Y_{1:1}, Y_{2:1} \ldots\ldots\ldots, Y_{n_2':1};\ Y_{1:2}, Y_{2:2} \ldots\ldots\ldots, Y_{n_2':2};\ \ldots\ldots\ldots;\ Y_{1:n_2'}, Y_{2:n_2'} \ldots\ldots\ldots, Y_{n_2':n_2'}$

The estimate of μ_2 is made by using the estimator

$$\bar{y}'_{2(rss)} = \frac{\sum_{j=1}^{n_2'} y_{(j:j)}}{n_2'}$$

$$E\left(\bar{y}'_{2(rss)}\right) = E\left(\frac{\sum_{j=1}^{n_2'} E(y_{(j:j)})}{n_2'}\right) = E[\bar{y}_2] = \mu_2$$

The RSS counter part of $\bar{\bar{y}}$ in SRS is

$$\bar{\bar{y}} = \frac{n_1}{n}\bar{y}_1 + \frac{n_2}{n}\bar{y}'_2 = w_1\bar{y}_1 + w_2\bar{y}'_{2(rss)}$$

It can be represented by

$$\bar{\bar{y}}_{(rss)} = (w_1\bar{y}_1 + w_2\bar{y}_2) + w_2(\bar{y}'_{2(rss)} - \bar{y}_2)$$

Its conditional variance is

$$V(\bar{\bar{y}}_{(rss)}|s) = \frac{\sigma^2}{n} + w_2^2\, V\,(\bar{y}'_{2(rss)} - \bar{y}_2\,|s)$$

Explicit expression of second term in the R.H.S. it is :

$$V(w_2(\bar{y}'_{2(rss)} - \bar{y}_2)|s) = w_2^2 E[(\bar{y}'_{2(rss)} - \mu_2) - (\bar{y}_2 - \mu_2)|s)^2]$$

$$= w_2^2[E(\bar{y}'_{2(rss)} - \mu_2)|s)^2 + E(\bar{y}_2 - \mu_2)|s)^2 - 2E\left((\bar{y}'_{2(rss)} - \mu_2)(\bar{y}_2 - \mu_2)\middle|s\right)]$$

The first term of the equation within brackets is equal to

$$E(\bar{y}'_{2(rss)} - \mu_2)|s)^2 = \frac{\sum_{j=1}^{n_2'} \sigma^2_{(j)}}{n_2'}$$

$$= \frac{\sigma^2}{n_2'} - \frac{\sum_{j=1}^{n_2'} \Delta_{(j)}^2}{n_2'}$$

where

$$\frac{\sum_{j=1}^{n_2'} \Delta_{(j)}^2}{n_2'} = \frac{\sum_{j=i}^{n_2'} (\mu_{(j)} - \mu)^2}{n_2'}$$

The second term is related to :

$$E(\bar{y}_2 - \mu_2)|s)^2 = \frac{\sigma_2^2}{n_2}$$

and

$$E\left(E\left(\bar{y}'_{2(rss)} - \mu_2\right)(\bar{y}_2 - \mu_2)\right)|s) = E(((\bar{y}_2 - \mu_2))^2|s) = \frac{\sigma_2^2}{n_2}$$

Hence the counter part of variance of SRS under non response in case of RSS is given below

$$V\left(w_2\left(\bar{y}'_{2(rss)} - \bar{y}_2\right)|s\right) = w_2^2\left(\frac{\sigma^2_{2(rss)}}{n_2'} - \frac{\sigma_2^2}{n_2}\right) = w_2^2\left(\frac{\sigma^2}{n_2'} - \frac{\sigma_2^2}{n_2} - \frac{\sum_{j=1}^{n_2'} \Delta_{(j)}^2}{n_2'}\right)$$

Substituting $n_2' = {}^{n_2}/_{K}$, we have

$$EV\left(\bar{\bar{y}}_{(rss)}\right) = \frac{\sigma^2}{n} + \frac{W_2(K-1)\sigma_2^2}{n} - W_2 E\left(\frac{K\sum_{j=1}^{n_2'} \Delta_{(j)}^2}{n}\right)$$

which is smaller than SRS.

Using the procedure of Hansen and Hurwitz (1946) ranked stratified set procedure based on McIntyre (1952) is applied to the existing stratified random sampling plan. For this a population is divided into *L* mutually exclusive and exhaustive strata, and a ranked set sample (RSS) of n_h elements is quantified within each stratum, h =1, 2, …, L. The sampling is performed independently across the strata. Therefore, we can think of a RSSS (Ranked set stratified sampling) scheme as a collection of *L* separate ranked set samples. Where $\sum_{h=1}^{L} N_h = N$ is population size and $\sum_{h=1}^{L} n_h = n$ is sample size, then the variance of the estimator of population $\bar{\bar{y}}_{(rssh)}$ based on Hansen and Hurwitz (1946) is given by :

$$V\left(\bar{\bar{y}}_{(rssh)}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)p_h^2 S_{0h}^2 + \sum_{h=1}^{L}\frac{(K_h-1)}{n_h}W_{h2} - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

where, $W_{h2} = \frac{N_{h2}}{N_h}$, S_{0h}^2 and S_{0h2}^2 are the mean squares of entire group and non-response group respectively in the h^{th} stratum and K is inverse ratio of sub-sample among non-response.

Under ranked set, the variance of proportional allocation becomes :

$$V\left[\bar{\bar{y}}_{(rssh)}\right]_{PA} = \sum_{i=1}^{k}\left(\frac{1}{n}-\frac{1}{N}\right)p_h^2 S_{0h}^2 + \sum_{h=1}^{L}\frac{(K_h-1)}{n_h}W_{h2}p_i - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right),$$

New allocation of sample sizes based on various combinations of non-response rate, stratum size and non-response variances under ranked set sampling are given below:

1- $n_h = \frac{np_hW_{h2}}{\sum_h^L p_hW_{h2}}$

2- $n_h = \frac{np_hW_{h2}S_{0h2}^2}{\sum_h^L p_hW_{h2}}$

3- $n_h = \frac{np_hS_{0h2}^2}{\sum_{h=1}^{L}\frac{p_h}{W_{h2}}}$

4 - $n_h = \frac{np_hS_{0h2}^2}{W_{h2}\sum_{h=1}^{L}\frac{p_hS_{0h2}^2}{W_{h2}}}$

5- $n_h = \frac{np_hW_{h2}S_{0h2}^2}{W_{h2}\sum_{h=1}^{L}\frac{p_hW_{h2}}{W_{h2}}}$

6- $n_h = \frac{np_hW_{h2}S_{0h2}^2}{W_{h2}\sum_{h=1}^{L}\frac{W_{h2}S_{0h}}{W_{h2}}}$

Putting the value of n_h in expression

$$\left(\bar{\bar{y}}_{(rssh)}\right) = \sum_{h=1}^{L}\left(\frac{1}{n_h}-\frac{1}{N_h}\right)p_h^2\sigma_{0h}^2 + \sum_{h=1}^{L}\frac{(K_h-1)}{n_h}W_{h2} - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right),$$

the variances based on above six new allocations will take the following form:

$$V[\bar{\bar{y}}rssh]_1 = \frac{1}{n}\left[\sum_{h=1}^{L} p_h W_{h2}\right]\left[\sum_{h=1}^{L}\left\{\frac{p_h S_{0h}^2}{W_{h2}} + \frac{(K_h - 1)}{W_{h2}} W_{h2} p_h S_{0h2}^2\right\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

$$V[\bar{\bar{y}}rssh]_2 = \frac{1}{n}\left[\sum_{h=1}^{L} p_h W_{h2} S_{0h}\right]\left[\sum_{h=1}^{L}\left\{\frac{p_h S_{0h}}{W_{h2}} + \frac{(K_h - 1)}{W_{h2}}\frac{S_{0h2}^2}{S_{0h}}\right\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

$$V[\bar{\bar{y}}rssh]_3 = \frac{1}{n}\left[\sum_{h=1}^{L}\frac{p_h}{W_{h2}}\right]\left[\sum_{h=1}^{L}\{p_h W_{h2} S_{0h}^2 + (K_h - 1) W_{h2}^2 p_h s_{0h}^2\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

$$V[\bar{\bar{y}}rssh]_4 = \frac{1}{n}\left[\sum_{h=1}^{L}\frac{p_h S_{0h}}{W_{h2}}\right]\left[\sum_{h=1}^{L}\left\{p_h W_{h2} S_{0h} + (K_h - 1) W_{h2}^2 p_h \frac{S_{0h}^2}{S_{0h}}\right\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

$$V[\bar{\bar{y}}rssh]_5 = \frac{1}{n}\left[\sum_{h=1}^{L}\frac{p_h W_{h1}}{W_{h2}}\right]\left[\sum_{h=1}^{L}\left\{\frac{p_h W_{h2} S_{0h}^2}{W_{h1}} + \frac{(K_h - 1) W_{h2}^2 p_h S_{0h2}^2}{W_{h2}}\right\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

$$V[\bar{\bar{y}}rssh]_6 = \frac{1}{n}\left[\sum_{h=1}^{L}\frac{p_h W_{h1} S_{0h}}{W_{h2}}\right]\left[\sum_{h=1}^{L}\left\{\frac{p_h W_{h2} S_{0h}}{W_{h1}} + \frac{(K_h - 1) W_{h2}^2 p_h S_{0h2}^2}{W_{h2} S_{0h}}\right\}\right] - S_{0h2}^2\left(\frac{Kh\sum_{h=1}^{n_2'}\Delta_{(h)}^2}{n}\right)$$

where S_{0h}^2 is mean square of response group, S_{0h2}^2 is mean square of non-response group and $(h = 1,2,3,4,5,6)$

6.5 Numerical illustration

In order to investigate the efficiency of the estimators of rank set sampling in situations of non response, while considering the problems of new allocation, apple data is considered. The data refers to yield of 260 orchards in metric tons from block Lar of district Ganderbal. For the purpose of illustration, we have randomly divided the 260 orchards into three strata consisting of 73, 70, 97 orchards and a sample of size 50 orchards is taken. Sample sizes among different allocations along with stratum size and stratum variances are given Table-24. A

set size "*m*" =2 is choosen with different cycles under ranked set procedure given in Table-25.

From Table-26, it is evident that under new allocation schemes, ranked set sampling causes reduction in variances in case non-response situations as compared to stratified random sampling. Same is the case in Table-27, there is reduction in variances with the increase in set size among under different non-response rates.

It is concluded from Tables-24-27 that under different combinations of non-responses rates and inverse ratio of sub-sample under non-response, new allocation schemes depending upon the knowledge of non-response rate, stratum sizes and stratum variances among non-response produce more précised estimates under ranked set sampling as compared to stratified random sampling, because of the reason that ranked set stratified sampling combines the variance reduction that arises from stratifying the population with the increased precision ranked set sampling holds over simple random sampling, also the ranked set samples are more regularly spaced.

Table-24 : Summary statistics among different allocations along with stratum size and stratum variance

Stratum (i)	Size (N_i)	Stratum Mean ($\bar{Y}_{0i}$)	Stratum variance (S_{0i}^2)	Sample Sizes in different allocations						
				Proportional allocation	1	2	3	4	5	6
					n_h	n_h	n_h	n_h	n_h	n_h
1	73	22.11	6.89	15	16	28	20	30	20	24
2	90	12.09	7.63	15	14	10	10	10	12	8
3	97	7.62	4.14	20	20	12	20	10	18	18

Table-25 : Set sizes along with different cycles under RSS

	Set size (m= 2)					
	1	**2**	**3**	**4**	**5**	**6**
	8	14	10	15	10	12
No of cycles in each strata	7	5	5	5	7	4
	10	6	15	5	8	9

Table-26 : Comparison of variances of new allocations under non-response situations in case of ranked set stratified sampling and stratified simple random sampling

Allocations								
				$V[\bar{Y}_{strs}]$			**Allocations**	$V(\bar{\bar{y}}_{(rss)})$ **where** $(h = 1,2,3,4,5,6)$
n_{2h}	m_i	W_{h2}	K_h	***Proportional allocation***	**11.64**	**Ranked set stratified sampling**	**Proportional allocation under RSS**	**8.02**
15	8	22	1.8	**Stratified sampling**	**8.61**		**1**	**7.58**
23	10	33	2.3				**2**	**7.63**
30	12	44	2.5				**3**	**7.44**
							4	**7.13**
							4	**7.54**
							6	**7.34**

Table-27 : Variances under different non-response rates

Non-response rate (%)	*M*	$V(\bar{\bar{y}}_{(rss)})$
44	5	6.668
	6	6.205
	7	5.184
	8	4.286
33	5	5.689
	6	5.276
	7	4.396
	8	3.753
22	5	2.727
	6	2.544
	7	2.27
	8	2.014

Chapter – 7

SUMMARY AND CONCLUSION

Cost-effective sampling methods are of major concern in statistics, especially when the measurement of the characteristic of interest is costly and time consuming. Environmental monitoring and assessment, Forest surveys etc.; require observational data as opposed to data obtained from controlled experiments. Obtaining such data requires identification of sample units to represent the population of interest, followed by selection of particular units to quantify the desired characteristics. The most expensive part of this process is laboratory analysis, while identification of potential sample units is comparatively simple. We can therefore achieve great observational economy if we are able to identify a large number of sample units to represent the population of interest, yet only have to quantify a carefully selected subsample. This potential for observational economy can be obtained by *ranked set sampling* (RSS). McIntyre (1952) developed the procedure of RSS to find a more efficient method to estimate the yield of pastures.The present study has been carried out on Pinus and apple data. Numerical and graphical analysis of Pinus data is carried out in R – software, while the univariate study of Apple data is carried out in SAS. Also regression analysis and Jacknifing of Pinus data is carried out in SAS using the function `PROC REG` and `JACK REG`. With the help of R-software new functions like `drss(m,r)`,`varwts(n,h)`,`makeAlloc(n,m)`,`ratio.est(n,N(x,y))`were developed, which are simple to execute.

The present work consists of seven chapters. Chapter first deals with complete introduction of rank set sampling and its historical background. Chapter second contain an introduction to R and SAS softwares. The regression analysis, correlation analysis and confidence interval are given in numerical summary. In graphic summary box plots, qqnorm plots are provided and also four new functions

drss(m,r),varwts(n,h),makeAlloc(n,m),ratio.est(n,N(x,y)) have been developed and complete codes along with illustrations have been given. Due effort is devoted to study important aspects of rank set sampling. In chapter three simple linear regression model are considered with respect to samples taken from the identified sampling techniques like simple random sampling (SRS), systematic sampling (SYS) including rank set sampling (RSS), also estimation of parameters of regression model and regression estimates of mean of response variable in context to rank set sampling is discussed. It is found that the coefficient of determination obtained from regression model based on rank set sample was higher than rest of two sampling schemes, also the parameters of comparison like root mean square error, p value and coefficient of variation (CV), were much lower in rank set based regression model than others. From density curves again the curves are more symmetric in case of rank set sample as compared to SRS and SYS. Also from validation technique (Jacknifing) using the function JACKREG in SAS, there was consistency in the measure of R^2, Adj R^2 and RMSE in case of RSS as compared to SRS and SYS. The above results occurred because rank set samples are more regularly spaced than those obtained from SRS and SYS and therefore more representative of the population. Because of ranking the RSS procedure induces stratification at sample level which involves the gained precision in this scheme. Chapter four deals with the study of rank set sampling under stratification. In this chapter ranked set sampling is introduced within the framework of stratified sampling. Rather than selecting a simple random sample within each stratum as is done in stratified simple random sampling (SSRS), a ranked set sample within each stratum is taken and the resulting technique is known as Ranked set stratified sampling (RSSS). This sampling design combines the variance reduction that arises from stratifying the population with the increased precision RSS holds over SRS. The variable used for ranking in the RSS procedure is also used to stratify the population. From the simulation results it is concluded that RSS, when used in place of SRS in the final stage of stratified sampling, can provide considerably more accurate

estimates of population means. The increased precision of an RSS estimator makes the procedure worthwhile when ranking is in-expensive relative to the cost of quantifying the variable of interest. Incorporating RSS into a stratified design improves the estimates of all variables quantified in the survey as proved from simulations. Chapter five is devoted to ratio estimation. RSS methodology is examined under ratio method. New ratio estimators for RSS are introduced based on various combinations of known values of Median, Quartile deviation, coefficient of Skewness, Kurtosis, and Correlation coefficient of auxiliary variable. Proposed ratio estimators under RSS proved to be more efficient as compared to ratio estimators under SRS. From the results it is concluded that the efficiency of RSS estimators decreases as the correlation coefficient decreases, also the efficiency increases as the set size m is increasing. Utilizing the knowledge of the deciles of the auxiliary variable X, a gain in efficiency is obtained in case of ratio estimators based on RSS with respect to SRS for estimating the population mean of the variable of interest Y. The bias of the new/proposed estimators based on RSS is found to be very less small. Modified ratio estimators under RSS achieve a worthwhile reduction in MSE over SRS. In chapter six the problem of estimating the population mean under non responses is studied under rank set sampling. The concept of inverse ratio of subsample under non response and its combination with non-response rate is introduced in this chapter. The introduction of this combinations has provided a new way to tackle the non-response in sample surveys under RSS. New allocation schemes are proposed in order to study their effect on sampling variance. Basic method of Non-Response estimation given by Hansen and Hurwitz (1946) is used in this chapter. Results suggest that under different combinations of non-responses rates and inverse ratio of sub-sample under non-response, new allocation schemes depending upon the knowledge of non-response rate, stratum sizes and stratum variances among non-response produce more precise estimates under ranked set sampling as compared to stratified random sampling, because of the reason that ranked set stratified sampling combines the variance reduction that arises from

stratifying the population with the increased precision ranked set sampling holds over simple random sampling, which is already proven in Chapter-4.

RSS procedures are more attractive than its counter parts as they increase the efficiency of population mean. It has been established that RSS has its practical implications. It can be applied to analyze data generated in a scientific investigation. R and SAS software packages facilitates a lot in implementation of RSS methodology which is very informative and applicable to sample surveys of horticultural and forestry crops which lack a proper sampling frame. These benefits of RSS need not be restricted to stratified sampling, replacing SRS with RSS in the final stage of any survey design will greatly improve the accuracy of the estimators.

LITERATURE CITED

Abu-Dayyeh, W. and Muttlak, H.A. 1996. Using ranked set sampling for hypothesis tests on the scale parameter of the exponential and uniform distributions. *Pakistan Journal of Statistics* **12** : 131-138.

Bai, Z.D. and Chen, Z. 2003. On the theory of ranked set sampling and its ramifications. *Journal of Statistical Planning and Inference* **109** : 81-99.

Becker, R.A. Chambers, J.M. and Wilks, A.R. 1988. The New S Language. Chapman and Hall, New York.

Bhoj, D.S. 1997. Estimation of parameters of the extreme value distribution using ranked set sampling. *Communications in Statistics-Theory and Methods* **26** : 653-660.

Bohn, L.L. and Wolfe, D.A. 1992. Non-parametric two sample procedures for ranked set sampling data. *J. Amer. Statist. Assoc*. **87** : 552-561.

Bouza, C. 2009. New ratio estimators of the mean using ranked set sampling methods. *Revista Investigation Operacional* **30** : 97-108

Chen, Z. and Shen, L. 2003. Two-layer ranked set sampling with concomitant variables. *Journal of Statistical Planning and Inference* **115** : 45–57.

Chen, Z. 1999. Density estimation using ranked-set sampling data. *Environmental and Ecological Statistics* **6** : 135-146.

Chen, Z. 2000. On ranked-set sample quantiles and their applications. *Journal of Statistical Planning and Inference* **83** : 125-135.

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

Chen, Z. 2001. The optimal ranked-set sampling scheme for inference on population quantiles. *Statist. Sinica*. **11** : 23-37.

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

Chen, Z. and Wang, Y. 2004. Efficient Regression Analysis with Ranked-Set Sampling. *Biometrics* **60** : 97-104.

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

Cochran, W.G. 1977. *Sampling Techniques.* John Wiley and Sons, New York.

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

Dell, T.R. and Clutter, V. 1952. Ranked set sampling theory with order statistics background. *Biometrics* **28** : 545-555.

Gaajendra, K.A. and Bouza, C. 2012. Double sampling with rank set selection in the second phase with non-response: Analytical results and Monte carlo experiments. *Journal of Probability and Statistics* **23** : 45-53.

Gilbert, R.O. 1995. Ranked set sampling. *DQO Statistics Bulletin: Statistical Methods for Data Quality Objective Process, PNL-SA-26377.* The Pacific Northwest Laboratory.

Gilbert, R.O. and Eberhardt, L.L. 1976. An evaluation of double sampling for estimating plutonium inventory in surface soil. **In** : *Radioecology and Energy*

Sources. [Ed. C.E. Cushing]. Stroudsburg, Pennsylvania: Dowden, Hutchison and Ross, pp. 157-163.

Hall, L.K. and Dell, T.R. 1966. Trial of ranked set sampling for forage yields. *Forest Science* **12** : 22-6.

Hani, M.S. 2002. On Regression Analysis with Random Regressors Using Ranked Samples. *Information and Management Sciences* **13** : 19-36.

Hansen, M.H. and Hurwitz, W.N. 1946. The problem of non-response in sampling surveys. *Jour. Amer. Stat. Assoc.* **41** : 517-529.

Hettmansperger, T.P. 1995. The ranked set sample sign test. *Journal of Nonparametric Statistics* **4** : 263-270.

Husby, C.N., Elizabeth, A., Stasny, N. and Douglas, A.W. 2005. An application of ranked set sampling for mean and median estimation using USDA crop production data. *Journal of Agricultural, Biological and Environmental Statistics* **10** : 354-373.

Iqbal, M.J., Maqbool, S. and Mir, S.A. 2013. Modified ratio estimators of population mean using linear combination of co-efficient of skewness and quartile deviation. *International Journal of Modern Mathematical Sciences* **6**(3) : 174-183.

Kadilar, C. and Cingi, H. 2004. Ratio estimators in simple random sampling, Applied Mathematics and Computation **151** : 893-902.

Kadilar, C. and Cingi, H. 2006. An Improvement in Esti-mating the Population mean by using the Correlation Co-efficient. *Hacettepe Journal of Mathematics and Statistics* **35**(1) : 103-109.

Kamarulzaman, I. 2011. On comparison of some variation of rank set sampling. *Sains Malaysiana* **40** : 397-401.

Kaur, A., Patil, G.P. and Taillie, C. 1997. Unequal allocation models for ranked set sampling with skew distributions. *Biometrics* **53** : 123-130.

Khan, A.A. and Mir, A.H. 2005. Applications of R-software in agricultural data analysis. *SKUAST Journal of Research* **7**(1) : 36-64.

Koyuncu, N. and Kadilar, C. 2009. Efficient Estimators for the Population mean. *Hacettepe Journal of Mathematics and Statistics* **38**(2) : 217-225.

Kvam, P.H. and Samaniego, F.J. 1993. On the inadmissibility of empirical averages as estimators in ranked set sampling. *Journal of Statistical Planning and Inference* **36** : 39-55.

Martin, W.L., Sharik, T.L., Oderwald, R.G. and Smith, D. 1980. Evaluation of ranked set sampling for estimating shrub Phytomass in Appalachian Oak Forests. *Publication No. FWS-4-80,* School of Forestry and Wildlife Resources, Virginia Polytechnic Institute and State University.

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

McIntyre, G.A. 1978. Statistical aspects of vegetation sampling: Measurement of Grassland Vegetation and Animal Production. *Commonwealth Bureau of Pastures and Field Crops*. Hurley, Berkshire, UK. **45** : 8-21.

Minzhu, K.M. 2005. Quantile Estimation from Ranked Set Sampling Data. *Sankhya: The Indian Journal of Statistics* **67** : 295-304.

Murray, R.A. 2000. The use of ranked set sampling in spray deposit assessment. *Aspects of Applied Biology* **43** : 16-20.

Murthy, M.N. 1967. Sampling theory and methods, Statistical Publishing Society, Calcutta, India.

Muttlak, H.A. 1995. Parameters Estimation in a simple linear regression using rank set sampling. *Biometrical. J.* **37** : 799-810.

Nussbaum, B.D. and Sinha, B.K. 1997. Cost effective gasoline sampling using ranked set sampling. In a Proceedings of the Section on Statistics and the Environment. *American Statistical Association* **41** : 34-39

Omer-Ozturk and Wolfe, D.A. 1998. Optimal ranked set sampling protocol for the signed rank test. Technical Report TR 630, Ohio State University Department of Statistics.

Patil, G.P. and Sinha, A.K. 1993. Observational economy of ranked set sampling: comparison with the regression estimator. *Environmetrics* **4** : 399-412.

Patil, G.P., Sinha, A.K. and Taillie, C. 1993. Relative precision of ranked set sampling: Comparison with the regression estimator. *Environmetrics* **4** : 399-412.

Prasad, B. 1989. Some improved ratio type estimators of population mean and ratio in finite population sample surveys. *Communications in Statistics: Theory and Methods* **18** : 379-392.

Presnell, B. and Bohn, L.L. 1999. U-statistics and imperfect ranking in ranked set sampling. *Journal of Nonparametric Statistics* **10** : 111-126.

Risch, N. and Zhang, H. 1995. Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268** : 1584-1589.

Samawi, H.M. 1997. On regression analysis using ranked set sample. *Journal of Statistical Research (JSR)* **35** : 93-105.

Samawi, H.M. 1999. On quantiles estimation with application to normal ranges and Hodges-Lehmann estimate using a variety of ranked set sample. Department of Statistics, Yarmouk University, Irbid, Jordan.

Samawi, H.M. and Muttlak, H.A. 1996. Estimating the population mean using extreme ranked set sampling. *Biometrical. J.* **38** : 577-586.

Samawi, H.M. and Muttlak, H.A. 1996. Estimation of ratio using rank set sampling. *Biometrical Journal* **38** : 753-764.

Sen, A.R. 1993. Some early developments in ratio estimation. *Biometric Journal* **35**(1) : 3-13.

Silva, P.L.D.N. and Skinner, C.J. 1997. Variable selection for regression estimation in finite populations. *Survey Methodology* **23** : 23-32.

Singh, D. and Chaudhary, F.S. 1986. Theory and Analysis of Sample Survey Designs, New Age International Publisher.

Singh, G.N. 2003. On the improvement of product method of estimation in sample surveys. *Journal of the Indian Society of Agricultural Statistics* **56**(3) : 267-265.

Singh, H.P. and Tailor, R. 2003. Use of known correlation co-efficient in estimating the finite population means. *Statistics in Transition* **6**(4) : 555-560.

Singh, H.P. and Tailor, R. 2005. Estimation of finite population mean with known co-efficient of variation of an auxiliary. *STATISTICA* **65**(3) : 301-313.

Singh, H.P., Tailor, R., Tailor, R. and Kakran, M.S. 2004. An Improved Estimator of population mean using Power transformation. *Journal of the Indian Society of Agricultural Statistics* **58**(2) : 223-230.

Stokes, S.L. 1977. Ranked set sampling with concomitant variables. *Communications in Statistics-Theory and Methods* **83** : 374-381.

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

Stokes, S.L. and Sager, T.W. 1988. Characterization of a ranked set sample with application to estimating distribution functions. *Journal of the American Statistical Association* **83** : 374-381.

Takahasi, K. and 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** : 1-31.

Upadhyaya, L.N. and Singh, H.P. 1999. Use of transformed auxiliary variable in estimating the finite population mean. *Biometrical Journal* **41**(5) : 627-636.

Venables, W.N. and Ripley, B.D. 2009. *Modern Applied Statistics with S-PLUS*, 4th edition, Springer Verlag, New York.

Walid, A.D., Dorvo, S.S. and Obaid, A.V. 2011. On Regression Analysis Using Extreme Ranked Set Sampling. *The First International Conference on Interdisciplinary Research and Development.* 31 May-1 June, Thailand.

Yan, Z. and Tian, B. 2010. Ratio Method to the Mean Estimation Using Co efficient of Skewness of Auxiliary Variable, ICICA 2010, Part II, CCIS **106** : 103-110.

Yaprak, A.W. 2007. Parameter estimation in multiple linear regression models using ranked set sampling. *Commun. Fac. Sci. Univ. Ank. Series*. **56** : 7-20.

You, G. 2009. Efficient designs for sampling and sub sampling in fisheries research based on ranked sets. *Journal of Marine Science* **66** : 928-934.

Yu, P.L.H. and Lam, K. 1997. Regression estimator in ranked set sampling. *Biometrics* **53** : 1070-1080.

Zehua, C. 2000. The efficiency of ranked-set sampling relative to simple random sampling under multi-parameter families. *Statistica Sinica* **10** : 247-263.

Zehua, C.N. 2008. General ranked set sampling for efficient treatment comparisons. *Statistica Sinica.* **18** : 91-104.

Zhao, X. and Chen, Z. 2000. The Ranked-Set Sampling M-Estimates for Symmetric Location Families. *Annals of the Institute of Statistical Mathematics* **54** : 626-640.

Sher-e-Kashmir
University of Agricultural Sciences & Technology of Kashmir
Division of Agricultural Statistics,
Shalimar Campus, Srinagar – 190 025
-::o::-

CERTIFICATE

Certified that all the corrections/amendments as suggested by External Examiner Dr. K.K. Tyagi, Principal Scientist, Sample Survey Division, IASRI, New Delhi during Viva-Voce examination held on 16th of April, 2014 have been incorporated in the manuscript entitled **"On Some Aspects of Rank Set Sampling and Non-Response Situations Utilizing R/SAS Softwares"** submitted by **Mr. M Iqbal Jeelani Bhat (Regd. No. 2011-380-D).**

(Dr. S.A. Mir)
Chairman
Advisory Committee