Department of Statistics
PRESIDENCY UNIVERSITY

RANKED SET SAMPLING
TOWARDS
THE ENHANCEMENT
OF
STATISTICAL PRECISION

ANNESHA MUSTAFI

Registration No.	:	18414210024
Examination Roll No.	:	18414024

KAZI SAHAJIT ISLAM

Registration No.	:	18414110020
Examination Roll No.	:	18414020

M.Sc. Even Semester, 2020

June 20, 2020

ACKNOWLEDGEMENTS

This dissertation is a milestone in our academic career. We have been fortunate to learn theories and concepts which would have been impossible if we had not extensively carried out the needed research. We are grateful to a number of people who have guided and supported us throughout the research process and provided assistance for our venture.

We would first like to thank our supervisor, Prof. Saurav De, Department of Statistics, Presidency University, whose expertise was invaluable in the formulating of the research topic and methodology in particular. His recommendations and instructions have enabled us to assemble and finish the dissertation effectively.

We would like to thank all our professors of the Department of Statistics, Presidency University, who have supported and encouraged us to believe in our abilities. We would also like to thank the non-teaching staff of our department.

Finally, our family and friends have supported and helped us along the course of this dissertation by giving encouragement and providing the moral and emotional support we needed to complete our thesis. To them, we are eternally grateful.

ABSTRACT

Ranked set sample (RSS) was introduced by McIntyre (1952)[1] as a method of selecting data if the sampling units can be ,easily ranked without any cost, but it is very difficult or expensive to measure them, They proposed to use the RSS mean as an estimator for the population mean instead of the usual estimator which is the mean of the simple random sample (SRS). In this paper we consider estimating parameters of Negative Binomial using RSS to show that RSS will give more precision in estimating parameter. Then we will try to implement this result to overcome an Overdispersive case. Then we will conduct some testing problem using the RSS sampling and it will appere that use of RSS gives much better results in terms of the empirical power function compared to the SRS.

Contents

List of Figures

List of Tables

List of Abbreviations

MLE	- Maximum Likelihood Estimator
MME	- Method of Moments Estimator
MSE	- Mean Squared Error
NB(r, θ)	- Negative Binomial Distribution with parameters r and θ, where $r > 0$ and $0 < \theta < 1$
RSS	- Ranked Set Sample
SRS	- Simple Random Sample

1 INTRODUCTION

1.1 Background

1.1.1 Simple Random Sampling

This is the simplest form of random sampling. From the list of N identifiable units, units are drawn one by one with replacement or without replacement. No auxiliary information is used in the drawing of samples.

A sample is said to be selected by *simple random sampling with replacement* by n draws from a population of size N if the sample is drawn by observing the following rules:

- At each draw, each unit in the population has the same chance of being selected.
- A unit selected at a draw is returned to the population before the next draw.

A sample is said to be selected by *simple random sampling without replacement* by n draws from a population of size N if the sample is drawn by observing the following rules:

- At each draw, each available unit in the population has the same chance of being selected.
- A unit selected at a draw is removed from the population before the next draw.

1.1.2 Ranked Set Sampling

The goal of ranked set sampling is to collect observations from a population that are more likely to span the full range of values in the population. The use of ranked set sampling increases the chance that the collected samples will yield representative measurements. This results in better estimates of the mean as well as improved performance of many statistical procedures.

To obtain an RSS of k observations from a population, we proceed in the following manner:

- First, an initial *SRS* of k units is selected from the population and rank ordered on the *attribute of interest*. A variety of mechanisms can be used to obtain this ranking, including visual comparisons, expert opinion, or through the use of auxiliary variables, but it cannot involve actual measurements of the attribute of interest on the selected units.
- The unit that is judged to be the **smallest** in this ranking is included as the first item in the *RSS* and the attribute of interest is formally measured for the unit. This initial measurement is called the *first judgment* order statistic and is denoted by $X_{[1]}$.
- A second *SRS* (independent of the first *SRS*) of size k is selected from the population and ranked in the same manner as the first *SRS*. From this second *SRS*, we select the item ranked as the **second smallest** of the k units i.e., the second judgment order statistic and add its attribute measurement, $X_{[2]}$, to the *RSS;* and so on...
- This entire process results in the k measured observations $X_{[1]}, X_{[2]}, ..., X_{[k]}$ and is called a *cycle*. The number of units, k, in each SRS is called the set size.

- Thus to complete a single ranked set cycle, we need to use a total of k^2 units from the population to separately rank k independent simple random samples of size k each. The measured observations $X_{[1]}, X_{[2]}, ..., X_{[k]}$ constitute a balanced ranked set sample of size k, where the descriptor "balanced" refers to the fact that we have collected one judgment order statistic for each of the ranks $1, 2, \ldots, k$.

To obtain a balanced RSS with a desired total number of measured observations i.e., sample size $n = km$, we repeat the entire process for m independent cycles, yielding the following balanced RSS of size n (see Table 1.1).

Table 1.1: m Cycles of Ranked Set Sampling

Cycle 1	$X_{[1]1}$	$X_{[2]2}$	$\cdots$	$X_{[k]1}$
Cycle 2	$X_{[1]2}$	$X_{[2]2}$	$\cdots$	$X_{[k]2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Cycle m	$X_{[1]m}$	$X_{[2]m}$	$\cdots$	$X_{[k]m}$

For simplicity, we consider only the case of a single cycle ($m = 1$), so that the total sample size n is equal to the set size k. Even if we consider m cycles, the ultimate results will be similar as in the case of a single cycle.

1.1.3 Overdispersion

Count observations often exhibit variability exceeding that predicted by the given statistical model. This phenomenon is called overdispersion. Suppose we assume that each person has the same probability of dying in a fatal accident in the next week. More realistically, these probabilities vary, due to factors such as amount of time spent driving, whether the person wears a seat belt, and geographical location etc. Such variation causes fatality counts to display more variation than predicted by the Poisson model.

Suppose that Y is a random variable with variance $var(Y|\mu)$ for given μ, but itself varies because of unmeasured factors such as those just described. Let $\theta = E(\mu)$. Then unconditionally,

$E(Y) = E[E(Y|\mu)]$ and $Var(Y) = E[var(Y|\mu)] + var[E(Y|\mu)]$.

For $Y \sim Poisson(given\ \mu)$, then $E(Y) = E(\mu) = \theta$ and $var(Y) = E(\mu) + var(\mu) > \theta$.

1.2 Aims and Objectives

Overdispersive case is not rare; rather it is the most common scenario in practice. Presence of overdispersion may cause standard errors of the estimates to be overestimated; i.e. a variable may appear to be a significant predictor when it is in fact not significant which will affect overall goodness-of-fit. We will try to find out how ranked set sampling works under the overdispersive situation. The efficiency of parameter estimation under ranked set sampling compared to simple random sampling will also be a concern to us. Not only that, rank set sampling can also be a very useful tool for testing purpose. So, we will compare the power of test under simple random sampling and ranked set sampling.

Now some questions may arise :

- How much precision do we get in parameter estimation under ranked set sampling ?
- Does this method really work in every overdispersive fitted model ?
- How to construct a test statistic in case of ranked set sampling ?
- Do we really need to construct a new test statistic under ranked set sampling ?

We will try to answer these questions in this project.

1.3 Structure of Report

The paper has been designed to follow the following structure of chapters :

1. Introduction
   - This is the introduction to the dissertation assisted by background and objectives of the research.
2. Literature Review
   - This chapter is a discussion of previous research studies related to comparison of simple random sampling and rank set sampling.
3. Methodology
   - This chapter is an elaboration of the research methodology adopted for the research study and the aspects of the research that the research methodology will cover.
4. Simulation Study and Analysis
   - This chapter is a collection of simulation studies based on the research methodology along with its analysis.
5. Conclusion
   - This chapter is an outline of the outcome of the critical analysis tool in the current research.

2 LITERATURE REVIEW

RSS was suggested and applied to the problem of estimating mean pasture yields by McIntyre (1952)[1]. Takahasi and Wakimoto (1968)[2] independently suggested the same method.

Consider two mutually independent sets of n observations each from a continuous population with distribution function F, density function f, finite mean μ, and finite variance σ^2.One set of n observations, $X_1, ..., X_n$, is collected as a simple random sample (SRS) and the second set of n observations is collected as a balanced ranked set sample (RSS), corresponding to set size k and m cycles, with $n = km$. The ranked set sample observations from *cycle* 1 are denoted by $\left(X_{[1]1}, X_{[2]1}, ..., X_{[k]1}\right)$; *cycle* 2 are denoted by $\left(X_{[1]2}, X_{[2]2}, ..., X_{[k]2}\right)$; . . ., and the final *cycle* m are denoted by $\left(X_{[1]m}, X_{[2]m}, ..., X_{[k]m}\right)$.

The SRS estimator for the population mean μ is just the sample mean $\hat{\mu}_{SRS} = \bar{X}$ and it is well known that $E[\hat{\mu}_{SRS}] = \mu$ and $Var(\hat{\mu}_{SRS}) = \frac{\sigma^2}{n}$.

The natural ranked set sample estimator, $\hat{\mu}_{RSS}$, for the population mean μ is

$$\hat{\mu}_{RSS} = \bar{X}_{RSS} = \sum_{j=1}^{m}\sum_{i=1}^{k} \frac{X_{[i]j}}{km} . \tag{2.1}$$

The balanced RSS estimator $\hat{\mu}_{RSS}$(2.1) is also an unbiased estimator for the population mean μ regardless of whether the judgment rankings are perfect or imperfect. Dell and Clutter (1972)[3] established this result in the general setting for set size k and m cycles without any restriction on the accuracy of the judgment rankings. We demonstrate the argument under the more restrictive assumption that the judgment rankings are perfect. Under this additional assumption of perfect rankings, the ranked set sample observations are, in fact, true order statistics from the underlying continuous population.

For simplicity in the argument, we consider only the case of a single cycle ($m = 1$), so that the total sample size n is equal to the set size k.

Under the assumption of perfect rankings, we can represent the RSS observations for this setting by $X^*_{(1)}, X^*_{(2)}, \ldots, X^*_{(k)}$, where these k variables are mutually independent and $X^*_{(i)}, i = 1, 2, \ldots k$, is distributed like the i^{th} order statistic for a random sample of size k from a continuous distribution with distribution function F and density f.

It follows immediately from properties of a simple average that

$$E\left[\hat{\mu}_{RSS}\right] = E\left[\bar{X}_{RSS}\right] = \frac{1}{k}\sum_{i=1}^{k} E\left[X^*_{(i)}\right] . \tag{2.2}$$

Moreover, since $X^*_{(i)}$ is distributed like the i^{th} order statistic for a random sample of size k from a continuous distribution with distribution function F and density f under perfect rankings, we have

$$E\left[X^*_{(i)}\right] = \int_{-\infty}^{\infty} x \frac{k!}{(i-1)!\,(k-i)!} \left[F\left(x\right)\right]^{i-1} \left[1 - F(x)\right]^{k-i} f(x)dx\,, \tag{2.3}$$

for $i = 1, 2, \ldots k$. Combining (2.2) and (2.3), we obtain

$$\begin{aligned} E\left[\bar{X}_{RSS}\right] &= \frac{1}{k}\sum_{i=1}^{k}\left\{\int_{-\infty}^{\infty} kx \left(\begin{array}{c} k-1 \\ i-1 \end{array}\right)\left[F\left(x\right)\right]^{i-1}\left[1-F(x)\right]^{k-i} f(x)dx\right\} \\ &= \int_{-\infty}^{\infty} xf(x)\left\{\sum_{i=1}^{k}\left(\begin{array}{c} k-1 \\ i-1 \end{array}\right)\left[F\left(x\right)\right]^{i-1}\left[1-F(x)\right]^{k-i}\right\} dx \,. \end{aligned} \tag{2.4}$$

Letting $q = i - 1$ in the summation in (2.4), we see that

$$\sum_{i=1}^{k}\left(\begin{array}{c} k-1 \\ i-1 \end{array}\right)\left[F\left(x\right)\right]^{i-1}\left[1-F(x)\right]^{k-i} = \sum_{q=0}^{k-1}\left(\begin{array}{c} k-1 \\ q \end{array}\right)\left[F\left(x\right)\right]^{q}\left[1-F(x)\right]^{(k-1)-q} = 1\,, \tag{2.5}$$

since the latter expression is just the sum over the entire sample space of the probabilities for a binomial random variable with parameters $k - 1$ and $p = F(x)$.

Using this fact in (2.3), we obtain

$$E\left[\hat{\mu}_{RSS}\right] = E\left[\bar{X}_{RSS}\right] = \int_{-\infty}^{\infty} xf(x)dx = \mu\,, \tag{2.6}$$

thus establishing the fact that $\hat{\mu}_{RSS}$ is an unbiased estimator for μ.

Takahasi and Wakimoto (1968) proved that the mean of RSS is an unbiased estimator of the population mean with a smaller variance than the variance of the sample mean of a SRS of the same sample size.

To obtain the variance of the RSS estimator $\hat{\mu}_{RSS}$, we note that the mutual independence of the $X^*_{(i)}$'s, $i = 1, 2, \ldots k$, enables us to write

$$Var\left(\bar{X}_{RSS}\right) = \frac{1}{k^2}\sum_{i=1}^{k} Var\left(X^*_{(i)}\right)\,. \tag{2.7}$$

Letting $\mu^*_{(i)} = E\left[X^*_{(i)}\right]$, for $i = 1, 2, \ldots k$, we note that

$$\begin{aligned} E\left[\left(X^*_{(i)}-\mu\right)^2\right] = E\left[\left(X^*_{(i)}-\mu^*_{(i)}+\mu^*_{(i)}-\mu\right)^2\right] &= E\left[\left(X^*_{(i)}-\mu^*_{(i)}\right)^2\right]+\left(\mu^*_{(i)}-\mu\right)^2 \\ &= Var\left(X^*_{(i)}\right)+\left(\mu^*_{(i)}-\mu\right)^2\,, \end{aligned} \tag{2.8}$$

since the cross product terms are zero. Combining (2.7) and (2.8) yields the expression

$$Var\left(\bar{X}_{RSS}\right) = \frac{1}{k^2}\sum_{i=1}^{k} E\left[\left(X^*_{(i)}-\mu\right)^2\right] - \frac{1}{k^2}\sum_{i=1}^{k}\left(\mu^*_{(i)}-\mu\right)^2\,. \tag{2.9}$$

Now, proceeding as we did with $E\left[\bar{X}_{RSS}\right]$, we see that

$$\begin{aligned}\sum_{i=1}^{k} E\left[\left(X^{*}_{(i)}-\mu\right)^{2}\right] =&\sum_{i=1}^{k}\int_{-\infty}^{\infty} k(x-\mu)^{2}\left(\begin{array}{c}k-1\\ i-1\end{array}\right)\left[F\left(x\right)\right]^{i-1}\left[1-F(x)\right]^{k-i} f(x)dx \\ &= k\int_{-\infty}^{\infty}(x-\mu)^{2}f(x)\left\{\sum_{i=1}^{k}\left(\begin{array}{c}k-1\\ i-1\end{array}\right)\left[F\left(x\right)\right]^{i-1}\left[1-F(x)\right]^{k-i}\right\}dx\ .\end{aligned} \tag{2.10}$$

Once again using the binomial distribution, the interior sum is equal to 1 and we obtain

$$\sum_{i=1}^{k} E\left[\left(X^{*}_{(i)}-\mu\right)^{2}\right] = k\int_{-\infty}^{\infty}(x-\mu)^{2}f(x)dx = k\sigma^{2}\ . \tag{2.11}$$

Combining (2.9) and (2.11), it follows that

$$Var\left(\bar{X}_{RSS}\right) = \frac{1}{k^{2}}\left\{k\sigma^{2}-\sum_{i=1}^{k}\left(\mu^{*}_{(i)}-\mu\right)^{2}\right\} = \frac{\sigma^{2}}{k}-\frac{1}{k^{2}}\sum_{i=1}^{k}\left(\mu^{*}_{(i)}-\mu\right)^{2}\ . \tag{2.12}$$

Thus, both $\hat{\mu}_{SRS}$ and $\hat{\mu}_{RSS}$ are unbiased estimators for the population mean. Moreover, from (2.12), it follows that

$$Var\left(\bar{X}_{RSS}\right) = \frac{\sigma^{2}}{k}-\frac{1}{k^{2}}\sum_{i=1}^{k}\left(\mu^{*}_{(i)}-\mu\right)^{2} = Var\left(\bar{X}\right)-\frac{1}{k^{2}}\sum_{i=1}^{k}\left(\mu^{*}_{(i)}-\mu\right)^{2} \leq Var\left(\bar{X}\right)\ , \tag{2.13}$$

since $\sum_{i=1}^{k}\left(\mu^{*}_{(i)}-\mu\right)^{2}\geq 0$. Hence, in the case of perfect rankings not only is $\bar{X}_{RSS}$ an unbiased estimator, but also its variance is always no larger than the variance of the SRS estimator $\bar{X}$ based on the same number of measured observations. In fact, this is a strict inequality unless $\mu^{*}_{(i)}=\mu$ for all $i=1,2,\ldots k$, which is the case only if the judgment rankings are purely random.

In many applications it is very difficult or expensive to measure the sampling units, but the units can be ranked without any cost. In agricultural and environmental studies, it is possible to rank the sampling units without actually measuring them.

3 METHODOLOGY

3.1 Error in Estimating Parameters

Suppose $X_1, X_2, ..., X_n$ be a random sample of size n from a Negative Binomial (r, θ), $r > 0$ and $0 < \theta < 1$

$$f(x) = \left(\begin{array}{c} r+x-1 \\ x \end{array} \right) \theta^r (1-\theta)^x, x = 0, 1, 2, ...$$

We know, the MSE of an estimator $\hat{\theta}$ with respect to an unknown parameter θ is defined as

$$MSE\left(\hat{\theta}\right) = E_\theta \left[\left(\hat{\theta} - \theta\right)^2 \right]$$

We have to find the Mean Squared Estimator (MSE) for both of the parameters r and θ using Simple Random Sampling and Ranked Set Sampling.

Consider m random samples each of size n and suppose $\hat{r}_1, \hat{r}_2, ..., \hat{r}_m$ and $\hat{\theta}_1, \hat{\theta}_2, ..., \hat{\theta}_m$ be estimated parameters of r and θ respectively.

Then using Monte Carlo method, we get

$$MSE(\hat{r}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{r}_i - r)^2 \tag{3.1}$$

$$MSE(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2 \tag{3.2}$$

Our target is to show that those values will be small under Ranked Set Sample than Simple Random Sample.

3.1.1 Method of Moments Estimator for NB(r,θ)

We know that if $X \sim NB(r, \theta)$,
$E(X) = \frac{r(1-\theta)}{\theta}$ and $Var(X) = \frac{r(1-\theta)}{\theta^2}$.

Now for $X_1, X_2, ..., X_n \overset{iid}{\sim} NB(r, \theta)$, using Method of Moments estimation,
$E(X) = \bar{X}$ and $var(X) = s^2$
where $\bar{X} = Sample\,Mean$ and $s^2 = Sample\,Variance$

$$\Rightarrow \bar{X} = \frac{r(1-\theta)}{\theta} \text{ and } s^2 = \frac{r(1-\theta)}{\theta^2}$$

$$\Rightarrow \hat{\theta} = \frac{\bar{X}}{s^2} \text{ and } \hat{r} = \frac{\bar{X}^2}{s^2 - \bar{X}} \tag{3.3}$$

Note that, as we are considering samples, it may happen that $\bar{X} \geq s^2$ and that will cause $\hat{\theta} \geq 1$ and $\hat{r} < 0$.

3.1.2 Maximum Likelihood Estimator for NB(r,θ)

Suppose $X_1, X_2, ..., X_n$ be a random sample of size n from a Negative Binomial (r, θ), $r > 0$ and $0 < \theta < 1$.

The likelihood function is

$$L = \prod_{i=1}^{n} \left(\begin{array}{c} r + x_i - 1 \\ x_i \end{array} \right) \theta^r (1-\theta)^{x_i}$$

from which we calculate the log-likelihood function

$$logL = \sum_{i=1}^{n} log \frac{\Gamma(r + x_i)}{\Gamma(x_i + 1)\Gamma(r)} + nr\, log\theta + log(1-\theta) \sum_{i=1}^{n} x_i \, .$$

To find the maximum we take the partial derivatives with respect to r and θ and set them equal to zero :

$$\begin{aligned} \frac{\partial}{\partial\theta} logL = 0 &\Rightarrow \frac{nr}{\theta} + \frac{1}{(1-\theta)}(-1) \sum_{i=1}^{n} x_i = 0 \\ &\Rightarrow \frac{1-\theta}{\theta} = \frac{\bar{x}}{r} \\ &\Rightarrow \frac{1}{\theta} = \frac{\bar{x}+r}{r} \\ &\therefore \hat{\theta} = \frac{r}{\bar{x}+r} \end{aligned} \tag{3.4}$$

and

$$\begin{aligned} &\frac{\partial}{\partial r} logL = 0 \\ \Rightarrow &\sum_{i=1}^{n} \psi(x_i + r) - n\psi(r) + n\, log\hat{\theta} = 0 \\ \Rightarrow &log\left(\frac{r}{\bar{x}+r}\right) = \psi(r) - \frac{1}{n} \sum_{i=1}^{r} \psi(x_i + r) \\ \Rightarrow &\frac{r}{\bar{x}+r} = e^u \, , u = \psi(r) - \frac{1}{n} \sum_{i=1}^{r} \psi(x_i + r) \\ &\text{where } \psi(x) = \frac{d}{dx} log(\Gamma(x)) = \frac{\Gamma'(x)}{\Gamma(x)} \text{ is the digamma function} \\ \therefore &\hat{r} = \frac{\bar{x}e^u}{1-e^u} \\ \Rightarrow &\hat{r} = \frac{\bar{x}}{e^u - 1} \end{aligned} \tag{3.5}$$

The maximum likelihood estimator only exists for samples for which the sample variance is larger than the sample mean.

3.2 Regression

3.2.1 Poisson Regression

Poisson regression is a generalized linear model form of regression analysis used to model count data and contingency tables. Poisson regression assumes the response variable Y has a Poisson distribution, and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. A Poisson regression model is sometimes known as a log-linear model, especially when used to model contingency tables.

The Poisson distribution models the probability of y events (i.e. failure, death, or existence) with the formula

$$P(Y=y|\mu)=\frac{e^{-\mu}\mu^{y}}{y!}\ ,y=0,1,2,...$$

If $x\in\mathbb{R}^{p}$ is a vector of independent variables, then the model takes the form

$$\begin{aligned} log(E(Y|x)) &= x^{T}\beta \\ \iff E(Y|x) &= e^{x^{T}\beta} \end{aligned} \tag{3.6}$$

where x is a covariate term and β is the vector of coefficients of the model. So, we use $\mu=E(Y|x)=e^{x^{T}\beta}$, and thus for given a data set consisting of n vectors $(y_i,x_i^{p\times 1})$,$i=1,2,...,n$, Poisson pmf is given by

$$p(y_1,y_2,...,y_n|x_1,x_2,...,x_n;\beta)=\prod_{i=1}^{n}\frac{e^{y_i x^{T}\beta}e^{-x^{T}\beta}}{y!}$$

The likelihood equations may be formed by taking the derivatives with respect to each regression coefficient and setting the result equal to zero. Doing this leads to a set of nonlinear equations that admits no closed-form solution.

3.2.2 Poisson Regression model for sample from $NB(r,\theta)$

Let $Y_1,Y_2,...,Y_n$ be a random sample from $NB(r,\theta)$ (assuming r is known to us) i.e.

$$p(y|\theta)=\begin{pmatrix} r+y-1 \\ y \end{pmatrix}\theta^{r}\left(1-\theta\right)^{y}\ ,y=0,1,2,...$$

Then we know $E(Y)=\frac{r(1-\theta)}{\theta}$.

Then consider the model as

$$\begin{aligned} E(Y|x) &= e^{x^{T}\beta} \\ \implies \frac{r(1-\theta)}{\theta} &= e^{x^{T}\beta} \\ \implies \frac{1-\theta}{\theta} &= e^{log(r)+x^{T}\beta} \\ \implies \frac{1-\theta}{\theta} &= e^{x^{*T}\beta^{*}} \\ \therefore\theta &= \frac{1}{1+e^{x^{*T}\beta^{*}}} \end{aligned} \tag{3.7}$$

,where $x^{*T}=\begin{pmatrix} log(r) & x^{T} \end{pmatrix}$ and $\beta^{*}=\begin{pmatrix} 1 \\ \beta \end{pmatrix}$.

Hence, for given a data set consisting of n vectors $(y_i,x_i^{p\times 1})$,$i=1,2,...,n$ the Negative Binomial pmf is given by,

$$p(y_1, y_2, ..., y_n|x_1, x_2, ..., x_n; \theta) = \prod_{i=1}^{n} \left(\begin{array}{c} r + y_i - 1 \\ y_i \end{array} \right) \left(\frac{1}{1 + e^{x^{*T}\beta^*}} \right)^r \left(1 - \frac{1}{1 + e^{x^{*T}\beta^*}} \right)^{y_i}$$

The likelihood equations may be formed by taking the derivatives with respect to each regression coefficient and setting the result equal to zero. Doing this leads to a set of nonlinear equations that admits no closed-form solution.

3.3 Hypothesis Testing

3.3.1 Testing for Normal Mean :

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, σ^2 unknown.

Consider the problem for testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$

Then using NP lemma we can easily state that the MP test for (H_0, H_1) is given by

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s} > c \\ 0 & \text{o.w.} \end{cases} \tag{3.8}$$

where $\bar{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} \left(X_i - \bar{X} \right)^2$

Let, $T_{SRS} = \frac{\sqrt{n}\left(\bar{X} - \mu_0\right)}{s}$ be the test statistic.

Note that if the sample is drawn using SRS from the population $N(\mu_0, \sigma^2)$, then $T_{SRS} \overset{H_0}{\sim} t_{n-1}$

Then a level $\alpha = 0.05$ will give us

$$P_{H_0} \left(T_{SRS} > c \right) = \alpha$$
$$\Rightarrow c = t_{\alpha, n-1}$$

Hence

$$\phi(x) = \begin{cases} 1 & \text{if } T_{SRS} > t_{\alpha,n-1} \\ 0 & \text{o.w.} \end{cases} \tag{3.9}$$

Power function of the test is

$$\beta_\mu \left(\phi \right) = P_\mu \left(T_{SRS} > t_{\alpha,n-1} \right)$$

Note that if the sample is drawn using RSS, then it is very difficult to compute the exact value of c, as the distribution is completely unknown to us.

So to test the same hypothesis using the RSS sample, we use the test statistic $T_{RSS} = \frac{\sqrt{n}(\bar{X}_{[\cdot]} - \mu_0)}{s^*}$ where $\bar{X}_{[\cdot]} = \frac{1}{n} \sum\limits_{\alpha=1}^{n} \bar{X}_{[\alpha]}$ and $s^{*2} = \frac{1}{n} \sum\limits_{\alpha=1}^{n} \left(X_{[\alpha]} - \bar{X}_{[\cdot]} \right)^2$

Then test function is

$$\phi^*(x) = \begin{cases} 1 & \text{if } T_{RSS} > c \\ 0 & \text{o.w.} \end{cases} \tag{3.10}$$

such that $P_{H_0} \left\{ T_{RSS} > c \right\} = \alpha$

3.3.2 Testing for Normal Mean (Using Heuristic Approach) :

Let $X_1, X_2, X_3, \ldots \overset{iid}{\sim} N(\mu, \sigma^2)$.

Let $\left\{X_{[1]}, X_{[2]}, ..., X_{[n]}\right\}$ be an RSS of size n from the above distribution. $X_{[i]}$'s are independently but not identically distributed.

Let $\bar{X}_{[\cdot]} = \frac{1}{n} \sum\limits_{\alpha=1}^{n} \bar{X}_{[\alpha]}$: sample mean for RSS(n) sample.

Then according to the properties of RSS, we know

$$E\left(\bar{X}_{[\cdot]}\right) = E\left(X_i\right) = \mu \text{ , and}$$
$$V\left(\bar{X}_{[\cdot]}\right) = \frac{\sigma^2}{n} - \frac{1}{n^2}\sum_{\alpha=1}^{n}\left(E\left(X_{[\alpha]} - \mu\right)\right)^2 .$$

Now pdf of $X_{[\alpha]}$:

$$f_{X_{[\alpha]}}(x) = \frac{n!}{(\alpha-1)!(n-\alpha)!}\Phi^{\alpha-1}\left(\frac{x-\mu}{\sigma}\right)\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{n-\alpha}\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx .$$

$$\begin{aligned}
\therefore E\left(X_{[\alpha]} - \mu\right) &= \int_{-\infty}^{\infty}\frac{n!}{(\alpha-1)!(n-\alpha)!}\left(x-\mu\right)\Phi^{\alpha-1}\left(\frac{x-\mu}{\sigma}\right)\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{n-\alpha}\phi\left(\frac{x-\mu}{\sigma}\right)d\left(\frac{x-\mu}{\sigma}\right)\\
&= \frac{n!\,\sigma}{(\alpha-1)!(n-\alpha)!}\int_{-\infty}^{\infty}\left(\frac{x-\mu}{\sigma}\right)\Phi^{\alpha-1}\left(\frac{x-\mu}{\sigma}\right)\left\{1-\Phi\left(\frac{x-\mu}{\sigma}\right)\right\}^{n-\alpha}\phi\left(\frac{x-\mu}{\sigma}\right)d\left(\frac{x-\mu}{\sigma}\right)\\
&= \frac{n!\,\sigma}{(\alpha-1)!(n-\alpha)!}\int_{-\infty}^{\infty} z\Phi^{\alpha-1}\left(z\right)\left\{1-\Phi\left(z\right)\right\}^{n-\alpha}\phi\left(z\right)dz \qquad \text{,} where\ z = \frac{x-\mu}{\sigma}
\end{aligned}$$

Let $\Phi\left(z\right) = u \Rightarrow du = \phi\left(z\right)dz$ and $z = \Phi^{-1}\left(u\right)$

$$\begin{aligned}
\therefore E\left(X_{[\alpha]} - \mu\right) =&\frac{n!\,\sigma}{(\alpha-1)!(n-\alpha)!}\int_{0}^{1}\Phi^{-1}\left(u\right)u^{\alpha-1}\left(1-u\right)^{n-\alpha}du\\
=&\sigma\int_{0}^{1}\Phi^{-1}\left(u\right)\frac{u^{\alpha-1}\left(1-u\right)^{\overline{n-\alpha+1}-1}}{\mathrm{B}\left(\alpha, n-\alpha+1\right)}du\\
=&\sigma E\left(\Phi^{-1}(u)\right), \qquad where\ \ U \sim Beta(\alpha, n-\alpha+1)\\
\cong&\sigma\frac{1}{N}\sum_{i=1}^{N}\Phi^{-1}(u_i)
\end{aligned}$$

,where u_i's are samples from $Beta(\alpha, n-\alpha+1)$ distribution,

N is a very large number, say $N = 5000$ or 10000 etc.

$$\begin{aligned}
c(\alpha, n) &= E\left(\Phi^{-1}(u)\right) \qquad , say\\
&\cong \frac{1}{N}\sum_{i=1}^{N}\Phi^{-1}(u_i)
\end{aligned}$$

$$\therefore \tfrac{1}{n^2}\sum_{\alpha=1}^{n}\left(E\left(X_{[\alpha]} - \mu\right)\right)^2 = \sigma^2\tfrac{1}{n^2}\sum_{\alpha=1}^{n}c^2\left(\alpha, n\right)$$

$$\therefore V\left(\bar{X}_{[\cdot]}\right) = \tfrac{\sigma^2}{n} - \sigma^2\tfrac{1}{n^2}\sum_{\alpha=1}^{n}c^2\left(\alpha, n\right) = \tfrac{\sigma^2}{n}\left[1 - \tfrac{1}{n}\sum_{\alpha=1}^{n}c^2\left(\alpha, n\right)\right]$$

Also note that, in the similar way,

$$
\begin{aligned}
E\left(X_{[\alpha]}-\mu\right)^2 &= \sigma^2 \int_0^1 \left\{\Phi^{-1}\left(u\right)\right\}^2 \frac{u^{\alpha-1}\left(1-u\right)^{\overline{n-\alpha+1}-1}}{\mathrm{B}\left(\alpha, n-\alpha+1\right)} du \\
&= \sigma^2 E\left(\Phi^{-1}(u)\right)^2, U \sim Beta(\alpha, n-\alpha+1) \\
\therefore E\left[\frac{\left(X_{[\alpha]}-\mu\right)^2}{c^*\left(\alpha, n\right)}\right] &= \sigma^2, \text{ where } c^*\left(\alpha, n\right) \cong \frac{1}{N}\sum_{i=1}^{N}\left\{\Phi^{-1}(u_i)\right\}^2
\end{aligned}
$$

Hence based on n RSS sample observations $X_{[1]}, X_{[2]}, ..., X_{[n]}$, we have

$$
\begin{aligned}
\sigma^2 =& \frac{1}{n}\sum_{\alpha=1}^{n} E\left[\frac{\left(X_{[\alpha]}-\mu\right)^2}{c^*\left(\alpha, n\right)}\right] \\
\Rightarrow \hat{\sigma^2}_{RSS} =& \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\left(X_{[\alpha]}-\mu\right)^2}{c^*\left(\alpha, n\right)} \quad \text{if } \mu \text{ is known (Unbiased for } \sigma^2 \text{)} \\
=& \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\left(X_{[\alpha]}-\bar{X}_{[.]}\right)^2}{c^*\left(\alpha, n\right)} \quad \text{if } \mu \text{ is unknown (Biased for } \sigma^2 \text{)}
\end{aligned}
$$

Now, for testing $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$ (say), we can define a heuristic test statistic

$$
T_{RSS} = \frac{\sqrt{n}\left(\bar{X}_{[.]}-0\right)}{\sqrt{\hat{\sigma^2}_{RSS}\left[1-\frac{1}{n}\sum_{\alpha=1}^{n} c^2\left(\alpha, n\right)\right]}} \tag{3.11}
$$

where

$$
\hat{\sigma^2}_{RSS} = \frac{1}{n}\sum_{\alpha=1}^{n}\frac{\left(X_{[\alpha]}-\bar{X}_{[.]}\right)^2}{c^*\left(\alpha, n\right)}
$$

Hence, our test function for testing $H_0 : \mu = \mu_0 \ vs \ H_1 : \mu > \mu_0$ is

$$
\phi^*(x) = \begin{cases} 1 & \text{if } T_{RSS} > c \\ 0 & o.w. \end{cases} \tag{3.12}
$$

This critical region c can be formed using the empirical approach with the help of computer software.

3.3.3 Testing for Normal Variance

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, $\mu \in \mathbb{R}, \sigma^2 > 0$, μ unknown.

Consider the problem for testing $H_0 : \sigma^2 = \sigma_0^2$ vs. $H_1 : \sigma^2 > \sigma_0^2$

Then using NP lemma we can easily state that the MP test for (H_0, H_1) is given by

$$\phi(x) = \begin{cases} 1 & \text{if } \frac{(n-1)s^2}{\sigma_0^2} > c \\ 0 & \text{o.w.} \end{cases}$$

where $\bar{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_i$ and $s^2 = \frac{1}{n-1} \sum\limits_{i=1}^{n} \left(X_i - \bar{X} \right)^2$

Let, $T_{SRS} = \frac{(n-1)s^2}{\sigma_0^2}$ be the test statistic.

Note that if the sample is drawn using SRS from the population $N(\mu, \sigma_0^2)$, then $T_{SRS} \overset{H_0}{\sim} \chi^2_{n-1}$

Then a level $\alpha = 0.05$ will give us

$$P_{H_0} \left(T_{SRS} > c \right) = \alpha$$
$$\Rightarrow c = \chi^2_{\alpha, n-1}$$

Hence

$$\phi_1(x) = \begin{cases} 1 & \text{if } T_{SRS} > \chi^2_{\alpha, n-1} \\ 0 & \text{o.w.} \end{cases} \tag{3.13}$$

Power function of the test is

$$\beta_\mu \left(\phi \right) = P_\mu \left(T_{SRS} > \chi^2_{\alpha, n-1} \right)$$

Note that if the sample is drawn using RSS, then it is very difficult to compute the exact value of c, as the distribution is completely unknown to us.

So to test the same hypothesis using the RSS sample, we use the test statistic $T_{RSS} = \frac{1}{\sigma_0^2} \sum\limits_{i=1}^{n} \left(X_{[i]} - \bar{X}_{[\cdot]} \right)^2$

Then test function is

$$\phi_1^*(x) = \begin{cases} 1 & \text{if } T_{RSS} > c \\ 0 & \text{o.w.} \end{cases} \tag{3.14}$$

such that $P_{H_0} \left\{ T_{RSS} > c \right\} = \alpha$

3.3.4 Testing for Exponential Mean

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim}$exponential distribution with mean $\frac{1}{\lambda}$, $\lambda > 0$.

Consider the problem for testing $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda > \lambda_0$

Then using NP lemma we can easily state that the MP test for (H_0, H_1) is given by

$$\phi(x) = \begin{cases} 1 & \text{if } 2\lambda_0 \sum\limits_{i=1}^{n} X_i < c \\ 0 & \text{o.w.} \end{cases}$$

Let, $T_{SRS} = 2\lambda_0 \sum_{i=1}^{n} X_i$ be the test statistic.

Note that if the sample is drawn using SRS from the population $Exp(mean = \frac{1}{\lambda_0})$, then, $T_{SRS} \overset{H_0}{\sim} \chi^2_{2n}$

Then a level $\alpha = 0.05$ will give us

$$\begin{aligned} &P_{H_0}\left(T_{SRS} < c\right) = \alpha \\ &\Rightarrow c = \chi^2_{1-\alpha,2n} \end{aligned}$$

Hence

$$\phi_2(x) = \begin{cases} 1 & \text{if } T_{SRS} < \chi^2_{1-\alpha,2n} \\ 0 & \text{o.w.} \end{cases} \tag{3.15}$$

Power function of the test is

$$\beta_\mu\left(\phi\right) = P_\mu\left(T_{SRS} < \chi^2_{1-\alpha,2n}\right)$$

Note that if the sample is drawn using RSS, then it is very difficult to compute the exact value of c, as the distribution is completely unknown to us.

So to test the same hypothesis using the RSS sample, we use the test statistic $T_{RSS} = 2\lambda_0 \sum_{i=1}^{n} X_{[i]}$

Then the test function is

$$\phi_2^*(x) = \begin{cases} 1 & \text{if } T_{RSS} < c \\ 0 & \text{o.w.} \end{cases} \tag{3.16}$$

such that $P_{H_0}\left\{T_{RSS} < c\right\} = \alpha$

4 SIMULATION STUDY AND ANALYSIS

Simulation studies are computer experiments that involve creating data by pseudo-random sampling from known probability distributions. They are an invaluable tool for statistical research, particularly for the evaluation of new methods and for the comparison of alternative methods.

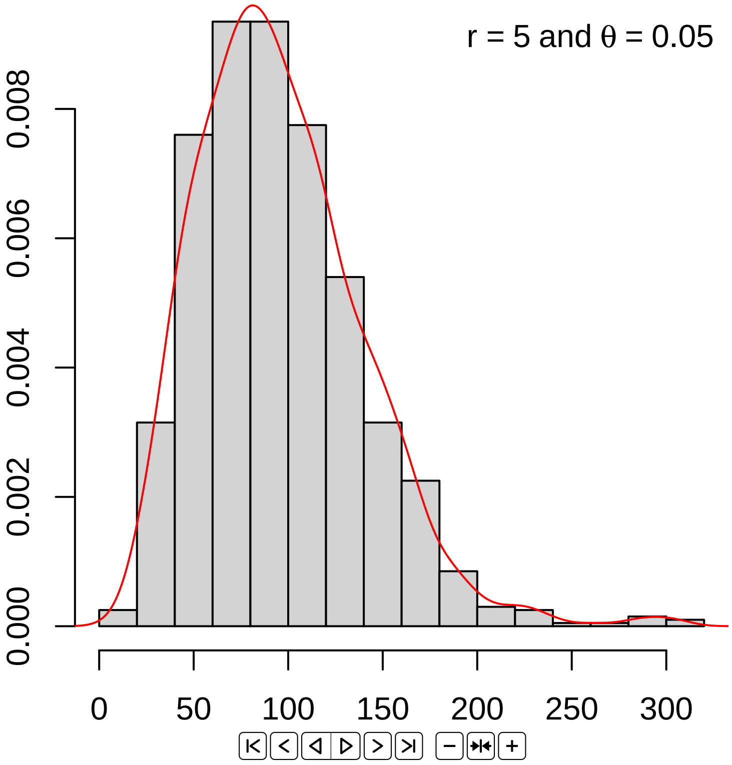
4.1 Error in Estimating Parameters

4.1.1 Method of Moments Estimator :

Negative Binomial to Poisson Distribution-

Before computing the MSE for $NB(r, \theta)$ using MME note that as θ increases, the distribution tends to Poisson distribution. The following animated graph[1] will clearly show that :

Figure 4.1: Animated Histogram of $NB(r, \theta)$ as θ increases



Hence, estimating two parameters will not be appropriate in this situation. So, we restrict ourselves to small values of θ for estimating the parameters of $NB(r, \theta)$.

[1] Not all pdf readers are capable of running animation. We recommend to use Adobe Reader.

Computation of MSE of r and θ -

To compute the MSE mentioned in Section (3.1.1) for $NB(r, \theta)$ using Method of Moments we will use computer simulation. The algorithm of the simulation is as follows :

1. We draw a random sample $\{X_1, X_2, ..., X_n\}$ of size $n = 50$ from $NB(r, \theta)$ using SRS and RSS sampling.
2. Estimate the parameters using the formula (3.3) to get $\hat{r}_1^{SRS}$ & $\hat{\theta}_1^{SRS}$ for SRS and $\hat{r}_1^{RSS}$ & $\hat{\theta}_1^{RSS}$ for RSS .
3. Repeat Step 1 and 2 thousand times to get $\{\hat{r}_1^{SRS}, \hat{r}_2^{SRS}, ..., \hat{r}_{1000}^{SRS}\}$ & $\{\hat{\theta}_1^{SRS}, \hat{\theta}_2^{SRS}, ..., \hat{\theta}_{1000}^{SRS}\}$ for SRS and $\{\hat{r}_1^{RSS}, \hat{r}_2^{RSS}, ..., \hat{r}_{1000}^{RSS}\}$ & $\{\hat{\theta}_1^{RSS}, \hat{\theta}_2^{RSS}, ..., \hat{\theta}_{1000}^{RSS}\}$ for RSS samples.
4. Then we calculate $MSE(\hat{r}^{SRS})$ & $MSE(\hat{\theta}^{SRS})$ for SRS and $MSE(\hat{r}^{RSS})$ & $MSE(\hat{\theta}^{RSS})$ for RSS using the Formula (3.1) and Formula (3.2).

The following table shows the outcome of our simulation :

Table 4.1: MSE under MME for $NB(r, \theta)$

Parameter Values	MSE Value		Parameter Values	MSE Value	
	SRS	RSS		SRS	RSS
$r = 5$	2.84	0.97	r =5	272.04	6.17
$\theta = 0.23$	0.0030	0.0013	$\theta = 0.53$	0.0175	0.0082
$r = 7$	4.03	0.0011	$r = 7$	1450.83	267.51
$\theta = 0.14$	1.52	0.0005	$\theta = 0.86$	0.0142	0.0138
$r = 10$	54.5	10.9	$r = 10$	23989.91	0.0179
$\theta = 0.46$	0.0107	0.0054	$\theta = 0.61$	39.63	0.0090
$r = 16$	51.63	15.05	$r = 16$	7738.25	4109.58
$\theta = 0.37$	0.0066	0.0031	$\theta = 0.72$	0.0183	0.0107

4.1.2 Maximum Likelihood Estimator :

Here, we compute the MSE mentioned in Section (3.1.2) for $NB(r, \theta)$ by Maximum Likelihood using computer simulation. The similar steps of section 4.1.1 will be followed except in Step 2 we will use the formula (3.4) and (3.5). But note that to compute $\hat{r}$ and $\hat{\theta}$ we use an iterative procedure.

The following table shows the outcome of our simulation :

Table 4.2: MSE under MLE for $NB(r, \theta)$

Parameter Values	MSE Value		Parameter Values	MSE Value	
	SRS	RSS		SRS	RSS
$r = 5$	3.38	0.26	r =5	285.29	0.69
$\theta = 0.23$	0.0036	0.0004	$\theta = 0.53$	0.0174	0.0019
$r = 7$	4.46	0.0012	$r = 7$	8746.3	1.54
$\theta = 0.14$	0.33	0.0001	$\theta = 0.63$	0.0242	0.0017
$r = 10$	54.80	1.32	$r = 10$	23583.78	14.92
$\theta = 0.46$	0.0112	0.0009	$\theta = 0.78$	0.0304	0.0030
$r = 16$	92.71	2.27	$r = 16$	20406.06	10.75
$\theta = 0.37$	0.0081	0.0005	$\theta = 0.72$	0.0273	0.0017

⋆ Remark

We can see from Table (4.1) and Table (4.2), RSS shows less error (Mean squared Error) for both ($\hat{r}$ & $\hat{\theta}$) the estimates. Hence, we can conclude that if we use RSS sampling, then we can get a better precision in estimating parameters than under SRS sampling.

4.2 RSS in Regression

4.2.1 Poisson Model

First of all we need to construct a Poisson Model using SRS and RSS sampling for our regression. For that we proceed with the following steps :

1. We draw a covariate matrix $X^{n\times p}$ from $N(0,1)$ and let $\beta^{p\times 1}$ be our true coefficients parameters.
2. Then, let $\lambda_i = e^{x_i^T\beta}$ for $i = 1, 2, ..., n$.
3. Now draw a single random sample from $poisson(\lambda_i)$ using SRS and RSS and call it y_i, $\forall i = 1(1)n$.
4. Fit a Poisson regression with response $y^{n\times 1}$ and covariates $X^{n\times p}$.
5. The standard error corresponding to the estimate $\hat{\beta}$ is recorded for both RSS and SRS.
6. Then we calculate the relative efficiency of RSS with respect to SRS i.e. we calculate $\xi = \frac{Var(\hat{\beta}_{SRS})}{Var(\hat{\beta}_{RSS})}$.

For this simulation study we take $p = 2$. The following Table 4.3 shows the relative efficiency of the RSS with respect to SRS for $n = 20, 50$ and 100.

Table 4.3: Relative Efficiency of RSS in Regression Estimate

True β	Relative Efficiency		
	$n = 20$	$n = 50$	$n = 100$
$\beta_1 = 1.5$	**0.943**	**1.066**	**1.092**
$\beta_2 = 0.3$	**0.924**	**1.022**	**1.053**
$\beta_1 = 1.63$	0.919	1.159	0.841
$\beta_2 = 1.03$	0.972	1.101	0.916
$\beta_1 = 1.6$	0.980	0.909	1.083
$\beta_2 = 0.52$	0.962	0.971	1.013
$\beta_1 = 0.55$	**0.828**	**1.182**	**1.154**
$\beta_2 = 0.83$	**0.891**	**1.099**	**1.006**
$\beta_1 = 0.801$	0.790	0.999	1.026
$\beta_2 = 0.388$	0.837	0.869	1.034
$\beta_1 = 0.93$	**1.055**	**1.081**	**1.235**
$\beta_2 = 0.27$	**1.081**	**0.919**	**1.195**

4.2.2 Negative Binomial Model

Here we need to construct a Negative Binomial Model using SRS and RSS sampling for our regression. Here we follow the steps mentioned in Section 4.2.1 but here we fix a value of r for $NB(r,\theta)$ and in Step 2 we calculate $\theta_i = \frac{1}{1+e^{x_i^T\beta}}$ $\forall i = 1(1)n$. Then we draw random sample from $NB(r,\theta_i)$. Then we fit a Poisson regression Model and check the efficiency of RSS.

For this simulation study we take $p = 2$, i.e. we consider two covariates. Then Table shows the relative efficiency of the RSS with respect to SRS for $n = 20, 50$ and 100.

Table 4.4: Relative Efficiency of RSS in Regression Estimate

r	True β	Relative Efficiency		
		$n = 20$	$n = 50$	$n = 100$
$r = 4$	$\beta_1 = 1.36$	0.888	1.009	1.045
	$\beta_2 = 0.93$	0.919	0.979	1.087
	$\beta_1 = 1.91$	0.863	0.934	1.018
	$\beta_2 = 1.36$	0.899	0.864	0.983
	$\beta_1 = 0.57$	0.901	0.948	1.173
	$\beta_2 = 1.41$	0.939	0.835	1.057
$r = 11$	$\beta_1 = 1.05$	0.960	1.021	1.028
	$\beta_2 = 0.403$	0.947	1.011	1.007
	$\beta_1 = 1.85$	1.099	0.884	0.971
	$\beta_2 = 1.47$	1.086	0.878	0.975
	$\beta_1 = 0.63$	**1.057**	**1.094**	**1.043**
	$\beta_2 = 0.06$	**1.068**	**1.081**	**1.075**
$r = 15$	$\beta_1 = 0.69$	0.982	1.039	1.015
	$\beta_2 = 1.79$	0.989	1.037	1.025
	$\beta_1 = 0.96$	0.938	0.989	1.047
	$\beta_2 = 0.51$	0.920	1.007	1.023
	$\beta_1 = 0.07$	**1.069**	**1.037**	**1.041**
	$\beta_2 = 0.15$	**1.071**	**1.007**	**1.018**
$r = 21$	$\beta_1 = 1.66$	0.926	0.996	0.998
	$\beta_2 = 1.38$	0.966	1.018	1.071
	$\beta_1 = 0.62$	0.945	0.987	1.015
	$\beta_2 = 0.26$	0.946	0.978	1.011
	$\beta_1 = 1.84$	**1.105**	**1.034**	**1.053**
	$\beta_2 = 0.92$	**1.097**	**1.012**	**1.068**

★ Remark

From the values of Table (4.3) and Table (4.4), we actually can not give any particular answer to our question. We can see that the relative efficiency of RSS is sometimes greater than 1 (**Marked with Bold font style**) and sometimes less than 1. But one thing is clear that RSS does not perform worse than SRS in any situation of regression estimate, they are more or less equally capable to capture the variability in the model. Further from the Table (4.4) we can see that for large sample most of the time efficiency is more than one, i.e. RSS performs better than SRS in those situation. So, to get more precision in estimating model parameter one can always prefer RSS sampling.

4.3 Hypothesis Testing

4.3.1 Testing for Normal Mean

For $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ we will conduct a test $H_0 : \mu = \mu_0 \; vs \; H_1 : \mu > \mu_0$ using the test statistic mentioned in Formula (3.8). Note that the cutoff c mentioned in the Formula (3.9) works for a SRS sample, but in case of RSS we will use a computer simulation to calculate the critical value c. The algorithm of the simulation to get the value of c is as follows:

1. We generate a sample of size n from $N(\mu_0, \sigma^2)$ using RSS sampling to get $X_{[1]}, X_{[2]}, ..., X_{[n]}$ and let $\bar{X}_1 = \frac{1}{n}\sum_{i=1}^{n} X_{[i]}$.
2. Then we compute $T_1 = \frac{\sqrt{n}(\bar{X}_1 - \mu_0)}{s_1} \; ; where \; s_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_{[i]} - \bar{X}_1)^2$
3. Repeat step 1 and 2 five thousand times $(k = 1, 2, ..., 5000)$ to find $T_1, T_2, ..., T_{5000}$.
4. Order T_k , $k = 1, 2, ..., 5000$ to get $T_{(1)}, T_{(2)}, ..., T_{(5000)}$.
5. With $\alpha = 0.05$ we get $c \equiv T_{(4750)}$.

Now to compute the empirical power of the test for SRS and RSS the following steps are used:

1. Generate a sample of size n from $N(\mu, \sigma^2)$ (where $\mu > \mu_0, i.e.$ under alternative hypothesis) using both SRS and RSS sampling.
2. Compute T_1^{SRS} for SRS sample and T_1^{RSS} for RSS sample as mentioned in Formula (3.8).
3. Repeat step 1 and 2 five thousand times to get $T_1^{SRS}, T_2^{SRS}, ..., T_{5000}^{SRS}$ and $T_1^{RSS}, T_2^{RSS}, ..., T_{5000}^{RSS}$.
4. Then find how many of the $T_k^{SRS} > t_{n-1;\frac{\alpha}{2}}$,say d_1 and $T_k^{RSS} > c$, say d_2.
5. Then the empirical power of the test for SRS is ,$\beta_\mu(\phi) = \frac{d_1}{5000}$ and RSS is ,$\beta_\mu(\phi^*) = \frac{d_2}{5000}$.

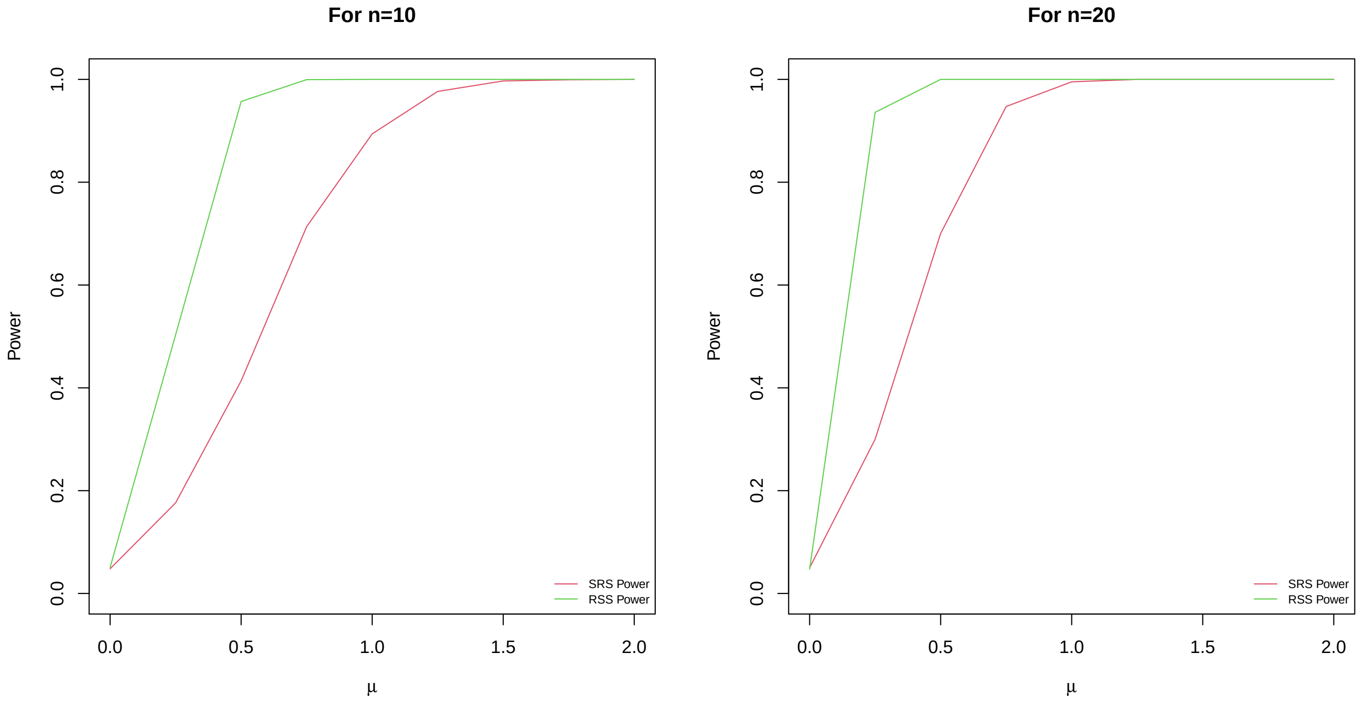
Consider the test $H_0 : \mu = 0 \; vs \; \mu > 0$. To compare the two tests ϕ and ϕ^*, we take $\alpha = 0.05$ and $n = 10$ and 20. Table (4.5) shows the power for both tests for $n = 10$ and 20 and $\alpha = 0.05$.

Table 4.5: $\beta_\mu(\phi)$ and $\beta_\mu(\phi^*)$ for testing of mean in $N(\mu, \sigma^2)$

σ^2	μ	$\beta_\mu(\phi)$		$\beta_\mu(\phi^*)$	
		$n = 10$	$n = 20$	$n = 10$	$n = 20$
1	2.00	1.0000	1.0000	1.0000	1.0000
	1.75	0.9993	1.0000	1.0000	1.0000
	1.50	0.9970	1.0000	1.0000	1.0000
	1.25	0.9766	1.0000	1.0000	1.0000
	1.00	0.8940	0.9953	1.0000	1.0000
	0.75	0.7136	0.9473	0.9996	1.0000
	0.50	0.4136	0.7003	0.9570	1.0000
	0.25	0.1763	0.3000	0.5033	0.9360
	0.00	0.0480	0.0500	0.0506	0.0473
4	2.00	0.8963	1.0000	1.0000	1.0000
	1.75	0.8053	1.0000	1.0000	1.0000
	1.50	0.7163	1.0000	0.9990	1.0000
	1.25	0.5640	1.0000	0.9906	1.0000
	1.00	0.4233	0.9940	0.9540	1.0000
	0.75	0.2796	0.9450	0.8203	1.0000
	0.50	0.1940	0.7076	0.5366	1.0000
	0.25	0.1000	0.2883	0.2180	0.9510
	0.00	0.0460	0.0533	0.0493	0.0583

The following Graph clearly shows Power of the tests for SRS and RSS :

Figure 4.2: Power Curve for different μ of $N(\mu, 1)$ when $\alpha = 0.05$



4.3.2 Testing for Normal Mean (Using Heuristic Approach):

Now we will use the test statistic mentioned in section (3.3.2) to conduct a test of $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$. The similar steps mentioned in section (4.3.1) except here we use the test statistic of Formula (3.11). The Table 4.6 shows the empirical power of the test ϕ_1^* for $\alpha = 0.05$ and $n = 10$ and 20.

Table 4.6: $\beta_\mu(\phi_1^*)$ in RSS Test for testing of mean in $N(\mu, \sigma^2)$

σ^2	μ	$\beta_\mu(\phi_1^*)$		σ^2	μ	$\beta_\mu(\phi_1^*)$	
		$n = 10$	$n = 20$			$n = 10$	$n = 20$
1	2.00	1.0000	1.0000	4	2.00	1.0000	1.0000
	1.75	1.0000	1.0000		1.75	1.0000	1.0000
	1.50	1.0000	1.0000		1.50	0.9996	1.0000
	1.25	1.0000	1.0000		1.25	0.9924	1.0000
	1.00	1.0000	1.0000		1.00	0.9628	0.9900
	0.75	0.9996	1.0000		0.75	0.8236	0.9988
	0.50	0.9416	1.0000		0.50	0.5588	0.9416
	0.25	0.4796	0.9432		0.25	0.2260	0.4896
	0.00	0.0516	0.0560		0.00	0.0472	0.0592

Remark : From the Table 4.5 and Table 4.6 , if we compare the values of Power of the test for RSS sampling, we can see that though the test statistics are different, the power of the tests are almost equal. In testing $H_0 : \mu = \mu_0$ *vs* $H_1 : \mu > \mu_0$ we can use either one of the test statistic. As the test statistic in Section (3.3.1) is computationally more easy to work with, we will can use that in this testing purpose.

4.3.3 Testing for Normal Variance :

Now for $X_1, X_2, ..., X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ we conduct a test of $H_0 : \sigma^2 = \sigma_0^2$ *vs* $H_1 : \sigma^2 > \sigma_0^2$. Here the test statistic is as mentioned in Formula (3.13). Here, we can also propose a test statistic for RSS separately, but in this case also the power of the test will be almost equal. Hence, here we proceed with the test using the test statistic for RSS.

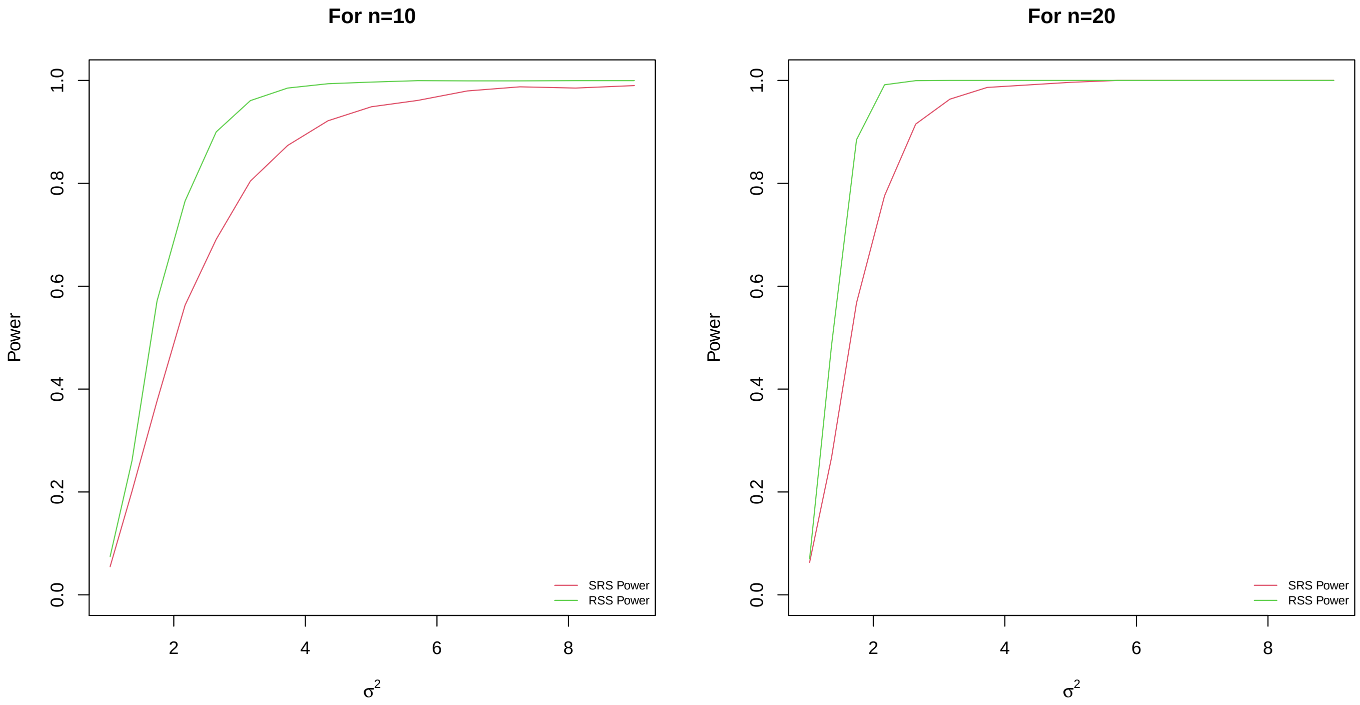
The similar steps mentioned in section (4.3.1) are carried out, except here we use the test statistic of Formula (3.14). For our simulation study we assume $\sigma_0^2 = 1$. The Table 4.7 shows the empirical power of the test ϕ_2 and ϕ_2^* for $\alpha = 0.05$ and $n = 10$, 15 and 20.

Table 4.7: $\beta_\mu(\phi_2)$ and $\beta_\mu(\phi_2^*)$ for testing of Variance in $N(\mu, \sigma^2)$

σ^2	$\beta_\mu(\phi_2)$			$\beta_\mu(\phi_2^*)$		
	$n = 10$	$n = 15$	$n = 20$	$n = 10$	$n = 15$	$n = 20$
1.03182825	0.0548	0.0576	0.0632	0.0744	0.0588	0.0700
1.36520776	0.2016	0.2300	0.2668	0.2612	0.3572	0.4848
1.74518006	0.3768	0.4772	0.5676	0.5712	0.7420	0.8848
2.17174515	0.5632	0.7004	0.7760	0.7656	0.9300	0.9916
2.64490305	0.6912	0.8244	0.9152	0.9000	0.9864	0.9996
3.16465374	0.8044	0.9152	0.9636	0.9608	0.9972	1.0000
3.73099723	0.8736	0.9564	0.9864	0.9852	1.0000	1.0000
4.34393352	0.9216	0.9828	0.9912	0.9936	1.0000	1.0000
5.00346260	0.9488	0.9912	0.9964	0.9968	1.0000	1.0000
5.70958449	0.9612	0.9944	1.0000	0.9996	1.0000	1.0000
6.46229917	0.9796	0.9968	1.0000	0.9992	1.0000	1.0000
7.26160665	0.9876	0.9980	1.0000	0.9992	1.0000	1.0000
8.10750693	0.9852	1.0000	1.0000	0.9996	1.0000	1.0000
9.00000000	0.9900	1.0000	1.0000	0.9996	1.0000	1.0000

The following Graph clearly shows the Power of the tests for SRS and RSS :

Figure 4.3: Power Curve for different σ^2 of $N(0, \sigma^2)$ when $\alpha = 0.05$



4.3.4 Testing for Exponential Mean :

Suppose $X_1, X_2, ..., X_n \overset{iid}{\sim} exp(mean = \frac{1}{\lambda})$ we conduct a test of $H_0 : \lambda = \lambda_0 \ vs \ H_1 : \lambda > \lambda_0$. Here the test statistic is as mentioned in Formula (3.15). The simile steps mentioned in section (4.3.1) are carried out, except here we use the test statistic of Formula (3.16) and as we consider the lower tail here, the value of critical region under RSS sample is $c = T_{(250)}$.

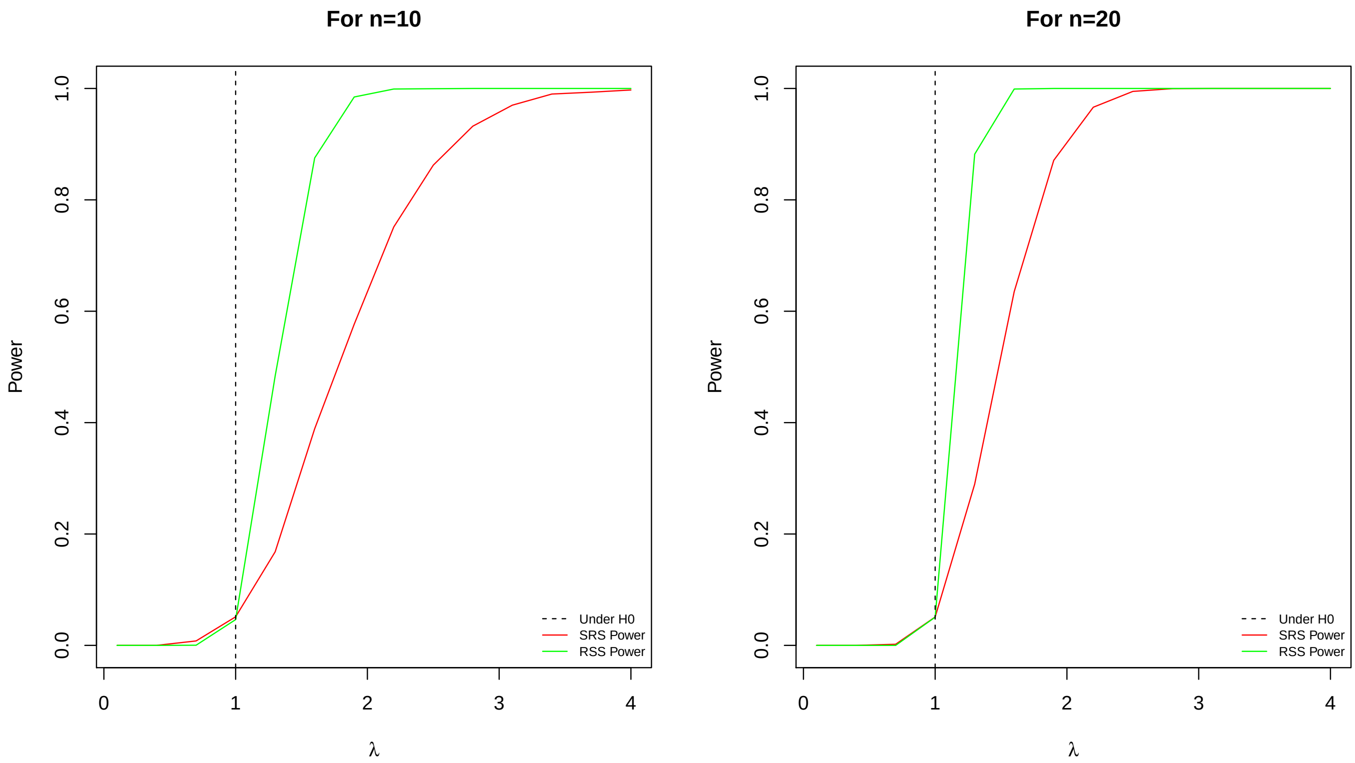
For this simulation study we take $\lambda_0 = 1$. Than the Table shows the empirical power of the test ϕ_3 and ϕ_3^* for $\alpha = 0.05$ and $n = 10, 15$ and 20.

Table 4.8: $\beta_\lambda(\phi_3)$ and $\beta_\lambda(\phi_3^*)$ for testing of mean in $exp(\frac{1}{\lambda})$

λ	$\beta_\lambda(\phi_3)$			$\beta_\lambda(\phi_3^*)$		
	$n = 10$	$n = 15$	$n = 20$	$n = 10$	$n = 15$	$n = 20$
0.7	0.0080	0.0017	0.0020	0.0003	0.0000	0.0000
1.0	0.0513	0.0563	0.0510	0.0463	0.0437	0.0507
1.3	0.1680	0.2267	0.2896	0.4837	0.7160	0.8820
1.6	0.3890	0.5280	0.6353	0.8753	0.9813	0.9990
1.9	0.5767	0.7680	0.8710	0.9847	0.9990	1.0000
2.2	0.7513	0.9187	0.9663	0.9990	1.0000	1.0000
2.5	0.8623	0.9727	0.9947	0.9996	1.0000	1.0000
2.8	0.9323	0.9947	0.9997	1.0000	1.0000	1.0000
3.1	0.9700	0.9993	1.0000	1.0000	1.0000	1.0000
3.4	0.9900	0.9997	1.0000	1.0000	1.0000	1.0000
3.7	0.9933	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	0.9973	1.0000	1.0000	1.0000	1.0000	1.0000

The following Graph clearly shows the Power of the test for SRS and RSS :

Figure 4.4: Power Curve for different λ of $exp(\frac{1}{\lambda})$ when $\alpha = 0.05$



⋆ Remark

We have seen in Section (3.3.2) that it's difficult to compute the distribution of test statistic for RSS sampling even in Normal setting but our empirical study[2] can show some results. Through empirical study, here we are able to conduct the power function for the testing problems. From the above Figures (4.2,4.3,4.4) we can clearly see that in each of the testing procedure we get a significant amount of increment in the power curve for using RSS compared to using SRS. Hence, though we aren't able to find the exact distribution of RSS test statistic we can conclude that using empirical approach we can get a more Powerful test for using RSS than SRS.

[2]

R-Code : Click Here to download the R Codes.

5 CONCLUSION

Ranked set sampling is a procedure which may be used to improve the precision of the estimator of the mean. It is useful in cases where the variable of interest is much more difficult to measure than to order. It should be noted that, in the estimation of variance, RSS is not necessarily more efficient than SRS when sample size is small, and the relative efficiency is much smaller than in the estimation of population mean even when RSS is beneficial. Therefore, if the estimation of variance is the primary purpose, it is not worthwhile to apply RSS. RSS is most useful when both the population mean and variance are to be estimated.

We see that the variances of estimated coefficients are same as those of their counterparts based on an SRS, which implies that, in the estimation of the regression coefficients, RSS and SRS are quite equivalent. RSS cannot do much for the improvement of the estimation of the regression coefficients.

We consider testing some hypotheses about μ and σ^2 of the Normal distribution and λ of Exponential distribution using RSS. It appears that the use of RSS gives much better results in terms of the empirical power function compared to the SRS. Though as the sample size increases the power of SRS increases, still RSS shows a significant improvement in terms of empirical Power of the test. Hence, RSS can be used wherever possible to achieve a more powerful test even in small sample.

Lastly, we say what started as a simple attempt by McIntyre [1] to utilize additional information to improve precision in the estimation of pasture yields through the selection of more representative sample observations has clearly grown into a major field of statistical methodology that continues to attract substantial research activity.

6 REFERENCES

1. G. A. McIntyre, “A method for unbiased selective sampling, using ranked sets,” Australian Journal of Agricultural Research, vol. 3, pp. 385–390, 1952..

2. K Takahasi, K Wakimoto, “On unbiased estimates of the population mean based on the sample stratified by means of ordering” Annals of the institute of statistical mathematics, 1968

3. R. Dell and J. L. Clutter, “Ranked set sampling theory with order statistics background,” Biometrics, vol. 28, pp. 545–555, 1972

4. K. Takahasi, “Practical note on estimation of population means based on samples stratified by means of ordering,” Annals of the Institute of Statistical Mathematics, vol. 22, no. 1, pp. 421–428, 1970

5. Douglas A. Wolfe, “Ranked Set Sampling: Its Relevance and Impact on Statistical Inference” ISRN Probability and Statistics Volume 2012, Article ID 568385, 32 pages.

6. Hassen A. Muttlak a & Walid Abu-Dayyeh, “Testing some hypotheses about the normal distribution using ranked set sample: a more powerful test” Journal of Information and Optimization Sciences, 19:1, 1-11.

7. [Book] Zehua Chen, Zhidong Bai, Bimal K. Sinha; “Ranked Set Sampling: Theory and Application” Springer, Berlin Heidelberg NewYork.