A Study on the Performance of Different Types of Estimators in Finite Population Sampling

Project Submitted to the P.G. Department of Statistics, Utkal University for the Partial fulfillment of Degree of Master of Science in Statistics

Submitted by
Aman Kumar Patel
Roll No – 12612V174002

Under the Supervision of
Dr. Priyaranjan Dash
Faculty, P.G. Department of Statistics,
Utkal University

P.G. DEPARTMENT OF STATISTICS
UTKAL UNIVERSITY, VANI VIHAR
BHUBANESWAR-751004
MAY 2019

Dr. Priyaranjan Dash.
Faculty,
P.G. Department of Statistics,
Utkal University
Vani Vihar, Bhubaneswar - 751004.

CERTIFICATE

This is to certify that the project entitled that ***"A Study on the Performance of Different Types of Estimators in Finite Population Sampling"*** *is being submitted by Aman Kumar Patel for the degree of Master in Science in Statistics of Utkal University bearing Roll No. 12612V174002 is a record of bonafide research work carried out under my guidance and supervision.*

The project has reached the standard, fulfilling the requirements for the regulation relating to M.Sc. of this University. This work has not been submitted to any other University or research institute for the award of any degree or diploma.

Dr. Priyaranjan Dash
(Supervisor)

DECLARATION

I do hereby declare that the thesis entitled ***"A Study on the Performance of Different Types of Estimators in Finite Population Sampling"*** *is an original work and that it has not been previously submitted to any University or Research Institute for the award of any Degree or Diploma.*

Aman Kumar Patel
Roll No. - 12612V174002
P.G. Department of Statistics
Utkal University, Vani Vihar
Bhubaneswar

ACKNOWLEDGMENT

I would like to take this opportunity of expressing gratefulness to my Professor and guide Dr. Priyaranjan Dash, P.G. Department of Statistics, Utkal University for his guidance and encouragement to complete piece of work. He also supported me with necessary relevant data related to my research for which I am equally indebted to him.
I express my sincere gratitude to Prof. K.B. Panda, Dr. R.K. Sahoo, Dr. P.K. Swain of P.G. Department of Statistics, Utkal University for their kind help and valuable suggestions.
Also I am thankful to all the office bearers of the P.G. Department of Statistics.
I am also thankful to all my friends in the Departments for their support and encouragement throughout my study.
Last but not the least, I would like to thank all my family members especially my sister whose moral support and constant inspiration helped me in completing this thesis.

Aman Kumar Patel

A Study on the Performance of Different Types of Estimators in Finite Population Sampling

Aman Kumar Patel[1]
Roll No - 12612V174002
P.G. Fourth Semister
P.G. Department of Statistics
Utkal University - 751004.

May 20, 2019

[1]E.Mail. amanthe001@gmail.com

Contents

Chapter 1

INTRODUCTION

1.1 NEED OF THE SAMPLE SURVEY

The theory of survey sampling plays an important role in statistical theory. This theory is extensively used to collect information on different socio-economic and demographic factor, man-power, cultivable land, material and oil etc. for an effective economic planning of the state. Considerable planning is also required to collect these information for taking any rationale decision regarding the resources and the needs. In the present world, the primary users of statistical data are the state, industry, business, scientific institutions, public organizations and international agencies. if the resources are unlimited, then various requirements can be easily met and a state in such a case does not face any difficulty in planning its economic condition. But very often a state falls short of its needs, because of limited resources and then arises the problem of an efficient planning for the best utilization of the available resources. This is because the state cannot afford to completely expand its resources in meeting present needs. But at the same time it must meet the future demands. So for a proper planning of the state fairly detailed data on the various available resources and on the needs are to be collected. For example, a country very often needs data on production and consumption of different types of food grains to enable it to take objective decision regarding its import of the population does not simply deal with an estimator but a sampling strategy.

1.2 USE OF AUXILIARY INFORMATION IN SURVEY SAMPLING

Sometimes information on a variable closely related to the study variable, is available or may be made available at low cost on all the units of the population. This variable is known as auxiliary variable or ancillary variable or a a concomitant variable.
In Sampling theory, the degree of emphasis laid on the use of auxiliary information for improving the precision of the estimates. For instance while estimating the yield of rice, the area under wheat may be known from official sources or may be made available diverting a part of the resources of the same. Then yield and area of rice may be the study variable and auxiliary variable respectively. A statistician may use this auxiliary information of area under rice to improve the estimator of parametric value of the population i.e. yield.

This auxiliary information may be used in different stages:

- Utilization of information at pre-selection stage i.e. for stratifying the population.
- Utilization of information at selection stage i.e. in selecting the units with probabilities proportional to some suitable measure of size (size being based on some auxiliary variables).
- Utilization of information at estimation stage i.e. in formulation of the ratio-type, regression, difference and product estimators etc.

An investigator uses the auxiliary information at the pre selection stage in classifying the a population as in the stratified sampling or uses those as size measure of the corresponding value of the study variable as selected with unequal probabilities in various PPS sampling schemes. Hansen and Hurwitz (1943) pioneered this idea and demonstrated the profitability of selecting sampling units with probability proportional to size of the of the units and suggested the use of auxiliary variable to achieve the same. Cochran (1940) developed the idea of using the auxiliary variable at the estimation stage by forwarding the ratio estimator. He assumed the existence of a positive correlation between the study variable and the auxiliary variable. In a landmark paper Cochran (1942) also discussed the regression model of estimation. Murthy (1964) later on forwarded the

product estimator when the auxiliary variable is found out to be negatively correlated with the study variable. Moreover, the main advantage with the regression estimator over the both ratio and product estimator is that, it can be applied for situations of both positive and negative correlation. The ratio estimator under Lahiri (1951) or Midzuno-Sen (1952) scheme of sampling is an example where the auxiliary character is used at both the selection and estimation stage.
Ratio, regression and product estimator and their modifications usually exploit the information on an auxiliary. In the passage of time, a host of estimators usually auxiliary information has been suggested by many authors which fall into any one of the above categories.
Sometimes information on several auxiliary variable are also available instead of a single auxiliary variable. Fruitful use of these information through multivariate ratio, regression and product estimators has been made by many authors. The significant contribution in this directions are due to Olkin (1958), Raj (1965), Shukla (1966), Singh (1967), Mohanty and Pattanaik (1984) and many others.
Usually the conventional ratio, regression, and product estimators are biased in nature. Attempts have therefore been made by many researchers in different ways to develop unbiased estimators, because this plays an important role while one has to deal with samples of small or moderate sizes.

1.3 DEFINITIONS

Population 1. *In Statistics, a population is a set of similar items or events which is of interest for some question or experiment. A Statistical population can be group of existing objects (e.g. the set of all stars within the milky way galaxy, set of all the peoples living in Sundargarh with title Patel)*

Population 2. *Any finite or infinite aggregation of individuals, not necessarily animate, subject to a statistical study is called as Population.*

Sample. *A finite part of a statistical population whose properties are studied to gain information about the whole*

Estimator. *An estimator is a statistic that estimates some fact about the population.(e.g. Sample mean $\overline{X}$ is an estimator for the population mean μ)*

1.4 SAMPLING

Sampling is a process used in statistical analysis in which a predetermined number of observations are taken from a larger population. The methodology used to sample from a larger population depends on the type of analysis being performed but may include simple random sampling or systematic sampling.

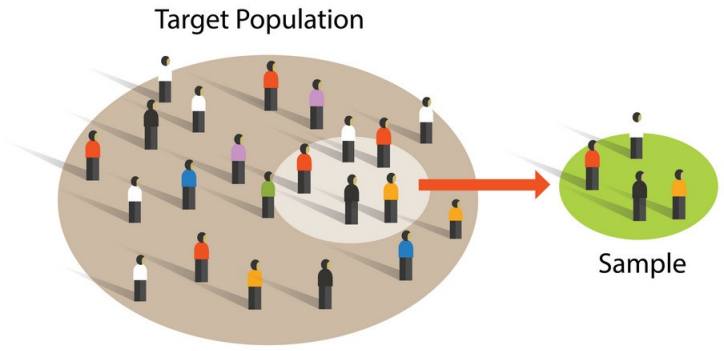


Figure 1.1: Selection of a Sample from a Population

A.C. Rosander, "The sample has many advantages over a census or complete enumeration. If carefully designed, the sample is not only considerably cheaper but may give results which are just accurate and sometimes more accurate than those of a census technique. Hence, a carefully designed sample may actually be better than a poorly planned and executed census."
Prof. R.A. Fisher, the sample technique has four important advantages over census technique of data collection. They are "Speed, Economy, Adaptability and Scientific approach."

Chapter 2

A REVIEW OF DIFFERENT ESTIMATORS

2.1 INTRODUCTION

In survey sampling, it is well established that the use of auxiliary information results in substantial gain in efficiency over the estimators which do not use such information.
Consider a finite population of N units and y is the variable under study taking value y_i for the i-th unit of the population($i = 1, 2, \cdots, N$). Suppose we want to estimate the population mean $\overline{Y}$ of the study variable y. For this we select a random sample of size n from the population and study the sample units only. Let y_i is the value of the study variable y for the i-th unit of the sample ($i = 1, 2, \cdots, n$).
Suppose we have the information on an auxiliary variable x with known Population mean $\overline{X}$, then we can efficiently use this information to estimate $\overline{Y}$.
Here, we have different kinds of estimators to estimate the population mean $\overline{Y}$, such as:

- Simple Mean Estimator
- Ratio Estimator
- Product Estimator
- Exponential Estimator
- Square-root Estimator
- Regression Estimator etc...

Here a question comes to our mind that which estimator will perform the best

2.2 SIMPLE MEAN ESTIMATOR

As we know that sample mean is an unbiased estimator of population mean. So we can consider the sample mean for estimating our population parameter $\overline{Y}$
The simple mean estimator is given by

$$t_0 = \overline{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} yi \tag{2.2.1}$$

Under SRSWOR scheme, $\overline{y}$ is an unbiased estimator of the population mean $\overline{Y}$ with a variance

$$V(t_0) \;=\; \left(\frac{1}{n} - \frac{1}{N}\right) S_y^2, \tag{2.2.2}$$

where $S_y^2 = \dfrac{1}{N-1}\displaystyle\sum_{i=1}^{N}(y_i - \overline{Y})^2$.

2.3 RATIO ESTIMATOR

The ratio estimator for estimating $\overline{Y}$ is given by :

$$t_r \;=\; \frac{\overline{y}}{\overline{x}}\overline{X}. \tag{2.3.3}$$

Ratio estimator is a biased estimator with absolute bias as

$$|Bias(t_r)| \;=\; |\rho_{\hat{R}\overline{x}}\sigma_{\hat{R}}\sigma_{\overline{x}}| \tag{2.3.4}$$

So the ratio estimator t_r will be an unbiased estimator of $\overline{Y}$ only when $cov(\hat{R}, \overline{x})$ is equal to 0 (zero).
As we know relative bias of the ratio estimator t_r cannot exceed the coefficient of variation of $\overline{x}$ and therefore the bias of t_r is a bounded quantity. It may be concluded that, if the coefficient of variation $\overline{x}$ is sufficiently small, say less than $\frac{1}{10}$ the bias as compared to the standard error, may be considered to be negligible, i.e. 'n' is so chosen that

$$CV(\overline{x}) = \left(\frac{1}{n} - \frac{1}{N}\right)^2 \frac{S_x}{\overline{x}} < \frac{1}{10} \tag{2.3.5}$$

Variance of Ratio estimator is given by

$$V(t_r) = \left(\frac{1}{n} - \frac{1}{N}\right) \overline{Y}^2 \left({C_y}^2 + {C_x}^2 - 2\rho C_y C_x\right) \tag{2.3.6}$$

Ratio estimator is applicable when ρ (correlation between auxiliary value and study variable) is grater than 0.5, when the population coefficient of variation of x and y are equal.

2.4 PRODUCT ESTIMATOR

The product estimator for estimating $\overline{Y}$ is given by

$$t_p \;=\; \frac{\overline{y}\,\overline{x}}{\overline{X}}. \tag{2.4.7}$$

Product estimator is a biased estimator with absolute bias

$$|Bias(t_p)| \;=\; \overline{Y}\frac{Cov(\overline{y},\overline{x})}{\overline{y}\,\overline{x}}. \tag{2.4.8}$$

Variance of product estimator is given by:

$$\hat{Var}(t_p) \;=\; \frac{1-f}{n}\left({S_y}^2 + \hat{R}^2{S_x}^2 + 2\hat{R}S_{yx}\right). \tag{2.4.9}$$

Product estimator is applicable when ρ (correlation between auxiliary value and study variable) is less than -0.5.

2.5 EXPONENTIAL ESTIMATOR

The exponential estimator for estimating $\overline{Y}$ is given by :

$$t_e \;=\; \overline{y}\exp\left(\frac{\overline{X}-\overline{x}}{\overline{X}+\overline{x}}\right). \tag{2.5.10}$$

Bias in exponential estimator is given below

$$Bias(t_e) = \theta\overline{Y}\left(\frac{3}{8}{C_x}^2 - \frac{1}{2}\rho C_x C_y\right), \text{where } \theta = \frac{N-n}{(N-1)n} \quad (2.5.11)$$

Variance of the exponential estimator t_e is given by

$$Var(t_e) = \theta\overline{Y}^2\left({C_y}^2 + \frac{1}{4}{C_x}^2 - \rho C_x C_y\right) \quad (2.5.12)$$

Exponential estimator is preferable when k lies between $\frac{1}{4}$ and $\frac{3}{4}$ i.e $\frac{1}{4} < k < \frac{3}{4}$, where k is $\rho\sqrt{\frac{{C_y}^2}{{C_x}^2}}$. Under the condition, $C_y = C_x$, we can simply prefer the exponential estimator t_e when $\frac{1}{4} < \rho < \frac{3}{4}$.

2.6 SQUARE ROOT ESTIMATOR

The square-root estimator for estimating $\overline{Y}$ is given by :

$$t_{sq} = \overline{y}\sqrt{\frac{\overline{X}}{\overline{x}}}. \quad (2.6.13)$$

Bias in square-root estimator is given by

$$Bias(t_{sq}) = \theta\overline{Y}\left(\frac{3}{8}{C_x}^2 - \frac{1}{2}\rho C_x C_y\right) \quad (2.6.14)$$

$$Where\ \theta = \frac{N-n}{(N-1)n}$$

Variance of square-root estimator is given by

$$Var(t_e) = \theta\overline{Y}^2\left({C_y}^2 + \frac{1}{4}{C_x}^2 - \rho C_x C_y\right) \quad (2.6.15)$$

Square-root estimator is applicable when k is grater than $\frac{1}{4}$ and less than $\frac{3}{4}$
i.e $\frac{1}{4} < k < \frac{3}{4}$
where k is $\rho\sqrt{\frac{{C_y}^2}{{C_x}^2}}$. Under the condition, $C_y = C_x$, we can simply prefer the square-root estimator t_e when $\frac{1}{4} < \rho < \frac{3}{4}$.
From the above, it can be seen that both the square-root estimator and the exponential estimator performs similarly in terms of bias and variance up to first order of approximation only.

2.7 REGRESSION ESTIMATOR

The Regression estimator for estimating $\overline{Y}$ is given by :

$$t_{reg} = \overline{y} + \hat{\beta}\left(\overline{X} - \overline{x}\right), \tag{2.7.16}$$

where $\hat{\beta}$ is the sample regression coefficient of y on x.
Variance of this regression estimator t_{reg} is given by:

$$Var(t_{reg}) = \frac{1-f}{n} {S_y}^2 (1-\rho^2) \tag{2.7.17}$$

Chapter 3

PERFORMANCE OF DIFFERENT ESTIMATORS

3.1 TEST USING A HYPOTHETICAL DATA

To generate the hypothetical data and calculations, *Rstudio* is going to help us in our experiments.
First we generate a Bivariate Normal Population

$$BVN(\mu_x = 15, \mu_y = 30, \sigma_x = 3, \sigma_y = 5, \rho),$$

where X is the auxiliary variable and Y is our study variable. Here we generate 5 different bivariate normal populations with 5 different ρ values i.e. 0.25, 0.5, 0.75, 0.85 and 0.95 respectively. We will generate 3 different types of population with population sizes N as 50, 75, and 100. So, we have considered 15 different populations for analysing the performance of these five types of estimators.
For generating the Bivariate normal population, we are going to use the library *mvtnorm*

3.1.1 Generating Bivariate Normal Population

```
library(mutnorm)
```

To generate a hypothetical bivariate population as desire we need to create a function.

```
Rbvn <- function(muX =15,muY =30,rho,sX=3,sY=5,N){
        mu <- c(muX,muY)
        sigma <- matrix(c(sX^2,sX*sY*rho,sX*sY*rho,sY^2),
        nrow = 2,ncol = 2)
        bvn <- mvrnorm(N, mu = mu, Sigma = sigma )
        colnames(bvn) <- c("X","Y")
```

```
        bvn
}
```

Above we already defined the mean X as 15, standard deviation of X as 3, mean Y as 30, standard deviation of Y as 5. so to generate the desired population we need to change the ρ values only.

```
rho5 <- c(0.25,0.5,0.75,0.85,0.95)# 5 Rho values
N3 <- c(50,75,100)      # 3 Population Sizes
Pop_list <- list(NULL)  # Blank list
k <- 10 # using the seed form 10 onwards
count <- 1

for(i in rho5){
        for (j in N3) {
        set.seed(k)
        Rbvn(rho = i,N = j) -> Pop_list[[count]]
        k <- k+1
        count <- count+1
        }
}
Pop_list # Here is the population list
```

The above code will show the 15 desired population with population size N as 50,75 & 100 each with different ρ values 0.25,0.5,0.75,0.85,0.95. Now the corresponding population parameters of each population can be viewed in a tabular format.

```
Pop_Mat <- matrix(0,15,7)
length(Pop_list[[1]][,1])
colnames(Pop_Mat) <- c("N","Rho","Xbar","Ybar","VarX",
        "VarY","B")
for (i in 1:15) {
        Pop_Mat[i,1] <- length(Pop_list[[i]][,1])
        Pop_Mat[i,2] <- cor(Pop_list[[i]][,1],Pop_list[[i]][,2])
        Pop_Mat[i,3] <- mean(Pop_list[[i]][,1])
        Pop_Mat[i,4] <- mean(Pop_list[[i]][,2])
        Pop_Mat[i,5] <- var(Pop_list[[i]][,1])
        Pop_Mat[i,6] <- var(Pop_list[[i]][,2])
        Pop_Mat[i,7] <- lm(Pop_list[[i]][,2]~
                           Pop_list[[i]][,1])$coeff[2]
}
Pop_Mat
write.csv(x =Pop_Mat ,"Population_Matrix.csv")
```

Here the above table contains the Population parameters like population size N, correlation ρ, means of auxiliary information $\overline{X}$,means of study variables $\overline{Y}$, variances of auxiliary information ${\sigma_X}^2$, and variances of

Table 3.1: Description of Bivariate Normal Populations

Pop. No.	N	ρ	$\overline{X}$	$\overline{Y}$	σ_x^2	σ_y^2	β
1	50	0.158164	14.43263	28.34911	8.10263	19.07774	0.242694
2	75	0.159856	14.82293	29.30495	8.607466	20.47613	0.246555
3	100	0.172094	14.93852	29.85139	8.593877	18.89153	0.255155
4	50	0.544475	15.17073	29.78225	8.514576	23.34358	0.90153
5	75	0.460483	15.29324	30.43901	8.689643	21.40752	0.722762
6	100	0.420181	15.45579	30.40135	9.682386	25.24028	0.67841
7	50	0.693292	15.45986	30.74358	9.496191	19.37059	0.990177
8	75	0.745438	15.19739	30.14524	10.55411	29.46595	1.245549
9	100	0.718557	14.68231	29.38599	8.375374	24.23983	1.222429
10	50	0.834339	15.45104	30.53736	7.44743	28.16395	1.622505
11	75	0.85949	14.70542	29.75169	9.334614	24.71413	1.398509
12	100	0.875487	15.32277	30.29044	9.249777	26.24263	1.474645
13	50	0.958901	15.32563	30.62833	9.338578	27.43437	1.643543
14	75	0.942701	15.0093	30.25368	7.770396	19.82867	1.505911
15	100	0.944071	14.70546	29.34387	7.875859	22.75528	1.604712

study variables ${\sigma_Y}^2$.

3.1.2 SAMPLING

Now from this hypothetical population we draw 3 samples from each population with sample sizes 5, 10, and 15.

```
Sample_list <- list(NULL)
cnt <- 1
for(s in c(5,10,15)){
        for (i in 1:15) {
          set.seed(i)
          sample(1:nrow(Pop_list[[i]]),size = s) -> n
          Pop_list[[i]][n,] -> Sample_list[[cnt]]
          cnt <- cnt+1
        }
}
Sample_list
```

Here we get the 45 biveriate samples from the 15 biveriate normal Populations

Now we need mean, variance, β,ρ values for the drawn 45 samples individually. So we generate a table to fulfill our requirement.

```
Sample_Mat <- matrix(0,45,8)
colnames(Sample_Mat) <- c("N","n","r","xbar",
                          "ybar","varx","vary","b")
for (i in 1:45) {
        Sample_Mat[i,2] <- length(Sample_list[[i]][,1])
        Sample_Mat[i,3] <- cor(Sample_list[[i]][,1],
```

```
                                    Sample_list[[i]][,2])
        Sample_Mat[i,4] <- mean(Sample_list[[i]][,1])
        Sample_Mat[i,5] <- mean(Sample_list[[i]][,2])
        Sample_Mat[i,6] <- var(Sample_list[[i]][,1])
        Sample_Mat[i,7] <- var(Sample_list[[i]][,2])
        Sample_Mat[i,8] <- lm(Sample_list[[i]][,2]~
                                    Sample_list[[i]][,1])$coeff[2]
}
Sample_Mat[,1] <- rep(Pop_Mat[,1],3)
Sample_Mat
```

Here we get the table with mean, variance of the auxiliary and study samples, also their correlation (ρ) ,slope (β)

Table 3.2: Table for Sample

Pop. No.	N	n	r	$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	b
1	50	5	0.22781	14.52554	30.05951	0.420047	22.34069	1.661393
2	75	5	-0.02845	14.88525	26.81912	9.967511	11.47383	-0.03052
3	100	5	-0.16039	16.32238	31.04808	6.792859	50.90957	-0.43907
4	50	5	0.29035	14.05295	27.29088	4.562836	21.76926	0.634201
5	75	5	0.553175	14.08013	29.12255	2.284957	49.76438	2.581564
6	100	5	0.627183	17.02544	33.28859	12.11235	16.75219	0.737592
7	50	5	0.900826	13.95764	29.584	11.44694	17.67461	1.119364
8	75	5	0.827219	14.4624	30.94865	6.685749	46.59677	2.183852
9	100	5	0.760023	12.39791	26.12486	4.628656	16.26931	1.424898
10	50	5	0.93636	15.70562	31.84712	6.169969	13.84229	1.402508
11	75	5	0.947383	15.99996	31.53334	6.702237	17.73035	1.5409
12	100	5	0.952206	16.4465	30.29625	7.966773	26.29153	1.729807
13	50	5	0.909112	15.99347	31.65311	5.034281	17.25533	1.683102
14	75	5	0.991952	15.22319	30.08262	6.700195	20.185	1.721715
15	100	5	0.986285	16.53549	31.77468	7.945164	29.98976	1.916186
16	50	10	0.30518	15.05384	29.00851	6.240386	19.6283	0.541242
17	75	10	0.011656	14.55483	27.39325	9.768439	7.029054	0.009887
18	100	10	0.087702	15.31117	31.40863	9.290635	40.5082	0.18313
19	50	10	0.653423	14.14218	29.57939	8.715609	26.37673	1.136726
20	75	10	0.094744	14.74743	29.66686	7.15733	27.73586	0.186507
21	100	10	0.762437	14.6345	28.95523	12.8638	29.71481	1.158793
22	50	10	0.905579	14.70334	30.62467	12.76617	22.76075	1.209176
23	75	10	0.748157	16.01493	31.28073	13.85106	35.79265	1.202674
24	100	10	0.749043	13.35647	26.99757	3.932468	23.08027	1.814658
25	50	10	0.928778	16.30657	32.71015	4.326659	9.034073	1.342076
26	75	10	0.908005	16.30742	31.62779	4.720002	11.44017	1.413622
27	100	10	0.92361	16.36689	31.00307	10.76611	28.06984	1.491349
28	50	10	0.958182	16.50763	32.39362	6.619692	23.15675	1.792124
29	75	10	0.980253	15.33961	30.5951	7.878097	21.27027	1.610698
30	100	10	0.97377	15.43447	30.22676	8.177349	25.56911	1.721901

Table 3.3: Table for Sample (Contd...)

Pop. No.	N	n	r	$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	b
31	50	15	-0.01268	14.62968	28.88788	5.245407	23.28428	-0.02672
32	75	15	-0.19024	14.53224	28.68278	8.600897	10.50279	-0.21022
33	100	15	0.049806	15.50075	30.31107	10.23424	32.17119	0.088306
34	50	15	0.620308	14.35124	28.67592	7.61489	23.02816	1.07871
35	75	15	0.374437	15.47052	30.87548	12.53277	25.58882	0.535032
36	100	15	0.508517	15.71332	30.31828	12.99799	35.28137	0.837799
37	50	15	0.798631	15.29635	30.8817	10.3354	15.17241	0.967631
38	75	15	0.800112	14.95701	29.59349	21.39278	44.48297	1.153756
39	100	15	0.740046	13.29305	26.87332	5.393013	22.17762	1.500723
40	50	15	0.765019	15.70232	31.55654	3.965074	14.80435	1.478227
41	75	15	0.924501	15.40529	29.66524	9.230049	21.87074	1.423106
42	100	15	0.878051	15.56073	30.5564	9.899324	24.96666	1.394432
43	50	15	0.951007	15.6014	30.99276	7.26338	20.44849	1.595678
44	75	15	0.969258	14.68381	29.44855	9.005974	23.92918	1.57993
45	100	15	0.957887	15.49164	30.59365	6.259506	18.99521	1.668654

Now we compare our Population mean for the study variable with various estimators discussed above in chapter.2 i.e. simple mean estimator,exponential estimator,ratio estimator, square root estimator, and regression estimator. To see which is appropriate among them, we have to see the absolute difference of the estimators from the population mean $\overline{Y}$

```
D <- rbind(Pop_Mat,Pop_Mat,Pop_Mat)
Comparison_mat <- matrix(0,45,10)
colnames(Comparison_mat) <- c("Rho","N","n","Y_bar",
        "y_bar","yexp_bar","yr_bar",
        "ysq_bar","yreg_bar","Appropriate")
for (i in 1:45) {
        Comparison_mat[i,1] <- D[i,2]
        Comparison_mat[i,2] <- D[i,1]
        Comparison_mat[i,3] <- Sample_Mat[i,2]
        Comparison_mat[i,4] <- D[i,4]
        Comparison_mat[i,5] <- Sample_Mat[i,5]
        Comparison_mat[i,6] <- Sample_Mat[i,5]*exp(
                (D[i,3]-Sample_Mat[i,4])/
                (D[i,3]+Sample_Mat[i,4]))
        Comparison_mat[i,7] <- (Sample_Mat[i,5]*D[i,3]) /
                Sample_Mat[i,4]
        Comparison_mat[i,8] <- Sample_Mat[i,5]*(D[i,3] /
                Sample_Mat[i,4])^0.5
        Comparison_mat[i,9] <- Sample_Mat[i,5] + (
                (coef(lm(Sample_list[[i]][,"Y"]~
                Sample_list[[i]][,"X"]))[2])*
                        (D[i,3]-Sample_Mat[i,4]))
        Comparison_mat[i,10] <- which.min(abs(
                Comparison_mat[i,5:9]-Comparison_mat[i,4]))
}
Comparison_mat
```

And here we get the table with our desired results. Here the column *Appro* is nominal representing 1 for simple mean estimator, 2 for exponential estimator. 3 for ratio estimator, 4 for square root estimator and 5 for regression estimator.

Table 3.4: Mean Absolute Deviations of Various Estimators from $\overline{Y}$

Sample No.	ρ	N	n	$\overline{Y}$	$\overline{y}$	t_e	t_r	t_{sq}	t_{reg}	Appropriate
1	0.16	50	5	28.349	30.060	29.963	29.867	29.963	29.905	3
2	0.16	75	5	29.305	26.819	26.763	26.707	26.763	26.821	5
3	0.17	100	5	29.851	31.048	29.704	28.416	29.703	31.656	2
4	0.54	50	5	29.782	27.291	28.355	29.462	28.355	28.000	3
5	0.46	75	5	30.439	29.123	30.350	31.632	30.351	32.254	4
6	0.42	100	5	30.401	33.289	31.718	30.220	31.717	32.131	3
7	0.69	50	5	30.744	29.584	31.134	32.768	31.135	31.266	2
8	0.75	75	5	30.145	30.949	31.725	32.521	31.725	32.554	1
9	0.72	100	5	29.386	26.125	28.424	30.939	28.430	29.380	5
10	0.83	50	5	30.537	31.847	31.588	31.331	31.588	31.490	3
11	0.86	75	5	29.752	31.533	30.232	28.982	30.231	29.539	5
12	0.88	100	5	30.290	30.296	29.243	28.226	29.243	28.352	1
13	0.96	50	5	30.628	31.653	30.985	30.331	30.985	30.529	5
14	0.94	75	5	30.254	30.083	29.871	29.660	29.871	29.714	1
15	0.94	100	5	29.344	31.775	29.967	28.258	29.965	28.268	4
16	0.16	50	10	28.349	29.009	28.404	27.811	28.404	28.672	4
17	0.16	75	10	29.305	27.393	27.644	27.898	27.644	27.396	3
18	0.17	100	10	29.851	31.409	31.024	30.644	31.024	31.340	3
19	0.54	50	10	29.782	29.579	30.636	31.731	30.636	30.749	1
20	0.46	75	10	30.439	29.667	30.211	30.765	30.211	29.769	4
21	0.42	100	10	30.401	28.955	29.756	30.580	29.757	29.907	3
22	0.69	50	10	30.744	30.625	31.402	32.200	31.403	31.539	1
23	0.75	75	10	30.145	31.281	30.472	29.684	30.472	30.297	5
24	0.72	100	10	29.386	26.998	28.305	29.678	28.306	29.404	5
25	0.83	50	10	30.537	32.710	31.841	30.994	31.841	31.562	3
26	0.86	75	10	29.752	31.628	30.035	28.521	30.034	29.363	4
27	0.88	100	10	30.290	31.003	29.998	29.025	29.998	29.446	2
28	0.96	50	10	30.628	32.394	31.213	30.074	31.212	30.275	5
29	0.94	75	10	30.254	30.595	30.264	29.936	30.264	30.063	4
30	0.94	100	10	29.344	30.227	29.504	28.799	29.504	28.971	4
31	0.16	50	15	28.349	28.888	28.693	28.499	28.693	28.893	3
32	0.16	75	15	29.305	28.683	28.968	29.257	28.968	28.622	3
33	0.17	100	15	29.851	30.311	29.756	29.212	29.756	30.261	2
34	0.54	50	15	29.782	28.676	29.483	30.313	29.483	29.560	5
35	0.46	75	15	30.439	30.875	30.698	30.522	30.698	30.781	3
36	0.42	100	15	30.401	30.318	30.069	29.821	30.069	30.103	1
37	0.69	50	15	30.744	30.882	31.046	31.212	31.046	31.040	1
38	0.75	75	15	30.145	29.593	29.830	30.069	29.830	29.871	3
39	0.72	100	15	29.386	26.873	28.242	29.682	28.243	28.958	3
40	0.83	50	15	30.537	31.557	31.303	31.052	31.303	31.185	3
41	0.86	75	15	29.752	29.665	28.984	28.318	28.984	28.669	1
42	0.88	100	15	30.290	30.556	30.322	30.089	30.322	30.225	4
43	0.96	50	15	30.628	30.993	30.718	30.445	30.718	30.553	5
44	0.94	75	15	30.254	29.449	29.773	30.101	29.773	29.963	3
45	0.94	100	15	29.344	30.594	29.807	29.041	29.807	29.282	5

In this whole experiment we can see the score of each estimator, i.e. the no of times experiment is considered as Appropriate for the given different population sizes, sample sizes and correlations.

```
tabulate(Comparison_mat[,"Appropriate"])
```

- And find that out of 45 cases
- Simple mean estimator is applicable for 8 cases
- Exponential estimator is applicable for only 4 cases
- Ratio estimator is applicable for 15 cases
- Square-root estimator is applicable for 8 cases
- Regression estimator is applicable for 10 cases

Now to see the overall performance of the estimator for the 15 populations, we need to create a table

```
k<-table(round(Comparison_mat[,"Rho"],3),
```

```
          Comparison_mat[,"Appropriate"])
k
```

Table 3.5: Comparison

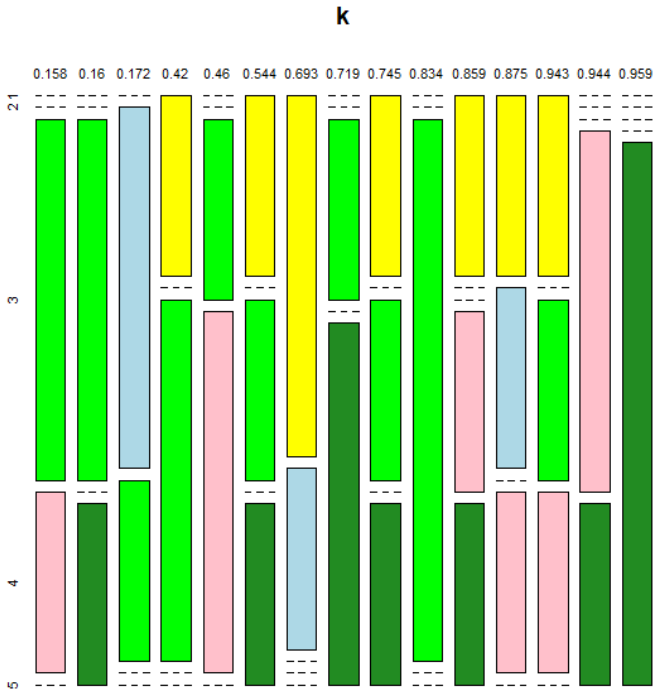
N	Correlations	Simple Mean	Exponential	Ratio .	Square root	Regression
50	0.158	0	0	2	1	0
100	0.16	0	0	2	0	1
150	0.172	0	2	1	0	0
50	0.42	1	0	2	0	0
100	0.46	0	0	1	2	0
150	0.544	1	0	1	0	1
50	0.693	2	1	0	0	0
100	0.719	0	0	1	0	2
150	0.745	1	0	1	0	1
50	0.834	0	0	3	0	0
100	0.859	1	0	0	1	1
150	0.875	1	1	0	1	0
50	0.943	1	0	1	1	0
100	0.944	0	0	0	2	1
150	0.959	0	0	0	0	3

Now to represent the above table, we can do a mosaic plot.

```
mosaicplot(k,color = c("yellow","lightblue",
        "green","pink","forestgreen"))
```

In this plot

- **yellow** color represents the simple mean estimator
- **light blue** color represents the exponential estimator
- **green** color represents the ratio estimator
- **pink** color represents the square-root estimator
- **yellow** color represents the regression estimator



3.2 TEST USING NATURAL DATA

A pilot scheme for studying milk yield, breeds, feeding and management practice of cattle and buffaloes were conducted in eastern Uttar Pradesh during 1957-1958. The given data below presents total number of milch cows in the 19 selected villages of dry region as enumerated in the rainy season 1957 and as given by Census, 1956.

```
Si.No.Villages <- 1:19
RainySeason_1957 <- c(35,38,71,4,63,4,14,7,66,44,8,229,
        27,30,29,29,97,30,40)
Census_1956 <- c(47,46,253,19,121,4,5,7,50,162,9,256,
        74,28,41,27,25,40,66)
Data <- data.frame(Si.No.Villages,RainySeason_1957,
        Census_1956);Data
write.csv(Data,"Naturaldata.csv")
```

Table 3.6: Add caption

	Si.No. Villages	Rainy Season_1957	Census_1956
1	1	35	47
2	2	38	46
3	3	71	253
4	4	4	19
5	5	63	121
6	6	4	4
7	7	14	5
8	8	7	7
9	9	66	50
10	10	44	162
11	11	8	9
12	12	229	256
13	13	27	74
14	14	30	28
15	15	29	41
16	16	29	27
17	17	97	25
18	18	30	40
19	19	40	66

Now lets suppose the given data of 19 villages is Population for us and out of which we are further going to draw all possible samples with sample size 5 representing the population parameter.

```
Xi = Census_1956;Yi = RainySeason_1957;  Xi;Yi
samples_X <- t(combn(x = Xi,m = 5))
samples_Y <- t(combn(x = Yi,m = 5))
```

$Sample\_X$ represents all 11628 possible samples from population X, similarly $Sample\_Y$ represents all 11628 possible samples from population Y. Here we get 11628 of samples. Now lets see how many of these 11628 samples are highly correlated.

```
s.cor <- c()
for(i in 1:nrow(samples_X)){
s.cor[i] <- cor(samples_X[i,],samples_Y[i,])
}
coR <- as.data.frame(table(round(s.cor,1)))
colnames(coR) <- c("sample_correlations","Frequency");coR
prop.table(table(s.cor >= 0.7)
write.csv(coR,"cor.csv"))
```

Here we can see that the probability of selecting a bivariate sample with correlation coefficient ρ grater than 0.7 is 0.678 i.e. most of the samples are highly correlated.

Table 3.7: Sample correlation and their frequencies

sample_correlations	Frequency
-0.9	4
-0.8	2
-0.7	19
-0.6	26
-0.5	33
-0.4	46
-0.3	63
-0.2	82
-0.1	145
0	277
0.1	359
0.2	404
0.3	468
0.4	623
0.5	368
0.6	417
0.7	1154
0.8	1703
0.9	3023
1	2412

Then we use our estimators to see which estimator is performing best in this case of high correlation values. Here we have simple mean, product, ratio, exponential, square-root and regression estimators.

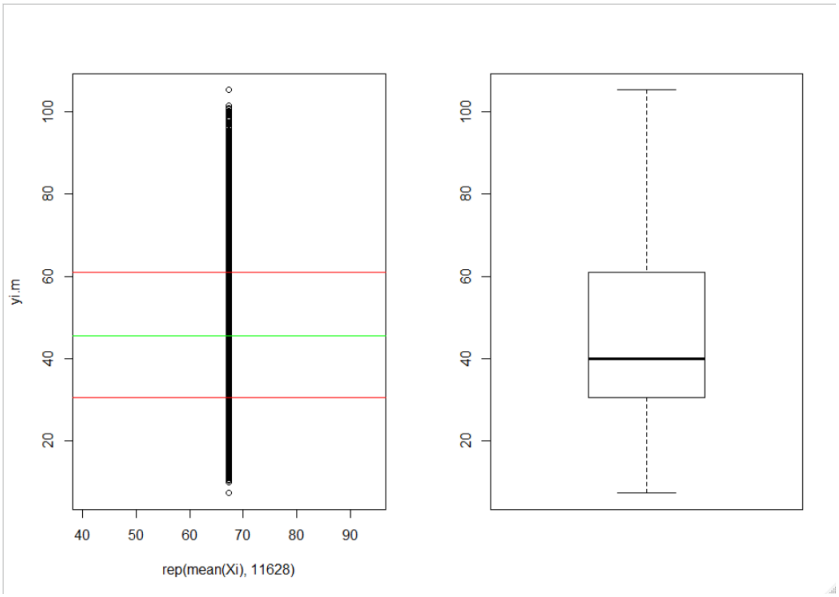
```
yi.s <- samples_Y
xi.s <- samples_X
xi.m <- rowMeans(xi.s)
beta <- c()
for (i in 1:nrow(yi.s)){
beta[i] <- coef(lm(yi.s[i,]~xi.s[i,]))[2]
}

yi.m <- rowMeans(yi.s);IQR(yi.m)#(Simple mean estimator)
yi.p <- (yi.m*xi.m)/mean(Xi);IQR(yi.p)#(product estimator)
yi.r <- (yi.m*mean(Xi))/xi.m ;IQR(yi.r)#(ratio estimator)
yi.e <- yi.m*exp((mean(Xi)-xi.m)/(mean(Xi)+xi.m));
IQR(yi.e) # ( For exponential estimator)
yi.sq <- yi.m*(mean(Xi)/xi.m)^.5  ;IQR(yi.sq) # --
yi.reg <- yi.m + (beta*(Pop_Xbar-xi.m));IQR(yi.reg)
#---------------------------------------
```

Here we are having all estimated values of the corresponding samples. Now its time to analyze it. Lets analyze the estimator using various plots and see distribution of the data for all samples. In the plots analyzed

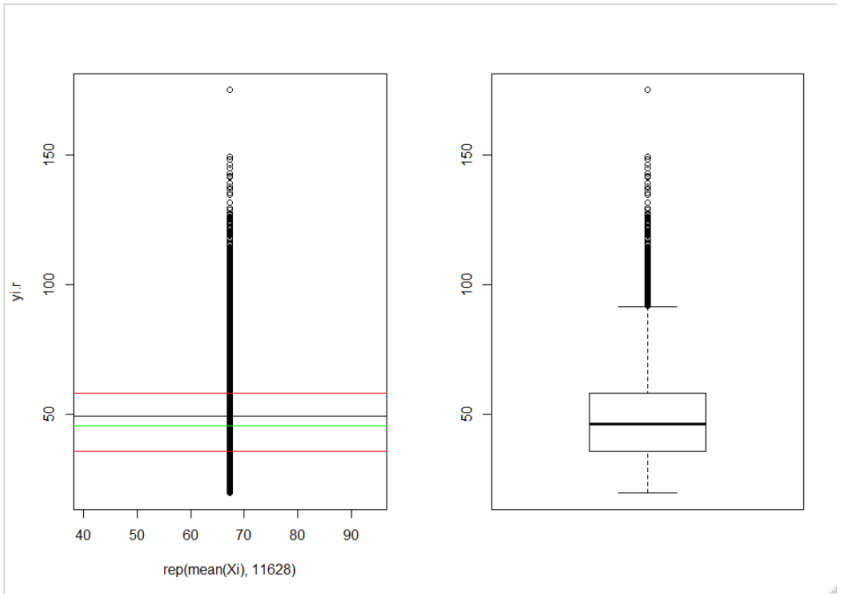
below having $3rd$ and $1st$ *quartile* in red lines, and green line representing the mean value of the estimated y values, black line closer to green line represents Population mean $\overline{Y}$. The case where there is no black line closer to green line, there the green line is above the black line.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.m)
abline(a=mean(yi.m),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.m)[2],0,col="red")
abline(a=quantile(yi.m)[4],0,col="red")
boxplot(yi.m)
```



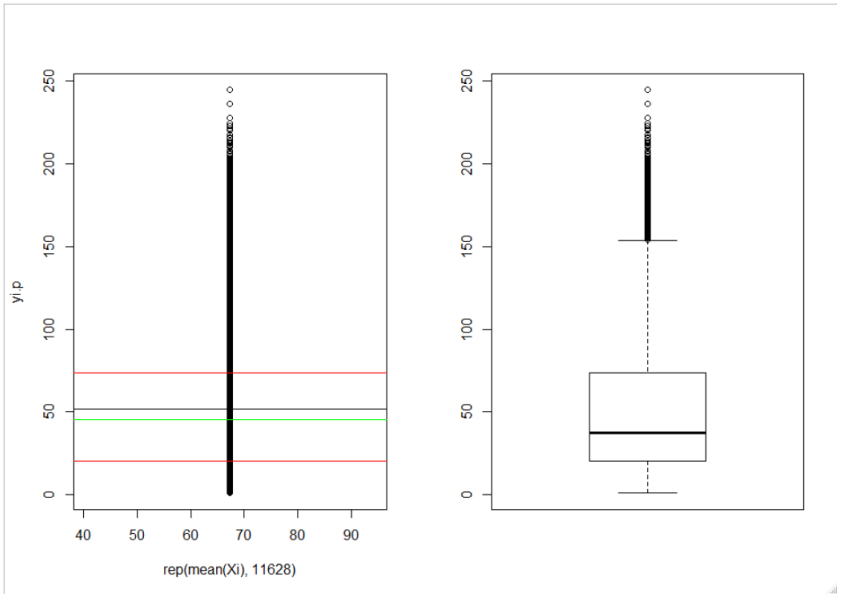
Here in the plot we can see that the Inter-Quartile-Range for simple mean estimator is 30.4 and the estimated population mean for all possible samples ranges from 7 to 105, and 50 percent of the estimated values are in between 30 to 61.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.r)
abline(a=mean(yi.r),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.r)[2],0,col="red")
abline(a=quantile(yi.r)[4],0,col="red")
boxplot(yi.r)
IQR(yi.r)
```



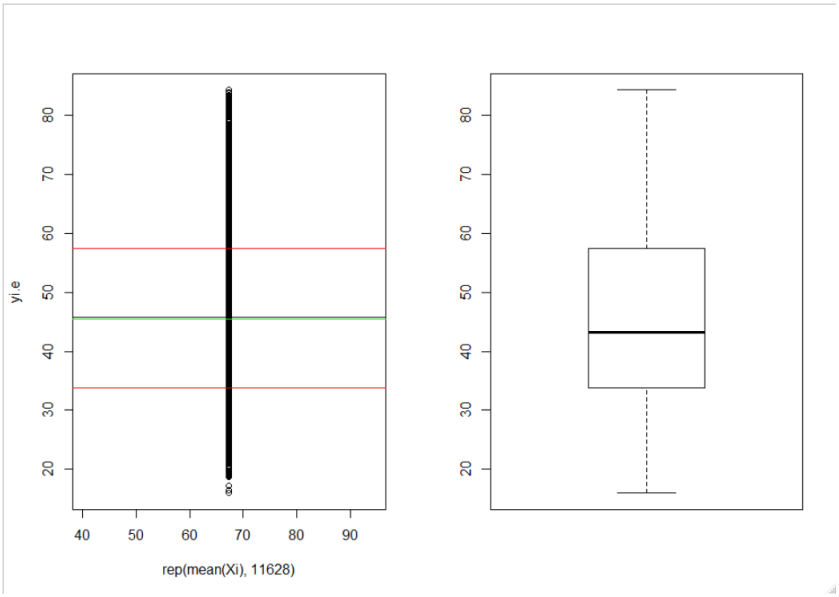
Here in the plot we can see that the Inter-Quartile-Range for ratio estimator is 22.39 and the estimated population mean for all possible samples ranges from 19 to 175, and 50 percent of the estimated values are in between 35 to 58.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.p)
abline(a=mean(yi.p),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.p)[2],0,col="red")
abline(a=quantile(yi.p)[4],0,col="red")
boxplot(yi.p)
IQR(yi.p)
```



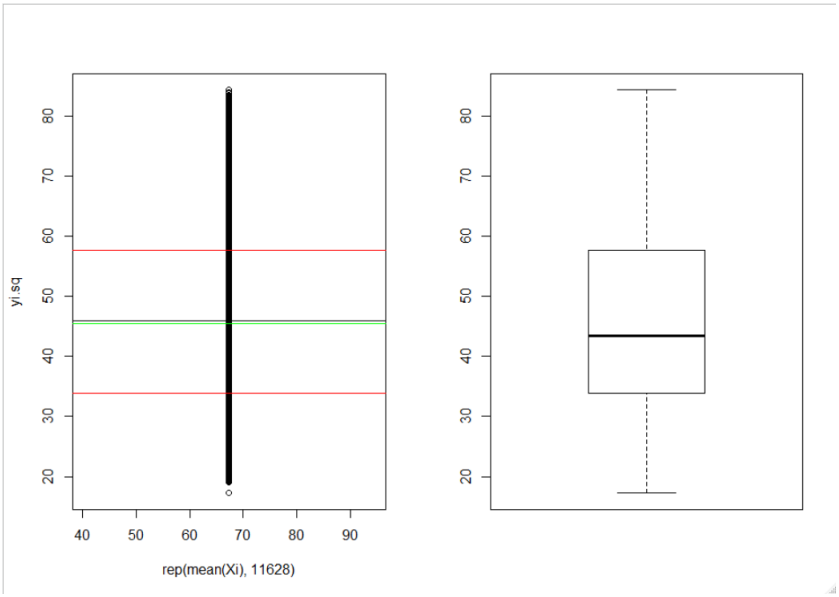
Here in the plot we can see that the Inter-Quartile-Range for product estimator is 53.34 and the estimated population mean for all possible samples ranges from 0 to 244, and 50 percent of the estimated values are in between 20 to 73.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.e)
abline(a=mean(yi.e),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.e)[2],0,col="red")
abline(a=quantile(yi.e)[4],0,col="red")
boxplot(yi.e)
IQR(yi.e)
```



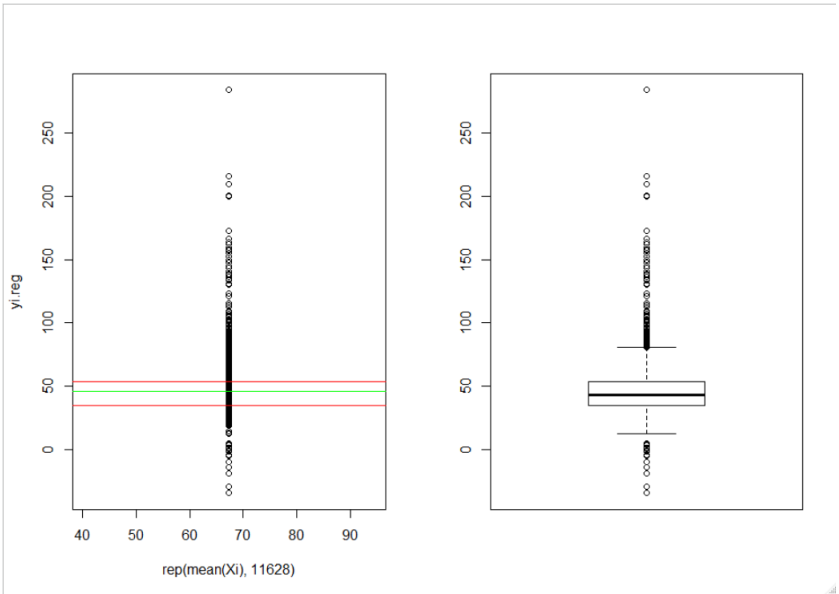
Here in the plot we can see that the Inter-Quartile-Range for exponential estimator is 23.62 and the estimated population mean for all possible samples ranges from 15 to 84, and 50 percent of the estimated values are in between 33 to 57.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.sq)
abline(a=mean(yi.sq),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.sq)[2],0,col="red")
abline(a=quantile(yi.sq)[4],0,col="red")
boxplot(yi.sq)
IQR(yi.sq)
```



Here in the plot we can see that the Inter-Quartile-Range for square-root estimator is 23.67 and the estimated population mean for all possible samples ranges from 17 to 84, and 50 percent of the estimated values are in between 34 to 57.

```
#---------------------------------------
par(mfrow=c(1,2))
plot(x = rep(mean(Xi),11628),yi.reg)
abline(a=mean(yi.reg),0)
abline(a=mean(Yi),0,col="green")
abline(a=quantile(yi.reg)[2],0,col="red")
abline(a=quantile(yi.reg)[4],0,col="red")
boxplot(yi.reg)
IQR(yi.reg)
```



Here in the plot we can see that the Inter-Quartile-Range for regression estimator is 18.34 and the estimated population mean for all possible samples ranges from -34 to 284, and 50 percent of the estimated values are in between 35 to 53.
Now lets have a look on the summary table for all the estimators including Inter-Quartile-Range.

```
estimates <- data.frame(yi.m,yi.r,yi.p,yi.e,yi.sq,yi.reg)
summary_est <- sapply(estimates, summary)
IQR_est <- sapply(estimates,IQR)
range_est <- diff(sapply(estimates, range))
estimate.tables <-rbind(summary_est,IQR_est,range_est)
write.csv(estimate.tables,"summaryestimator.csv")
```

Table 3.8: Summary of Estimators

	yi.m	yi.r	yi.p	yi.e	yi.sq	yi.reg
Min.	7.40	19.68	0.97	15.97	17.27	-34.67
1st Qu.	30.60	35.75	20.32	33.76	33.93	34.94
Median	40.00	46.29	37.57	43.16	43.37	43.36
Mean	45.53	49.28	51.74	45.76	45.96	45.78
3rd Qu.	61.00	58.15	73.66	57.38	57.60	53.29
Max.	105.20	175.16	244.49	84.27	84.26	284.16
IQR	30.40	22.40	53.35	23.63	23.67	18.35
Range	97.80	155.48	243.52	68.30	67.00	318.83

We can see that, in case of *ratio* and *regression* and *product* estimator the range is very high as compare to others, but in case of *product* estimator the IQR is very high as compare to all others, where *ratio* and *regression* estimator having the least IQR. The IQR is taking an important role for choosing the appropriate estimator as it explains the 50 percent of the total data which lies between the $3rd$ and the $1st$ *quartile* .

Chapter 4

CONCLUSION

We have analyzed the performance of several estimators on the basis of hypothetical populations, specifically bivariate normal population by specifying the parameters μ_x, μ_y, σ_x, σ_y and ρ. We first generate populations of different sizes $N = 50,\ 75$ and 100 by changing the values of $\rho = 0.25, 0.5, 0.75, 0.85, 0.95$. So, we generate 15 such populations and from this we generate samples of size $n = 5, 10$ and 15. On th basis of all possible samples from these populations, we have observed the following points:

- For the samples having high degree of positive correlation i.e., ($\rho > 0.7$) between the Auxiliary and Study values, its better to use Ratio and regression estimator,
- Simple mean estimator is to be preferred where ρ takes the value closer to 0.5.
- For very low correlation it is quite difficult to decide which estimator will perform better.

Again, we consider a Natural population as described in Section 3.2, we have seen that the probability of getting a sample with high degree of correlation ($\rho > 0.7$) from the given population is 0.67. In these samples, it is preferable to go for the regression and ratio estimator for estimating $\overline{Y}$.
Looking at the summary table of all estimators we have seen that:

- Regression estimator is having the least Inter Quartile Range i.e. 18.34 means 50 percent of all estimated values are indicating to the cluster between the range 34.9 to 53.2 which is closer to our population mean ie 45.5 as compare to other estimators.

- Next to regression estimator, we are having the ratio estimator, which is having the IQR value 22.39. Although regression estimator is having the least IQR, it also having some negative estimated values. Hence its good to go for the Ratio Estimator instead of regression estimator.

- Next to ratio estimator, we are having the exponential and square-root estimator, both having IQR 23.63 and 23.67 respectively, there is negligible difference in IQR between the exponential and square-root estimator as they both having equal amount of bias.

- Finally we have our usual simple mean estimator, having an IQR 30.4. Here 50 percent of all estimated values are indicating to the cluster between the range 30.60 to 61. As the width of the interval between 1st and 3rd quartile is more as compare to the above estimators, so it is less preferable for estimating population mean $\overline{Y}$

- Looking at the product estimator, having IQR 53.34 i.e. highest IQR among all, we can conclude that product estimator are useless for the bivariate samples having the high positive correlation value.

Bibliography

1. Bansal, A. (2017). *Survey Sampling*, Narosa Publishing House, Ch. 10, pp. 10.20-10.23.

2. Deming, W.E. (1950). *Some Theory of Sampling*, Dover Publications, NY, Ch - 1, pp. 02.

3. Swain, A.K. P.C. (2003). *Finite Population Sampling: Theory and Methods*, South Asian Publishers, Ch. 9, pp. 288-322.

4. Sukhatme, P.V., Sukhatme, B.V., Sukhatme, S. and Asok, C. (1984). *Sampling Theory of Surveys with Applications.* Iowa State University Press, Ames and Indian Society of Agricultural Statistics, New Delhi.

5. Cochran, W.G. (1963). *Sampling Techniques.* Second edition, Wiley Publications in Statistics, John Wiley & Sons.