**Topic – 4**

IRIS

Name :- Akash Mandal

Roll No. :- 16MA20007

Course :- CS40003

Problem Statement –

a) Consider the 150 observations as very close to population data. Find the population mean.

b) Assume a sample of size 50 chosen at random, find the population variance.

c) Compare the sample variance with that of population variance?

Reference –

IRIS.xlsx (Included in the project folder for reference)

Observation of data –

The database IRIS.csv reveals the following basic structure-

1. 150 observations of 6 variables.
2. It is a database about flowers of iris family and measurements of its parts.
3. There are 4 numerical variables, SepalLengthCm, SepalWidthCm, PetalLengthCm, PetalWidthCm. We carry out the problem statement for each of these variables.

Population Means and Variances–

We calculate the population means of the four variables using the mean() function of R. We have –

1. Sepal Length – 5.843333cm
2. Sepal Width – 3.054cm
3. Petal Length – 3.758667cm
4. Petal Width – 1.198667cm

Similarly for variances, we have –

1. Sepal Length – 0.6856935
2. Sepal Width – 0.188004
3. Petal Length – 3.113179
4. Petal Width – 0.5824143

Code Snippet

> pop\_sep\_len\_mean <- mean(IRIS$SepalLengthCm)

> pop\_sep\_len\_var <- var(IRIS$SepalLengthCm)

> pop\_sep\_wid\_mean <- mean(IRIS$SepalWidthCm)

> pop\_sep\_wid\_var <- var(IRIS$SepalWidthCm)

> pop\_pet\_len\_mean <- mean(IRIS$PetalLengthCm)

> pop\_pet\_len\_var <- var(IRIS$PetalLengthCm)

> pop\_pet\_wid\_mean <- mean(IRIS$PetalWidthCm)

> pop\_pet\_wid\_var <- var(IRIS$PetalWidthCm)

> print(pop\_sep\_len\_mean)

[1] 5.843333

> print(pop\_sep\_len\_var)

[1] 0.6856935

> print(pop\_sep\_wid\_mean)

[1] 3.054

> print(pop\_sep\_wid\_var)

[1] 0.188004

> print(pop\_pet\_len\_mean)

[1] 3.758667

> print(pop\_pet\_len\_var)

[1] 3.113179

> print(pop\_pet\_wid\_mean)

[1] 1.198667

> print(pop\_pet\_wid\_var)

[1] 0.5824143

Comparison with sample of 50 –

A random sample of 50 is chosen from the data, and the means and variances are recalculated for the sample.

The means turn out to be –

1. Sepal Length – 5.896cm
2. Sepal Width – 3.044cm
3. Petal Length – 3.822cm
4. Petal Width – 1.234cm

And the variances are –

1. Sepal Length – 0.6771265
2. Sepal Width – 0.1808816
3. Petal Length – 3.191139
4. Petal Width – 0.5847388

Code Snippet

> getSample <- function(tab, size){

+ return (tab[sample(1:nrow(tab), size, replace = FALSE),])

+ }

> samp <- getSample(IRIS, 50

> samp\_sep\_len\_mean <- mean(samp$SepalLengthCm)

> samp\_sep\_len\_var <- var(samp$SepalLengthCm)

> samp\_sep\_wid\_mean <- mean(samp$SepalWidthCm)

> samp\_sep\_wid\_var <- var(samp$SepalWidthCm)

> samp\_pet\_len\_mean <- mean(samp$PetalLengthCm)

> samp\_pet\_len\_var <- var(samp$PetalLengthCm)

> samp\_pet\_wid\_mean <- mean(samp$PetalWidthCm)

> samp\_pet\_wid\_var <- var(samp$PetalWidthCm)

> print(samp\_sep\_len\_mean)

[1] 5.988

> print(samp\_sep\_len\_var)

[1] 0.8969959

> print(samp\_sep\_wid\_mean)

[1] 3.082

> print(samp\_sep\_wid\_var)

[1] 0.1696694

> print(samp\_pet\_len\_mean)

[1] 3.84

> print(samp\_pet\_len\_var)

[1] 3.497959

> print(samp\_pet\_wid\_mean)

[1] 1.196

> print(samp\_pet\_wid\_var)

[1] 0.5767184

Different-sized samples, Central Limit Theorem

To verify the central limit theorem, we take 500 samples of each of sample sizes 1, 2, 10, 20, 50, 80, 100, 140, and 149. Using them we plot the density distribution of the sample mean, and compare it to a normal distribution.  
We plot each of the four variables, and the red line depicts the population mean, to which the distribution converges.

Code Snippet

> getSampDist <- function(size){

+ dist.sep.len=c()

+ dist.sep.wid=c()

+ dist.pet.len=c()

+ dist.pet.wid=c()

+ for(i in 1:500)

+ {

+ samp <- getSample(IRIS, size)

+ mean.sep.len <- mean(samp$SepalLengthCm)

+ mean.sep.wid <- mean(samp$SepalWidthCm)

+ mean.pet.len <- mean(samp$PetalLengthCm)

+ mean.pet.wid <- mean(samp$PetalWidthCm)

+ dist.sep.len <- c(dist.sep.len, mean.sep.len)

+ dist.sep.wid <- c(dist.sep.wid, mean.sep.wid)

+ dist.pet.len <- c(dist.pet.len, mean.pet.len)

+ dist.pet.wid <- c(dist.pet.wid, mean.pet.wid)

+ }

+ par(mfrow = c(2, 2))

+ hist(dist.sep.len, breaks = 15, xlim = c(min(IRIS$SepalLengthCm), max(IRIS$SepalLengthCm)), freq = FALSE, xlab = "Sample mean of Sepal Length(cm)", main = paste("Distribution Sepal Length-Sample size = ", size))

+ curve(dnorm(x,mean(dist.sep.len),sd(dist.sep.len)),lwd=2,add=TRUE)

+ abline(v = pop\_sep\_len\_mean, col = "red", lwd = 2)

+ hist(dist.sep.wid, breaks = 15, xlim = c(min(IRIS$SepalWidthCm), max(IRIS$SepalWidthCm)), freq = FALSE, xlab = "Sample mean of Sepal Width(cm)", main = paste("Distribution Sepal Width-Sample size = ", size))

+ curve(dnorm(x,mean(dist.sep.wid),sd(dist.sep.wid)),lwd=2,add=TRUE)

+ abline(v = pop\_sep\_wid\_mean, col = "red", lwd = 2)

+ hist(dist.pet.len, breaks = 15, xlim = c(min(IRIS$PetalLengthCm), max(IRIS$PetalLengthCm)), freq = FALSE, xlab = "Sample mean of Petal Length(cm)", main = paste("Distribution Petal Legth-Sample size = ", size))

+ curve(dnorm(x,mean(dist.pet.len),sd(dist.pet.len)),lwd=2,add=TRUE)

+ abline(v = pop\_pet\_len\_mean, col = "red", lwd = 2)

+ hist(dist.pet.wid, breaks = 15, xlim = c(min(IRIS$PetalWidthCm), max(IRIS$PetalWidthCm)), freq = FALSE, xlab = "Sample mean of Petal Width(cm)", main = paste("Distribution Petal Width-Sample size = ", size))

+ curve(dnorm(x,mean(dist.pet.wid),sd(dist.pet.wid)),lwd=2,add=TRUE)

+ abline(v = pop\_pet\_wid\_mean, col = "red", lwd = 2)

+ }

Code Snippet

> getSampDist(1)

> getSampDist(2)

> getSampDist(10)

> getSampDist(20)

> getSampDist(50)

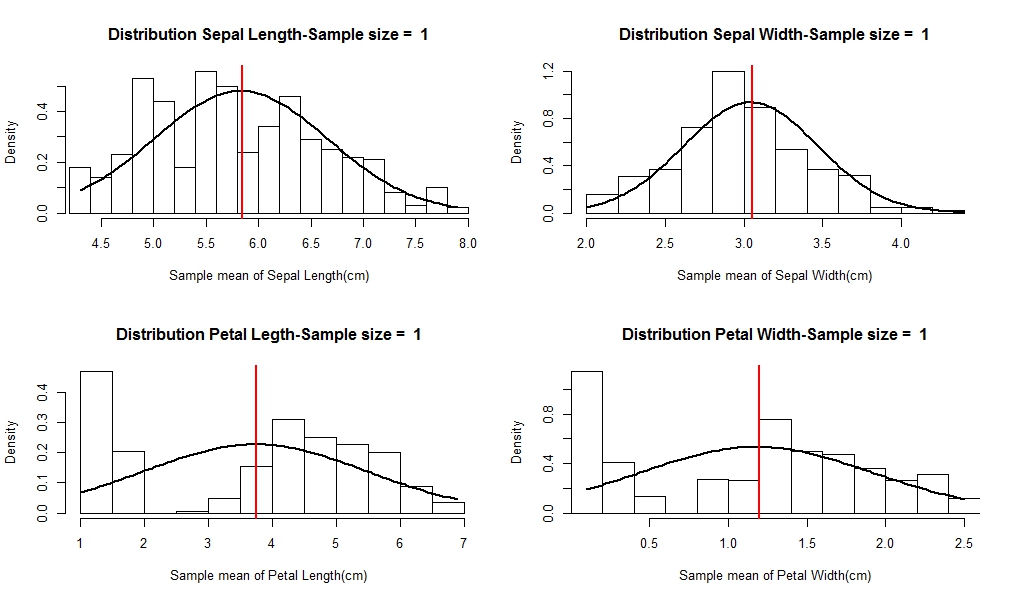
> getSampDist(80)

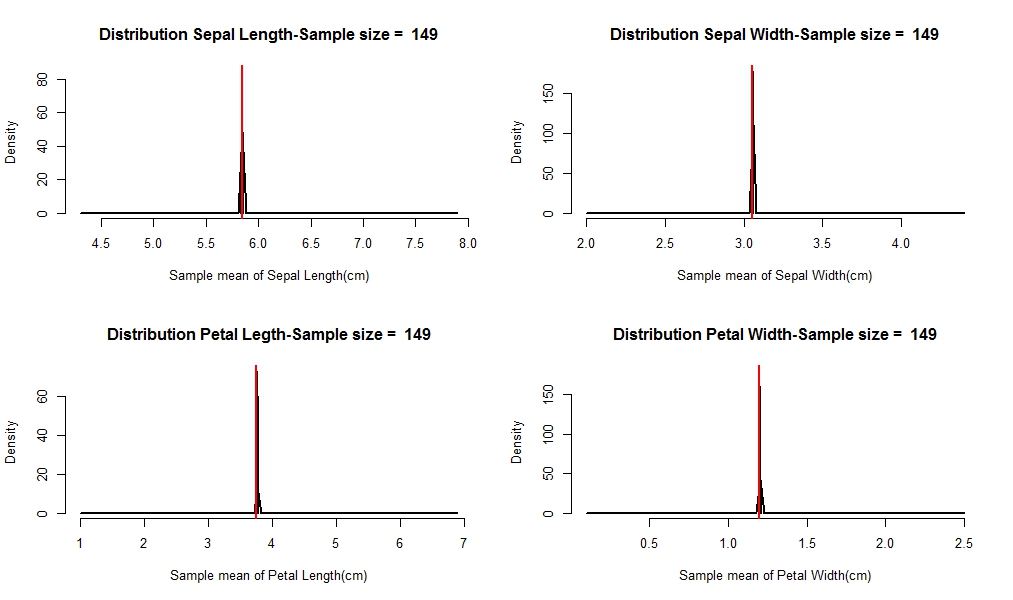
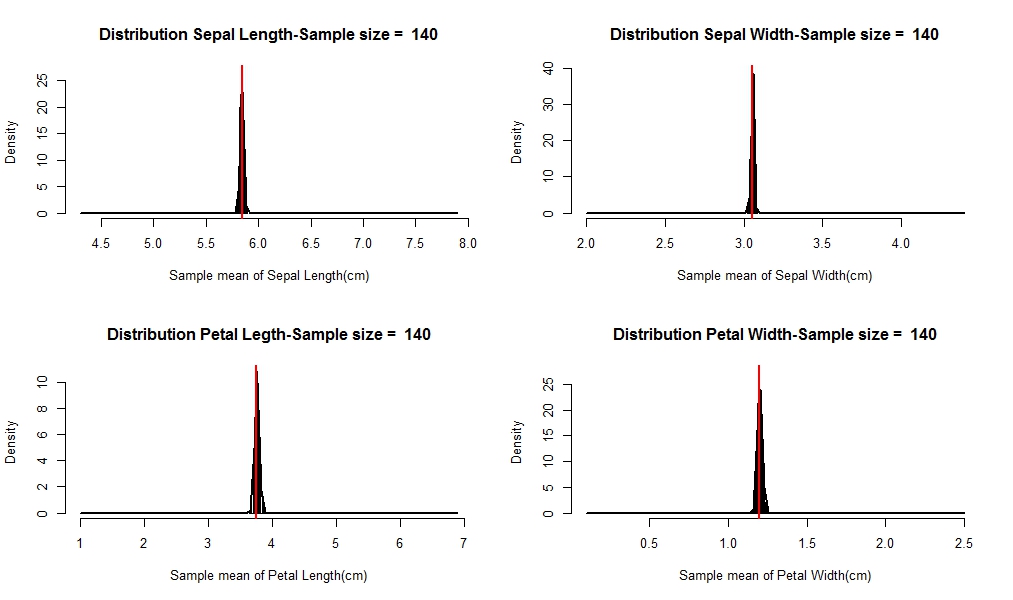
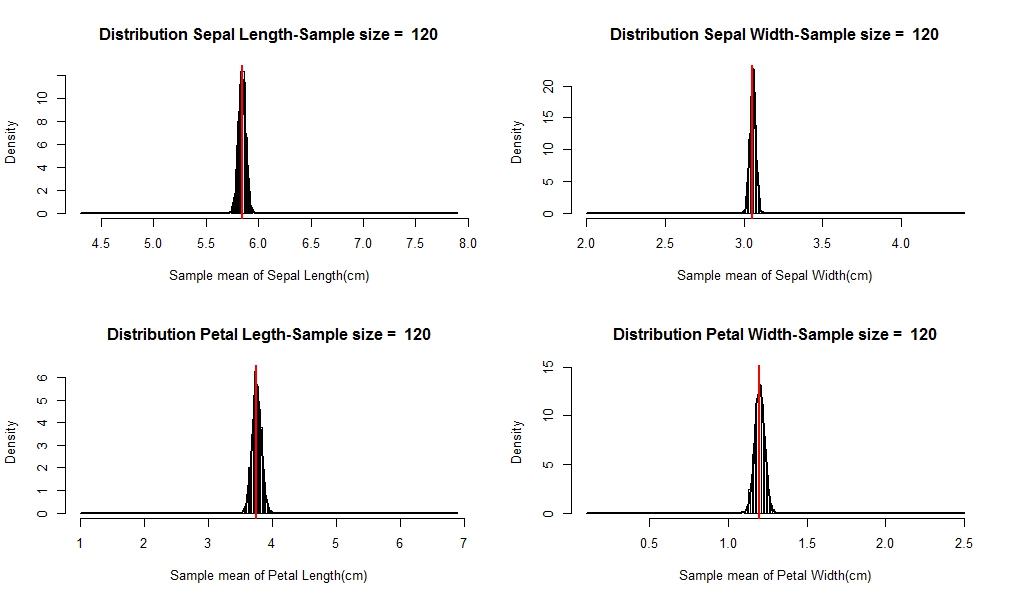
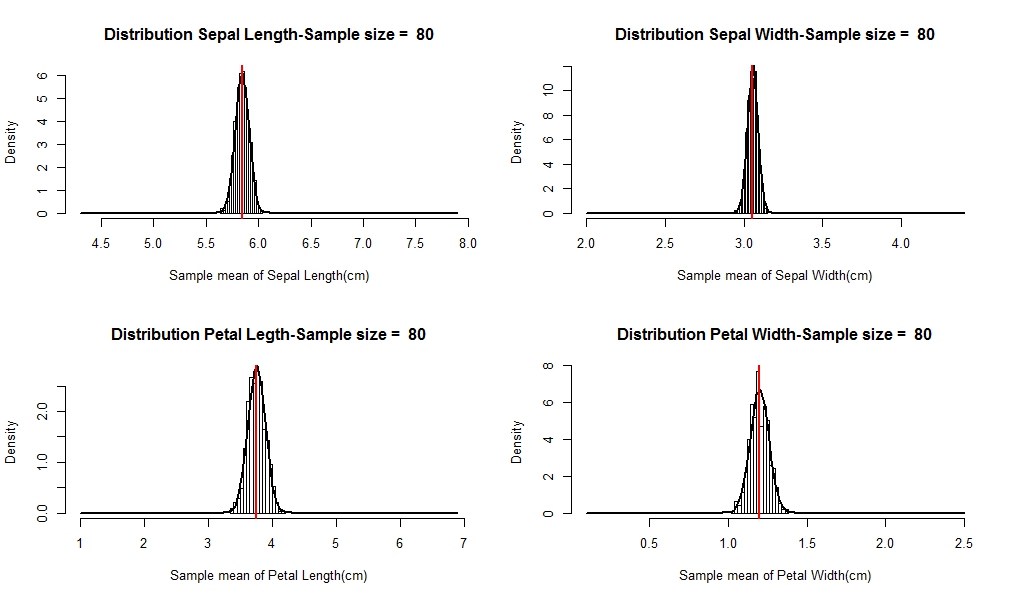
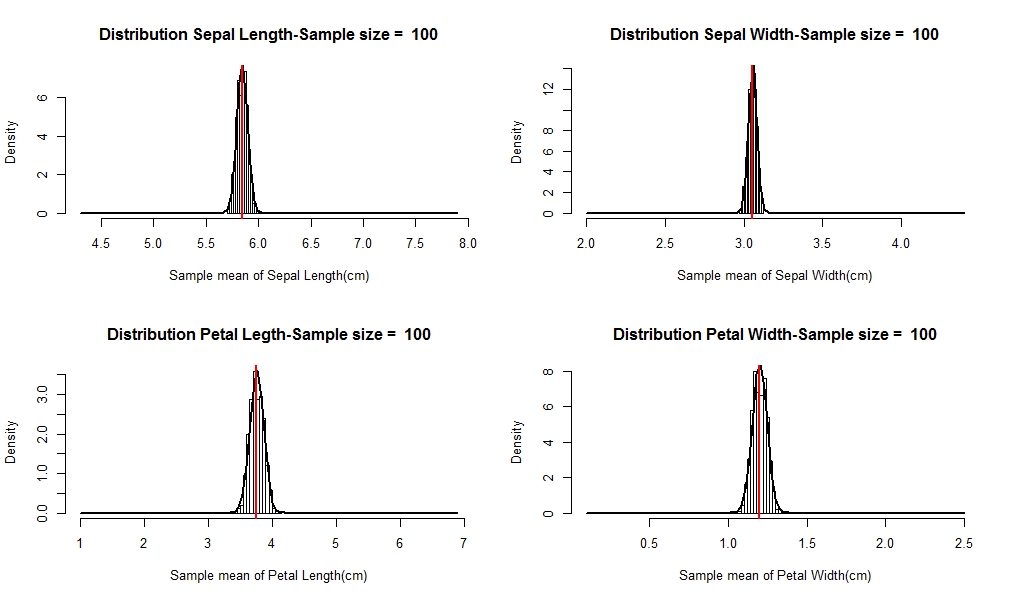
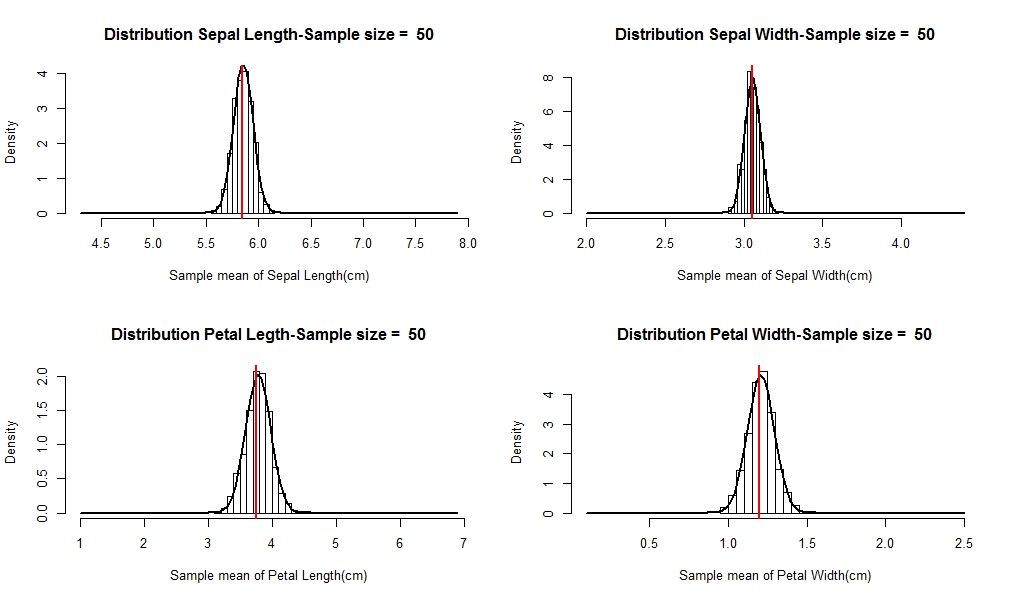
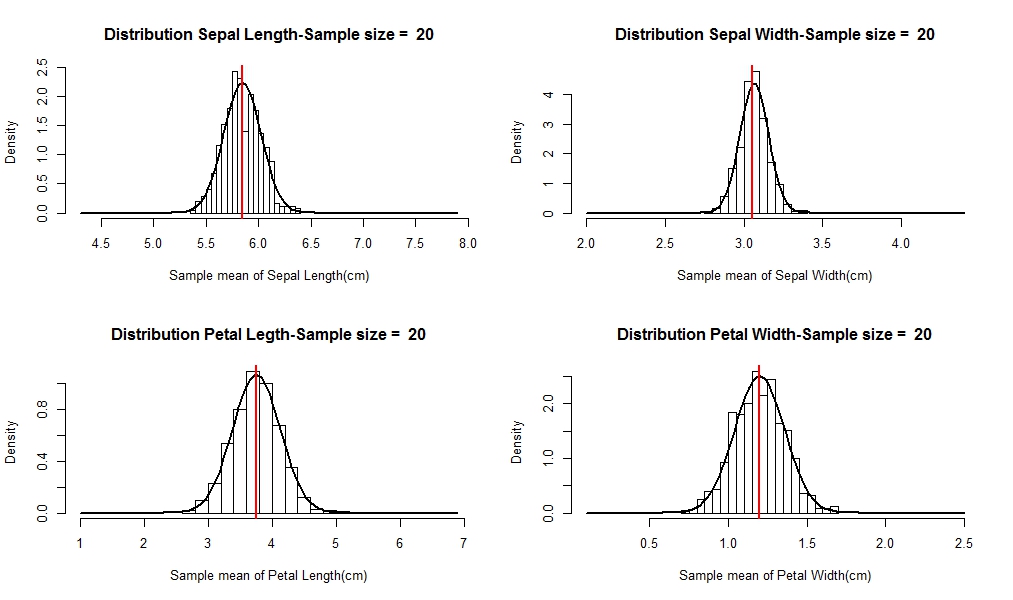
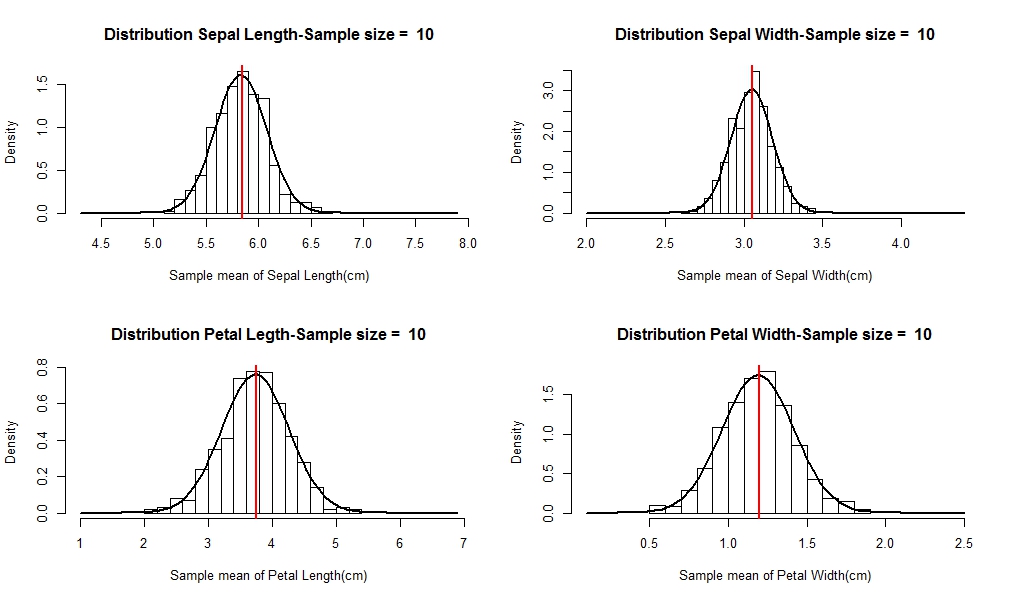
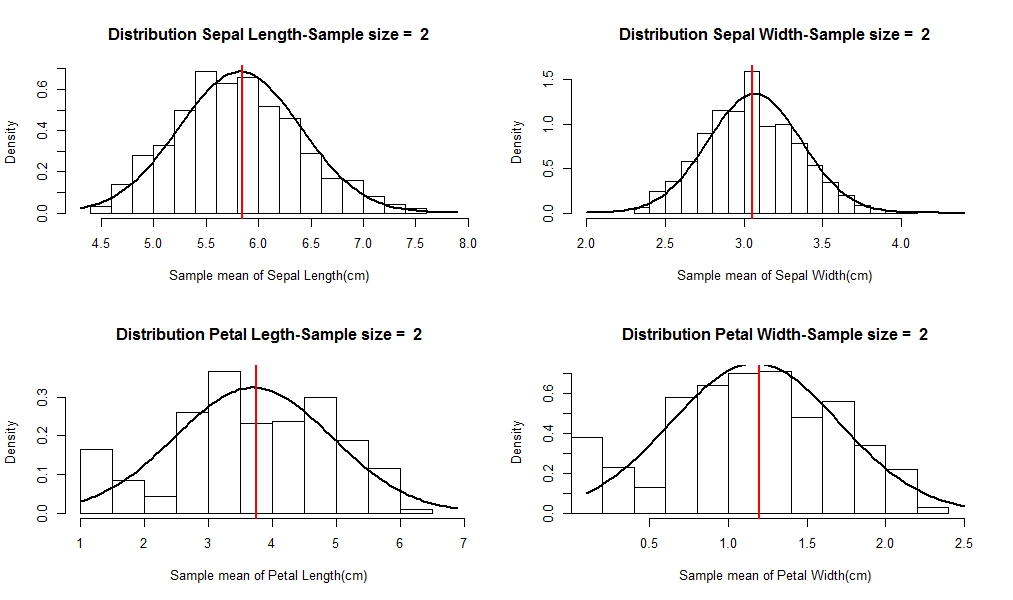
> getSampDist(100)

> getSampDist(120)

> getSampDist(140)

> getSampDist(149)





Conclusion-

We compare our sample mean and variance with that of the population’s, and for different samples verify the central limit theorem, we find a close approximation of the sampling distribution to the normal distribution as sample size increases, and it converges to the population mean, with decreasing variance.