Visualisation: Assignment 1

Students Name: Sucheta Jhunjhunwala Roll Number: MDS202151

Dead Line : 23 Nov 2021

Instruction:

- Work on the `Assignment 1.Rmd' file. Compile the file as pdf. Submit only the pdf file in moodle.
- If you want to do the work on Google colab, then please share the Colab link on the moodle.
- There are four problems.
- **Total 10 points**

Problem 1 (3 points)

Problem Statement: Write an R function which will test Central Limit Theorem.

- Assume the underlying population distribution follow Poisson distribution with rate parameter λ
- We want to estimate the unknown λ with the sample mean

$$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- The exact sampling distribution of $\hat{\lambda}$ is unknown
- But CLT tells us that as sample size n increases the sampling distribution of $\hat{\lambda}$ can be approximated by Gaussian distribution.

Input in the function: * n: sample size * λ : rate parameter * N: simulation size

Output from the function:

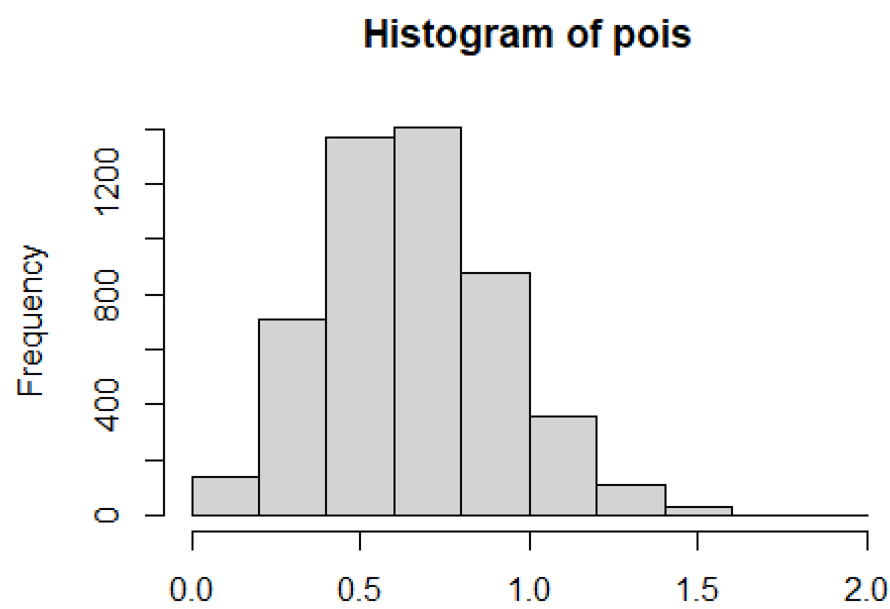
- Histogram of the sampling distribution
- QQ-plot

Test cases: * case 1 a: $\lambda = 0.7$, n=10, N=5000 * case 1 b: $\lambda = 0.7$, n=30, N=5000 * case 1 c: $\lambda = 0.7$, n=100, N=5000 * case 1 c: $\lambda = 0.7$, n=300, N=5000

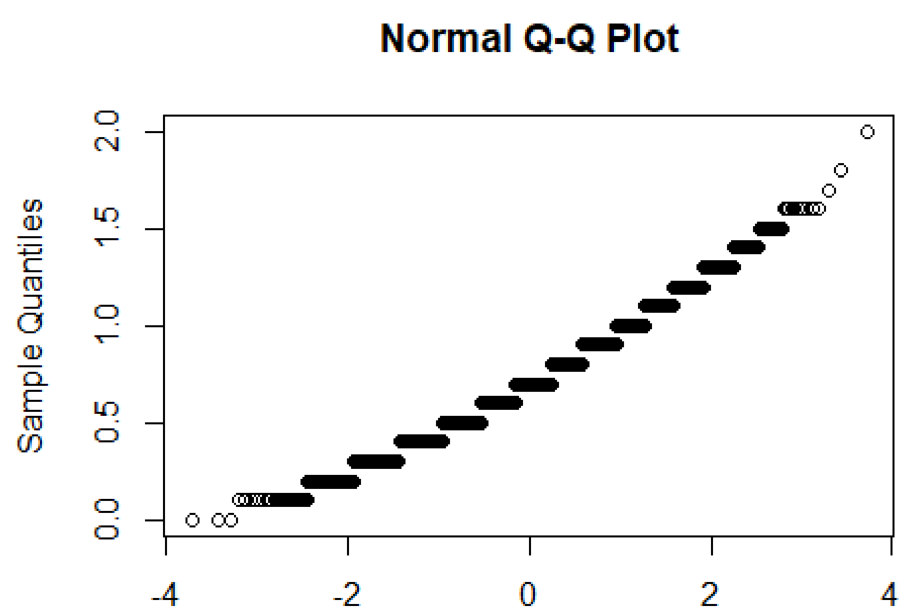
- case 2 a: $\lambda = 1.7$, n=10, N=5000
- case 2 b: $\lambda = 1.7$, n=30, N=5000
- case 2 c: $\lambda = 1.7$, n=100, N=5000
- case 2 c: $\lambda = 1.7$, n=300, N=5000

```
## write your R-function for problem 1 here
##
##
```

```r
clt = function(lambda, n, N){
  pois = c()
  for (i in c(1:N)){
    pois[i] = mean(rpois(n,lambda))
  }
  hist(pois,xlab = paste("Poisson Distribution with lambda:",lambda,"and
sample size:",n))
  qqnorm(pois,xlab = paste("Poisson Distribution with lambda:",lambda,"and
sample size:",n))
}

clt(0.7,10,5000)
```

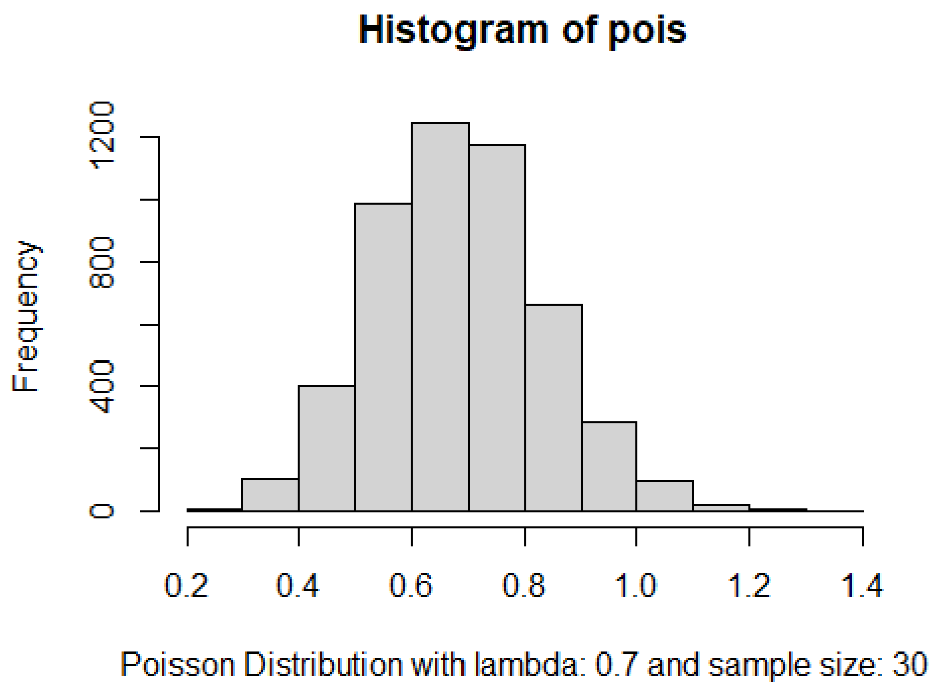


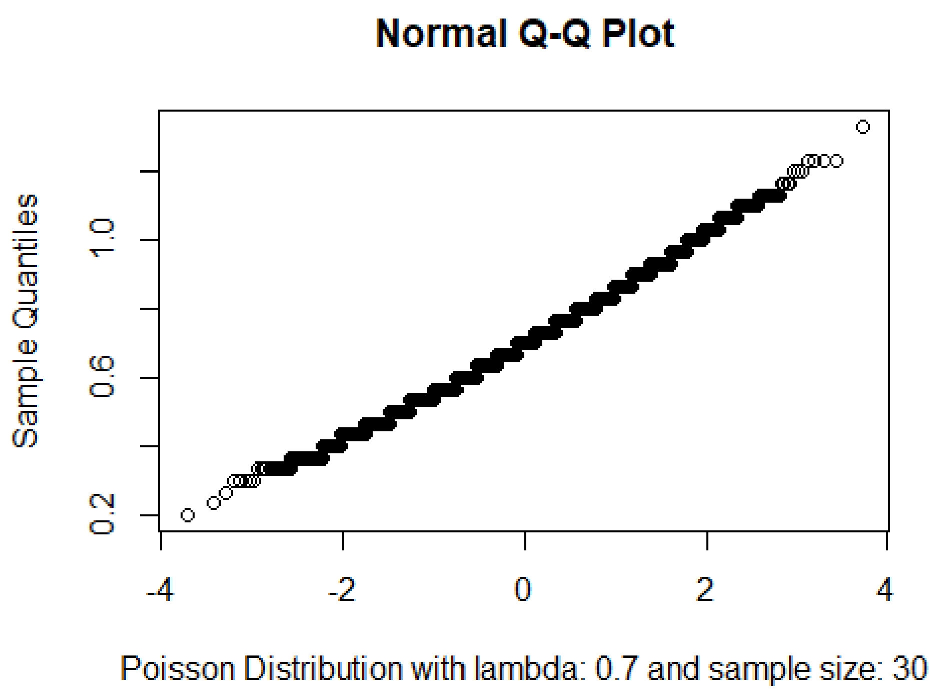
Poisson Distribution with lambda: 0.7 and sample size: 10



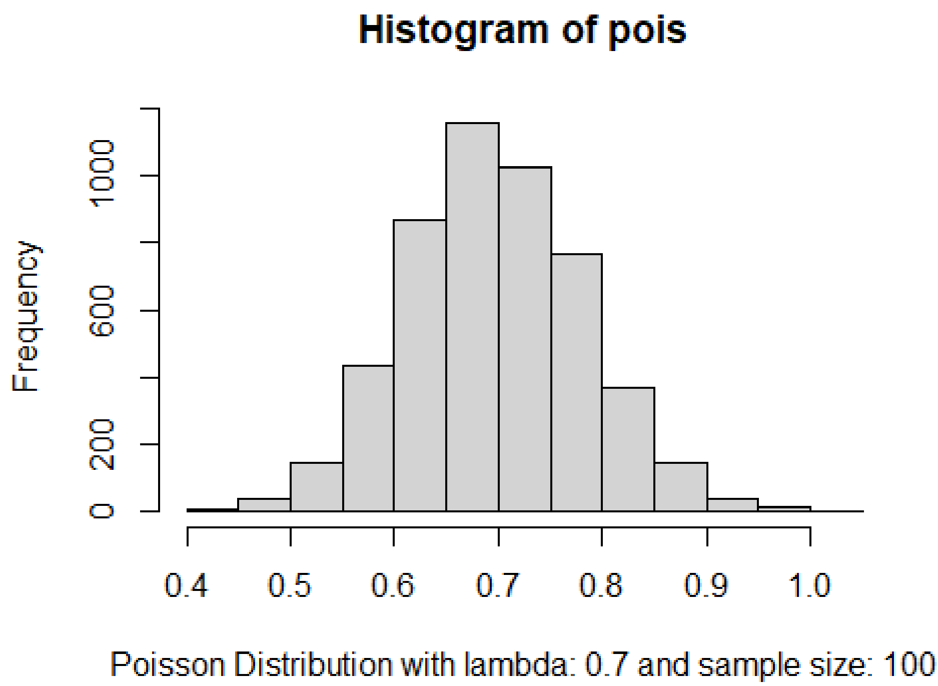
Poisson Distribution with lambda: 0.7 and sample size: 10

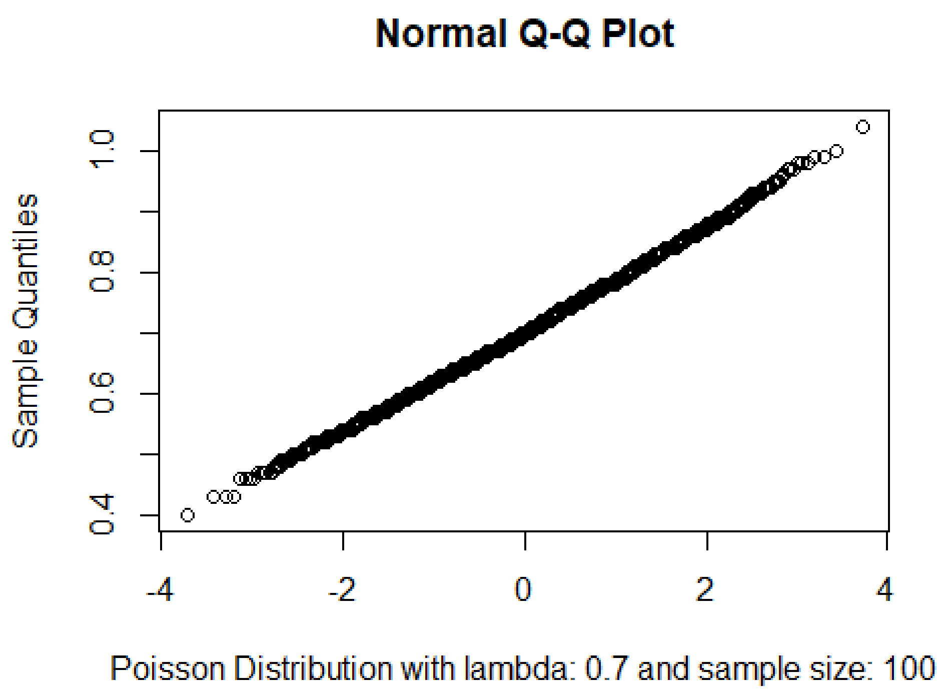
```
clt(0.7,30,5000)
```



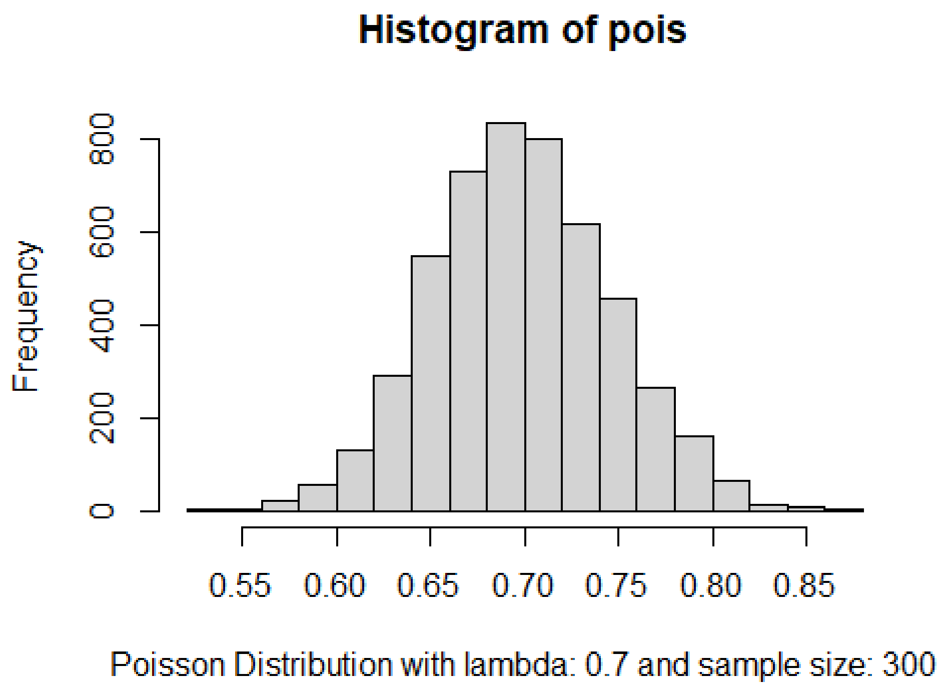


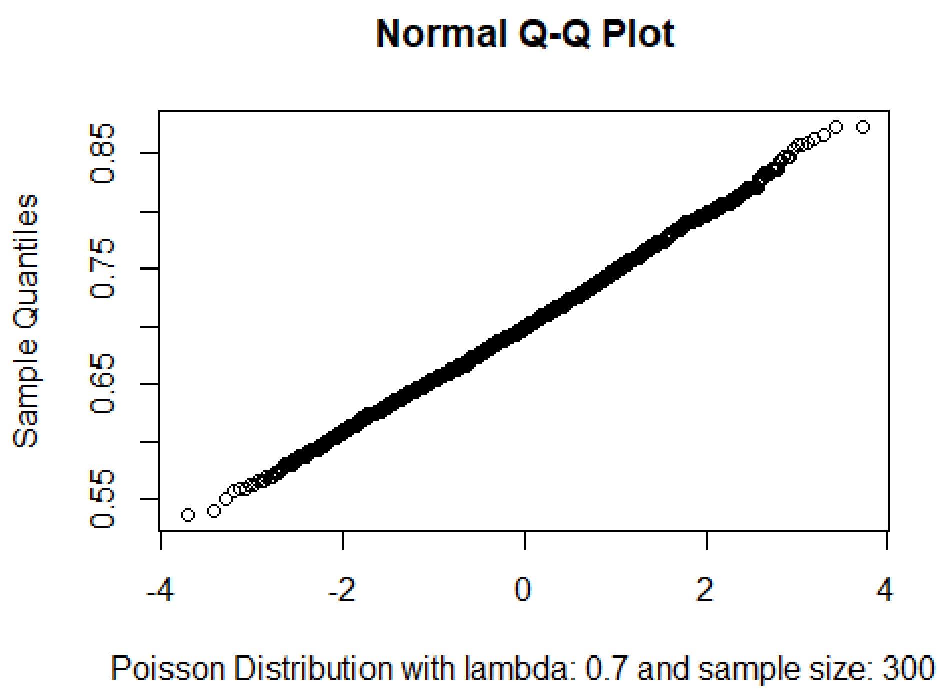

```
clt(0.7,100,5000)
```



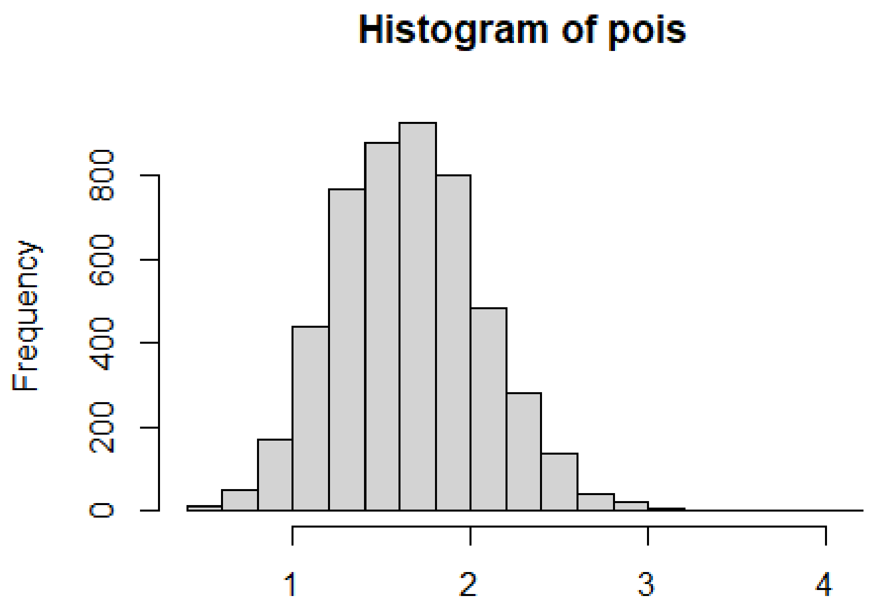



```
clt(0.7,300,5000)
```

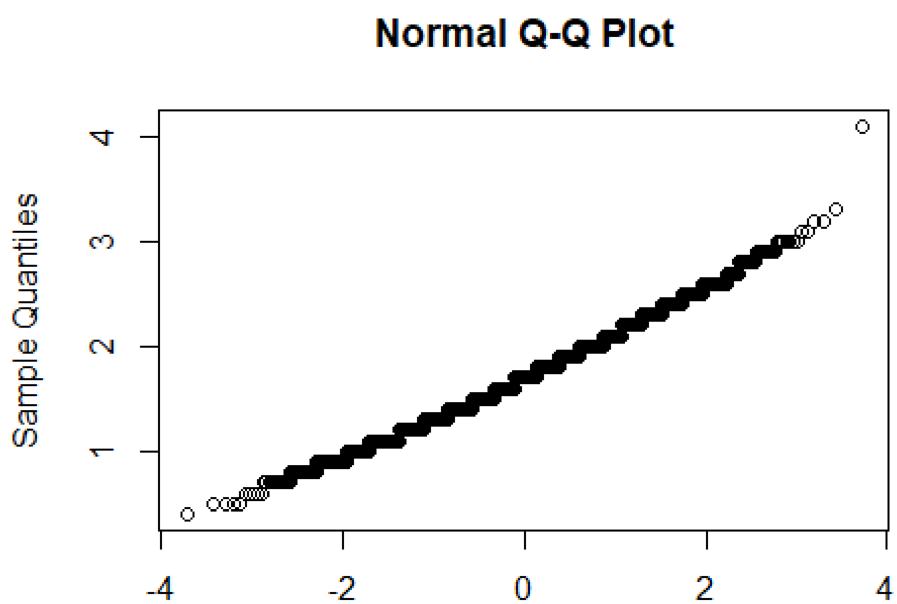





```
clt(1.7,10,5000)
```

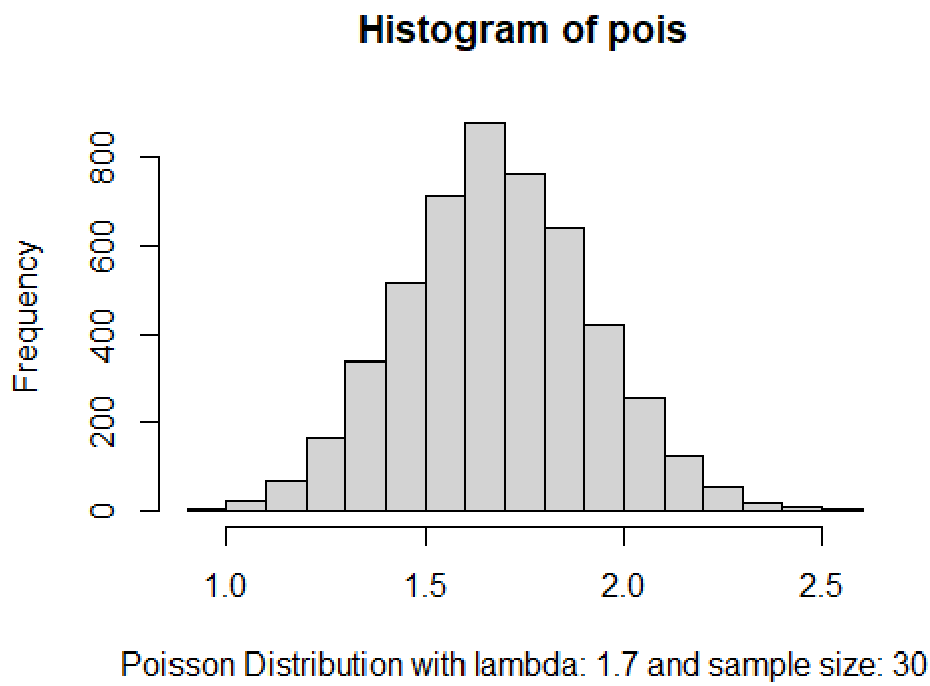


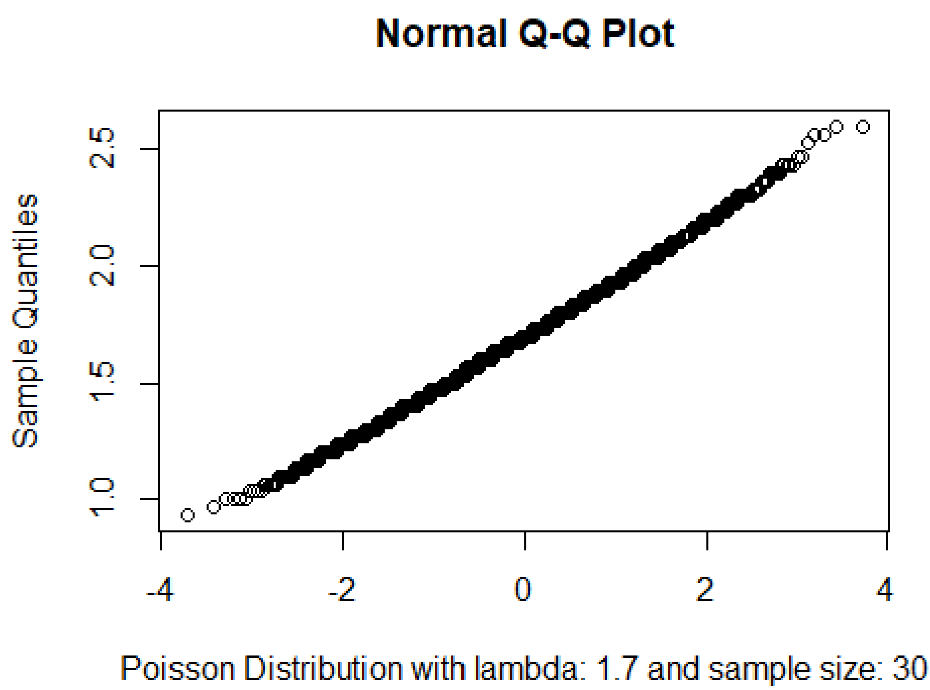
Poisson Distribution with lambda: 1.7 and sample size: 10



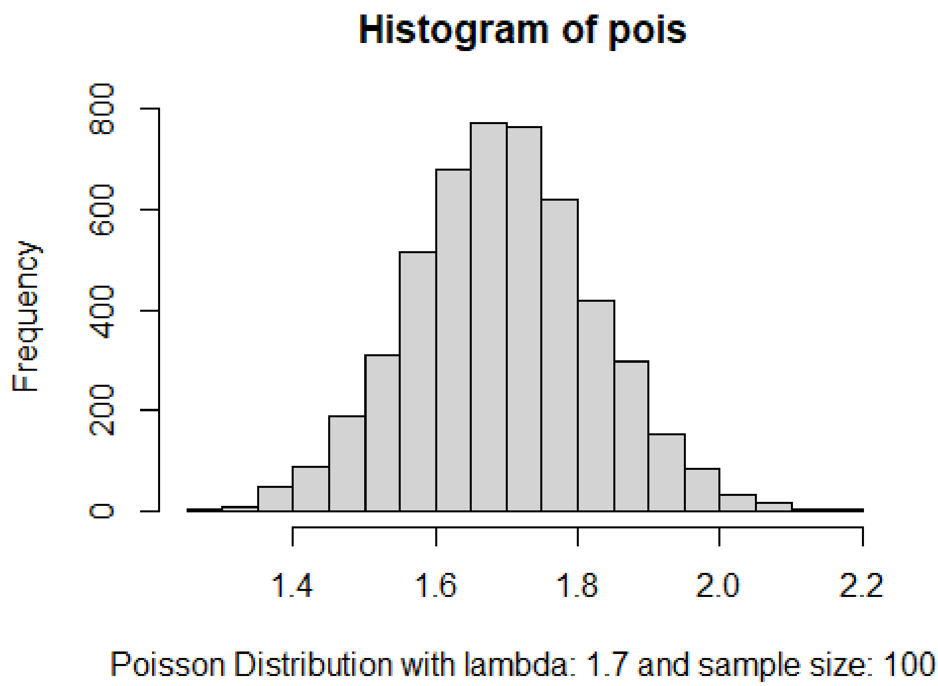
Poisson Distribution with lambda: 1.7 and sample size: 10

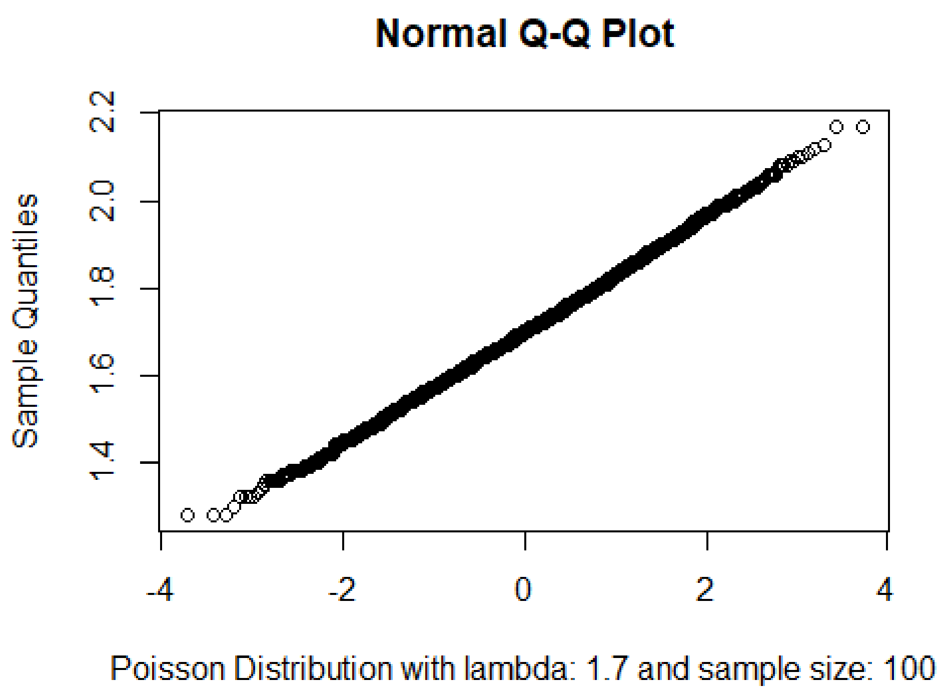
```
clt(1.7,30,5000)
```



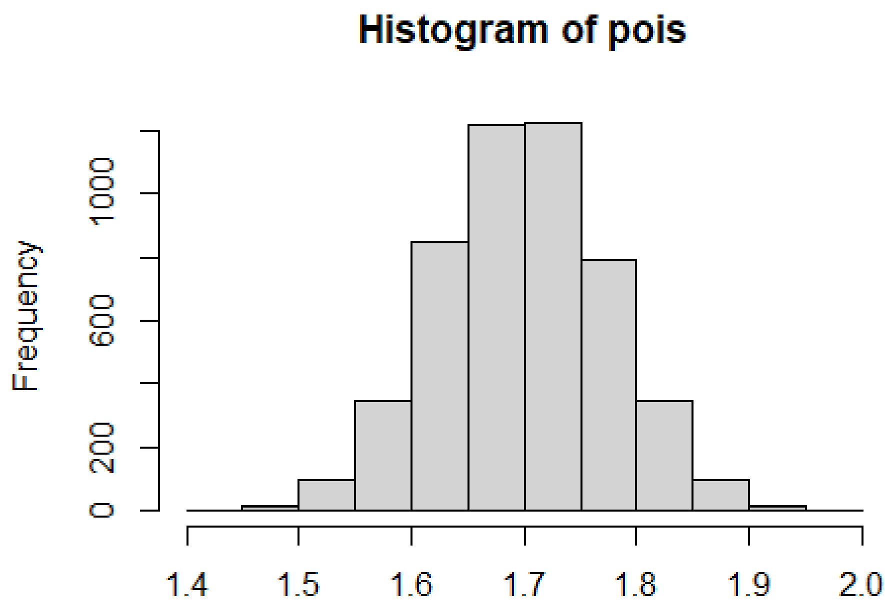



```
clt(1.7,100,5000)
```

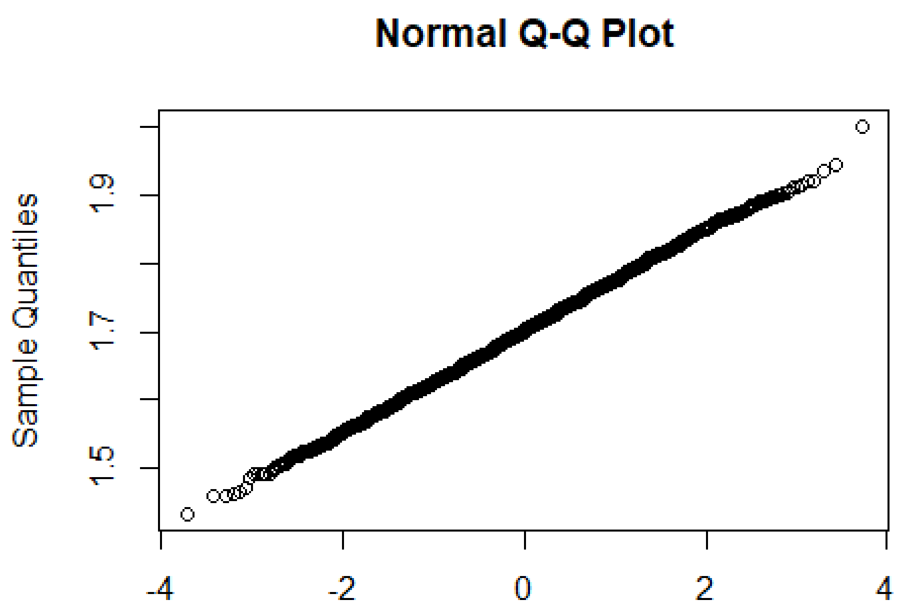





```
clt(1.7,300,5000)
```



Poisson Distribution with lambda: 1.7 and sample size: 300



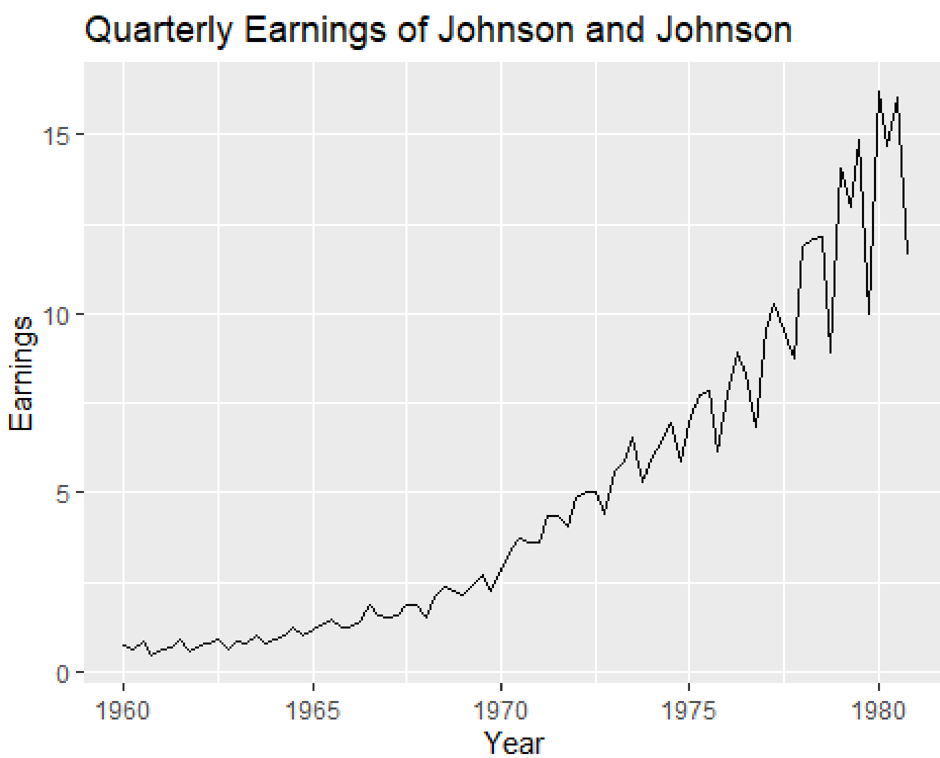
Poisson Distribution with lambda: 1.7 and sample size: 300

Problem 2: (1 point)

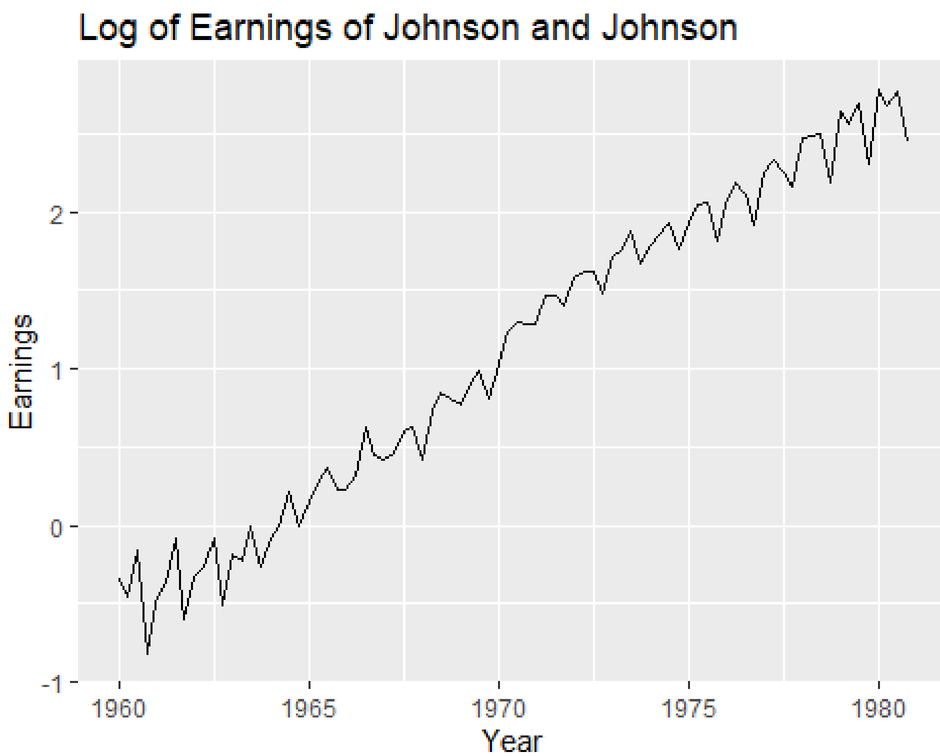
Consider the `JohnsonJohnson` dataset. The datset contains the Quarterly earnings (dollars) per Johnson & Johnson share 1960–80.

a) Draw the time series plot of Quarterly earnings in regular scale and log-scale using the ggplot (1 point)

```
head(JohnsonJohnson)

## [1] 0.71 0.63 0.85 0.44 0.61 0.69

library(ggplot2)

jjdf = data.frame(earnings = JohnsonJohnson, time = time(JohnsonJohnson))
ggplot(data = jjdf,aes(time,JohnsonJohnson))+geom_line()+ggtitle("Quarterly
Earnings of Johnson and Johnson")+xlab("Year")+ylab("Earnings")

## Don't know how to automatically pick scale for object of type ts.
Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts.
Defaulting to continuous.
```



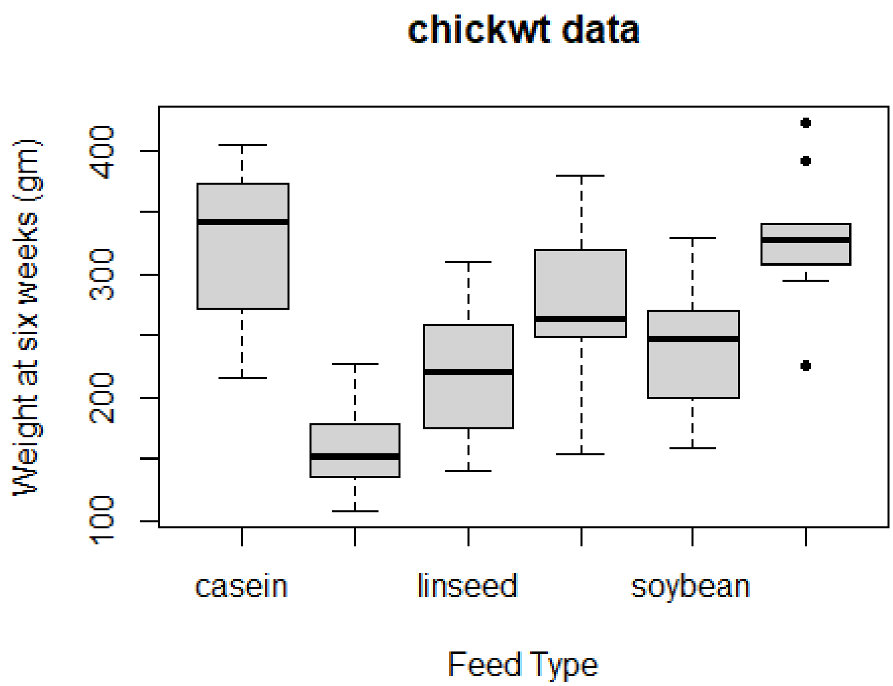

```
jjdflog = data.frame(earnings = log(JohnsonJohnson), time =
time(JohnsonJohnson))
ggplot(data = jjdflog,aes(time,log(JohnsonJohnson)))+geom_line()+ggtitle("Log
of Earnings of Johnson and Johnson")+xlab("Year")+ylab("Earnings")

## Don't know how to automatically pick scale for object of type ts.
Defaulting to continuous.
## Don't know how to automatically pick scale for object of type ts.
Defaulting to continuous.
```



Problem 3: (2 points)

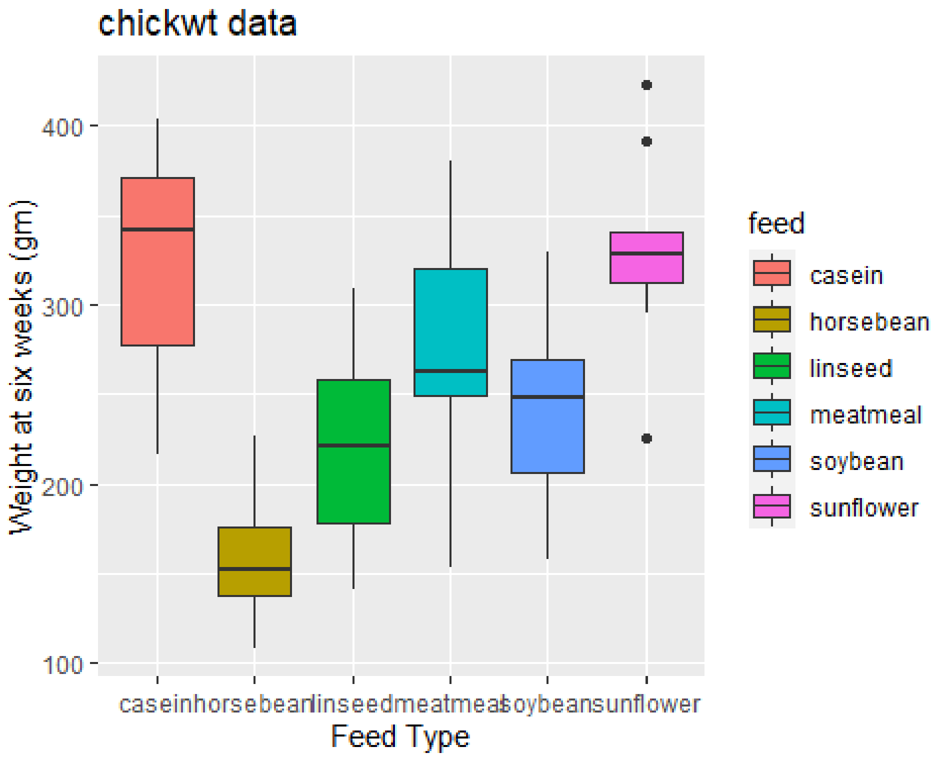
- An experiment was conducted to measure and compare the effectiveness of various feed supplements on the growth rate of chickens.
- Following R-code is a standard side-by-side boxplot showing effect of feed supplements on the growth rate of chickens.

```
boxplot(weight~feed,data=chickwts,pch=20
        ,main = "chickwt data"
        ,ylab = "Weight at six weeks (gm)"
        ,xlab = "Feed Type")
```



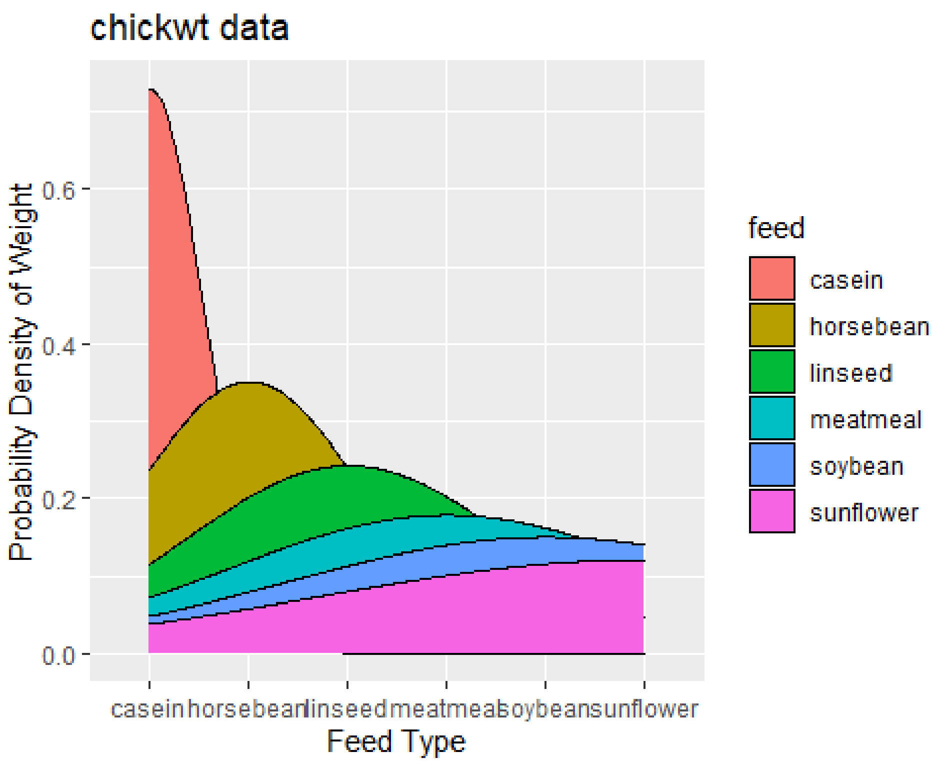
a) Reproduce the same plot using the `ggplot`; while fill each boxes with different colour. (1 point)

```
ggplot(data = chickwts,mapping = aes(feed,weight,fill =
feed))+geom_boxplot()+ylab("Weight at six weeks (gm)")+xlab("Feed
Type")+ggtitle("chickwt data")
```



b) In addition draw probability density plot for weights of chicken's growth by each feed seperately using the `ggplot`. Draw this plot seperately. (1 point)

```
ggplot(data = chickwts,mapping =
aes(feed,fill=feed))+geom_density()+ylab("Probability Density of
Weight")+xlab("Feed Type")+ggtitle("chickwt data")
```

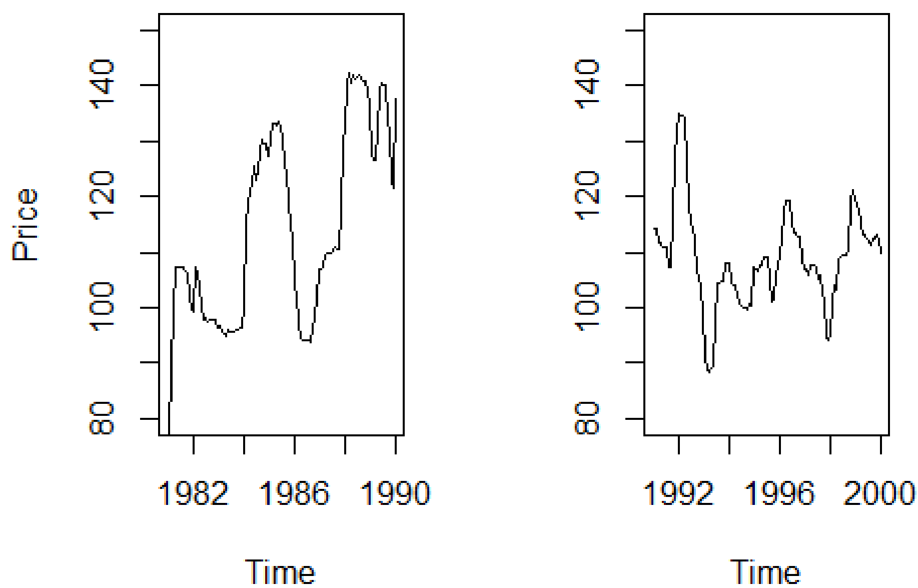


Problem 4: (4 points)

- Consider the monthly data on the price of frozen orange juice concentrate in the orange-growing region of Florida.
- The data is available in `FrozenJuice` dataset of the `AER` package.
- We want to compare the average of price between decade of 1980's and 1990's. So we split the data into two

```
library(AER)

## Loading required package: car

## Warning: package 'car' was built under R version 4.1.2

## Loading required package: carData

## Loading required package: lmtest

## Warning: package 'lmtest' was built under R version 4.1.2

## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: sandwich

## Loading required package: survival

data("FrozenJuice")

data_80_90=window(FrozenJuice,start=1981,end=1990)
data_90_2K=window(FrozenJuice,start=1991,end=2000)
par(mfrow=c(1,2))
plot(data_80_90[,'price'],ylim=c(80,150),ylab='Price')
plot(data_90_2K[,'price'],ylim=c(80,150),ylab='')
```



- Generally it is believed that the price of the product increases over time due to inflation effect. So we expect that the average price during 1991-2000 would be higher than the 1981-1990.

The mean and standard deviation of price is estimates as

```
n1 = nrow(data_80_90)
cat('number of samples in 80s decade: ',n1,'\n')

## number of samples in 80s decade:  109

m1 = mean(data_80_90[,'price'])
s1 = sd(data_80_90[,'price'])
cat('mean and sd for 80s decade','\n')

## mean and sd for 80s decade
```

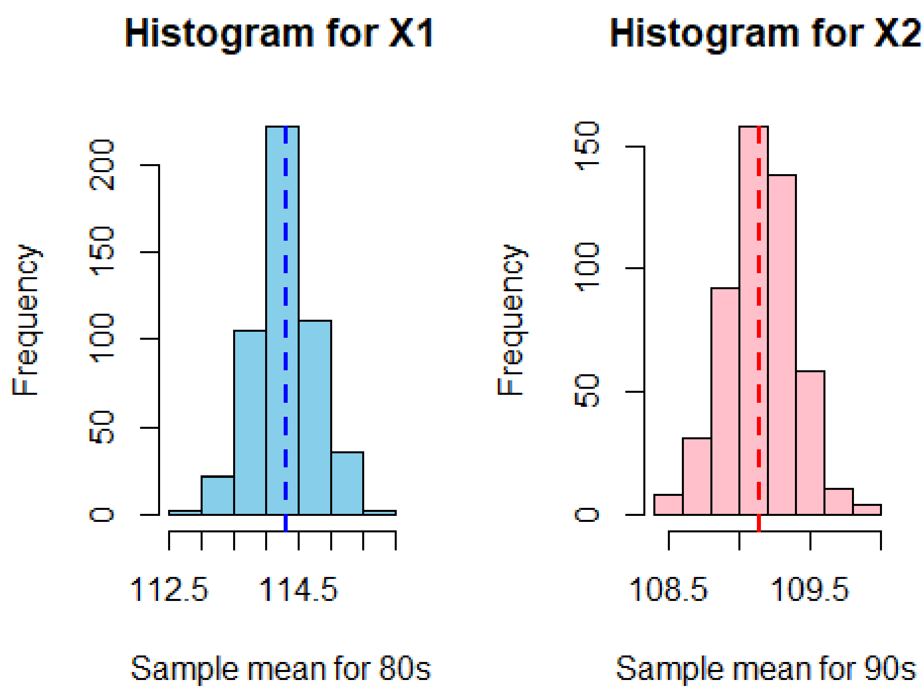
```
round(c(mean = m1,sd = s1),2)

##   mean     sd 
## 114.32  16.88

n2 = nrow(data_90_2K)
cat('number of samples in 90s decade: ',n2,'\n')

## number of samples in 90s decade:  109

m2 = mean(data_90_2K[,'price'])
s2 = sd(data_90_2K[,'price'])
cat('mean and sd for 90s decade','\n')

## mean and sd for 90s decade

round(c(mean = m2,sd = s2),2)

##   mean     sd 
## 109.14   9.25

round(c(mean = m2,sd = s2),2)

##   mean     sd 
## 109.14   9.25
```

- The sample size for both decades are more than 100. So we can assume that CLT will kick-in.

a) If $\bar{X}_1$ and $\bar{X}_2$ are the sample mean of the price the two decades, plot the sampling distributions of sample mean for both decades on the same graph. (1 point)

```
n=100
sim.size = 500
sample_mean1 = rep(NA,n)
sample_mean2 = rep(NA,n)
for (i in 1:sim.size){
  sam1 = sample(data_80_90[,'price'],n,prob = NULL)
  sample_mean1[i]=mean(sam1)
}

for (i in 1:sim.size){
  sam2 = sample(data_90_2K[,'price'],n,prob = NULL)
  sample_mean2[i]=mean(sam2)
}


par(mfrow=c(1,2))
h1 =hist(sample_mean1,col='skyblue',xlab='Sample mean for 
80s',main='Histogram for X1')
abline(v=m1,lwd=2,col='blue',lty=2)
h2 = hist(sample_mean2,col='pink',xlab='Sample mean for 90s',main='Histogram
```

```
for X2')
abline(v=m2,lwd=2,col='red',lty=2)
```



b) Simulate the $\bar{X}_1$ and $\bar{X}_2$ from respective sampling distribution, then calculate the difference.

$$d = \bar{X}_1 - \bar{X}_2$$

c) Simulate d; `5000` times. (1 point)

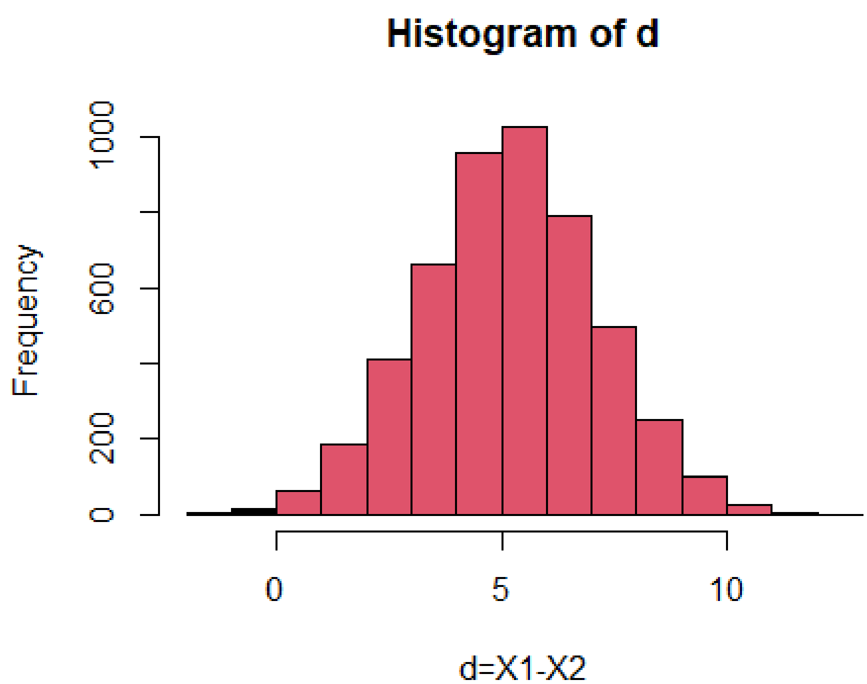
```
n=50
sim.size = 5000
sample_mean1 = rep(NA,n)
sample_mean2 = rep(NA,n)
for (i in 1:sim.size){
  sam1 = sample(data_80_90[,'price'],n,prob = NULL)
  sample_mean1[i]=mean(sam1)
}

for (i in 1:sim.size){
  sam2 = sample(data_90_2K[,'price'],n,prob = NULL)
  sample_mean2[i]=mean(sam2)
}
```

c) Calculate $P(d < 0)$ as

$$\hat{P}(d < 0) = \frac{\text{number of d<0}}{5000}$$

d) and draw the histogram of d and marked the area where $d < 0$ (1 point) http://127.0.0.1:11992/help/library/graphics/help/plot.histogram

```
d = sample_mean1-sample_mean2
probd = sum(d<0)/5000
cat("The probability that d is less than 0 is :")

## The probability that d is less than 0 is :

print(probd)

## [1] 0.0038

d_hist = hist(d, plot = F)
cuts = cut(d_hist$breaks, breaks = c(-Inf, -0.01, Inf))
plot(d_hist,col=cuts,xlab='d=X1-X2')
```



d) Based on the analysis, what is the chance that the average price of Juice for decade 1981-90 was same or less than the decade of 1991-2000? (1 point)

```
#There is approximately 0.38% chance that the average price for Orange Juice in the
#80s was same or less than the average price of Orange Juice in the 90s.
#From the Histogram also we can observe almost none of the values taken by d are negative.
```