Visualisation: Assignment 1

Students Name: Aniket Santra Roll Number: MDS202106

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

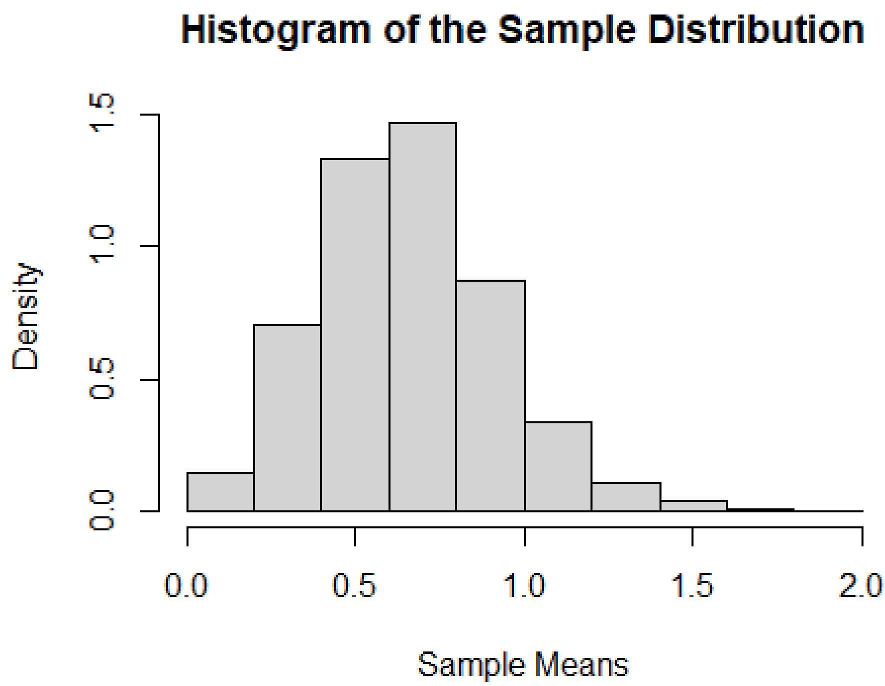
- case 2 a: $\lambda = 1.7$, n=10, N=5000
- case 2 b: $\lambda = 1.7$, n=30, N=5000
- case 2 c: $\lambda = 1.7$, n=100, N=5000
- case 2 c: $\lambda = 1.7$, n=300, N=5000

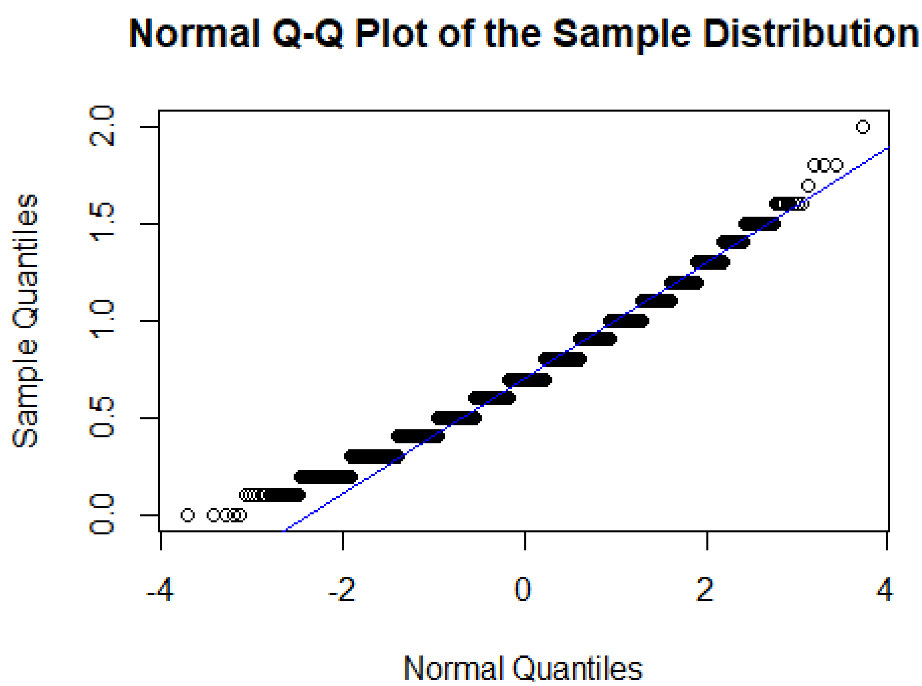
```r
## write your R-function for problem 1 here
poiss_clt=function(l,n,N){
  s_means=rep(NA,N)
```

```
  for (k in 1:N){
    p_sam=rpois(n,l)
    e_lamda=mean(p_sam)
    s_means[k]=e_lamda}
  hist(s_means, probability=TRUE, main="Histogram of the Sample
Distribution", xlab="Sample Means")
  qqnorm(s_means, main="Normal Q-Q Plot of the Sample Distribution",
xlab="Normal Quantiles")
  qqline(s_means,col="blue")}
```

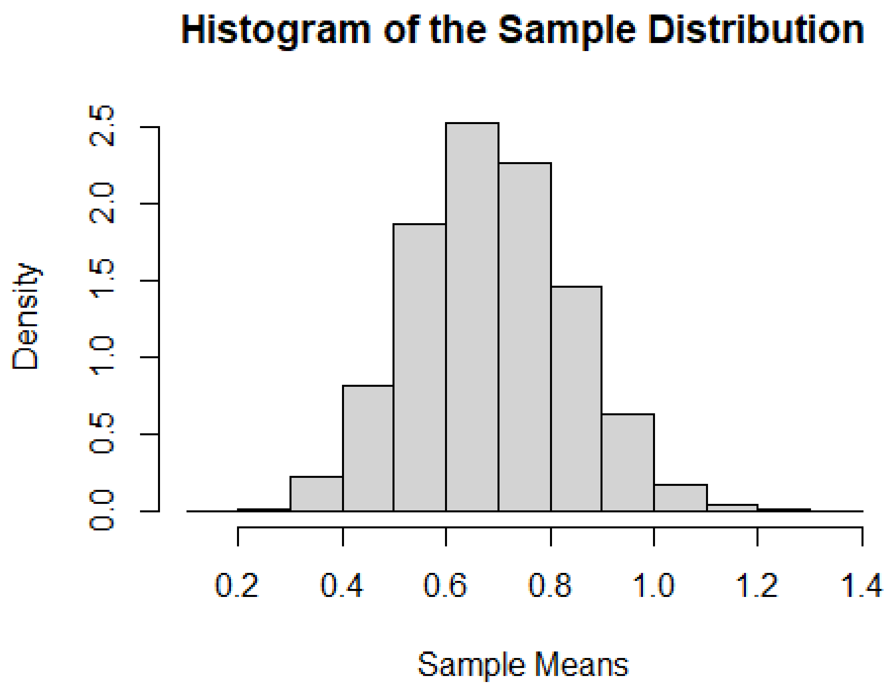
Test cases case 1

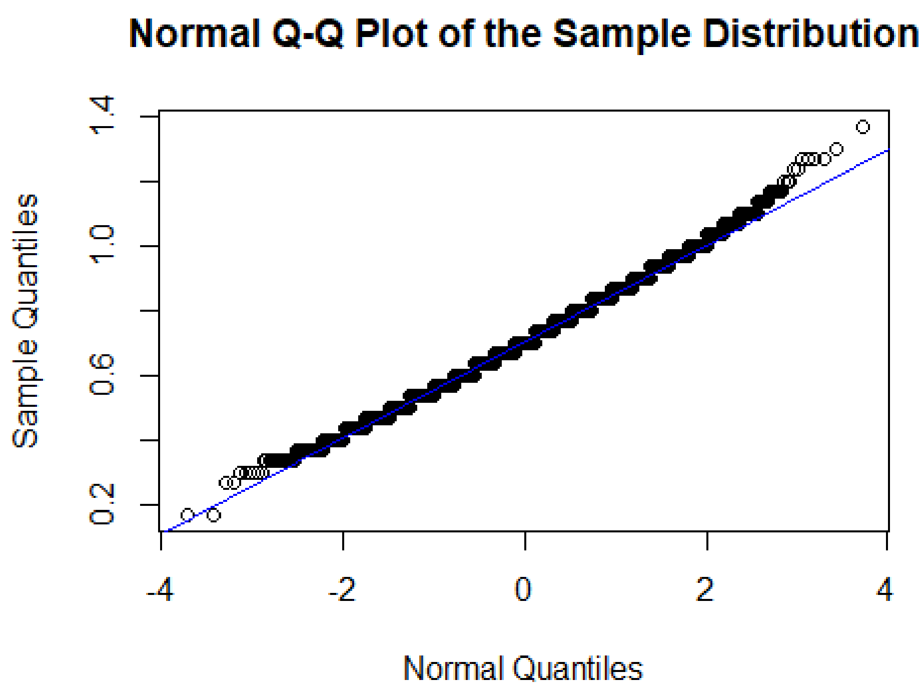
```
poiss_clt(0.7,10,5000)
```



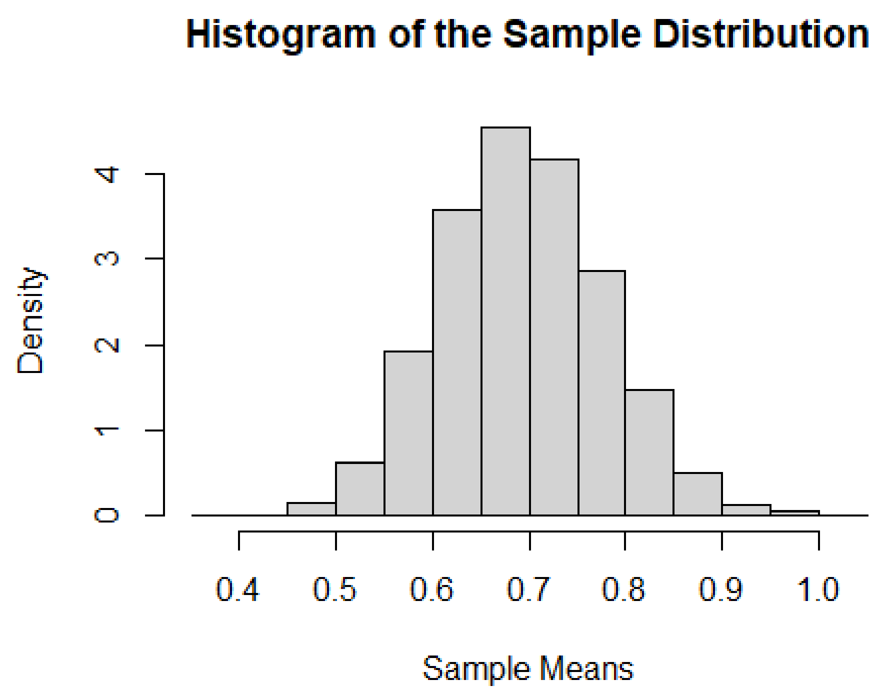


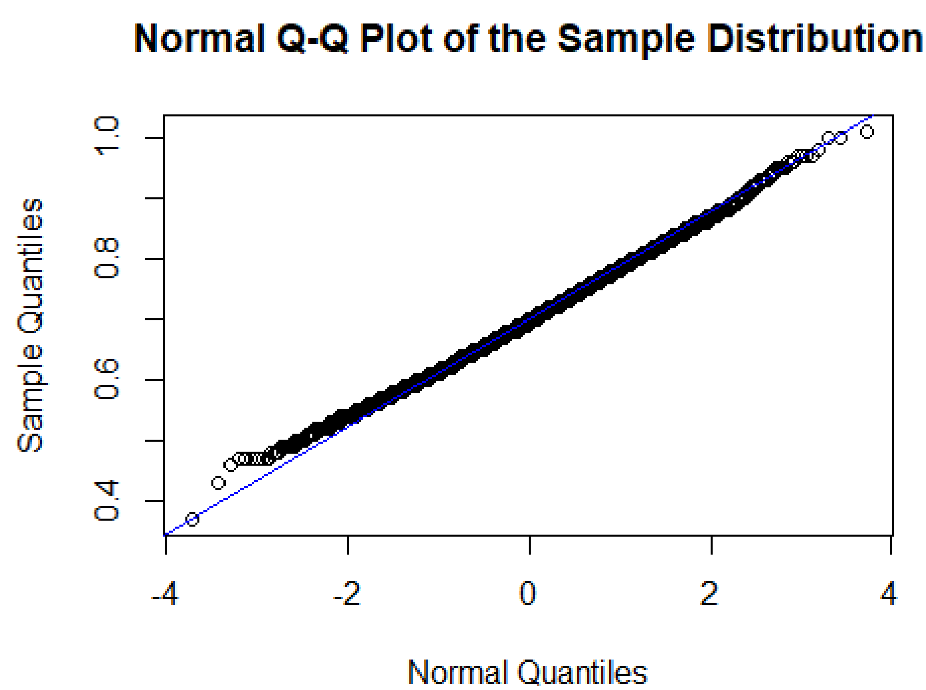

```
poiss_clt(0.7,30,5000)
```



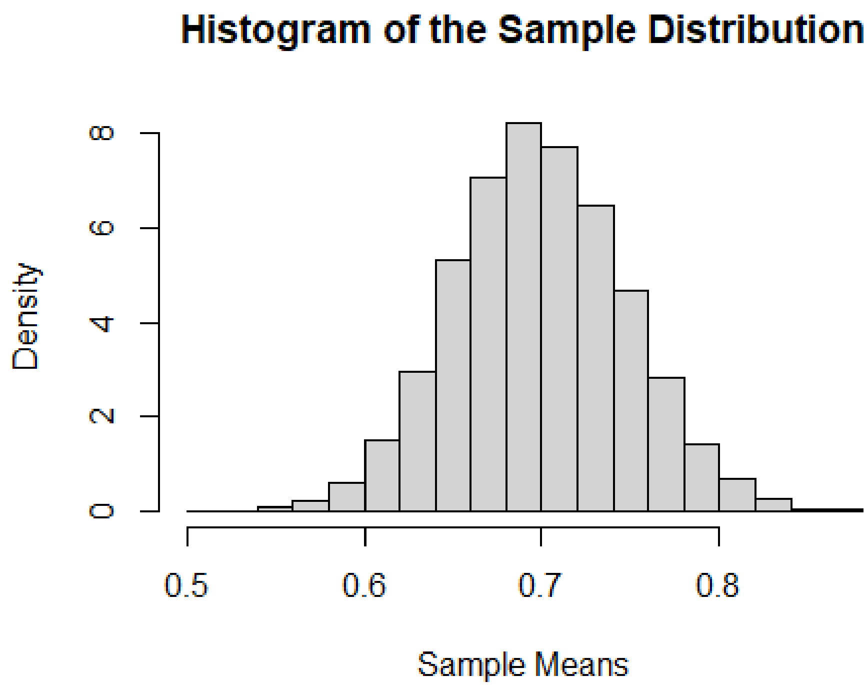


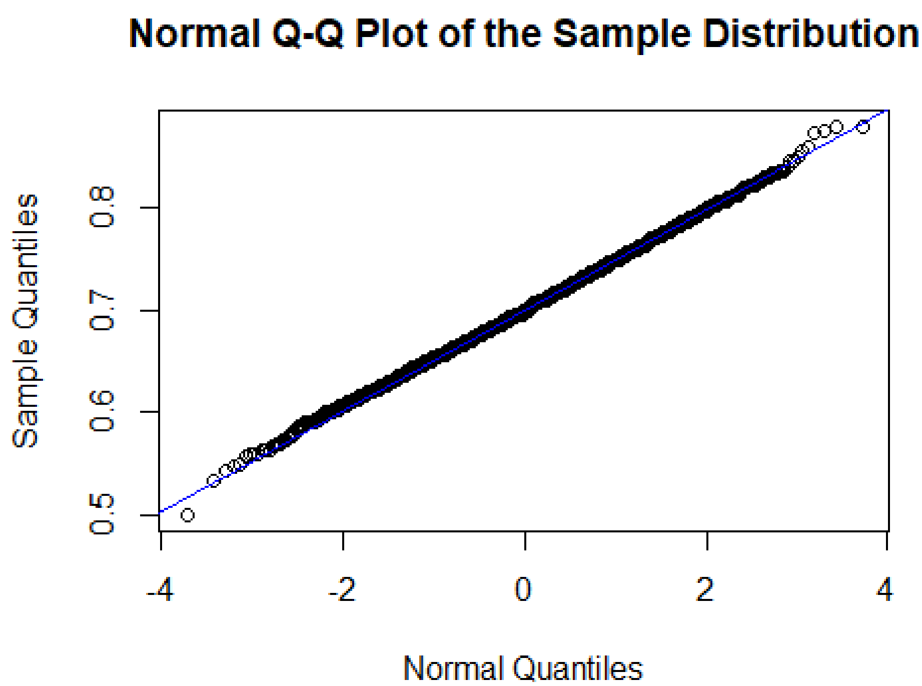

```
poiss_clt(0.7,100,5000)
```





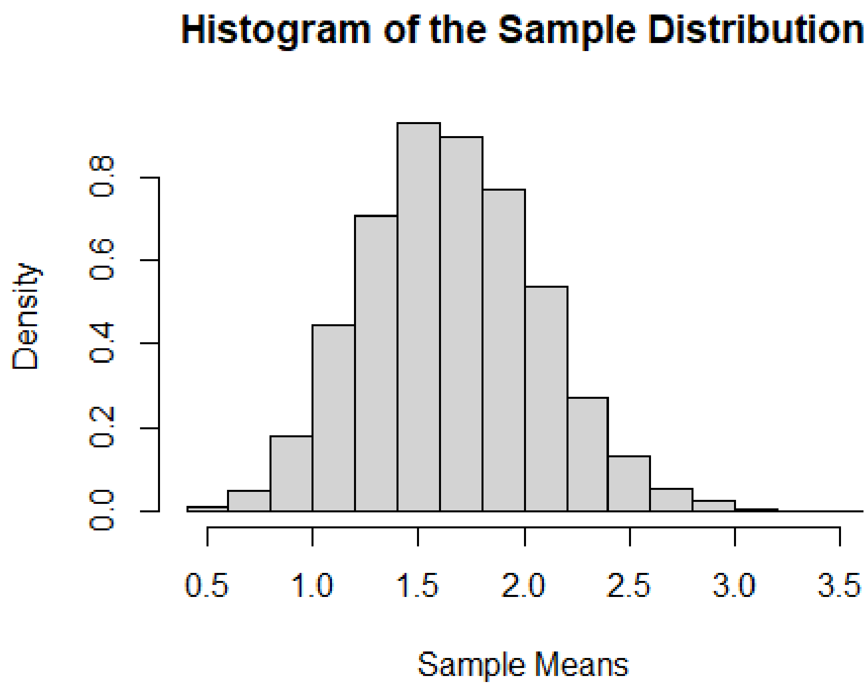

```
poiss_clt(0.7,300,5000)
```

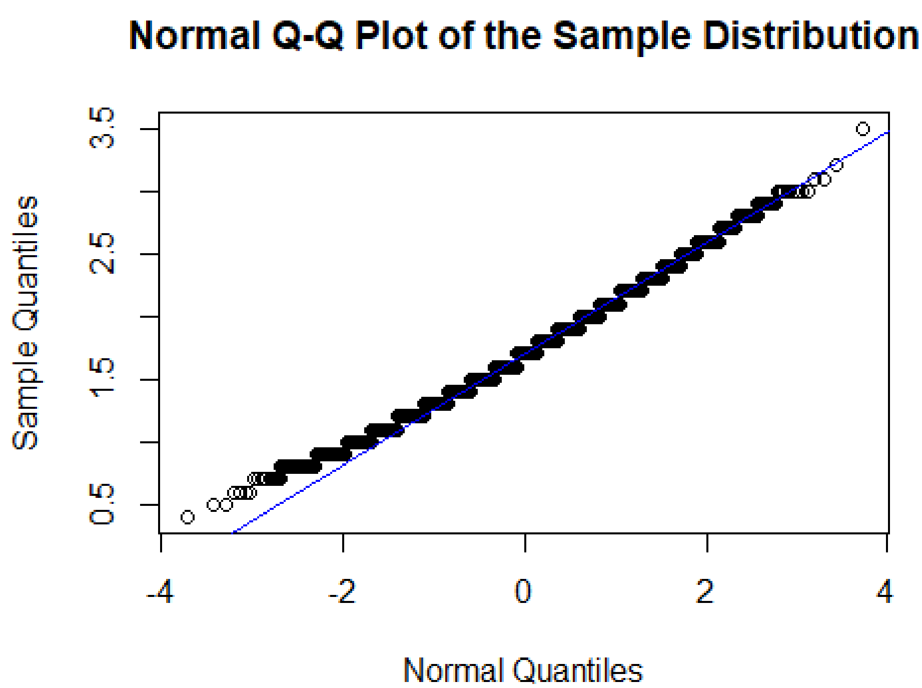




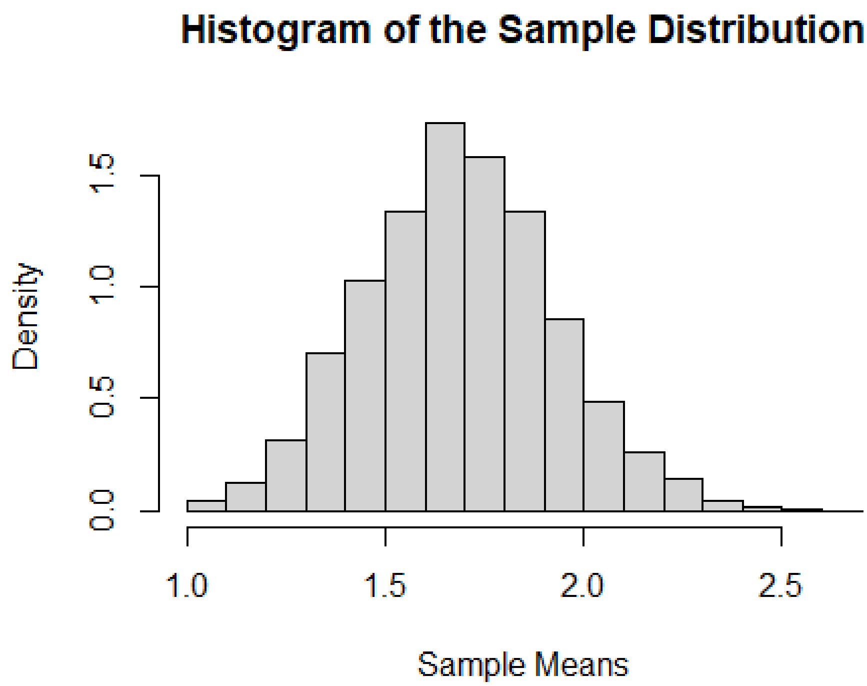
case 2

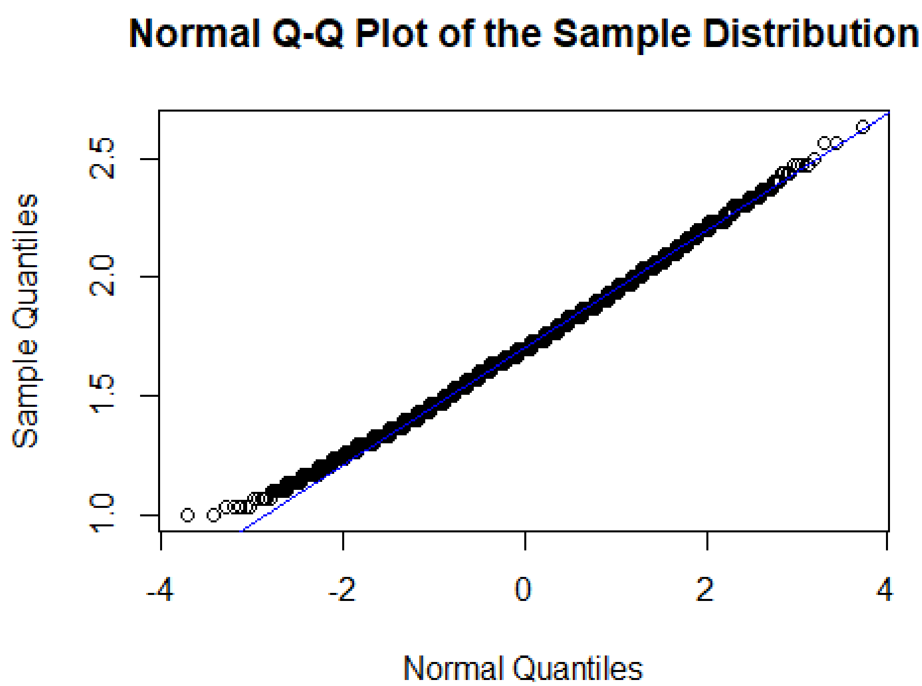
```
poiss_clt(1.7,10,5000)
```



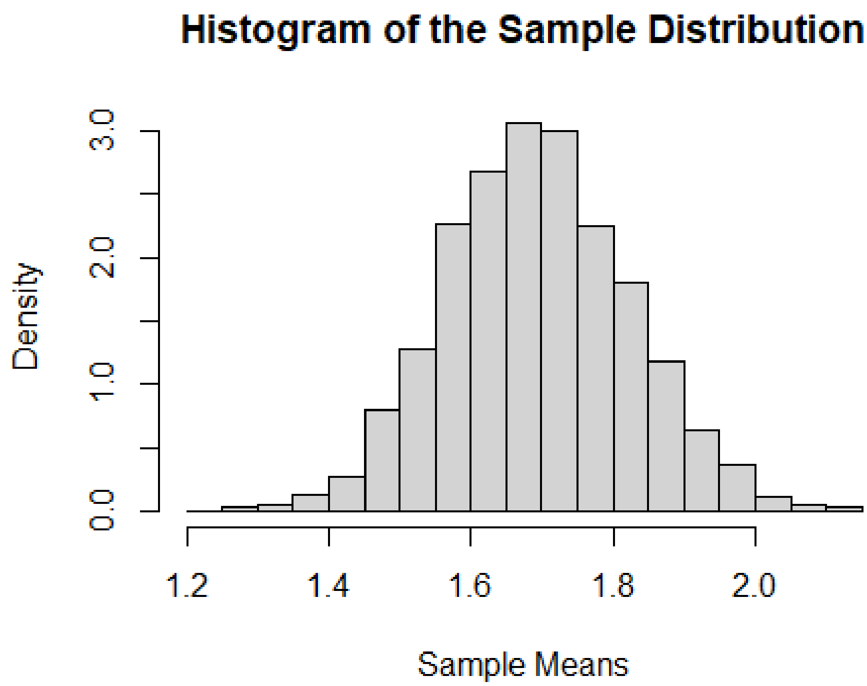


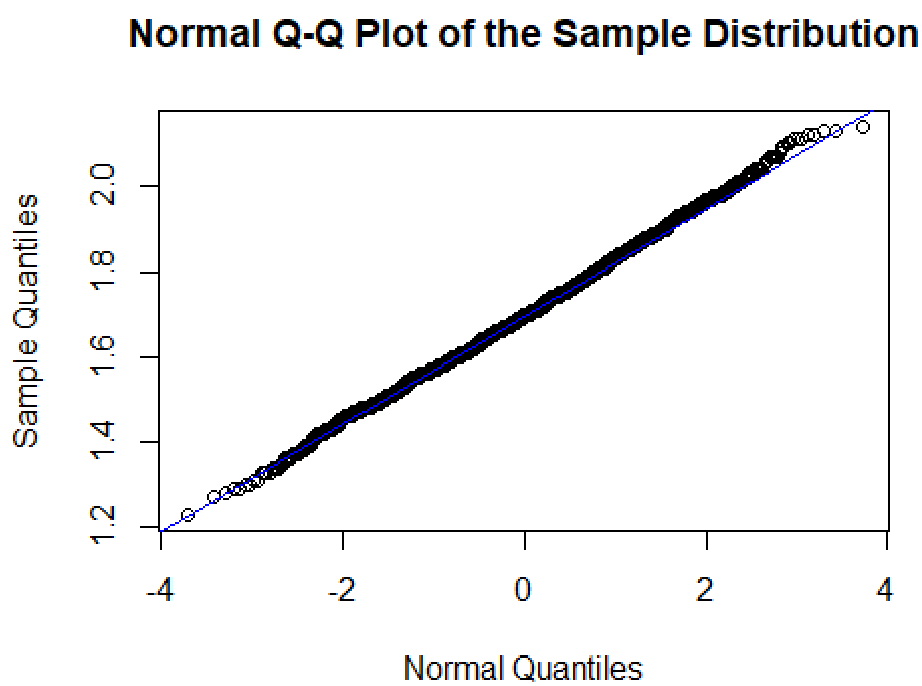

```
poiss_clt(1.7,30,5000)
```



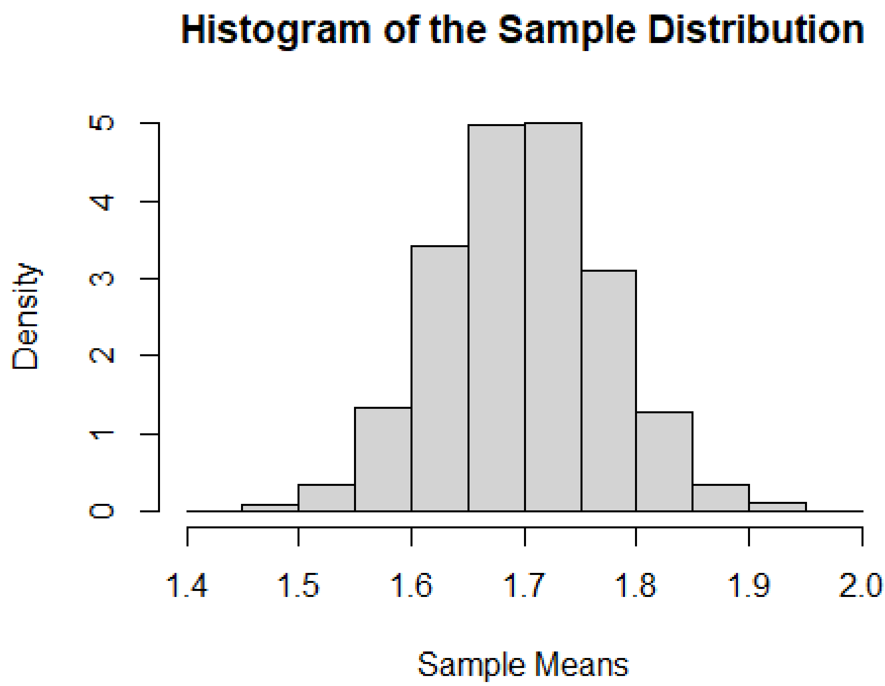


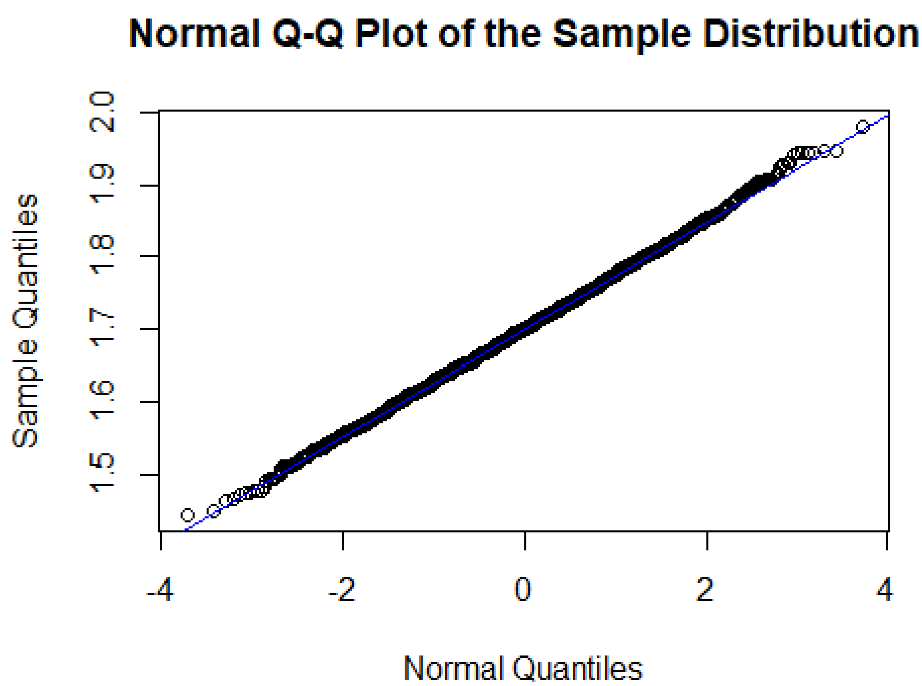

```
poiss_clt(1.7,100,5000)
```






```
poiss_clt(1.7,300,5000)
```



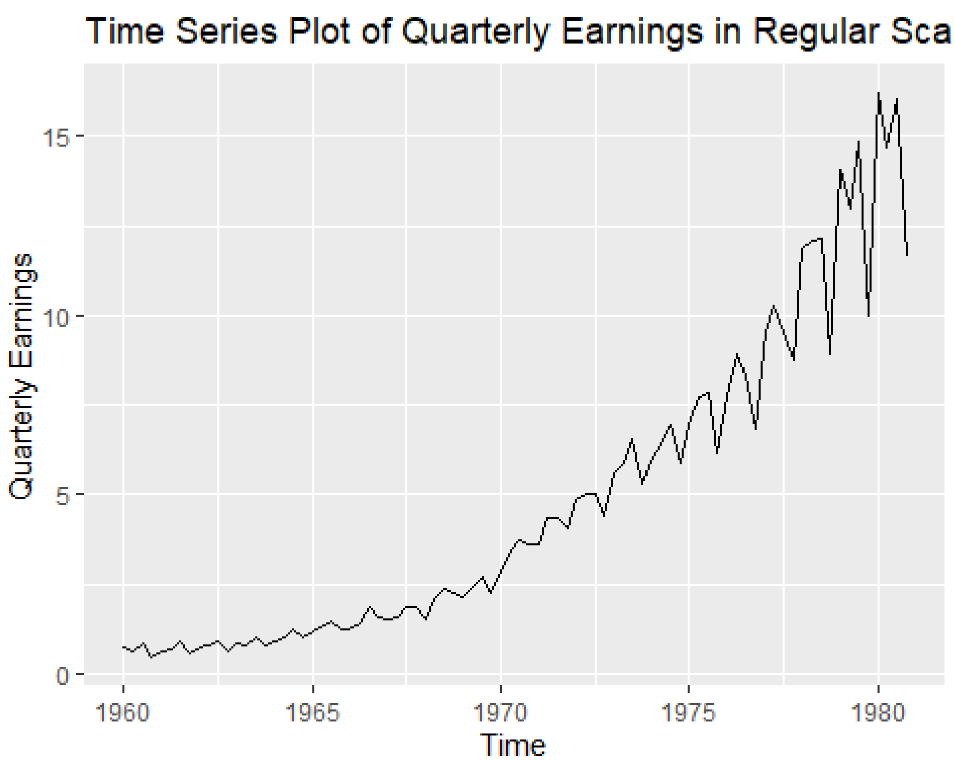


Problem 2: (1 point)

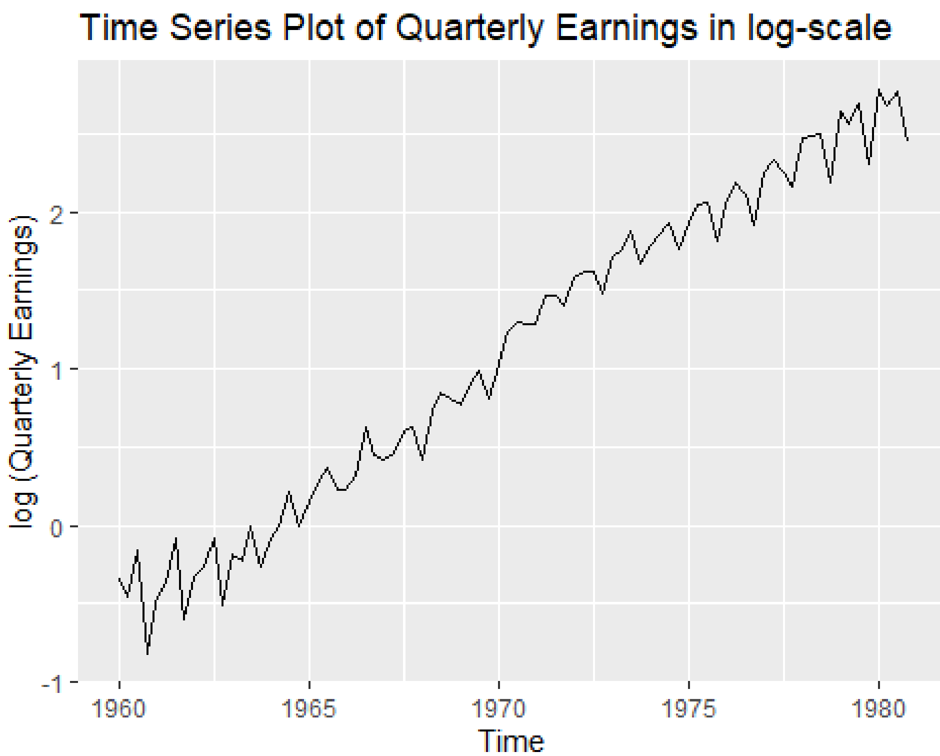
Consider the `JohnsonJohnson` dataset. The datset contains the Quarterly earnings (dollars) per Johnson & Johnson share 1960–80.

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)
JJ_data=data.frame(cbind(quarterly_earnings=JohnsonJohnson,time=time(JohnsonJ
ohnson)))
ggplot(JJ_data)+geom_line(aes(x=time,y=quarterly_earnings))+labs(title="Time
Series Plot of Quarterly Earnings in Regular Scale",x="Time",y="Quarterly
Earnings")
```



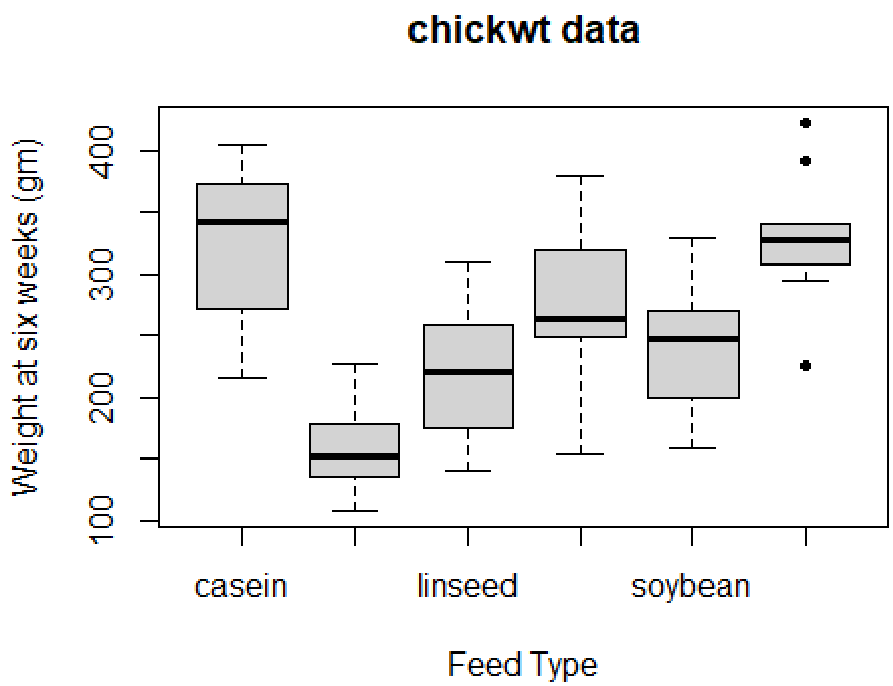

```
ggplot(JJ_data)+geom_line(aes(x=time,y=log(quarterly_earnings)))+labs(title="
Time Series Plot of Quarterly Earnings in log-scale",x="Time",y="log
(Quarterly Earnings)")
```



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.

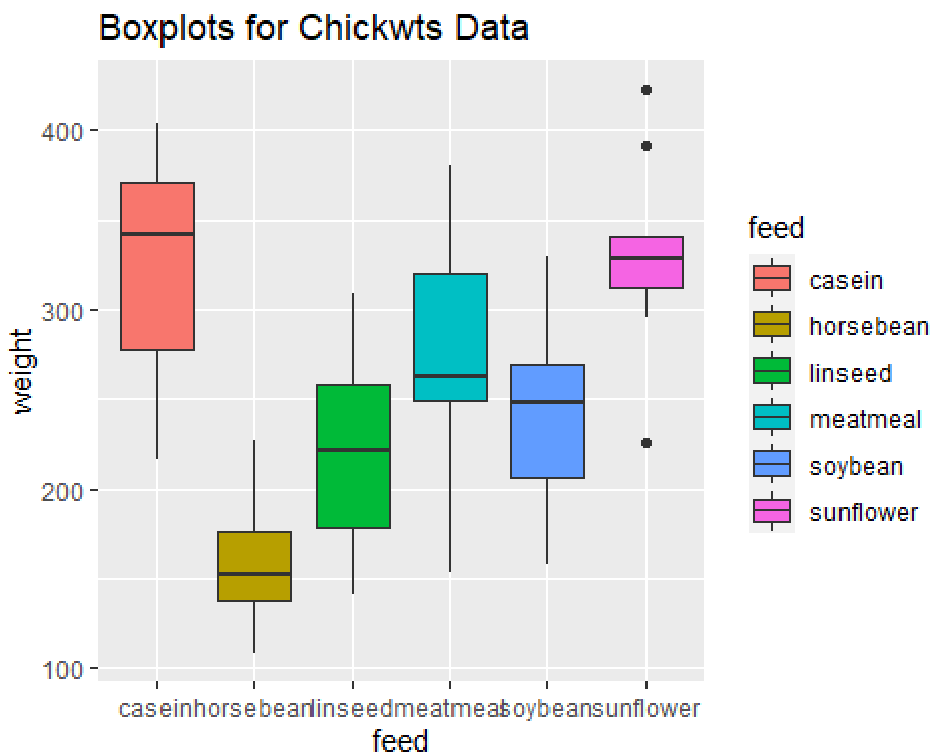
```
boxplot(weight~feed,data=chickwts,pch=20
        ,main = "chickwt data"
        ,ylab = "Weight at six weeks (gm)"
        ,xlab = "Feed Type")
```



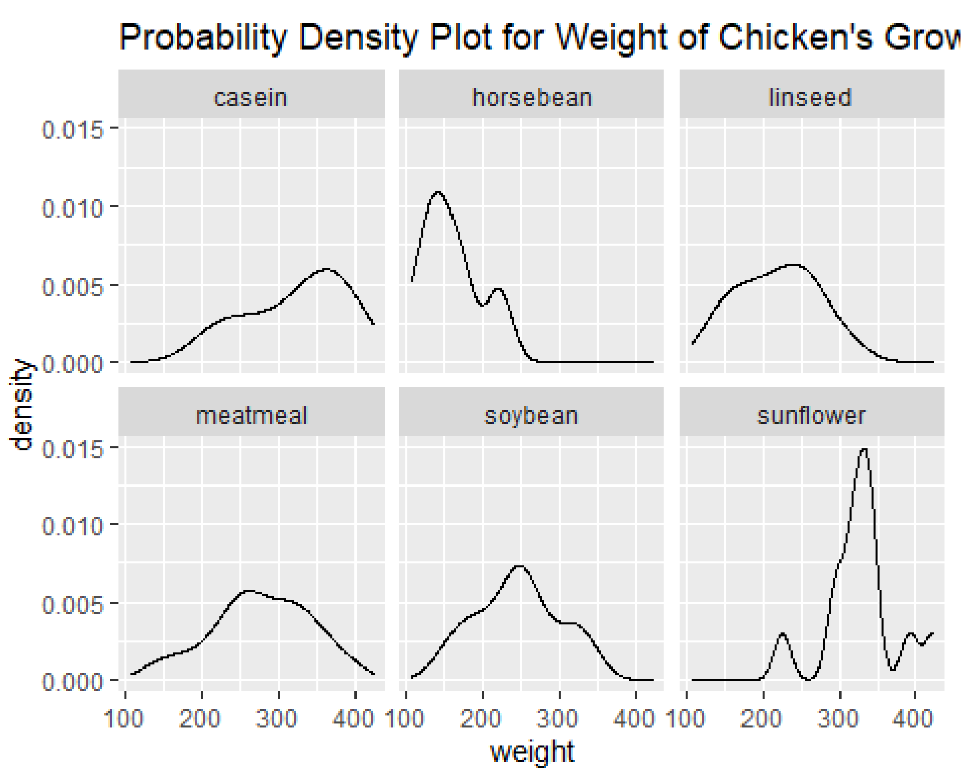
a) Reproduce the same plot using the `ggplot`; while fill each boxes with different colour. (1 point)

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)

```
library(ggplot2)
ggplot(chickwts)+geom_boxplot(mapping=aes(x=feed,y=weight,fill=feed))+labs(ti
tle="Boxplots for Chickwts Data")
```



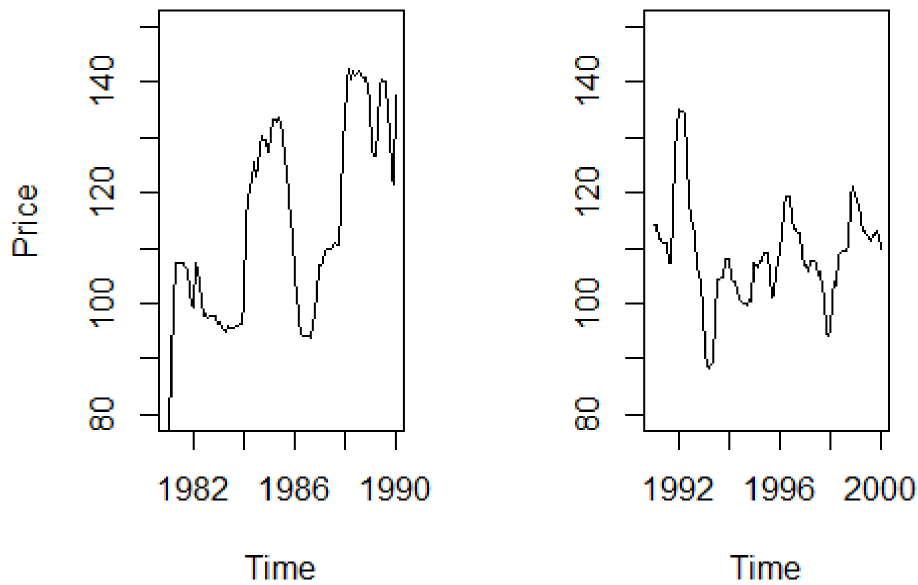

```
ggplot(chickwts)+geom_density(aes(x=weight))+facet_wrap(.~feed)+labs(title="P
robability Density Plot for Weight of Chicken's Growth By Each Feed
separately")
```



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
```

```r
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
```

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

```
##   mean     sd
## 109.14   9.25
```

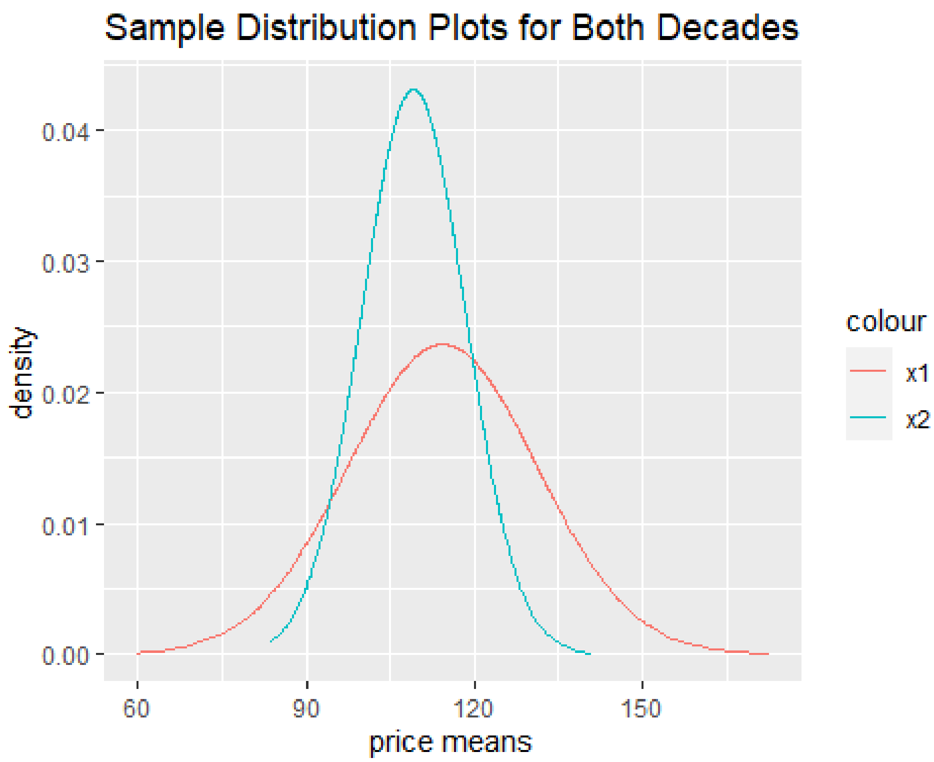
```r
round(c(mean = m2,sd = s2),2)
```

```
##   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)

```r
library(ggplot2)
x1=sort(rnorm(1000,mean=m1,sd=s1))
density_x1=dnorm(x1,mean=m1,sd=s1)
x2=sort(rnorm(1000,mean=m2,sd=s2))
density_x2=dnorm(x2,mean=m2,sd=s2)
df=data.frame(cbind(x1,density_x1,x2,density_x2))
ggplot(df)+geom_line(aes(x=x1,y=density_x1,col="x1"))+geom_line(aes(x=x2,y=de
nsity_x2,col="x2"))+labs(x="price means",y="density",title="Sample
Distribution Plots for Both Decades")
```



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

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

Simulate d; 5000 times. (1 point)

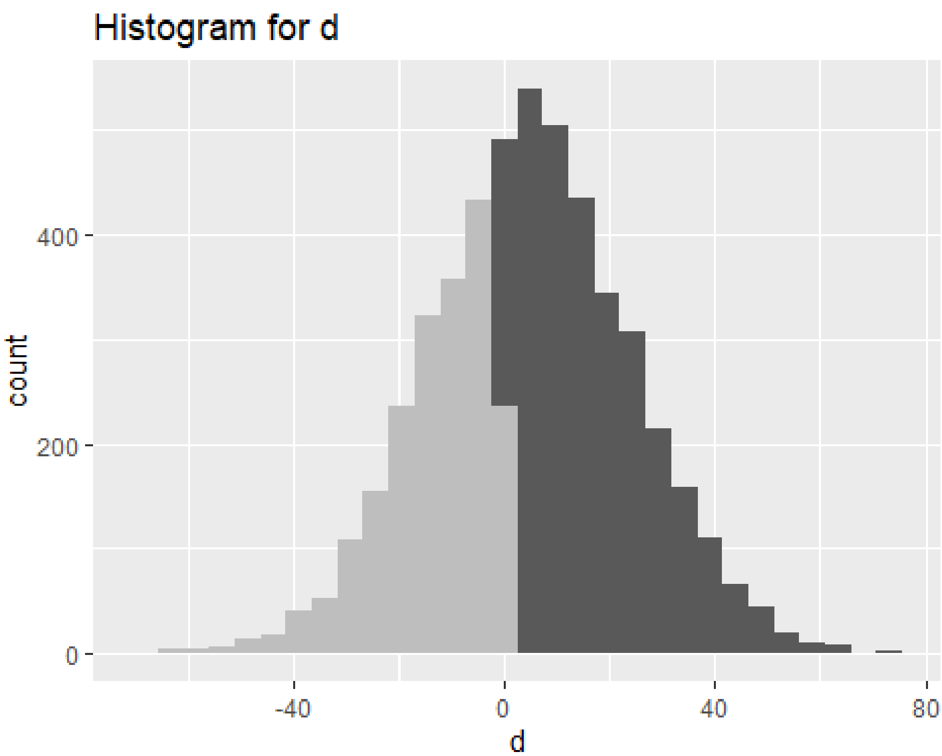
```
x1_sim=rnorm(5000,m1,s1)
x2_sim=rnorm(5000,m2,s2)
d=x1_sim-x2_sim
```

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)

```
#The probability d<0 is:-
p=sum(d<0)/5000
p

## [1] 0.3968

df1=data.frame(d)
ggplot(df1,aes(d))+geom_histogram()+geom_histogram(data=subset(df1,d<0),fill=
"grey")+labs(title="Histogram for d")

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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)

```
#Therefore based on the analysis we can say the chance that average price of
Juice for decade 1981-90 was same or less than the decade of 1991-2000 is:-
p

## [1] 0.3968
```