Statistical inference course project

Synopsis

This is the project for the statistical inference class. In it, you will use simulation to explore inference and do some simple inferential data analysis.The project consists of two parts:

1. A simulation exercise.

2. Basic inferential data analysis.

This project investigates the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution is simulated with rexp(n, lambda) where lambda is the rate parameter, theoretical mean of exponential distribution is 1/lambda and theoretical standard deviation is also 1/lambda. This project performs a thousand simulations to get the distribution of averages of 40 exponentials, where the lambda is set to 0.2 for all of the simulations. The simulated samples are used to illustrate and explain the properties of the distribution of the mean of 40 exponentials in the following ways:

1.Show the sample mean and compare it to the theoretical mean of the distribution.

2.Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

3.Show that the distribution is approximately normal

Task

We start by simulating a thousand sets of 40 exponentials using lambda 0.2 and calculate the mean for each set.

R Code:

```
set.seed(123456) # For reproducible results

lambda <- 0.2        # Rate for all simulations

n <- 40                    # No of samples

s <- 1000                 # No of simulations

sim <- data.frame(mean=numeric(s)) # data frame to store all    means
```

```
for (i in 1:s) { # simulate s times
    samples <-rexp(n,lambda) # simulate n samples
    sim[i,1]<-mean(samples) # sample mean
}
```

1. Show the sample mean and compare it to the theoretical mean of the distribution.

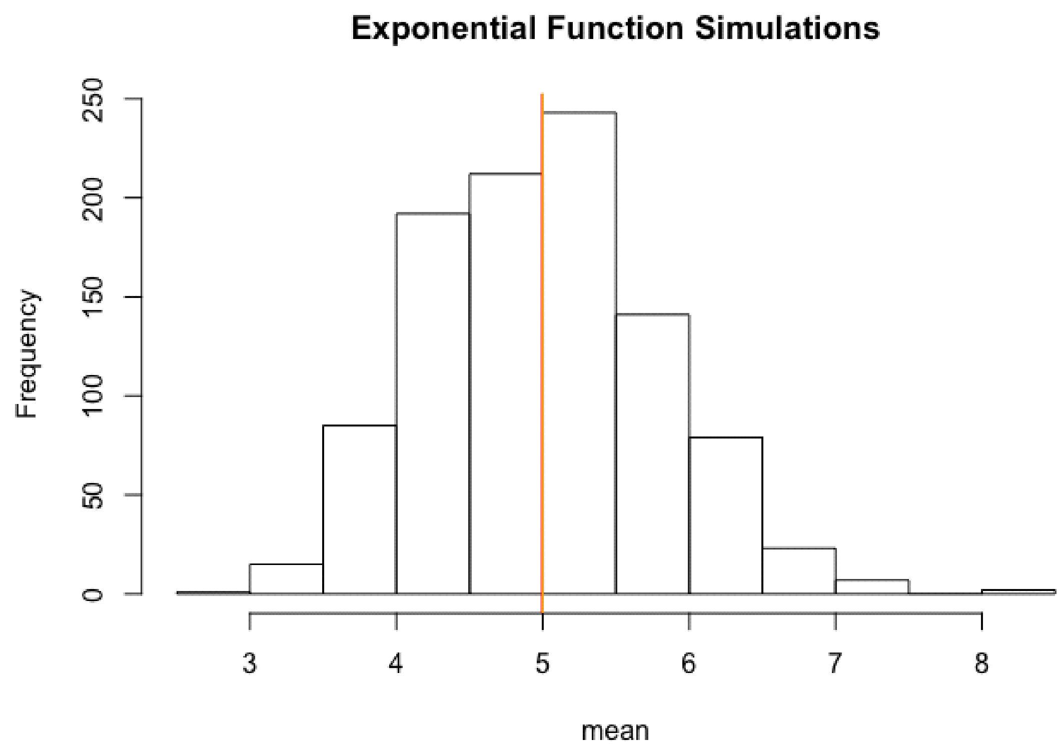
With the simulated data, we calculate the theoretical mean and sample mean for the exponential distribution.

```
tm <- 1/lambda             #    theoretical mean
sm <- mean(sim$mean) # avg sample mean
```

Check the figure below,the sample mean is very close to the theoretical mean at 5:

R Code:

```
hist(means_exponentials, xlab = "mean", main = "Exponential Function Simulations")
abline(v = analytical_mean, col = "red")
abline(v = theory_mean, col = "orange")
```



"Theoretical Mean: 5.00"

"Average Sample Mean: 5.02"

2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

R Code:

```
g1<-((1/lambda)^2)/n     # Theoretical Variance
g1
##     0.625
```

Compute the variance for simulated samples,

R Code:

```
g2<- var(sim$mean)        # Sample variance
g2
```

0.6570391

The result shows that the theoretical variance is very close to the sample variance.

3. Show that the distribution is approximately normal.

When we overlay the histogram of the theoretical mean of the exponential distribution with the histogram of simulated sample means (as shown below), the two distribution curves are very similar and normally distributed. We observe that the histogram for the mean of 1000 simulated 40 random exponential values is symmetric around the mean with a bell shape.

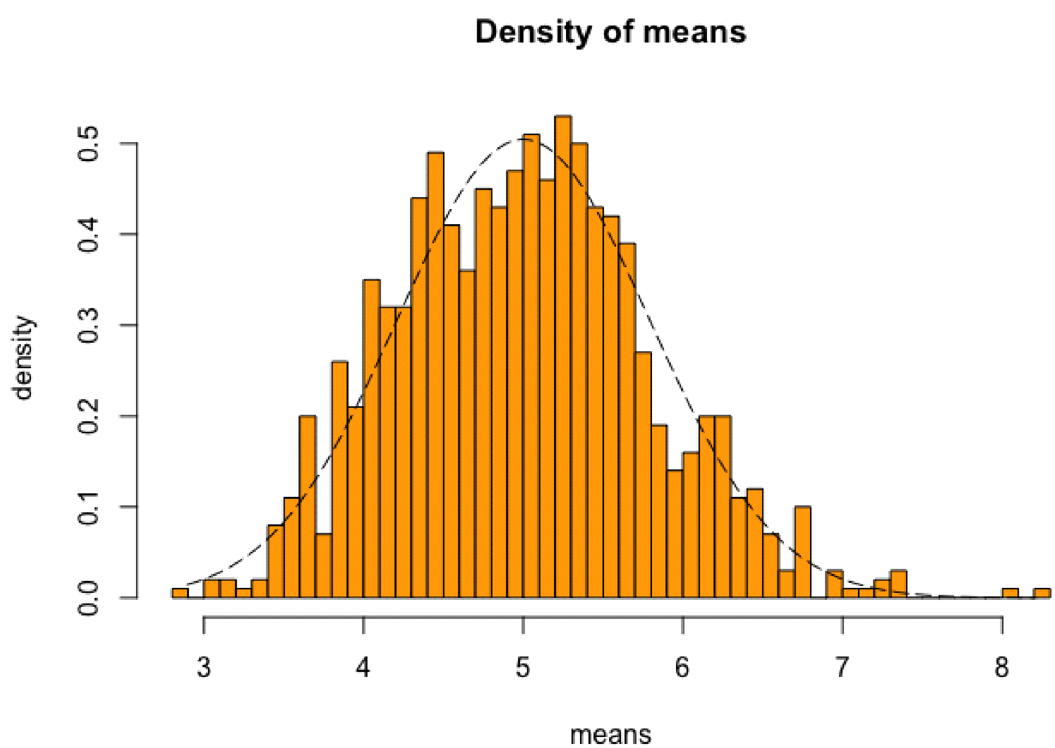
R Code:

xfit <- seq(min(means_exponentials), max(means_exponentials), length=100)

yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))

hist(means_exponentials,breaks=n,prob=T,col="orange",xlab = "means",main="Density of means",ylab="density")

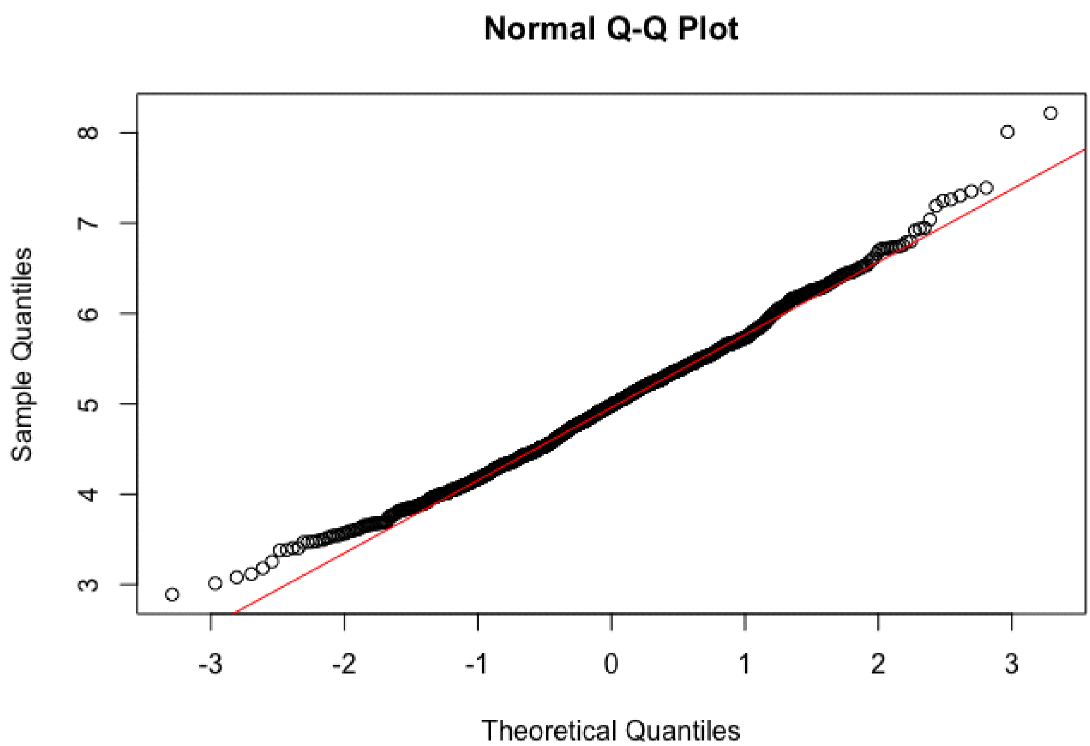
lines(xfit, yfit, pch=22, col="black", lty=5)



R Code:

qqnorm(means_exponentials)

qqline(means_exponentials, col = 2)



Answer:

Due to the central limit theorem (CLT), the distribution of averages of 40 exponentials is very close to a normal distribution.