Statistical Inference Course Project: A simulation exercise

MY

Sunday, October 25, 2015

Assignment Description

Investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should
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

Simulation

The code in the snippet below performs the simulations to collect necessary data.

```
#load the plotting library
library(ggplot2)
#variables that control the simulation
no_sim <- 1000
lambda <- 0.2
n <- 40
set.seed(234)
#Create a matrix of 1000 rows with the columns corresponding to random
simulation 40 times
sim_matrix <- matrix(rexp(no_sim * n, rate=lambda), no_sim, n)
sim_mean <- rowMeans(sim_matrix)
```

The simulation data is now plotted to explore it a bit. See the plot in Appendix Figure 1.

Sample Mean Comparison

The actual mean for the sample data and theoretical mean are calculated below:

```
mean_data <- mean(sim_mean)
theory_mean <- 1/lambda
```

Actual center of the distribution based on the simulations is 5.0015729 while the theoretical mean for lambda = 0.2 is 5. This implies that the actual mean from sample data is very close to the theoretical mean of normal data.

Variance Comparison

The actual variance for the sample data and theoretical variance are calculated below:

```
actual_var <- var(sim_mean)
theory_var <- (1/lambda)^2/n
```

Actual variance for the sample data is 0.6631504 while the theoretical variance is 0.625. Both these values are again quite close to each other.

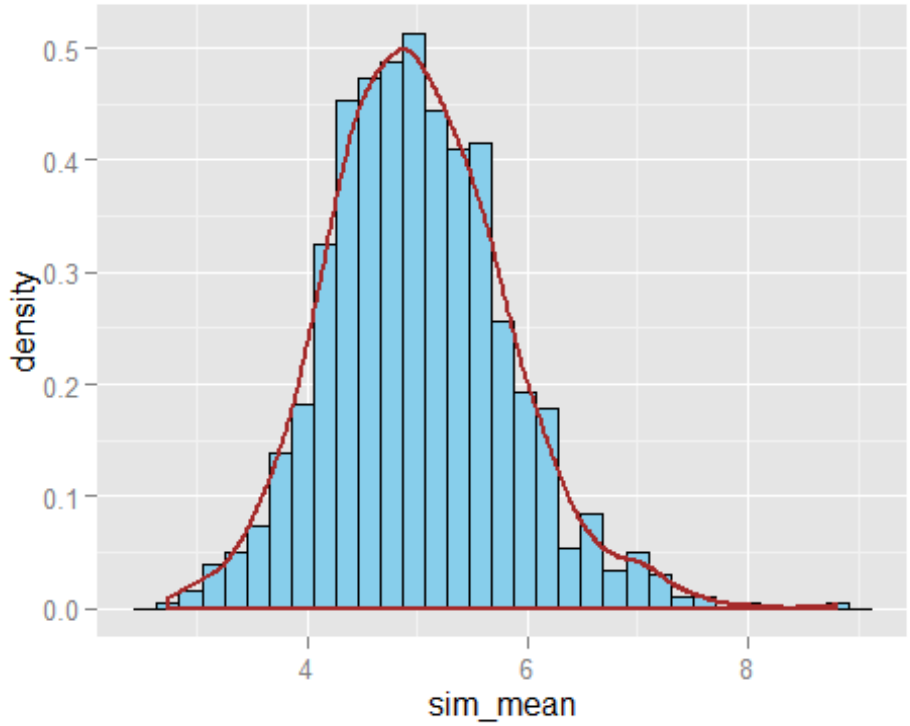
Approximately Normal Distribution

To prove the concept, we use the following three steps: 1. Create an approximate normal distribution and see how the sample data aligns with it. 2. Compare the confidence interval along with the mean and variance with normal distribution. 3. q-q plot for quantiles

Step 1

```
plotdata <- data.frame(sim_mean)
m <- ggplot(plotdata, aes(x =sim_mean))
m <- m + geom_histogram(aes(y=..density..), colour="black",
fill = "sky Blue")
m + geom_density(colour="brown", size=1)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust
this.
```



The figure above shows that the histogram can be adeqautely approximated with the normal distribution.

Step 2

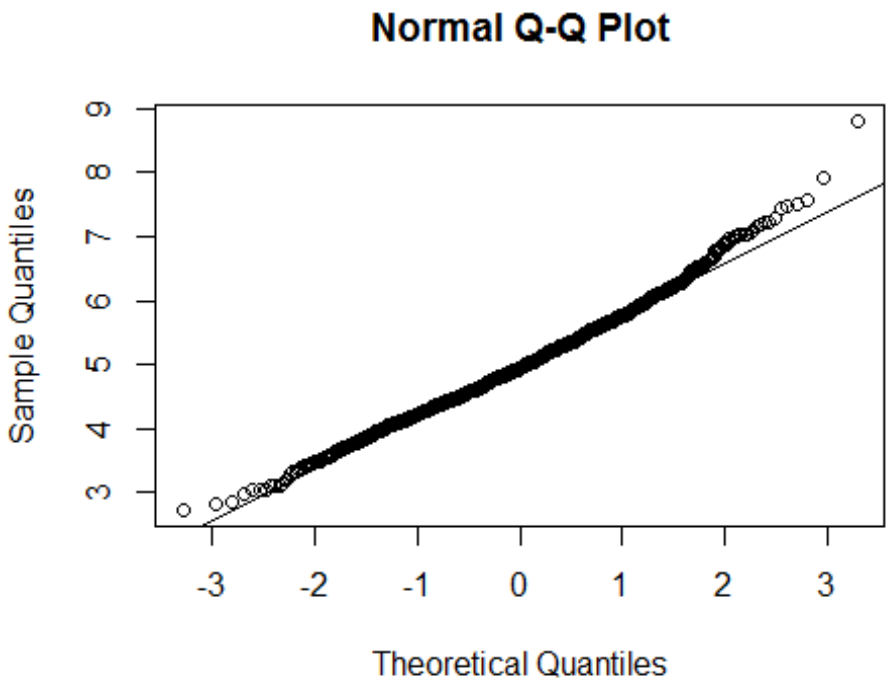
In the sections above, we have proved that the variance and mean of sample data closely resemble those of normal distribution. Lets also try to match the confidence intervals which are calculated below:

```
actual_conf_interval <- round (mean(sim_mean) + c(-
1,1)*1.96*sd(sim_mean)/sqrt(n),3)
theory_conf_interval <- theory_mean + c(-
1,1)*1.96*sqrt(theory_var)/sqrt(n)
```

Actual 95% confidence interval [4.749, 5.254]. Theoretical 95% confidence interval [4.755, 5.245]

Step 3

```
qqnorm(sim_mean)
qqline(sim_mean)
```

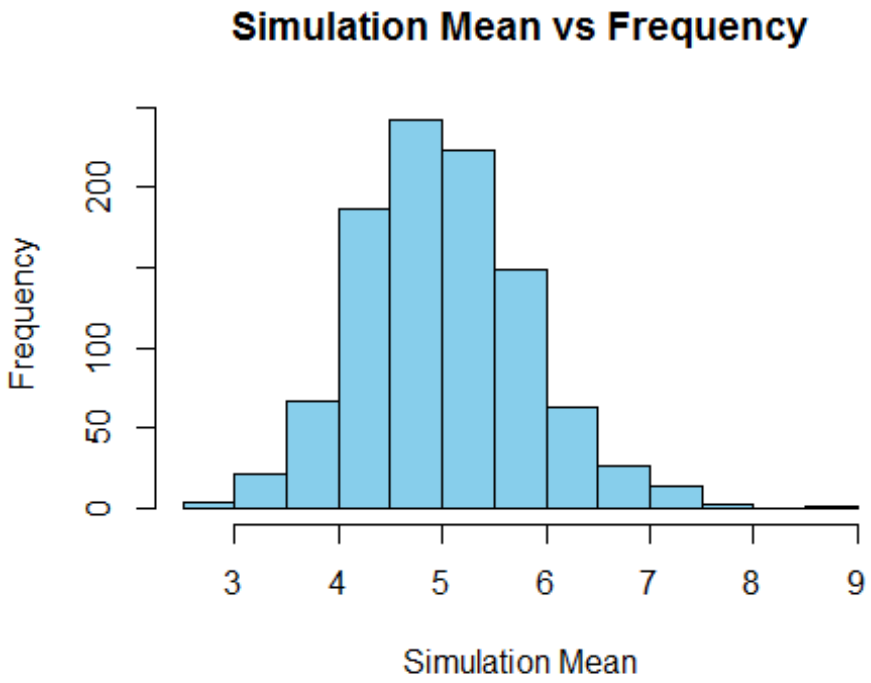


The theoretical quantiles also match closely with the actual quantiles. These three evidences prove that the distribution is approximately normal.

Appendix:

Figure 1: Histogram

```
hist(sim_mean, col = "sky blue",xlab = "Simulation Mean", main =
"Simulation Mean vs Frequency")
```


Simulation Mean vs Frequency
Frequency
0
50
100
200
3
4
5
6
7
8
9
Simulation Mean