Statistical Inference Course Project

Owusu

March 27, 2016

1. Project Overview

In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem (CLT). 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. I will set lambda = 0.2 for all of the simulations. I will investigate the distribution of averages of 40 exponentials. Note that I 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.

In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

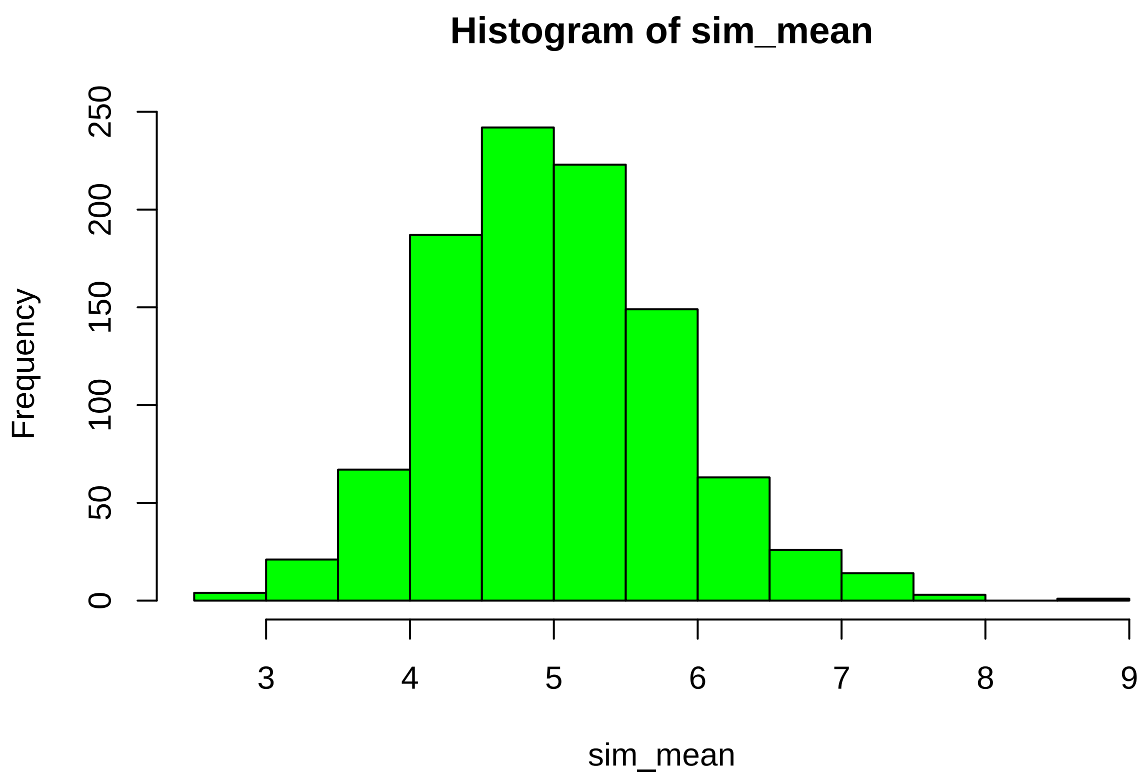
2. Simulations

```r
# Load Labrary
library(ggplot2)
```

```r
# Set of constants
n <- 40  # number of exponentials
lambda <- 0.2  #lambda for rexp
nos <- 1000  # number of stimulation of test

# Set the seed to create reproducible
set.seed(234)

# Run the test to get matrix
sim_matrix <- matrix(rexp(nos * n, rate=lambda), nos, n)
sim_mean <- rowMeans(sim_matrix)
hist(sim_mean, col="green")
```



3. Sample Mean versus Theoretical Mean

The expected mean μ of a exponential distribution of rate λ is $\mu = \frac{1}{\lambda}$

```
mu <- 1/lambda
mu
```

```
## [1] 5
```

Let $\bar{X}$ be the average sample mean of 1000 simulation randomly sampled exponential distributions.

```
mean_data <- mean(sim_mean)
mean_data
```

```
## [1] 5.001573
```

Actual center of the distribution based on the simulations is 5.001573 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.

4. Sample Variance versus Theoretical Variance

The expected standard deviation σ of a exponential distribution of rate λ is $\sigma = \frac{1/\lambda}{\sqrt{n}}$ and the variance Var of standard deviation σ is $Var = \sigma^2$

```
sd <- 1/lambda/sqrt(n)
sd
```

```
## [1] 0.7905694
```

```
Var <- sd^2
Var
```

```
## [1] 0.625
```

Let Var_a be the variance of the average sample mean of 1000 simulations of 40 randomly sampled exponential distribution, and σ_a the represent the standard deviation.

```
sd_a <- sd(sim_mean)
sd_a
```

```
## [1] 0.8143405
```

```
#Var_a <- var(sim_mean)
Var_a <- sd_a^2
Var_a
```

```
## [1] 0.6631504
```

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

5. Approximately Normal Distribution

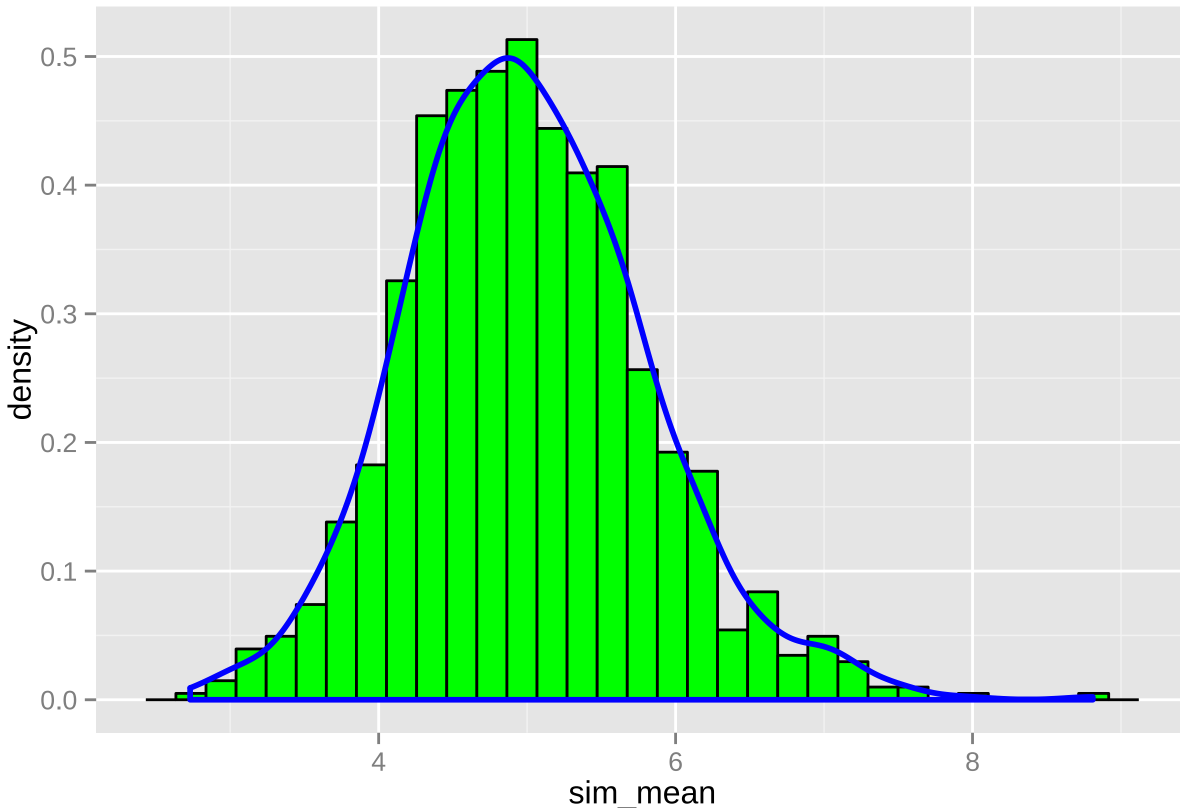
To prove the concept, we use the following three steps:

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);
u <- ggplot(plotdata, aes(x =sim_mean))
u <- u + geom_histogram(aes(y=..density..), colour="black", fill = "green")
u + geom_density(colour="blue", 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_a/sqrt(n),3)
theory_conf_interval <- mu + c(-1,1)*1.96*sqrt(Var)/sqrt(n);
actual_conf_interval
```

```
## [1] 4.749 5.254
```

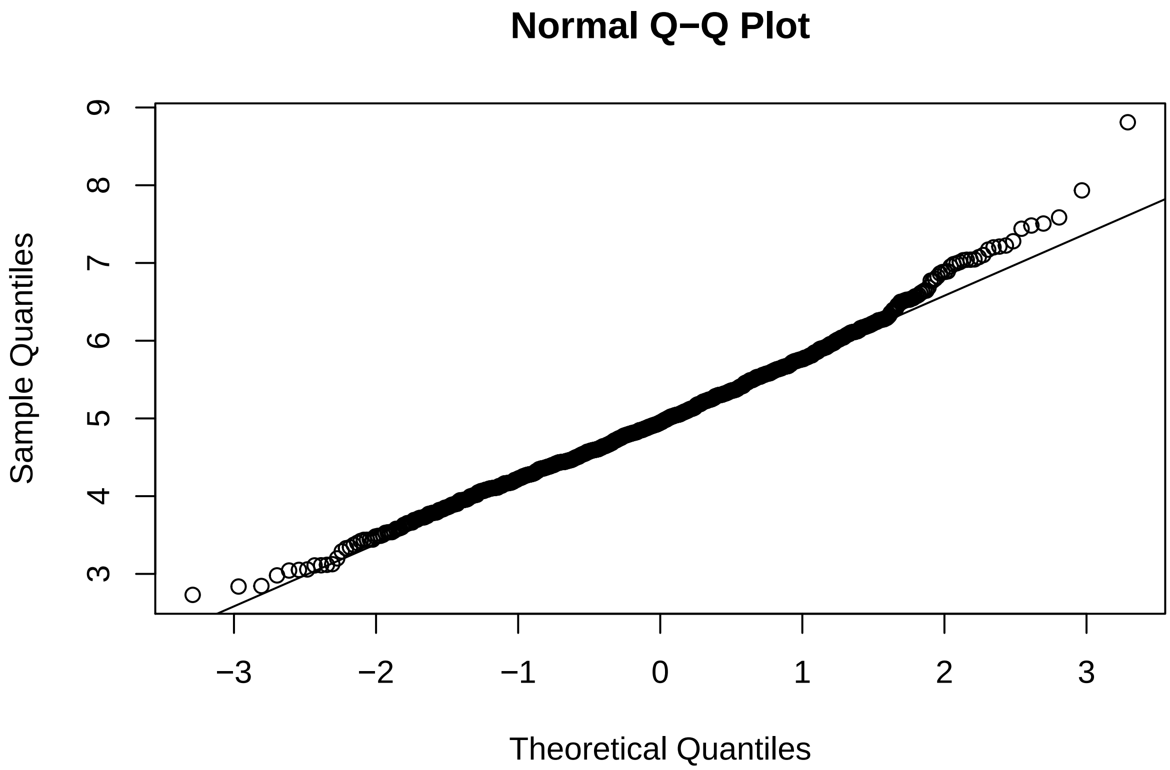
```
theory_conf_interval
```

```
## [1] 4.755 5.245
```

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.