Coursera Satistical Inference - Course Project PART 1

Author: Dino Chioda Date: May 23, 2015

Overview

In this project, which is the first part of the Course Project for the Statistical Inference course, we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. We will demonstrate this by comparing means, variances and standard deviations to show that the distribution is approximately normal.

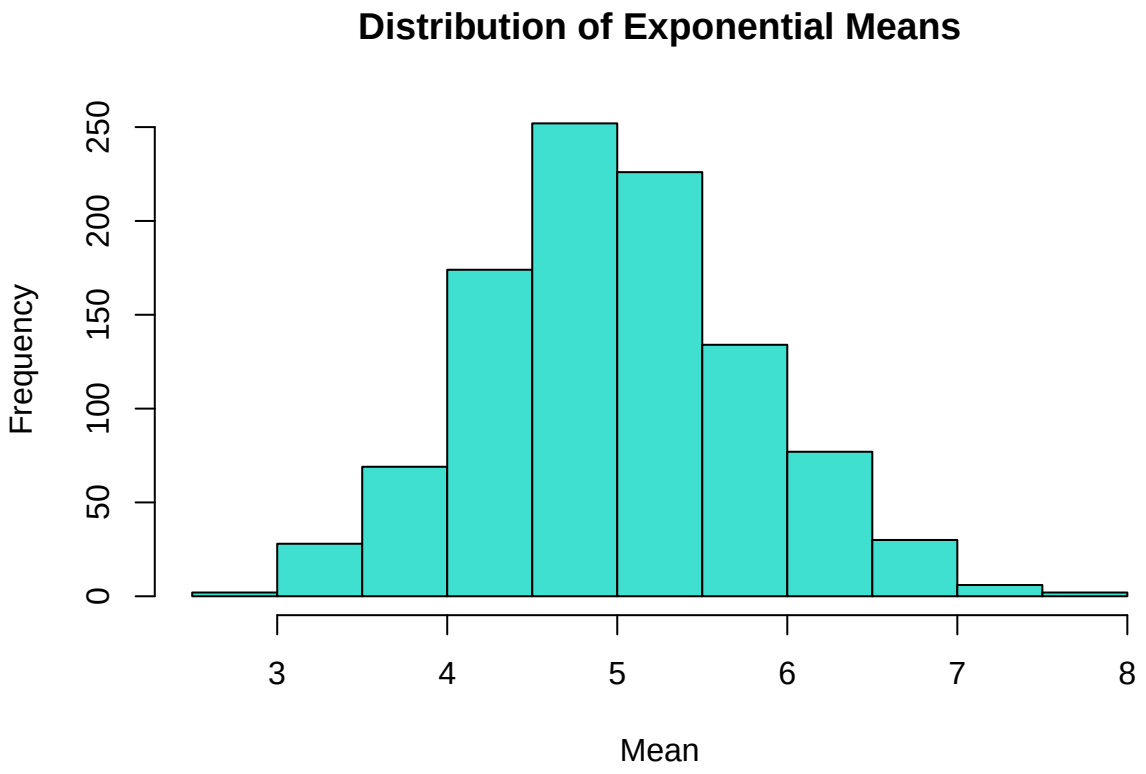
Prepping and Running the Simulations

```
n <- 40
set.seed(1972)
lambda <- .2
sims <- 1000

# Run Simulations
sim_exp <- matrix(rexp(sims*n,rate=lambda),sims)
sim_exp_means <- apply(sim_exp,1,mean)
```

The following histogram shows the distribution of the means of the exponential distributions.

```
hist(sim_exp_means, xlab="Mean", ylab="Frequency", main="Distribution of Exponential Means",
     col="turquoise")
```



Question 1: Show where the distribution is centered and compare it to the theoretical centre of the distribution.

We will begin by calculating the relevant statistics of the distribution. Then we will compare each statistic separately.

```
# Sample statistics
sample_mean <- mean(sim_exp_means)
sample_var <- var(sim_exp_means)
sample_sd <- sd(sim_exp_means)

# Theoretical statistics
theory_mean <- 1/lambda
theory_var <- 1/(lambda^2*n)
theory_sd <- sqrt(theory_var)
```

The sample mean is **4.9892808**. The theoretical mean is **5**. The difference is negligible.

Question 2: Show how variable it is and compare it to the theoretical variance of the distribution.

The sample variance is **0.6531348**. The theoretical variance is **0.625**. The values are very close as well.

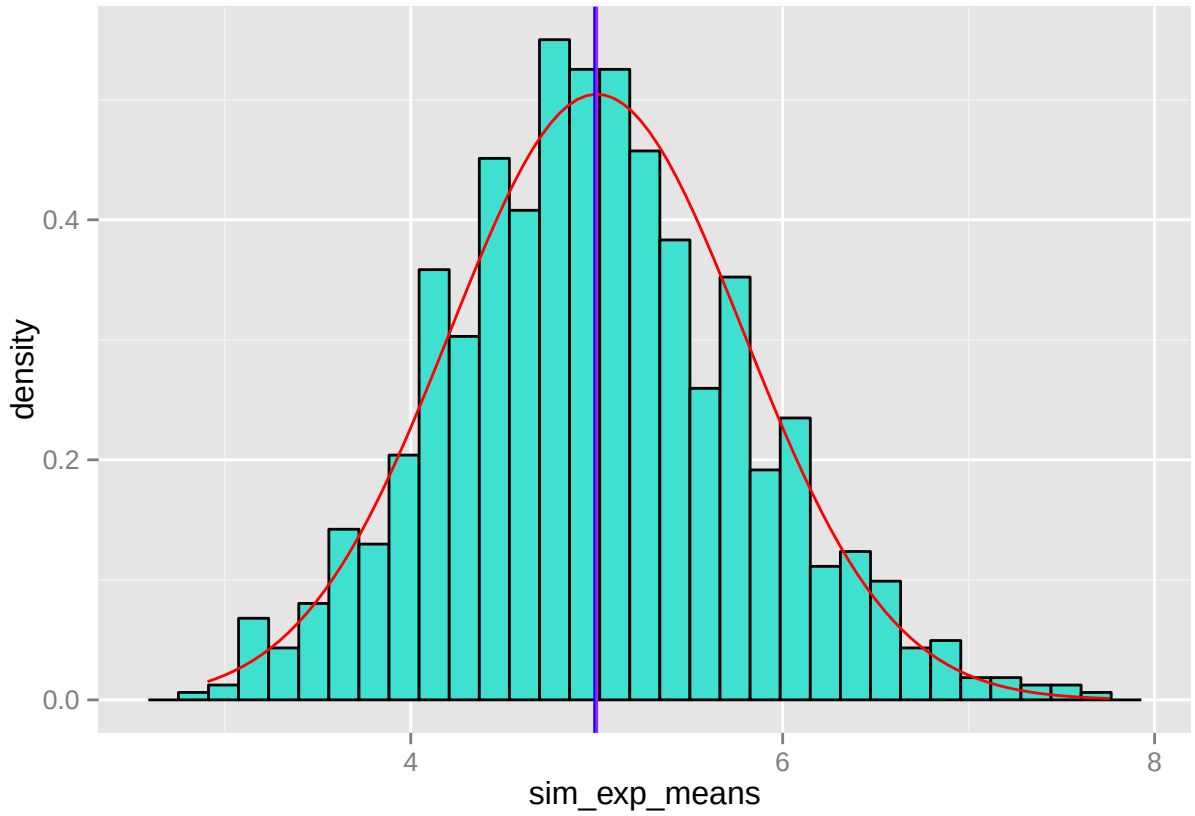
The sample standard deviation is **0.8081676**. The theoretical standard deviation is **0.7905694**. As expected, the values are also very close.

Question 3: Show that the distribution is approximately normal.

We will use 2 tests to validate the normal approximation of the distribution:

1. Plot the histogram of the means and superimpose a normal distribution with theoretical mean and standard deviation previously calculated.

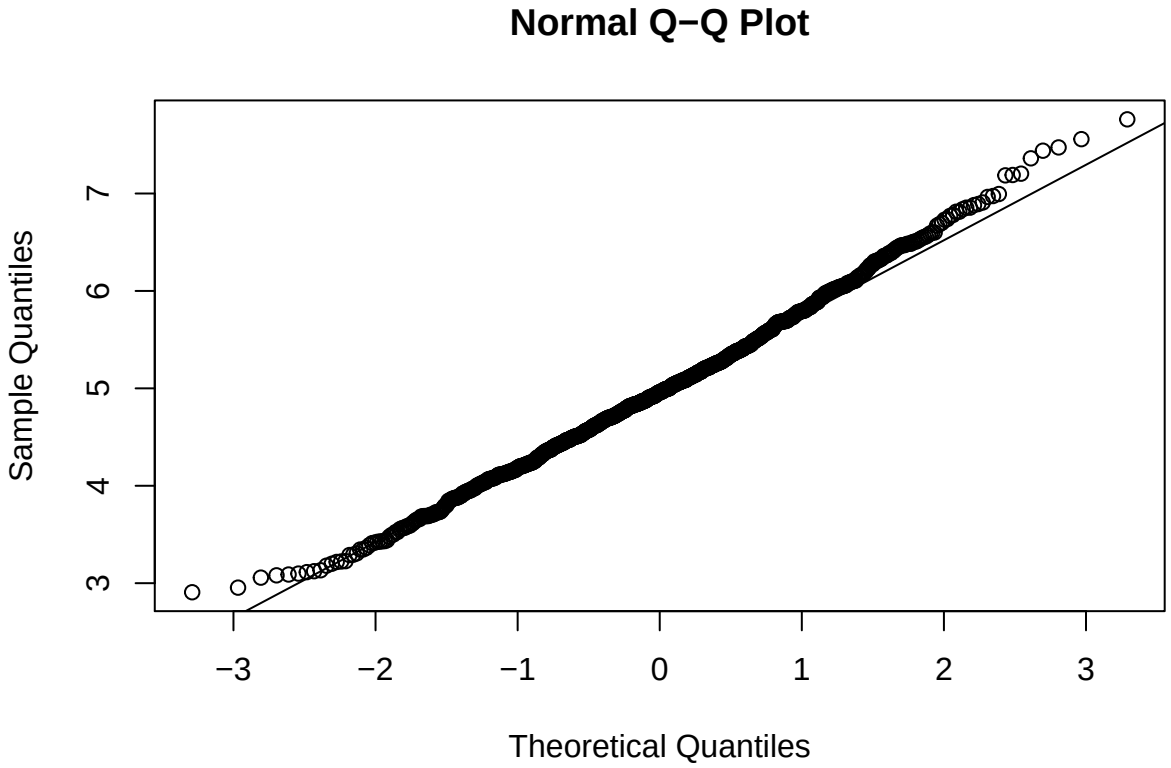
```
library(ggplot2)
compare_plot_data <- data.frame(sim_exp_means)
compare_plot <- ggplot(compare_plot_data, aes(x=sim_exp_means)) +
        geom_histogram(aes(y=..density..), color="black", fill="turquoise") +
        stat_function(fun = dnorm, args = list(mean = theory_mean, sd = theory_sd), color="red") +
        geom_vline(xintercept = sample_mean, color="blue") +
        geom_vline(xintercept = theory_mean, color="purple")
print(compare_plot)
```



As can be seen from the graph, the "fit" of the two distributions is good.

2. Compare the actual quantiles of the distribution to normal quantiles.

```
qqnorm(sim_exp_means)
qqline(sim_exp_means)
```



The quantiles match closely to each other.

Based on these 2 tests, we can say that the normal distribution closely approximates the sample distribution.