Statistical Inference: Course Project Part 1

S. Servaes

17 August 2017

Part 1: Simulation Exercise Instructions

The project investigates the exponential distribution in R and compares it with the Central Limit Theorem. The project illustrates via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials

1. Simulations

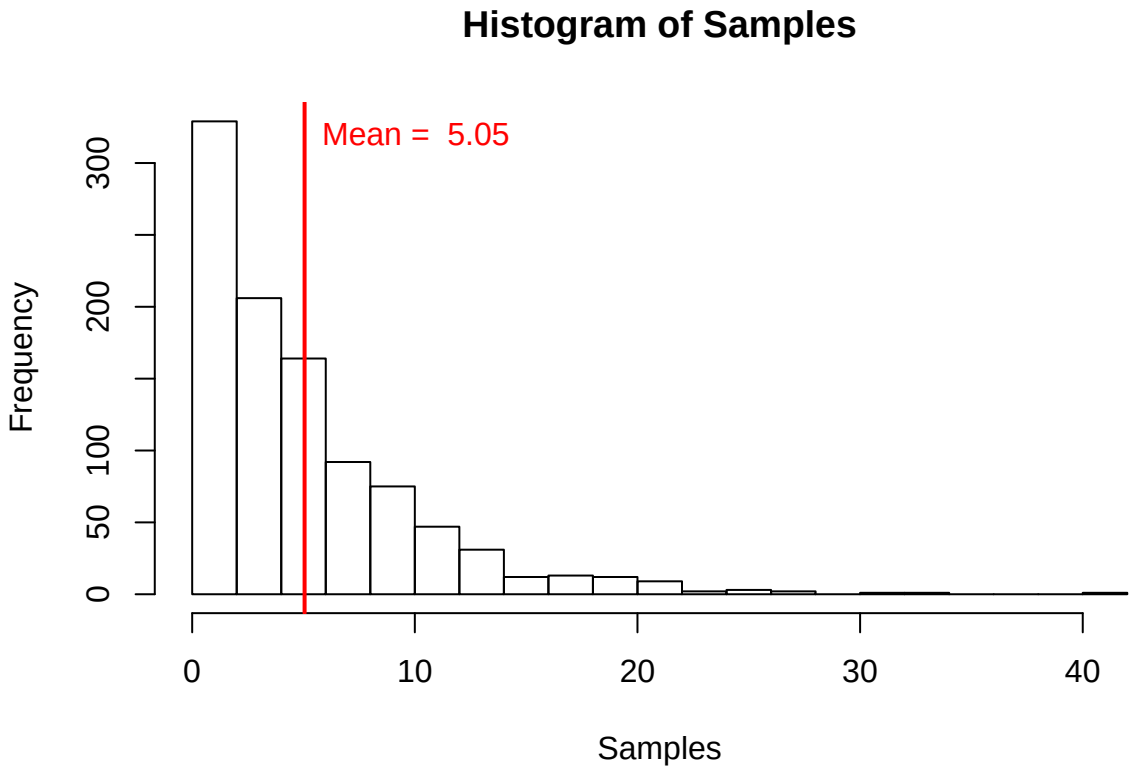
First we will generate the required distributions. One exponential distribution with a lambda of 0.2 containing 1000 samples named `samples` will be generated. Furthermore, a distribution containing 1000 averages of 40 samples from an exponential distribution with a lambda of 0.2, will be generated as well.

```
# Assign the variables
set.seed(2017)
n = 40
lambda = 0.2
mns = NULL

# Simulate 1000 samples of an exponential distribution with lambda
samples <- rexp(1000, lambda)

# Simulate 1000 times the average of 40 samples from an exponential distribution with lambda
for(i in 1:1000){
        mns = c(mns, mean(rexp(n, lambda)))
}
```

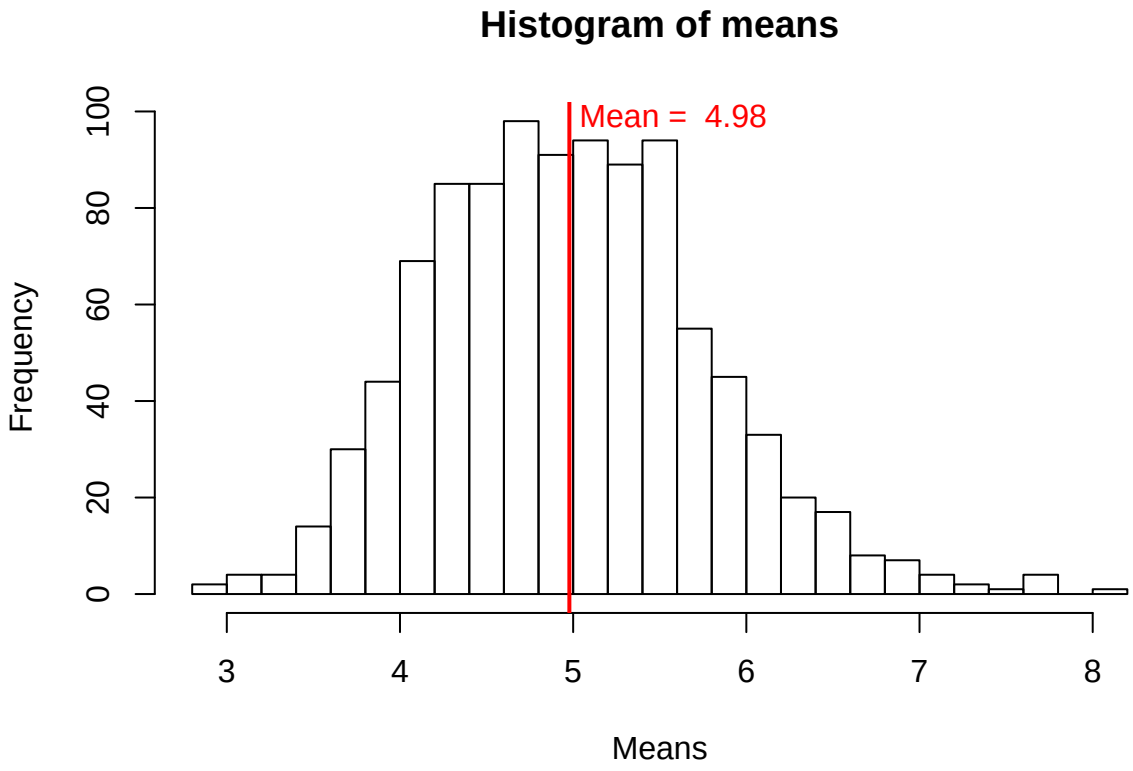
2. Sample Mean versus Theoretical Mean.

In order to answer the first question, the means are calculated for the distributions.
Also a few plots are quickly made

```
# Calculate the means of the 2 distributions
mean_samples <- mean(samples)
mean_mns <- mean(mns)

hist(samples, breaks = 20, main = "Histogram of Samples", xlab = "Samples")
abline(v = mean(samples), col = "red", lwd = "2")
text(mean(samples) + 5, 320, paste("Mean = ", round(mean(samples),2)), col = "red")
```




```
hist(mns, breaks = 20, main = "Histogram of means", xlab = "Means")
abline(v = mean(mns), col = "red", lwd = "2")
text(mean(mns) + 0.6, 99, paste("Mean = ", round(mean(mns),2)), col = "red")
```



The theoretical mean of the samples distribution is 5.
The simulated samples have a mean of 5.05.

The theoretical mean of the means distribution has a mean of 5.
The simulated means distribution has a mean of 4.98.

3. Sample Variance versus Theoretical Variance.

In order to answer the second question, the variances are calculated for the distributions.

```
var_samples <- var(samples)
var_mns <- var(mns)
```

The theoretical variance of the exponential distribution is 25.
The simulated variances have a mean of 25.08.

The theoretical mean of the means distribution has a mean of 0.62.
The simulated means distribution has a mean of 0.62.

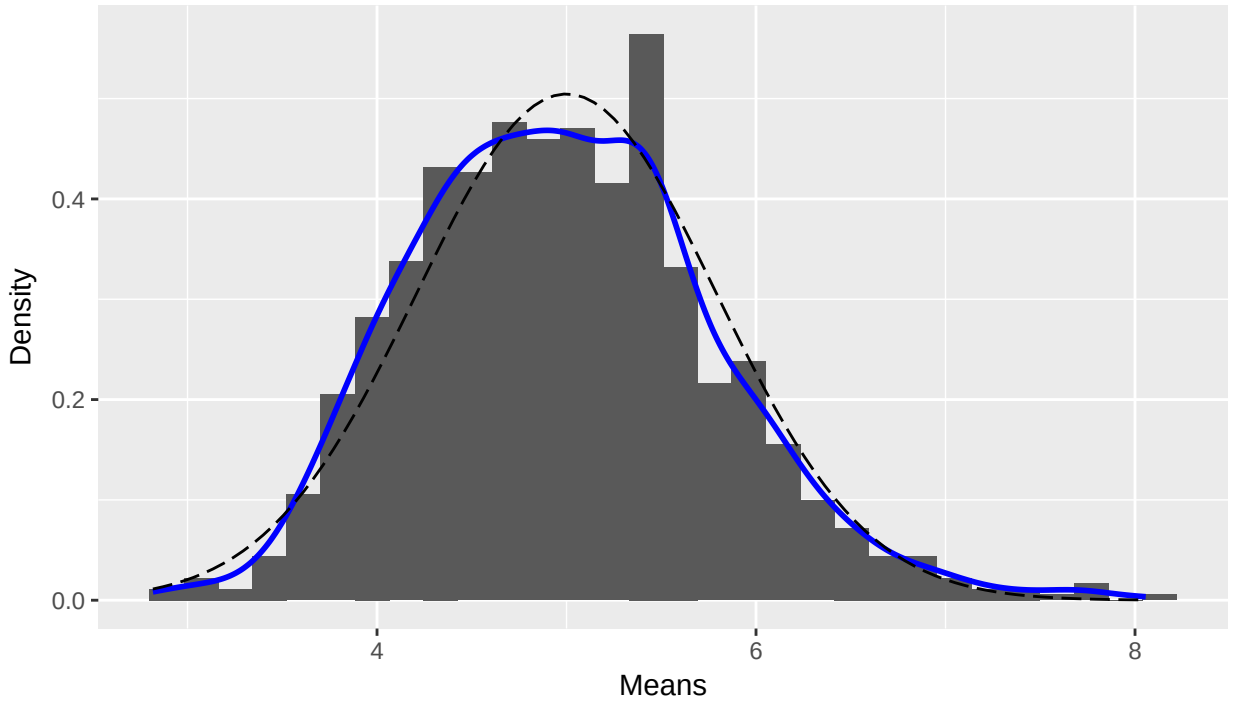
3. Distribution.

The code below generates normal distribution with a similar mean and sd to compare to the distribution of the means. This distribution will be plotted (dashed line) over the fitted distribution of the means (blue line). Furthermore, this code will plot a qqplot of the distribution of the means to assess its normality.

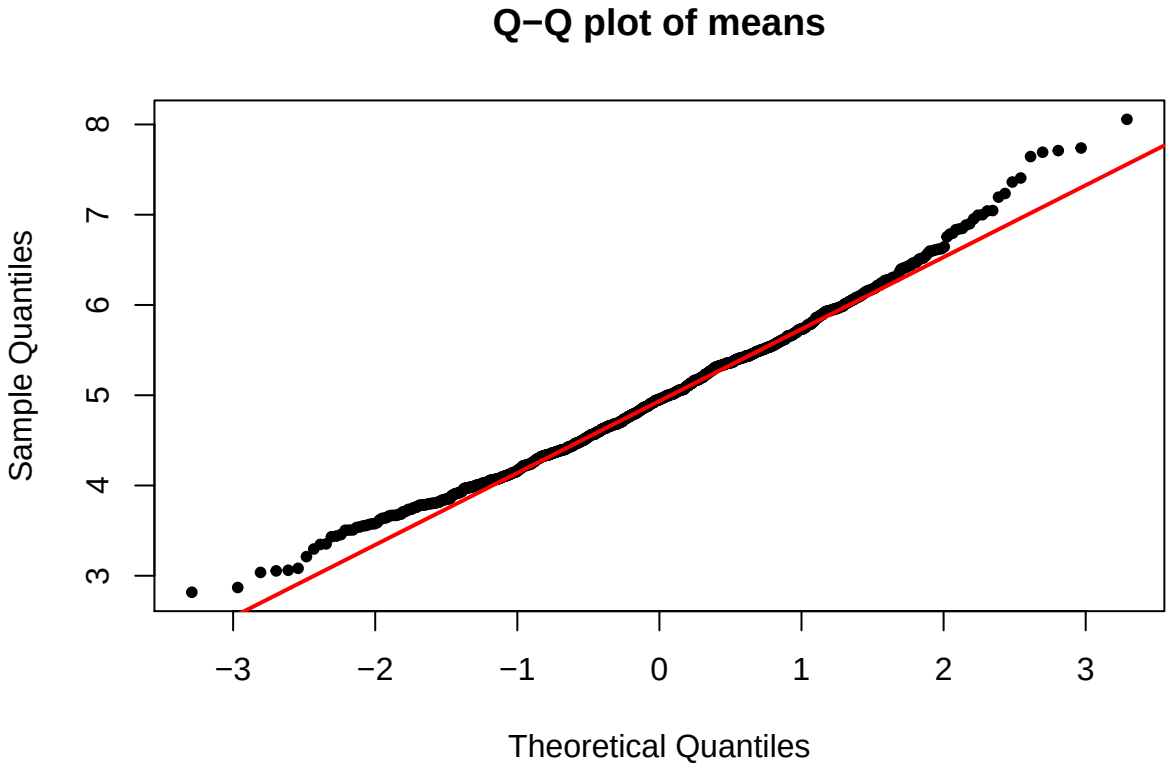
```
# generate a sequence of lenght 100
xfit <- seq(min(mns), max(mns), length=100)

# generate a normal distribution of this sequence
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))

# Make a histogram of the means distribution and compare the fit to the normal distribution
caption <- "This plot shows the density of the means in 30 different bins. The blue line represents the
caption <- paste(strwrap(caption, 100), sep="", collapse=" \n ")

ggplot(data = as.data.frame(mns), aes(mns)) +
        geom_histogram(bins = 30, aes(y = ..density..)) +
        labs(title = "Density of means", subtitle = caption) +
        xlab("Means") +
        ylab("Density") +
        geom_line(stat = "density", aes(mns), col = "blue", size = 1) +
        geom_line(data = as.data.frame(xfit), aes(xfit, yfit), linetype = 5)
```

Density of means

This plot shows the density of the means in 30 different bins. The blue line represents the density, while the black dashed line shows how a perfect normal distribution with mean 1/lambda an standard deviation of 1/lambda/sqrt(n) would look like.




```
# Generate a qq plot
qqnorm(mns, pch = 20, main = "Q-Q plot of means")
qqline(mns, col = "red", lwd = "2")
```



From the histograms and the qqplot it is clearly visible that the distribution of means has a gaussian distribution that approaches a normal distribution. This is to be expected because of the Central Limit Theorem stating that, in most situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (a bell curve) even if the original variables themselves are not normally distributed.