Exploring the Central Limit Theorem (CLT)

Chiara Di Gravio

15 August 2016

Overview

The distribution of averages of 40 exponentials (with rate parameter $\lambda = 0.2$) was investigated to better understand how the CLT works in practice. 1,000 simulations were run and the results were compared to the ones expected given the CLT.

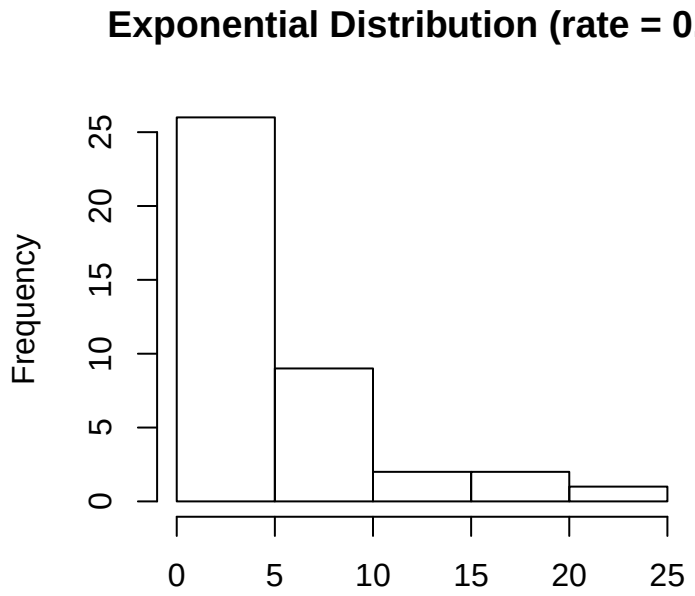
Specifically, we expected the distribution of averages of exponentials to be normally distributed with mean = $1/\lambda$ and standard deviation = $\sqrt{n}/\lambda$.

Simulations

First, we set the parameters for the simulation (rate parameter λ, numbers of exponentials and simulations, seed for reproducibilty) and we plotted an exponential distribution with $\lambda = 0.2$. The plot of the exponential distribution showed our starting point.

```
# rate parameter
lambda <- 0.2
# number of exponential
n <- 40
# numer of simulations
nrepl <- 1000
# set seed for reproducibility
set.seed(341)

hist(rexp(n, lambda), main = "Exponential Distribution (rate = 0.2)", xlab = "")
```



Then, we ran the simulation. 40 exponential distributions were generated and the mean was computed. This process was repeated for 1,000 times:

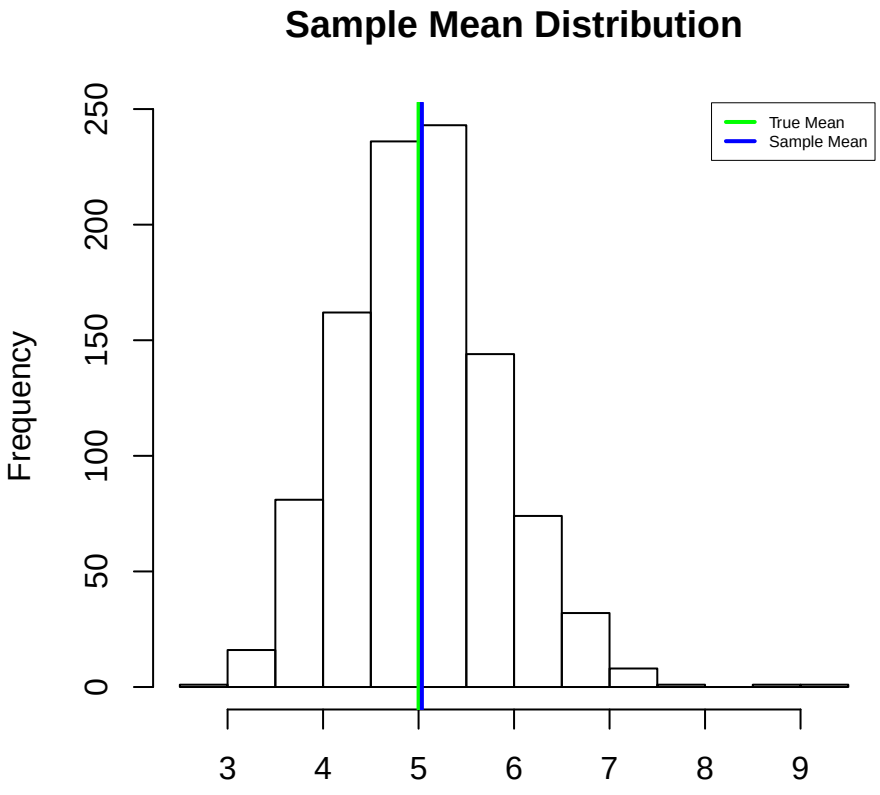
```
# initialise vector of means
meanexp <- c()
for(i in 1:nrepl){
      # take random number from exponential
      randomexp <- rexp(n, lambda)
      # compute mean
      meanexp[i] <- mean(randomexp)
}

# sample mean
mean(meanexp)
# theoretical mean
1/lambda
```

Sample vs Theoretical Mean

The sample mean was 5.03, while the theoretical mean was 5. The following histogram plots the distribution of averages of exponentials and better summarises how close the sample (green line) and the theoretical mean (blue line) are:

```
# plot the mean
hist(meanexp, main = "Sample Mean Distribution", xlab = "")
abline(v = 1/lambda, col = "green", lwd = 2)
abline(v = mean(meanexp), col = "blue", lwd = 2)
legend("topright", c("True Mean", "Sample Mean"), col = c("green", "blue"),
       lwd = c(2,2),box.lwd = 0, cex = 0.5)
```



Sample vs Theoretical Variance

As expected from the CLT, sample variance and theoretical variance are also quite close (difference: 0.03 without considering possible variability due to the simulation):

```
# sample and theoretical variance
variances <- round(c(var(meanexp),((1/lambda)/sqrt(n))^2),2)
names(variances) <- c("Sample Variance", "Theoretical Variance")
kable(variances)
```

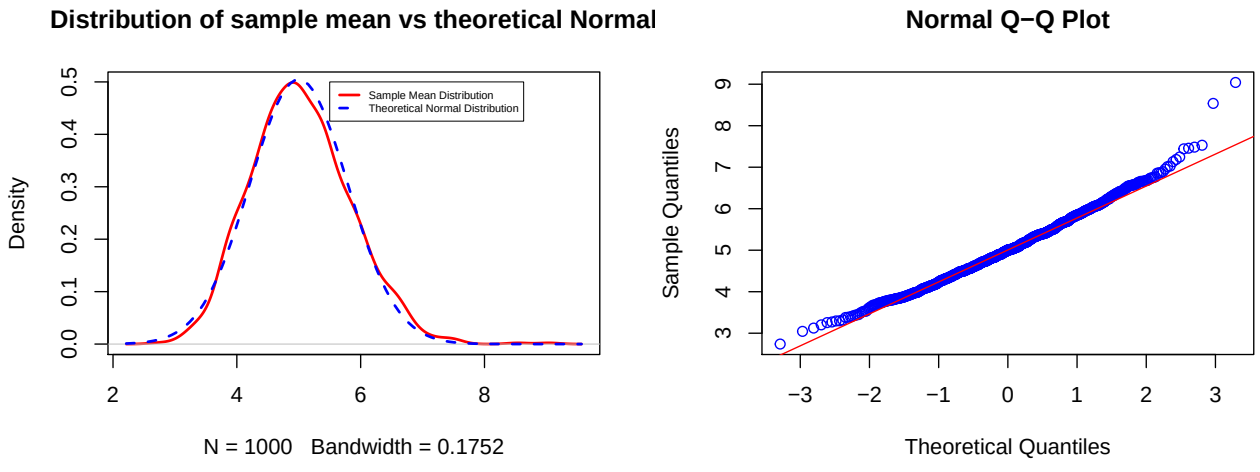
Sample Variance	0.65
Theoretical Variance	0.62

Distributions

The exponential distribution (Figure 1) had a right skewed distribution, whereas the distibution of the average of exponentials (Figure 2) was symmetric and centered at 5.03. The distributions of average of exponentials (blue) and the one of a Normal distribution (red) were compared using both a density plot and a QQplot.

```
# graphically check normality

par(mfrow = c(1,2))
# plot density
plot(density(meanexp), lwd=2, col = "red",
     main = "Distribution of sample mean vs theoretical Normal")
curve(dnorm(x, mean = 1/lambda, sd = ((1/lambda)/sqrt(n))), col="blue", lwd=2,
      add=TRUE, lty = 2)
legend(x = 5.5, y = 0.5, c("Sample Mean Distribution",
                           "Theoretical Normal Distribution"), col = c("red", "blue"),
       lwd = c(2,2), lty = c(1, 2),box.lwd = 0, cex = 0.5)
# QQplot
qqnorm(meanexp, col = "blue")
qqline(meanexp, col = "red")
```



The blue and the red densities were quite similar with the biggest differences observed in the tails of the distribution.