# Statistical Inference Course Project Part 1

**Johns Hopkins University - Coursera**

**by Massimo Malandra**

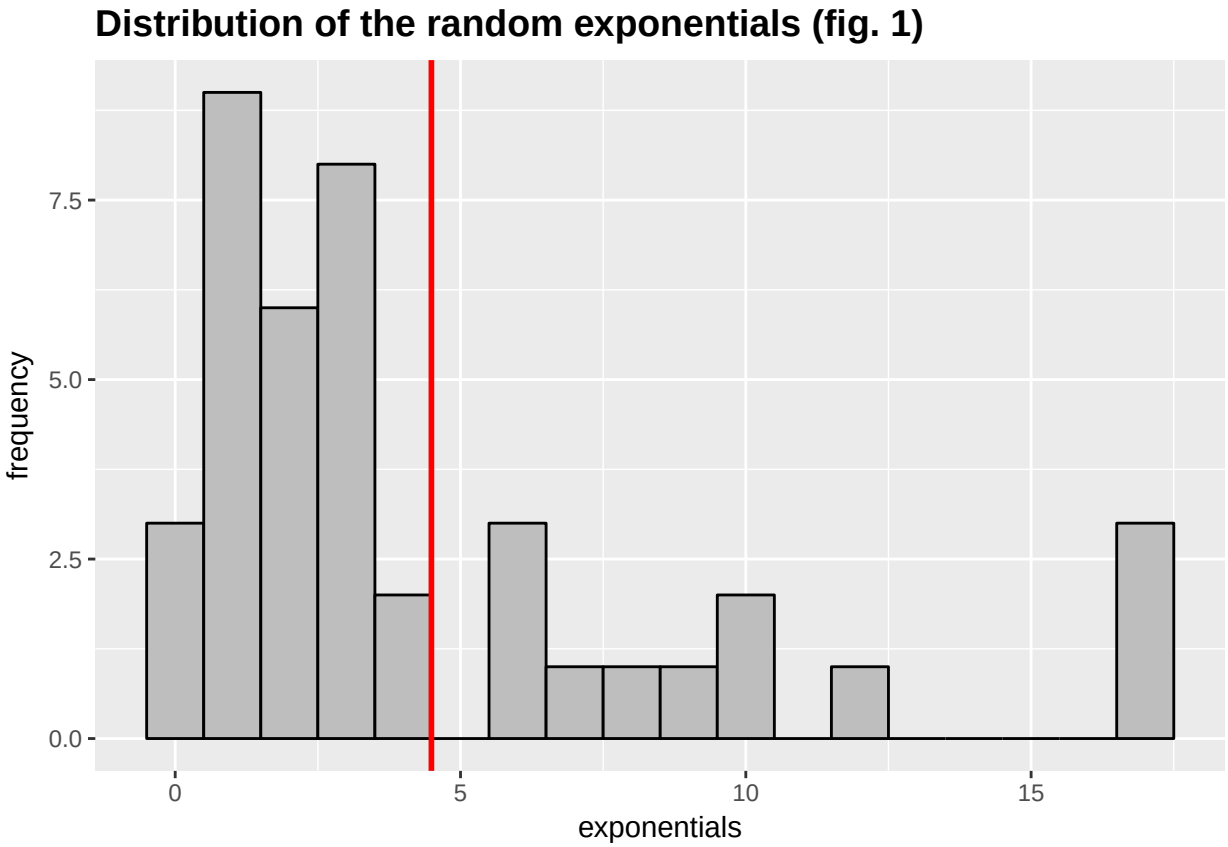
---

**Overview:**

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We will simulate the exponential distribution using the rexp(n, lambda) function in R, where lambda represents the rate parameter. The mean and the standard deviation of exponential distribution have the same value which is 1/lamba. We will set lambda = 0.2 for all the simulations and investigate the distribution of averages of 40 exponentials, with 1000 simulations.

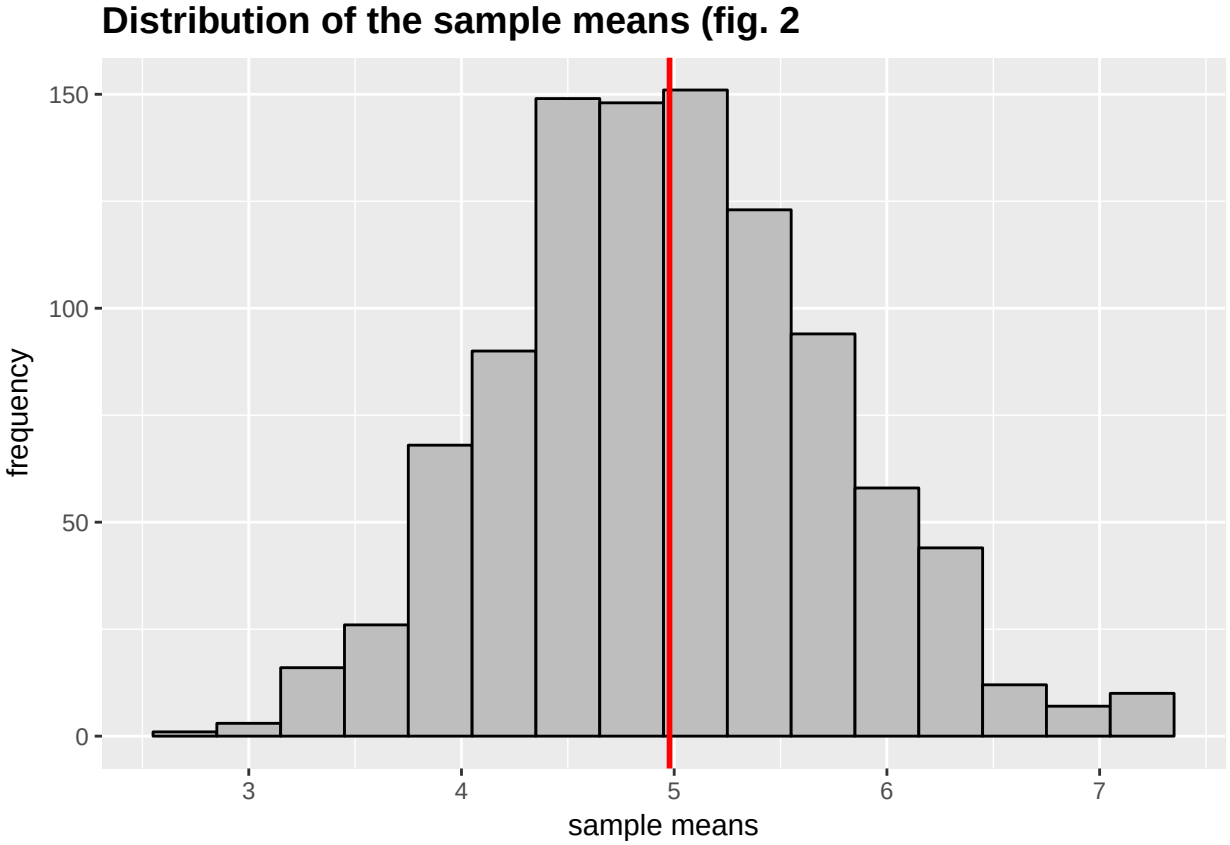
**Simulations:**

The histogram above (fig. 1) represents the 1000 simulations of the random exponentials, generated with the rexp(n, lambda) function in R, with lambda set to 0.2. As indicated by the red vertical line, the mean is approximately 4.49.




The figure above (n.2) represents the distribution of the means of the sample of random exponentials: as it can be easily seen, the shape is quite different from the one in fig. 1 (which approximately follows the typical

distribution of an exponential) and it is very close to a normal distribution. This is in line with what the Central Limit Theorem states.




**Sample Mean versus Theoretical Mean:**

Also, it can be noted that the mean of the sample means distribution is equal to 4.98 (see red vertical line), very close to the mean of the randome exponentials distribution - another affirmation of the CLT. Both means are very close to the theoretical mean of the exponential distribution, which is 1/lambda, and so – for lambda = 0.2 – it is equal to 5.

```
## Theoretical mean: 5
```

```
## Mean of the sample means: 4.978
```

```
## Difference between means: -0.022
```

**Sample Variance versus Theoretical Variance:**

Regarding the variamce, it can be noted that the theoretical variance for the exponential distribution is (1/lambda)^2, hence - for lambda = 0.2 - equal to 25. We just saw that the Central Limit Theorem applies for the mean, so we would expect that the variance of the sample of the means will be equal to the variance divided by the number of the observations in the exponentials sample - which is 40: hence the variance we expect according to the CLT is 25/40=0.625. The calculation of the actual variance shows that it is equal to 0.587 confirming what expected from the CLT.

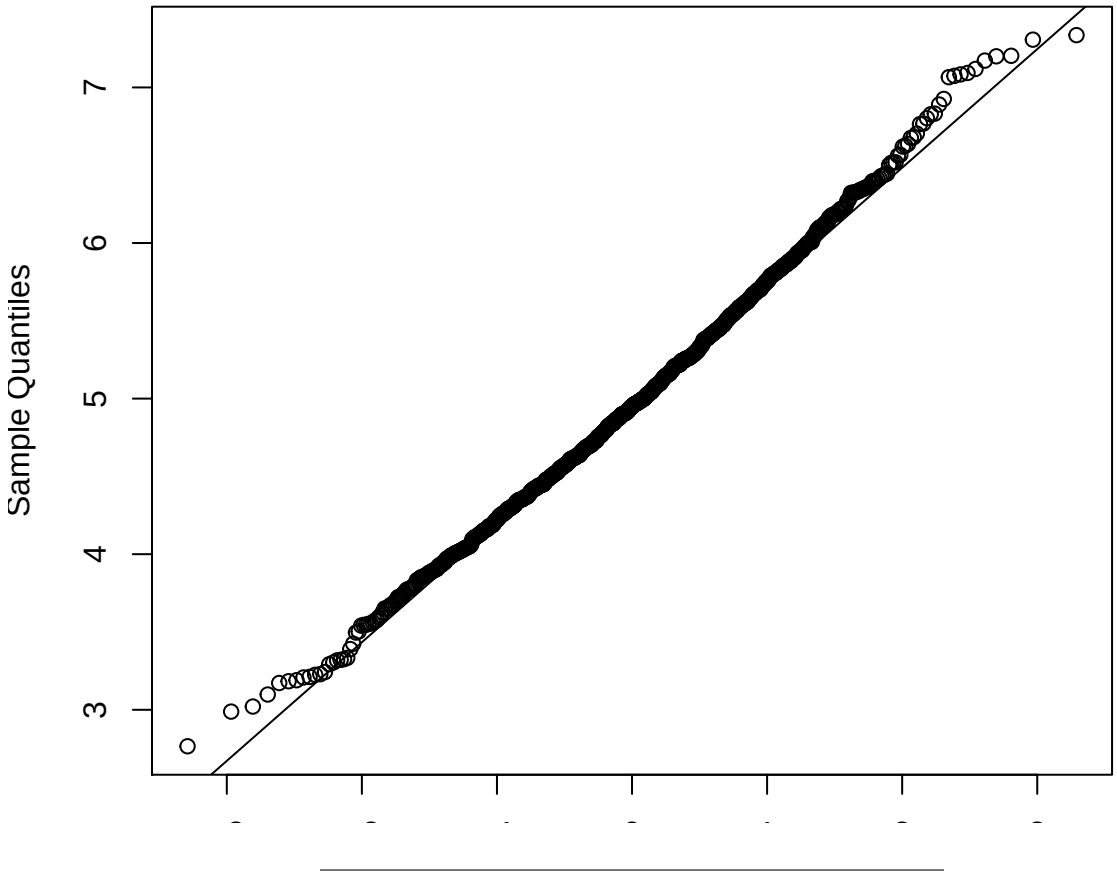
```
## Expected variance: 0.625
```

```
## Actual variance: 0.592
```

```
## Difference between variances: -0.033
```

**Distribution:**

As the above Q-Q plot suggests, we can confirm that the distribution of the sample means is approximately normal, given that the points are located along the normal line.




**Appendix:**

**R code:**

```
set.seed(137)
lambda <- 0.2
n <- 40
sim <- 1000
exp <- rexp(n, lambda)
expDf <- as.data.frame(exp)
meanExp <- mean(exp)
library(ggplot2)

# fig. 1
ggplot(expDf, aes(x = exp)) +
    geom_histogram(col="black", fill="grey", binwidth=1) +
    geom_vline(aes(xintercept=mean(expDf$exp)),color="red", size=1) +
    ggtitle("Distribution of the random exponentials (fig. 1)") +
    labs(x="exponentials", y="frequency") +
    theme(plot.title=element_text(size=14, face="bold"))
```

```r
means <- NULL
for (i in 1:sim) means = c(means, mean(rexp(n, lambda)))
meansDf <- as.data.frame(means)
meanMeans <- mean(means)

# fig.2
ggplot(meansDf, aes(x = means)) +
    geom_histogram(col="black", fill="grey", binwidth=0.3) +
    geom_vline(aes(xintercept=mean(meansDf$means)), color="red", size=1) +
    ggtitle("Distribution of the sample means (fig. 2") +
    labs(x="sample means", y="frequency") +
    theme(plot.title=element_text(size=14, face="bold"))

theoreticalMean <- 1/lambda
cat(c("Theoretical mean:", round(theoreticalMean, 3)))
cat(c("Mean of the sample means:", round(meanMeans, 3)))
meanDifference <- meanMeans - theoreticalMean
cat(c("Difference between means:", round(meanDifference, 3)))

expStdv <- 1/lambda
expVar <- expStdv^2
theoreticalVar <- expVar/n
cat(c("Expected variance:", round(theoreticalVar, 3)))
varMeans <- var(means)
cat(c("Actual variance:", round(varMeans, 3)))
varDifference <- varMeans - theoreticalVar
cat(c("Difference between variances:", round(varDifference, 3)))

# fig. 3
par(pin=c(5,4))
qqnorm(means)
qqline(means)
```