Statistical Inference Course Project Part 1

by Szymon Tomczyk

```
library(tidyverse)
library(datasets)
```

Part 1: Simulation Exercise

In this part we are supposed to generate 1000 averages of 40 random exponentials from a continous exponential distribution with rate lambda = 0.2. We will use this data to test the assumptions of LLN and CLT in practice.

Simulation

Declare the simulation parameters

```
set.seed(666)
n <- 1000
lambda <- 0.2
```

Simulated 1000 averages of 40 exponentials from the expponential distribution with the rate lambda = 0.2

```
s.mean = NULL
for (i in 1 : n) { s.mean <- c(s.mean, mean(rexp(40, lambda))) }

head(s.mean)
```

```
## [1] 4.267297 3.640230 5.646912 4.717544 4.463866 4.974047
```

Theoretical statistics vs. sample statistics

Calculate the theoretical mean and sample mean

```
1/lambda ## Theoretical mean
```

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

```
mean(s.mean) ## Sample mean
```

```
## [1] 4.987818
```

Calculate the theoretical variance and sample variance

```
(1/lambda/sqrt(40))^2 # Theoretical variance
```

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

```
var(s.mean) # Sample variance
```

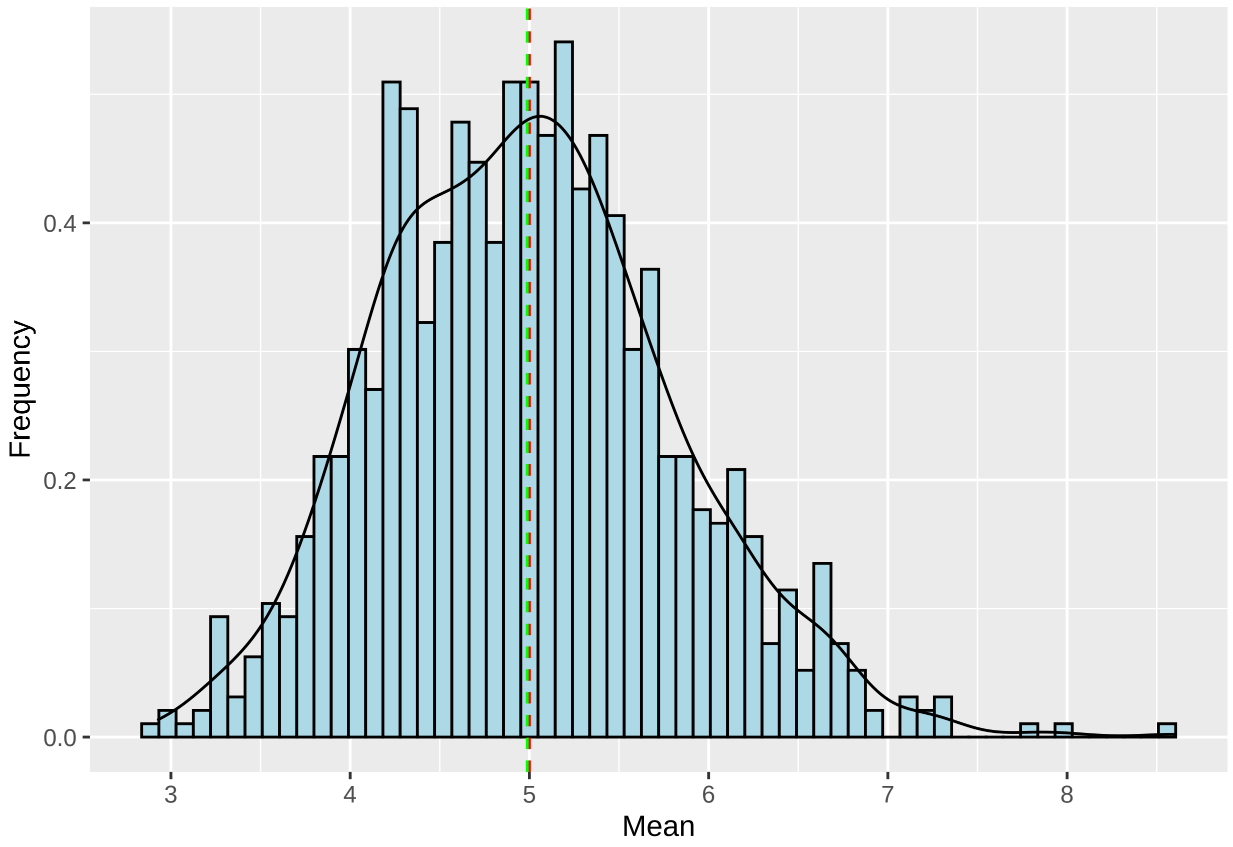
```
## [1] 0.6646822
```

Conclusion: both sample mean and variance are very close to the predicted theoretical parameters. This result is in accordance with Law of Large Numbers

Normality of the sampling distribution of the sampling mean

Plot the histogram of the sampling distribution of the mean. The red dashed line corresponds to the theoretical mean and the grean one to the sample mean.

```
s.mean <- as.data.frame(s.mean)
plot <- ggplot(s.mean, aes(s.mean)) +
        geom_histogram(aes(y=..density..), bins = 60,
                       colour = "black", fill = "lightblue")+
        geom_density()+
        geom_vline(aes(xintercept=c(1/lambda)),
                   color="red", linetype="dashed",
                   size = 0.5)+
        geom_vline(aes(xintercept=mean(s.mean)),
                   color="green", linetype="dashed",
                   size = 0.5)+
        xlab("Mean") + ylab("Frequency")
plot
```



Test with Shapiro–Wilk test for normality

```r
shapiro.test(s.mean$s.mean)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  s.mean$s.mean
## W = 0.99187, p-value = 2.584e-05
```

Conclusion: Under the CLT we can assume that the sampling distribution of the mean is approximately normal. However, our specific sample does not pass Shapiro–Wilk test so the normality assumtion is not valid.