Statistical Inference Course Project - Part 1

Ekaterina Abramova

21 January 2017

A simulation exercise

In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda. Set lambda = 0.2 for all of the simulations. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials

1.Show the sample mean and compare it to the theoretical mean of the distribution.

We will create a simulation sample called simulated_samp each of size 40 drawn from exponential distribution with rate parameter $\lambda = 0.2$. Seed will be set at 333 to make the simulation reproducible.

```
library(ggplot2)
# set seed for reproducability
set.seed(333)
# set lambda to 0.2
lambda <- 0.2
# 40 samples
n <- 40
# 1000 simulations
n.sim <- 1000
# simulate
simulated_samp <- matrix(rexp(n*n.sim, rate = lambda), nrow = n.sim, ncol = n)
# calculate mean of exponentials
simulated_samp_mean <- apply(simulated_samp, 1, mean)
mean(simulated_samp_mean)
```

```
## [1] 5.045337
```

```
# analytical mean
theory_mean <- 1/lambda
theory_mean
```

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

The sample mean is 5.0453, which is close to the theoretical mean, $\mu=1/\lambda=5$

2.Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.

```r
var(simulated_samp_mean)
```

```
## [1] 0.6260539
```

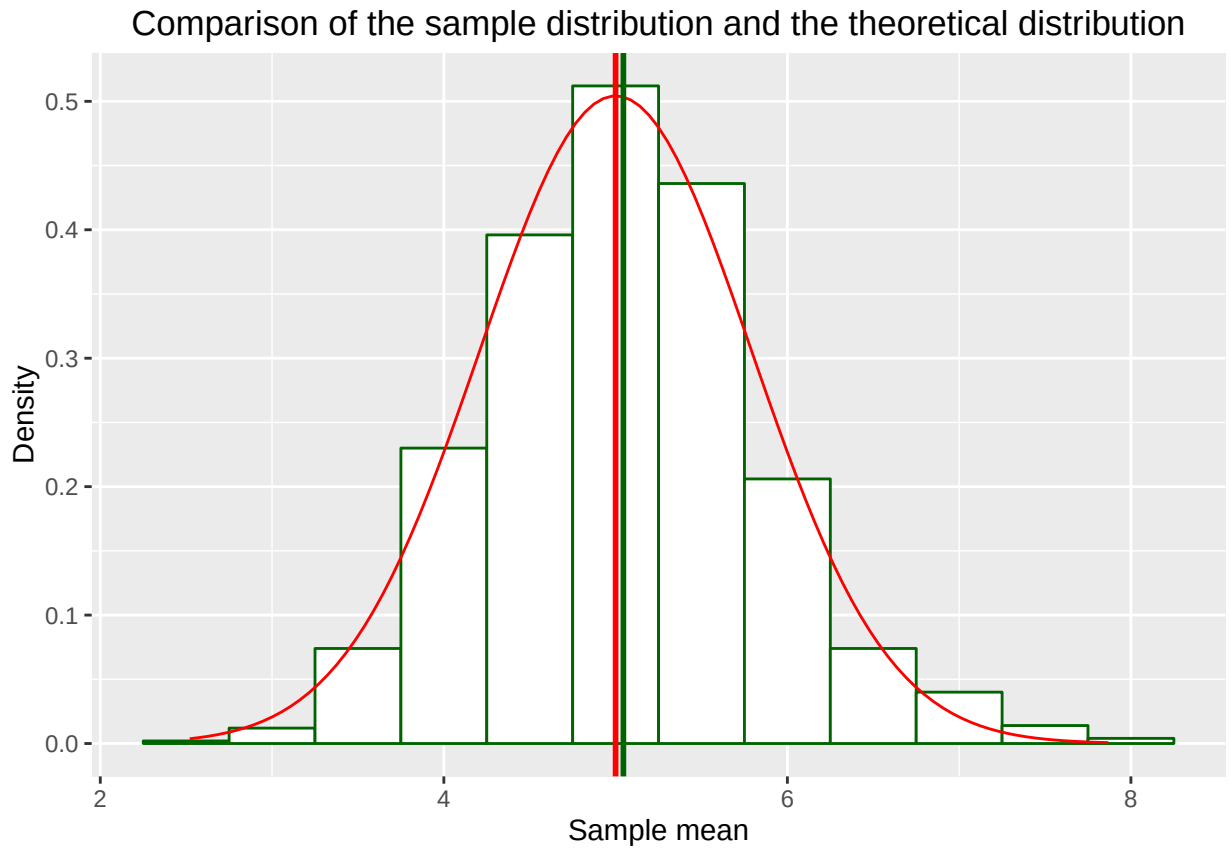
```r
theory_var <- (1 / lambda ^ 2) / n
theory_var
```

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

The sample variance is 0.626, which is also close the theoretical variance 0.625

3.Show that the distribution is approximately normal

```r
data <- as.data.frame(simulated_samp_mean)
ggplot(data, aes(x = simulated_samp_mean)) +
  geom_histogram(binwidth = 0.5, color = 'darkgreen', fill = 'white', aes(y = ..density..)) +
  stat_function( fun = dnorm, color = 'red',
                 args = list(mean = 5, sd = sqrt(0.626))) +
  xlab('Sample mean') +
  ylab('Density') +
  ggtitle('Comparison of the sample distribution and the theoretical distribution')+
  geom_vline(xintercept = c(mean(simulated_samp_mean), 5), size = c(1,1),
               colour = c("darkgreen", "red"))
```



The above figure shows the distribution of the sample mean. It is approximately normal. The red density curve corresponds to N(5,0.626) density.