Statistical Inference: Peer Assessment Part 1

Part 1: Simulation excercise

Overview

In this project the exponential distribution in R will be investigated and compared 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. The distribution of averages of 40 exponentials will be investigated. A thousand simulations will be done.

1. Show the sample mean and compare it to the theoretical mean of the distribution.
2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.

Simulations

load libraries

```
library(knitr)
library(ggplot2)
```

set seed to create reproducability

```
set.seed(123)
```

set the variables

```
n <- 40
lambda <- 0.2
num_sim <- 1000
```

run the simulations

Each element of c() is the mean of 40 randomly exponentials. There are 1000 simulation-runs, so c() consists of 1000 elements.

```
means_row = NULL
for (i in 1 : 1000) means_row = c(means_row, mean(rexp(n,lambda)))
```

Sample mean versus Theoretical mean

The sample mean is the mean over the 1000 row means.The theoretical mean is 1/lambda.

```
mean_sample <- mean(means_row)
mean_theor <- 1/lambda
mean_sample

## [1] 5.011911

mean_theor

## [1] 5
```

The sample mean (first value) is very close to the theoretical mean (second value). See also the plot in the paragraph "Distribution".

Sample variance versus Theoretical variance

The sample standard deviation is the standard deviation over the 1000 row means.The theoretical standard deviation is 1/lambda/sqrt(n). The variance is the squared standard deviation.

```
sd_sample <- sd(means_row)
sd_theor <- 1/lambda/sqrt(n)
sd_sample

## [1] 0.7749147

sd_theor

## [1] 0.7905694
```

The sample standard deviation (first value) is close to the theoretical standard deviation (second value).

The variance is the squared standard deviation:

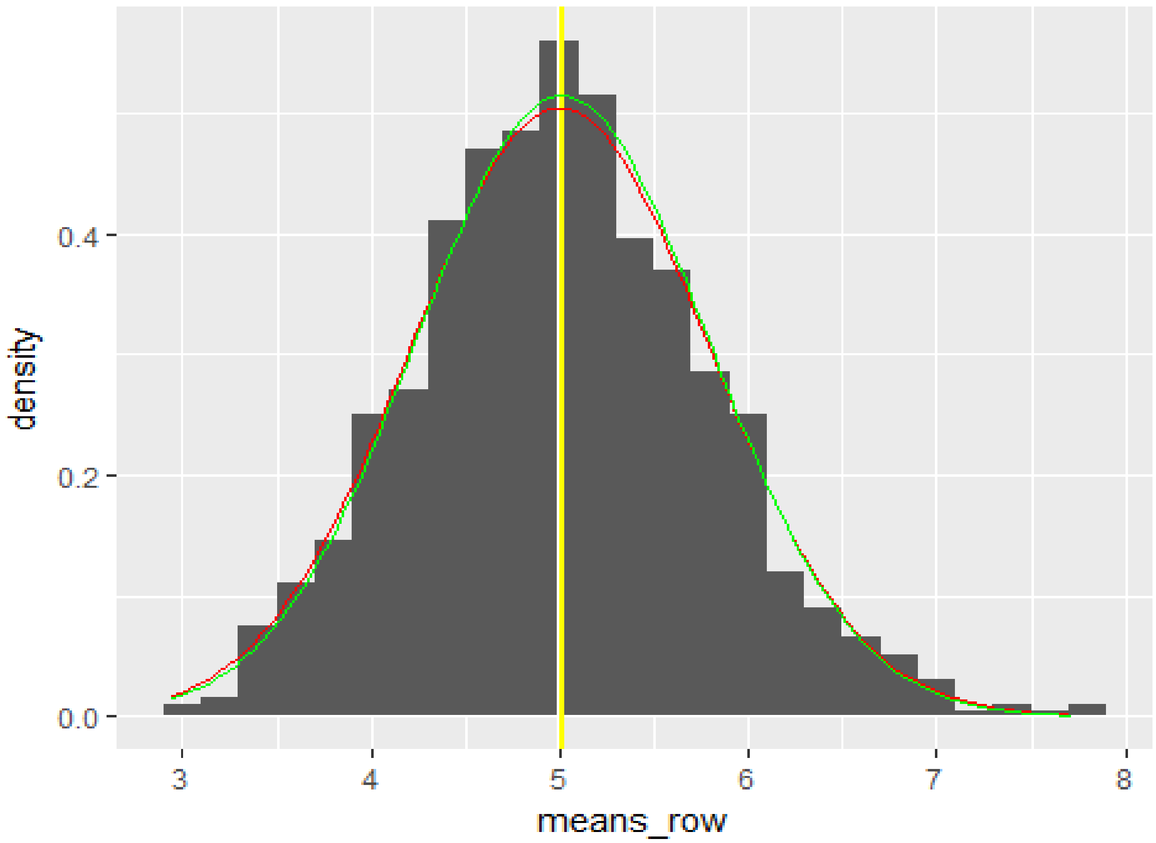
```
var_sample <- var(means_row)
var_theor <- sd_theor^2
var_sample

## [1] 0.6004928

var_theor

## [1] 0.625
```

The sample variance (first value) is close to the theoretical variance (second value).

Distribution

Via a graph will be shown that the distribution of averages of 40 exponentials is approximately normal. The graph is developed with ggplot() as shown in the code below:

```
graph_data <- data.frame(means_row)
distr_graph <- ggplot(graph_data, aes(x = means_row)) +
  geom_histogram(binwidth = lambda, aes(y = ..density..)) +
  geom_vline(xintercept = mean_theor, colour = "white", size = 1) +
  geom_vline(xintercept = mean_sample, colour = "yellow", size = 1) +
  stat_function(fun = dnorm, args = list(mean = mean_theor, sd = sd_theor), colour = "red") +
  stat_function(fun = dnorm, args = list(mean = mean_sample, sd = sd_sample), colour = "green")
  distr_graph
```



In the graph, the yellow line is the sample mean and the white line is the theoretical mean. The red curve is the normal curve formed by the theoretical mean and theoretical standard deviation. The normal curve formed by the sample mean and sample standard deviation is shown in green (n=40 and number of simulations=1000). The curves are close to each other so the distribution of averages of 40 exponentials is approximately normal for large n (1000).This proves the CLT (central limit theory).