Course project: Part 1

Anastasia Kuprina

7/5/2019

Overview

The following report covers Task 1 in the final course project of course "Statistical Inference" available on Coursera. In this project I will investigate the exponential distribution in R and compare it with the Central Limit Theorem.

Simulations

A thousand simulations are performed. The value for lambda is set at 0.2 and the distribution of means of 40 exponential distributions are used.

The simulations illustrate: - comparison of the sample mean with the theoretical mean of the distribution, - how variable the sample is and compare it to the theoretical variance of the distribution, - that the distribution is approximately normal.

Sample vs Theoretical means

Setup

```
set.seed(2612)
library(tidyverse)
library(ggrepel)

# given by the assignment
lambda <- 0.2

# number of exponential distributions to use in simulations
n_of_distributions <- 40

#number of simulations
number_of_simulations <- 1000

#Running the simulations
simulations <- replicate(number_of_simulations,
                         rexp(n_of_distributions, lambda))
glimpse(simulations)
```

```
##  num [1:40, 1:1000] 0.505 6.886 6.863 21.047 0.674 ...
```

Mean comparison

Theoretical mean is computed as 1 over lambda:

```
theoretical_mean <- 1/lambda
paste0('Theoretical mean: ', theoretical_mean)
```

```
## [1] "Theoretical mean: 5"
```

For a sample mean, I need to compute a mean for each simulation and then find a mean of the sample means.

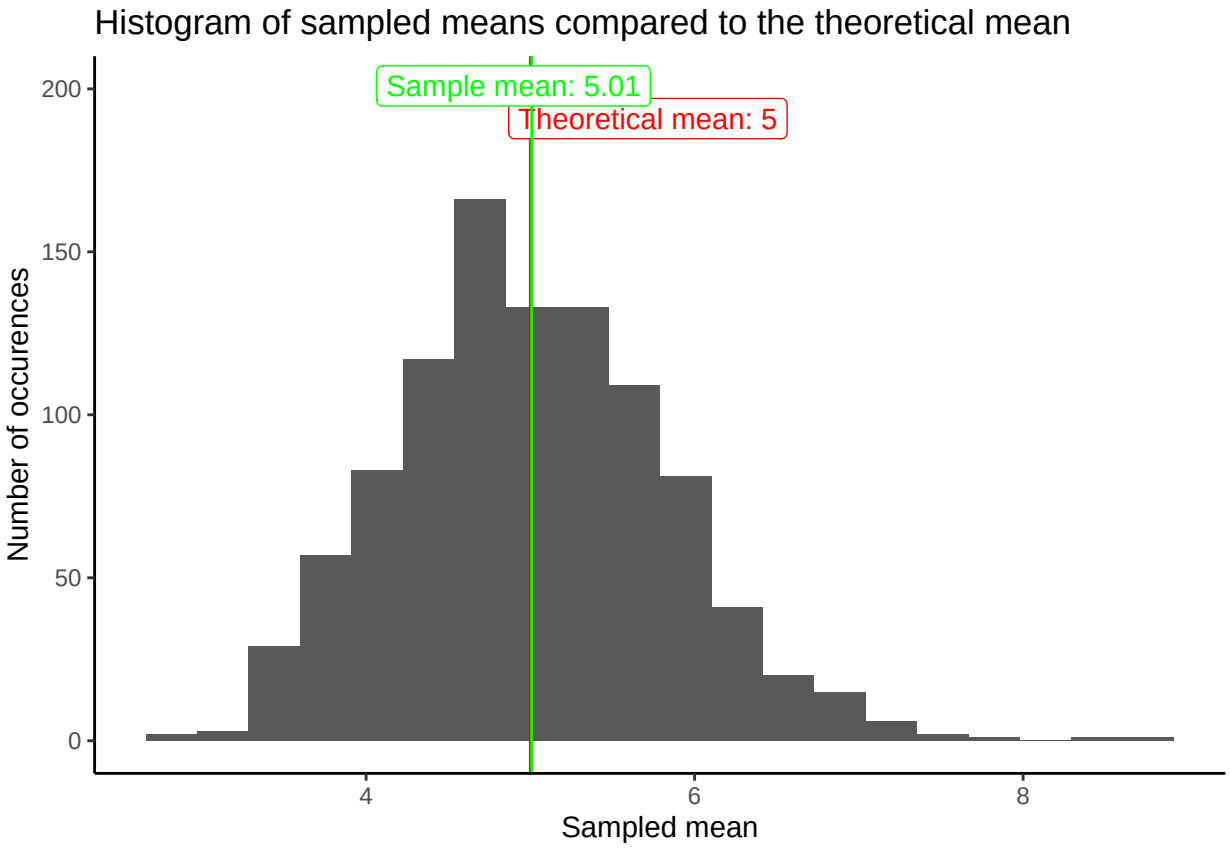
```r
simluated_means <- apply(simulations, 2, mean)

mean_of_sampled_means <- mean(simluated_means)
paste0('Sample mean: ', mean_of_sampled_means)
```

```
## [1] "Sample mean: 5.00839620389703"
```

Visually this can be displayed as:

```r
ggplot() +
  geom_histogram(aes(x = simluated_means), bins = 20) +
  geom_vline(xintercept = theoretical_mean, color = 'red') +
  geom_label_repel(aes(x = theoretical_mean + 1, y = 200,
                label = paste0('Theoretical mean: ', round(theoretical_mean,2))),
            color = 'red') +
  geom_vline(xintercept = mean_of_sampled_means, color = 'green') +
  geom_label_repel(aes(x = mean_of_sampled_means - 1, y = 200,
                       label = paste0('Sample mean: ', round( mean_of_sampled_means,2))),
                   color = 'green') +
  theme_classic() +
  labs(title = 'Histogram of sampled means compared to the theoretical mean',
       y = 'Number of occurences',
       x = 'Sampled mean')
```



Sample vs Theoretical variances

Theoretical variance of the exponential distribution is 1/lambda^2:

```r
theoretical_variance <- 1/(lambda^2)
paste0('Theoretical variance: ', theoretical_variance)
```

```
## [1] "Theoretical variance: 25"
```

In the same way I calculated means for the simulated dataset, variance calculation is possible:

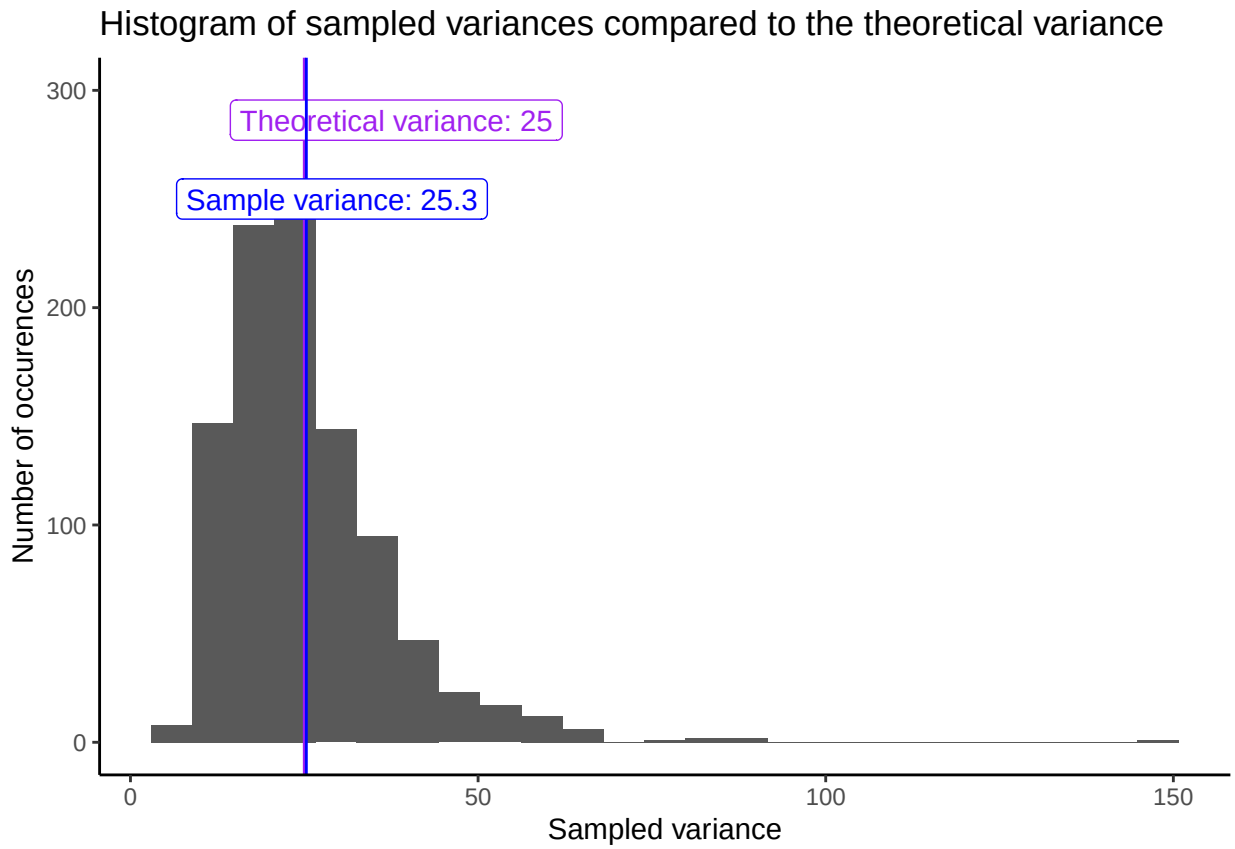
```r
simulated_variances <- apply(simulations, 2, var)

mean_of_sampled_variances <- mean(simulated_variances)
paste0('Sample data variance: ', mean_of_sampled_variances)
```

```
## [1] "Sample data variance: 25.2987318536376"
```

Visually the differences can be displayed for example as following:

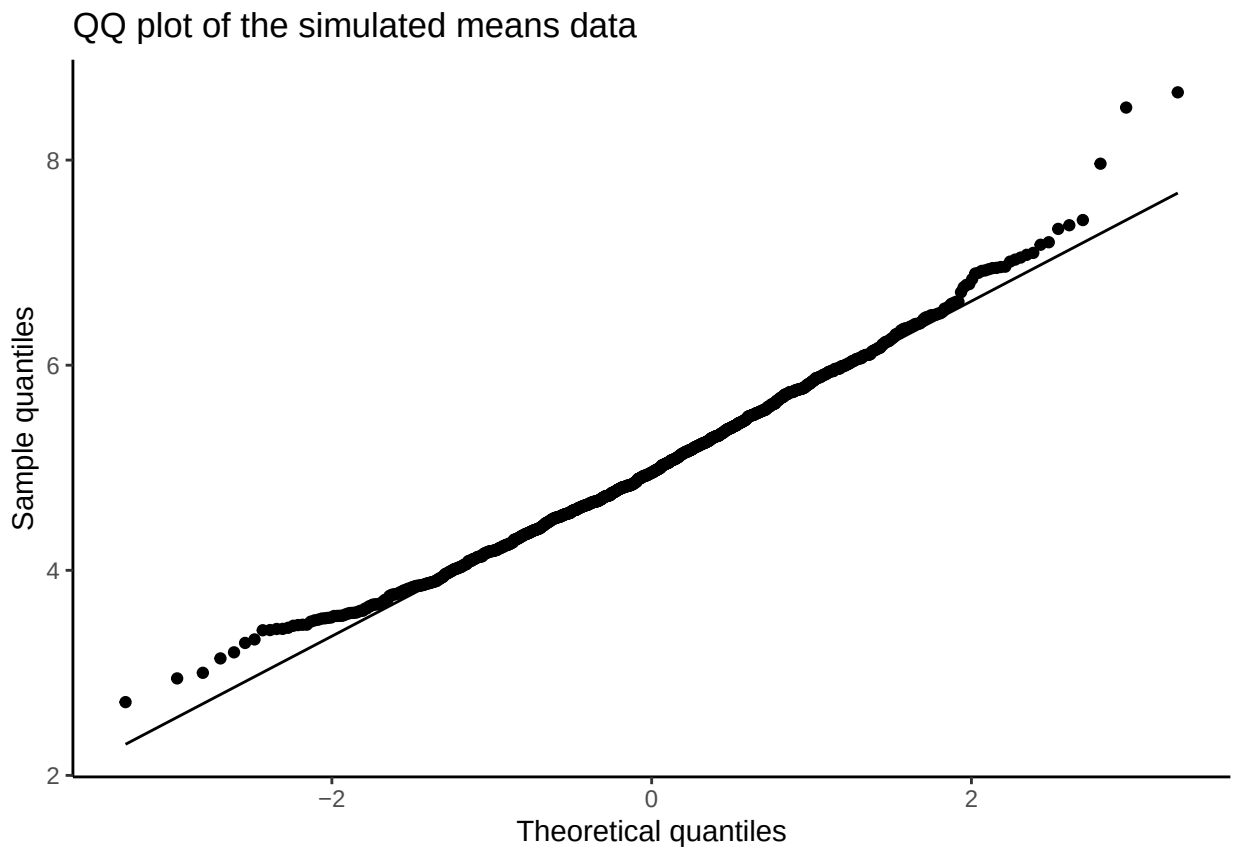
```r
ggplot() +
  geom_histogram(aes(x = simulated_variances), bins = 25) +
  geom_vline(xintercept = theoretical_variance, color = 'purple') +
  geom_label_repel(aes(x = theoretical_variance + 20, y = 300,
                label = paste0('Theoretical variance: ', round(theoretical_variance,2))),
             color = 'purple') +
  geom_vline(xintercept = mean_of_sampled_variances, color = 'blue') +
  geom_label_repel(aes(x = mean_of_sampled_variances - 20, y = 250,
                       label = paste0('Sample variance: ', round( mean_of_sampled_variances,2))),
                   color = 'blue') +
  theme_classic() +
  labs(title = 'Histogram of sampled variances compared to the theoretical variance',
       y = 'Number of occurences',
       x = 'Sampled variance')
```



Distribution of the means is normally distributed

1) quantile-quantile (qq) plot

```
ggplot(data = data.frame(simluated_means),
       aes(sample = simluated_means)) +
  stat_qq() +
  stat_qq_line() +
  theme_classic() +
  labs(title = 'QQ plot of the simulated means data',
       x = 'Theoretical quantiles',
       y = 'Sample quantiles')
```



2) distribution density compared to normal distribution's density if it had the same mean

If we find normal distribution with the same standard deviation and mean as our resampled means and compare the densities, we can "eyeball" the plot to see how close the densities are.

```
min_for_range <- min(simluated_means)
max_for_range <- max(simluated_means)
variance_of_means <- var(simluated_means)


ggplot() +
  stat_function(aes(x = simluated_means),
                fun = dnorm, n = 1000, args = list(mean = mean_of_sampled_means,
                                                   sd = sqrt(variance_of_means)),
                colour = 'black',
                size = 2)  +
  geom_density(aes(x = simluated_means),
               fill = 'orange',
               alpha = 0.5) +
  geom_label_repel(aes(x = mean_of_sampled_means,
                       y = 0.2,
                       label = 'Sampled means density'),
                   color = 'orange') +
  geom_label_repel(aes(x = mean_of_sampled_means,
                       y = 0.55,
                       label = 'Normal distribution density')) +
  labs(title = 'Normal distribution and sampled means densities',
```

```
       subtitle = 'Normal distribution and sample means distributions have the same parameters',
       x = 'Sampled means',
       y = 'Density')
```