Statistical Inference - Coursera Part 1

Loc Nguyen

9/24/2021

Overview

In this project I 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. I will investigate the distribution of averages of 40 exponentials. Note that I will need to do a thousand simulations.

Loading libraries

Loading necessary libraries for the report

```
library(tidyverse)
library(ggplot2)
```

Simulations

Firstly, Set seed and other parameters for reproducability later

```
set.seed(11)
lambda <- 0.2
n <- 40
```

Then, I generation a list of 40 vectors consisting of 1000 exponential random variables

```
sim_exp <- replicate(1000, rexp(n, lambda))
```

Sample Mean versus Theoretical Mean (Question 1)

Show the sample mean and compare it to the theoretical mean of the distribution.
Calculate mean of each exponentials

```
each_mean <- apply(sim_exp, 2, mean)
```

Calculate sample mean

```
sam_mean <- mean(each_mean)
sam_mean
```

```
## [1] 4.987157
```

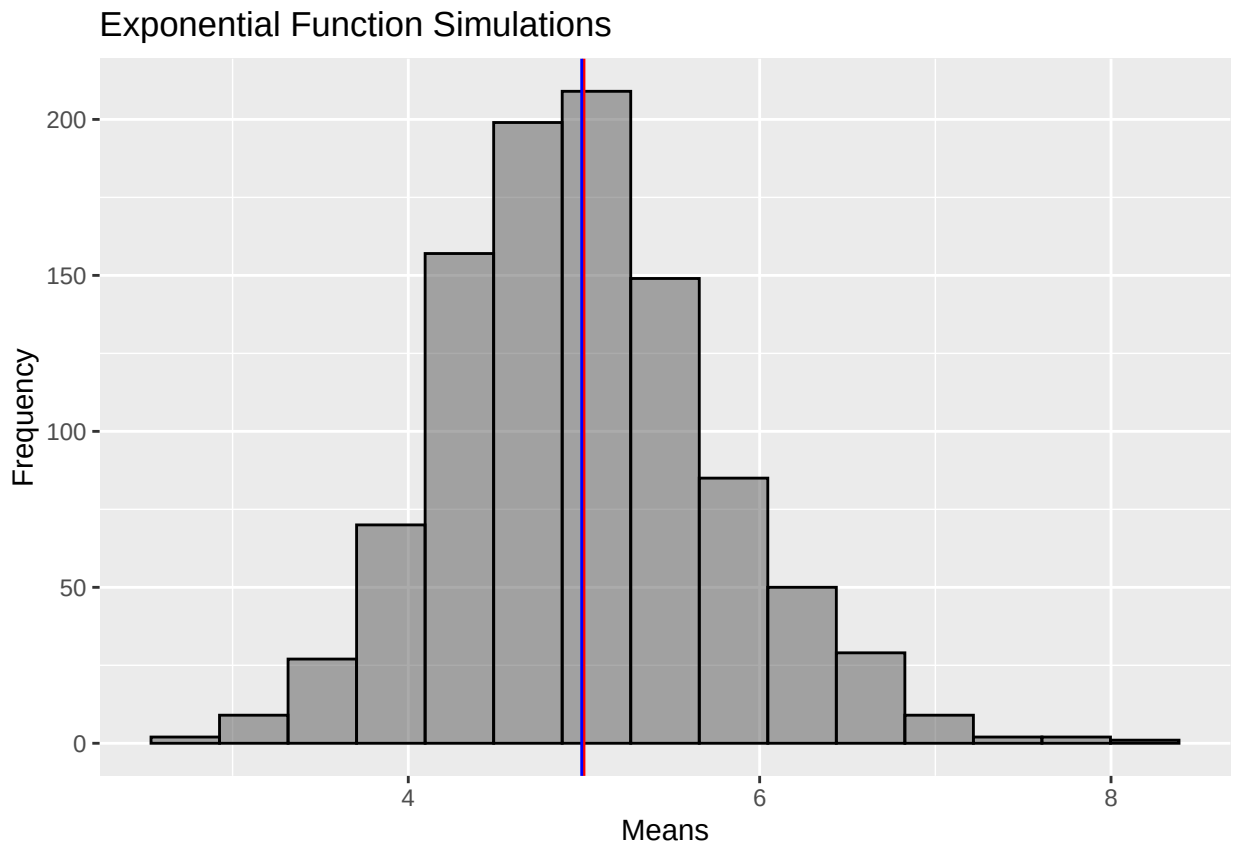
Calculate theoretical mean

```
theory_mean <- 1/lambda
theory_mean
```

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

Compare and show in figure. The sample mean is shown in blue line and theoretical mean is shown in red line

```
      ggplot() + geom_histogram(aes(x = each_mean), bins = 15, alpha = 0.5
                                , color="black") +
            geom_vline(aes(xintercept = theory_mean),col = 'red') +
            geom_vline(aes(xintercept = sam_mean), col = "blue") +
            labs(x = "Means", y = "Frequency", title = "Exponential Function Simulations")
```



The analytics mean is 4.9871567 and the theoretical mean is 5. So, The center of distribution of averages of 40 exponentials is near to the theoretical center of the distribution.

Sample Variance versus Theoretical Variance (Question 2)

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

```
stan_sd_dist <- sd(each_mean)
var_dist <- stan_sd_dist^2
var_dist
```

```
## [1] 0.6009383
```

Theoretical variance of the distribution

```
var_theory <- ((1/lambda)*(1/sqrt(n)))^2
var_theory
```

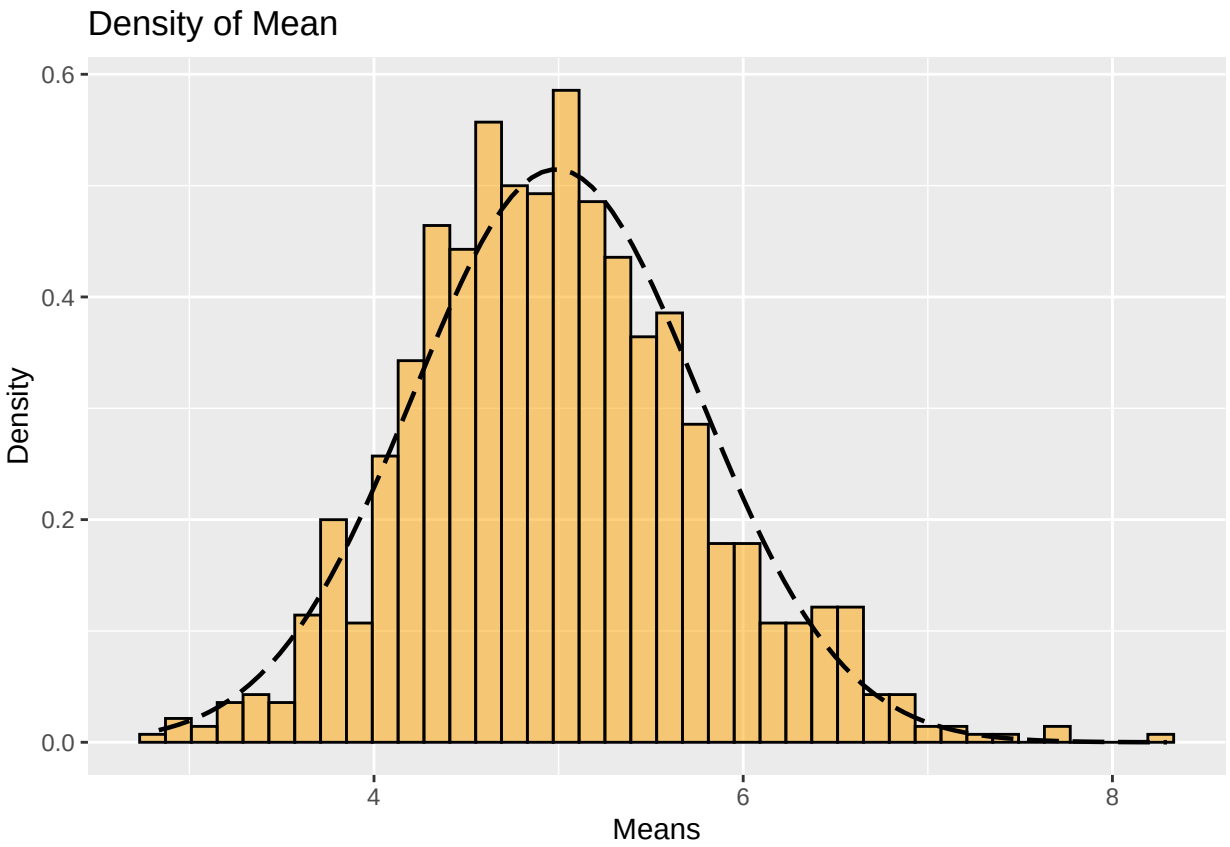
```
## [1] 0.625
```

The Variance of distribution is 0.6009383 and Theoretical variance of the distribution is 0.625

Distribution. (Question 3)

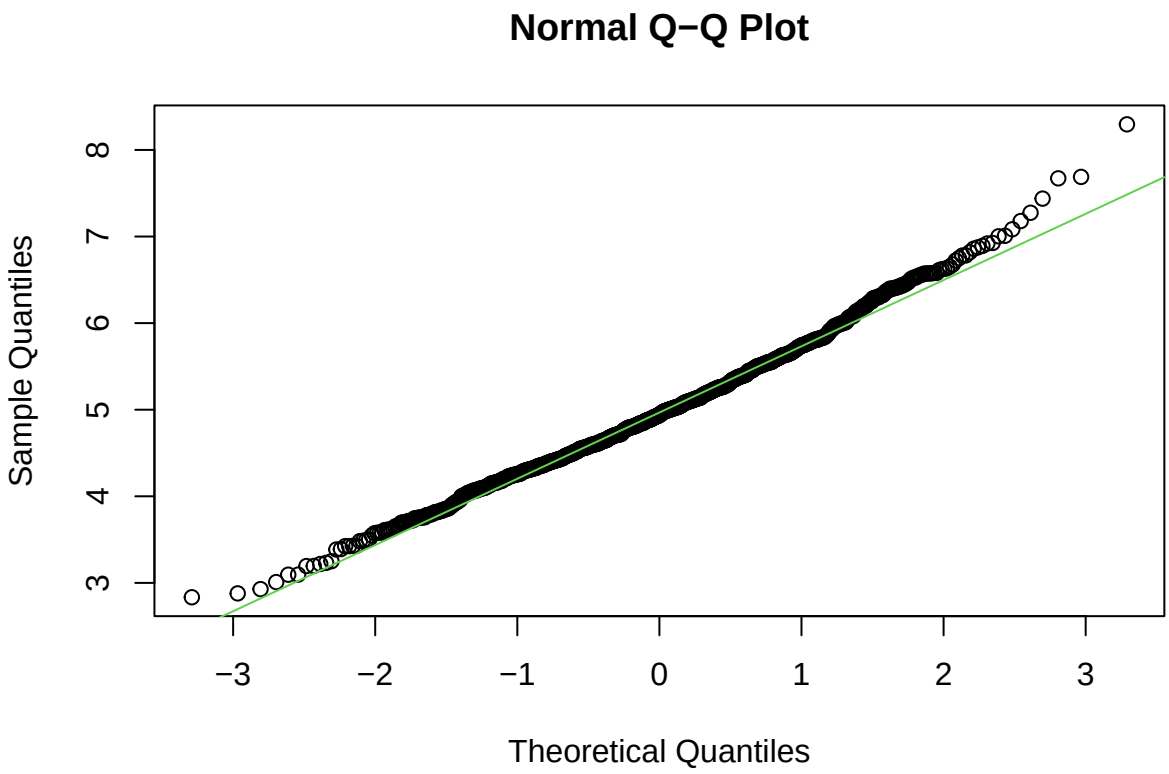
Show that the distribution is approximately normal
Let's plot the distribution of a thousands means in ggplot2

```
# Show that the distribution is approximately normal.
ggplot() + geom_histogram(aes(x = each_mean, y = ..density..), bins = 40,
                          alpha = 0.5, color="black", fill = "orange") +
      stat_function(fun = dnorm, args = list(mean = sam_mean,
                        sd = stan_sd_dist), size = 0.8, lty = 5) +
      labs(x = "Means", y = "Density", title = "Density of Mean")
```



Also qqplot are used to compare sample distributions with a normal distribution. If the qqplot draws a straight line it can be said that the distribution of the sample is normal.

```
# compare the distribution of averages of 40 exponentials to a normal distribution
qqnorm(each_mean)
qqline(each_mean, col = 3)
```



The normal Q-Q Plot also shows that the distribution of the samples means has a normal distribution.

Due to the central limit theorem (CLT), the distribution of averages of 40 exponentials is approximately equal to a normal distribution.