Statistical Inference - Simulation Exercise

Dinh Tuan Phan

4/29/2021

Synopsis

This is a project for the Coursera's Statistical Inference Class from Johns Hopkins University. The project consists of two parts:

1. A simulation exercise.
2. Basic inferential data analysis.

Instructions

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. You should

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. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

Loading packages

```r
library("data.table")
library("ggplot2")
```

Create sample data

```r
# set seed for reproducibility
set.seed(0)
# set lambda to 0.2
lambda <- 0.2
# we need 40 samples
n <- 40
# set 1000 simulations
simulations <- 1000
# simulate the samples
simulated_exponentials <- replicate(simulations, rexp(n, lambda))
```

```
# calculate mean of samples
means_exponentials <- apply(simulated_exponentials, 2, mean)
```

Question 1

Show where the distribution is centered at and compare it to the theoretical center of the distribution.

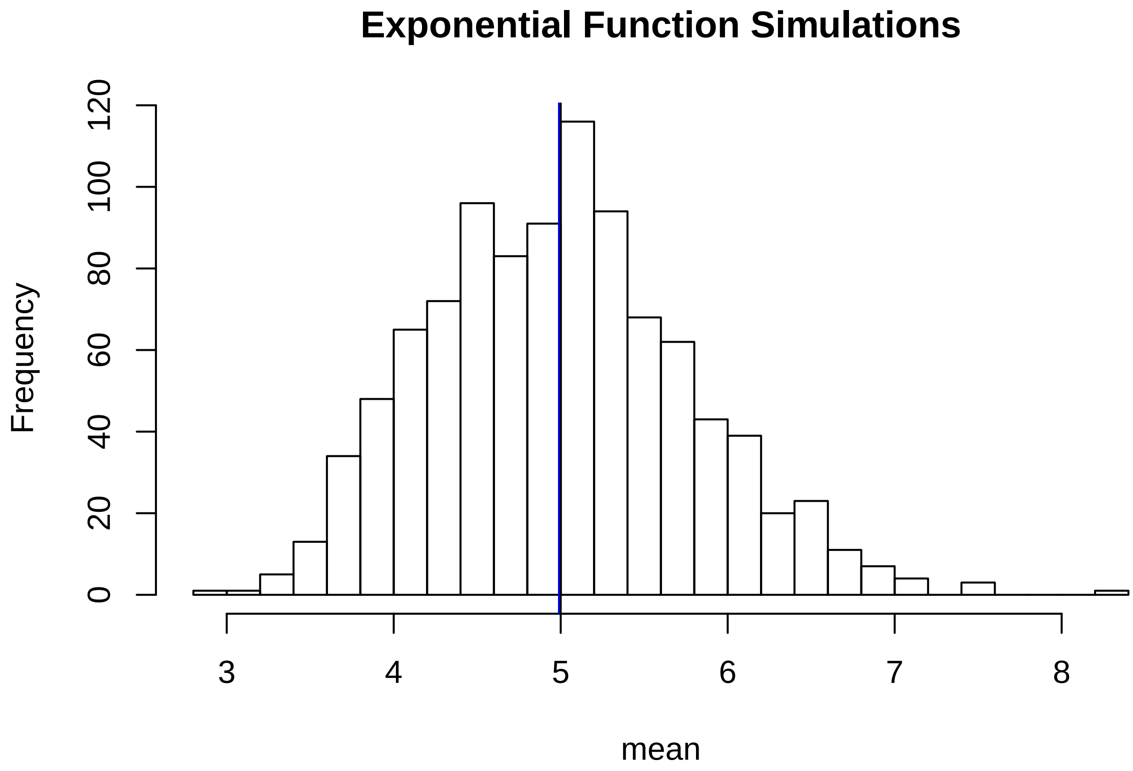
```
analytical_mean <- mean(means_exponentials)
analytical_mean
```

```
## [1] 4.989678
```

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

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

```
# plot data
hist(means_exponentials, xlab = "mean", main = "Exponential Function Simulations",20)
abline(v = analytical_mean, col = "blue")
abline(v = theory_mean, col = "black")
```



The analytics mean is 4.989678 and the theoretical mean 5. The center of distribution of averages of 40 exponentials is close to the theoretical center of the distribution.

Question 2

Show how variable it is and compare it to the theoretical variance of the distribution..

```
# standard deviation
standard_deviation_dist <- sd(means_exponentials)
standard_deviation_dist
```

```
## [1] 0.7862304
```

```
# standard deviation
standard_deviation_theory <- (1/lambda)/sqrt(n)
standard_deviation_theory
```

```
## [1] 0.7905694
```

```
# variance of distribution
variance_dist <- standard_deviation_dist^2
variance_dist
```

```
## [1] 0.6181582
```

```
# variance from expression
variance_theory <- ((1/lambda)*(1/sqrt(n)))^2
variance_theory
```

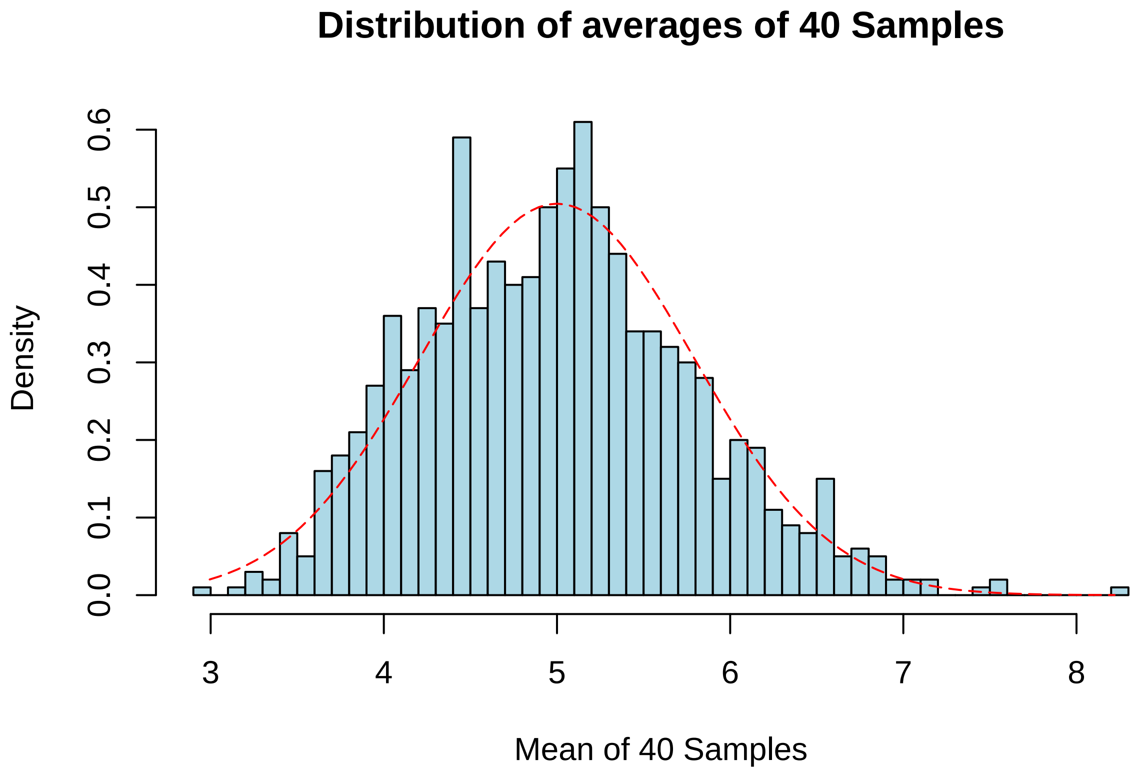
```
## [1] 0.625
```

Standard Deviation of the distribution is 0.78624 with the theoretical SD calculated as 0.79056. The Theoretical variance is calculated as 0.61815. The actual variance of the distribution is 0.625

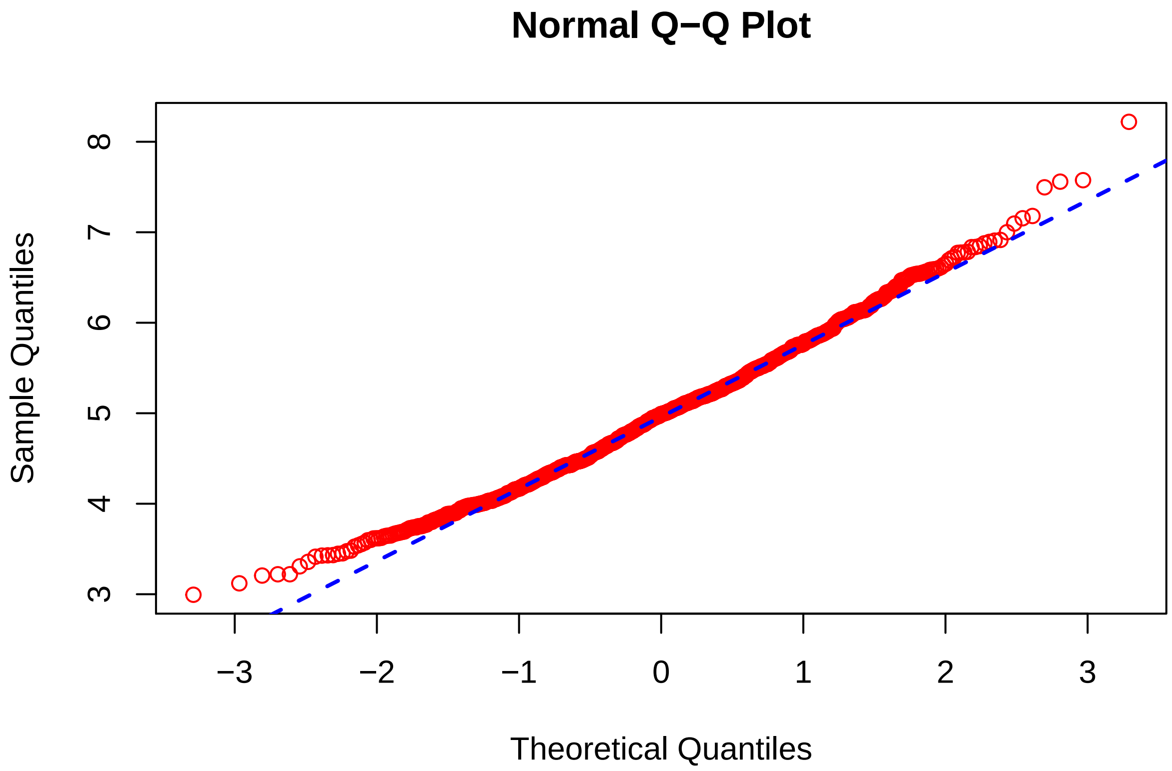
Question 3

Show that the distribution is approximately normal.

```
xfit <- seq(min(means_exponentials), max(means_exponentials), length=100)
yfit <- dnorm(xfit, mean=1/lambda, sd=(1/lambda/sqrt(n)))
hist(means_exponentials,breaks=n,prob=T,col="lightblue",xlab = "Mean of 40 Samples",main="Distribution
lines(xfit, yfit, pch=10, col="red", lty=5)
```



```
# compare the distribution of averages of 40 samples to a normal distribution
qqnorm(means_exponentials, col = 'red', lwd = 1)
qqline(means_exponentials, col = 'blue', lwd = 2, lty=2)
```



Due to the central limit theorem (CLT), the distribution of averages of 40 samples is very close to a normal distribution.