

Part1: Simulation Exercise

Charles

August 7, 2017

Overview

In this report, we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. We compare the sample mean and sample variance with its theoretical values. Then show that the distribution is approximately normal.

Simulation

We first set up working environment:

```
setwd("C:/Study/Coursera/1 Data-Science/2 RStudio/6 Class 6/Coursera_DataScience_Class6_FinalProject")
set.seed(135246987)
library(ggplot2)
library(gridExtra)
library(grid)
```

Then, as we would like to investigate the distribution of average of 40 exponentials, and 1000 samples with $\lambda = 0.2$. So we set up the parameters:

```
n = 40
lambda = 0.2
numSample = 1000
```

We sample from an exponential distribution by creating a 1000×40 matrix and take mean for every row to get the sample means:

```
dataExp <- matrix( rexp(n*numSample,lambda), nrow = numSample)
dataMean <- data.frame(Mean = apply(dataExp,1,mean))
```

Here's the sample means:

```
head(dataMean)
```

```
##           Mean
## 1 3.828543
## 2 5.361409
## 3 4.052994
## 4 3.785670
## 5 5.589950
## 6 5.016944
```

Sample Mean versus Theoretical Mean

We calculate the sample mean:

```
mu <- mean(dataExp)
```

```
## [1] 4.997655
```

The theoretical mean is $1/\lambda = 5$. As we can see, it's pretty close.

Sample Variance versus Theoretical Variance

We calculate the sample variance:

```
sd <- sd(dataExp)/sqrt(n)
```

```
## [1] 0.7846227
```

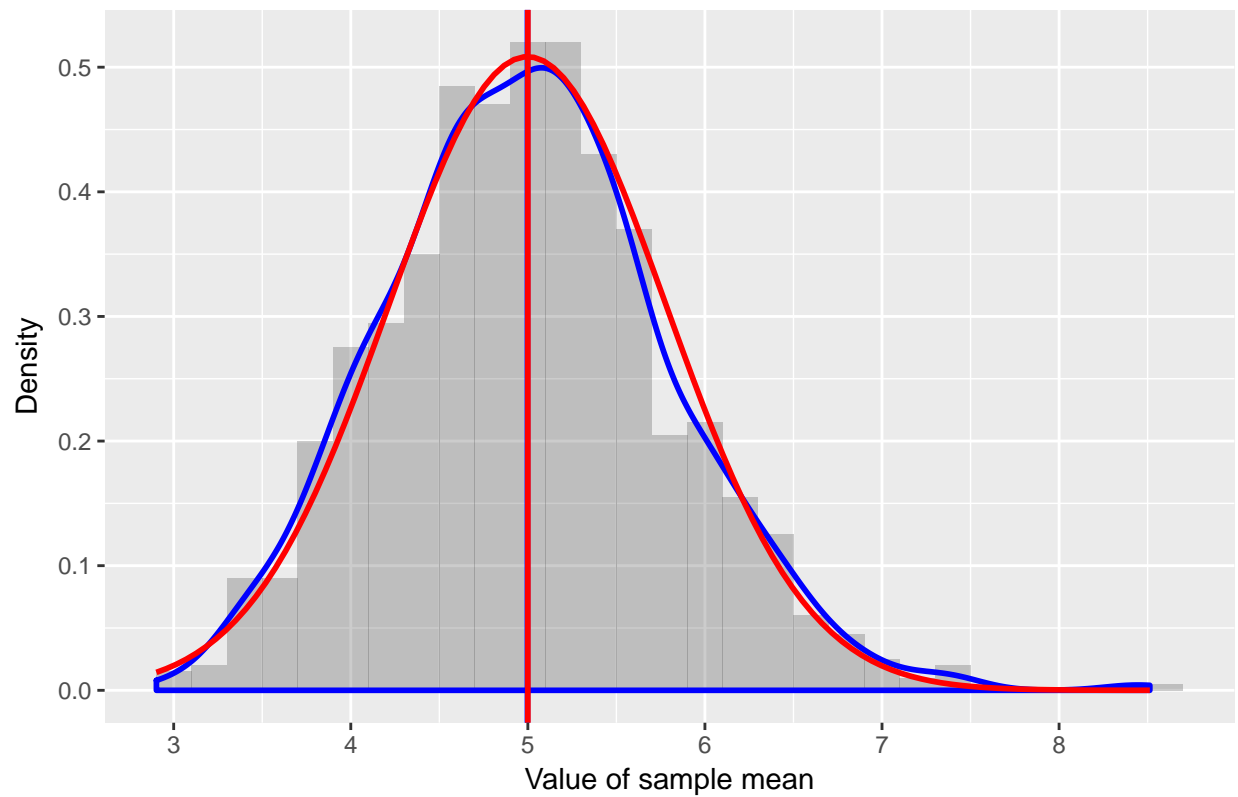
The theoretical mean is $1/\lambda = 5$. As we can see, it's still very close.

Distribution

We use ggplot to draw the distribution of the sample means. Use blue curve to draw the density of sample means and blue line to draw the mean of sample means. Use the red curve to draw the normal distribution and the red line to draw its mean.

```
g <- ggplot(data = dataMean, aes(x=Mean)) +  
  geom_histogram(binwidth = 0.2, aes(y=..density..), alpha=0.3) +  
  geom_vline(xintercept = mean(dataExp), size = 1, color = "blue") +  
  geom_density(color="blue",size=1) +  
  stat_function(fun = dnorm, args = list(mean=mu, sd=sd), color="red",size=1) +  
  geom_vline(xintercept = 1/lambda, color="red",size=1) +  
  ggtitle("Density Function of Sample distribution and normal distribution") +  
  xlab("Value of sample mean") +  
  ylab("Density")  
plot(g)
```

Density Function of Sample distribution and normal distribution



As we can see above, the distribution is approximately normal, and the mean of them two is pretty close to each other.