Statistical Inference Course Project - Part 1

Trieu Tran

September 21, 2015

Synopsis

This is the first part of the course project for the statistical inference class. The goal of this project is performing some statistical inferential works such as, exploring inference by using simulation functionality of R to study the exponential distribution then comparing it with Central Limit Theorem.

Simulations

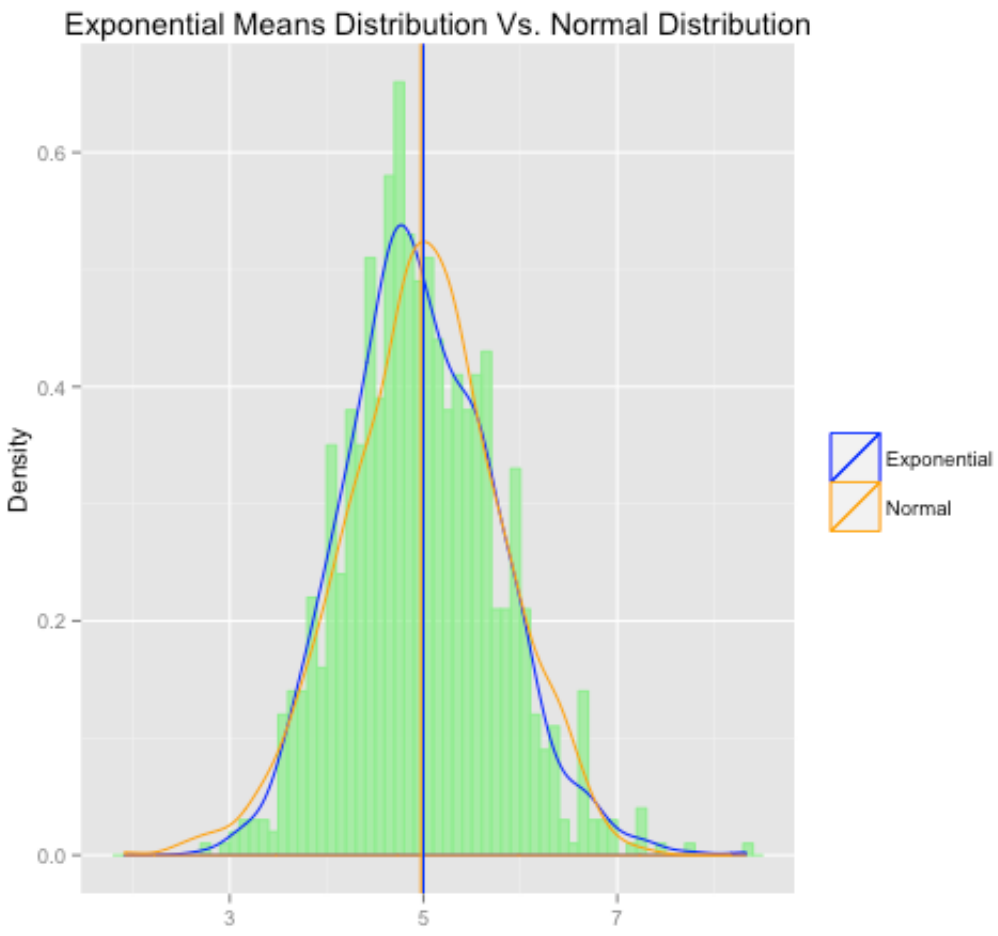
Random sampling from the exponential distribution

The function **rexp** of R stat package creates random values belong to the exponential distribution; and it requires two input parameters, n - number of observations (n = 40) and rate (lambda = 0.2). My simulation plan is running this function one thousand times, a fairly large number of trials, then collecting arithmetic mean of each trial.

```
library(knitr)
library(ggplot2)

## checking existence of a folder named "figure", if not then creating one to store plot figures
figureDir <- 'figure'
if (!file.exists(figureDir)){
    dir.create(figureDir)
}

## initializing variables
set.seed(12345)
lambda <- 0.2
n <- 40
numSim <- 1000

## replicating random exponential sampling with sampling size n = 40, and repeat
## nSim = 1000 times, and storing sample data in a matrix
simMatrix <- replicate(numSim, rexp(n, rate = lambda))

## getting average for each column of the matrix
simColMeans <- colMeans(simMatrix)

## determining sample mean, sd and variance
sampleMean <- mean(simColMeans)
sampleSD <- sd(simColMeans)
sampleVar <- sampleSD^2

## calculating normal (theoritical) mean, normal standard deviation and variance
normalMean <- 1/lambda
normalSD <- 1/lambda/sqrt(40)
normalVar <- normalSD^2

## making a dataframe storing sample means in "x"" column; and the "y" column contains random values
## from the normal distribution with mean equals to the above calculated "normalMean"
## and standard deviation equals to "normalSD"

data <- data.frame(x = simColMeans, y = rnorm(1000, mean = normalMean, sd = normalSD))
```

Plotting histogram and PDF of sample and PDF of theoritical (normal) distribution

```
p <- ggplot(data, aes(x = x))
## adding histogram
p <- p +  geom_histogram(aes(y = ..density..), fill = I("lightgreen"), col = I("lightgreen"), alpha = 0.75, bin
width = 0.1)
## adding density line of the sample distro
p <- p + geom_density(aes(colour = "1")) + labs(x = "", y = "Density", title = "")
## adding density line of the normal distro
p <- p +  geom_density(aes(x = y, colour = "2"))
## drawing sample mean
p <- p + geom_vline(color = "blue", xintercept = 1/lambda)
## drawing normal mean
p <- p + geom_vline(color = "orange", xintercept = sampleMean)
## labeling
p <- p + labs(x = "", y = "Density", title = "Exponential Means Distribution Vs. Normal Distribution")
## adding legend
p <- p + scale_colour_manual(
        values=c("1"="blue", "2"="orange"),
        name="",
        labels = c("Exponential", "Normal"))
```



Observations

```
round(sampleMean, 4)
```

```
## [1] 4.972
```

```
round(normalMean, 4)
```

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

```
round(sampleVar, 4)
```

```
## [1] 0.5954
```

```
round(normalVar, 4)
```

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

1. The sample mean **4.9720** (the vertical blue line in the plot) is very close to the theoretical (normal) mean **5.0000** (the vertical orange line in the plot), as predicted
2. Similiarity, the variance of the sample is **0.5954**, which is also approximately equal to the variance of the normal distribution **0.6250**

Conclusions

From the above plot, the probability density functions of both sample and theoritical distribution almost overlap each other. Or we can say that the distribution of the sample means follows the normal distribution, as stated in the Central Limit Theorem.