Statistical Inference Part 1: Simulation Exercise

JackHo

30 October 2016

Table of Contents

Before Cooking

To keep the article is shorter than 3 pages, some codes have been shaded by setting `echo=FALSE' in the chunk option.

```
# Loading necessary packages
require(ggplot2)
require(dplyr)
```

Overview

In this project I investigated 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 investigated the distribution of averages of 40 exponentials.To get better illsutrations, I simulated the expoential distribution `100000` times.

Part 1: Simulation Exercise Instructions

1.1 Compare the sample mean with the theoretical mean of expoential distribution

```
# Set seed- keeping its reproducible
set.seed(1)
# Sim 100000 times and calculate means
SMeanRexp <- vector(length = 100000, mode = "numeric")
CMeanRexp <- vector(length = 100000, mode = "numeric")
SVarianceRexp <- vector(length = 100000, mode = "numeric")
CVarianceRexp <- vector(length = 100000, mode = "numeric")
vec <- vector(length = 40, mode = "numeric")
for(i in 1:100000){
      vec<-rexp(n = 40, rate = .2);      SMeanRexp[i] <- mean(vec)
      SVarianceRexp[i] <- var(vec);      CMeanRexp[i] <- sum(SMeanRexp)/i
      CVarianceRexp[i] <- sum(SVarianceRexp)/i
}
```

```r
print(sprintf("The mean of the sample is:  %f The Theoretical mean is: %f", mean(SRexp[
,1]), 1/.2))

## [1] "The mean of the sample is:  5.002796 The Theoretical mean is: 5.000000"
```

The `sample mean` (5.0027956) of the distribution is a little larger than the `theoretical mean` (5) of the distribution.

1.2 Compare the sample variance with the theoretical variance of expoential distribution

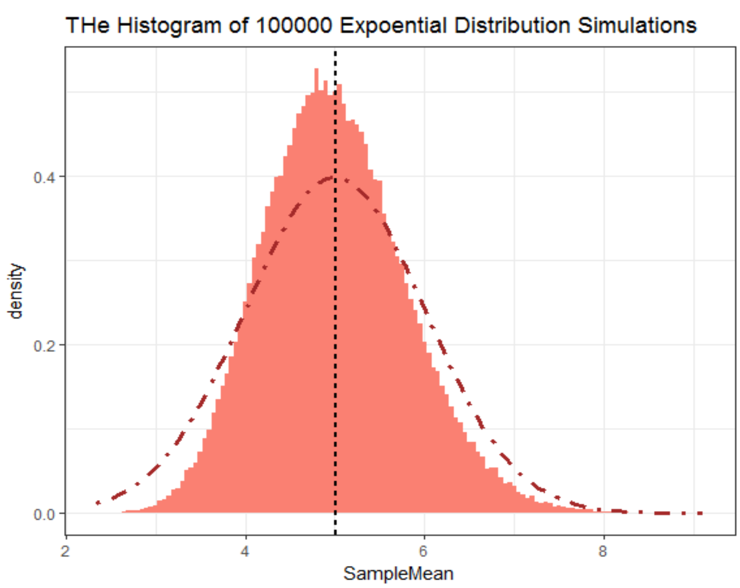
```r
print(sprintf("The mean of the variance is:  %f The Theoretical variance is: %f", mean(
SRexp$SampleVariance),(1/.2)^2))

## [1] "The mean of the variance is:  25.021734 The Theoretical variance is: 25.000000"
```

The `sample variance` (25.0217336) of the distribution is a little larger than the `theoretical mean` (25) of the distribution.

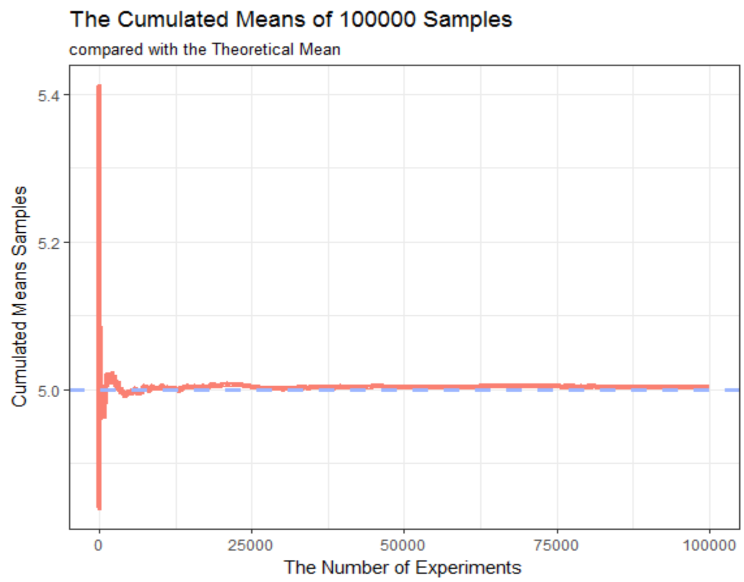
1.3 Visulaze the distributions and trends

```r
# Draw histogram- checking the distribution of 1000 times exponential distribution
ggplot(SRexp,aes(SampleMean)) + geom_histogram(fill="salmon",aes(y= ..density..),binwid
th = .05)+stat_function(fun = dnorm, color = "brown", size =1.1, linetype =4,args = lis
t(mean=5))+ggtitle("THe Histogram of 100000 Expoential Distribution Simulations")+ geom
_vline(xintercept = c(5,mean(SRexp$SampleMean)),size=.5,col="black",linetype=2)+theme_b
w() # the difference is too small to distinguish the distance between two vertical line
s
```



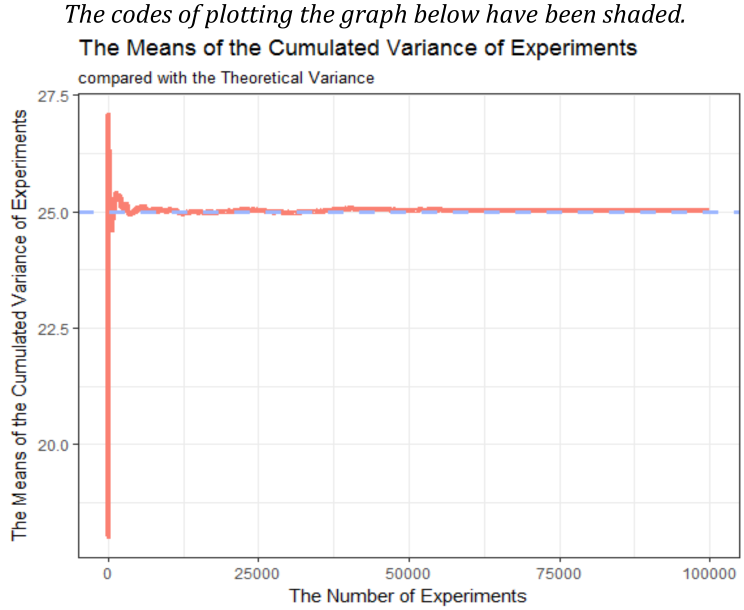
From the plot `THe Histogram of the simulation of 100000 Expoential Distribution`, the distribution is highly close to Gaussian (very look like a bell-shape).

```r
# Plot the trends of the cumulated sample means
ggplot(SRexp,aes(y=CMeanRexp,x=1:100000))+geom_line(col="salmon",size = 1.5)+geom_hline
(yintercept = 1/.2,col="#9cb6ff",linetype=2.2,size=1.2) + xlab("The Number of Experimen
ts") + ylab("Cumulated Means Samples")+ggtitle(label = "The Cumulated Means of 100000 S
amples", subtitle ="compared with the Theoretical Mean")+theme_bw()
```



From the plot The Cumulated Mean of 100000 Expoential Distribution compared with the Theoretical Mean of 100000 Expoential Distribution, the cumulated tends to be close to the theoretical mean after around 25000 times experiments.

The codes of plotting the graph below have been shaded.



Bascially, like the things between the mean of the sample means, the variance of the sample will ultimately be close to the population variance. See the plot below.

1.4 Summary

As a whole, from the 2 sections above, based on a large number of experiments, the mean of the sample means will be gravitated toward to the population mean. As a result, we could concluded that the mean of the samples can predict the mean of the population, and the variance of the sample, can also predict the variance of the population variance.