Coursera - Statistical Inference Course Project - Part 1

Goh LW

Saturday, October 24, 2015

Overview

In the below project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The mean of exponential distribution used is 1/lambda and the standard deviation is also 1/lambda. A thousand simulation will be run, the lambda will be set as 0.2 as required in project, the distribution of average of 40 exponentials will be explore.

Simulation

First we run 1000 simulation using random number with lambda=0.2, and take the average of 40 exponentials.

```
lambda <- 0.2
sample <- 40
set.seed(6) #a seed is set to ensure the result is reproducable

# start simulate
simulated_exponentials=NULL
for (i in 1 : 1000)
{
  simulated_exponentials = c(simulated_exponentials, mean(rexp(sample, lambda)))
}
```

Result

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

```
sam_mean <- mean(simulated_exponentials) #sample mean = 4.950171
sam_mean
```

```
## [1] 4.950171
```

```
theor_mean <- 1/lambda #theoretical mean = 5
theor_mean
```

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

As we can see from the output of R code above, the distribution is centered at 4.95, and it is close to the theoretical center of distribution, which is 5. With sample size increase, the value of theoretical and sample value will be getting closer and closer to each other.

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

```r
sd(simulated_exponentials) #standard deviation of distribution of averages of 40 exponentials
```

```
## [1] 0.8148381
```

```r
(1/lambda)/sqrt(sample) #standard deviation from analytical expression
```

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

```r
var(simulated_exponentials) # Variance of the sample mean
```

```
## [1] 0.6639611
```

```r
1/((lambda*lambda) * sample) # Theoritcal variance of the distribution
```

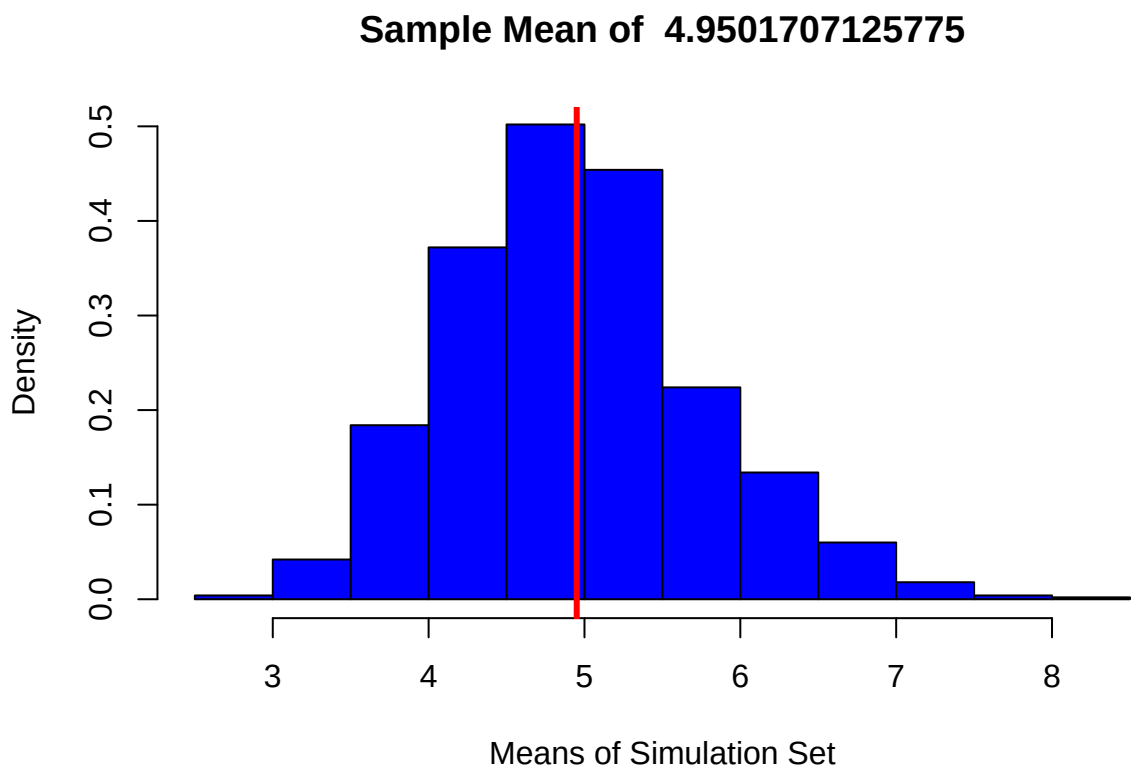
```
## [1] 0.625
```

As observation above, the standard deviation for distribution of averages of 40 exponentials versus analytical expression is not far apart (about 2% difference), while variance of the sample mean versus theoritical is close to each other too (4% difference)

3.) Show that the distribution is approximately normal

The CLT states that averages are approximately normal, with distributions - centered at the population mean - with standard deviation equal to the standard error of the mean - CLT gives no guarantee that n is large enough

So let's put the data that we gathered in previous steps of this project to test by visualize it :



as illustrated by the histrogram above, the distribution of our simulation mean data is center at it's mean, and forming a bell curve.