Statistical Inference: Peer Assessment, Part 1

S C NG

20 Jun, 2015

Overview

In this project we will investigate the exponential distribution in R and compare it with the Central Limit Theorem(CLT).

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.

We will investigate the distribution of averages of 40 exponentials with thousand simulations.

Setup

First of all, the following default settings and libraries are loaded

```
#preset default options for Rmd, codes not shown in report
require(knitr)
```

```
## Loading required package: knitr
```

```
opts_chunk$set(cache=TRUE, echo=TRUE)

#load required libraries for data analysis
require(ggplot2)
```

```
## Loading required package: ggplot2
```

Simulation

The following R codes are used for performing 1000 rounds of simulations. For each round, a sample size of 40 random variables under exponential distribution with rate equals lambda (0.2) are generated, and the means for each round are captured in vector exp_sample_means.

```
set.seed(111)
sample_size <- 40
lambda <- 0.2
exp_sample_means = NULL
for(i in 1:1000) exp_sample_means = c(exp_sample_means, mean(rexp(sample_size, rate=lambda)))
```

Show the sample mean and compare it to the theoretical mean of the distribution

The overall sample mean is calculated after the simulation

```r
set.seed(111)
sample_size <- 40
lambda <- 0.2
exp_sample_means = NULL
for(i in 1:1000) exp_sample_means = c(exp_sample_means, mean(rexp(sample_size, rate=lambda)))
mean(exp_sample_means)
```

```
## [1] 5.02562
```

In this assignment, we assume the mean of exponential distribution is 1/lambda, where lambda = 0.2. Therefore, the theorical mean of the exponential distribution is calculated as follows:

```r
exp_theo_mean <- 1/lambda
exp_theo_mean
```

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

The sample mean is almost the same as the theorical mean of the distribution

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution

The variance of the sample can be calculated from below

```r
var(exp_sample_means)
```

```
## [1] 0.6069798
```

According to Central Limit Theorem (CLT), The theorical variance of the distribution equals the square of theorical standard deviation divided by sample size. In this assignment, the standard deviation is also 1/lambda

```r
(1/lambda)^2/sample_size
```

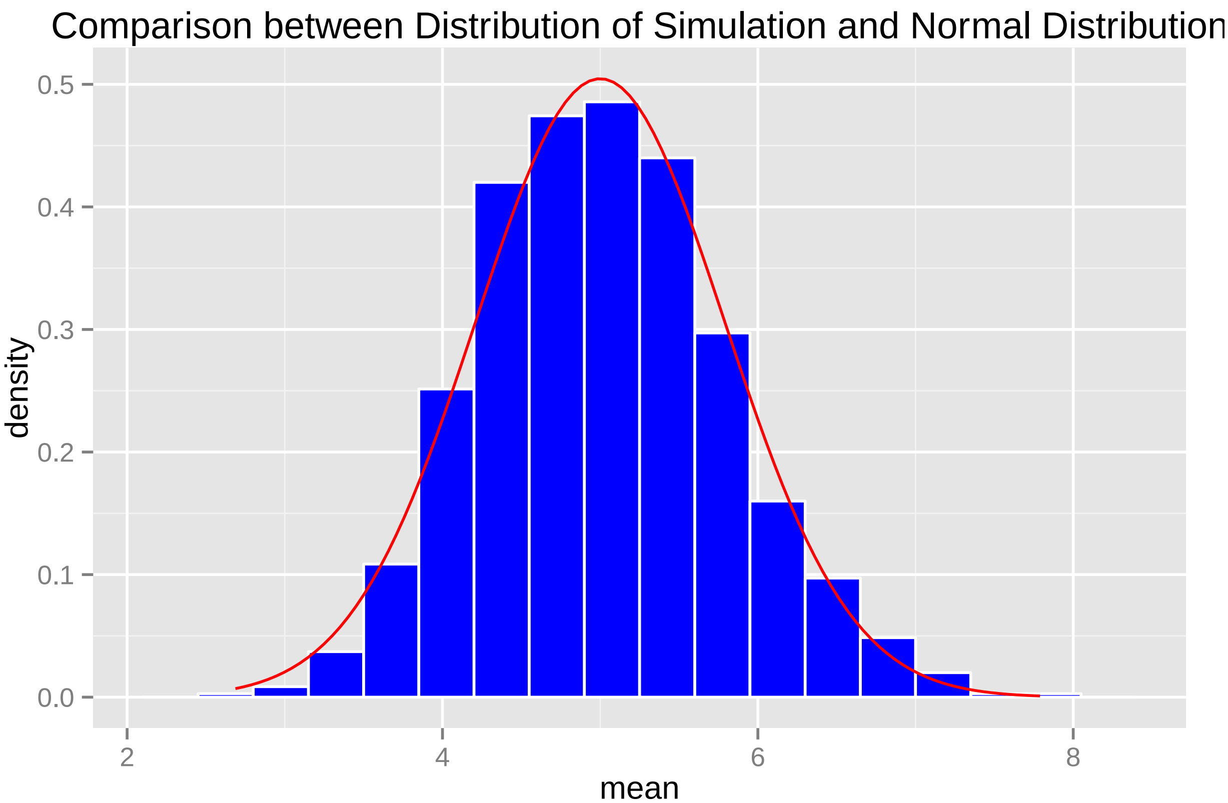
```
## [1] 0.625
```

The sample variance is very close to the theorical variance of the distribution

Show that the distribution is approximately normal

For this point, we 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. In order to illustrate this comparison, a histogram of the distribution for simulation has been plotted. An overlay of density with the theorical normal distribution according to CLT (with mean=0.2, standard deviation=0.2/sqrt(40)) is added for comparison

```r
sample_df <- as.data.frame(exp_sample_means)
g <- ggplot(sample_df, aes(x=exp_sample_means))
g <- g + geom_histogram(aes(y = ..density..), binwidth=0.35, fill='blue', color='white')
g <- g + stat_function(fun = dnorm,
    args=list(mean=1/lambda, sd=(1/lambda)/sqrt(sample_size)),
    bin_width=0.5, color='red')
g <- g + ggtitle("Comparison between Distribution of Simulation and Normal Distribution")
g <- g + xlab("mean") + ylab("density")
g
```



From the diagram, we can see that the distribution of our sample is approximately normal