Coursera Statistical Inference Project

Part 1: Simulation Exercise

Junyang Liu

6/2/2017

Coursera Statistical Inference Course Project

This is a project for the Coursera Statistical Inference Class. The project consists of two parts:

- Simulation Exercise to explore inference
- Basic inferential analysis using the *ToothGrowth* data in the R datasets package

Part 1

Overview

Investigate 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. You will investigate the distribution of averages of 40 exponentials. Note that you will need to do a thousand simulations.

Objectives

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

- Show the sample mean and compare it to the theoretical mean of the distribution.
- Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution.
- Show that the distribution is approximately normal.

In point 3, 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.

Process

```
#set seed
set.seed(100)

#set Lambda to 0.2
lambda<- 0.2

#set sample size to 40
n<- 40

#set simulation counts to 1000
simulation<- 1000

#simulation
```

```r
simulated_exp<- replicate(simulation, rexp(n, lambda))

#calculate mean
exp_mean<-apply(simulated_exp, 2, mean)
```

Question 1

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

```r
#overall distribution mean
overall_mean<- mean(exp_mean)
overall_mean
```

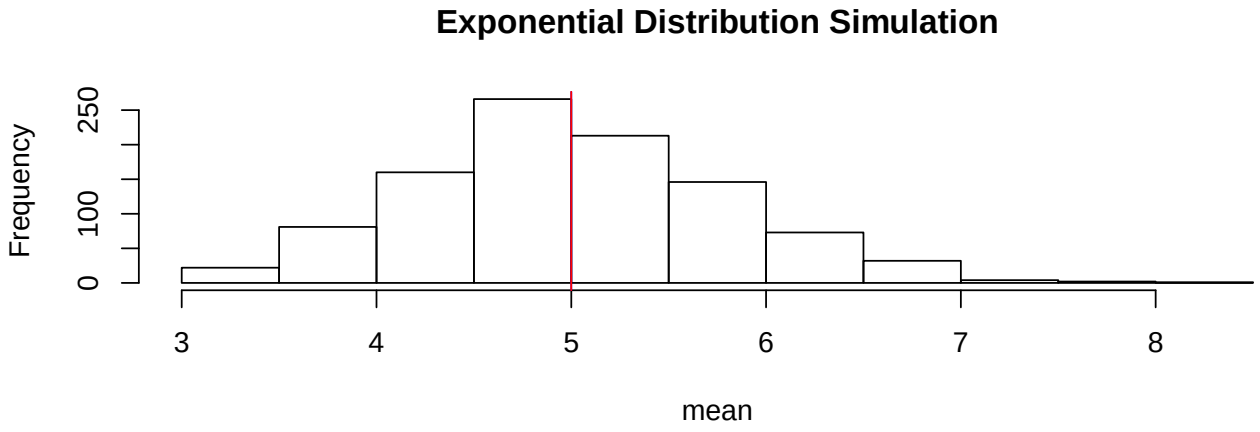
```
## [1] 4.999702
```

```r
#true mean of exponential distribution
true_exp_mean <- 1/lambda
true_exp_mean
```

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

Plots

```r
hist(exp_mean, xlab="mean", main = "Exponential Distribution Simulation")
abline(v = true_exp_mean, col = "blue")
abline(v = overall_mean, col = "red")
```



Question 1 Conclusion

The simulated sample overall_mean is 4.999702, the simulated true mean 5. The center of distribution of averages of 40 exponentials is very close to the theoretical center of the distribution. it is so close that even makes it hard to see it from the plot.

Question 2

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

```r
#standard deviation of distribution
exp_mean_sd<- sd(exp_mean)
exp_mean_sd
```

```
## [1] 0.8020251
```

```r
#true standard deviation
true_mean_sd<- (1/lambda)/sqrt(n)
true_mean_sd
```

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

```r
#variance
exp_mean_var<-exp_mean_sd^2
exp_mean_var
```

```
## [1] 0.6432442
```

```r
#true variance
true_mean_var<-((1/lambda)*(1/sqrt(n)))^2
true_mean_var
```

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

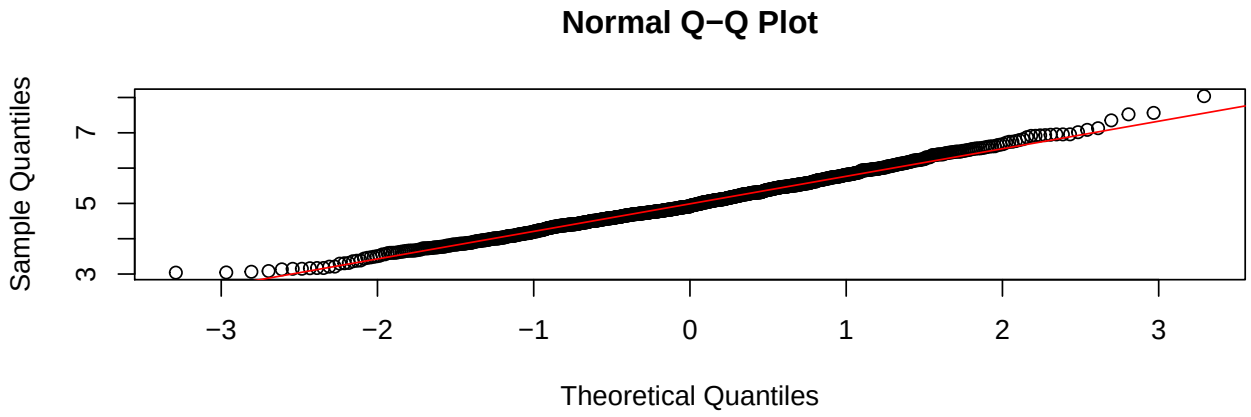
Question 2 Conclusion

SD of the distribution is 0.8020251 with the theoretical SD = 0.7905694. The theoretical variance is 0.625. The sample variance of the distribution is 0.6432442

Question 3

Show that the distribution is approximately normal.

```r
# compare the distribution of averages of 40 exponentials to a normal distribution
qqnorm(exp_mean)
qqline(exp_mean, col = 2)
```



Question 3 Conclusion

By central limit theorem (CLT), the distribution of averages of 40 exponentials is very close to a normal distribution.