Statistical Inference Project Part 1

Ryan Wissman

Saturday, January 24, 2015

Coursera Statistical Inference Course Project

This project consists of two parts:

1) Data simulations
2) Data Analysis

More information regarding this project:

In this project you will 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.

This project will cover 3 areas:

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should 1. Show the sample mean and compare it to the theoretical mean of the distribution. 2. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. 3. Show that the distribution is approximately normal.

Setup and Simulation

First we will need set the parameters and run a simlulation of 1000 means for 40 samples.

```
#set the number of simulations
n_sim <- 1000

#set the sample size
n_samples <- 40

#set the random seed
set.seed(1313)

#set the value of lambda
lambda <- 0.2

#create dataframe containing simulated means
mean_data <- data.frame(mean = apply(matrix(rexp(n_samples*n_sim, lambda), n_sim), 1, mean))
head(mean_data)
```

```
##       mean
## 1 7.028397
## 2 5.321861
```

```
## 3 6.276993
## 4 4.088829
## 5 5.591798
## 6 5.513257
```

Question 1

Show the sample mean and compare it to the theoretical mean of the distribution. We see that sample mean (5.003448) is extremely close to the theoretical mean (1/0.2=5).

```
#Theoretical mean
1/lambda
```

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

```
#Sample mean
mean(mean_data$mean)
```

```
## [1] 5.003448
```

Question 2

Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. The sample standard deviation (.7628219) is also very clost to the theoretical standard deviation (.7905694)

```
#Theoretical standard deviation - CLT
(1/lambda)/sqrt(n_samples)
```

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

```
#Sample standard deviation
sd(mean_data$mean)
```

```
## [1] 0.7628219
```

We also notice that sample variance (.5818972) and theoretical variance (.625) are close, but the simulation has a lower variance than the theoretical variance of the normal distribution.

```
#Theoretical variance
((1/lambda)/sqrt(n_samples))^2
```

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

```
#Sample variance
var(mean_data$mean)
```

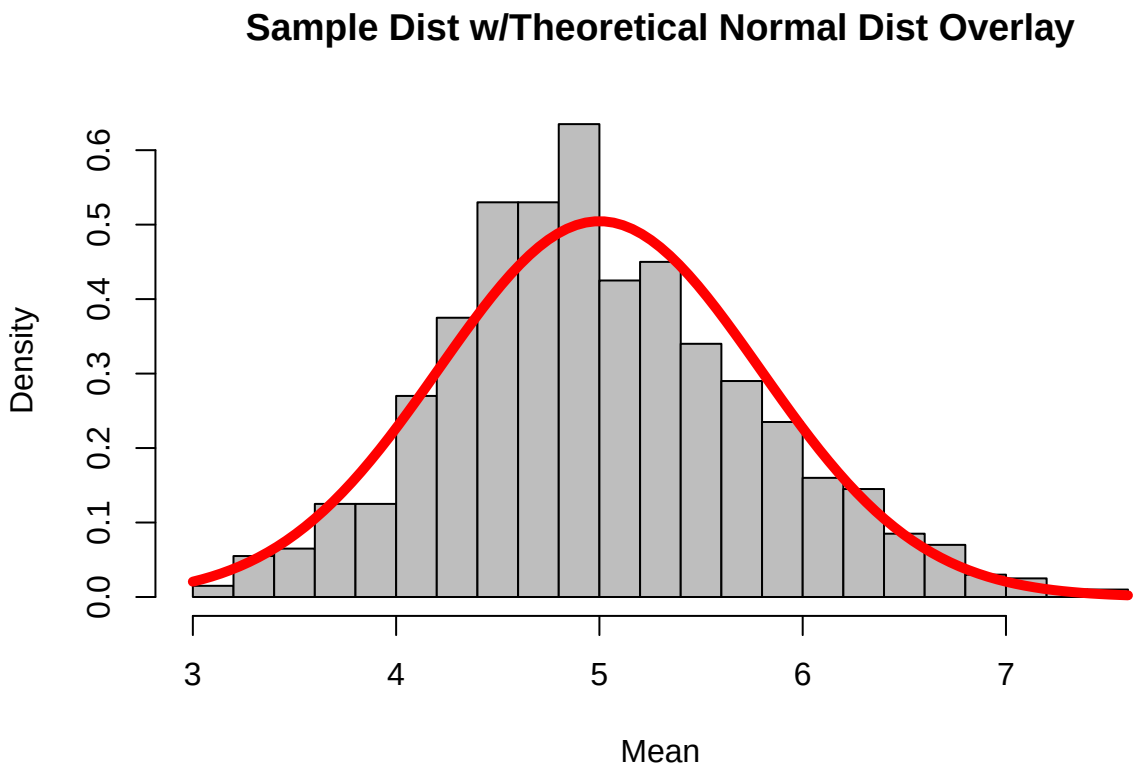
```
## [1] 0.5818972
```

Question 3

Show that the distribution is approximately normal. We will plot a histogram of the sample population and a overlay the simulated normal distribution to visualize how close they appear.

```
#plot the sample distribution
hist(mean_data$mean, breaks=16, freq=FALSE, col="gray", main="Sample Dist w/Theoretical Normal Dist Ove

#add the theoretical normal distribution curve on top in red
curve(dnorm(x, 1/lambda, (1/lambda/sqrt(n_samples))), col="red", lty=1, lwd=5, add=TRUE)
```



As we can see, the sample distribution follows the normal distribution fairly well. However, it is not an exact fit. Some of the simulation falls outside of the normal distribution (red overlay). Perhaps with more simulations we would get even closer.