Coursera Statistical Inference Assignment 1

Anthony Iannolo

4/27/2018

Introduction

In this assignment we are investigating the exponential distribution and comparing it to the Central Limit Theorem. The exponential distribution will be simulated in R using rexp(n, lambda) where lambda is the rate parameter.

The mean of the exponential distribution is 1/lambda and the standard deviation is the same 1/lambda. Lambda will be set to 0.2 for all the simulations. The requirement is to investigate the distribution of averages of 40 exponentials and perform 1000 simulations.

The assignment consists of three key parts:

1: Simulate the sample mean and then compare it to the theoretical mean

2: Show how variable the sample is compared to the theoretical variance of the distribution

3: Investigate whether the exponential distribution is approximately normal

Key Assumptions for this exercise:

- Simulated sample means are random and Independent Identically Distributed (iid)

- Normality improves with increased sample size.

Part 1 - Simulate the sample mean and then compare it to the theoretical mean

The code below executes a simulation of 40 samples for the exponential distribution and takes the mean of each sample. This is repeated 10,000 times.

- Key observations are the simulated mean is very close to the theoretical mean of the distribution. The upcoming histogram illustrates how close the means are and centered in the distribution.

```
set.seed(10)
num_of_sims <- 1000
lambda <-0.2
theory_mean <- 1/lambda
n <- 40

sim_data <- data.frame(ncol=2, nrow=num_of_sims)
names (sim_data) <- c("simulation#", "simulation_mean")

for (i in 1:num_of_sims) {
  sim_data[i,1] <- i
  sim_data[i,2] <- mean(rexp(n,lambda))
}
```

```
sim_mean <- mean(sim_data$simulation_mean)
```

```
## [1] "Simulated mean: 5.045"
```

```
## [1] "Theoretical mean: 5"
```

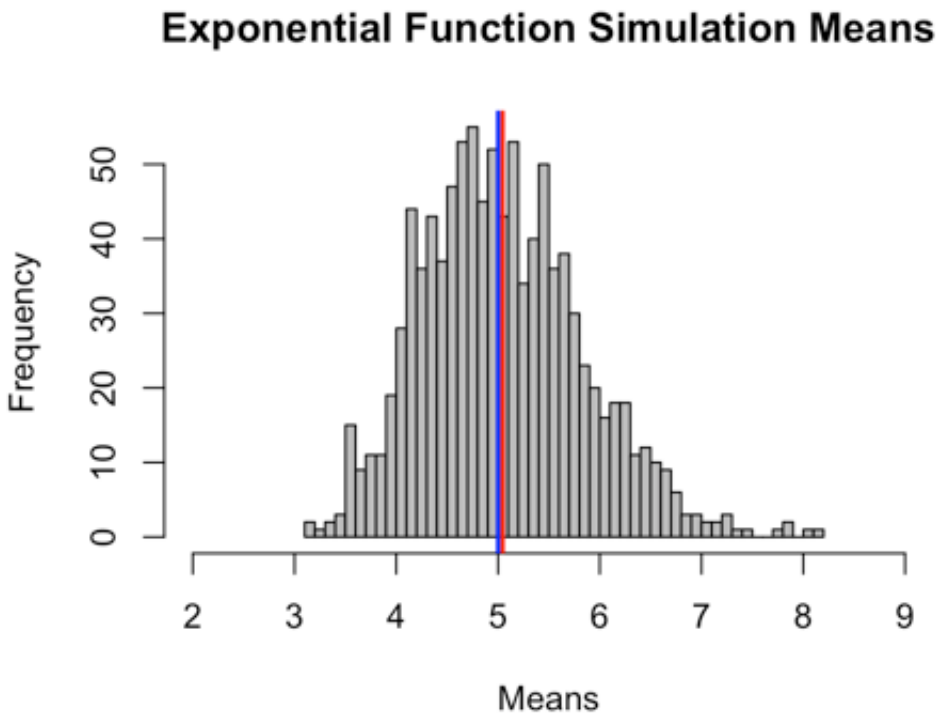
Create a histogram of the simulated means and plot the Sample and Theoretical Means against the Distribution

- Key observations in the Histogram below is it's general shape and how it seems to form the outlines of a normal distribution. Note how the simulated mean (red-line) is very close to the theoretical mean (blue-line).

- The Central Limit Theorem (CLT) states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases. The upcoming analyses will furhter highlight how the results of the simulation are inline with expectations of the CLT.

```
hist(sim_data$simulation_mean, breaks=40, xlim = c(2,9), main="Exponential Function S
imulation Means",
     xlab="Means", ylab="Frequency", col ="gray")

# plot a vertical red line at the mean of the sample means
    abline(v=mean(sim_mean), lwd="2", col="red")
    abline(v=mean(theory_mean), lwd="2", col="blue")
```



Part 2 - Show how variable the sample is compared to the theoretical variance of the distribution

Variance of Simulated Means versus Theoretical Variance

- Key observation shown in the output below, the simulated variance of the means and the theoretical variance are close alond with their standard deviations, further validating CLT.

```
## [1] "Theoretical standard deviation:  0.791"
```

```
## [1] "Simulated standard deviation:  0.798"
```

```
## [1] "Theoretical variance:  0.625"
```

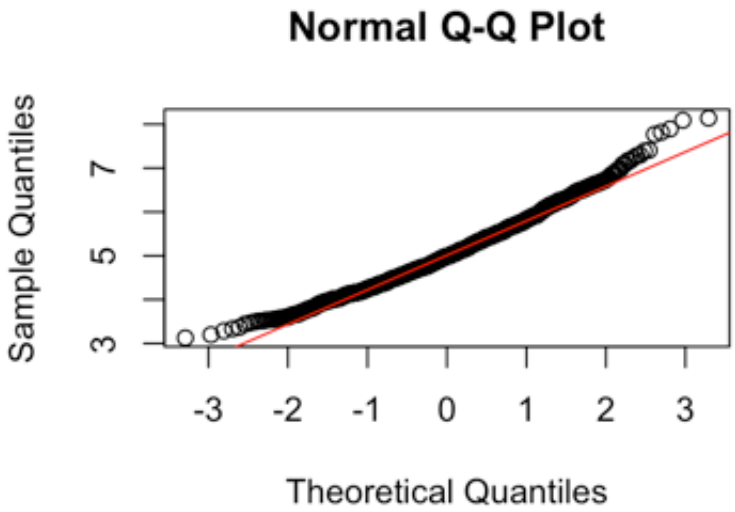
```
## [1] "Simulated variance:  0.637"
```

Part 3 - Investigate whether the exponential distribution is approximately normal.

With the Central Limit Theorem, the means of the sample simulations should follow a normal distribution.

- Key observation using the QQ Plots below is one can see the simulated data points are approximately normally disributed, as the data points follow the normal line. The upcoming confidence intervals serve as furhter proof.

```
qqnorm(sim_data$simulation_mean)
qqline(sim_data$simulation_mean, col = 2)
```



Confidence Intervals for the Simulated and Theoretical Mean

- Key observation using the confidence interval is the results are very close and indicate that 95% of the densities fall within 2 standard deviations of the means providing furhter evidence.

First is the Simulation confidence interval

```
sim_conf_interval <- (mean(sim_data$simulation_mean) +
                            c(-1,1)*1.96*sd(sim_data$simulation_mean)/sqrt(n))
```

```
## [1] "Simulated Confidence Interval:"
```

```
## [1] 4.798 5.292
```

Next is the Theoretical confidence interval

```
theory_variance <- ((1/lambda)/sqrt(n))^2
theory_conf_interval <- theory_mean + c(-1,1)*1.96*(sqrt(theory_variance)/sqrt(n))
```

```
## [1] "Theoretical Confidence Interval:"
```

```
## [1] 4.755 5.245
```

Summary

- The Central Limit Theorem (CLT) states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases.

In the above analyses the simulated means are close to the theoretical means and through furhter illustrations and confidence intervals it can be observed the averages of iid means become that of a standard normal and the CLT applies in this case.

Appendix including Code

1: Simulate the sample mean and then compare it to the theoretical mean

```
set.seed(10)
num_of_sims <- 1000
lambda <-0.2
theory_mean <- 1/lambda
n <- 40

sim_data <- data.frame(ncol=2, nrow=num_of_sims)
names (sim_data) <- c("simulation#", "simulation_mean")

for (i in 1:num_of_sims) {
  sim_data[i,1] <- i
  sim_data[i,2] <- mean(rexp(n,lambda))
}
```

```
sim_mean <- mean(sim_data$simulation_mean)
```

```
paste0("Simulated mean: ",round(sim_mean,3))
```

```
paste0("Theoretical mean: ", theory_mean)
```

Create a histogram of the simulated means and plot the Sample and Theoretical Means against the Distribution

```
hist(sim_data$simulation_mean, breaks=40, xlim = c(2,9), main="Exponential Function S
imulation Means",
     xlab="Means", ylab="Frequency", col ="gray")

# plot a vertical red line at the mean of the sample means
    abline(v=mean(sim_mean), lwd="2", col="red")
    abline(v=mean(theory_mean), lwd="2", col="blue")
```

Show how variable the sample is compared to the theoretical variance of the distribution

```
print(paste("Theoretical standard deviation: ", round( (1/lambda)/sqrt(n) ,3)))
theory_sd <- (1/lambda)/sqrt(n)


print(paste("Simulated standard deviation: ", round(sd(sim_data$simulation_mean) ,3))
)
sim_sd <- sd(sim_data$simulation_mean)


print(paste("Theoretical variance: ", round( ((1/lambda)/sqrt(n))^2 ,3)))
theory_variance <- ((1/lambda)/sqrt(n))^2


print(paste("Simulated variance: ", round(sd(sim_data$simulation_mean)^2 ,3)))
sim_variance <- sd(sim_data$simulation_mean)^2
```

Investigate whether the exponential distribution is approximately normal using QQ PLOTS

```
 qqnorm(sim_data$simulation_mean)
 qqline(sim_data$simulation_mean, col = 2)
```

Confidence Intervals for the Simulated and Theoretical Mean

First is the Simulation confidence interval

```
sim_conf_interval <- (mean(sim_data$simulation_mean) +
                            c(-1,1)*1.96*sd(sim_data$simulation_mean)/sqrt(n))
```

```
print("Simulated Confidence Interval:")
round(sim_conf_interval,3)
```

Next is the Theoretical confidence interval

```
theory_variance <- ((1/lambda)/sqrt(n))^2
theory_conf_interval <- theory_mean + c(-1,1)*1.96*(sqrt(theory_variance)/sqrt(n))
```

```
print("Theoretical Confidence Interval:")
round(theory_conf_interval,3)
```