Exponential Distribution vs the Central Limit Theorem

Wayne Hockensmith

January 31, 2017

Overview

The purpose of the project is to investigate the exponential distribution and compare it with the Central Limit Theorem (CTL) using R. The exponential distribution will be simulated in R with rexp(n, lambda) where λ is the rate parameter. The mean of exponential distribution is $1/\lambda$ and the standard deviation is $1/\lambda$. For all simulations within this project $\lambda = 0.2$. This project will investigate the distribution of averages of fourty exponetials with one thousand simulations.

Three specifics were given to compare.

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 approximimately normal.

Simulation

The simulation is set up with the following code:

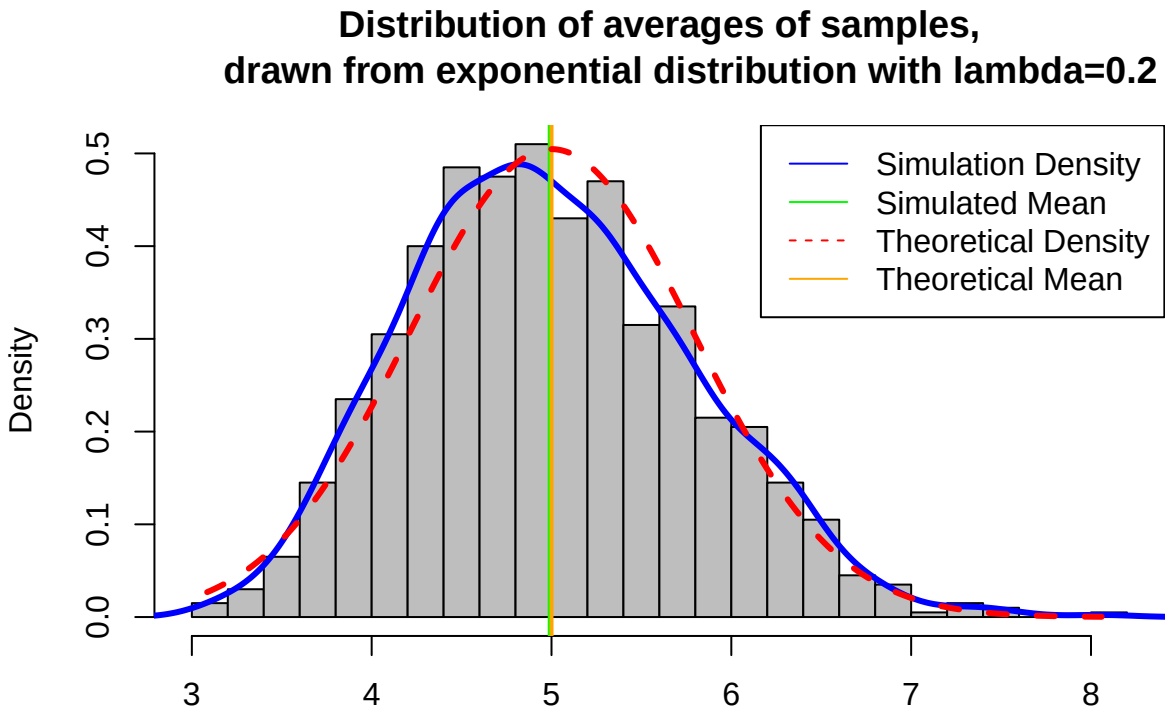
```
# Seed is set at 1.  Any number can be used, using a seed ensures the same starting point
#  each time.
set.seed(1)
# Prepair for the simulation as out lined in the overview:
n<-40
lambda <- 0.2
Simulations <- 1000
# Run the simulation:
 Exponential_simulations <-  matrix(rexp(Simulations*n, rate=lambda), Simulations, n)
 rmeans <- rowMeans(Exponential_simulations)
```

Results

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

```
# Build the histogram of averages:
 hist(rmeans, breaks=25, prob=TRUE,col="gray",
      main="Distribution of averages of samples,
      drawn from exponential distribution with lambda=0.2", xlab="")
# Add the density of the averages of the samples line:
 lines(density(rmeans), lwd=3,col="blue")
# Add the line for the theoretical center of distribution "means" (CTL):
abline(v=mean(rmeans), col="green",lwd=2)
# Add the line for the simulated mean:
abline(v=1/lambda, col="orange",lwd=2)
```

```r
# Add the theoretical density of the average of samples line (CTL):
x <- seq(min(rmeans), max(rmeans), length=1000)
y <- dnorm(x, mean=1/lambda, sd=(1/lambda/sqrt(n)))
lines(x, y, lwd=3, col="red", lty=8)
# Add the legend:
legend('topright', c("Simulation Density", "Simulated Mean","Theoretical Density",
       "Theoretical Mean"), lty=c(1,1,8,1), col=c("blue", "green","red", "orange"))
```




```r
# Calculate the means of the simulation (exponential distribution):
 mean(rmeans)
```

```
## [1] 4.990025
```

```r
# Calculate the means of the Theoretical Distribution (CTL):
 1/lambda
```

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

The graph above shows the distribution of averages of 40 exponentials (sample means "Green") is centered at 4.990025 which is close to the theoretical center of the distribution of 5 ("Orange"). Eventhough the two profiles are not excaclly the same the means of both are only off by 0.009975.

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

```r
 # Calculate the standard deviation of distribution of averages of 40 exponentials:
 sd(rmeans)
```

```
## [1] 0.7859435
```

```
# Calculate the standard deviation from analytical expression:
(1/lambda)/sqrt(n)
```

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

```
# Calculate the Variance of the sample mean:
var(rmeans)
```

```
## [1] 0.6177072
```

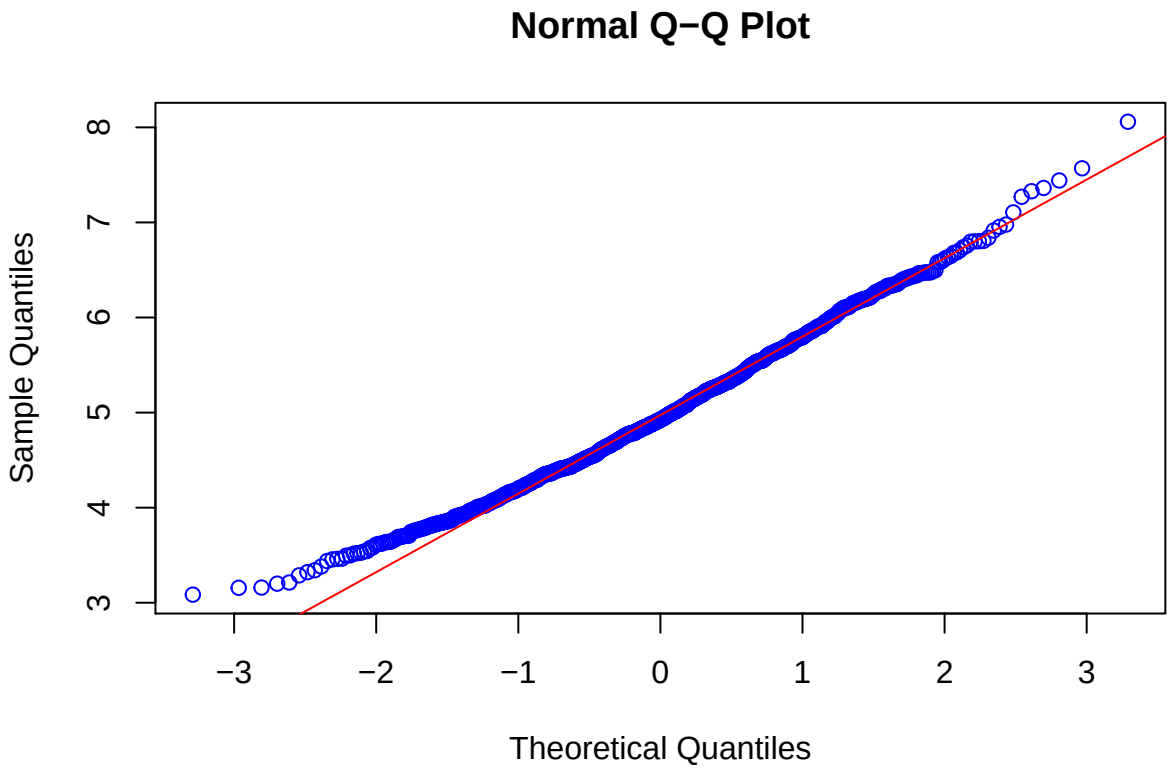
```
# Calculate thetTheoritcal variance of the distribution:
1/((lambda*lambda) * n)
```

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

The variance or variabilllity in the distribution of the 40 exponentials is almost the same as the theoretical vaariance of the distribuition. The sample means variance is 0.6177072 and the theoretical variance of the distribution is 0.625.
This is difference of 0.0072928.

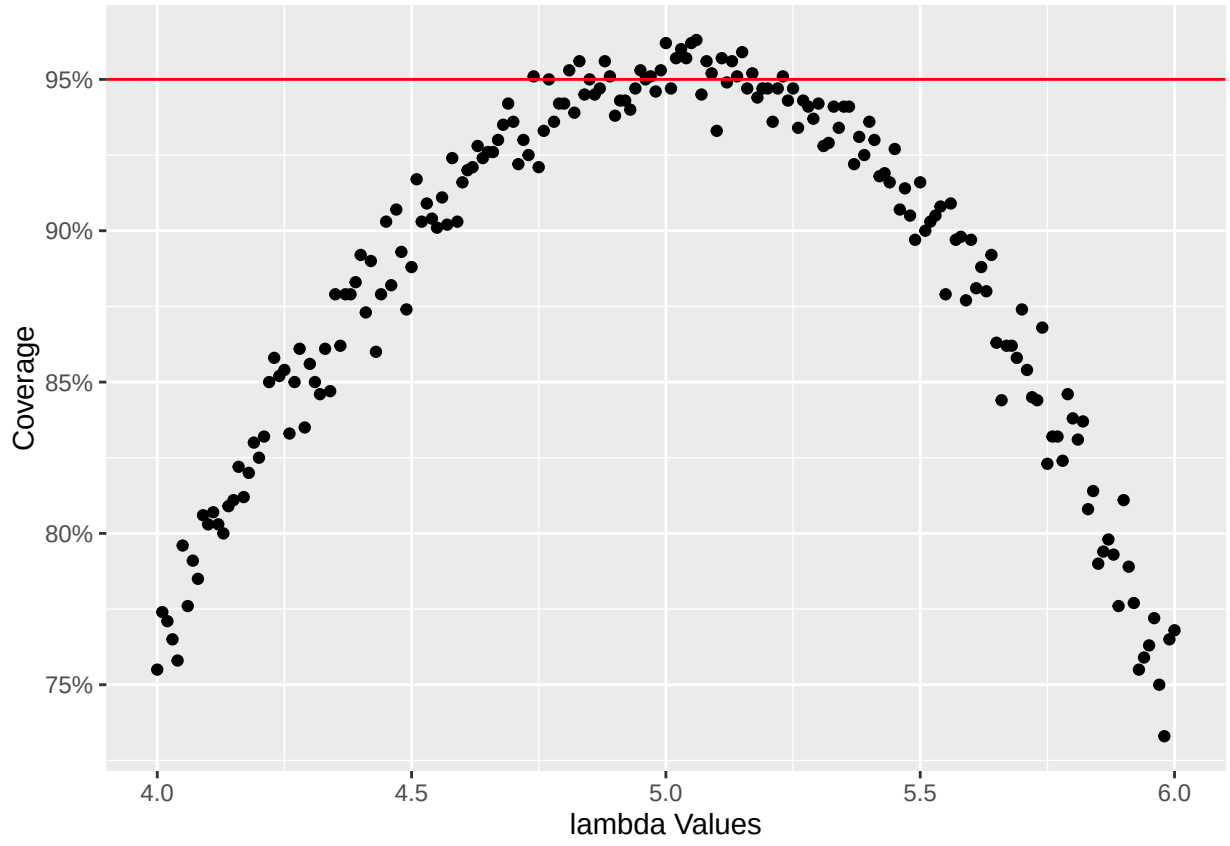
3. Show that the distribution is approximimately normal.

```
qqnorm(rmeans,col="blue"); qqline(rmeans,col="red")
```



The qplot varifies the data is distibuted close to normal.
The more circles that land on the line the closer to normally distributed the data is.

```r
# Evaluate coverage of the confidence interval:
lambda_Values <- seq(4, 6, by=0.01)
Coverage <- sapply(lambda_Values, function(lamb) {
    mu_hat <- rowMeans(matrix(rexp(n*Simulations, rate=0.2),Simulations, n))
    lowlim <- mu_hat - qnorm(0.975) * sqrt(1/lambda**2/n)
    uplim <- mu_hat + qnorm(0.975) * sqrt(1/lambda**2/n)
    mean(lowlim < lamb & uplim > lamb)
})
# Plot coverage:
# Load two needed libraries first:
library(ggplot2)
library(scales)
qplot(lambda_Values, Coverage) + geom_hline(yintercept=0.95, col=2)  +
  scale_y_continuous(breaks=seq(0, 5, 0.05),labels = scales::percent) +
  xlab("lambda Values")
```



The 95% confidence intervals for the rate parameter (λ) to be estimated ($\hat{\lambda}$) are $\hat{\lambda}lower = \hat{\lambda}(1 - \frac{1.96}{\sqrt{n}})$ and $\hat{\lambda}upper = \hat{\lambda}(1 + \frac{1.96}{\sqrt{n}})$. As can be seen from the plot above, for selection of $\hat{\lambda}$ around 5, the average of the sample mean falls within the confidence interval at least 95% of the time. Note that the true rate, λ is 5.