Statistical Inference Course Project

Kie Gouveia

June 20, 2015

Project Purpose

The purpose of this report is to investigate the exponential distribution in R and to compare it with the Central Limit Theorum. this will be done through a series of calculations and simulations to illustrate the behaviour of the distribution and theorum.

The Exponential Distribution describes the time between events in a Poisson process in which event occur continuously and independently at a constant average rate.

First, we can illustrate the difference between the theoretical mean of the exponential distribution and the average of sample means from simulated data.

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

```
lambda = 0.2 # setting lambda equal to 0.2
nexp = 40 # number of exponentials being generated is equal to 40
set.seed(124) # allows for reproducibility of results

# Theoretical Mean
## This is based upon the theoretical calculation for mean
## of an exponential function.
theoreticalMean <- (1/lambda)
  print(theoreticalMean)
```

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

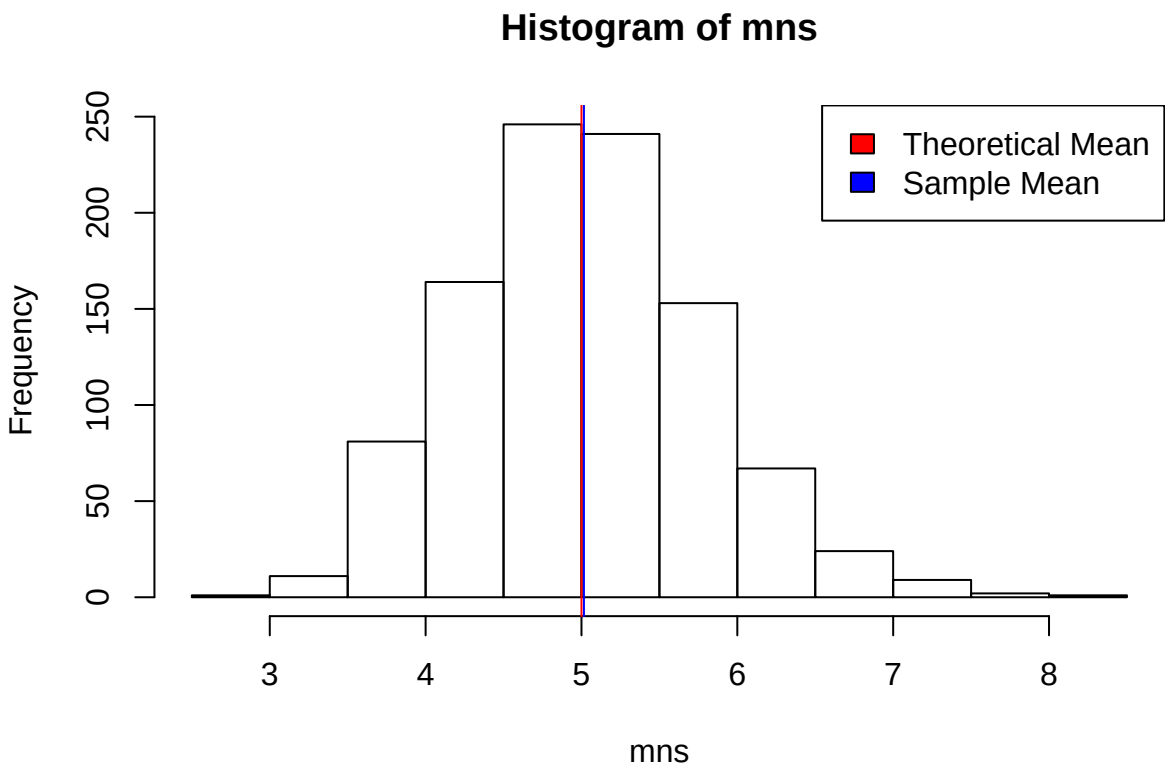
```
# Sample Mean
## This is obtained by simulating 40 random exponential variables 1000 times,
## taking the mean of each group. Finally, we obtain the mean of these averages
mns = vector() # set means equal to an epty vector.

for(i in 1:1000)
  mns = c(mns, mean(rexp(nexp, lambda)))

# Calculate the mean of the distribution that we just created.
sampleMean <- mean(mns)
print(sampleMean)
```

```
## [1] 5.015703
```

```
# Plot a histogram of the sample means. Use vertical lines to
# indicate where the theoretical and sample means lie.
hist(mns)
abline(v = sampleMean, col = "blue")
abline(v = theoreticalMean, col = "red")
legend("topright",
       legend=c("Theoretical Mean", "Sample Mean"),
       fill=c("Red", "Blue"))
```



This execise shows us that the average of these sample means is incredibly close to the predicted (theoretical) mean. In this case the average of our sample means is 5.02 compared with a theoretical mean of 5. This is due to the fact that the arithmetic mean is an unbiased predictor, meaning that the eventually converges on the true (theoretical) mean, given enough simulations.

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

Theoretical Variance

This is based upon the theoretical forumala for variance of an exponential distribution.

```
theoreticalVariance <- 1/(lambda^2)
print(theoreticalVariance)
```

```
## [1] 25
```

Sample Variance

This is obtained by calculating the variance of our previous distribution of sample mean (mns).

```
sampleVariance <- var(mns)
print(sampleVariance)
```

```
## [1] 0.5985168
```

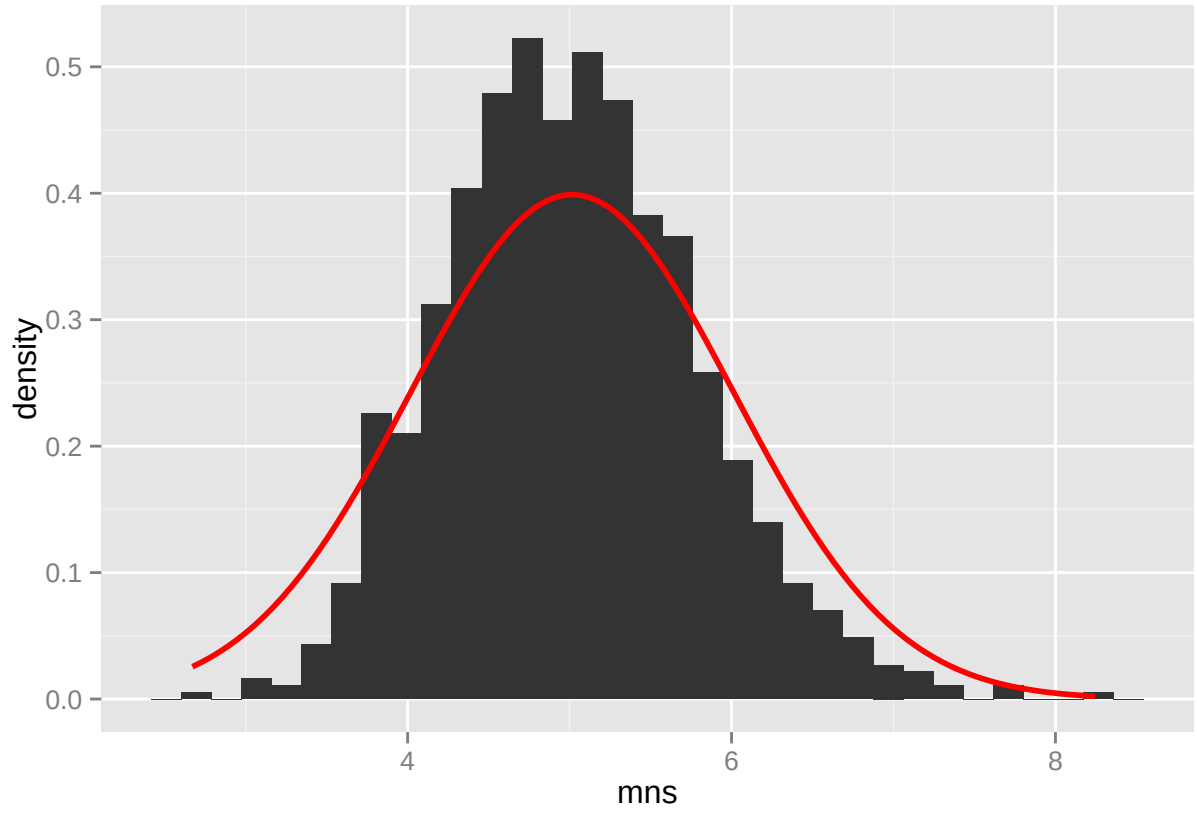
Here, we can see that the theoretical variance of the exponential distribution is much higher than the variance among our sample means. This is due to the fact that the average distance away from the mean decreases as the number of simulations increases.

3. Show that the distribution is approximately normal.

Finally, we can compare the our distibution of sample means with the normal distibution to confirm their similarity. Below, the red curve depicts a normal curve overlayed on the histogram of sample means for comparison.

```
library(ggplot2)
mns <- as.data.frame(mns)

ggplot(data = mns, aes(x = mns)) +
    geom_histogram(aes(y = ..density..)) +
  stat_function(fun = dnorm, colour = "red", size=1, args = list(mean = sampleMean))
```



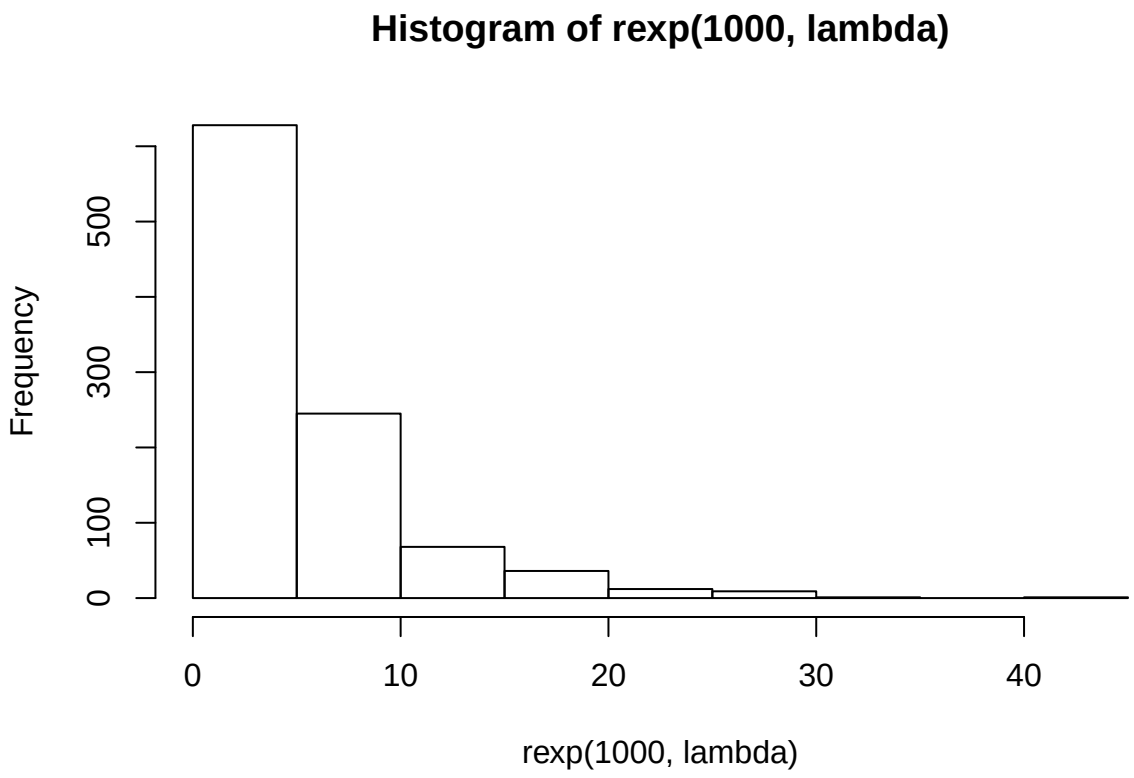
This depicts the distribution of the means of 1000 simulations which each contained 40 random exponential variables. It is evident that this histogram has taken on the approximate form of the normal distiribution (characterized by a symmetrical, bell-shape). This behaviour is predicted by the Central Limit Theorum.

The reason this quality emerges is because for each of the 1000 distributions, the average of the 40 exponential variables is taken. With the theoretical mean functioning as an anchor point, the means of the simulation averages will flunctuate around this central value, creating the highest probability density at the theoretical mean. As you move further from the theoretical mean value, denisty diminishes, creating the shape of a normal distribution.

Appendix

Appendix A: The Exponential Function Visualized

```
hist(rexp(1000, lambda))
```



This chart depicts the distribution of 1000 exponential variables, contains most of its density on the left hand side of the distribution, with the density dwindling as you look further to the right.

Intuitively, this makes sense, the higher the rate of an event (lambda), the shorter the time will be before the next event occurs.