The Exponential Distribution in R and the Central Limit Theorem

James Sheldon

May 23, 2016

```r
sessionInfo()
```

```
## R version 3.3.0 (2016-05-03)
## Platform: x86_64-apple-darwin13.4.0 (64-bit)
## Running under: OS X 10.11.5 (El Capitan)
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] ggplot2_2.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5      digest_0.6.9     plyr_1.8.3       grid_3.3.0
##  [5] gtable_0.2.0     formatR_1.4      magrittr_1.5     evaluate_0.9
##  [9] scales_0.4.0     stringi_1.0-1    rmarkdown_0.9.6  tools_3.3.0
## [13] stringr_1.0.0    munsell_0.4.3    yaml_2.1.13      colorspace_1.2-6
## [17] htmltools_0.3.5  knitr_1.13
```

Overview

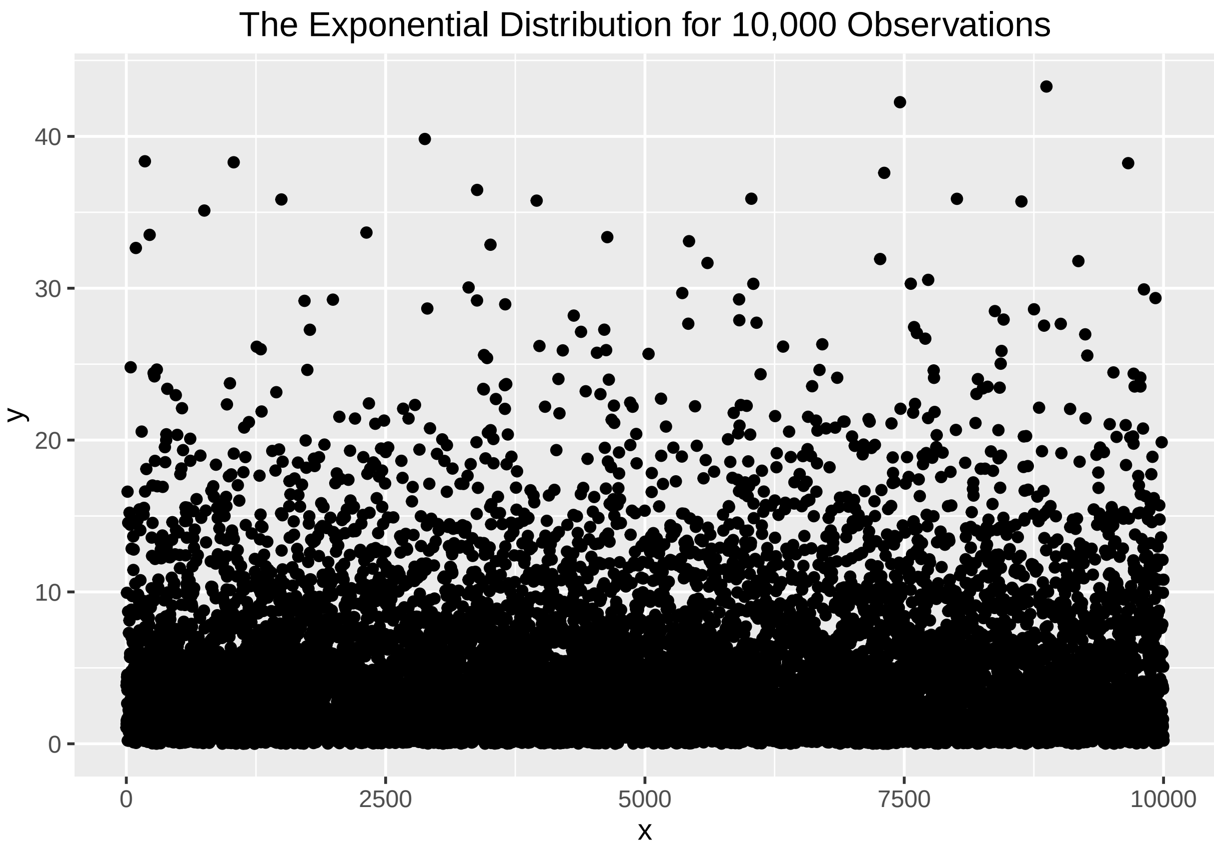
This is part 1 of the final course project for the statistical inference class from Coursera presented by Johns Hopkins University. The purpose of this exercise is to explore how the Exponential Distribution relates to the Central Limit Theorem using one thousand mean simulations.

Simulation

Let's begin by generating and plotting 10,000 observations based on the exponential distribution.

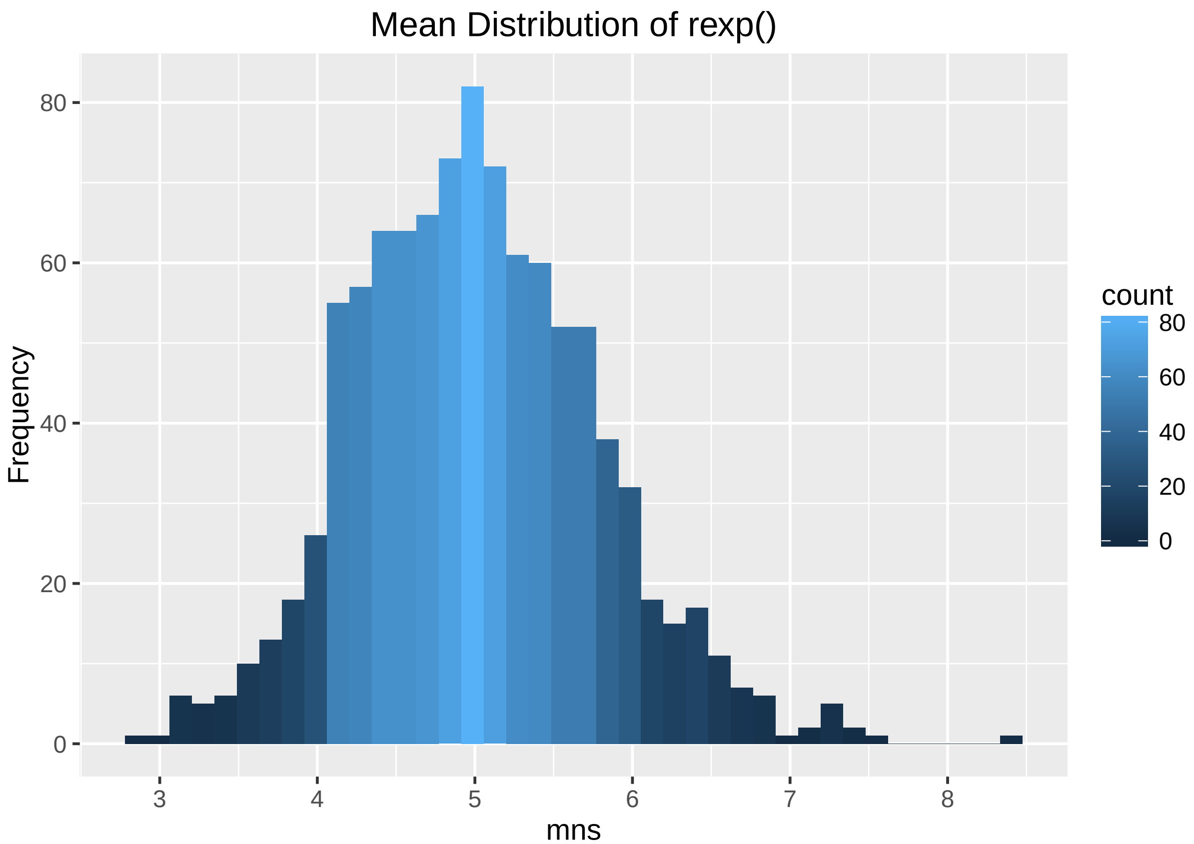
```r
set.seed(10736)
lambda <- 0.2 # fixed rate parameter
n <- 40 # number of exponentials
sim <- 1000 # fixed number of simulations per instruction

# the exponential distribution
qplot(x = c (1:10000), y = rexp(10000,lambda), geom = "point", main = "The Exponential Distribution for
```



That's fun. Now, how is this data distributed? Based on the assignment notes on the submission page, let's plot 1,000 averages of these randomly generated observations.

```
# generate the collection of means for 1,000 simulations of the exponential distribution
mns <- NULL
for (i in 1 : sim) mns <- c(mns, mean(rexp(n, lambda)))
qplot(mns, geom = "histogram", bins = 40, fill = ..count.., ylab = "Frequency", main = "Mean Distributi
```

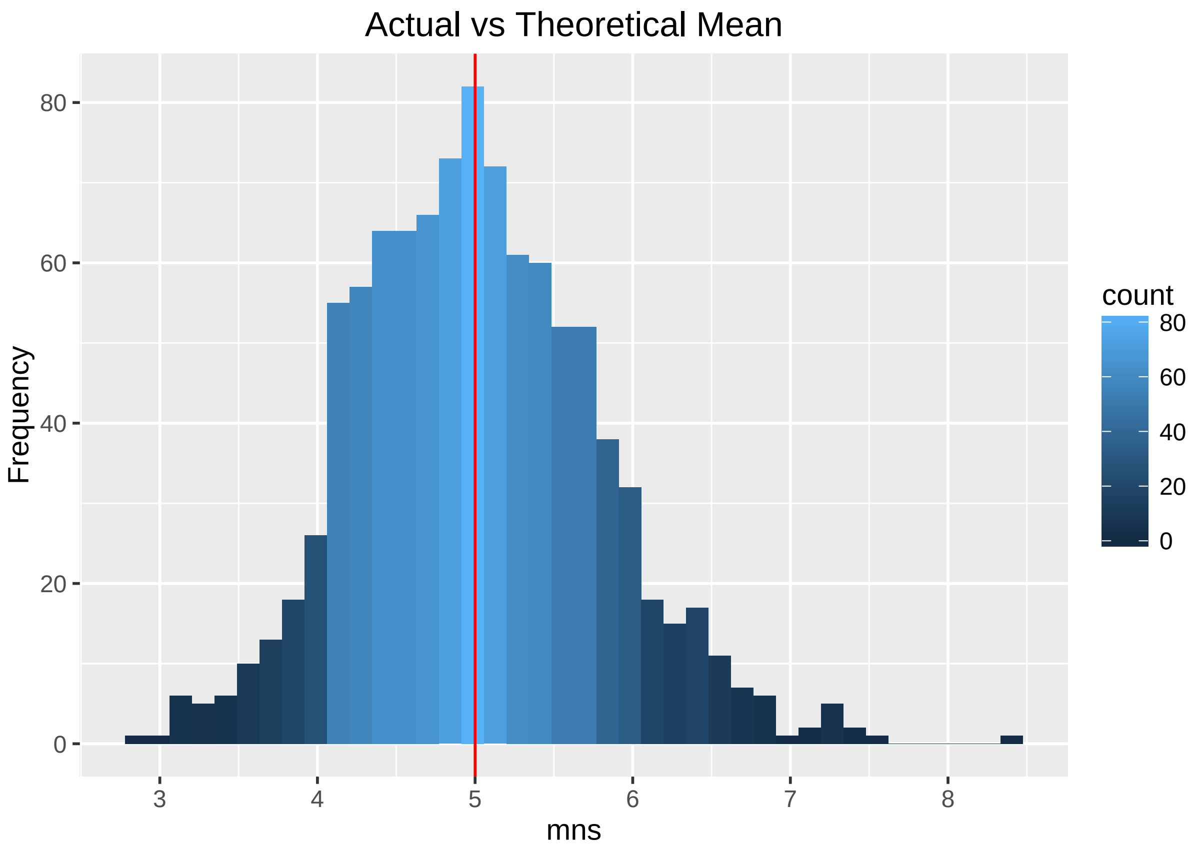


We can see the normal distribution of these means taking shape. And based on the color density it looks like the central tendency is at or near a mean of 5. We'll explore that further next.

Sample Mean Versus Theoretical Mean

Since the theoretical mean of the exponential distribution is 1/lambda and in our examples we are using lambda = 0.2, the sample and theoretical mean should both be at or near 5. Let's see what happens.

```
qplot(mns, geom = "histogram", bins = 40, fill = ..count.., ylab = "Frequency", main = "Actual vs Theore
```



It looks like we're in pretty good shape. In mathland, the sample mean we have is

```
round(mean(mns),2)
```

```
## [1] 5.01
```

so we're right on track.

Sample Variance Versus Theoretical Variance

Since the standard deviation of the exponential distribution is (1/lambda)/sqrt(n) we see the following for the standard deviation:

```
print(paste("Theoretical: ", round((1/lambda)/sqrt(n),4), ", Sample", round(sd(mns),4)))
```

```
## [1] "Theoretical:  0.7906 , Sample 0.7727"
```

and for the variances we have

```
print(paste("Theoretical: ", (1/lambda)^2/n, ", Sample", round(var(mns),4)))
```

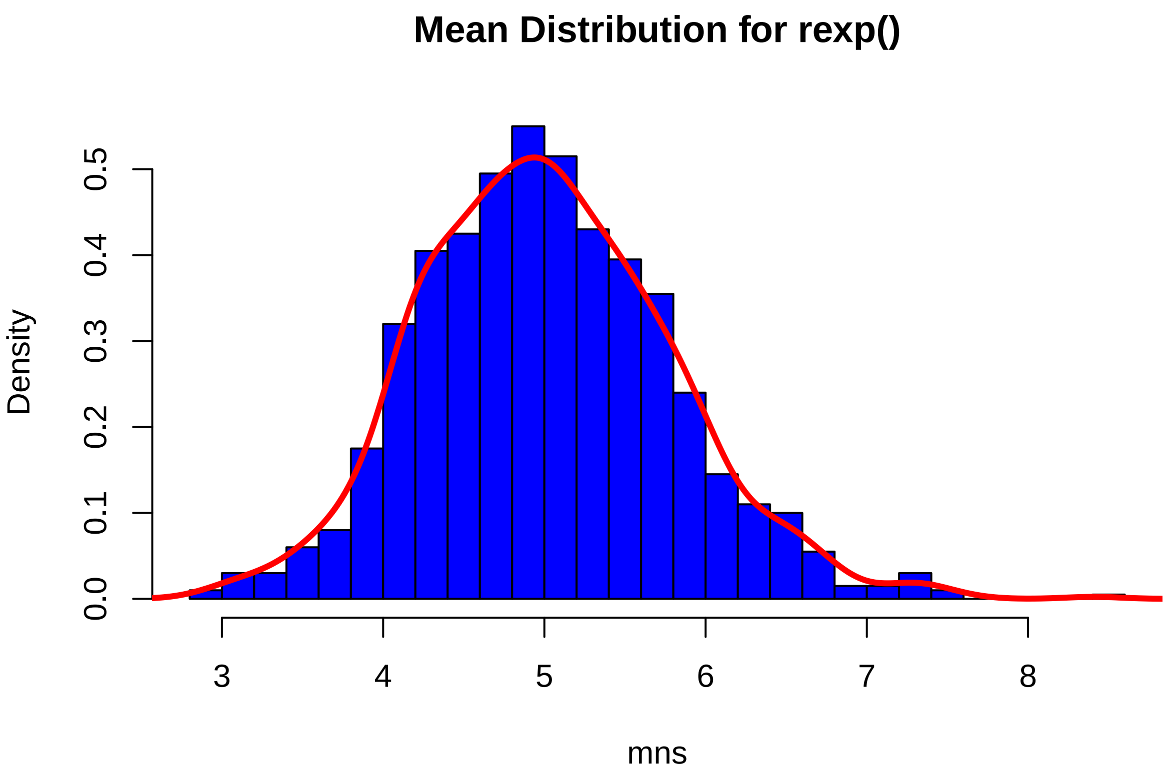
```
## [1] "Theoretical:  0.625 , Sample 0.5971"
```

Based on the coursework, this suggests we're on track with the theoretical values.

Is the Distribution of Means Normal?

Finally we were asked to explore whether or not the exponential distribution is close to normal. It certainly would seem as though it should be based on the Central Limit Theorem.

```r
hist(mns, prob=TRUE, col="blue", main="Mean Distribution for rexp()", breaks=20)
lines(density(mns), lwd=3, col="red")
```



Based on this graph it certainly seems as though the theory matches up with our work.