# Simulations and Inferential Data Analysis, Statistical Inference: Simulations

*Scott Zuehlke*

*December 3, 2016*

This report was produced as part of the Coursera Data Science Specialization, Statistical Inference course. All code was produced on a Mac running OS X 10.11.6, with RStudio version 0.99.902. This report was produced as part of the Coursera Data Science Specialization, Statistical Inference course. The purpose of this report is to twofold: Part one is to run simulations against the exponential distribution, therefore confirming the bootstrap method works, by checking sample distribution mean and spread, obtained by simulation, against the theoreteical mean and spread. Part two is to run basic inferential analysis by testing the effects of dosage and supplmenet on Guinea Pig tooth growth.

## Part 1: Simuliation of Exponential Distribution

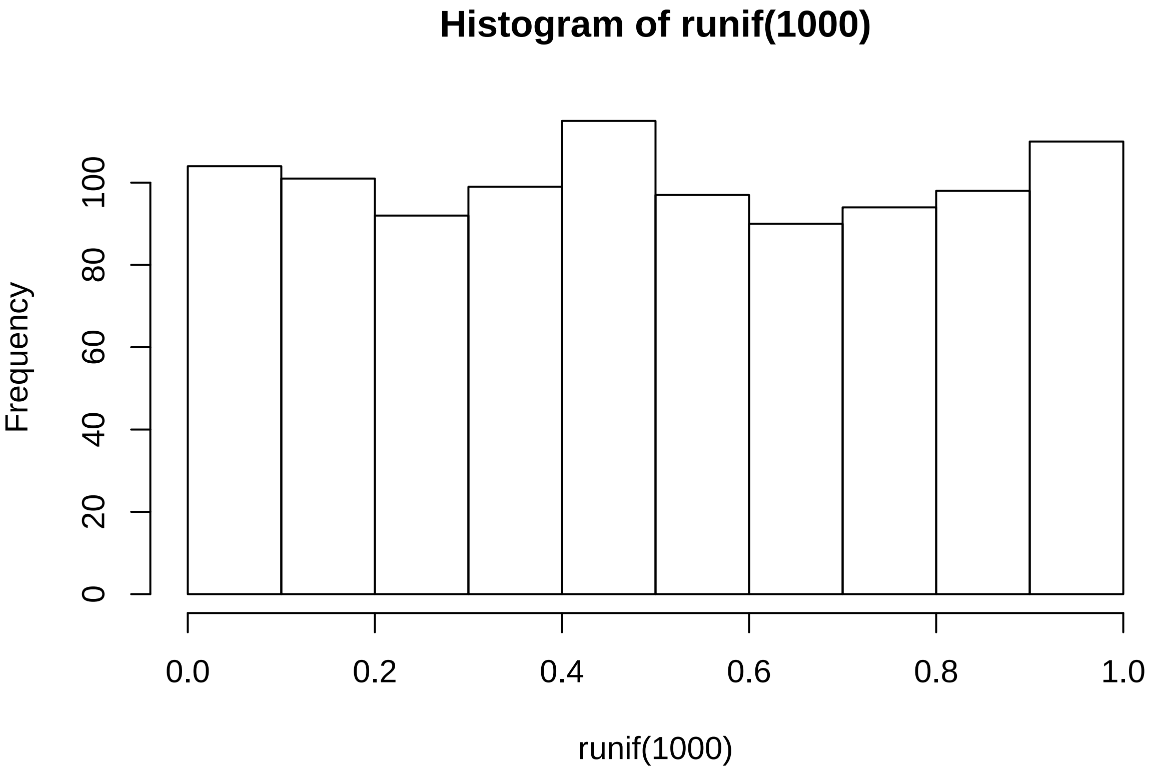
In this project you will investigate the exponential distribution in R and compare it with the Central Limit Theorem. The exponential distribution can be simulated in R with rexp(n, lambda) where lambda is the rate parameter. The mean of exponential distribution is 1/lambda and the standard deviation 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.

Illustrate via simulation and associated explanatory text the properties of the distribution of the mean of 40 exponentials. You should

Show the sample mean and compare it to the theoretical mean of the distribution. Show how variable the sample is (via variance) and compare it to the theoretical variance of the distribution. Show that the distribution is approximately normal. In point 3, focus on the difference between the distribution of a large collection of random exponentials and the distribution of a large collection of averages of 40 exponentials.

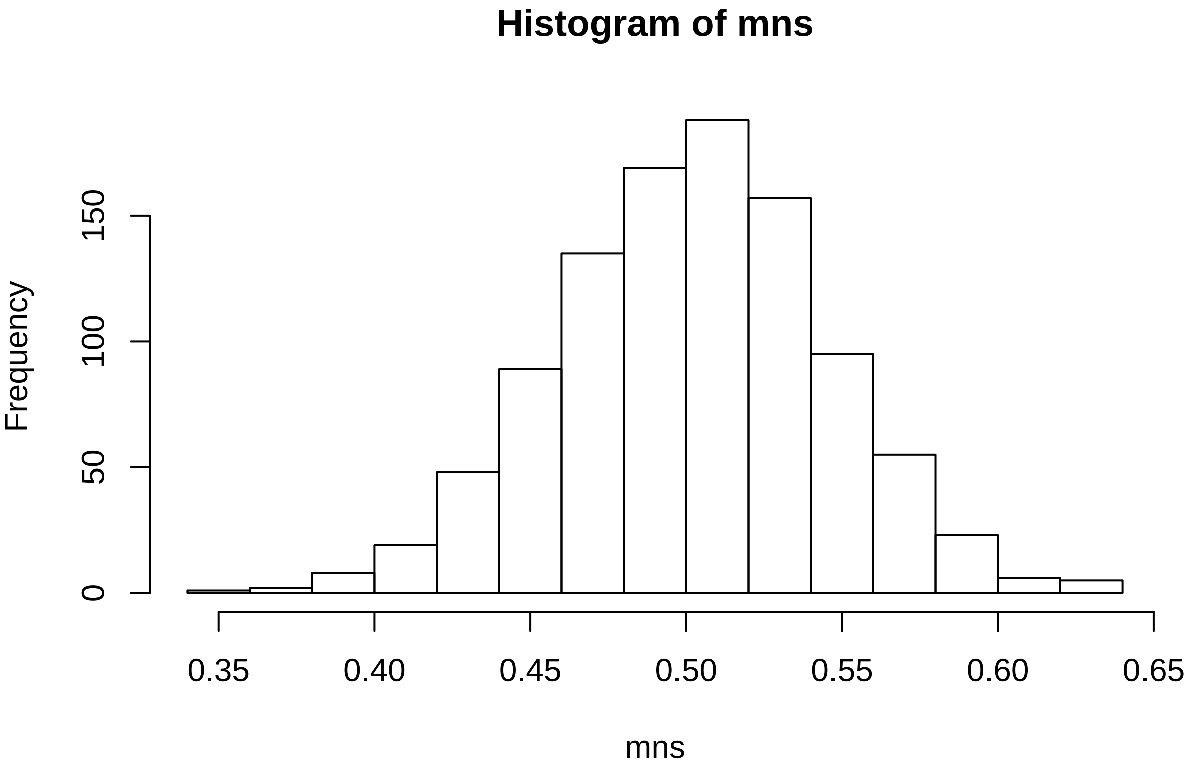
As a motivating example, compare the distribution of 1000 random uniforms

```
hist(runif(1000))
```



and the distribution of 1000 averages of 40 random uniforms

```
mns = NULL
for (i in 1 : 1000) mns = c(mns, mean(runif(40)))
hist(mns)
```



This distribution looks far more Gaussian than the original uniform distribution! This exercise is asking you to use your knowledge of the theory given in class to relate the two distributions.

**From here out, the work completed is that of Scott ZUehlke, the author. The previosu was taken from the Assignment page.**

First, load ggplot2 to take advantage of its plotting capabilities.

```
library(ggplot2)
```

Then, start with running 1000 simulations of 40 exponentials with lambda 0.2

```
set.seed(35)
lambda <- 0.2
sims <- 1000
n <- 40
```

Create a matrix of exponenetial simulations.

```
expMatrix <- data.frame(x = sapply(sims, function(x) {mean(rexp(n, lambda))}))
expSims <- matrix(rexp(n*sims,lambda),nrow=sims)
expSimsMeans <- data.frame(means=apply(expSims,1,mean))
```

```
simMean <- mean(expSimsMeans$means)
simMean
```

```
## [1] 4.98931
```

```
trueMean <- 1/lambda
trueMean
```

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

```
pct_diff <- (simMean - trueMean) / trueMean
pct_diff
```

```
## [1] -0.002138084
```

```
expSimsSD <- sd(expSimsMeans$means)
expSimsSD
```

```
## [1] 0.7978989
```

```
expSimsVar <- var(expSimsMeans$means)
expSimsVar
```

```
## [1] 0.6366427
```

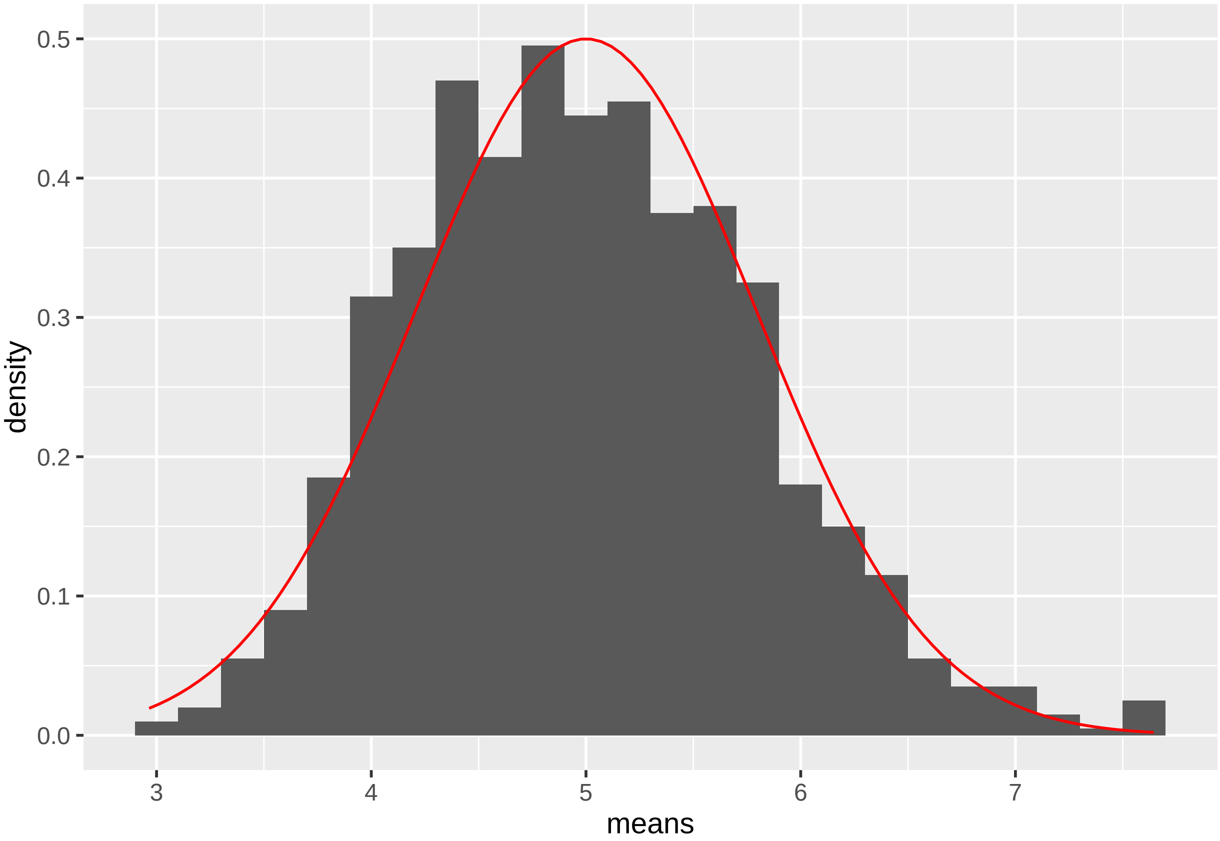
```
trueSD <- 1/lambda/sqrt(n)
trueSD
```

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

```
trueVar <- trueSD^2
trueVar
```

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

```
library(ggplot2)
ggplot(data = expSimsMeans, aes(x = means)) +
  geom_histogram(aes(y=..density..), binwidth = 0.20) +
  stat_function(fun = dnorm, args = list(mean = trueMean, sd = sd(expSimsMeans$means)), col='red')
```



Looking at the plot, the mean of the histogram, which is our simulated run of exponentials with lambda = .2, is center very close to 5. This is no coincidence, as this is the theoretical mean. The spread is also very close to the imposed normal curve. This confirms that, with enough simulations, the mean and variance of the exponential approaches the normal distribution, as the bootstrapping method describes.

Conclusion:

In the simulations in part 1, we did confirm that, given enough simulations, the distribution of means and standard deviations, does conform to the normal distribution, as described by the central limit theorem.