# Statistical Inference Course Project:

## *A Comparison of the Exponential Distribution and Central Limit Theorem*

*Erik Cornelsen*

*March 25, 2016*

# Project Overview:

We will use R to investigate the exponential distribution and the Central Limit Theorem. First we will simulate data for the exponential distribution. Then we'll use that data to compare the actual statistics with the expected theoretical statistics to:

1. compare the sample mean and the theoretical mean of the distribution
2. compare the sample variance and the theoretical variance of the distribution
3. show that the distribution is approximately normal

# Simulate Data:

First we generate/simulate data for the exponential distribution. We run a series of 1000 simulations with 40 observations in each simulation and using rexp() to generate random deviates.

```
library(ggplot2)

nsim            <- 1000         #number of simulations
nobs            <- 40           #number of observations per simulation
lambda          <- 0.2          #lambda is the rate parameter.
set.seed(84043)                 #setting the seed is needed when generating random deviat
es

#generate a matrix with 40 cols for the observations and 1000 rows for the simulations
#use rexp() to simulate the exponential distribution
simdata <- matrix( rexp(n=nsim*nobs, rate=lambda), nsim, nobs)
df <- data.frame(simdata)
#str(simdata)
```

## Sample Mean versus Theoretical Mean:

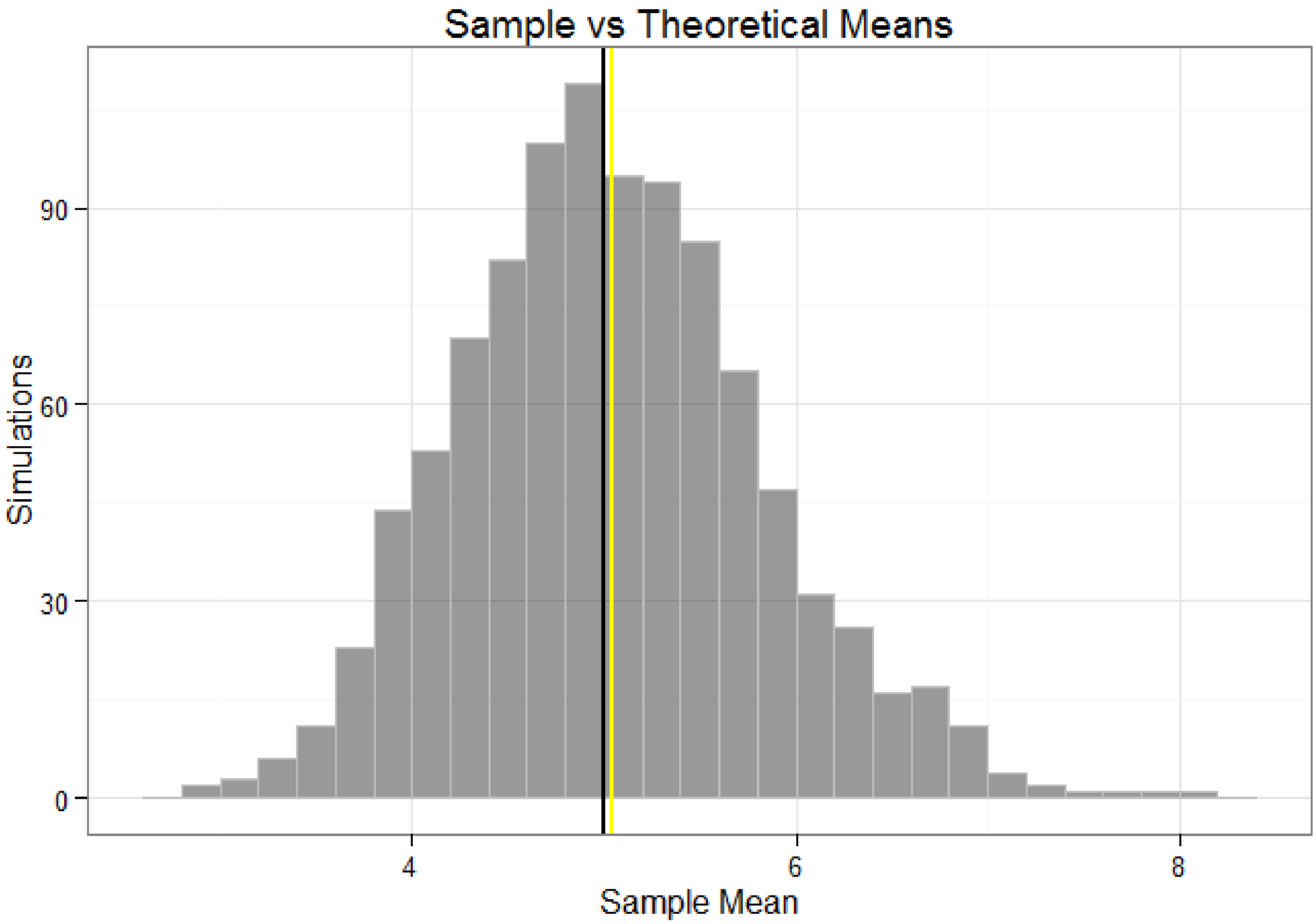
Include figures with titles. In the figures, highlight the means you are comparing. Include text that explains the figures and what is shown on them, and provides appropriate numbers.

```
theoryMean      <- 1/lambda     # (1/λ) = theoretical mean
df['rowMean']   <- rowMeans(simdata)
actualMean      <- mean(df$rowMean)
```

A quick comparison between the theoretical mean (5) and sample mean (5.0403413) show that they are very close after 1000 simulations.

Below is a histogram showing all of the sample means, the theoretical mean (black) and the mean of the sample means (yellow).

```
library(ggplot2)
g <- ggplot(df,  position="identity", aes(rowMean)) +
    geom_histogram(alpha=.5,color="gray", binwidth=8/40 ) +
    geom_vline(xintercept=theoryMean, colour="black", size=1) +
    geom_vline(xintercept=actualMean, colour="yellow", size=1) +
    theme_bw() +
    labs(title="Sample vs Theoretical Means", x="Sample Mean", y="Simulations")
print(g)
```



# Sample Variance versus Theoretical Variance:

We’ll now compare the variance of the sample means to the theoretical variance of the population. The theoretical variance will be determined with the formula: σ2=((1/λ)^2)/n. Where n is the number of observations.

```
theoryVar <- ((1/lambda)^2)/nobs
actualVar <- var(df$rowMean)
```

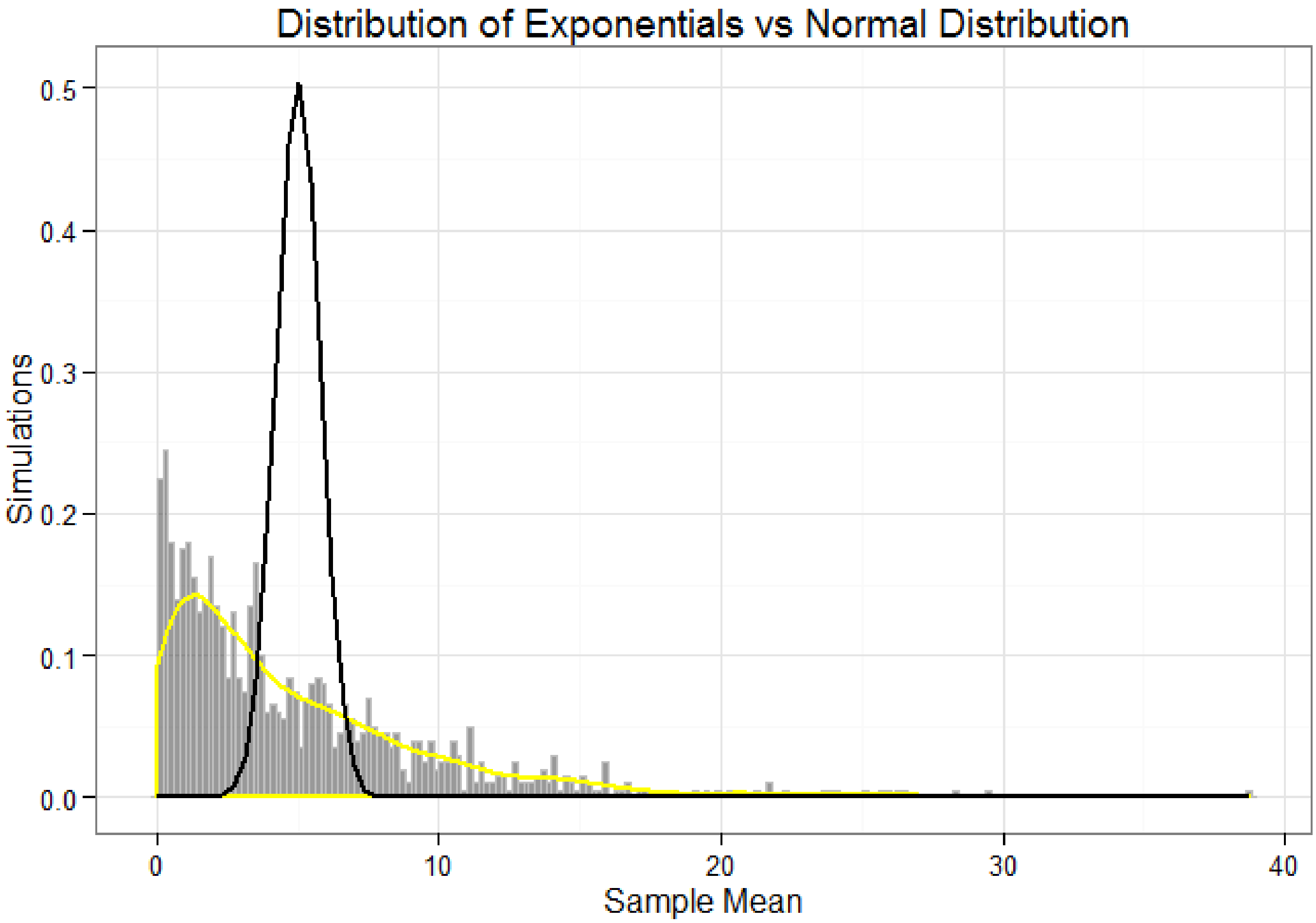
The theoretical variance (0.625) and the sample variance (0.6264955) are very close. This is a good thing, and indicates that the actual distance of the sample means from the central mean is consistent with the theoretical variance.

# Distribution Comparisons:

If you simply took a large collection of random exponentials you would not get a normal distribution. However, The Central Limit Theorem (CLT) states that the distribution of averages is often normal, even if the distribution that the data is being sampled from is very non-normal. Said another way, if you take the mean of many samples of a population (x-bars) and then plot those x-bars on a histogram you will find that their distribuition is aproximatly normal. The plots below illustrate how this plays out with the samples we’ve generated.

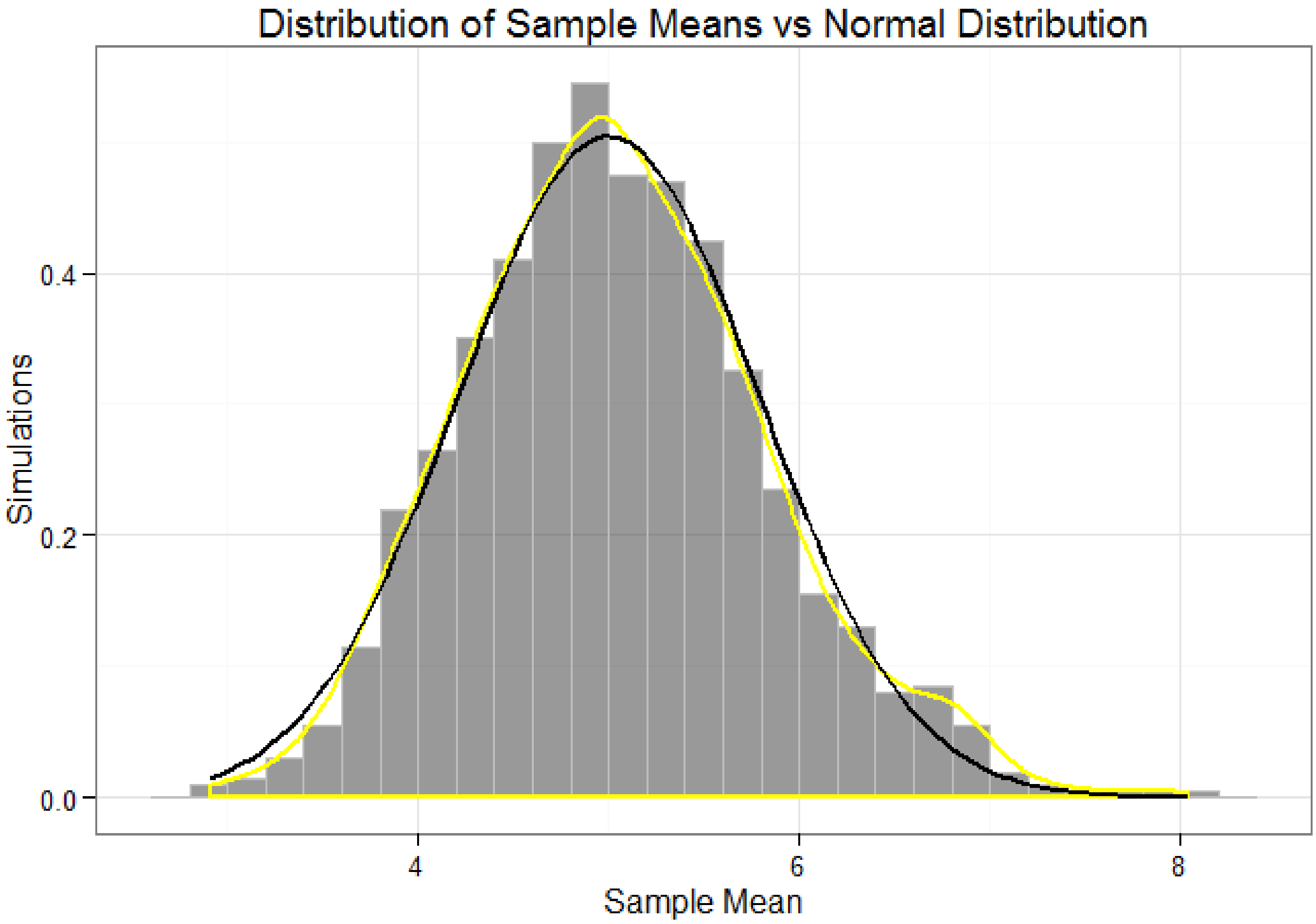
In our first plot the yellow line shows a 1000 random exponentials while the black line shows a normal distribution. You can see that they are not at all similar.

```
g <- ggplot(df,  position="identity", aes(X1)) +
    geom_histogram(aes(y=..density..), alpha=.5,color="gray", binwidth=8/40 ) +
    geom_density(color="yellow",size=1) +
    stat_function(fun=dnorm, colour="black", size=1, args=list(mean=theoryMean, sd=sqrt(t
heoryVar))) +
    theme_bw() +
    labs(title="Distribution of Exponentials vs Normal Distribution", x="Sample Mean", y
="Simulations")
print(g)
```



In our second plot the yellow line shows the sample mean distributions (x-bars) while the black line shows a normal distribution. You can see that they are very similar, and a good illustration of the Central Limit Theorem.

```
g <- ggplot(df,  position="identity", aes(rowMean)) +
    geom_histogram(aes(y=..density..), alpha=.5,color="gray", binwidth=8/40 ) +
    geom_density(color="yellow",size=1) +
    stat_function(fun=dnorm, colour="black", size=1, args=list(mean=theoryMean, sd=sqrt(t
heoryVar))) +
    theme_bw() +
    labs(title="Distribution of Sample Means vs Normal Distribution", x="Sample Mean", y
="Simulations")
print(g)
```


Distribution of Sample Means vs Normal Distribution
Simulations
0.4
0.2
0.0
4
6
8
Sample Mean