Central Limit Theorem

Introduction

The Central Limit Theorem (CLT)states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the *sample size increases*. This means that it is possible to get an approximation of mean and standard deviation for the whole distribution with only one observed average and without knowing the population distribution.

The result of increasing sample size is that

$$\frac{\bar{X}_n - k}{r/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - k)}{r} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

converges on a distribution similar to that of the standard normal range for large n

The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(k, r^2/n)$

```
library(knitr)
library(dplyr)
library(stats)
library(reshape2)
library(ggplot2)
```

Test the Central Limit Theorem by simulation:

Some Initial Assumptions

The theoretical mean (k) and variance Var of exponential distribution with parameter j are respectively $1/j$ and $1/j^2$,

If the CLT is true, then:

- The mean of our simulation $\bar{X}_n$ should approach k, i.e., $\frac{1}{0.2}$ or: 5
- The variance of our simulation, Var_n should approach $1/j^2$, i.e, $\frac{1}{0.2^2}$ or 25
- The variance of our sample mean is the Var_n divided by the our sample number* $Var_n(\bar{X})$ should approach r^2/n, i.e. $\frac{1}{0.2^2/40}$ or 0.625
- The standard deviation r is the square root of the Var so $Var = r^2$ ans should approach

or 0.7905694.

Let's find out if this is true by using simulation of **random exponentials** with n =40 drawn 1,000 times.

$$\frac{1}{0.2^2/40}$$

Set Constants

- set the options to display four digits after the decimal.
- The rate parameter, j , as prescribed by the assignment, is 0.2
- The number of random exponentials to means test, n is 40.
- The seed is 127
- The number of simulations of 40 random exponents is 1000
- Theoretical constants:
  - Theoretical means - Standard deviation
  - Variance of sample mean

```
options(digits = 4)

set.seed(127)

l = 0.2 #lambda, as prescribed in assignment
ex = 40 #number of exponentials examined
sims = 1000 #number of simulations


tmean<-1/l #theoretical mean
tsd<-1/l/sqrt(ex) #theoritical standard deviation
tvar<-(1/l/sqrt(ex))^2 #theoretical variance
```

Draw Samples

We'll draw 40 samples of exponentials, 1,000 times, producing a matrix with 1000 rows and 40 columns

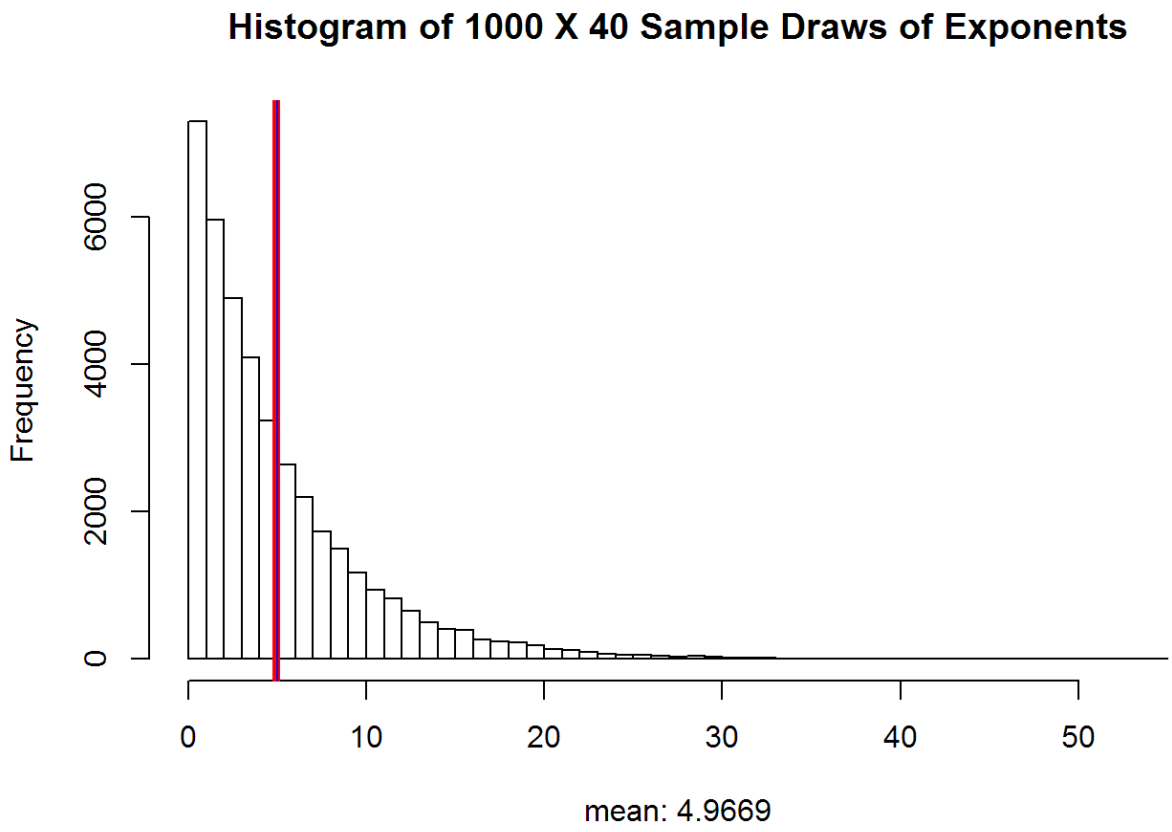
```
smatrix <- matrix(rexp(ex * sims, .2), nrow = sims, ncol=ex)
```

Compare Means:

The theoretical mean for a sample this size is calculated as $1/\lamba$ is 5.

A Histogram of the sample set values presents a right-tailed plot with a mean of 4.9669:

```
hist(smatrix, breaks = 40, main="Histogram of 1000 X 40 Sample Draws of Exponen
ts", xlab=paste("mean:", round(mean(smatrix), digits = 4), sep=" "))
abline(v=mean(smatrix), col="red", lwd=4)
abline(v=tmean, col="blue", lwd=1)
```



Now let's take a look at the mean of the each of the 1000 X 40 sample draws:

```
 # find the mean of each of the 1000 rows of 40 values and put into a 1000 x 1
data frame.
smatrixmeans<-data.frame(value=rowMeans(smatrix))
tbl_df(smatrixmeans)
```

```
## Source: local data frame [1,000 x 1]
##
##    value
##    (dbl)
## 1  4.961
## 2  4.880
## 3  5.153
## 4  3.658
## 5  4.313
## 6  3.807
## 7  5.486
## 8  4.550
## 9  6.467
## 10 5.098
## ..   ...
```

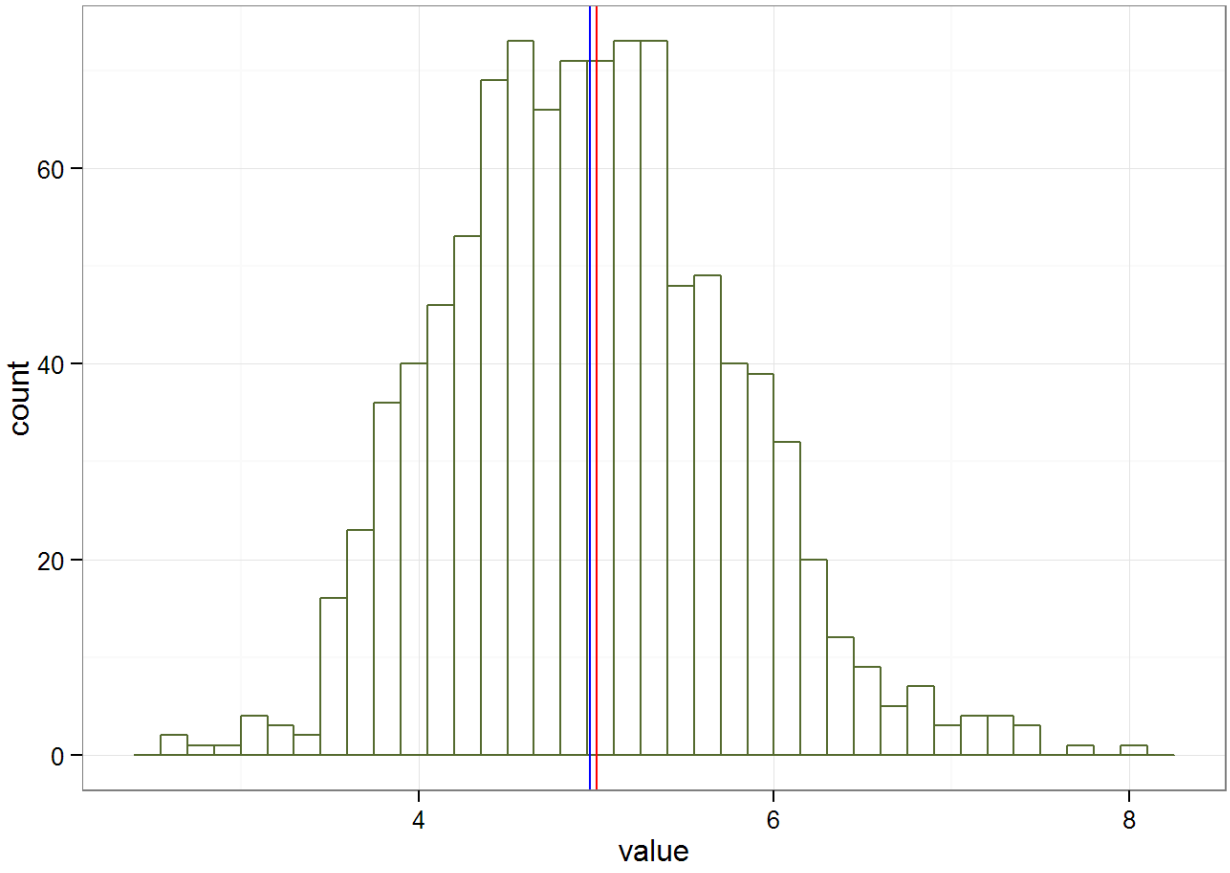
```
# calculate the overall mean of the new data frame
smean<-mean(smatrixmeans$value)


# calculate the standard deviation of the new data frame values
ssd<-sd(smatrixmeans$value)


# calculate the variance of the new data frame values
svar<-var(smatrixmeans$value)
```

Once again, our theoretical mean is **5** (red) and our mean of sample means is **4.9669** (blue):

```
fig1<-ggplot(smatrixmeans, aes(value))+
        geom_histogram( col="darkolivegreen", fill=NA, binwidth=0.15)+
        geom_vline(xintercept=smean, stat="vline", col="blue")+
        geom_vline(xintercept=tmean, stat="vline", col="red") +
        theme_bw()
fig1
```



The theoretical mean, μ 5 and the sample mean $\bar{X}$ 4.9669 are very close.

Compare Variance of the Sample Mean:

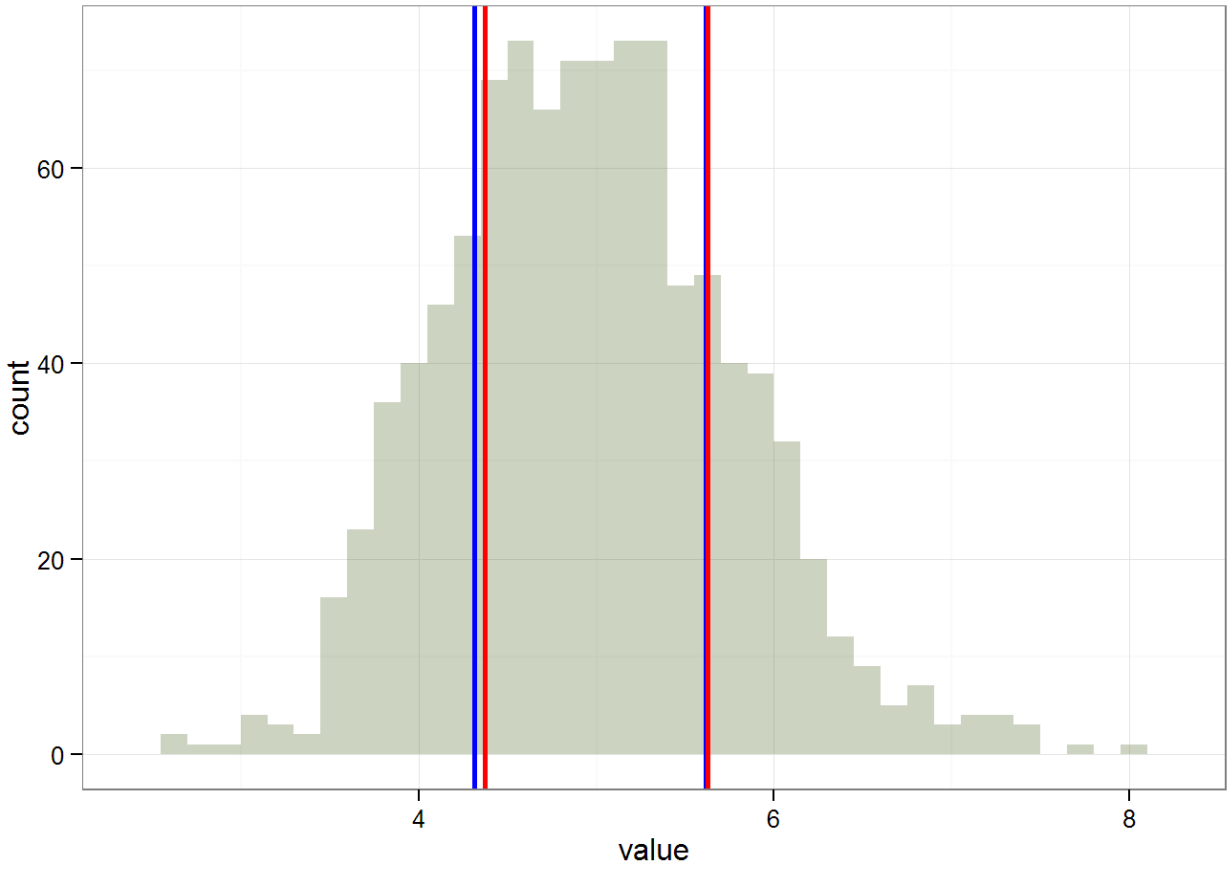
The theoretical variance of the sample mean is equal to $1/(j^2/n)$, where n =sample number.

Plugging in the values, j = .2, n =40:

The **Theoretical** Var , r^2 , of sample means is **0.625** (in red, below)

The **Calculated** Var , $Var_n(\bar{X})$, of the sample mean is **0.6509** (in blue, below)

```
fig2<-ggplot(smatrixmeans, aes(value))+
        geom_histogram(col=NA, fill="darkolivegreen",alpha=.3, binwidth=0.15)+
        geom_vline(xintercept=c(smean-svar, smean+svar), col="blue", lwd=1)+
        geom_vline(xintercept=c(tmean-tvar, tmean+tvar), col="red", lwd=1)+
        theme_bw()
fig2
```

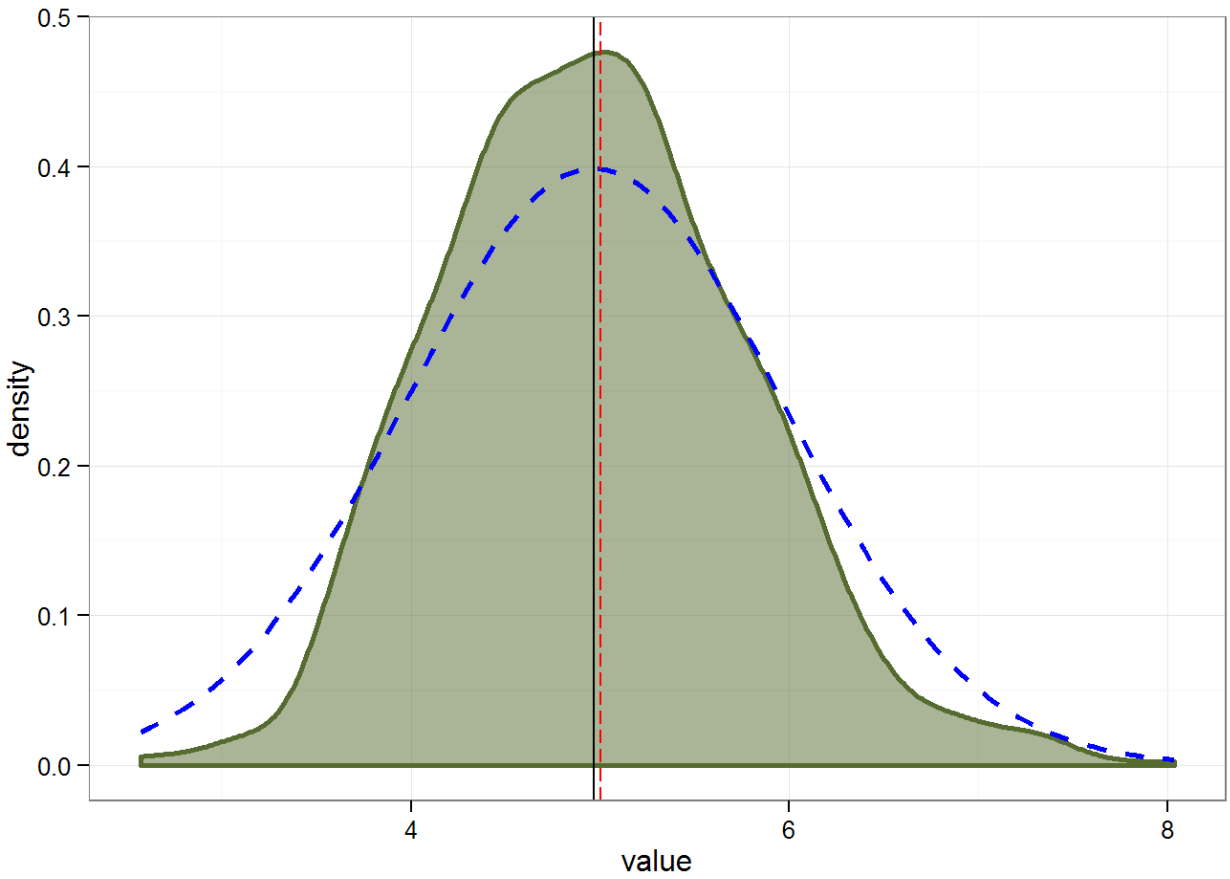


Compare Density Distribution:

As the number of means of sample means (n times the number of simulations) increases, the density plot should more closely resemble the Normal density plot.

To compare densities, we can overlay a density plot for the distribution of the calculated sample means with a theoretical normal distribution density (the classic bell curve) to test this.

```
fig3<-ggplot(smatrixmeans, aes(value))+
        geom_density(alpha=.5, col="darkolivegreen", fill="darkolivegreen" , lw
d=1)+
        stat_function(fun=dnorm, args=list(mean(smatrixmeans$value, sd=sd(smatr
ixmeans$value))), col="blue",lwd=1, lty=2)+
        geom_vline(xintercept=mean(smatrixmeans$value), stat="vline")+
        geom_vline(xintercept=1/l, stat_vline=1/l, col="red", linetype = "longd
ash")+
        theme_bw()
fig3
```



Make a Q-Q plot

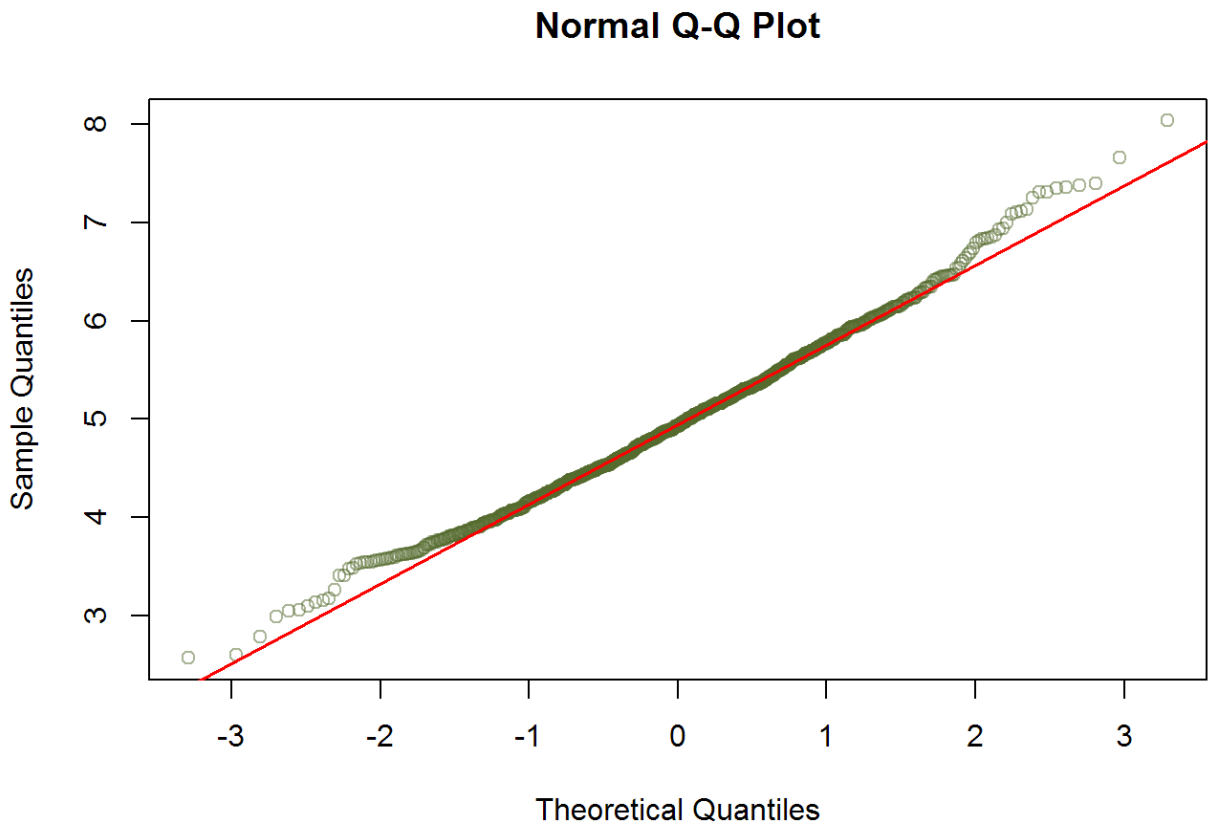
Eric Cai (http://www.r-bloggers.com/author/eric-cai-the-chemical-statistician/) from r-bloggers Exploratory Data Analysis (http://www.r-bloggers.com/exploratory-data-analysis-quantile-quantile-plots-for-new-yorks-ozone-pollution-data/) gives an excellent explanation of the usefulness of Quantile-Quantile (Q-Q) plots for testing how close distributions fit:

What is a Quantile-Quantile Plot?

A quantile-quantile plot, or Q-Q plot, is a plot of the sorted quantiles of one data set against the sorted quantiles of another data set. It is used to visually inspect the similarity between the underlying distributions of 2 data sets. Each point (x, y) is a plot of a quantile of one distribution along the vertical axis (y-axis) against the corresponding quantile of the other distribution along the horizontal axis (x-axis). If the 2 distributions are similar, then the points would lie close to the identity line, y = x.

A Quantile-Quantile plot, displaying both the theoretical normal mean for our sample range as well as the actual sample means shows a very tight fit, yet more evidence that the Central Limit Theorem is useful for working with very large data sets..

```
par(mfrow=c(1,1))
qqnorm(smatrixmeans$value, col=rgb(.333, 0.42, .18, 0.5))
qqline(smatrixmeans$value, col="red", lwd=1.5)
```



We can see that the behavior of large samples approaches that of the theoretical normal for Mean, Variance and Density, the closer we get to the theoretical Mean, Variance and Density.

Theoretical	**Sample**
k = 5	$\bar{X}_n$ = 4.9669
r^2 = 0.625	$Var_n(\bar{X})$ = 0.6509