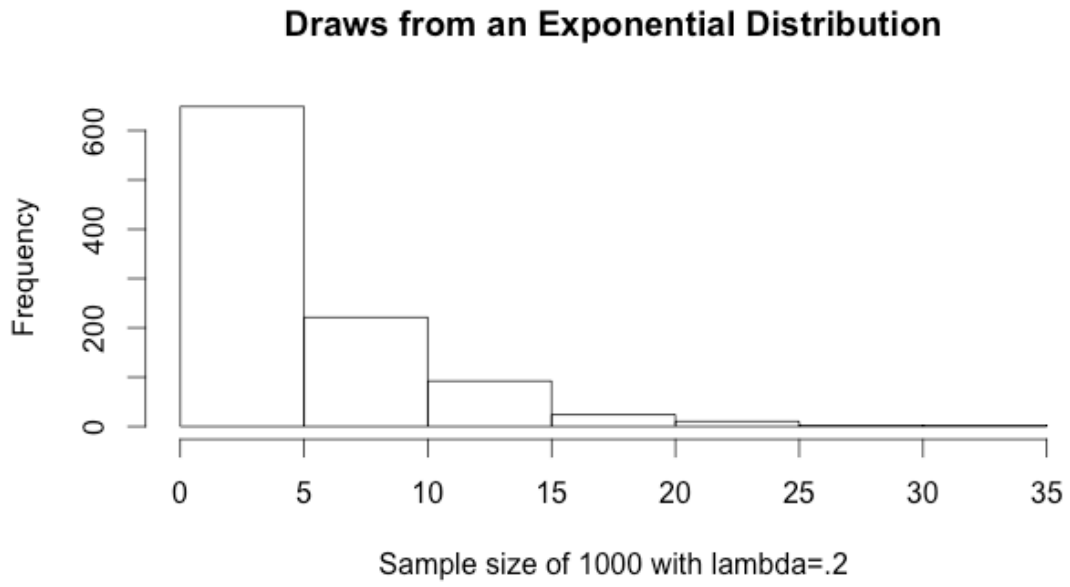
The goal of this document is to explore the Central Limit Theorem using the exponential distribution.

Per below, the exponential distribution has a very different shape than the well-known bell-shaped normal distribution:



The population mean of an exponential distribution with a lambda of .2 is: **5**. This is calculated using the formula: 1/lambda.

The sample mean of the sample shown above is: **4.8**. Here's the code used for this exercise

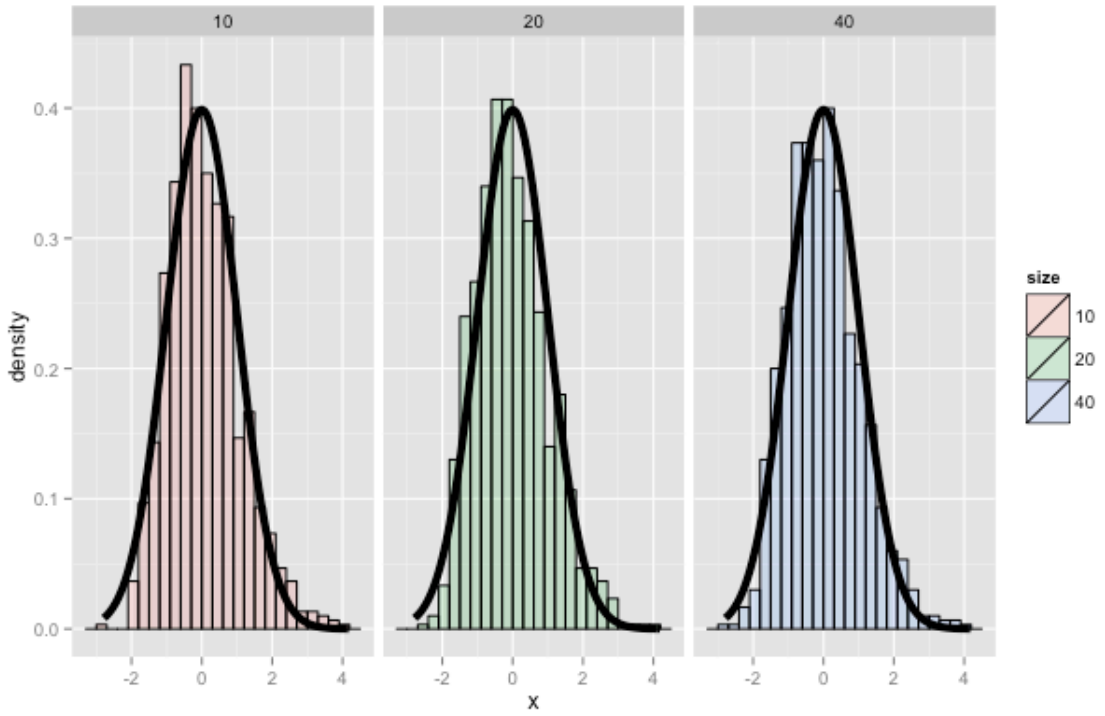
```
x<-rexp(1000,.2)

hist(x, xlab="Sample size of 1000 with lambda=.2", main="Draws from an Exponential Distribution")

mean(x)
```

I reran this exercise a handful of times and even with a large sample size of 1000, the sample mean bounced around a lot. Here are the values that I calculated for sample mean during my first five trials using this method**: 4.8, 5.0, 4.9, 5.1, 4.9**.

A better way to approximate the mean of this distribution is to use the Central Limit Theorem (CLT). Instead of drawing one large sample from the population, we will use the the distribution of many small samples. The CLT tells us that if we take many smaller samples, their means will be normally distributed with a mean equal to: *population mean*, and a standard deviation equal to: *population standard deviation divided by the square root of the sample size* .

Note that the population mean for the exponential distribution is 1/lambda or 5 in this case. And the population standard deviation is also 1/lambda, thus also 5 in this case. Let's take a sample of size n, take its mean, subtract off 5 and divide by 5/sqrt(n) and repeat this over and over. If the CLT is right, this should look exactly like a standard bell curve.

The histograms below display the distribution of the different normalized averages from the simulation.



As you can see the approximation is quite good. The shape of each histogram is close to a standard normal and it becomes a closer approximation as n increases from 10 to 20 and finally to 40. With a sample size equal to 40, we have a very nice Gaussian distribution.

Here's how this simulation has addressed each of the questions posed by the assignment:

1. **Show where the distribution is centered at and compare it to the theoretical center of the distribution.**
    - The histograms of sample means shown above were normalized using the population mean.
    - Because the resulting distributions turn out to be standard normals--with a mean of zero and a standard deviation of 1--we can confirm our hypothesis that the mean of the sample means is centered around the population mean (which is 5).

2. Show how variable it is and compare it to the theoretical variance of the distribution.

- Because the resulting distributions turn out to approximate standard normals--with a mean of zero and standard deviation of 1--we can confirm our hypothesis that the standard deviation of each sample is(1/lambda)/ (sqrt(n)), e.g. 5/sqrt(n).

3. Show that the distribution is approximately normal.

- The density curve for the standard normal distribution is overlayed onto each histogram. You can see that the histograms approximate this density curve, And this approximation improves as the sample size increases from 10 to 20 and finally to 40. At a sample size of 40, the histogram adheres closely to the standard normal density curve.

FYI, here's the code I used to perform this exercise:

```
library(ggplot2)
nosim <- 1000
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5
dat <- data.frame(
      x = c(apply(matrix(rexp(nosim*10,.2),
                  nosim), 1, cfunc, 10),
          apply(matrix(rexp(nosim*20,.2),
                  nosim), 1, cfunc, 20),
          apply(matrix(rexp(nosim*40,.2),
                  nosim), 1, cfunc, 40)
      ),
      size = factor(rep(c(10, 20, 40), rep(nosim, 3))))
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(alpha = .20, binwidth=.3,
colour = "black", aes(y = ..density..))
g <- g + stat_function(fun = dnorm, size = 2)
g + facet_grid(. ~ size)
```