Comparison of the Central Limit Theorem vs. the exponential distribution.

Author: Craig Anthony

Overview: This R Knitr pdf file use simulations, text, and graphical output to illustrate a comparision of the Central Limit Theorem vs. an exponential distribution. This document compares:

1. The sample mean to the theoretical mean of the distribution.
2. The sample variance to the theoretical variance of the distribution.
3. That the distribution is normal.

Background Information:

The Central Limit Theorem (CLT)

The CLT states that If random samples, each of size n, are taken from (nearly) any population with a mean MU and a standard deviation SIGMA , the sampling distribution of the sample means (averages) will:

1. Have True Mean (Mean of Sample mean values) = Population Mean
2. Have Standard Deviation (of the distribution of Sample Mean) equal to the standard error.
3. The sample mean will be approximately normally distributed when the sample size n is large regardless of the shape of the population distribution.

This means is that the CLT lets you apply statistical techniques that assume a normal population distribution to non-normal population distributions.

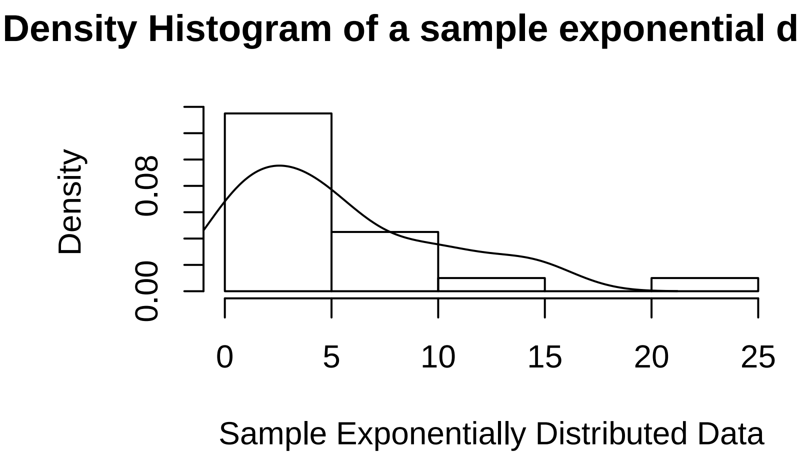
Exponential Distribution

The exponential distribution is an asymetric, non-normal that describes the mean time of occurance between continouously occuring, independent events that happen at a constant rate.The exponential distribution has the following properties.

1. Non-normal distribution and asymetric distribution.
2. Expected value and standard deviation equals 1/lambda where lambda equals the rate parameter.
3. Variance equals 1/(lambda*lambda)

The exponential distribution is used to predict time related events such as time between hardware failures and wait times for customer service. Exponentially distributed data typically looks like this.

```
lambda <- .2
n = 40
sim_means = NULL
for (i in 1 : 1000) sim_means = c(sim_means, mean(rexp(n,lambda)))
sample_mean <- mean(sim_means)
par(pin=c(3,1))
hist(rexp(n,lambda),prob=TRUE,main="Density Histogram of a sample exponential distribution",xlab="Sa
lines(density(rexp(n,lambda)))
```



1. Sample Mean vs. Theoretical Mean

Theoretical Mean = 1/lambda = 1/.2 = 5 Calculate the sample mean where lamba = .2, n = 40, number of simulations = 1000

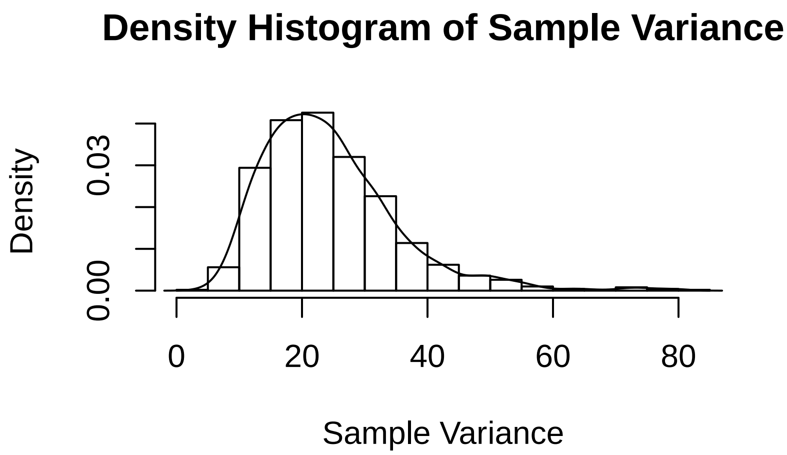
```r
lambda <- .2
n = 40
sim_means = NULL
for (i in 1 : 1000) sim_means = c(sim_means, mean(rexp(n,lambda)))
sample_mean <- mean(sim_means)
```

The theoretical mean equals **5**. The sample mean equals **4.9684303**. The theoretical mean and the sample mean are very nearly equal and have a more normalized distribution at a sample size of 40. Smaller sample sizes would have more skewed results.

2. Theoretical Variance vs. Sample Variance.

Theoretical variance = 1/lambda*lambda = 1/.04 = 25, lambda <- .2 Calculate the sample variance

```r
sample_size = 40
sim_vars = NULL
for (i in 1 : 1000) sim_vars = c(sim_vars, var(rexp(sample_size,lambda)))
sample_variance <- mean(sim_vars)
par(pin=c(3,1))
hist(sim_vars,prob=TRUE,main="Density Histogram of Sample Variance",xlab="Sample Variance")
lines(density(sim_vars))
```



The theoretical variance equals 25 The sample variance equals 24.3838517 The variance has a longer upper limit tail which makes sense because none of the generated interval data can be less than 0.

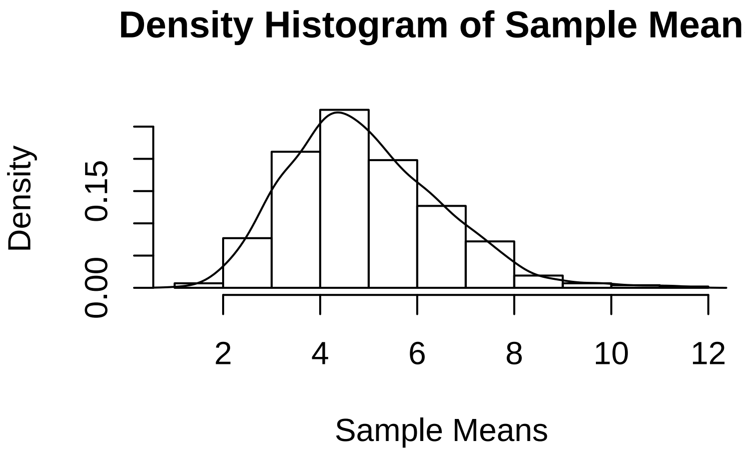
3. Illustration of normal distribution

As described in the Background Section, the CLT states that he sample mean will be approximately normally distributed when the sample size n is large regardless of the shape of the population distribution. Normally distributed data is graphically represented by a forming a bell shaped curve. Another way to determine if data is normally distributed is if:

1) mean = median = mode.
2) There are the same number of samples below the mean as above it.

Because the numbers carry out to six decimal places the odds of finding a mode number is quite small and I will not calculate it. The following code will generate 1000 sample means of an exponential distribution of sample size = 40 and a rate parameter lambda = .2. The code will generate a histogram with curve of results as well as the mean, median of the results. The code will also generate a count of the number of samples below and above the mean.

```
n = 10
less_mean = 0
greater_mean = 0
sim_means = NULL
for (i in 1 : 1000)sim_means = c(sim_means, mean(rexp(n,lambda)))
sample_mean <- mean(sim_means)
sample_median <- median(sim_means)
for (i in 1: 1000){
   if(sim_means[i] > sample_mean) greater_mean = greater_mean + 1
   if(sim_means[i] > sample_mean) lesser_mean = greater_mean + 1
}
par(pin=c(3,1))
hist(sim_means,prob=TRUE,main="Density Histogram of Sample Means",xlab="Sample Means")
lines(density(sim_means))
```



The sample mean equals 4.925837 The sample median equals 4.7461415 There are 452 means below the sample mean. There are 451 means above the sample mean.