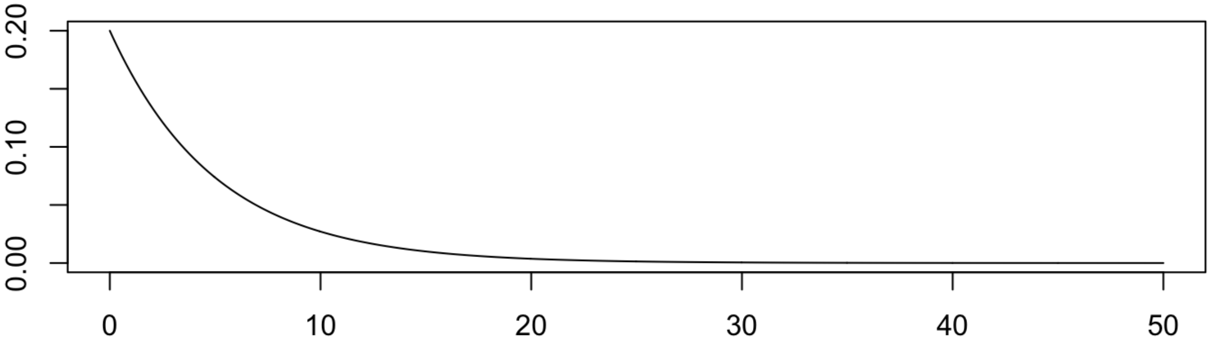
Central Limit Theorem and Inferential Data Analysis

Andrew Witherspoon

Part 1: Central Limit Theorem Simulation

The Central Limit Theorem is an important theorem in statistics, which states, that a distribution of means of independent independently distributed random variables, will take on the shape of a normal distribution if our sample size big enough. This is extremely useful, as so many of our statistical tools rely on the assumption of normality.

Let’s start by looking at an exponential distribution, which is decidedly not a normal shape. Here is a look at the shape of exponential distribution, with a rate (lambda) of 0.2:



If The Central Limit Theorem is correct, then if we take a random sample from the above distribution, take the mean of that random sample, subtract off the population mean (which is 1/lambda), divide by the standard error of the estimate, and do this n times, the distribution of *these* sample means will be approximately normal for large values of n.

To put the above paragraph into a formula notation,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\text{Estimate} - \text{Mean of the Estimate}}{\text{Standard Error of the Estimate}}$$

Let’s run a simulation to see if this is correct. We will take a 1000 samples of 40 random numbers from the above exponential distribution:

```
lambda = .2 #this is the rate from the above distibution
n = 40 #sample size
nSamps = 1000 #the number of samples

samps <- matrix(, nrow = nSamps, ncol = n)
for(i in 1:nSamps){
       samps[i,] <- rexp(n, rate = lambda)
}
```

Now we can take the mean of each of our 1000 samples to get the sample mean:

```
sampleMeans <- apply(samps, 1, mean)
```

Now let’s take the CLT formula piece by piece. Subtract the population mean (1/lambda) from the sample mean:

$$\bar{X}_n - \mu$$

```
CLTtop <- (sampleMeans - 1/lambda)
```

Next divide the sample standard deviation (also 1/lambda) by the square root of the sample size, n (this is the standard error of the mean):

$$\sigma/\sqrt{n}$$

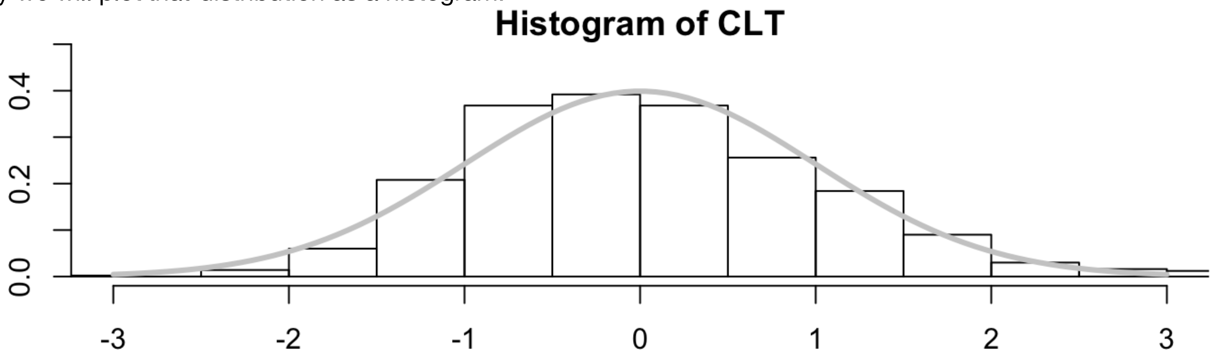
```
CLTbottom <- (1/lambda)/sqrt(n)
```

Now divide former by the latter:

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

```
CLT <- CLTtop / CLTbottom
```

And finally we will plot *that* distribution as a histogram:



We've overlayed a normal distribution density plot to show that our histogram is a very close fit to a standard normal. The Central Limit Theorem works!

To further illustrate, the mean of a standard normal is 0, and the variance is 1. Let's compare that to the mean and variance of our CLT calculation from our simulation samples:

```
mean(CLT)
```

```
## [1] 0.01898086
```

```
var(CLT)
```

```
## [1] 0.9636892
```

Looks like we're pretty close! With even larger values of n, these would be get even closer.

Part 2: Inferential Data Analysis

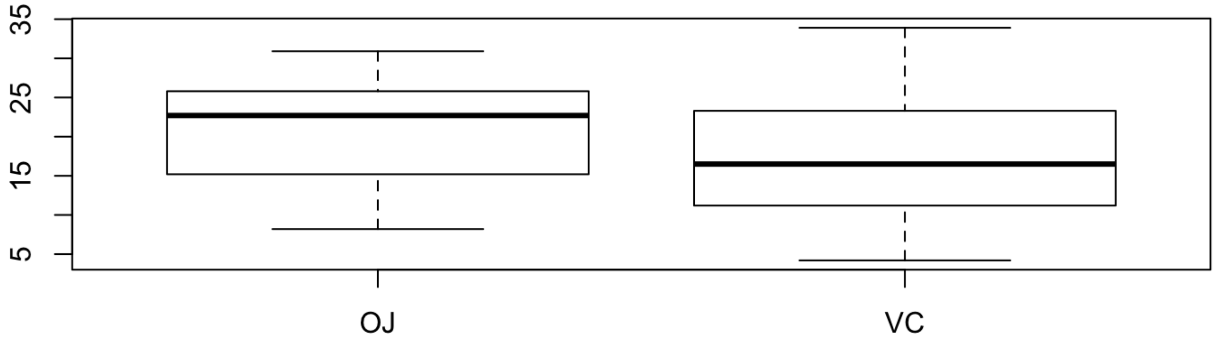
For this exercise, we will use the **ToothGrowth** data, which is in the R datasets package.The R documentation gives a description of the data:

"The response is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC)."

	0.5	**1**	**2**
OJ	10	10	10
VC	10	10	10

As the table above shows, there are 10 observations for each combination of dose, and supplement type.

Let's start by comparing the mean tooth length of guinea pigs treated with orange juice to the tooth length of those treated with ascorbic acid:



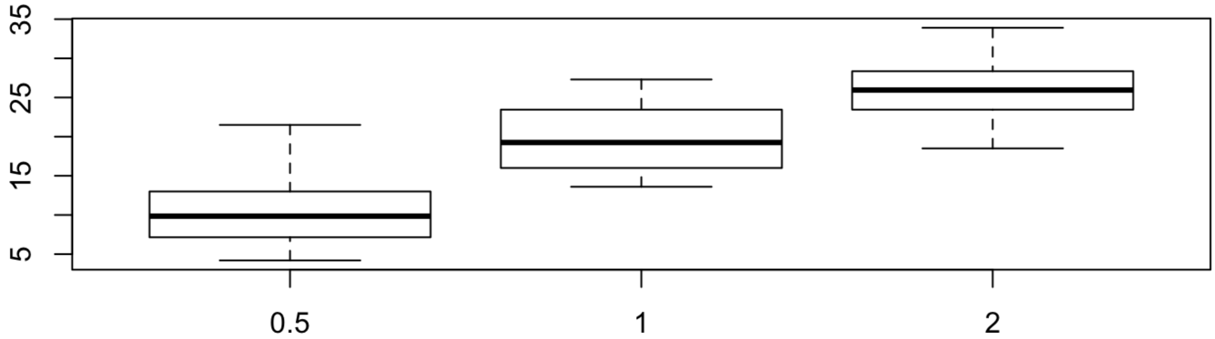
Our null hypothesis is that the mean tooth length of the OJ group not different than the mean tooth length of the VC group. We'll run a t-test to see if we can reject this hypothesis:

```
OJ <-  ToothGrowth$len[ToothGrowth$supp=="OJ"]
VC <-  ToothGrowth$len[ToothGrowth$supp=="VC"]
t.test(OJ, VC, alternative = "two.sided", conf.level = .95)$conf.int
```

```
## [1] -0.1710156  7.5710156
## attr(,"conf.level")
## [1] 0.95
```

The 95% confidence interval contains the value 0, therefore we cannot reject the null hypothesis. The data does not show that tooth length is affected by supplement type.

Let's do the same type of hypothesis testing based on dose:



A dose of 0.5 and a dose of 2 have the largest mean difference, so let's use these values for our hypothesis testing. Once again, our null hypothesis will be that the mean of dose0.5 and the mean of dose2 are not different.

```
t.test(dose0.5, dose2, alternative = "two.sided", conf.level = .95)$conf.int
```

```
## [1] -18.15617 -12.83383
## attr(,"conf.level")
## [1] 0.95
```

We can reject the null hypothesis, as the confidence interval does not contain 0. With 95% confidence, the data shows that the dose does have an effect on tooth length.

Link for full markdown document code (https://github.com/elAndrew/Course6FinalProject/blob/master/Project.Rmd)