Lab 8

March 6-7, 2018

In today's lab we will discuss the Normal Distribution and the Central Limit Theorem using R.

The Normal Distribution

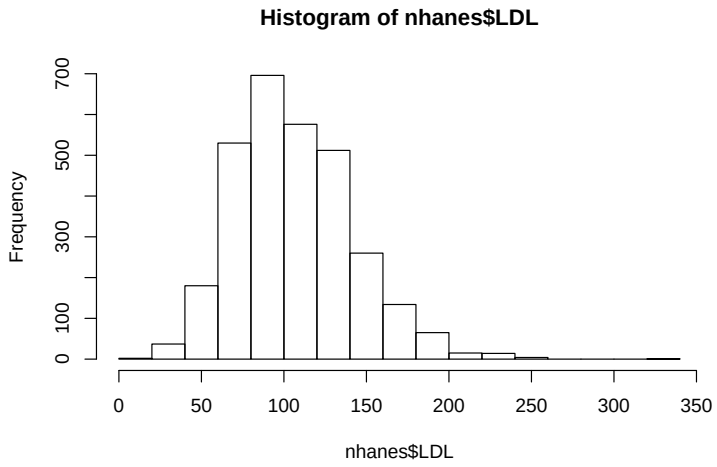
We will be using the lipids.txt dataset from the course website. Here we are going to look at various lipid levels of the 3026 adults in the study.

As reference, a reading of Triglycerides (TRG) is a measurement of total cholesterol and includes the concentration of LDL and HDL in the blood. A reading of low-density lipoprotein (LDL) is a more specific measurement of total cholesterol and is typically referred to as "bad" cholesterol.

```
nhanes <- read.delim("http://myweb.uiowa.edu/pbreheny/data/lipids.txt")
```

Let's begin by looking at a histogram of LDL measurements.

```
hist(nhanes$LDL)
```



Although LDL values are slightly right-skewed, let's assume that it is close enough to the normal to answer some health questions using the normal distribution.

First, to use the normal distribution we must find the mean and standard deviation of the data

```
xbar<- mean(nhanes$LDL)
std.dev<- sd(nhanes$LDL)
```

Suppose we are interested in comparing the NHANEs LDL data to the following guidelines:

LDL cholesterol levels should be less than 100 mg/dL. Levels of 100 to 129 mg/dL are acceptable for people with no health issues but may be of more concern for those with heart disease or heart disease risk factors. A reading of 130 to 159 mg/dL is borderline high and 160 to 189 mg/dL is high. A reading of 190 mg/dL or higher is considered very high.

What is the probability of observing an LDL measurement that is 160 or greater(LDL readings that are considered high and very high)?

To visualize this we will draw a bell-shaped curve on the board

As the default, the pnorm function will give us the probability of the lower tail. We must change the direction of the tail if we want to find a probability that is greater than a number of interest.

```
pnorm(160, mean = xbar, sd=std.dev, lower.tail = FALSE)
```

```
## [1] 0.06872828
```

```
#OR simply find the z-score first and plug it into the pnorm function

z<- (160-xbar)/std.dev

pnorm(z, lower.tail = FALSE)
```

```
## [1] 0.06872828
```

Note: The normal distribution is used for continuous data, unlike the binomial distribution that is used for discrete data. As discussed in class last week, you must not include the number of interest in the 'pbinom' function when finding the complement. This is unlike the 'pnorm' function where you should include the number of interest even when finding the complement. Essentially, this is because the distribution is continuous. If you were to find the probability of having a LDL 160 or greater using: 1-pnorm(159, mean=xbar, sd=std.dev) you would see that the probability is different from the one above and is incorrect because you are including values such as 159.9, 159.8, etc. Therefore **include** 160 as shown below and you will get the same answer as above:

```
1-pnorm(160, mean = xbar, sd=std.dev, lower.tail = TRUE)
```

```
## [1] 0.06872828
```

How would we interpret our findings?

Now, let's compare this to the probability of 160 or greater using the actual data

Do you expect the probability of observing a LDL 160 or greater using the actual data to be exactly the same, less than, or greater than the probability found using the normal distribution? Why?

```
sum(nhanes$LDL>=160)/length(nhanes$LDL)
```

```
## [1] 0.07865169
```

As illustrated, the normal distribution allows us to make inferences about a population, assuming that the true population distribution follows a bell-shaped perfectly symmetric curve. This may not always be true, so it is important to consider how well the data fits the assumed distribution and what the expected distribution of the population is (i.e. what have other studies found?)

What is the probability of observing a LDL that would be classified as borderline high (160-189)?

Again we will draw a bell-shaped curve on the board

```
pnorm(189, mean = xbar, sd=std.dev, lower.tail = TRUE)-
pnorm(160, mean = xbar, sd=std.dev, lower.tail = TRUE)
```

```
## [1] 0.05785488
```

```
#OR

pnorm(160, mean = xbar, sd=std.dev, lower.tail = FALSE)-
pnorm(189, mean = xbar, sd=std.dev, lower.tail = FALSE)
```

```
## [1] 0.05785488
```

```
#OR

1-(pnorm(160, mean = xbar, sd=std.dev, lower.tail = TRUE)+
pnorm(189, mean = xbar, sd=std.dev, lower.tail = FALSE))
```

```
## [1] 0.05785488
```

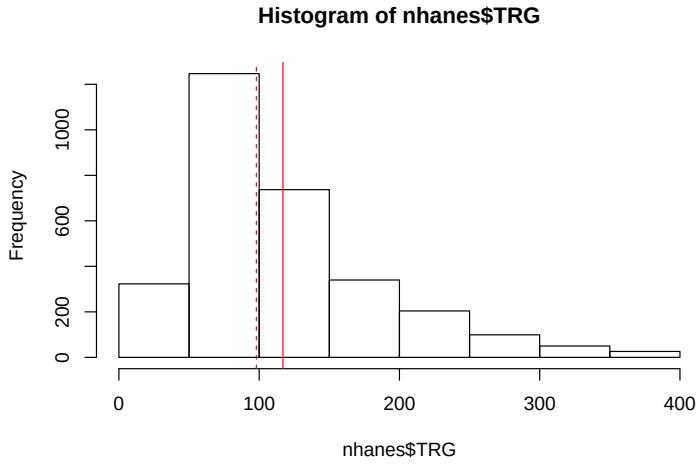
As you can see you can find the answer many ways as long as you understand the direction in which you are finding the probability.

Central Limit Theorem

We are going to perform a hands-on exercise to understand the central limit theorem. We will be looking at the triglycerides (TRG) readings. Let us assume that this data represents the **population**. First, let's look at the distribution.

```
x <- seq(0,max(nhanes$TRG), len=101)

hist(nhanes$TRG)
abline(v=mean(nhanes$TRG), lty=1, col="red")
abline(v=median(nhanes$TRG), lty=2, col="red")
```



From looking at the histogram we can see that the data is skewed right and does not follow a normal distribution very closely.

Now let's say that we are going to randomly select from the population to conduct a study. Each of you will conduct your own study. The first time you are able to recruit 50 people, and the second time you are able to recruit 100 people.

Standard Error

Recall in class that the standard error is the standard deviation of the population divided by the squareroot of the total number of observations in a study (n). In this case since we know the population, we can find the population standard deviation and therefore the standard error for each of the experiments. *Find the standard error for each study with 50 and 100 subjects.*

Distribution of Means

To show you all how the central limit theorem works we are going to have everyone in the class run the sample with 50 and 100 subjects. Each person will report their mean in a CSV file. The TA will save the document as a CSV file and upload the file using read.csv. We will look at the average of all of the means to see how they compare to that of the actual data.

R lets you randomly select samples from a dataset using the sample() function. Each of you will perform your own studies using the sample function to draw random TRG measurements for 50 and 100 subjects. Go ahead and save these draws and find the mean of your studies as follows:

```
sample50 <- sample(nhanes$TRG, 50)
sample100 <- sample(nhanes$TRG, 100)

mean(sample50)
mean(sample100)
```

```
TRGmeans<- read.csv("TRGmeans.csv")
```

```
x11(width = 12, height = 8)
par(mfrow=c(1,2))

hist(TRGmeans$fifty, main = "Means of 50 subject study")
hist(TRGmeans$hundred, main = "Means of 100 subject study")
```

What do you observe about the distribution of the means compared to the population distribution?

Now, let's take a look at the standard deviation of the means. What is this value comparable to: the population standard deviation or the standard error?

```
sd(TRGmeans$fifty)
sd(TRGmeans$hundred)
```