Math 338 **Lab Assignment #16** Fall 2019

In this lab, we will be working with a subset of a much larger data set commonly known as Fisher's or Anderson's Iris data set, see Fisher, R.A. (1936), *The use of multiple measurements in taxonomic problems*. The data set describes the petal length, petal width, sepal length, and sepal width of 50 *setosa*, 50 *versicolor*, and 50 *virginica* iris flowers. It is already pre-loaded into R; you can view the dataset using the command:

```
> View(iris)
```

We will only be working with the petal length for the *setosa* species. Let's extract that data from the original set by using some nifty functions in the **dplyr** package. If you don't have the package already installed, install it, then run:

```
> library(dplyr)
> setosa.petal.length <- iris %>% filter(Species == "setosa") %>%
select(Petal.Length)
```

Here we are doing two things with the dataset. The **%>%** operator is called a "pipe." It means, "take the thing on the left side, then perform the command on the right side on it." It is used in the **dplyr** package and a few other packages designed to make data science easier with R.

So in our command, we first take the data set iris, then we **filter** the data set to get only the cases with the value "setosa" for the variable Species, then we **select** the variables of interest (in this case we are only interested in *Petal.Length*). The resulting data frame is stored in the variable **setosa.petal.length**.

Question #1 You have used the **ggplot2** package to construct histograms. Use it to plot the histogram showing petal length of the *setosa* species and insert the histogram below. (Look at Lab 14 and Lab 15 for example code)

Please see attached document for the graph.

If we believe the Central Limit Theorem is going to be accurate (and assuming the data were collected in a random fashion), we can perform a hypothesis test to test the claim that the population mean petal length is 1.3 centimeters. Assume the population standard deviation of petal lengths is 0.5 cm.

Question #2 Describe the distribution of the sample. Does the shape of the distribution suggest that the Central Limit Theorem will be roughly accurate and so we can model the sampling distribution of all possible sample means as a normal distribution?

The graph is not skewed right or left so it is normally distributed.

Question #3 Write the null hypothesis for this test using an appropriate symbol for the parameter.

Our population sample mean will be mu which is 1.3 and the total number of elements in the sample is going to be 50.

Question #4 What is the distribution of sample means under the null hypothesis? Remember that we need to specify the family/shape of the distribution and all relevant parameters of the distribution.

Our sigma will be 0.5/\sqrt{50} which is directly proportional to the amount of elements in our sample.

\bar{X} \sim N(1.3, \frac{0.5}{\sqrt{50}})

First let's work in the Neyman-Pearson framework. Recall that in this framework we need to specify an alternative value of our parameter. Let's suppose that a meaningful difference is an increase of 0.2 cm; that is, under the alternative hypothesis, our parameter value is 1.5 cm.

The critical region is found in a similar way as in Lab 8, except we are now working with the `pnorm`/`qnorm` commands (because our sampling distribution is normal, not binomial).

```
> n <- # the sample size from question 3
> mu0 <- # the value of mu if the null hypothesis is correct
> mu1 <- # the value of mu if the alternative hypothesis is correct
> sigma <- # the assumed value of sigma
> alpha <- # the conventional maximum value of alpha
> critical.value <- qnorm(alpha, mean = mu0, sd = sigma/sqrt(n), lower.tail = FALSE) #  because mu1 > mu0; otherwise we'd use lower.tail = TRUE
```

Note that because we are working with continuous random variables here, we don't worry about the distinction between ≥ and > when computing the critical value, so we don't need to check our α value with `pnorm` like we did with `pbinom`.

Question #5 Assuming α = 0.05, find the critical region for this test.

The critical value is going to be 1.416309

To get the observed sample mean, we simply take the mean of the specific variable we're interested in. There are a couple of different ways to do this:

```
> mean(setosa.petal.length$Petal.Length)
```

OR

```
> setosa.petal.length %>% summarize(mean = mean(Petal.Length)) # the dplyr way; this way works much better when we're computing means of several different groups
```

Question #6 Is our observed sample mean (from the 50 flowers) in the critical region? What should we conclude about the population mean petal length (in the N-P framework)?

Our observed sample mean from the 50 flowers is 1.462. Since this value does fall in the critical region and by N-P framework, we can conclude that the alternative hypothesis is true and we can reject the null hypothesis.

Question #7 What is the sampling distribution of our sample means if in fact the alternative hypothesis H_1 is correct?

\bar{X} \sim N(1.5, \frac{0.5}{\sqrt{50}})

Recall that to find power, we use our critical value and the sampling distribution if H_1 is correct:

```
> power <- pnorm(critical.value, mean = mu1, sd = sigma/sqrt(n), lower.tail
= FALSE) #  because mu1 > mu0; otherwise we’d use lower.tail = TRUE
```

Question #8 What is the power of our test to detect the specific alternative value of 1.5 cm?

Our power is going to be 0.881709 which is quite good.

Now let’s work in the Null Hypothesis Significance Testing framework. Recall that in this framework our alternative hypothesis is an inequality. Let’s assume that if the population mean is not 1.3 cm, then it must be larger than 1.3 cm.

Question #9 Write the alternative hypothesis H_a for this test using an appropriate symbol for the parameter.

Our population sample mean will be mu which is 1.5 and the total number of elements in the sample is going to be 50.

Recall that the p-value is the probability of observing a sample with our observed value of the test statistic, or a value more favorable to the alternative hypothesis. Use the `pnorm` command with the appropriate sample mean and `mean`, `sd`, and `lower.tail` arguments to find the p-value.

Question #10 What is the p-value for our test? Using a 5% significance level, what should we conclude about the population mean petal length (in the NHST framework)?

Our p-value is going to be 0.01098096 which means we are going to be able to reject the null hypothesis and accept the alternative hypothesis.