**Lab 9 – Hypothesis testing**

**To submit: answers to all numbered questions. When the question asks you to write code or create graphs, submit the code and/or graphs in the Word document as part of your answer. Also submit a single .R file that contains all of your code.**

In Lab 8, we saw how to use the **t.test** command to construct confidence intervals for means. For example, here we used the **t.test** function in R on thirty randomly-selected net race times from the **TenMileRace** dataset to construct confidence intervals in a single line:

> t.test(racetimes30)

One Sample t-test

data: racetimes30

t = 35.126, df = 29, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

5194.265 5836.535

sample estimates:

mean of x

5515.4

In this lab, we are going perform some hypothesis tests by making use of the rest of the output returned by the **t.test** command. Once again we will be using the **survey** datasets, which is in the **MASS** library.

Let’s examine the ages of students in the **survey** dataset:

> t.test(survey$Age)

One Sample t-test

data: survey$Age

t = 48.447, df = 236, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

19.54600 21.20303

sample estimates:

mean of x

20.37451

We see some familiar variables and labels from hypothesis tests: df, the number of degrees of freedom, p-value, t (which is ttest from hypothesis tests), and a mention of the alternative hypothesis. Notice that the alternative hypothesis here is that the true mean age is not equal to zero – which is pretty obvious for our data set! Let’s give it a more reasonable null hypothesis: say, that the mean student age is 20.

> t.test(survey$Age, mu=20)

One Sample t-test

data: survey$Age

t = 0.89053, df = 236, p-value = 0.3741

alternative hypothesis: true mean is not equal to 20

95 percent confidence interval:

19.54600 21.20303

sample estimates:

mean of x

20.37451

That’s more like it. We are telling R to test the claim that the mean student age is equal to 20. This is the null hypothesis here. Note that the confidence interval is 0.95, which indicates a default significance of α=0.05.   
  
The **p.value** output is the key to evaluating this claim. If the null hypothesis were true – that is, if the mean student age were 20 - we would expect to get a sample mean at least as far from 20 as the one we got (20.37451), 37.41% of the time. That’s pretty likely! Clearly, it’s safe to say that there’s no evidence that the mean student age differs from 20. (In general, if the p-value is greater than alpha, which in our case is 1-0.95=0.05, it’s safe to say there’s no evidence that the null hypothesis is false.)

We can also do one-sided hypothesis tests using the optional **alternative** argument.

> t.test(survey$Age, mu=19, alternative="greater")

One Sample t-test

data: survey$Age

t = 3.2683, df = 236, p-value = 0.0006215

alternative hypothesis: true mean is greater than 19

95 percent confidence interval:

19.68004 Inf

sample estimates:

mean of x

20.37451

1. At the default significance level of α=0.05, do we have evidence that the mean student age is greater than 19?

Note that the upper bound of the confidence interval is “Inf”, for “infinity”. This is a *one-sided confidence interval*, which we did not cover in class. The **alternative=“greater”** argument tells R that for both the hypothesis test and the confidence interval, we only care about whether the mean student age is greater than 19. R tells is that we are 95% sure that the mean student age is 19.68004 years or older.

We can also make claims about how one mean relates to another. Consider, for instance, the claim that male students are taller on average than female students.

1. Read the help file for **t.test** and then give a sequence of commands to test the claim at α=0.05 that male students are on average taller than female students. Write a sentence stating your conclusion. Hint: you need to call **t.test** with three arguments; you can use default values for the rest.
2. Create a function that reads in a list of data, significance level, information about the mean (ie, we are measuring student ages), units, a value of the mean for the null hypothesis, and an inequality (“greater than”, “less than”, “not equal to”). Your function should test the claim given by your input, and return a sentence conclusion. For example:  
     
   > hypothesis(survey$Age, 0.01, "student age", "years", "20", "greater")

At alpha=0.01, we do not have sufficient evidence that the mean student age is greater than 20 years .

Use your function to test the claim that at α=0.1, the mean student pulse is greater than 73 beats per minute. Give both your code and your output as answers to this question.

1. Use the **t.test** command to test the claim at α=0.05 that the span of students’ writing hands differs from the span of their non-writing hands, and include a sentence conclusion with your output. Again you will need to call the **t.test** command with just three arguments, but this is slightly different from the last question. When we compared male and female heights, we selected the male and female students independently of one another. But the hand spans are not independent of each other, because each student (well…the vast majority anyway) has a writing hand and a non-writing hand. Read the help file and experiment if necessary to see how to handle this kind of data. The output is below for you to compare.

data: survey$Wr.Hnd and survey$NW.Hnd

t = 2.1268, df = 235, p-value = 0.03448

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

0.006367389 0.166513967

sample estimates:

mean of the differences

0.08644068