Homework #5: One and two-sample t-tests.
Due: In lab, March 4-7

DIRECTIONS: For this homework, please turn in hard copies in class on Wed, Feb 13. This homework counts double. Write answers out in a separate Microsoft Word document (note that some questions may not require an answer – so you can skip these or just say "completed" – but we'll try to put what the actual question to be answered, or what you need to do to get credit, by placing that in **bold** in the question). We encourage you to work in groups for R-related material, but make sure to do the work yourself too and *always* come up with independent answers. Homeworks that have identical answers will receive F's.

Part 1: The One Sample T-Test:

Q1) Studying outside of class is a fundamental part of the learning process in college. One question that we'd like to understand is whether college students today (in particular, Psychology majors at CU) tend to study more or less than the typical college student did 40 years ago. According to Babcock & Marks (2011), college students in 1961 studied an average of 24 hours per week outside of class. In the first lab, you completed a survey on how many hours per week you typically studied during the semester. Use information collected in this class to test whether CU Psychology majors tend to study more or less than college students did in 1961.

First, load the datasets that I have stored on an external server. To do that, use the command below:

```
load(url("http://www.matthewckeller.com/Stats3101/Stat3101.Datasets-2013.RData"))
```

Consider this class to be a random sample of CU Psychology students. The variable we are interested in is "study.hours.weekly." Use R to conduct a one-sample t-test to answer your question. Note that we're using a t-test here because WE DO NOT KNOW THE POPULATION STANDARD DEVIATION (SIGMA), RATHER WE HAVE TO ESTIMATE THAT USING THE STANDARD DEVIATION OF THE SAMPLE. To conduct a t-test, you can use the menu system. First, make "survey2013" the active dataset as we've done before. Then use either syntax or the menu system to conduct a t-test. For using the Menu, go to Statistics -> Means -> Single-sample t-test. You'll need to place what the null hypothesis mean is in the box where it asks you, and make sure you're conducting a two-tailed test (i.e., that the "Alternative Hypothesis" has that the population mean != mu0, i.e., $\mu_{2013,CU} \neq \mu_{1961}$). Or you can use syntax like this:

```
t.test(survey2013$study.hours.weekly, alternative='two.sided', mu=24)
```

A) What are your alternative and null hypotheses? Write these in mathematical notation, as we've done in class.

B) What is your t-value and p-value? What do you conclude?

C) How "big" is this effect? Remember that there is a difference between knowing how unlikely something was to have happened given the null (which is given by the test statistic, a t-value in this case, and a corresponding p-value) and how important or big the effect was, which is given

by an effect size measure, such as the Cohen's d. Get the Cohen's d by using the formula we've discussed in class, except that instead of dividing the mean difference by the population standard, which we don't know, divide by its estimate (the standard deviation in the sample):

Cohen's d = $\frac{\bar{X} - \mu_0}{SD}$

So all you need to do to figure this out is to find the mean and the standard deviation of "study.hours.weekly." We've done this before in past homeworks, so if you can't remember how to get that information, look at your past homeworks and/or R notes you're taking. **Give the Cohen's d for this effect, and interpret this effect in word grandma could understand**.

D) It's often useful to have readers take a quick look at your data. To do this, create a boxplot:
boxplot (survey2013$study.hours.weekly)

Describe the shape of the distribution.

E) Write a four sentence summary of your findings. Make sure to include an estimate of the "effect size" (i.e., Cohen's d) of the effect you found, and include a boxplot of the scores below your 4-sentence summary.

F) Why could you not use a z-test in your analysis above? (I.e., what additional information would you have needed to conduct a z-test)?

G) Describe in plain English the difference between a normal distribution and a t-distribution. Why does this difference occur?

H) Babcock & Marks (2011) also found that college students in the U.S. studied an average of 14 hours per week in 2010. We are interested in whether Psychology majors at CU tend to study more or less than the typical college student across the U.S. does today**. Perform the analysis in R and write a four-sentence summary of your results.** Include a measure of the effect size in your summary.

Part 2: The Two-Sample T-Test

Q2) We are interested in whether students who have higher GPA's (the independent variable) tend to study more than those with lower GPA's. We collected information on GPA and typical hours studied/week in lab the first week of class. Consider this class to be a sample of all college students (admittedly, it's not a very representative sample of all college students, and so we have to be careful about generalizing our results). We want to know whether the population mean of hours studied is higher for low than for high GPA students.

We first need to create a new variable in survey2013 that splits students into either "high gpa" or "low gpa" groups. We can somewhat arbitrarily do this by creating a "median" split in the data using this syntax:
survey2013$high.gpa <- survey2013$gpa.major > median(survey2013$gpa.major,na.rm=TRUE)

People with "TRUE" in this column have self-reported a major GPA that is above the median in the sample.

A) Use a two independent samples t-test to answer whether time spent studying (the dependent variable) differs depending on whether students have higher or lower GPAs (the quasi-independent variable). **What are your alternative and null hypotheses?** Write these out using math notation introduced in class (E.g., Ho: mu1 – mu2 = 0).

B) Use R to conduct a two-sample t-test. To do this using syntax (which I suggest), do this:

```
t.test(survey2013$study.hours.weekly ~ survey2013$high.gpa,var.equal=TRUE)
```

You can do the above through the Menu, but it requires several additional steps, so I haven't put that down here. The syntax above is common for a lot of tests in R – inside the t.test() parentheses is a formula. The dependent variable goes on the left of the tilda (~) and the independent variable on the right. In words, the syntax for the formula reads, "study.hours.weekly is a function of high.gpa."

What is your t-value and p-value? What do you conclude?

C) How "big" is this effect? Use a Cohen's d again to find this. A Cohen's d for a two sample t-test is a bit more complicated unfortunately. First, we have to find the "pooled standard deviation" – that is, roughly, the average of the standard deviation of the high gpa group and the standard deviation of the low gpa group. (Technically, it is the square root of the weighted average of the two variances). The formula for finding the pooled variances, also on p. 319 in the book, is:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{SS_1 + SS_2}{df_1 + df_2}$$

Let's use the first of the two equalities (i.e., the one with the variances). So first step here is to find the pooled variance of study.hours.weekly (the dependent variable), which you can do in R using syntax. First find the two variances:

```
tapply(survey2013$study.hours.weekly,survey2013$high.gpa,var,na.rm=TRUE)
```

Then find the sample sizes of the two groups:

```
summary(survey2013$high.gpa)
```

Now just multiply the variances by the respective degrees of freedom (their sample sizes -1), and divide the whole thing by the total degrees of freedom (total sample size -2). In R, we can do it algebraically:

```
pooled.var <- ((23*35.77)+(20*187.8))/(45-2)
pooled.sd <- sqrt(pooled.var)
pooled.sd  #to see what the actual value is
```

You can find the means of the two groups from the output of the t.test() function above. Finally, find the Cohen's d for two-samples:

$$\text{Cohen's d} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}}$$

Make sure that you FIRST subtract the two means, then divide that difference by the pooled standard deviation. Whew!! **What is your Cohen's d? Describe in words what this means.**

D) Create a side-by-side boxplot of the two groups using R syntax like this:
boxplot(survey2013$study.hours.weekly ~ survey2013$high.gpa)

Describe the shapes of the two groups. There is a substantial difference between the two distributions, but it isn't in the central tendency. What is it?

E) Write a four sentence summary of your findings. Make sure to include an estimate of the "effect size" (i.e., Cohen's d) of the effect you found, and include a side-by-side boxplot of hours studied below your 4-sentence summary.

F) Was this an experiment or not? Can we make causal inference based on this study (this is a hypothetical, so answer irrespective of whether or not your results were significant)?

G) If you collected data on 100s of additional students (the exact number is kept intentionally vague), but your effect size estimate (Cohen's d estimate) stayed the same, what would happen to your p-value? Explain, intuitively, why this occurs.

H) The assumptions of a two-sample t-test are (in order of importance):
a) the datapoints are independent of one another
b) the two populations from which the samples are drawn have equal variances
c) the scores from the two populations are normally distributed

The first assumption is critical – if we get this wrong, our p-value will be incorrect. The second assumption is moderately important, but there are easy ways around it. The final assumption isn't that important if n>30 due to the central limit theorem (the distribution of sample means will become normally distributed, and therefore the distribution of t-values will become t-distributed).

Usually, we can get information on the first assumption only through knowing the design of the study. Did scores of one student influence another student (i.e., maybe students were copying answers)? Did students come in groups of friends (in which case our degrees of freedom for the test are wrong)? In this case, let's assume the scores are independent of one another.

The final two assumptions can be guessed at by looking at the sample data, in particular, the side-by-side boxplot of the two groups in our sample give us some indication of whether the two latter assumptions are likely to have been violated. **Comment on the status of the final two assumptions for this t-test.**

Q3) You will need to use the formulas presented in the book, Chapter 10, and the lectures to answer this question**. <u>Do this problem by hand and show your work.</u>**

Here are the scores on a statistical reasoning test from people who had taken a stats class:
GROUP1: 10, 18, 16, 15

Here are the scores on a statistical reasoning test from people who had not taken a stats class:
GROUP2: 12, 10, 8, 6

We are interested in whether statistical reasoning test scores are better for people who have taken a stats class.

A) What is the null and alternative hypothesis?

B) What is your alpha level? Give an intuitive explanation behind what this number means.

C) What is the mean of Group 1? Of Group 2?

D) What is the sum of squares (i.e., the sum of squared deviations from the mean) of Group 1? Of Group 2? Show your work.

E) What is the pooled variance for Group1 and 2 together (denoted s_p^2 in the book)? The formula for this is in the problem above. **Report also the square-root of this, which is the pooled standard deviation of Group1 and 2 together.**

F) What is the standard error of the mean for the mean difference between Group 1 & 2 (denoted $s_{(M1-M2)}$ in your book)? The formula for the standard error of the mean difference is:

$$s_{(M1-M2)} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

G) Put into words what the standard error of the mean is above.

H) What is your t-value? How many degrees of freedom does this test have?

I) Use Table B2 in the back of your book to understand whether to reject the null hypothesis or not, or use R using the following syntax (pretending that we want a p-value for a t-value of 2.32 and had 10 degrees of freedom for our test – obviously, the specific numbers you put in will be different!):

```
pt(2.32,df=10,lower.tail=FALSE)*2
```

What p-value corresponds to your t-value? What is your conclusion with respect to the null hypothesis?

J) What is your estimate of effect size (Cohen's d)? Use your estimated mean difference in the numerator and the square root of the pooled variance in the denominator.