Stat 021 Homework 1

Colin Perkins-Taylor

Due: Fri, Sept 13

Instructions: A hard copy of your homework must be handed in to me at the end of class on the due date or I must have recieved via email a **pdf** version of your homework by **noon** on the due date. If you are having trouble getting your *.Rmd* file to compile, you need to get help with this **before** the due date.

You are allowed to hand in **only one** late homework assignment throughout the semester. If you need to hand in this particular assignment late, you must let me know via email by noon on the due date.

You are encouraged to study with your peers to help complete the homework assignments but no copying is allowed. If I see that two or more homework assignments are copied, all students involved will recieve a grade of 0 on that assignment and will forfeit (perhaps retroactively) the opportunity to hand in a late homework.

Q 1) Download and install R and R Studio following the instructions in class. Install the package *swirl()* using the command "install.packages("swirl")". Once the package is installed, call the package to your R session using the command "library("swirl"). Follow the instructions that pop up in your console. Select the course option"1: R Programming: The basics of programming in R" and complete the following lessons (about 20 minutes to complete each)

- 1: Basic Building Blocks
- 2: Workspace and Files
- 8: Logic

Even though you don't get points for doing this, it will really help you to learn how to program in R. If the tutorial is covering something that you already know how to do, use the *skip()* command to move through the tutorial faster, but note that there are some questions which you will not be able to skip and you'll be forced to think through. (0 points)

```
library("swirl")
```

```
##
## | Hi! Type swirl() when you are ready to begin.
```

Q 2) Design your own experimental study (made up) on a population of your choice. Provide details on what is the population under study, how will you collect a sample, what are potential biases keeping your sample from truely being random and define different treatment/control groups and describle how you will randomly assign treatments to your observational units. (5 points)

Answer:

The focus of my study would be to test a newly developed drug that provides energy to college students so that they can stay up later doing homework without feeling fatigue or losing focus. In order to test this new drug, college students at various colleges and universities across the United States would be recruited. Sampling students from different size colleges (small liberal arts to large state universities) and who have different study habits would be essential for understanding the effects of the drug because these factors may explain the results. Students from at least one small college (1,000-2,000 students) and one large university (3,000-10,000+ students) would be tested in each state, and the baseline of each student's study and sleep behaviors would be recorded before they were tested. During testing, the students would not be allowed to consume any coffee or other substances that they may normally use to stay awake at night and do homework because this maximizes the likelihood that any behavioral changes would be due to my drug. Electronic and paper advertisements for the study would be distributed across each campus both physically and electronically

to maximize the probability of students learning about the study and participating in it since it would be on a volunteer basis.

Once the enrollment period for the study ended, within each college or university half of the students would randomly be assigned to take the drug while the other half would take a supplement that mirrored the drug, but did not actually do anything. The randomization would be done using a computer program with a random number generator that could produce a 1 (drug) or 0 (no drug). Therefore, each college or university would have an energy group (the test group of people taking the drug) and a control group(the test group of people taking the supplement, not the drug). In addition, there would be an overall energy group that included everyone taking the drug at all of the colleges and universities as well as an overall control group of people taking the supplement rather than the drug. Both the overall and college/university-specific energy and control groups could further be separated into colleges or universities exclusively, as there may be some differences between the two (although there shouldn't be).

The main potential bias preventing my sample from being truly random is that the drug is most likely more appealing to students who work at night and prefer staying up late rather than those who do work during the day, in the morning, or go to bed early. Although the drug should provide energy to those who wake up in the morning to do work, since they are well-rested they may already have energy or have other means of getting energy in the morning. This means that the sample may be inherently biased since it most likely only applies to one demographic of college students work behaviors rather than all of them. However, extra effort would be made to include students with various work behaviors to minimize these biases.

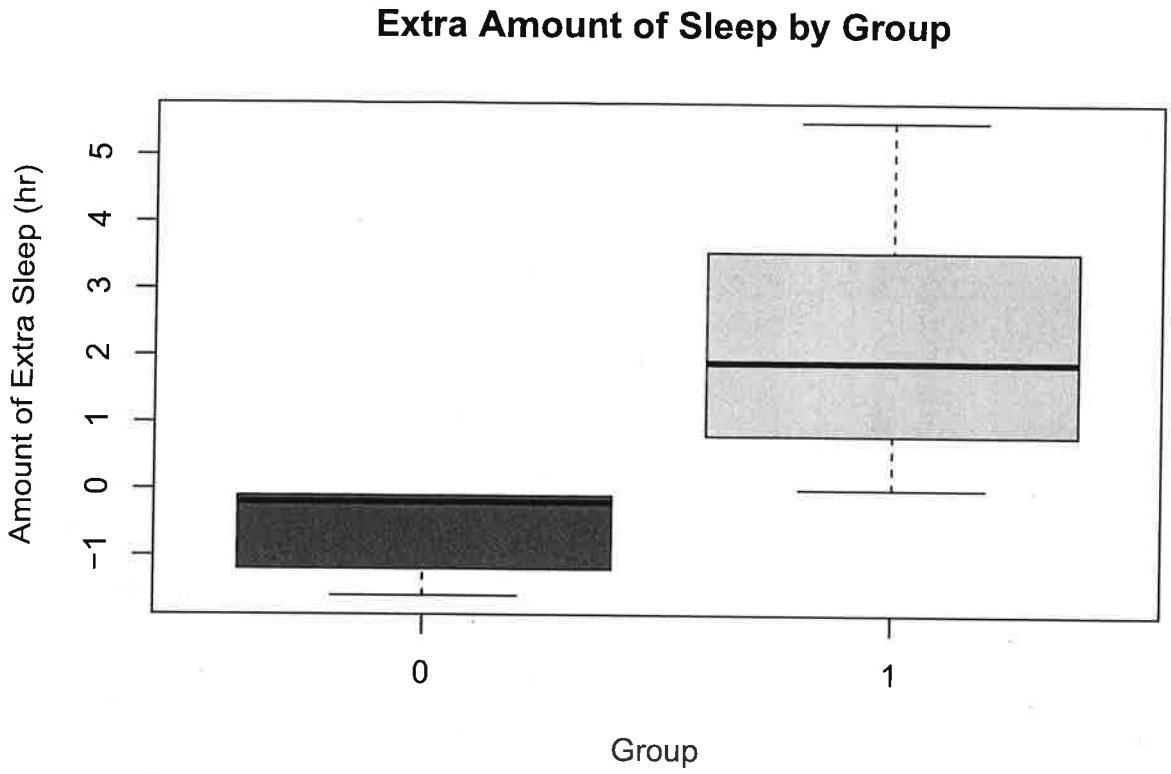
Q 3)

```
group1.sleep <- filter(sleep, group == 1)

t.test(x = group1.sleep$extra, alternative = "less", mu = 0.5, conf.level = 0.90)

##
##  One Sample t-test
##
## data:  group1.sleep$extra
## t = 0.4419, df = 9, p-value = 0.6655
## alternative hypothesis: true mean is less than 0.5
## 90 percent confidence interval:
##      -Inf 1.532427
## sample estimates:
## mean of x
##      0.75
sleep.cat1 <- mutate(sleep, extra1.cat = ifelse(extra >= 0, 1, 0))
count(sleep.cat1, extra1.cat == 1)

## # A tibble: 2 x 2
##   `extra1.cat == 1`     n
##   <lgl>             <int>
## 1 FALSE                 5
## 2 TRUE                 15
boxplot(extra ~ extra1.cat, data = sleep.cat1, ylab = "Amount of Extra Sleep (hr)",
        xlab = "Group", main = "Extra Amount of Sleep by Group",
        col = c("red", "orange"))
```



Access the data set called *sleep* in R. Note this data set is loaded into R automatically (you do not need to import it or install a package to access it). (10 points)

1. Create a new data set called *group1.sleep* that only contains data for individuals from group 1.
2. Perform a t-test on "group1.sleep" to test of the extra hours slept by group 1 is smaller than or equal to 0.5 hours at an $\alpha = 0.1$ level. I.e. Test the hypothesis $H_0 : \mu >= 0.5$ vs. $H_1 : \mu < 0.5$ at an $\alpha = 0.1$ significance level.
3. Report and interpret the 90% CI for the average extra hours of sleep for group 1.
4. Form a new categorical variable called *extra1.cat* that categorizes the variable *extra* into two groups, the first where extra hours slept is at least zero hours and the second where extra hours slept is less than zero hours. Print a table that counts the total number of observations in each group. (You may want to use the *ifelse()* function.)
5. Produce two boxplots for the variable *extra*, one corresponding to each group. Make sure each plot has a title. (You may find the function *grid.arrange()* in the package *gridExtra* useful for displaying two plots at once.)

Solution:

3. The 90% confidence interval for the average extra hours of sleep for group 1 is (-infinity, 1.53). This means that we are 90% confident that the population mean, μ, for the average extra hours of sleep for group 1 is within (-infinity, 1.53). Since we are 90% confident that μ is between -infinity and 1.53, the evidence does not support the alternative hypothesis that $\mu < 0.5$ extra hours of sleep for group 1, and therefore we cannot reject the null hypothesis (p = 0.6655).

Q 4) Suppose $X \sim N(\mu, \sigma^2)$. Show that the random variable defined as $Z = \frac{X-\mu}{\sigma}$ has $E[Z] = 0$ and $Var[Z] = 1$. Show all of your steps (you may handwrite your answer to this question). (5 points)

Hint: Recall/look up some common properties of normal random variables and the rules of the expectation and variance operations.

4) $X \sim N(\mu, \sigma^2)$ $\quad E[Z] = 0$ $\quad E[X] = \mu$

$\text{Var}[Z] = 1$ $\quad \text{Var}[X] = \sigma^2$ $\quad Z = \frac{X - \mu}{\sigma}$

Show random variable Z has $E[Z] = 0$:

$$E[Z] = E\left[\frac{X-\mu}{\sigma}\right] = E\frac{[X-\mu]}{\sigma} = \frac{E[X]-\mu}{\sigma} = \frac{\mu-\mu}{\sigma}$$

$$= \frac{0}{\sigma} = \boxed{0}$$ $E[Z]$ does equal 0.

Show random variable Z has $\text{Var}[Z] = 1$:

$$\text{Var}[Z] = \text{Var}\left[\frac{X-\mu}{\sigma}\right] = \text{Var}\left[\frac{X}{\sigma} - \frac{\mu}{\sigma}\right]$$

$$= \left(\frac{1}{\sigma}\right)^2 \text{Var}(X) = \frac{\text{Var}(X)}{\sigma^2} = \frac{\text{Var}(X)}{\text{Var}(X)} = \boxed{1}$$

$\text{Var}[Z]$ does equal 1.