Stat 021 Homework 1

Christina Holmgren

Due: Friday, 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)

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)

AnswerThis expirmental study will test whether eating a healthy breakfast has an affect on percieved energy levels. The test will occur 3 times, on different campuses in Philedelphia. The tested individuals will be randomly split into two groups (using computer). First, individuals from both groups will be instructed not to eat breakfast for a week and will report their energy levels every hour. Next the individuals from one of the groups will be given the same breakfast (chosen by nutritionist) every day at 8 am for a week while the other group will continue to skip breakfast. Everyone will report energy levels every hour and are instructed not to have lunch until noon.

The population being studied is full-time undergrad students (aged ~18-22) in the US. The sample will collected from a study of full-time students in the city of Philedelphia at three different universities using a small ($30) cash incentive as well as the potential for free breakfast. The students can record their percieved energy levels themselves and either send them in or bring them to the facility at the end of the expiriment. The food will be handed out to the students at the university at a chosen locations (the students will have

to arrive there and pick up/eat the food). There will be 200 students selected per location, half of them will be given breakfast in the second week of the expiriment.

The sample will not be truley random because the sample will only be taken from students in Philedelphia, rather than from all of the US. Additionally, only students from the 3 selected Universities where the expirment is taking place will be given the opportunity to participate. Additionally, the study will attract students who need the money, live close to the location that hands out the food, or are interested in the potential for free food. These students may not be representative of the population.

The control group will be the individuals that do not eat a breakfast for the two consecutive weeks. The "treated" group will be the one that is given breakfast after the first week. The week 1 and week 2 data collected on inviduals who were given the treatment will allow researchers to see difference in percieved energy levels before and after eating breakfast. The control group allows a comparison to a control group who has not undergone this change in eating habits.

Q 3) 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. Define a new variable called *group1.sleep* that includes only the values of the variable *extra* 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. $H_1 : \mu \neq 0.5$ at an $\alpha = 0.1$ significance level.
3. Report and interpret the 90% CI for the average extra hours of sleep for group 1.

A 90% confidence interval is a range of values in which yu can be 90% certain the true mean of extra hours of sleep for group 1 falls in.

1. 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.)
2. 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.)

// TO DO plot1 <- geom_boxplot(data = group1.sleep[extra1.cat == "more"], data = sleep, main = "Extra hours of sleep at least 0 hours") plot2 <- geom_boxplot(group1.sleep[extra1.cat == "less"], data = sleep, main = "Extra hours of sleep less than 0 hours") grid.arrange(p1, p2 , ncol = 2)

Do me a favor and write your solutions to the different parts of Q 3 all in the same space (not between each bulletted list item). You can include a chunk of R code like this:

90 percent confidence interval: -Inf 1.532427 A 90% confidence interval is a range of values in which you can be 90% certain the true mean of extra hours of sleep for group 1 falls in. It suggests that the number of hours slept by group 1 has a one-sided 90% upper bound of 1.532427 hours.

```
data(sleep)
sleep <- sleep
sleep %>% head()
```

```
##   extra group ID
## 1   0.7     1  1
```

```
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

```
group1.sleep <- sleep[sleep[, "group"] == 1,]
t.test(group1.sleep$extra, mu = 0.5, alternative = "less", paired= FALSE, 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
```

```
extra1.cat <- ifelse(group1.sleep$extra < 0, "less_hours", "more_hours")
table(extra1.cat)
```

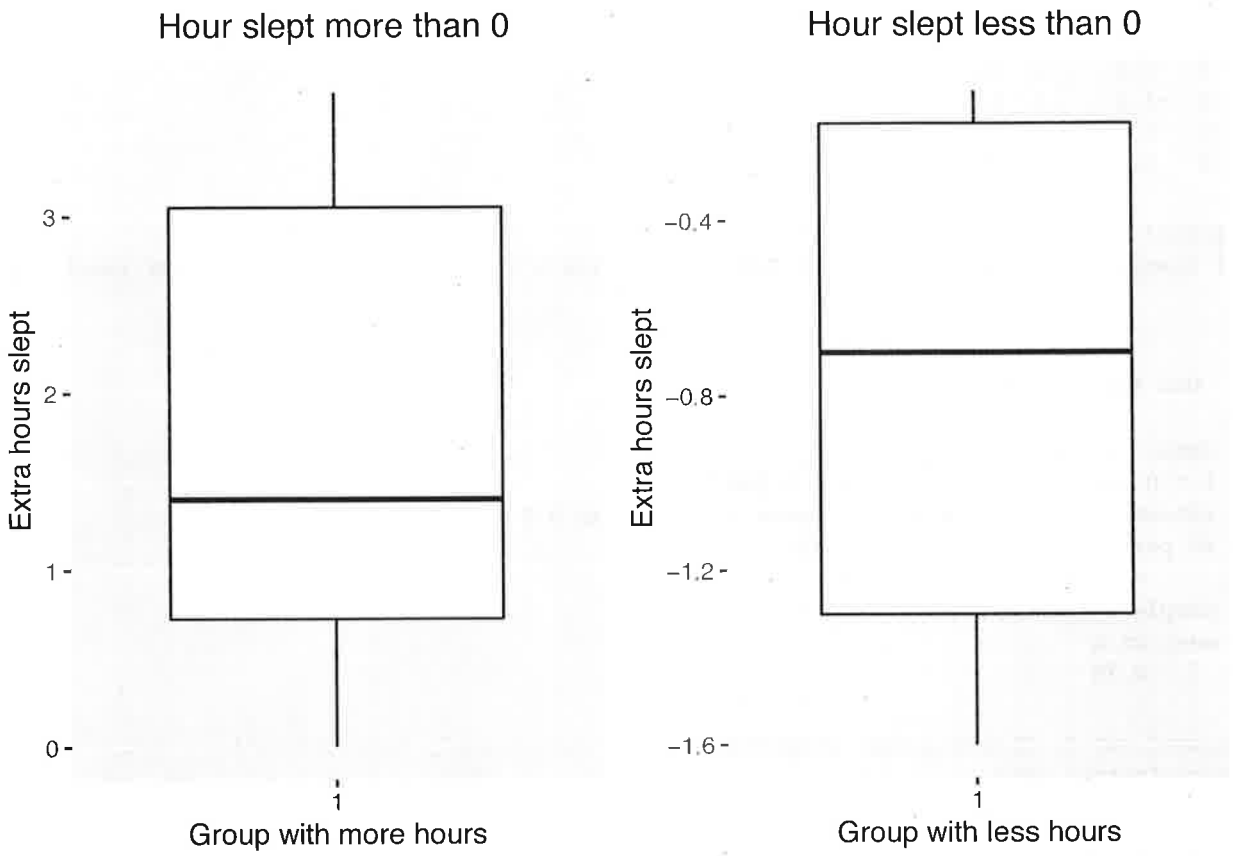
```
## extra1.cat
## less_hours more_hours
##          4          6
```

```
more_hours <- group1.sleep[group1.sleep[, "extra"] >= 0,]
less_hours <- group1.sleep[group1.sleep[, "extra"] < 0,]

plot_one_more <- ggplot(more_hours, aes(x= group, y = extra)) +
geom_boxplot()+
ggtitle("Hour slept more than 0") + ylab("Extra hours slept") +
xlab("Group with more hours") + theme(plot.title = element_text(hjust = 0.5))

plot_two_less <- ggplot(less_hours, aes(x= group, y = extra)) +
geom_boxplot()+
ggtitle("Hour slept less than 0") + ylab("Extra hours slept") +
xlab("Group with less hours") + theme(plot.title = element_text(hjust = 0.5))

grid.arrange(plot_one_more, plot_two_less, nrow = 1)
```

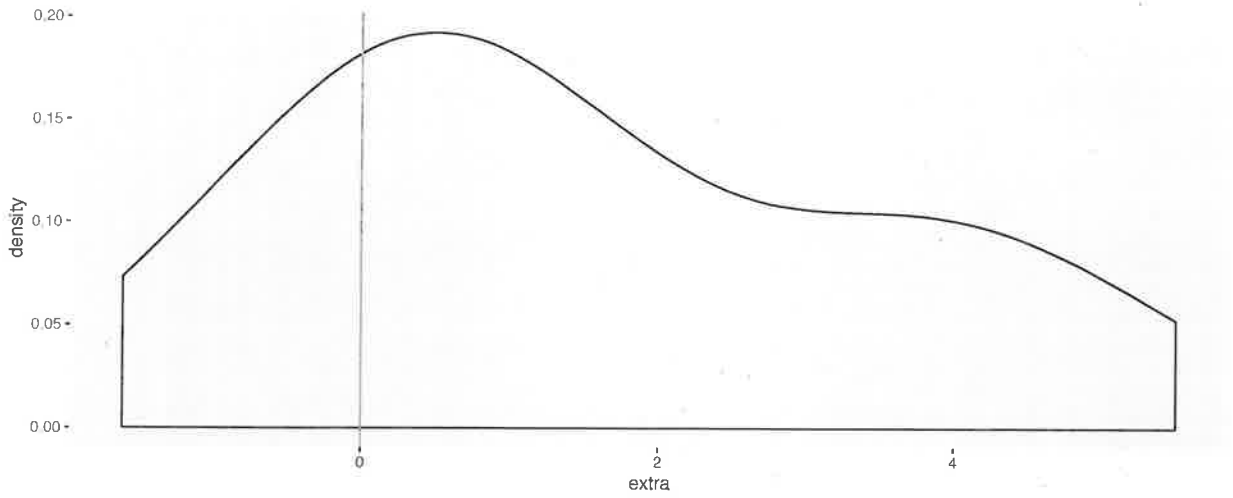


Note that the deliminators for r code are **not** apostraphes but are the tick marks found in the upper left hand corner of your keyboard. You will never need to print out an entire data set for me in your homework, just the fist few rows using the *head()* function is fine.

Another note, the "echo=TRUE" and "warning=FALSE" options in your R code chunk are settings that will make the incorporation of your code into your document a lot neater. These options tell R to print the output of the code to your document and to not print any warning signs that may come up in the console, respectively.

To include a plot, I recomend the following options for your R code chunk:

```
myPlot <- ggplot(sleep, aes(x=extra)) + geom_density() +
                                        geom_vline(xintercept=0, col="red")
myPlot
```



In the code above, *aes()* is short for aesthetic which doesn't make a whole lot of sense to me, regardless, it is the function that enables you to define your x (and y) variable(s).

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.

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

$E[\mu] = \mu$

So $E[Z] = \frac{\mu-\mu}{E[\sigma]}$ $\quad \mu - \mu = 0$ so $\frac{0}{E[\sigma]} = 0$

$E[Z] = 0$

$Var[Z] = Var\left[\frac{X-\mu}{\sigma}\right]$ pull out $Var\ \frac{1}{\sigma} \rightarrow \frac{1}{\sigma^2}$

$\frac{1}{\sigma^2} \cdot Var[X-\mu]$ $\quad Var[X-\mu] \rightarrow Var[X]$

$\frac{1}{\sigma^2} Var[X]$ $\quad Var[X] = \sigma^2$

$\frac{1}{\sigma^2}\sigma^2 \rightarrow \frac{\sigma^2}{\sigma^2} = 1$ $\quad Var[Z] = 1$

Stat 021 Homework 1

Daniel Lee

Due: Wed, Sept 11

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)

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)

Solution: I want to test whether giving children candy after they receive shots at the hospital makes their overall experience better. The population under study will be children who are coming to the hospital to get immunization. The control group will be children who don't get candy while the treatment group gets candy after receiving shots. Both groups will give a numerical rating between one and ten regarding their hospital visit.

To start a new paragraph, make sure you leave enough line breaks between your text. To include mathematical expressions in a R Markdown document, use the same format as you would for a LaTex document and surround the equation with dollar signs like this: $\sum_{i=1}^{n}(y_i^2 + \bar{y})^2 = 5$ for inline expressions and with double dollar signs for expressions centered on their own line such as

$$\sum_{i=1}^{n}(y_i^2 + \bar{y})^2 = 5.$$

Q 3) 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. Define a new variable called *group1.sleep* that includes only the values of the variable *extra* 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. $H_1 : \mu \neq 0.5$ 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.)

Do me a favor and write your solutions to the different parts of Q 3 all in the same space (not between each bulletted list item). You can include a chunk of R code like this:

```
data(sleep)
sleep %>% head()
```

```
##   extra group ID
## 1   0.7     1  1
## 2  -1.6     1  2
## 3  -0.2     1  3
## 4  -1.2     1  4
## 5  -0.1     1  5
## 6   3.4     1  6
```

```
#Part 1)
group1.sleep <- sleep %>% filter(group==1) %>%
select(extra)
#Part 2)
t.test(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
```

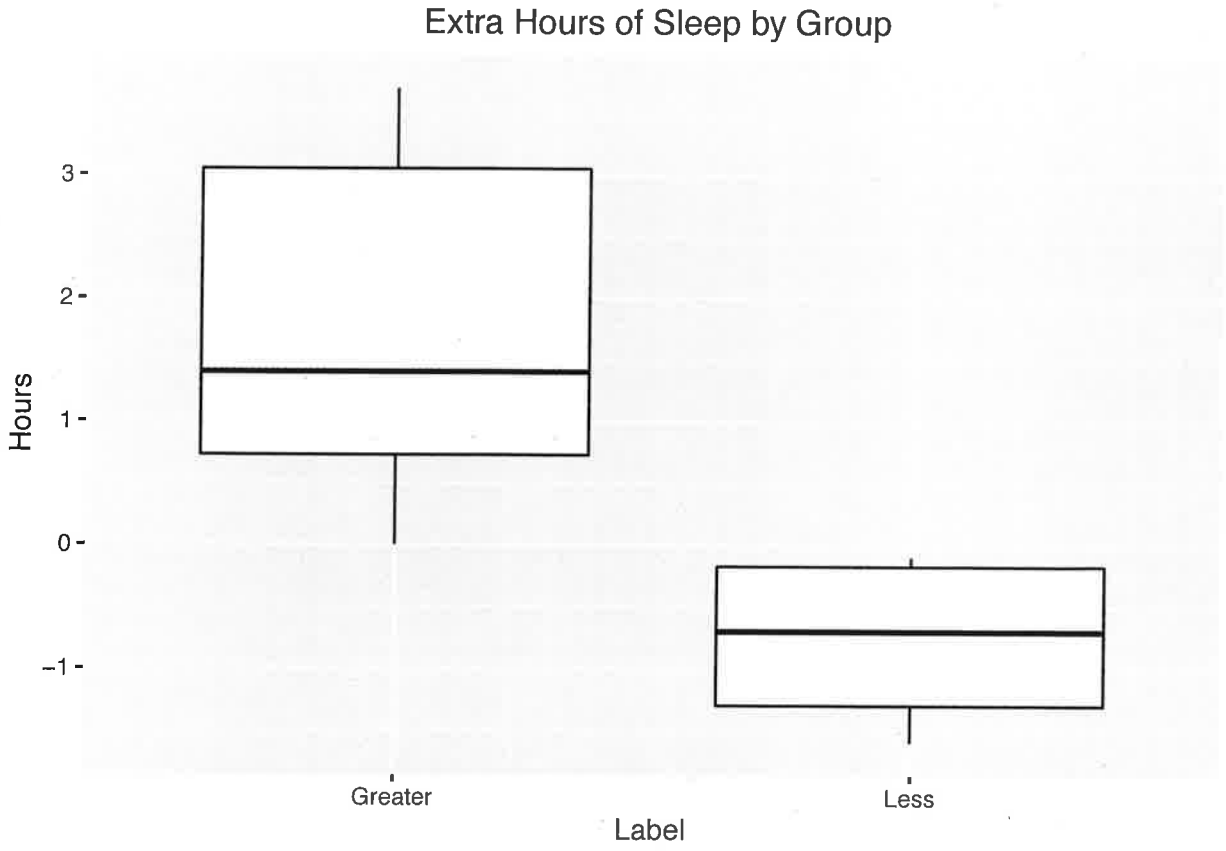
Part 3) Because the p-value (0.6655) is less than alpha (0.10), we fail to reject the null hypothesis in favor of the alternative hypothesis. The 90 percent confidence interval shows that the true mean is located between -Inf to 1.53 90 percent of the time.

```
#Part 4
group1.sleep <- group1.sleep %>%
  mutate(extra1.cat = ifelse(group1.sleep$extra >= 0,"Greater","Less"))
group1.sleep
```

```
##    extra extra1.cat
## 1    0.7    Greater
## 2   -1.6       Less
## 3   -0.2       Less
## 4   -1.2       Less
## 5   -0.1       Less
## 6    3.4    Greater
## 7    3.7    Greater
## 8    0.8    Greater
## 9    0.0    Greater
## 10   2.0    Greater
```

```
#Part 5)
library(gridExtra)
library(ggplot2)
colnames(group1.sleep)[colnames(group1.sleep)=="extra"]  <- "Hours"
colnames(group1.sleep)[colnames(group1.sleep)=="extra1.cat"]  <- "Label"

myPlot <- ggplot(group1.sleep, aes(x=Label,y=Hours)) + geom_boxplot()
myPlot <- myPlot + labs(title = "Extra Hours of Sleep by Group") + theme(plot.title = element_text(h
myPlot
```



Q 4) Suppose $X \sim N(\mu, \sigma^2)$. Show that the random variable defined as $Z = \frac{Z-\mu}{\sigma}$ 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.

Stat 21 - Homework 1

Daniel Lee
9/13/19

4) $E(Z) = E\left(\frac{X-\mu}{\sigma}\right) = \frac{E(X) - E(\mu)}{\sigma} = \frac{\mu - \mu}{\sigma} = 0$ ✓ (given)

$\text{var}(Z) = \text{var}\left(\frac{X-\mu}{\sigma}\right) = \frac{\text{var}(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$ ✓ (given)