```
---
title: "Stat 021 Homework 1"
author: "Tyler Soutendijk"
date: "Due: Fri, Sept 13"
output: pdf_document
---

```{r setup_pres, include=FALSE, echo=FALSE}
rm(list=ls())
library('tidyverse')                                    ##All of
this code here is called the "preamble"
library('gridExtra')
library('cowplot')

setwd("~/Desktop/Fall 2019 Courses/STAT")  ##fill in with your
own directory instead
```

19 celcius temp

<center>**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. </center>

***

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:

My experimental study will investigate a population of college students at Swarthmore (n=100). The sample will be selected as the first 100 students to sign up that sleep on the standard Swarthmore housing spring mattresses. If eligable for the study, college students will be compensated for their time. The experimental study tests the correlation between sleeping on a memory foam mattress and alertness. A hypothesis would be that those who sleep on a memory foam mattress have greater alertness in the morning than those who sleep on a spring mattress. One bias that would be evident here is volunteer or self-selection bias from those who will opt to participate in the study because of the compensation. Another bias is for those with either pre-existing insomnia (sleeping issues) OR a

deficit in alertness. These biases will absolutely keep the sample from being truely random.

Once a sample is selected, treatment groups will be assigned by random number generation within a set of 100 numbers. If a student (basic observational unit) recieves an odd number then they are assigned to the control group. If a student receives an even number then they are assigned to the treatment group. The control group will sleep on a spring mattress that is masked as a foam mattress. The treatment group will sleep on a foam mattress. Both groups will spend the night on their mattresses and then tested for alertness through cognitive testing in the morning. The cognitive testing will involve exercises similar to the Acute Concussion Evaluation (ACE) test. Results will be recorded and a one-sided z-test will determine if there is ample evidence to favor a memory foam mattress to a spring mattress for alertness.

If you use **double asterisks** to surround a phrase or word to make it bold. Use *single asterisks* to make italics.

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.

1. 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.

1. Report and interpret the 90% CI for the average extra hours of sleep for group 1.

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.)

1. 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:

```{r myLabelForThisChunk, echo=TRUE, warning=FALSE}
data(sleep)
sleep %>% head()

# Question 1
group1.sleep <- sleep %>%
  filter (group == 1) %>%
  select(extra)

# Question 2
t.test(group1.sleep, alternative="less", mu=0.5,
conf.level=0.90)

# Question 3
# From the one sample t-test of group1.sleep, I don't have enough evidence to reject the null hypothesis in favor of the alternative hypothesis (p-value (0.6655) is not less than the alpha level (0.10)). The 90 percent confidence interval shows that the true mean is 90 percent likely to be located within the interval (-Inf to 1.53).

# Question 4
group1.sleep <- group1.sleep %>%
  mutate (extra1.cat <-
```

```
ifelse(group1.sleep>=0,"Greater","Less"))
colnames(group1.sleep) <- c("extra", "extra1.cat")
table(group1.sleep$extra1.cat)

# Question 5
extra_hours_boxplot1 <- ggplot(data = group1.sleep,
aes(group1.sleep$extra1.cat, group1.sleep$extra)) +
geom_boxplot() + labs(title = "Extra Hours of Sleep", x =
"Greater or Less than Zero Hours?", y = "Number of Hours")
#extra_hours_boxplot2 <- ggplot(data = group1.sleep,
aes(x=group1.sleep$`extra hours`)) + geom_density() +
 #
geom_vline(xintercept=0, col="red")
 # I prefer the above formatting for my boxplot
extra_hours_boxplot1
```

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:

```
```{r aDifferentCodeChunk, echo=TRUE, warning=FALSE, fig.height
= 4, fig.width = 10, fig.align = 'center'}
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 $Z = \frac{Z-\mu}{\sigma}$ 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.

For $E[Z]=0$

$$Z = \frac{x-\mu}{\sigma}$$

$$E[Z] = \frac{E[x]-E[\mu]}{E[\sigma]}$$

$X \sim N(\mu,\sigma^2)$ $E[\mu]=\mu$ & $E[\sigma]=\sigma$ because they are parameters

$$\therefore E[Z] = \frac{E[x]-\mu}{\sigma}$$

$E[x]=\mu$ because it is an estimate of μ (x)

$$\therefore E[Z] = \frac{\mu-\mu}{\sigma}$$

$$E[Z] = \frac{0}{\sigma}$$

$$E[Z] = 0 \checkmark$$

For $Var[Z]=1$

$$Z = \frac{x-\mu}{\sigma}$$

$$Var[Z] = \frac{Var[x]-Var[\mu]}{Var[\sigma]}$$

$Var[x]=\sigma^2$

$Var[\mu]=0$ w/ $X \sim N(\mu,\sigma^2)$

$Var[\sigma]=\sigma^2$

$$Var[Z] = \frac{\sigma^2-0}{\sigma^2}$$

$$Var[Z] = 1 \checkmark$$

Stat 021 Homework 1

Chongkyung Kim

Due: Fri, Sept 13

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)

Does healthy soil help Swarthmore trees grow?

Population: trees on the Swarthmore campus

Sample Selection: Use random generator (using python, for example) to select 200 trees.

Random Assignment: Use random generator and select 100 trees; they are the treatment group, they are provided with healthy soil, engineered from a lab. Rest of the 100 trees do no receive any treatment and we let them grow in their natural setting and state.

Treatment Group: the randomly selected trees that grow in the artificially engineered, healthy soil

Control Group: the other 100 trees that just grow in their natural, original setting

Potential Biases: There's a chance that really healthy trees, that perhaps might not have needed the help of healthy soil, were assigned to the treatment group, thus keeping our sample from being truely random. There's also a chance that the very unhealthy, weak trees were assigned to the control group, which could actually have benefited from the healthy soils. These types of unexpected factors might have prevented us from having a truely random sample.

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.

We can be 90 percent confident that the true mean of the extra hours of sleep by group 1 is in the interval of (-inf , 1.53427]

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.)

```
data(sleep)

new_sleep <- sleep %>% filter(group==1) %>% mutate(group1.sleep = extra)

new_sleep
```

```
##    extra group ID group1.sleep
## 1    0.7     1  1          0.7
## 2   -1.6     1  2         -1.6
## 3   -0.2     1  3         -0.2
## 4   -1.2     1  4         -1.2
## 5   -0.1     1  5         -0.1
## 6    3.4     1  6          3.4
## 7    3.7     1  7          3.7
## 8    0.8     1  8          0.8
## 9    0.0     1  9          0.0
## 10   2.0     1 10          2.0
```

```
t.test(new_sleep$group1.sleep, alternative= "less" , mu=0.5,conf.level = .9)
```

```
##
##  One Sample t-test
##
## data:  new_sleep$group1.sleep
## 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
```

```
categorized_sleep1 <-sleep %>% filter(extra <0) %>% mutate(extra1.cat = "less than zero")
categorized_sleep2 <- sleep %>% filter(extra>=0) %>% mutate(extra1.cat ="at least zero")

categorized_sleep = rbind(categorized_sleep1,categorized_sleep2)

categorized_sleep
```

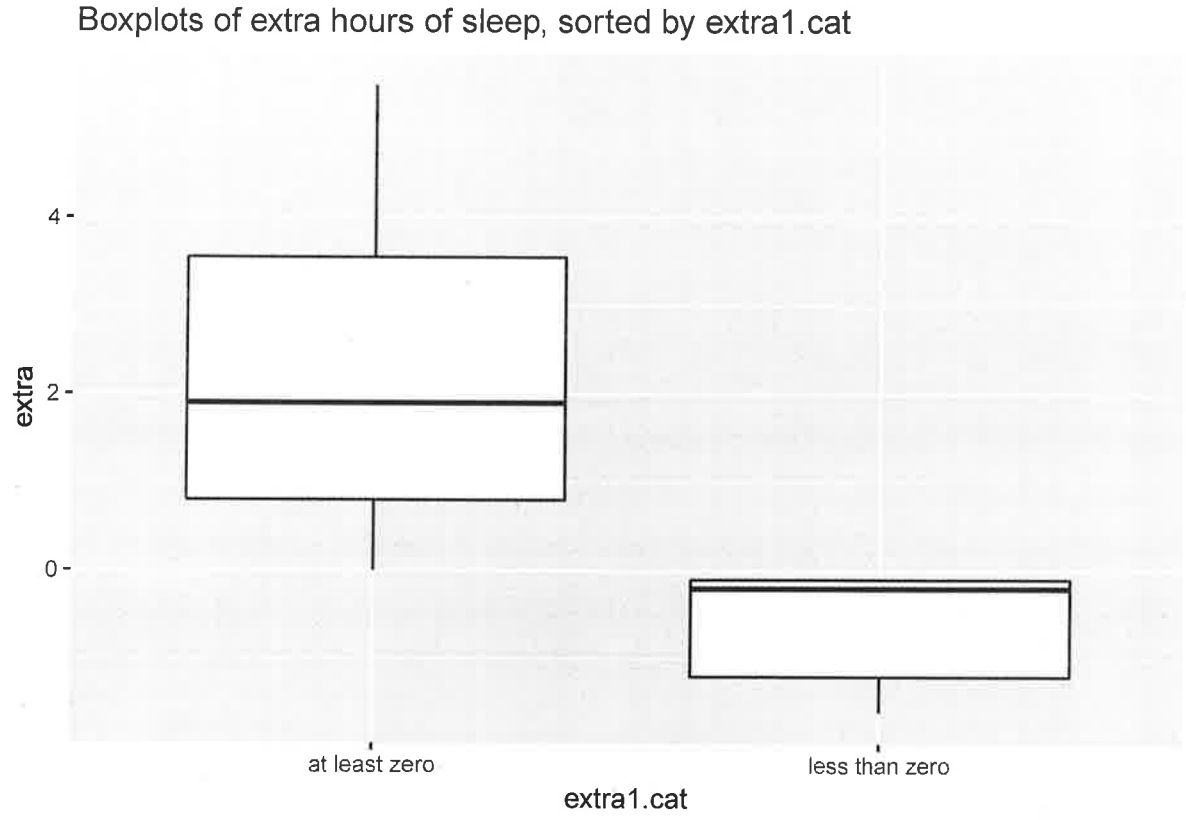
```
##    extra group ID     extra1.cat
## 1   -1.6     1  2 less than zero
## 2   -0.2     1  3 less than zero
## 3   -1.2     1  4 less than zero
## 4   -0.1     1  5 less than zero
## 5   -0.1     2  5 less than zero
## 6    0.7     1  1  at least zero
## 7    3.4     1  6  at least zero
## 8    3.7     1  7  at least zero
## 9    0.8     1  8  at least zero
## 10   0.0     1  9  at least zero
## 11   2.0     1 10  at least zero
## 12   1.9     2  1  at least zero
## 13   0.8     2  2  at least zero
## 14   1.1     2  3  at least zero
## 15   0.1     2  4  at least zero
## 16   4.4     2  6  at least zero
## 17   5.5     2  7  at least zero
## 18   1.6     2  8  at least zero
## 19   4.6     2  9  at least zero
## 20   3.4     2 10  at least zero
```

```
categorized_sleep %>% count(extra1.cat)
```

```
## # A tibble: 2 x 2
##   extra1.cat         n
##   <chr>          <int>
## 1 at least zero     15
## 2 less than zero     5
```

```
count_sleep <- categorized_sleep %>% count(extra1.cat)

p1 <- ggplot(data=categorized_sleep)+geom_boxplot(mapping = aes(extra1.cat,extra)) +
       labs(title = "Boxplots of extra hours of sleep, sorted by extra1.cat")
p1
```



Definition that $E(x) = \mu$ and $Var(x) = \sigma^2$.

then,

$$E(z) = E\left(\frac{X-\mu}{\sigma}\right) = E\left(\frac{x}{\sigma} - \frac{\mu}{\sigma}\right)$$

$$= \frac{E(x)}{\sigma} - \frac{\mu}{\sigma}$$

$$= \frac{\mu}{\sigma} - \frac{\mu}{\sigma} \quad (\text{since } E(x) = \mu)$$

$$= 0$$

$$Var(z) = Var\left(\frac{x-\mu}{\sigma}\right) = Var\left(\frac{X}{\sigma} - \frac{\mu}{\sigma}\right)$$

$$= \frac{1}{\sigma^2} Var(x) \quad (\text{since } Var(x) = \sigma^2)$$

$$= 1.$$

Name _____ Class _____ Date _____

I. Identifying someone or something, using ～は　～です

Imagine that you have a pen pal in Japan and that the following people live in your dormitory. Write a short description of each student to your pen pal. You may use **hiragana** for the words written in **katakana**.

なまえ Name	～じん Nationality	～ねんせい Year in school	せんこう Major
ブラウン ぶらうん	アメリカじん あめりか	いちねんせい	こうがく
キム きむ	かんこくじん	だいがくいんせい	ビジネス びじねす
チョー ちょ	ちゅうごくじん	にねんせい	えいご
モネ もね	カナダじん かなだ	よねんせい	アジアけんきゅう あじあ
スミス すみす	オーストラリアじん おーすとらりあ	さんねんせい	ぶんがく

■ Example ブラウンさんは　アメリカじんです。いちねんせいです。
（ぶらうん　あめりか）
せんこうは　こうがくです。

1. キムさんは _____。 _____。
（きむ）

せんこうは _____。

2. チョーさんは _____。 _____。
（ちょ）

せんこうは _____。

3. モネさんは _____。 _____。
（もね）

_____。

4. スミスさんは _____。 _____。
（すみす）

_____。

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.

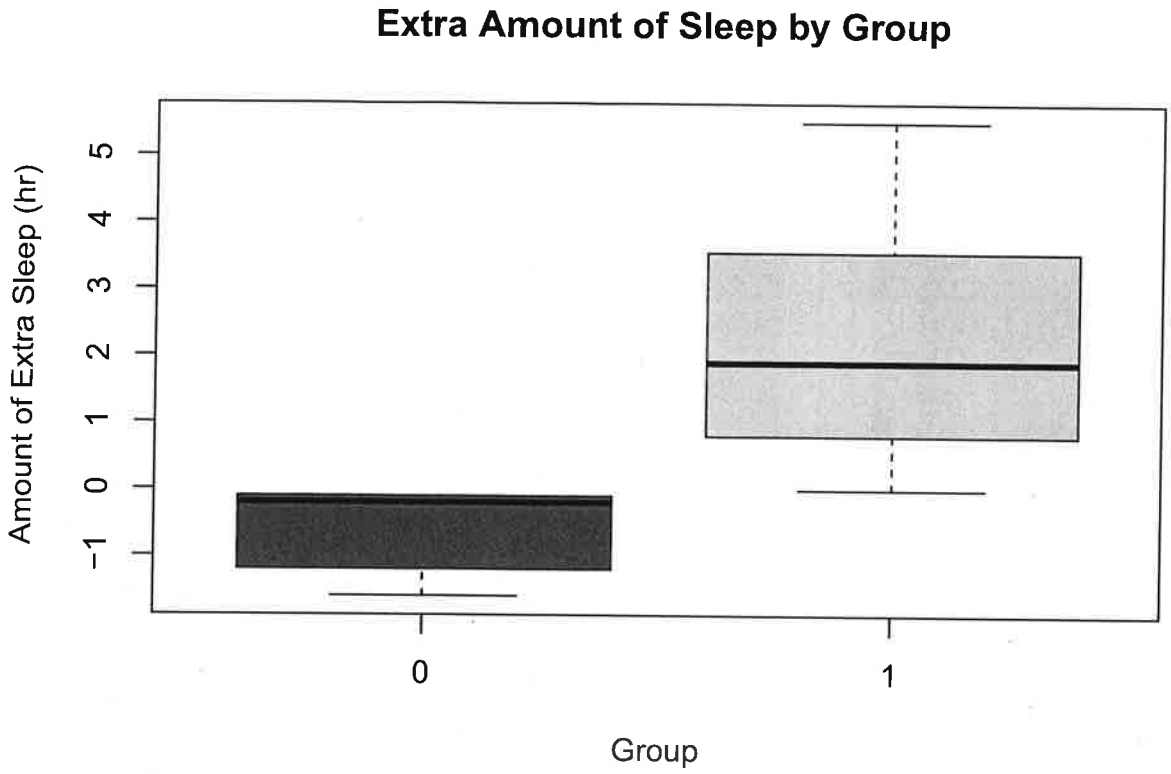
Q 3)

```
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.

Stat 021 Homework 1

Misha Mubashar Khan

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)

Answer This experimental study tests whether caffeine in coffee has an affect on heart rate. Individuals tested will be asked not to eat 6 hours prior to the testing time, and will be given an 8oz serving of dark roast coffee. Their heart rate will be recorded immediately before drinking the coffee and 15 minutes after drinking the coffee. Each individual will repeat this in 3 seperate trials.

The population under study are all full-time undergraduate students between the ages of 18-22. The sample of the study will be collected by opening the study up for all full-time college students in the state of PA, using a $40 cash prize as an incentive, with the caveat that the students can travel to the testing facility. Up to the first 100 registered students will be chosen for the sample.

However, several biases may keep the sample from being truly random. Firstly, only students in the state of PA are offered the study, and hence the sample size will not include any college students from the rest of the states. Secondly, even within the state, students living nearby the testing facility are more likely to be

included in the sample as compared to students living further away, which makes commuting to the facility harder.

The study will have one treatment group that drinks coffee, and one control group that does not drink coffee. The sample can randomly be split into two groups using an algorithm that places each student into either group. Both groups will follow the same process as outlined in the first paragraph, except the control group will be given hot water instead of coffee.

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. 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 $H_0 : \mu \geq 0.5$ 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.)

Answer

90 percent confidence interval: -Inf 1.532427

A 90% confidence interval is a range of values that you can be 90% certain contains the true mean of extra hours of sleep for group 1. The CI provides the tightest upper bound on the sample mean, suggesting that the number of extra hours slept by group one is has a one-sided upper 90% confidence bound of 1.532427 hours.

```
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
```

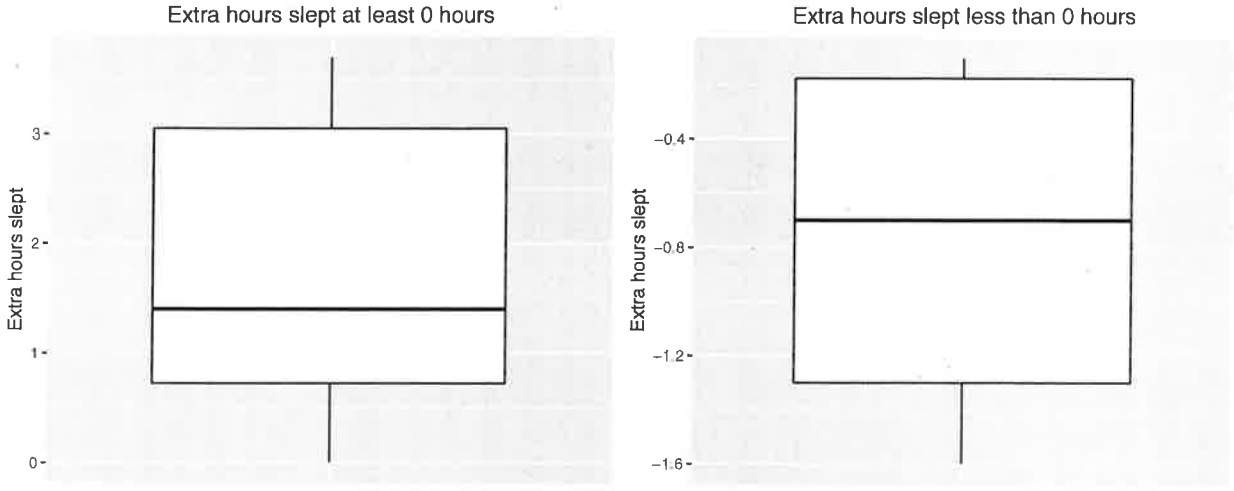
```
## alternative hypothesis: true mean is less than 0.5
## 90 percent confidence interval:
##      -Inf 1.532427
## sample estimates:
## mean of x
##      0.75

extra1.cat <- ifelse(group1.sleep$extra >= 0, "more", "less")
table(extra1.cat)


## extra1.cat
## less more
##    4    6

more <- group1.sleep[group1.sleep[, "extra"] >= 0,]
less <- group1.sleep[group1.sleep[, "extra"] < 0,]

pmore <- ggplot(more, aes(x= group, y = extra)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=4) +
  ggtitle("Extra hours slept at least 0 hours") + ylab("Extra hours slept") +
  xlab("More") + theme(plot.title = element_text(hjust = 0.5),
                       axis.title.x=element_blank(), axis.text.x=element_blank(),
                       axis.ticks.x=element_blank())
pless <- ggplot(less, aes(x= group, y = extra)) +
  geom_boxplot(outlier.colour="red", outlier.shape=8, outlier.size=4) +
  ggtitle("Extra hours slept less than 0 hours") + ylab("Extra hours slept") +
  xlab("Less") + theme(plot.title = element_text(hjust = 0.5),
                       axis.title.x=element_blank(), axis.text.x=element_blank(),
                       axis.ticks.x=element_blank())
grid.arrange(pmore, pless, nrow = 1)
```



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.

Answer $X \sim N(\mu, \sigma^2)$

~~$Z = \frac{Z - \mu}{\sigma}$~~ ~~$[Z] = 0$~~

Given : $Z = \frac{X - \mu}{\sigma}$

① Show $E[Z] = 0$.

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

Substitute $Z = \frac{X - \mu}{\sigma}$

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

$E[\mu] = \mu$.
$E[\sigma] = \sigma$

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

$E[X] = \mu$. Known property.

$$= \frac{\mu - \mu}{\sigma}$$

$$= \frac{0}{\sigma} = 0.$$

Hence $E[Z] = 0$ ✓ Shown

② Show $\text{var}[Z] = 1$.

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

Substitute $Z = \frac{X - \mu}{\sigma}$

$$= \frac{\text{var}[X] - \text{var}[\mu]}{\text{var}[\sigma]}$$

$\text{var}[\mu] = 0$. Known

$$= \frac{\text{var}[X] - 0}{\text{var}[\sigma]}$$

Pull $\text{var}[\sigma]$ out $= \sigma^2$

$$= \frac{\text{var}[X]}{\sigma^2}$$

$\text{var}[X] = \sigma^2$ Known property

$$= \frac{\sigma^2}{\sigma^2} = 1.$$

Hence $\text{var}[Z] = 1$ ✓ Shown.

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)
```

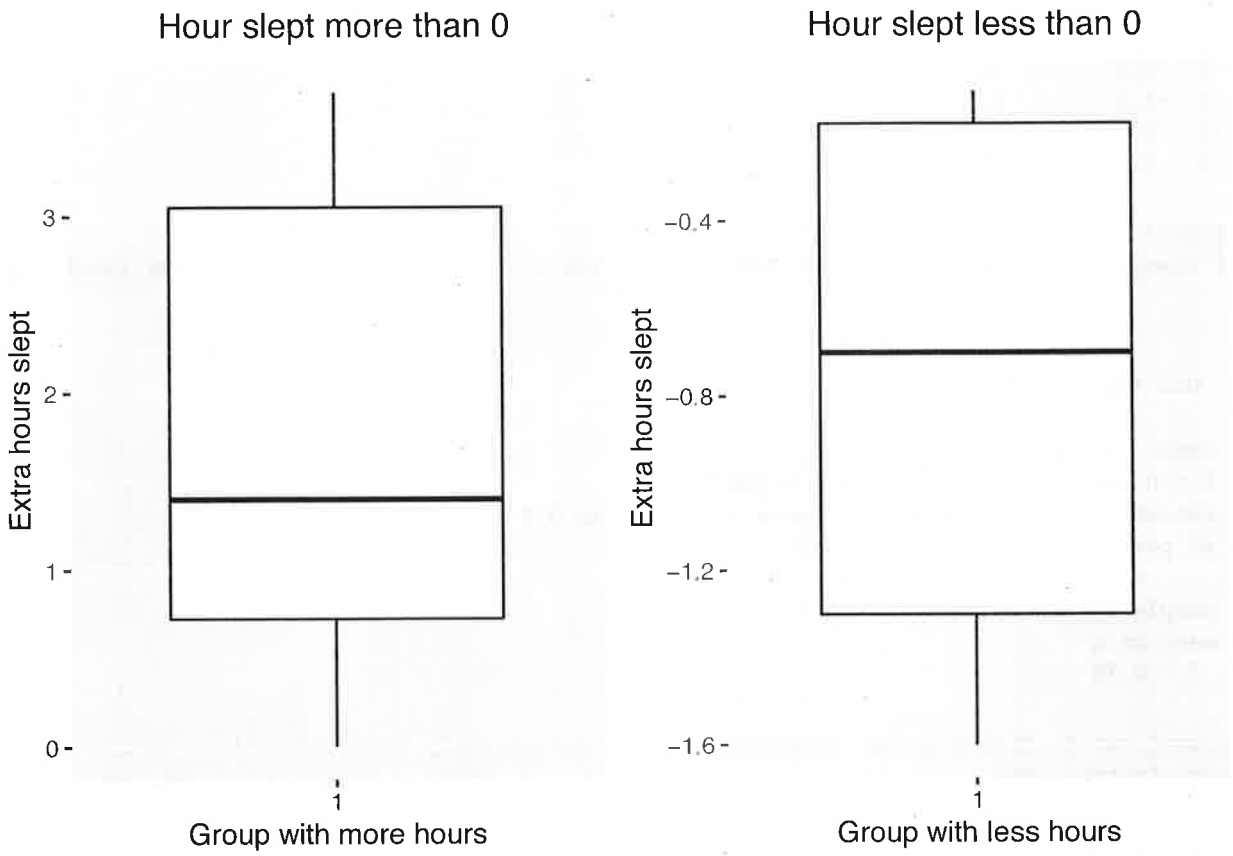
```
## extra1.cat
## less_hours more_hours
##          4          6
```

```
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)
```

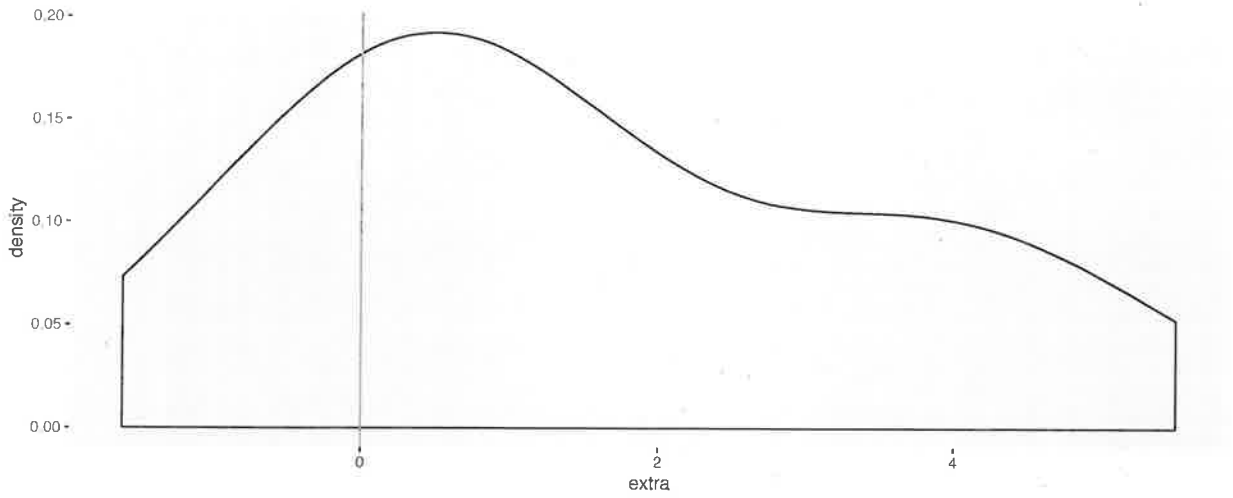


Note that the deliminators for r code are **not** apostraphes but are the tick marks found in the upper left hand corner of your keyboard. You will never need to print out an entire data set for me in your homework, just the fist few rows using the *head()* function is fine.

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
```

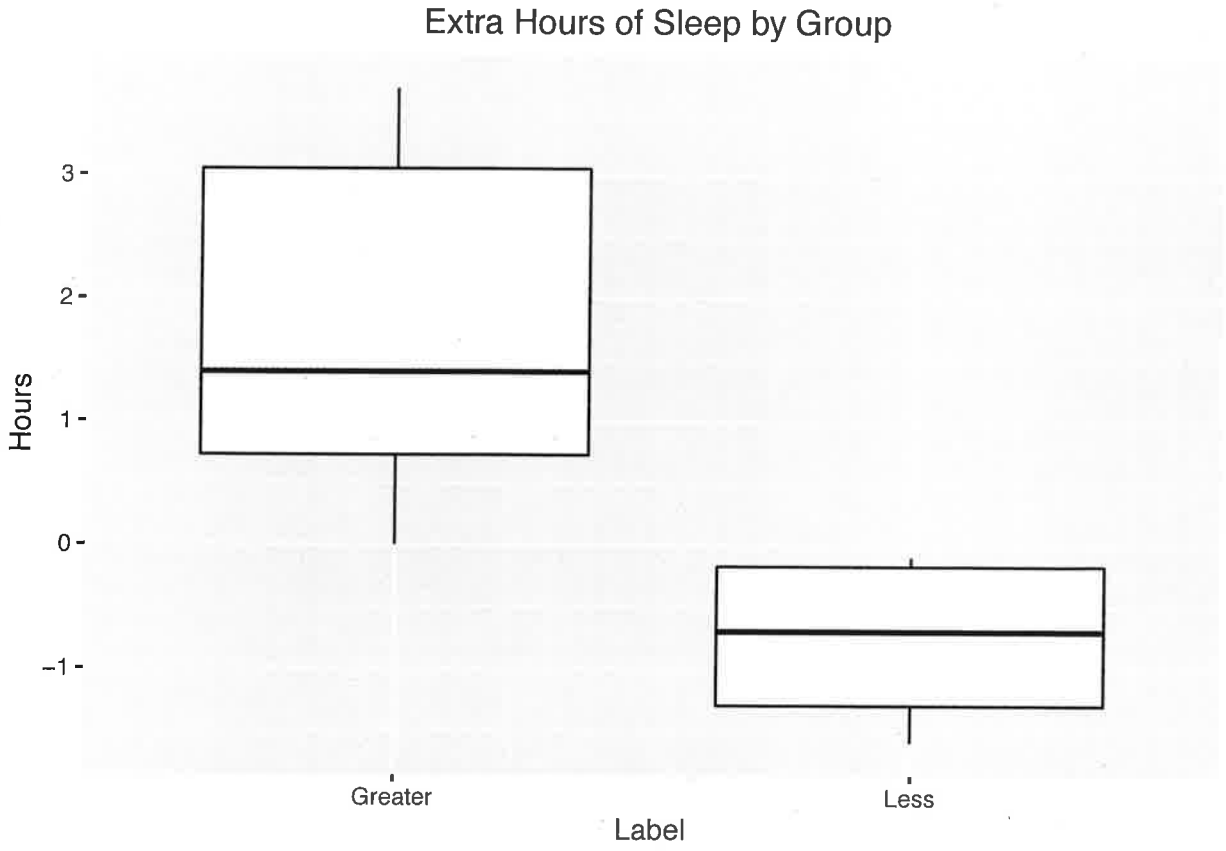
Part 3) Because the p-value (0.6655) is less than alpha (0.10), we fail to reject the null hypothesis in favor of the alternative hypothesis. The 90 percent confidence interval shows that the true mean is located between -Inf to 1.53 90 percent of the time.

```
#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)