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
```

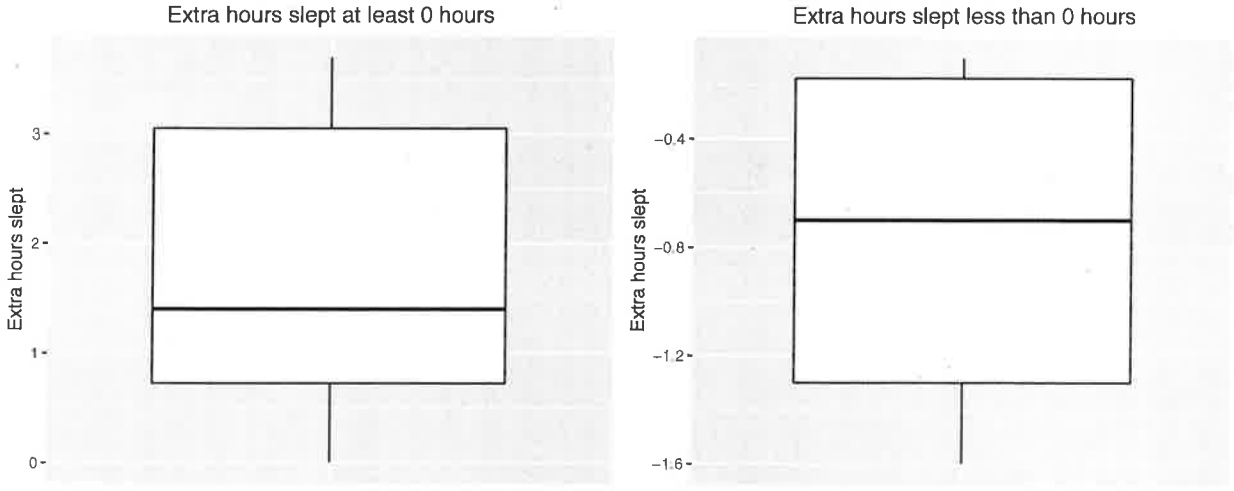
```
## alternative hypothesis: true mean is less than 0.5
## 90 percent confidence interval:
##      -Inf 1.532427
## sample estimates:
## mean of x
##      0.75

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.