STA130H1F – Fall 2021

Module 9 Problem Set - Sample Answers

S. Caetano and L. Bolton

Instructions

How do I hand in these problems for the 11:59 a.m. ET, December 2 deadline?

Your complete .Rmd file that you create for this problem set AND the resulting .pdf (i.e., the one you 'Knit to PDF' from your .Rmd file) must be uploaded into a Quercus assignment (link:https://q.utoronto.ca/courses/235890/assignments/742423) by 11:59 a.m. ET, on December 2. Late problem sets or problems submitted outside of Quercus (e.g., by email) are *not* accepted.

Problem set grading

There are two parts to your problem set. One is largely R-based with short written answers and the other is more focused on communication. We recommend you use a word processing software like Microsoft Word to check for grammar errors in your written work. Note: there can be issues copying from Word to R Markdown so it may be easier to write in this file first and then copy the text to Word. Then you can make any changes flagged in Word directly in this file.

Part 1

Question 1

```
library(tidyverse)
# If polite or rvest aren't't found, uncomment and run the following
# install.packages("rvest")
# install.packages("polite")
library(polite)
library(rvest)
```

The following code chunk sets up a session for scraping the U of T COIVD-19 dashboard.

```
# Save URL
url <- "https://www.utoronto.ca/utogether/covid19-dashboard"

# 'Bow' to the website
bow_to_dashboard <- bow(url,
    user_agent = "STA130 at U of T, contact sta130@utornto.ca if there are any issues")

# This gives us information about what we're allowed to do as well,
# as well as introducing us to the host
bow_to_dashboard
```

```
## <polite session> https://www.utoronto.ca/utogether/covid19-dashboard
##     User-agent: STA130 at U of T, contact sta130@utornto.ca if there are any issues
##     robots.txt: 68 rules are defined for 1 bots
##    Crawl delay: 10 sec
##   The path is scrapable for this user-agent
```

a) Write 1–2 sentences describing what we need to do to scrape this website ethically. Use your general knowledge from this module as well as making specific reference to the information above.

ANSWER

We should check that we are permitted to scrape this site. There is no specific Terms and Conditions page for the U of T site, but if we check the robots.txt and/or let the `polite` package check for us, we check if it is okay to scrape this site. "`The path is scrapable for this user-agent`" in the output from `bow()` shows us that there is not rule against scraping this site.
We should also see if there is an API to use instead (there doesn't appear to be one) and if there is a crawl delay (there is, 10 seconds for all users—this will be done automatically when we scrape based on the bow session we set up.).

```
# You're not required to understand the following code
# It scrapes the table of historical data on the dashboard
# The output is a tibble
uoft_covid <- scrape(bow_to_dashboard) %>%
  html_nodes("table") %>%
  html_table() %>%
```

```
  purrr::pluck(1) %>%
  rename(period = "Time period",
         cases = "Number of confirmed cases in our community",
         outbreaks = "Number of new outbreaks confirmed on our campuses")
```

```
### ANSWER ###
glimpse(uoft_covid)
```

b i) Use `glimpse()` and `head()` in the dataset `uoft_covid`.

```
## Rows: 88
## Columns: 3
## $ period    <chr> "November 13 to November 19", "November 6 to November 12", "~
## $ cases     <int> 6, 0, 1, 1, 3, 5, 4, 3, 4, 2, 5, 6, 4, 1, 1, 0, 0, 0, 0, 1, ~
## $ outbreaks <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ~
```

```
head(uoft_covid)
```

```
##                       period cases outbreaks
## 1 November 13 to November 19     6         0
## 2  November 6 to November 12     0         0
## 3   October 30 to November 5     1         0
## 4   October 23 to October 29     1         0
## 5   October 16 to October 22     3         0
## 6    October 9 to October 15     5         0
```

```
### ANSWER ###
```

b ii) How many rows and columns are in `uoft_covid`?

ANSWER
There are 88 rows and 3 columns.

b iii) Describe what an observation in this dataset.

ANSWER
One observation is one week.

b iv) Is this data set tidy? Yes or no. If no, explain why.

ANSWER
Yes, it is tidy.

c) Write code to find the total number of historical COVID cases in the U of T community since March 14, 2020 at U of T. This code should be able to be run on an updated scrape of the data without needing to be changed. It should output the number, but doesn't need to be saved or formatted.

```
### ANSWER ###
uoft_covid %>%
  summarise(total = sum(cases))
```

```
##   total
## 1   458
```

```
### ANSWER ###
```

```
fake_data <- read_csv("fake_raw.csv")
head(fake_data)
```

Suppose the table we scraped from the U of T site was made from raw data like the `fake_data` loaded below.

```
## # A tibble: 6 x 4
##   period                     gender            age   campus
##   <chr>                      <chr>             <chr> <chr>
## 1 November 13 to November 19 f                 19    UTSC
## 2 November 13 to November 19 prefer not to say 23    UTSG
## 3 November 13 to November 19 m                 21    UTSG
## 4 November 13 to November 19 f                 22    UTSG
## 5 November 13 to November 19 f                 19    UTSG
## 6 November 13 to November 19 f                 20    UTSG
```

Data dictionary for fake U of T COVID-19 data

Variable	Description
period	time period in which case was recorded, Saturday—Friday (1 week)
gender	Gender with levels: m, f, non-binary, not listed, prefer not to say
age	Age in years, with levels: 17 and younger, 18, 19, 20, 21, 22, 23, 24, 25 and older
campus	Which U of T campus the person belongs to, with levels: UTSG, UTSC, UTM (these are the University of Toronto St. George, Scarborough and Mississauga campuses, respectively)

d) The following two chunks are two versions of finding out how many U of T community members who reported having COVID shared the same age, gender and campus. One version reports a median of 12 for the variable `n` and the other a median of 5.

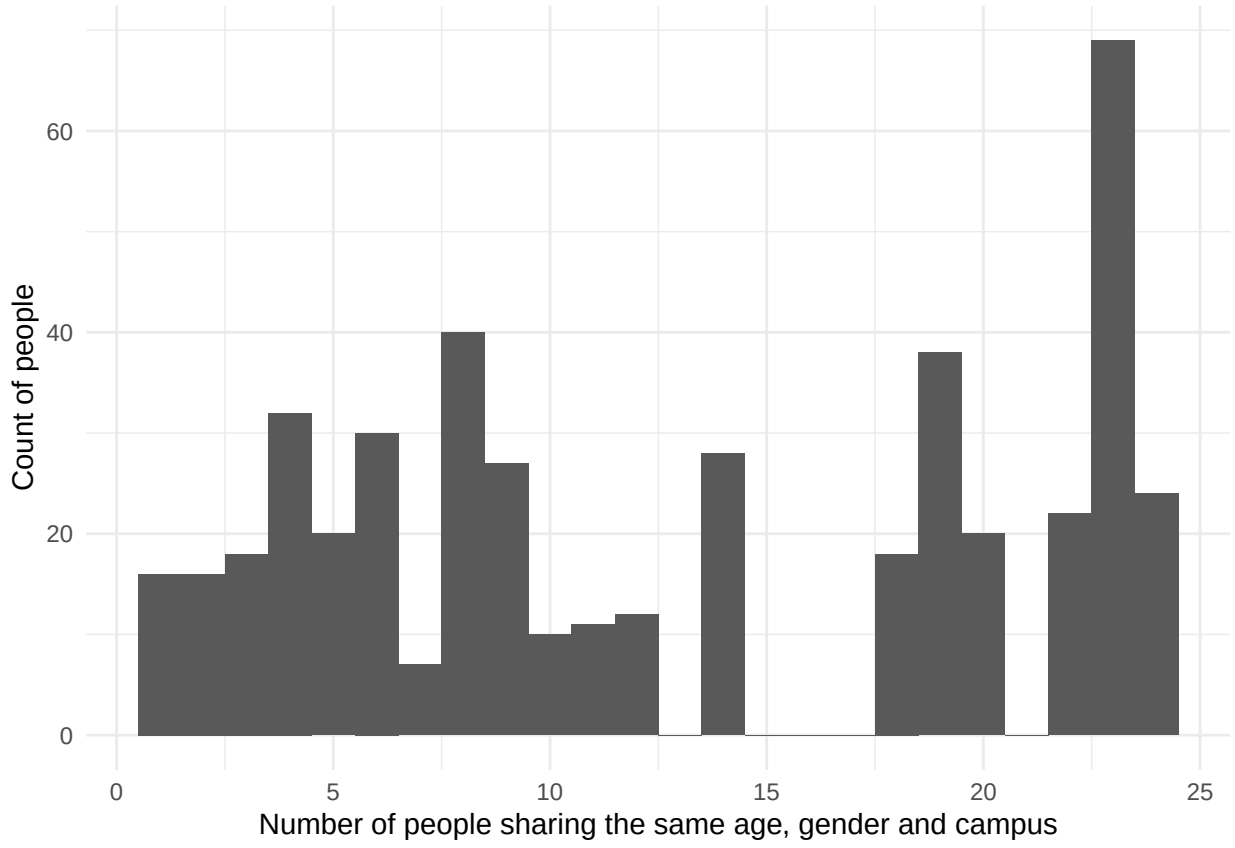
```
fake_counts_1 <- fake_data %>%
  group_by(gender, age, campus) %>%
```

```
  mutate(n = n()) %>%
  arrange(desc(n))

# You can uncomment
# View(fake_counts_1)

summary(fake_counts_1$n)
```

i) Create a histogram of `n` for each of `fake_counts_1` and `fake_counts_2`. Pick a sensible bin width or number of bins.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    6.00   12.00   13.04   21.50   24.00
```

```
## Add your histogram here

### ANSWER ###
fake_counts_1 %>%
  ggplot(aes(n)) +
  geom_histogram(binwidth = 1) + # lots of other sensible choices here, too
  theme_minimal() +
  labs(x = "Number of people sharing the same age, gender and campus",
       y =  "Count of people")
```

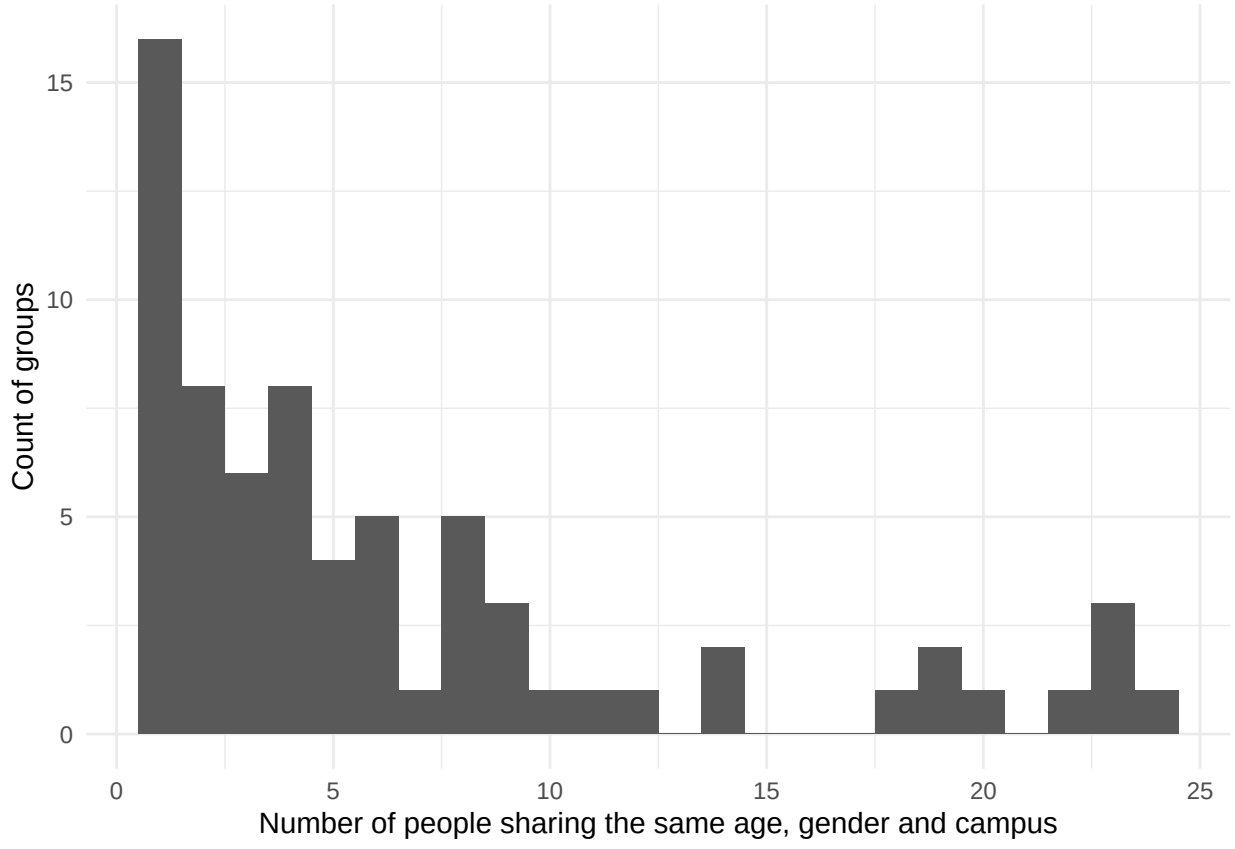


```
### ANSWER ###
```

```
fake_counts_2 <- fake_data %>%
  group_by(gender, age, campus) %>%
  summarize(n = n()) %>%
  arrange(desc(n))

# View(fake_counts_2)

summary(fake_counts_2$n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   4.000   6.543   8.000  24.000
```

```
### ANSWER ###
fake_counts_2 %>%
  ggplot(aes(n)) +
  geom_histogram(binwidth = 1) + # lots of other sensible choices here, too
  theme_minimal() +
  labs(x = "Number of people sharing the same age, gender and campus",
       y =  "Count of groups")
```




```
### ANSWER ###
```

ii) Explain why there is a difference, and which version you'd prefer to report. It may help for your to start by describing what each row represents (what is an 'observation' in each case).

ANSWER

In `fake_counts_1`, each row is a community member, and the `n` value represents how many other people (including them) have this combination of age, gender and campus.
In `fake_counts_2`, each row represents a combination of age, gender and campus and `n` represents how many community members had that particular combo.
The median is higher for `fake_counts_1,` because it is from the community member perspective. So, 50% of community members with COVID shared their identity combination with 12 or more other people, The median for `fake_counts_2` is lower because while in the first summary, when there are 25 people in a variable combination (e.g., f, UTSG, 20), that is observed 25 times, while in the this summary, when 25 people share the same identity characteristics, it is recorded one instance of a group of 25.
You could reasonably argue either way for which of these summaries is more informative. It will depend on the context. If I want to best communicate the experience of most people, I would prefer the approach for `fake_counts_1` as it tells me what the common experience is. If I am more worried about re-identification risk, I would look at `fake_counts_2` as it tells me there are quite a few small groups

f) Which members of the U of T community might be at risk for identification in this dataset? Please be thoughtful about the language you use and reach out on in office hours (or on Piazza) if you're not sure about how to write about age or gender, for example.

ANSWER

U of T community members that are in gender minorities (non-binary or not listed) might be at the most risk for re-identification as there may not be very many other people that share their characteristics in the full U of T community (not just this COVID group). For example, there may not be very many non-binary 19-year olds at UTM, so this could be very identifiying for them, while there are probably thousands of 19 year old women/feminine people at UTSG.
We may not have a good sense of why someone would prefer not to state their gender, so this does not seem likely to be as risky for re-id.

Question 2

(Note: Question 2 also provides some revision of topics learned in previous weeks, as you prepare to submit your project and the final assessment)

Lumosity is a brain training app thought to help cognitive skills - for example, memory, reasoning and focus. A large randomized trial was conducted to evaluate the impact of Lumosity training on cognitive skills. The study and results are presented in: Hardy, JL, Nelson, RA, Thomason, ME, Sternberg, DA, Katovich, K, Farzin, F, et al. (2015) "Enhancing Cognitive Abilities with Comprehensive Training: A Large, Online, Randomized, Active-Controlled Trial". *PLoS ONE* 10(9): e0134467. doi:10.1371/journal.pone.0134467.

Thousands of participants were recruited from Lumosity's free users (i.e., people who set up free Lumosity accounts but did not pay for full access) and randomly assigned to either:

- Lumosity training (Treatment) - complete Lumosity training online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks, or
- Crossword puzzles (Control) - complete crossword puzzles online for approximately 15 minutes at a time, at least 5 times a week for 10 weeks.

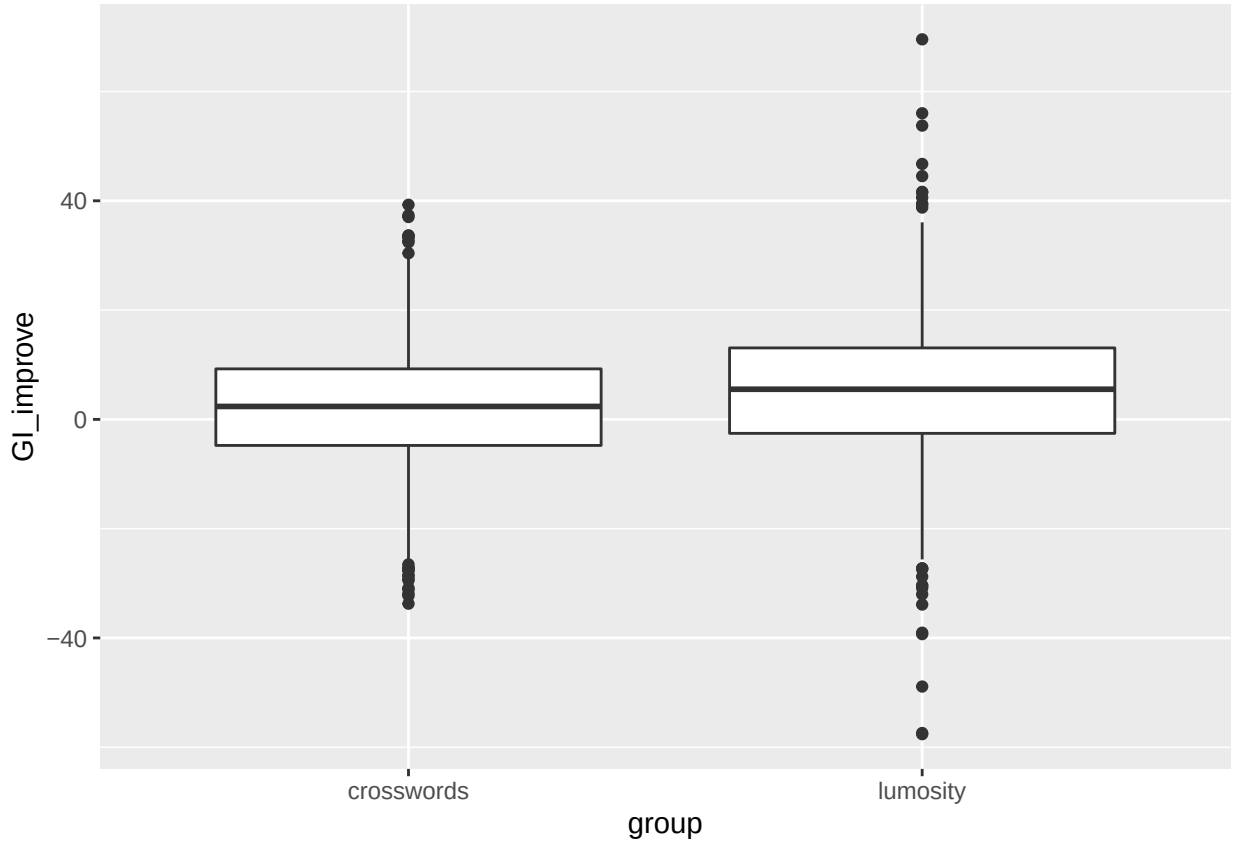
The main measure of cognitive skills was called the Grand Index (GI) Score; higher values mean better cognitive skills. The cognitive skills of the participants who completed the study were scored before and after the 10-week study period. We will store data on the improvement (i.e., after-before) in GI Scores (GI_improve) for the 5045 Lumosity users who participated in the study as well as several other the variables that may be useful are in a data frame called `study_dat`.

```
library(tidyverse)

study_dat<-read_csv("lumosity_study_data.csv")
glimpse(study_dat)
```

```
## Rows: 5,045
## Columns: 6
## $ participant_id     <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ~
## $ group              <chr> "crosswords", "lumosity", "crosswords", "crosswords~
## $ age_round          <dbl> 51, 24, 19, 21, 31, 25, 45, 27, 40, 50, 46, 52, 21,~
## $ GI_improve         <dbl> -5.992334, 9.030539, -15.030917, -9.265834, 3.18986~
## $ concentration_post <dbl> 1, 3, 5, 2, 4, 1, 4, 5, 5, 4, 4, 4, 5, 4, 4, 2, 4, ~
## $ active_days        <dbl> 56, 54, 6, 21, 43, 15, 0, 40, 39, 69, 64, 32, 21, 4~
```

Let's consider how Grand Index score improvements vary by type of online training:

```
ggplot(study_dat, aes(x=group, y=GI_improve)) + geom_boxplot()
```



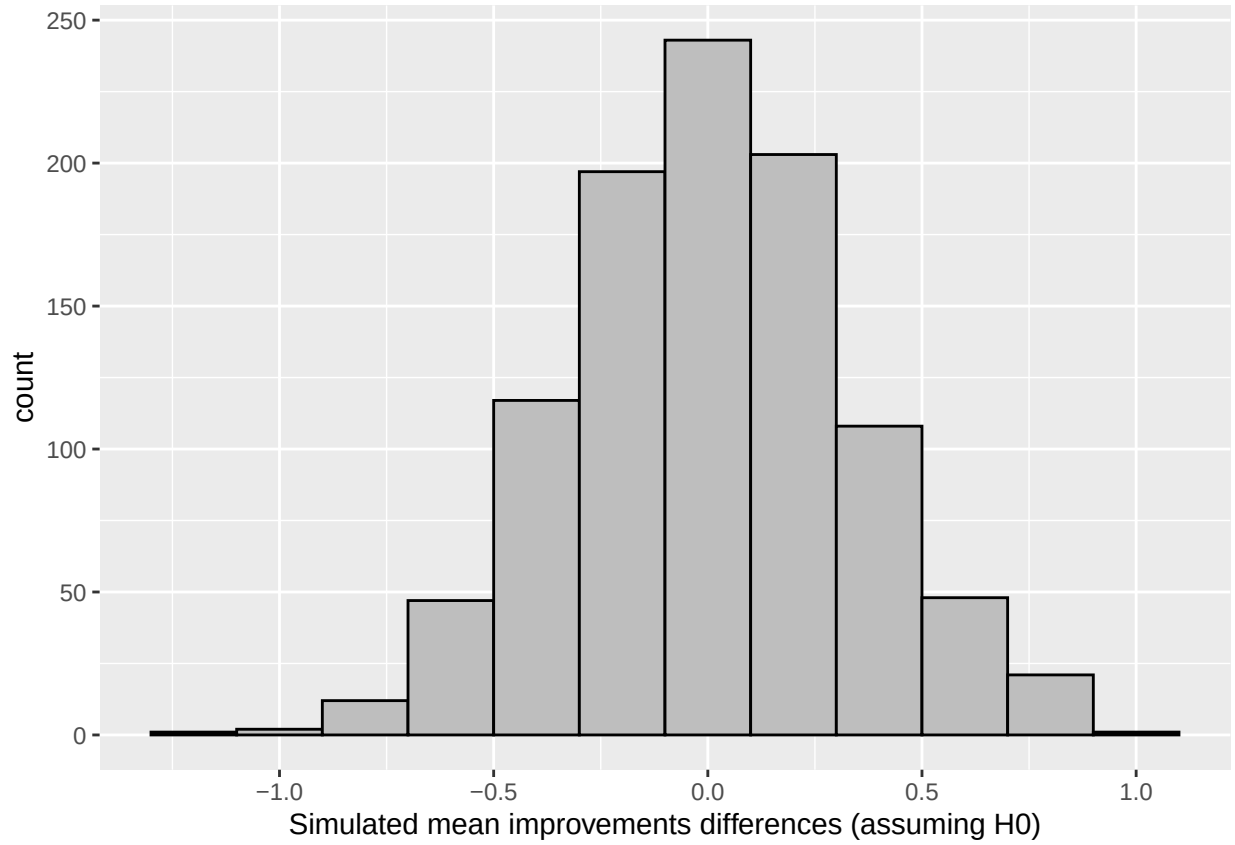

```
group_by(study_dat, group) %>%
  summarise(mean = mean(GI_improve),
            sd = sd(GI_improve),
            n = n())
```

```
## # A tibble: 2 x 4
##   group       mean    sd     n
##   <chr>      <dbl> <dbl> <int>
## 1 crosswords  2.14  10.6  2378
## 2 lumosity    5.24  12.0  2667
```

A hypothesis test on two groups can be conducted to compare mean Grand Index score improvements after online training with Lumosity and crosswords. (This might take a few moments to run.)

```
# compute test statistic
test_stat<-as.numeric(study_dat %>%
  group_by(group) %>%
  summarise(means = mean(GI_improve), .groups='drop') %>%
    #.groups='drop' is included to avoid a warning message being
    # printed, but doesn't change behaviour
  summarise(value = diff(means)))

# conduct randomization test
simulated_values <- rep(NA, 1000)
```

```
for (i in 1:1000) {
  sim <- study_dat %>% mutate(group = sample(group))
  sim_value <- sim %>%
    group_by(group) %>%
    summarise(means = mean(GI_improve), .groups='drop') %>%
    summarise(value = diff(means))
  simulated_values[i] <- as.numeric(sim_value)
}

sim <- tibble(mean_diff = simulated_values)

ggplot(sim, aes(x=mean_diff)) +
  geom_histogram(col="black",fill="gray", binwidth=0.2) +
  labs(x = "Simulated mean improvements differences (assuming H0)")
```




```
sim <- tibble(mean_diff = simulated_values)
sim %>%
  filter(mean_diff >= abs(test_stat) |
           mean_diff <= -1*abs(test_stat)) %>%
  summarise(p_value = n() / 1000)
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

The estimated p-value based on this randomization test is 0 so there is very strong evidence against the hypothesis that the mean Grand Index Score improvement is the same for those training using the Lumosity app and those completing online crossword puzzles.

a) Consider using a simple linear regression model instead to test for a difference in mean Grand Index score improvements for those who train using Lumosity and those who complete online crossword puzzles.

a i) Write down the appropriate regression model. Be sure to define any terms you use.

ANSWER

GIimprove$_i = \beta_0 + \beta_1 \text{group}_i + \epsilon_i$ Where: GIimprove_i=improvement in Grand Index (i.e., cognitive scores) from beginning
β_0 = average GI score improvement for reference group
β_1 = difference in GI score improvement between the group labelled as 1 and the reference group
ϵ_i = error term representing difference between observed GI score improvement for the ith selected user and the average sco

---

a ii) State the hypotheses to compare mean `GI_improve` when training using Lumosity and online crossword puzzles based on the model you specified in the previous part of this question.

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0$$

a iii) Use R to fit this model and interpret the estimated regression coefficients.

```
### ANSWER ###
mod1 <- lm(study_dat$GI_improve~study_dat$group)
summary(mod1)
```

```
##
## Call:
## lm(formula = study_dat$GI_improve ~ study_dat$group)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -62.848  -7.324   0.244   7.481  64.310
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                2.1416     0.2334   9.178   <2e-16 ***
## study_dat$grouplumosity    3.0972     0.3209   9.650   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.38 on 5043 degrees of freedom
## Multiple R-squared:  0.01813,    Adjusted R-squared:  0.01794
## F-statistic: 93.12 on 1 and 5043 DF,  p-value: < 2.2e-16
```

```
### ANSWER ###
```

ANSWER

$\hat{\beta}_0 = 2.1416$: This represents the estimated mean improvement in Grand Index score when completing online crossword puzzles.
$\hat{\beta}_1 = 3.0972$: This represents the estimated difference in mean improvement in Grand Index score after training with the Lumosity versus after completing online crossword puzzles.

a iv) We'll assume for the purposes of this question that the necessary assumptions for valid inference based on the tests conducted on the regression parameters by R are reasonable here.Interpret the p-value of this test to compare the mean improvement for Lumosity versus crossword puzzles. How does it compare to the p-value estimated using the randomization test earlier in this question? Is this surprising? Why or why not.

ANSWER

The p-value for the two-sided test of $H_0 : \beta_1 = 0$ is extremely small (i.e., <2e-16). There is very strong evidence against the hypothesis that mean improvement is the same when training is done using Lumosity versus crossword puzzles. The results suggest more improvement with Lumosity compared to crossword puzzles. This is the same conclusion as we would have made based on the randomization test, although the p-value obtained in that test was 0. It is not surprising that the estimated p-values from the randomization test and for the "treatment" parameter (β_1) of the linear regression model are similar, since they are both testing whether there is an association between Lumosity training (vs crossword puzzles) and improvement in Grand Index scores over the course of the study.

b) Consider the study design used by Hardy et al. (2015).

b i) Why type of study did Hardy et al. (2015) conduct? Use vocabulary from the course and justify your answer based on how the researchers did the study.

ANSWER

Hardy et al. (2015) conducted an experiment because they manipulated (i.e., assigned) a training type - Lumosity or crossword puzzles, then compared the GI score improvements across training types at the end of the study. They didn't simply observe/measure variables. They intentionally varied the explanatory variable (training type) to determine its affect on the response (GI score improvement).

b ii) Can we conclude that Lumosity training leads to more improvement in cognitive skills than completing crossword puzzles online based on these results? Explain your answer.

ANSWER

Since Hardy et al. (2015) conducted a large randomized experiment, confounding shouldn't be an issue (i.e., we should be able to conclude causation). The two user groups (Lumosity and crosswords) should look pretty similar to each other with respect to any potential confounding variables at the beginning of the study so that the main difference between them is which online activity they did during the study period. Therefore, it is reasonable to conclude that the difference we are seeing between the improvements in the Lumosity group and the crossword puzzles groups should be a result of their training type.

b iii) As reported in Hardy et al. (2015), 9919 participants consented to participate and were randomly assigned to a training type. However, only 5045 study participants actually completed the study. The dataset only included data on study participants who completed the study. How might this limit our conclusions?

ANSWER

Almost 50% of the participants did not complete the study. If the individuals who did (or did not) complete the study were different in some meaningful way between the Lumosity or crossword groups and that difference translates into different cognitive skills improvements on average, then there would be confounding - i.e., we wouldn't be able to tell whether the difference in mean GI score improvements was due to training type, or some other reason that is linked to not completing the study for one or both training types.

---

c) Perhaps age of the user is related to cognitive improvement as well.

c i) Do you think user ages would be different between the Lumosity group and crossword groups? Why or why not?

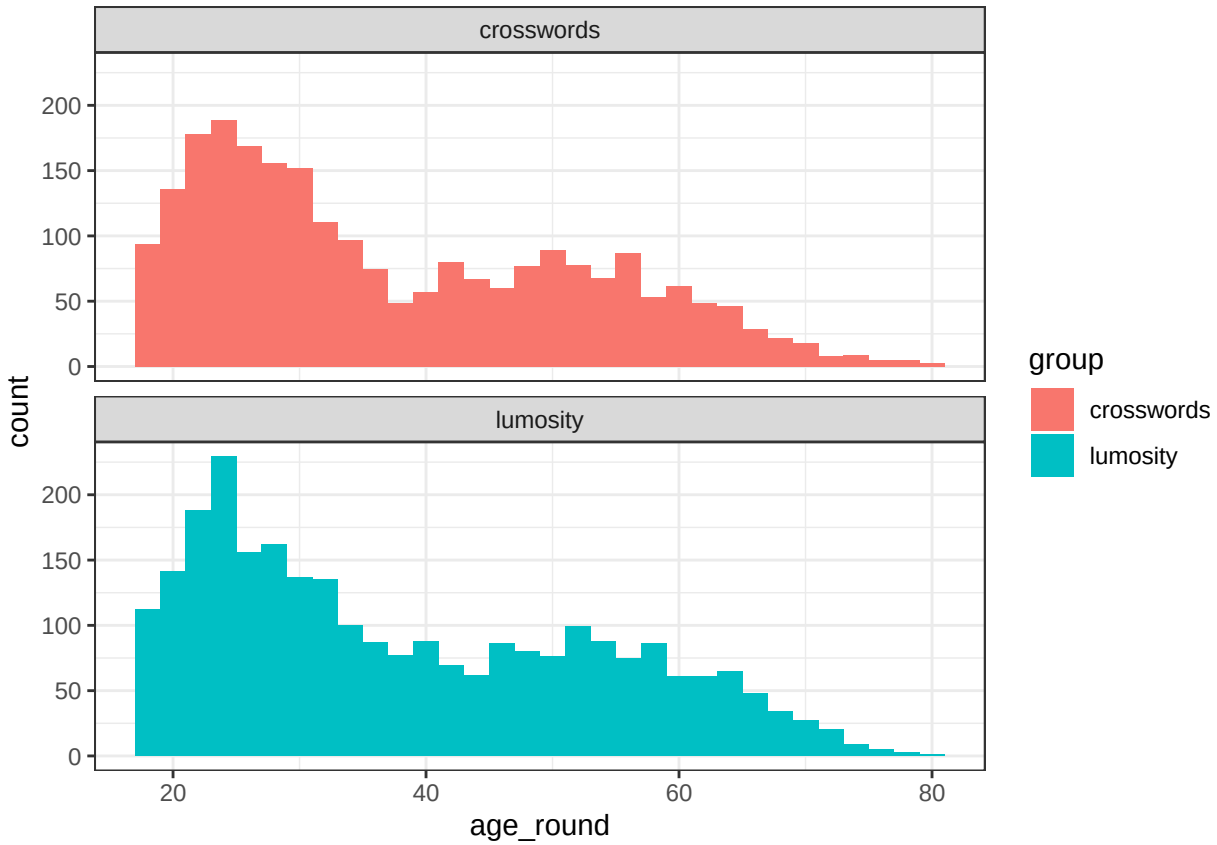
ANSWER

No. They should be similar. Since this study was a large randomized experiment the two user groups (Lumosity and crosswords) should look pretty similar to each other with respect to any potential confounding variables (like age). However, this assumes that everyone completes the study or if they don't, that people of certain ages are not more (or less) likely to complete the study if assigned to one training type rather than the another.

---

c ii) Produce an appropriate data summary to see if ages of the users differ for the Lumosity and crossword groups. Interpret your summary and comment on how this compares to your prediction about how ages would compare in *c(i)*.

```
### ANSWER ###
ggplot(study_dat, aes(x = age_round, fill = group)) +
  geom_histogram(binwidth=2) +
  facet_wrap( ~ group, ncol = 1) +
  theme_bw()
```




```
group_by(study_dat, group) %>%
  summarise(mean = mean(age_round),
            median = median(age_round),
            sd = sd(age_round),
            n = n())
```

```
## # A tibble: 2 x 5
##   group       mean median    sd     n
##   <chr>      <dbl>  <dbl> <dbl> <int>
## 1 crosswords  38.1     34  14.8  2378
## 2 lumosity    39.0     35  15.2  2667
```

```
### ANSWER ###
```

ANSWER

The distribution of ages of users look quite similar for the two groups (i.e. right skewed, perhaps bimodal with a high concentration of younger users in both groups, but a second smaller concentration of users between 45 and 55 years of age). On average, the Lumosity group was about 1 year older, though, and the standard deviation of ages in the Lumosity group was a little higher than the standard deviation of ages in the crosswords group (i.e., 15.22 years versus 14.8 years).
This seems generally consistent with the prediction in the last part of this question. However, since the groups are so large (both over 2000 users!), it's a little surprising that the mean ages aren't closer to each other and, therefore, the overall average age of the study participants considered together. Perhaps older individuals assigned to Lumosity training were more likely to complete the study than older individuals assigned to crosswords.

c iii) Suppose that there was a big difference in the age distributions of the two treatment groups - for example, suppose that younger users were much more likely to drop out of the Lumosity group than to drop out of the crossword puzzle group, and so the mean age of individuals who completed the study was 50 for the Lumosity group and 38 for the crossword puzzle group. How might this limit (if at all) the conclusions of the analysis?

ANSWER

We would have to recognize age as a potential confounding variable. If age is associated with GI_improve and varies across training groups, without including it in our model, we would not be able to separate out effects of training type and age on cognitive skill improvement.

(d) Hardy et al. (2015) included their ethics protocol with their paper. In that document, they described their study consent process: "Participants will give their informed consent prior to beginning the training study. They will read the informed consent form online and indicate their consent by clicking a radio button. . . .Participants will have unlimited time to read through the online form, and will have the opportunity to contact the researchers to ask questions before consenting to participate." (Hardy et al, 2015 ethics protocol, available at https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0134467.s007&type=supplementary)

Briefly explain why this process is necessary.

ANSWER

Research involving humans needs to abide by ethical principles (in Canada, the TCPS). Ethics protocols need to be submitted and evaluated by an Institutional Research Ethics Board (REB) before research can proceed in order to protect participants from risk and/or harm. An essential feature of all research involving humans is that participants must provide free and informed consent to participate (i.e., that they know exactly what participation in the study involves before they consent and they are not pressured to consent). Hardy et al. (2015) explained how information on the study was shared with potential participants and they were given enough time and invited to contact the researchers with any questions. Consenting to the study was completely voluntary and they just had to click on a button to consent to participate.

Part 2 (oral submission)

Watch the following short video on the Tuskegee syphilis study (link: [https://www.youtube.com/watch?v=afwK2CVpc9E&feature=share&fbclid=IwAR3r8TMnlNGIObYtp7NWg-krrBFMe6Ui9k4zqbcsGflh8g2Jxv_ymlMFkk8]). If you are not able to access the video using the first link, a version has been uploaded to MyMedia **here** (link:[https://mymedia.library.utoronto.ca/play/a80e488a0d9bd09193c9833336515ca5]). If possible, please watch the original video on Youtube to support the creators of this content. When watching the video, consider what are the main ethical concerns of the Tuskeege syphillis trial.

Your assignment is as follows: Identify at least 2 ethical concerns and describe them. Then, explain how you could conduct a similar trial today while avoiding some of the same ethical pitfalls that you identified from the original study.

Deliverable: Submit a video OR voice clip (aim for 4–5 minutes, with 5 minutes a hard max in either format) of you answering the above prompt. You should include a LINK (i.e., url) that the TA can click into to view/listen to the video/clip.

URL for my submission:

You can delete this the below once you've read it. Somethings to keep in mind:

- Do not feel the need to do tons of 'takes'. Rather, you can repeat yourself if you make a mistake, or feel you are unclear. This is not meant to be an additional burden, but rather to provide you with the opportunity to practice your oral communication skills and get a break from writing.
- You might be wondering how can I record this? One way to do this would be to schedule a Zoom meeting and record yourself in it. You can record the video to the cloud, or even directly on your computer! There will be many file types, including a video version, and one that is just a voice recording.
- You MUST upload a link (aka a URL) for your TA to watch your video. You can do this by uploading your video to mymedia, MS Stream, YouTube, etc. Alternatively, you can provide the zoom cloud link from your recording. ONLY links will be reviewed by the TA, we are not accepting mp4 or clip uploads.
- Please ensure that there is NO password protection on the video/link. The TA should be able to just "click the link" and "watch the video" (i.e., they should NOT need to type in a password). You can test this using an incognito window or new browser. There is more information about Zoom sharing settings here.
- If you are looking for more ideas of how to record yourself for this assignment or run into issues on how to upload your assignment, please post to Piazza.

Example

For example, include a line, such as the following, but change the link to your own video link:

URL for my submission: https://www.youtube.com/watch?v=dQw4w9WgXcQ.

This is an acceptable submission because:

- There is a link directly in your Rmd and pdf submission.
- The link is not a password/passcode needed to watch/listen to the video.