STA130H1F – FAll 2021

Module 4 Problem Set

YOUR NAME

Instructions

How do I hand in these problems for the 11:59 a.m. ET, October 7 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/705700) by 11:59 a.m. ET, on October 7. Late problem sets or problems submitted another way (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]

For questions 1 and 2 you'll be using the gratitude data introduced in Module 4.

Research question: Is there a difference between the **median** adjustment scores for the treatment and control groups?

Note: You can just write *median* for the population parameter and $\tilde{x}$ for the sample statistic.

(a) State your hypotheses

The below hypotheses are *not quite right*. Please **correct** them to be appropriate for this investigation.

$$H_0 : \tilde{x}_{\text{treatment}} - \tilde{x}_{\text{control}} = 0$$
$$H_A : median_{\text{treatment}} - median_{\text{control}} = 0$$

(b) Calculate your test statistic (real world)

There are three sub-steps for (b).

(b i) Load `tidyverse` and data (call it `gratitude`) The gratitude.csv should be available in the folder with this file.

```
# answer here
```

```
# answer here
```

(b ii) Save two objects, `n_control` and `n_treatment` that, respectively, store the number of observations in the control group and in the treatment group.

(b iii) Calculate the test statistic The following lines of code are OUT OF ORDER. Calculate the test statistic by putting them together correctly in the code chunk below (or by writing your own code for the calculation)

```
as.numeric()

group_by(treatment) %>%

summarize(median = median(adjustment)) %>%

summarize(test_stat = diff(median)) %>%

test_stat <- gratitude %>%

ungroup() %>%
```

(c) Simulate under the null hypothesis

(c i) Set values for simulation

- Set the seed to be the last three digits of YOUR student ID number. Your student ID number is the the one with no letters in it (i.e., it is *not* the one you use to log into your email or Quercus, which will have some letters from your name and may end in a number or numbers).
- Set the number of repetitions to be 1000.
- Set up a storage vector called `simulated_stats` to store our simulated statistics in later.

(c ii) Automate simulation with a `for` loop (simulation world)

IMPORTANT! In the first line of the chunk below, change `eval=FALSE` to `eval=TRUE`! If you don't, your code WILL NOT BE RUN WHEN YOU KNIT. If you *don't* do this and then ask a question about this on Piazza (where the answer is just you need to set `eval=TRUE`), you will have to tell your TA or Prof that they are the coolest statistician in the whole wide world! I'm sorry, I don't make the rules.^[Well, I guess I do... and this is a joke. But please read carefully.]

```
for (i in 1:repetitions){
  # this is were you create one simulation
  new_sim <- _____

  # you'll basically use whatever code you used for test_stat above,
  # but on the new_sim data/vector
  sim_val <- _____

  # store this stat in its own little slot in the storage vector you made above
  simulated_stats[i] <- sim_val
}
```

Turn your results into a data frame so we can use ggplot for plotting.

IMPORTANT! In the first line of the chunk below, change `eval=FALSE` to `eval=TRUE`.

```
sim_tibble <- tibble(simulated_statistics = simulated_stats)
```

(d) Evaluate the evidence against the null hypothesis

(d i) What is your hypothesized value?

Set it in the code chunk below AND change `eval=FALSE` to `eval=TRUE`.

```
hypothesized_value <- ____
```

(d ii) Plot

Visualize your simulated sampling distribution. Mark the two cutoffs that represent your test statistic (and the mirror/complement of your test statistic). Add an appropriate title and axes labels.

(d iii) Calculate your p-value

5. Make a conclusion

(5 i) Write 1–2 sentences discussing what we learn from this test.

(5 ii) If we were to set a statistical significance level of 0.05, which of the following would be true? Briefly (1–2 sentences), explain why.

A) We would fail to reject the null hypothesis.

B) We would fail to reject the alternative hypothesis.

C) We would reject the null hypothesis.

D) We would reject the alternative hypothesis.

(5 ii) Based on your answer to the above, if we were to use A) 0.05 and B) 0.10, as thresholds for statistical significance, briefly explain if your are at risk of making a Type I or Type II error and what that would mean in the context of this research question.

Answer for both of these thresholds. Is the answer the same for both? Different?

[Question 2]

Research question: Is there a difference in the proportion of students who successfully adjusted to university life between those who used a gratitude journal and those that did not?

(a) Create a new variable

To the `gratitude` dataset, add a new categorical variable called `adjust_status` that takes the value "adjusted" if a student scored above (greater than) 144 on their adjustment scale (`adjustment`), and "not adjusted" if they scored 144 or lower.

(b) Create a bar plot for `adjust_status`.

Create a bar plot of the new `adjust_status` variable you created. Write 1–2 sentences to describe the key finding in the plot. Add an appropriate title and axes labels.

(d) Conduct a hypothesis test so see if the proportion of adjusted students differs between the treatment and control groups

You don't have to break your code up into segments like in the examples, but you can if it helps you keep track of what you're doing.

(e) Write a conclusion, in context, based on your findings in (d)

[Question 3]

A Scottish woman noticed that her husband's scent changed. Six years later he was diagnosed with Parkinson's disease. His wife joined a Parkinson's charity and noticed that odour from other people. She mentioned

this to researchers who decided to test her abilities. They recruited 6 people with Parkinson's disease and 6 people without the disease. Each of the recruits wore a t-shirt for a day, and the woman was asked to smell the t-shirts (in random order) and determine which shirts were worn by someone with Parkinson's disease. She was correct for 11 of the 12 t-shirts! You can read about this here.

(a) Without conducting a simulation, describe what you would expect the sampling distribution of the proportion of correct guesses about the 12 shirts to look like if someone was just guessing.

(b) Carry out a test using simulation to determine if there is evidence that this woman has some ability to identify Parkinson's disease by smell, or if she was a lucky guesser.

Set the random number seed to the last three digits of your student number before carrying out your simulation. Use 10,000 repetitions. (This simulation is similar to the code in Question 1, but with many more simulated values of the test statistic under the null hypothesis. 10,000 is a lot of repetitions - more that is likely needed - but we'll do this many repetitions this time anyways.)

(c) Initially, the woman correctly identified all 6 people who had been diagnosed with Parkinson's but incorrectly identified one of the others as having Parkinson's. Eight months later he was was diagnosed with the disease. So the woman was actually correct 12 out of 12 times. Are you able to get the p-value for the test using the updated data (i.e., 12 correct instead of 11 correct), without running a new simulation? What would you change from your answer to (b)? What wouldn't you change?

(d) Which of the following statements is/are valid description(s) of the p-value you computed in (b):

A) The probability that the Scottish woman's high proportion of correctly identified of Parkinson's cases was just chance.

B) The probability that the Scottish woman's high proportion of correctly identified of Parkinson's cases was NOT just chance.

C) The probability of obtaining 11 correct 'sniffs' in a sample of 12 if someone was just guessing.

D) The probability of obtaining 11 correct 'sniffs' in a sample of 12 if someone is not just guessing (i.e., something more than just chance is acting).

Part 2 (Choice of oral or written submission)

For this week, you can complete the required task as a written assignment, OR you can submit an oral response (i.e. just sound, or sound and video). While it is optional *this time*, there will be two other problem set part #2s that you will be required to complete an oral response, and therefore it is a good idea to try and practice it now!

Suppose you were given the following dataset and told it was a survey conducted with a random sample of 12 student's in Prof Bolton's 4th year class on statistical consultation, communication, and collaboration (STA490). They were asked if they were currently living in Toronto (defined as Toronto, Etobicoke, Scarborough, York, North York and East York) or not and how long it took them to travel to the place they/their family usually purchases groceries, using whatever travel method they usually used to get there. **Student** is just an ID they were randomly assigned when the data was anonymized.

Student	Location	Travel time (mins)
113	Toronto	20
120	Not Toronto	17
499	Not Toronto	35
703	Toronto	7
190	Toronto	12
286	Toronto	45
203	Not Toronto	45
503	Toronto	15
154	Not Toronto	50
334	Toronto	25
417	Not Toronto	22
303	Toronto	20

In this writing activity you will practice writing instructions and demonstrate your understanding of simulation based hypothesis tests, the focus of Module 4.

Your task

Write instructions so that someone could perform ONE of the following hypothesis tests, using only the resources listed below. You do not have to use all the materials.

A) Are 75% of the students taking STA490 this semester living in Toronto?
B) Do students who live in Toronto have the same average travel time to their local supermarket as student who don't live in Toronto?

Resources

- A pack of 52 playing cards, half red, half blue
- A reusable shopping back
- A friend who loves doing mental arithmetic and is very good at it
- A printed sheet with all the travel times listed in a column, but nothing else
- A white board
- A whiteboard marker
- A white board eraser

Key things to include

- Briefly introduce which test you are choosing from above (A or B).
- Describe the hypotheses for your chosen test in plain words.
- Briefly describe why we would want to create a simulated sampling distribution.
- Provide instructions that would enable a physical simulation of a sampling distribution for your chosen test—**using only the resources listed above**. Your final instruction should be "And repeat steps X to Y 99 more times to get 100 total repetitions". Replace X and Y with whatever the relevant steps are.
- *Include **at least 2** vocabulary words from this module and explain/connect them with what you're doing in this physical simulation.*

Some things to keep in mind

- Try to not spend more than 20 minutes on the prompt.
- Aim for more than 200 but less than 350 words (if **written**) and no more than 4 minutes (if **oral**)
- Use full sentences (writing or speaking).
- Grammar is not the main focus of the assessment, but it is important that you communicate in a clear and professional manner (i.e., no slang).
- You do NOT need to have the resources (cards, etc.) above to do this task. There is no requirement to use them or actually carry out your instructions, just to create an appropriate instruction set.
- Be specific. A good principle when responding to a writing prompt in STA130 is to assume that your audience is not aware of the subject matter (or in this case has not read the prompt). You would want another first year university students to be able to follow your instructions.

For those who choose to do an oral submission

- If you choose to make a video or voice clip for the assignment, 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
- 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 provide a link/URL to your recording (e.g. through Zoom or from uploading to YouTube or MyMedia). Uploading/processing can be slow, so give yourself plenty of time to prepare this. **Paste the URL directly into your answers to this problem set.**
  - NOTE: Your link must NOT be password protected (this is often the default in Zoom sharing). You can test this by seeing if you can open the file in a new browser or in incognito mode. **Files that we cannot access to mark will receive a 0.**
- 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.

Vocabulary

- Statistical inference
- Population

- Random sample
- Sampling distribution
- Simulation
- Parameter
- Simulation statistic
- Test statistic
- P-value
- Type 1 and 2 errors (also written as Type I and II errors)
- Comparing two population means/medians/proportions
- One- and two-group hypothesis tests
- for() loops
- sample()
- diff()