STA130H1S – Winter 2021

Week 4 Problem Set

Yixing Xu

Instructions

How do I hand in these problems for the 11:59 a.m. ET, February 4 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/206597/assignments/544000) by 11:59 a.m. ET, on February 4. 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 writing. 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]

Approximately 22% of the general population use the social media platform Twitter. Suppose that the Department of Statistical Sciences (DoSS) is conducting a study to see if this percentage is the same among their undergraduate students (that is, all students in an undergraduate DoSS statistics program). This would help them to better promote social, academic and networking events to its students. Suppose 400 students in statistics programs are randomly selected and asked whether or not they use Twitter. Suppose that 103 of these 400 students respond that they use Twitter.

(a) What are appropriate null and alternative hypotheses to test the claim? Make sure you define the parameter in context.

The appropriate null hypothesis or H0, which is the hypothesis or quantity to be test, is the proportion of the general population who use the social media platform Twitter is 0.22.

The appropriate alternative hypothesis or Ha, which is the alternative hypothesis to the null hypothesis and is essentially just the hypothesis if the null hypothesis is incorrect, is the proportion of the general population who uses the social media platform Twitter is not 0.22.

(b) Use the `sample()` function to simulate the number of students who use Twitter in a random sample of 400 DoSS students, under the assumption that the prevalence of Twitter usage is the same among DoSS students as it is in the general population. How many Twitter users did you have in your simulated sample of 400 students? How does this simulated count compare to the results of the Twitter study (i.e., that 103 of the 400 students sampled use Twitter in the sample of real DOSS students)? How does it compare to the assumption that 22% of students use Twitter.

Note: the probabilities assigned to the values in the vector from which you're sampling using the `sample()` function are considered equal by default. For example, consider simulating flipping a fair coin 10 times:

```
sample(c("Head","Tail"),size=10,replace=TRUE)
```

```
##  [1] "Head" "Head" "Head" "Tail" "Tail" "Tail" "Head" "Tail" "Head" "Tail"
```

```
# will do the same thing as:
sample(c("Head","Tail"),size=10, prob=c(0.5, 0.5), replace=TRUE)
```

```
##  [1] "Tail" "Head" "Tail" "Head" "Head" "Tail" "Tail" "Tail" "Tail" "Head"
```

```
# Even though the exact counts of "Head" and "Tail" differ each time you
# run this code, if you simulate enough coin flips (by increasing
# the value of 'size', you'll get approximately the same proportion
# of "Head" and "Tail" outcomes)

# To modify the code to make Tails much more likely than Heads,
# we could change the probs:
sample(c("Head","Tail"),size=10,prob=c(0.1, 0.9), replace=TRUE)
```

```
##  [1] "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail" "Tail"
```

Set the random number seed to the last 2 digits of your student number (not your UTORID, your 9-10 digit student number) before carrying out your simulation. We set the seed so that the results won't change each time this code is run or knitted. If you don't do this, your interpretations and conclusions may not be relevant to the new run of the code (or when you knit your Rmd file!).

(c) Use R to estimate the sampling distribution of the test statistic under the assumption that the prevalence of Twitter usage among DoSS students matches that of the general population. Use 1000 repetitions and set the seed to the last 2 digits of your student number. Generate the plot of this estimated sampling distribution and describe the distribution in a few sentences.

Use `labs(x="your title here")` to give the x-axis a better name than the default. You can split your axis title across different lines to make it easier to read by adding `\n` OR typing an "enter"/carriage return.

(d) Use R to compute the p-value of this hypothesis test based on the sampling distribution that you estimated in part (c).

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

i. The probability that the proportion of DoSS students who use Twitter matches the general population.

ii. The probability that the proportion of DoSS students who use Twitter does not match the general population.

iii. The probability of obtaining a number of students who use Twitter in a sample of 400 students at least as extreme as the result in this study.

iv. The probability of obtaining a number of students who use Twitter in a sample of 400 students at least as extreme as the result in this study, if the prevelance of Twitter usage among all DoSS students matches the general population.

(f) Write a conclusion to this hypothesis test based on the p-value you computed in part (d).

[Question 2]

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 12 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 two 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 initial data (i.e., 11 correct instead of 12 correct), without running a new simulation? What would you change from your answer to (b)? What wouldn't you change?

[Question 3]

A 1920s tea party
British statistician Ronald Fisher was at a tea party in the 1920s. One of the other guests was algae scientist Dr Muriel Bristol, who refused a cup of tea from Fisher because he put milk in BEFORE pouring the tea in. Bristol was convinced she could taste the difference, and much preferred the taste of tea where the milk was poured in afterwards. Fisher didn't think that there could be a difference and proposed they tested this.

The test was set up as follows: 8 cups of tea were made, 4 with milk in first and 4 with tea in first.

The result: Bristol correctly identified whether the tea or milk was poured first for all 8 of the cups.

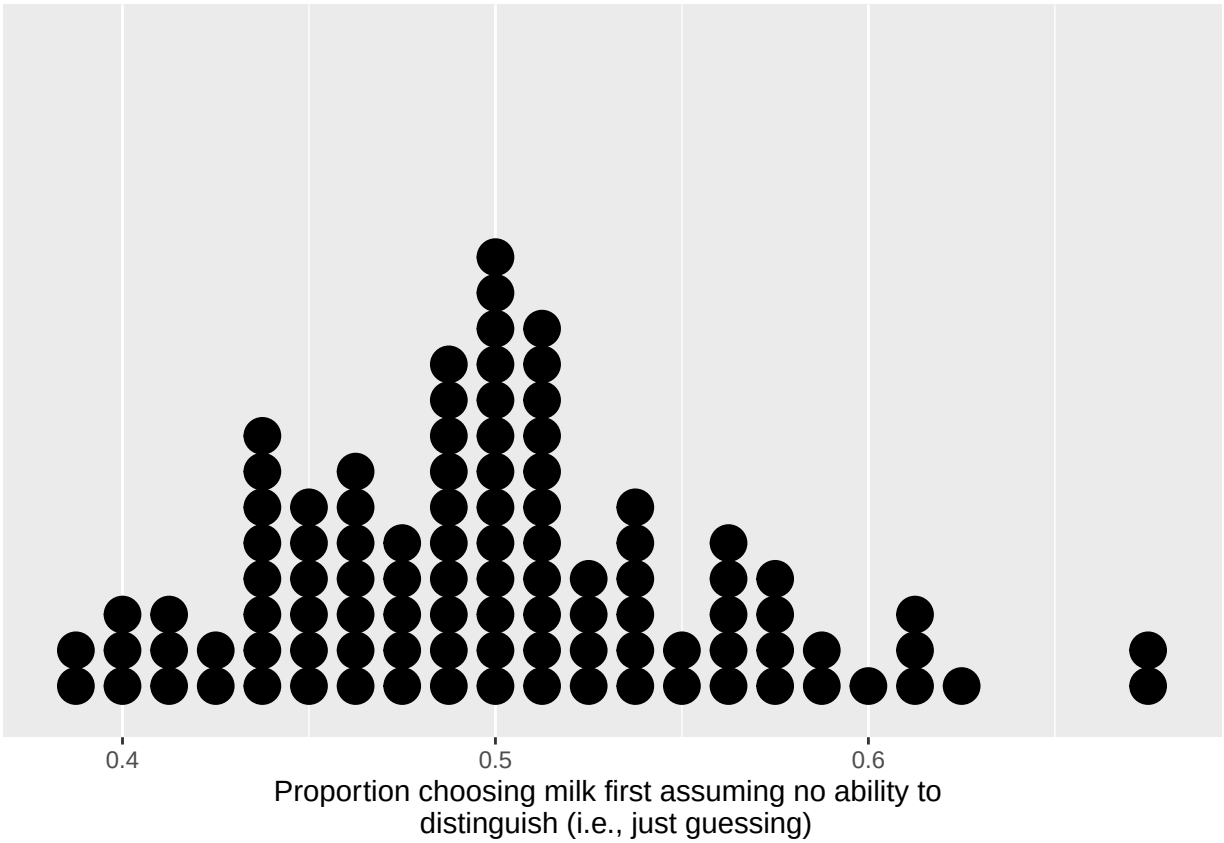
Fisher, being a good Statistician, wondered if this happened just by chance (Bristol was just guessing, 50/50), or whether it seemed more likely that Bristol was not guessing.

Your turn
Suppose you run an experiment like this with students in STA130. You get a random sample of 80 STA130 students to each taste one British-style cup of tea and tell you whether they think the milk or tea was poured first. 33 students correctly state which was poured first. Go through the steps to test whether students are just guessing or not.

(a) What are appropriate null and alternative hypothesis to test the claim? Make sure you define the parameter in context.

(b) Assume you conduct a test of significance using simulation and get the following estimated sampling distribution of the test statistic assuming the null hypothesis is true. For simplicity, this distribution shows the results of only **100** simulations. There are 100 dots on the plot, one for each simulation. (In practice, 100 simulations is not sufficient to obtain a good estimate of the sampling distribution.)

```
ggplot(sim, aes(p_correct)) +
  geom_dotplot() +
  labs(x="Proportion choosing milk first assuming no ability to \n distinguish (i.e., just guessing)") +
 scale_y_continuous(NULL, breaks = NULL)          # get rid of strange y-axis label
```



(i) What does each single dot in the plot represent?

(ii) Based on this plot, what is your estimate of the p-value?

(iii) What conclusion can you make based on the p-value you calculated in part b(ii)?

(iv) Suppose the analysis described in (b) is repeated but this time 1000 simulations are used to get a better estimate of the p-value, and the resulting p-value is 0.034. Do not conduct this simulation. What is an appropriate conclusion based on this p-value?

You may enjoy *this article* that provides more details about the tea party and the experiment. Optional.

Part 2

Your peers just finished participating in a new version of the British Tea experiment and had so much fun that they want you to come up with an entirely new experiment (it should NOT involve drinking tea)! In essence, your task is to come up with a hypothesis that you could test using a simulation test. You should do the following:

- Write the hypothesis out in plain words,
- Write the hypothesis using appropriate scientific/mathematical notation (e.g. Ho = 0),
- Prepare a small methods section that explains what the simulation test you will undertake to answer your question is doing for a lay audience,

- Include **at least 2** vocabulary words from this week and explain what they mean for a non-technical audience.

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.

- Use full sentences.

- At this point we have only learned how to do hypothesis test for one proportion. Keep that in mind with your designs.

- 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 or emojis should appear).

- 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).

Vocabulary

- statistical inference
- population
- random sample
- sampling distribution
- simulation
- parameter
- simulation statistic
- test statistic
- p-value