Homework 11

Insert Name

Math 141, Week 10

Due: 11:59pm, Friday April 15

Instructions

Work through the problems below and submit this document as a knitted .pdf to the Math 141 S22 Wells Lecture gradescope page.

For each problem, put your solution between the bars of red stars.

Acknowledgements

If you work with a classmate, please write a note acknowledging this.

Exercise 1

Recall that on Wednesday, April 6, we conducted a taste test to determine whether students can reliably distinguish between the lemon and lime flavors of LaCroix. The results of the test are contained in the data frame loaded using the following code:

The `correct` variable records whether each participant correctly identified the blue cup as containing the flavor which was different than the other two flavors.

The `preferences` variable identifies whether a participant indicated a preference for the blue cup (flavor = lemon), or the pink/green cup (flavor = lime). Many students indicated that they didn't enjoy either cup (usually accompanied by a comment like "They both taste nasty"), in which case their preference was coded as "none".

For this problem, we are viewing the survey participants as a random sample from the population of Reed students, and trying to estimate the proportion of students who would prefer to drink lemon Lacroix over lime.

a) Two tasks:

- When trying to estimate the Lacroix preferences of Reed students, explain why were are justified in only considering the responses from students who correctly identified the blue cup as different than the other two cups.
- Create a new data frame from the `lacroix` data set that only contains the survey responses from students who identified the correct cup, and contains a variable called `lemon` which takes the value `"yes"` for participants who preferred lemon, and `"no"` otherwise.

---

b) Out of the 29 students in our sample who identify the correct cup, find the proportion $\hat{p}$ of them who prefer lemon Lacroix.

c) Imagine we take repeated samples of size 29 from the population of Reed students who can identify the blue cup as being distinct from the others, and for each sample we calculate the value $\hat{p}$ of students who prefer lemon.

The Central Limit Theorem says that the distribution of $\hat{p}$ will be approximately Normal. Use our observed statistic from part (c) to estimate the mean and standard error for the distribution of $\hat{p}$.

d) Use theoretical tools[1] to find a 90% confidence interval for the proportion p of Reed students who prefer lemon, given that they are able identify the blue cup as containing the different flavor of LaCroix.

e) Use simulation to generate a 90% confidence interval for the proportion p. More specifically, use `infer` and your data frame from part (a), generate a bootstrap distribution for $\hat{p}$ and then use it to calculate a 90% confidence interval for p using the "percentile" method.

Exercise 2

This problem again uses the `lacroix` data set. We'll reload it here, just in case you made changes to it above.

Some students who participated in the test noted the temperature, both of the LaCroix and the weather outside, when discussing the differences between cups. Since the weather warmed over the course of the morning, the purpose of this problem is to investigate whether the time of the test played a role in the proportion of students who correctly identified the blue cup. So, we are interested in understanding two proportions:

- The proportion p_1 of Reed students who can correctly identify the blue cup during th 9am section.
- The proportion p_2 of Reed students who can correctly identify the blue cup during the 10am or 11am sections.

The `section` variable in the `lacroix` data frame tracks the section during which the participant completed the form, while the `correct` variable records whether a student's answer was correct. (i.e., did they correctly identify that the blue cup contained the different flavor of LaCroix).

Let's view our survey participants simple random sample of all Reed students who might complete this test during their designated section.

a) We have two samples: A 9am sample, and a 10 or 11am sample. How large are the samples?

b) Suppose we are interested in conducting a hypothesis test to investigate whether the time of the test makes a difference, by studying the difference in proportions $p_1 - p_2$. What are the null and alternative hypotheses?

[1] "Theoretical Tools" here means to use techiniques similar to what we did in class on Monday, April 11

c) Calculate the observed sample proportions $\hat{p_1}$ and $\hat{p_2}$, as well as the observed pooled proportion $\hat{p}$.

d) Are we justified in assuming that the null distribution for the sample statistic $\hat{p}_1 - \hat{p}_2$ is normally distributed? Explain whether we meet the two necessary conditions!

e) As you checked in part (d), we can expect that the null distribution for $\hat{p_1} - \hat{p_2}$ is approximately normal. Using theoretical tools, what are the mean and standard error for the null distribution?

f) Using the null distribution, what is the z-score for our observed statistic $\hat{p}_1 - \hat{p}_2$?

g) Use the `pnorm` function to determine the P-value for this hypothesis test. (Important: Be careful about whether you are conducting a one- or two-tailed test.)

h) Does there seem to be strong evidence that the time a student takes the test affects their ability to distinguish between the lemon and lime LaCroix? Why or why not?

i) Up to this point you calculated the P-Value for this test using theoretical techniques. Use a permutation test in `infer` to calculate the P-value for this hypothesis test using simulation.

j) In this problem, you've calculated a p-value both using theoretical techniques (calculating SE's and Z-scores and areas under normal curves), and also using simulation. Which do you prefer? Include a reason!

Exercise 3

A Kaiser Family Foundation poll surveyed 347 Democrats, 298 Republicans, and 617 Independents. Of those polled, 79% of Democrats, 55% of independents, and 24% of Republicans supported a generic "National Health Plan." Of course, we might expect these numbers to be different if we had conducted this poll a second time, just due to variation between samples.

a) Using theoretical techniques similar to those from class on Monday, April 11, find a 95% confidence interval for the difference of the proportion of all Democrats and all Independents ($p_D - p_I$) who support a National Health Plan.

b) True or false: If we had picked a random Democrat and a random Independent at the time of this poll, it is more likely that the Democrat would support the National Health Plan than the Independent. Use your work from part (a) to give evidence for this claim.

c. Suppose Kaiser Family Foundation would like to perform a new poll to estimate the proportion of independents who support a generic "National Health Plan", but this time, with a margin of error of only 0.01 for a 95% confidence interval. Calculate the minimum size necessary to obtain such margin of error (use the previous poll's proportion as your initial estimate for $\hat{p}$ in your calculation).