Lab 6 - Math 58b: error rates

your name here

due February 28, 2023

Lab Goals

By assessing our own comfort with randomness, we investigate 0.05. Additionally, we analyze data without a clear decision as to whether or not the null hypothesis should be rejected.

- why 0.05?
- what does a p-value really mean?
- how / why is there (is there not) always a single result for a research questions?

Getting started

Load packages

In this lab we will continue to use **infer** syntax and the `xpnorm()` function which is in the **mosaic** package.

Let's load the packages.

```
library(tidyverse)  # ggplot and %>%
library(mosaic)     # xpnorm and xqnorm
library(infer)      # simulation inference code
```

Preliminary personal level of significance Follow the link here and click through the scatterplot images: https://www.openintro.org/stat/why05.php?stat_book=os

(a) What is your personal level of significance?

(b) When you hear new information, do you consider yourself on the skeptical side or on the believing side? (There is no right answer!)

The data

(The data is only for Q2. The other questions do not require data.)

Researchers have conjectured that the use of the word "forbid" is more off-putting than the word "allow" (in affecting people's responses to survey questions). In particular, the suggestion is that people do not like to "forbid" anything. Students in an introductory statistics class were randomly assigned to answer one of the following questions:

- Should your college allow speeches on campus that might incite violence?
- Should your college forbid speeches on campus that might incite violence?

Of the 14 students who received the first question, 8 responded yes. Of the 15 students who received the second question, 13 said no. Think carefully about the response variable. It should *not* be coded as "yes" and "no" as answered on the questionnaire.

What do the data look like? Using the information above, create a 2x2 table (with your pencil) describing the dataset. What are your two variables? How many people are in each group? Again, do not code with "yes" and "no", use descriptive words.

Hypothesis test In order to formally test the researchers' conjecture (that the words can be off-putting), a null sampling distribution must be created (in order to compare the observed data against). You will create the null sampling distribution in two ways:

1. Using the **infer** syntax.
2. Using the Central Limit Theorem

   The Central Limit Theorem: the sample average (e.g., sample mean, sample proportion, etc.) in repeated random samples taken from a population will be normally distributed if the sample size is large enough.

Note that for the two proportion case, if the sample sizes are large enough the CLT says:

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right)$$

It isn't a hard and fast rule, but generally the CLT approximation is good if there are at least 10 (or possibly 5) successes and at least 10 (possibly 5) failures in each group.

To Turn In

Q1. Learning Community Q Describe one thing you learned from someone in your learning community this week (it could be: content, logistical help, background material, R information, etc.) 1-3 sentences.

Q2. Speeches and Violence In this first question, you will analyze the data described above using some old ideas and some new ideas.

To create a dataset, use the `rep()` function to replicate an appropriate number of times. Use the `c()` function to create a column. You should be able to do this on your own. For now, fill in the blanks (and then change to `eval = TRUE`).

```
# first create a data frame with the survey data
decision <- data.frame( ___ = c(rep("___", 14), rep("___", 15)),
                        ___  = c(rep("___", 8), rep("___", 6),
                                 rep("___", 13), rep("___", 2)))

table(decison)
```

a. Plot the observed data using `geom_bar()` and use `fill = response` to fill the bars in with appropriate colors, where the word `response` represents whatever you called the variable representing how the students responded to the survey.

b. Use **infer** to analyze the data. Report the one-sided p-value (you will report the conclusion in words below in part d.).

c. Use the following formula to create a Z-score for the same test (different p-value calculation) as was done with **infer**. Use R as a calculator to find the relevant Z-score, and find the one-sided p-value (you will report the conclusion in words below in part d.).

$$\text{Z score} = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

d. Give a complete conclusion to the data analysis / hypothesis test (that is, conclude what you think is most appropriate). State the null and alternative hypotheses, provide what you believe is the most accurate significance result (compare parts b. and c. above), and give a sense of to whom (what population, if any) the results can be applied.

Q3. p-values Read the ASA's statement on p-values: http://www.tandfonline.com/doi/pdf/10.1080/00031305.2016.1154108

Choose **two different** principles (pg 131-132) and explain each (separately) as if to a peer in a science class who is making conclusions about a recent study. Explain in your own words.

Q4. p-values, take two Why use p-values at all? That is, what is benefit of having a p-value (as opposed to simply descriptive statistics or graphs of the data)?

If you are still curious about the ideas in this lab (not part of the assignment):

None of the queries below are part of the lab. I offer them here for people who are intrigued by the ideas we've covered and want to know more. Indeed, the article linked below (which has been cited 3000+ times and viewed almost 3 million times) has a provocative title (and is incredibly well written).

Q5. Read Ioannidis (2005), "Why Most Published Research Findings are False" http://www.plosmedicine.org/article/fetchObject.action?uri=info%3Adoi%2F10.1371%2Fjournal.pmed.0020124&representation=PDF

(a) Consider table 1. Suppose that the level of significance is taken to be 0.05 and the power is 0.8. Also, set R (the number of true to not true relationships) to be 2 (for every 3 experiments, one is null). What percent of research findings (i.e., "significant" findings) are actually true (i.e., Ha is true)? [hint: for ease of calculation, you can set c to be something like 10,000.]

Solution:

Finding	True Yes	True No	Total
Yes	5333	167	5500
No	1334	3166	4500
Total	6667	3333	10000

Out of the 5500 significant findings, it turns out the 5333 of them are true. Therefore, the proportion of significant findings is $5333/5500 = 0.9696364$.

(b) Consider table 1. Suppose that the level of significance is taken to be 0.05 and the power is 0.3. Also, set R (the number of true to not true relationships) to be 0.1 (for every 11 experiments, 10 are null). What percent of research findings (i.e., "significant" findings) are actually true (i.e., Ha is true)? [hint: for ease of calculation, you can set c to be something like 10,000.]

Solution:

Finding	True Yes	True No	Total
Yes	273	455	728
No	636	8636	9272
Total	909	9091	10000

Out of the 728 significant findings, it turns out the 5333 of them are true. Therefore, the proportion of significant findings is $273/728 = 0.375$.

We notice that the proportion of significant findings is *very* dependent on not only the power, but also on the proportion of experiments which are null to start with. The 6 important corollaries in the Ioannidis paper all follow from the tables (table 1 and others like it) which reflect on the state of testing beyond just controlling for type I errors using a level of significance of 0.05. The 6 corollaries contain important ideas to think about in doing science.

Q6. In the Dance of the p-values (https://www.youtube.com/watch?v=5OL1RqHrZQ8), what is the narrator arguing?

Solution:

The point being made in the Dance of the p-values is that the p-value gives virtually no information about the effect of interest. A confidence interval gives not only the magnitude of the effect but also the variability associated with the estimate. That is, the degree of uncertainty in the effect is captured by the width of the interval. Additionally, in replicates of the same experiment, the p-value can vary widely (from significant to non-significant) whereas a confidence interval captures the true parameter 95% of the time (albeit sometimes overlapping the null value, too).