Lab 3 - MATH 141

Due Date: Tuesday 10/26/2021, 11:59 PM

Instructions:

- Please provide complete answers/solutions for each question/problem.
- If it involves mathematical computations, please provide your reasoning and/or detailed solutions.
- Please use the `echo = TRUE` option in each R code snippet to show your code.
- **Please save your work as pdf file(s), don't put your name in any part of the document, and submit it to the Gradescope page for this course. Your document upload will correspond to your name automatically in Gradescope.**

Example R Code Snippets:

```
# loading packages
library(tidyverse)
library(openintro)
```

```
# plotting
numbers <- c(1, 3, 6, 4, 9)
plot(numbers, type="o", col="blue")
title(main="Numbers", col.main="red", font.main=4)
```

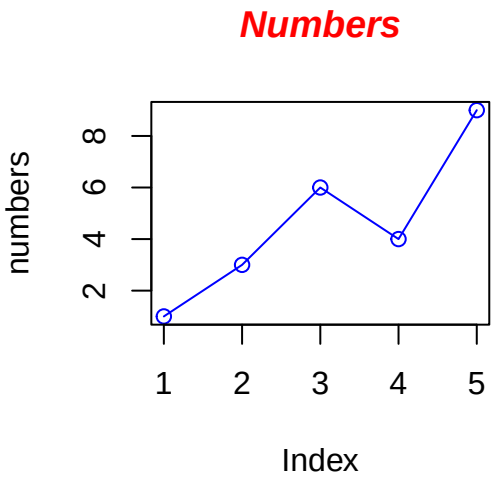

Figure 1: Figure caption

To reference a figure, use `\@ref(fig:#)` where the `#` indicates the snippet name. Fig. 1 is an example.

A. Hypothesis Testing with Randomization

I. Gender Discrimination

As the first step of any analysis, you should look at and summarize the data. Categorical variables are often summarized using proportions, and it is always important to understand the denominator of the proportion.

The discrimination study data are available as `gender_discrimination` using the `openintro` package.

1. Using the `count()` function, tabulate the variables gender and decision. Group the data by `gender`. Calculate the proportion of those who were and were not promoted in each gender and call this variable `prop_row`. Print out your results.
2. Using the `count()` function, tabulate the variables gender and decision. Group the data by `decision`. Calculate the proportion of those who males and females in each decision and call this variable `prop_col`. Print out your results.
3. From your calculations in the previous two points, what is the numerator and denominator for each?

II. Opportunity Cost

One-hundred and fifty students were recruited for the study, and each was given the following statement:

"Imagine that you have been saving some extra money on the side to make some purchases, and on your most recent visit to the video store you come across a special sale on a new video. This video is one with your favorite actor or actress, and your favorite type of movie (such as a comedy, drama, thriller, etc.). This particular video that you are considering is one you have been thinking about buying for a long time. It is available for a special sale price of $14.99. What would you do in this situation? Please circle one of the options below."

Half of the 150 students were randomized into a control group and were given the following two options:

a. Buy this entertaining video.

b. Not buy this entertaining video.

The remaining 75 students were placed in the treatment group, and they saw a slightly modified option (B):

a. Buy this entertaining video.

b. Not buy this entertaining video. Keep the $14.99 for other purchases.

Would the extra statement reminding students of an obvious fact impact the purchasing decision?

In this exercise, use the `opportunity_cost` data set, which is available through the `openintro` package. Answer the following questions.

1. How many rows and columns does this data set have? What are the variables? If you have categorical variables, state the levels. What type of study is this particular example?
2. Using the `count()` function, tabulate the variables group and decision. Group the data by `group`. Calculate the proportion of those who bought and not bought the video in each group and call this variable `prop_row`. Print out your results.
3. Suppose a success in this study is a student who chooses not to buy the video. Construct a point estimate for this difference as (T for treatment and C for control). State the null and alternative hypothesis.
4. Perform a randomization test (also known as the permutation test). Use at least 1000 shuffles and plot the distribution of the difference in proportions using a histogram. What can you conclude based on your results? Can we make causal statement in this study?

B. Confidence Intervals with Bootstrapping

I. Climate Change

From the Case Study - Climate Change (Proportion) section, use similar codes to answer the following questions. These problems requre a combination of what we have learned in lab since the first week (e.g. for loops, while loops, conditional statements etc).

For convenience, below is the R code used in this section.

```
# "the population" or "the data"
us_adults <- tibble(
  climate_change_affects = c(rep("Yes", 62000), rep("No", 38000))
)
```

1. Obtain a random sample of 60 from the `us_adults` data. Calculate the sample proportion, and use these to calculate and store the lower and upper bounds of the 95% confidence interval (use 1000 trials or resamples). Repeat these steps 100 times.
2. From problem 1, you should have 100 95% confidence intervals. Compute the number of intervals that contain the true population proportion. Remember that we assumed that the true population proportion is 0.62 (62%).
3. A 95% confidence interval gives us a region where, had we redone the same data, then 95% of the time, the true value p will be contained in the interval. What proportion of your confidence intervals include the true population proportion? Is this proportion exactly equal or close to the confidence level? Explain your reasoning.
4. Choose a different confidence level than 95%. Would you expect a confidence interval at this level to be wider or narrower than the confidence interval you calculated at the 95% confidence level? Explain your reasoning.
5. Find a confidence interval for the proportion of US Adults who think climate change is affecting their local community with a confidence level of your choosing (other than 95%) and interpret it.