Homework 7

Problem 1: Armfolding

A professor at an Australian university ran the following experiment with her students in a data science class. Everyone in the class stood up, and the professor asked everyone to fold their arms across their chest. Students then filled out an online survey with two pieces of information:

1. Did they fold their arms with the **left arm on top** or the **right arm on top**?
2. Were they **male or female**?

The professor wanted to know whether the way people fold their arms differs between males and females. According to the survey results, males were more likely to fold their arms with the left arm on top. But was this just random variation? Or does it reflect a consistent gender difference in the population?

The dataset `armfold.csv` includes:

- `LonR_fold`: a binary variable (1 = left arm on top, 0 = right arm on top)
- `Sex`: either `male` or `female`

Your task is to assess whether there's evidence of a difference in **population proportions** — that is, whether the true proportion of males who fold with left on top differs from the true proportion of females who do so. (By "population," let's assume we mean the population from which the undergraduate students at this university come from, not necessarily the entire human population.) You'll also explain and interpret the statistical concepts that underlie this analysis.

A. Load and examine the data. Report:

- The number of male and female students in the dataset.
- The sample proportion of males who folded their left arm on top.
- The sample proportion of females who folded their left arm on top.

B. What is the observed difference in proportions between the two groups (males minus females)?

C. Compute a **95% confidence interval** for the difference in proportions (males minus females). Report the result from R's built-in function, but also show your work by writing out:

- The formula for the **standard error** for the difference in proportions. *(Look this up in the textbook or slides, since you're not expected to memorize this.)*
- The values you plugged into the formula.
- The z* value you used and why.

Make sure the "hand-calculated" version and R's built-in function agree, up to minor rounding differences.

D. Interpret your confidence interval in context by completing the blanks in this sentence: "If we were to (blank 1), then we would expect that (blank 2).

E. In your own words, what does the **standard error** you calculated above represent? What is it measuring?

F. What does the term **sampling distribution** refer to in this context? Be specific about, what is varying from sample to sample, and what stays fixed.

G. What mathematical result or theorem justifies using a **normal distribution** to approximate the sampling distribution of the difference in sample proportions? Explain this result briefly in your own words.

H. Suppose your 95% confidence interval for the difference in proportions was [-0.01, 0.30]. Based on this, what would you say to someone who claims "there's no sex difference in arm folding"?

I. Imagine repeating this experiment many times with different random samples of university students. Would the confidence interval be different across samples? Why? What should be true about the collection of all those intervals?

Problem 2: Get out the vote

The data in `turnout.csv` contain information from a major party's voter database about a "get out the vote" campaign in advance of the 1998 midterm Congressional elections. The question of interest is whether receiving a "get out the vote" (GOTV) call from a volunteer in advance of the 1998 election increased the chances that someone actually voted that year. But from the standpoint of causal identification, the issue is that voters were not called randomly. Some voters were more likely to receive a GOTV call than others, and the recipients and non-recipients might differ in their underlying propensity to vote.

Each row in turnout.csv is about a single person. The variables relevant to our purposes are:

- `voted1998`: whether the person voted in the 1998 Congressional election. This is our outcome variable (1=yes, 0=no).
- `GOTV_call`: whether the person received a "get out the vote" call prior to the 1998 election (1=yes, 0=no). This is our treatment variable of interest.
- `voted1996`: whether the person voted in the 1996 Congressional election (1=yes, 0=no)
- `AGE`: the person's age in years
- `MAJORPTY`: whether the person is registered as a member of either one of the two major U.S. political parties (1=yes, 0=no)

Part A. How much more likely are GOTV call recipients to have voted in 1998? As a preliminary analysis, calculate the following quantities.

- The proportion of those receiving a GOTV call who voted in 1998.
- The sample proportion of those *not* receiving a GOTV call who voted in 1998.
- A large-sample 95% confidence interval for the **difference in these two proportions**: that is, the proportions of voting in 1998 (`voted1998==1`) for those who received a GOTV call versus those who didn't.

Part B. Consider the `voted1996`, `AGE`, and `MAJORPTY` variables. Provide evidence that at all three of these variables are **confounders** that prevent the difference you observed in Part A from representing the true causal effect of the GOTV call on the likelihood that a person voted in 1998. Confounders here would be factors that make someone more likely to receive a GOTV call *and* to have voted in 1998. Your evidence here can consist of any appropriate plot, table, or set of summary statistics, **together with an appropriate large-sample confidence interval.**

Part C. Now let's get a better estimate of the effect of the GOTV call on the likelihood that a person voted. Use matching to construct a data set with `GOTV_call` as our treatment variable, and with `voted1996`, `AGE`, and `MAJORPTY` as our "matching" or "balancing" variables. Use 5 control cases for each treated case in your matching (`ratio=5`). (Remember the `greenbuildings.R` walkthrough on matching from class before spring break.)

Provide evidence that your "matched" data set is, indeed, balanced with respect to the three confounders of `voted1996`, `AGE`, and `MAJORPTY`. (That is, show that these variables are no longer confounders for the matched data, by producing appropriate summary statistics and associated large-sample confidence intervals.) Then repeat your analysis from Part A, except using the matched data only. For this matched data set, calculate:

- The proportion of those receiving a GOTV call who voted in 1998.
- The sample proportion of those *not* receiving a GOTV call who voted in 1998.
- A large-sample 95% confidence interval for the **difference in these two proportions**: that is, the proportions of voting in 1998 (`voted1998==1`) for those who received a GOTV call versus those who didn't.

What do you conclude about the overall effect of the GOTV call on the likelihood of voting in the 1998 election?