Homework 7

Kavya Malgi

2025-04-07

Github Link: https://github.com/kavyamalgi/SDS315HW7

Problem 1: Armfolding

Part A: Load and examine the data.
- The number of male and female students in the dataset: **111 female** and **106 male** students

- The sample proportion of males who folded their left arm on top: **0.4716981.**

- The sample proportion of females who folded their left arm on top: **0.4234234.**

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

The observed difference in proportions between the two groups is **0.04827469.**

Part C: Compute a 95% confidence interval for the difference in proportions (males minus females).

- The formula for the standard error for the difference in proportions:

$$SE = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- The values you plugged into the formula:

- p_1: **0.4716981**
- p_2: **0.4234234**
- n_1: **106**
- n_2: **111**
- (1-p_1): 1 - 0.4716981=**0.5283019**
- (1-p_2): 1 - 0.4234234=**0.5765766**

The Standard Error for the Difference in Proportion: 0.06746

Formula for CI:

CI = Observed Diff +/- Z* x SE

Our z* number:

- **z* = 1.96**
  - It's 1.96 because we are doing a 95% CI, meaning we want to capture the middle 95% of the standard normal distribution (100% - 95% = 5%)

95% Confidence Interval: (-0.08393973, 0.1804891)

Part D: Interpret your confidence interval in context

If we were *to repeat this study many times using random samples of similar siz*es, then we would expect *that 95% of the resulting CIs would contain the true diff in props between males and females who fold their left arm on top of their right.*

Part E: what does the standard error you calculated above represent? What is it measuring?

The SE calculated represents the approximate variability in the diff between two sample proportions and it measures how much the observed diff in props varies from sample to sample and if we were to repeatedly draw random samples of the same size from the population.

Part F: What does the term sampling distribution refer to in this context?

In this context, the sampling distribution refers to the distribution of the differences in sample proportions one would get if they repeatedly took random samples of 106 males and 111 females from the same population and calculated the proportion of each group that folded their left arm on top.

More specifically, from sample to sample the sample proportions may change due to random sampling. **On the other han**d, the true population proportions of males and females who fold their left arm on top are fixed and the sample sizes stay the same.

Part G: What mathematical result or theorem justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions?

In this context, a theorem that justifies using a normal distribution to approx the sample dist of diff in sample props is **CLT.** When sample sizes are sufficiently large, sample dist of diff between two sample proportions will be approx normally distributed, regardless of the shape of the population distribution and both groups have at least 10 expected successes and 10 expected failures.

Part 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"?

Although we can't rule out the possibility that there's no difference, since 0 is in the interval, the data suggests that a meaningful diff is possible, almost as high as 30%. Therefore, we don't have strong enough evidence to confirm a gender difference, but we can't dismiss the possibility that one exists.

Part 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?

Yes, the CI would differ across samples because the sample proportions vary because of random sampling. However, if we repeated the experiment many times, 95% of the CI we construct would capture the true diff in props between males and females who fold their left arm on top.

Problem 2: Get out the vote

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: **0.6477733**
  - The sample proportion of those not receiving a GOTV call who voted in 1998: **0.4442449**
  - 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:
    - **Standard Error: 0.03077**
    - **CI: (0.1432104, 0.2638463)**

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.

For all three variables, I observed that they are associated with receiving a GOTV call and with voting in 1998. This is evident from the summary stats, the plots, and the **95% CI of (0.006443461, 0.107284916)** calculated for the differences. These relationships indicate that each variable is a confounder, and a failure to control for them would bias our estimate of the GOTV effect.

Part C: 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).

- The proportion of those receiving a GOTV call who voted in 1998: **0.6477733**
- The sample proportion of those not receiving a GOTV call who voted in 1998: **0.5692308**
- 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:

  We are 95% confident that if we were to take repeated samples, the true difference in proportions of voting in 1998 for those who received a GOTV call versus those who didn't is in between **0.01288268 and 0.14420234**.

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

The CI doesn't include 0 and the p-val is < 0.05, therefore it is safe to conclude that the GOTV call had a statistically significant effect on voter turnout.