SDS 315 Homework 7

Elizabeth 'Betsy' Sherhart UT EID: eas5778
Click here for link to GitHub repository

April 7, 2025

Problem 1: Armfolding

A. Load and examine the data

```
##
## Female   Male
##    111    106
```

```
## [1] 0.4716981
```

```
## [1] 0.4234234
```

There is 111 females students, 106 males students, meaning a total of 217 students in the data set. The proportion of males who folded their left arm on top was 0.472 and the proportion of females who folded their left arm on top was 0.423.

B.

```
## [1] 0.04827469
```

The observed difference in proportions between the groups is 0.048, which indicates in the sample about 5% more males folded their left on top.

Part C

The built in r function (prop.test) found the 95% confidence interval for for the difference in proportions (males minus females) is from -0.093 to 0.190.

The formula for standard error (SE) for difference of proportions is:

$$SE(p1 - p2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Where:

p1 = proportion of males with left arm on top in sample (0.4717)

n1 = number of males in sample (106)

p2 = proportion of females with left arm on top in sample (0.4234)

n2 = number of females in sample (111)

I used 1.96 for the z* value because that is the z* value for a 95% confidence level for a normal distribution.

Part D

If we were to take many random samples from this student population, we would expect the 95% confidence interval to contain the true difference in proportions of males and females who fold with the left arm on top 95% of the time.

Part E

The standard error (SE) is the standard deviation of the sampling distribution, which represents how much the sample differences could vary from the true population difference. It measures how much the sample difference in proportions would vary from sample to sample due to random sampling variability. It reflects the uncertainty in estimating the true population difference. A smaller standard error means more precise estimates and a narrower confidence interval.

Part F

In this context, the sampling distribution refers to the distribution of the difference in sample proportions (p1 - p2) we would observe if we repeatedly took random samples from the population.

Varies: The sample proportions (p1, p2) and their difference, as they change with each random sample. Fixed: The true population proportions and their difference, which remain constant across samples.

Part G

The Central Limit Theorem (CLT) justifies using the normal distribution to approximate the sampling distribution of the difference in sample proportions. The Central Limit Theorem (CLT) states that if we take enough random samples, the averages or proportions from those samples will form a normal distribution, even if the original data isn't normal. As long as the sample size is large enough (usually n >= 30) and there are at least 10 successes and 10 failures, we can use normal distribution tools to make estimates about the population, such as confidence intervals.

Part H

If the 95% confidence interval is [-0.01, 0.30], it means the difference could range from -1% (favoring females) to +30% (favoring males) for the sample. While the upper end of the interval suggests that males may fold their arms more often with the left arm on top, the inclusion of zero indicates we cannot rule out the possibility that there is no difference at all. Therefore, we don't have strong evidence of a significant difference between the sexes. So, you could say, "I agree, we can't confidently conclude there's a difference in arm folding between sexes based on this sample — the data is inconclusive."

Part I

Yes, the confidence intervals would differ across samples because each random sample varies slightly in its composition, leading to different estimates of the difference in proportions. This variability in the data affects the confidence intervals. However, if the experiment is repeated many times, the following would hold:

Coverage principle: About 95% of 95% confidence intervals would contain the true difference in proportions, reflecting the meaning of '95% confidence.' Spread: The intervals' widths would vary, with larger samples typically yielding narrower intervals.

In the long run, while individual intervals vary, we expect 95% of them to capture the true population value.

Problem 2: Get out the vote

Part A

The proportion of those receiving a GOTV call who voted in 1998: 0.648

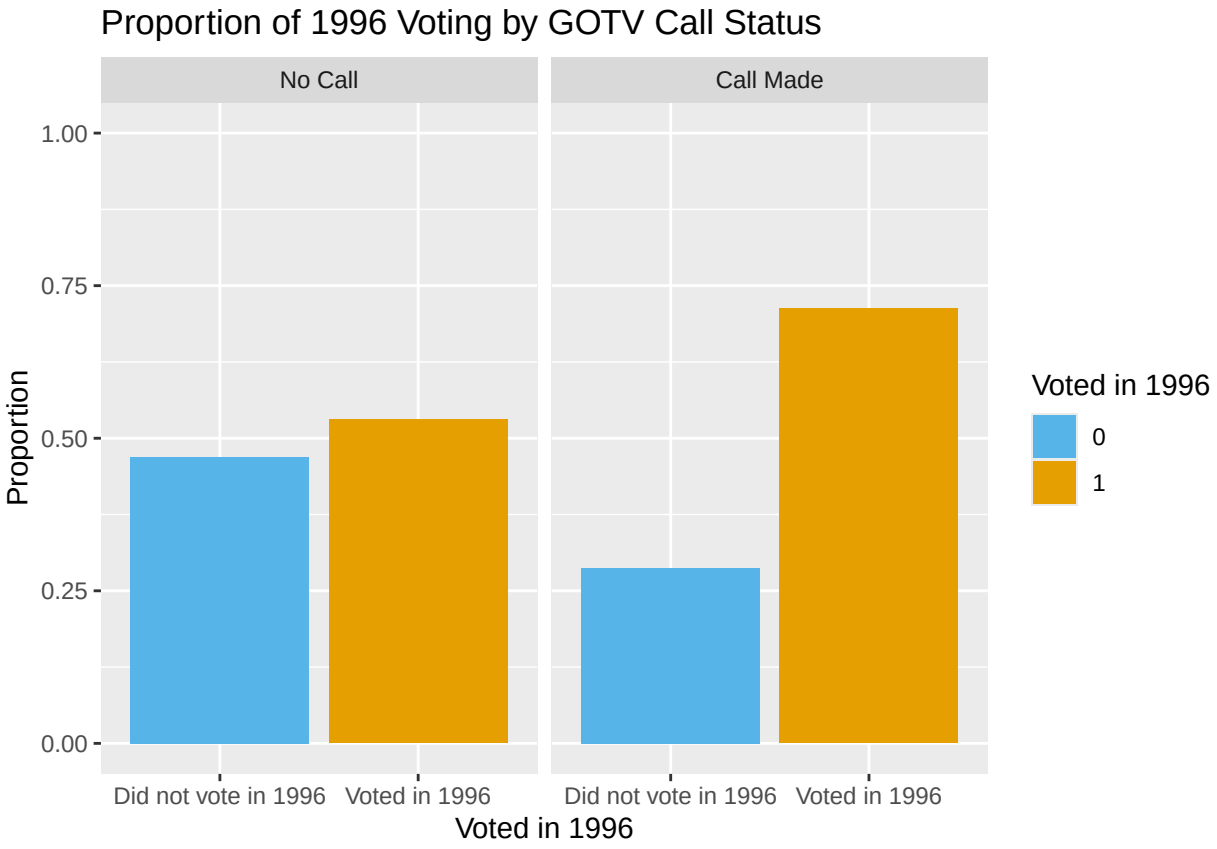
The sample proportion of those *not* receiving a GOTV call who voted in 1998: 0.444

The large-sample 95% confidence interval for the difference in the proportions of voting in 1998 (voted1998==1) for those who received a GOTV call versus those who didn't (received - didn't receive) was 0.141 to 0.266. This suggests a strong association between receiving a call and voting in 1998.

Part B

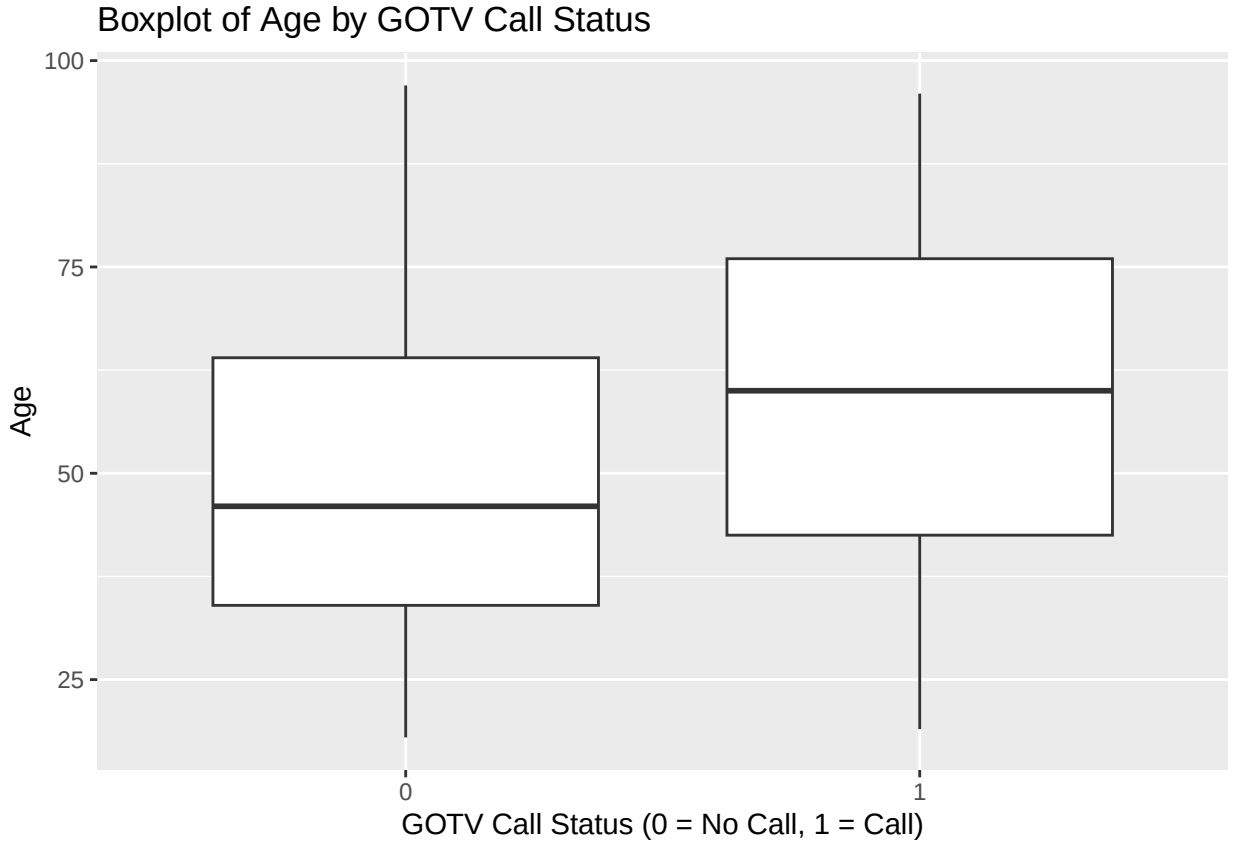
Since it was proven above that GOTV_call and voted1998 are related I'm just looking at the possible confounding variables relation to GOTV_call which then implies they are also related to voted1998.

voted1996



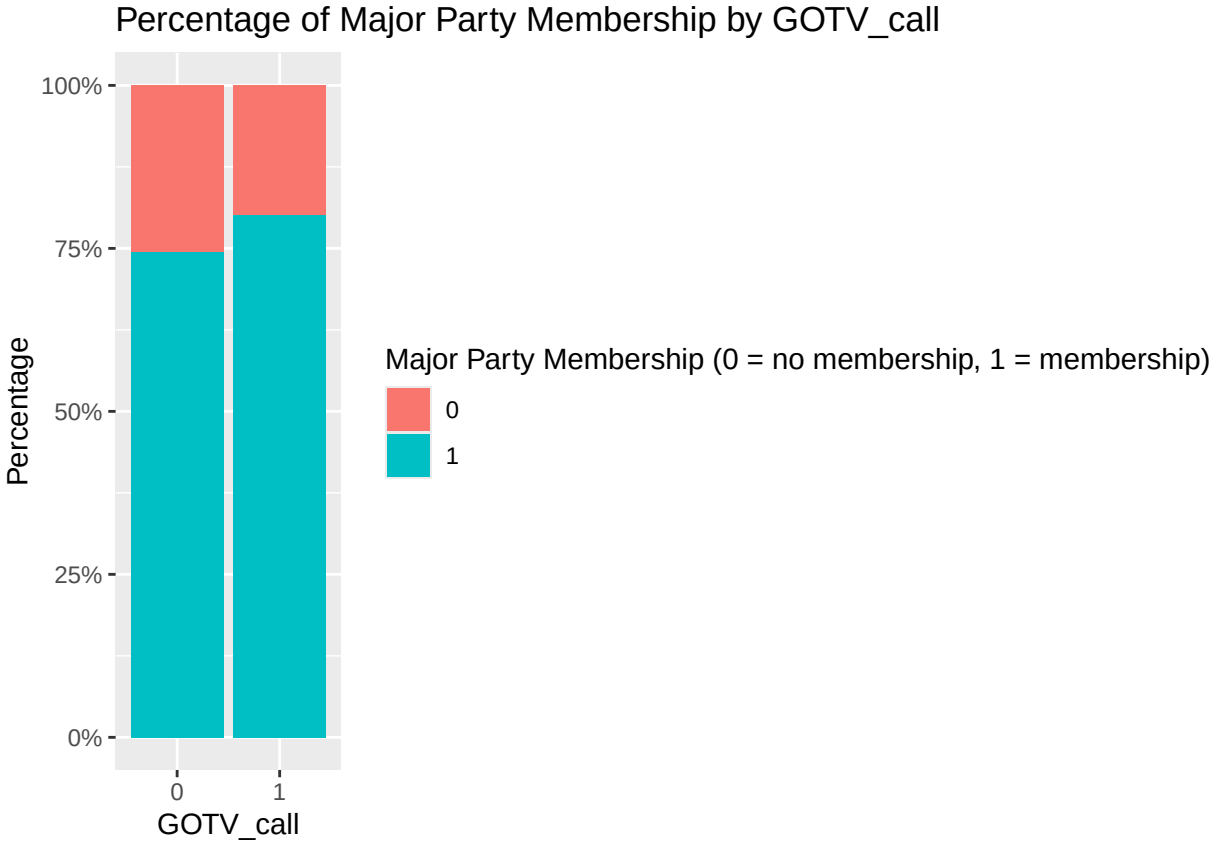
Whether a person voted in 1996 is a potential confounding variable in the relationship between GOTV_call and voting in 1998. Individuals who voted in 1996 are more likely to have been targeted for a GOTV call, and prior voting behavior also makes them more likely to vote again in 1998. This dual association suggests that voting in 1996 could confound the estimated effect of the GOTV intervention. Supporting this, the 95% confidence interval for the difference in proportions of receiving a GOTV call between those who voted and those who did not in 1996 ranges from 0.122 to 0.241, indicating a statistically significant difference.

AGE



A person's age is a potential confounding variable in the relationship between GOTV_call and voting in 1998. Older individuals were more likely to be targeted for a GOTV call, and age is also positively associated with the likelihood of voting. This dual association suggests that age may confound the estimated effect of the GOTV intervention. Supporting this, the 95% confidence interval for the difference in mean age between those who did not receive a GOTV call and those who did (did not - did receive) ranges from -11.4 to -6.4, indicating a statistically significant difference in age between the groups.

MAJORPTY



Whether a person is registered as a member to either major U.S. political party is a potential confounding variable in the relationship between GOTV_call and voting in 1998. Individuals who are members to a party are more likely to have been targeted for a GOTV call, and membership in a party makes them more likely to vote in 1998. This dual association suggests that membership in a major political party could confound the estimated effect of the GOTV intervention. Supporting this, the 95% confidence interval for the difference in proportions of receiving a GOTV call between those part of a major political party and those who are not part of a major political party ranges from 0.004 to 0.109, indicating a statistically significant difference.

Part C

After matching the data, the data set is indeed matched with the three confounders of voted1996, AGE, and MAJORPTY. The new summary statistic for voted1996 is a proportion of 0.713 for both groups (got call and did not get call) with a 95% confidence interval of -0.062 to 0.062 meaning it is centered on zero and hence statistically insignificant. The new summary statistic for AGE is a mean of 58.3 for both groups (got call and did not get call) with a 95% confidence interval of -2.76 to 2.68 meaning it is roughly centered on zero and hence statistically insignificant. The new summary statistic for MAJORPTY is a proportion of 0.8 for both groups (got call and did not get call) with a 95% confidence interval of -0.062 to 0.051 meaning it is roughly centered on zero and hence statistically insignificant.

For the matched data set the proportion of those receiving a GOTV call who voted in 1998 was 0.648, the sample proportion of those not receiving a GOTV call who voted in 1998 was 0.569 and the large-sample 95% confidence interval for the difference in these two proportions for those who received a GOTV call versus those who didn't is from 0.01 to 0.15.

Overall, it can be concluded that GOTV call had a positive effect on the likelihood of voting in the 1998 election. While the initial difference in turnout between those who received a call and those who did not was 20.4%, this difference was largely inflated by the confounding factors of voted1996, AGE, and MAJORPTY.

After matching on voted1996, AGE, and MAJORPTY, the estimated causal effect of the GOTV call was reduced to 7.9%, suggesting a smaller but statistically significant increase in voter turnout due to the GOTV call.