HW7Markdown

Ahantya Sharma

2025-04-02

UT EID: as236366

Github Link: https://github.com/Ahantya/SDS315/blob/main/HW7/Homework7Markdown.Rmd

Problem 1 - Arm Folding

A.

- There are 106 male students and 111 female students.
- The sample proportion of males who folded their left arm on top is 0.4717.
- The sample proportion of females who folded their left arm on top is 0.4234.

B.

- The observed difference in proportions between the two groups is about 0.0483 (males minus females).

C.

```
p1 = proportionMale
p2 = proportionFemale

minusP1 = 1 - proportionMale
minusP2 = 1 - proportionFemale
zValue = 1.96

handSE <- sqrt((p1 * (minusP1) / maleCount) + (p2 * (minusP2) / femaleCount))
marginError <- zValue * handSE
lowerBound <- (p1 - p2) - marginError
upperBound <- (p1 - p2) + marginError

confint(diffProps, level=0.95)
```

```
##      prop 1    prop 2       lower     upper level
## 1 0.5765766 0.5283019 -0.09315879 0.1897082  0.95
```

- The formula for the **standard error** for the difference in proportions is $\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}$.

- For this formula, I plugged in $\hat{p}_1$ as 0.4716981, $\hat{p}_2$ as 0.4234234, N_1 as 106, and N_2 as 111. So, the SE was sqrt((0.4716981 * (0.5283019) / 106) + (0.4234234 * (0.5765766) / 111)), which got me 0.0674563. Then, I calculated the lower and upper bounds of the confidence interval by adding and subtracting the margin of error from the sample estimate, p1 - p2. The sample estimate was 0.0482747, and the margin of error was 0.1322144. So, the lower bound was calculated (rounded to 4 digits) as -0.0839 and the upper bound was 0.1805. So, my confidence interval was (-0.0839, 0.1805).
- For the 95% confidence interval, I used 1.96 as my z* value because the true difference (in difference of proportions) is within 1.96 standard deviations from the mean in a normal distribution.

D.

If we were to repeatedly take random samples of male and female students and calculate the difference in proportions of those who fold their left arm on top, then we would expect that 95% of the resulting confidence intervals to contain the true difference in proportions between male and female students in the population who fold their left arm on top.

E.

The standard error (0.0675) above represents the expected amount of varying in the sample proportion differences between males and females who fold their left arm on top after repeatedly sampling (or bootstrap) students from this population.

F.

The term sampling distribution refers to the varying values or the distribution of the difference in sample proportions between male students and female students who fold their left arm on top when you repeatedly random sample students from this population. The sample proportions, and as a result, their differences will be changing from sample to sample as each sample has different students in the sample, but the the true population proportion difference will remain the same (even if we don't know the value).

G.

The mathematical result or theorem justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions is called the **Central Limit Theorem**. It states that when sample sizes are large enough, any sampling distribution of a sample proportion or sample mean has a normal distribution, disregarding the original shape of the population distribution. In this case, since the sample sizes are both over 30 from the female sample size and the male sample size, we can assume that both sample sizes are sufficiently large enough to assume a normal distribution for the sampling distrubution.

H.

If my 95% confidence interval for the difference in proportions was [-0.01, 0.30], it would make sense for someone to say that "there's no sex difference in arm folding" (for university students) as the confidence interval contains a difference of zero, meaning there is a chance of no true proportion difference between males and females in arm-folding which we cannot rule out.

I.

If you repeated this experiment many times with different random samples of university students, the confidence interval would be different across samples because each random sample has different students, which would lead to slight changes in the sample proportions for males and females who fold their left arm on top. As a result, since the confidence interval depends on the sample data, the variability in sample proportions would be captured and seen as the change in the confidence interval.

Problem 2 - Get out the vote

Part A

```
##      prop 1    prop 2     lower     upper level
## 1 0.5557551 0.3522267 0.1411399 0.2659167  0.95
```

The proportion of those receiving a GOTV call who voted in 1998 was 0.6478 and the proportion of those not receiving a GOTV call who voted in 1998 was 0.4442. According to my confidence interval, receiving a GOTV call increased the likelihood of voting by between 14.114 and 26.5917 percentage points, compared to not receiving a call.

Part B

Table 1: By GOTV Call

Variable	Received GOTV Call	Did Not Receive GOTV Call
Mean Age	58.3077	49.4253
Major Party Membership	0.8016	0.7448
1996 Voting Proportion	0.7126	0.5308

Table 2: By Voting in 1998

Variable	Voted in 1998	Did Not Vote in 1998
Mean Age	55.4153	44.9140
Major Party Membership	0.8019	0.7006
1996 Voting Proportion	0.7624	0.3497

```
##      prop 1   prop 2        lower      upper level
## 1 0.9821818 0.975492 0.0004615944 0.01291801  0.95

## [1] -11.395051  -6.369644
## attr(,"conf.level")
## [1] 0.95

##      prop 1    prop 2      lower      upper level
## 1 0.9859015 0.9696185 0.01060767 0.02195834  0.95
```

First Confidence Interval (Major Party Membership): This estimates the difference in major party membership proportions between those who received a GOTV call and those who did not. Since the confidence interval does not contain 0 as it's from (0.0044, 0.0129), there is a significant relation between party membership and the chance of receiving a GOTV call.

Second Confidence Interval (Age): This estimates the difference in mean age between GOTV call recipients and non-recipients. Since the confidence interval does not include 0 as it's from (-11.3951, -6.3696), it indicates that the age is significantly different between the two groups, meaning older individuals may have been more or less likely to receive a call.

Third Confidence Interval (1996 Voting): This estimates the difference in 1996 voting proportions between those who received a GOTV call and those who did not. Since the confidence interval does not contain 0 as it's from (0.1061, 0.2196), it suggests that past voting behavior (in 1996) is significantly related to the chance of receiving a GOTV call.

```
##      prop 1    prop 2     lower     upper level
## 1 0.6498182 0.5175145 0.1111651 0.1534422  0.95

## [1] -11.182008  -9.820602
## attr(,"conf.level")
## [1] 0.95

##      prop 1    prop 2     lower     upper level
## 1 0.7706513 0.3602624 0.3932429 0.4275349  0.95
```

First Confidence Interval (Major Party Membership and 1998 Voting): This estimates the difference in major party membership proportions between those who voted in 1998 and those who did not. Since the confidence interval does not contain 0 as it's from (0.1111, 0.1534), it suggests that major party membership is significantly associated with the chance of voting in 1998.

Second Confidence Interval (Age and 1998 Voting): This estimates the difference in mean age between those who voted in 1998 and those who did not. Since the confidence interval does not include 0 indicates as it's from (-11.182, -9.8206), it means that the age is significantly different between the two groups, meaning that age affects whether someone voted in 1998.

Third Confidence Interval (1996 Voting and 1998 Voting): This estimates the difference in 1996 voting proportions between those who voted in 1998 and those who did not. Since the confidence interval does not contain 0 as it's from (0.3932, 0.4275), it suggests that voting behavior in 1996 is signficiant in predicitng whether someone would vote in 1998.

Also, from Table 1, we see that older individuals, major party members, and past voters were more likely to be targeted for GOTV calls. From table 2, these same characteristics also made someone more likely to vote in 1998 (regardless of the GOTV calls). For example, we can see that the majority of people who had voted in 1998 also voted in 1996 and the majority of people who had received a GOTV called also voted in 1996, which shows the con-founder variable of voting in 1996.

This means that major party membership, age, and 1996 voting behavior are confounder variables that prevent my result in Part A from representing the true causal effect of the GOTV call on the likelihood that a person voted in 1998.

Part C

```
## Mean of voted1996 for GOTV Call Recipients:  0.7125506

## Mean of voted1996 for Non-GOTV Call Recipients:  0.530807

## Mean Age for GOTV Call Recipients:  58.30769

## Mean Age for Non-GOTV Call Recipients:  49.42534

## Mean Major Party Membership for GOTV Call Recipients:  0.8016194

## Mean Major Party Membership for Non-GOTV Call Recipients:  0.7447552
```

Before performing matching, here are the mean values for the three confounders (voted1996, AGE, and MAJORPTY) for both the treatment group (those who received a GOTV call) and the control group (those who did not receive a GOTV call).

The first two values are about voted1996, with the treatment mean followed by the control mean. The second two values are about age, with the treatment mean followed by the control mean. The third two values are about MAJORPTY, with the treatment mean followed by the control mean.

```
## Mean of voted1996 for GOTV Call Recipients (Matched Data):  0.7125506

## Mean of voted1996 for Non-GOTV Call Recipients (Matched Data):  0.7125506

## Mean Age for GOTV Call Recipients (Matched Data):  58.30769

## Mean Age for Non-GOTV Call Recipients (Matched Data):  58.2664

## Mean Major Party Membership for GOTV Call Recipients (Matched Data):  0.8016194

## Mean Major Party Membership for Non-GOTV Call Recipients (Matched Data):  0.8072874
```

After performing matching, here are the mean values for the three confounders (voted1996, AGE, and MAJORPTY) for both the treatment group (those who received a GOTV call) and the control group (those who did not receive a GOTV call).

The first two values are about voted1996, with the treatment mean followed by the control mean. The second two values are about age, with the treatment mean followed by the control mean. The third two values are about MAJORPTY, with the treatment mean followed by the control mean.

As the values are much more similar now, this suggests that the matching process has successfully balanced the confounders between the two groups. This implies that the differences between the treatment and control groups in terms of voted1996, AGE, and MAJORPTY are now minimized, and the groups are comparable with respect to these variables.

Now, the matched data set is balanced and these confounders will no longer affect the relationship between the GOTV call and the likelihood of voting in 1998 (in this dataset).

```
##      prop 1    prop 2       lower      upper level
## 1 0.2874494 0.2874494 -0.06182709 0.06182709  0.95

## [1] -2.760374  2.677783
## attr(,"conf.level")
## [1] 0.95

##      prop 1    prop 2      lower      upper level
## 1 0.1927126 0.1983806 -0.0624769 0.05114086  0.95
```

First Confidence Interval (1996 Voting and GOTV Call): This estimates the difference in 1996 voting proportions between those who received a GOTV call and those who did not. Since the confidence interval includes 0 as it's from (-0.6183, 0.6183), it suggests that the difference in voting behavior from 1996 between the two groups is not statistically significant, indicating that 1996 voting behavior may not strongly influence whether someone received a GOTV call.

Second Confidence Interval (Age and GOTV Call): This estimates the difference in mean age between those who received a GOTV call and those who did not. Since the confidence interval contains 0 as it's from (-2.7604, 2.6778) , it suggests that age is not significantly different between the two groups post-matching, implying that age is not a key differentiator in receiving a GOTV call once matching is applied.

Third Confidence Interval (Major Party Membership and GOTV Call): This estimates the difference in major party membership proportions between those who received a GOTV call and those who did not. The confidence interval containing 0 as it's from (-0.0625, 0.5111) indicates that the difference in major party membership between the two groups is not statistically significant, suggesting that being a member of a major party is not a strong factor in determining whether a person received a GOTV call after matching.

This further shows that after applying the matching procedure, the three confounders (1996 voting, age, and major party membership) do not significantly differ between those who received a GOTV call and those who did not, so they now won't affect the new 95% confidence interval.

```
##     diffprop
## -0.005668016
```

```
##   diffmean
## 0.04129555
```

```
## diffprop
##        0
```

Lastly, you can see the difference in proportions (difference in mean for age) between the control and treatment groups is close to 0 across the 3 confounder variables, suggesting that the matching was successful in balancing them (the treatment and control groups).

```
##      prop 1    prop 2      lower     upper level
## 1 0.4307692 0.3522267 0.01045353 0.1466315  0.95
```

As the new 95% confidence interval (0.1045, 0.1466) does not contain the value 0 within its interval, this means the null hypothesis can be rejected (GOTV call has no effect on the likelihood of voting in 1998). This means that receiving a GOTV call had a statistically significant effect on voter turnout in 1998. Since the difference in proportions is positive, we can conclude that individuals who received a GOTV call were more likely to vote in 1998 compared to those who did not receive a GOTV call.