HW7

Brian Liang

2025-04-02

Problem 1

A.

There are 106 males and 111 females in the dataset. The sample proportion of males who folded their left arm on top was around 0.4717, and the sample proportion of females who folded their left arm on top was around 0.4234.

B

There was an observed difference in proportion between the two groups (male - female) of around 0.0483.

C

Table 1: Confidence Interval for GOTV Call by 1996 Vote

lower	upper	level
-0.0931588	0.1897082	0.95

These values found for the confidence interval were found utilizing R's built-in function but this function can be double-checked using the "hand-calculated" version using the following steps.

The formula for the standard error of a difference in proportions is:

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

The values I plugged in were:

$$(p_1 = 0.47), (n_1 = 106), (p_2 = 0.42), (n_2 = 111)$$

The z* value I used was 1.96 as this represents the equivalent to looking for a 95% confidence interval. This is because 1.96 is the amount of standard deviations above and below the center of a normal distribution where 95% of the data will lie.

I was then able to use all of these values in this formula:

$$(p_1 - p_2) \pm z^* \cdot SE$$

to create a confidence interval for the difference in proportions of males and females and how they fold their arms.

This gives us a 95% confidence interval of -0.0839 to 0.1805 which is quite close to the values we found from the built-in R function.

D

If we were to find the true difference in proportion of males and females who fold their arms with the left arm over the right, we would expect there to be no difference with 95% confidence as 0 is contained within our confidence interval, with 0 signifying no difference.

E

The standard error represents the variability we have from our data of males and females who fold their arms with the left arm over the right. It helps to account for and measure how the observed difference could fluctuate due to the variations from how the data was obtained.

F

The sampling distribution in this context refers to the the distribution we would see if we were to take many random samples of males and females and the difference in proportion between the groups based on how they fold their arms. In each sample the proportions of males and proportion of females who fold their arms with the left arm over their right would vary from sample to sample which in turn would affect the difference in proportions. The parts that would stay fixed are the sample sizes, sampling method, and the true population proportions for males and females as we sample from the same population.

G

The theorem that justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions is the Central Limit Theorem. This theorem helps us as regardless of the shape of the original population distribution, the statistical fluctuations from sampling with a large enough sample size will mean the sample difference in proportions will look approximately normal. It will also be approximately centered around the true population proportion which will help determine our best estimate as to what the true population proportion would look like.

H

I would say to someone who claims "there's no sex difference in arm folding" based on a confidence interval of [-0.01, 0.30] that the interval does support their claims. This is because 0, which signifies there would be no difference, lies within the range of the interval and that the observed difference was not statistically significant. However, the values from the confidence interval suggest that there could be a difference as -0.01 is close to 0 and 0.30 is high in terms of proportions. It would likely require a larger sample or more data to find more definitive evidence of whether or not there is a difference.

I

Yes, the confidence intervals would be different across samples as the process is intrinsically variable as different samples will create different ranges for the confidence intervals. However it will be true that the collection of all these intervals will contain the true population statistic at the levels at which the confidence intervals were taken (95% of the time if it were a 95% confidence interval).

Problem 2

Part A

```
##  prop_1.0  prop_1.1
## 0.4442449 0.6477733
```

Above are the proportions of those who received a GOTV call who voted in 1998 and those who did not receive a GOTV call and voted in 1998. We see that the proportion who did receive a call was around 0.6478 and those who did not receive a call was around 0.4442.

I then created a confidence interval based on these proportions to find the difference in proportions of votes who voted in 1998 by if they received a GOTV call:

Table 2: Confidence Interval for GOTV Call by 1996 Vote

lower	upper	level
0.1411399	0.2659167	0.95

It was found that with 95% confidence, the true difference in proportion of those who voted in 1998 by if they received a GOTV call or not lies between the intervals of around 0.1411 and 0.2659. This suggests that there is a difference as those who received a call had a higher likelihood of voting.

Part B

However, the difference we found may have been affected by other, confounding variables.

```
## # A tibble: 2 x 2
##   voted1996 mean_GOTV_call
##       <int>          <dbl>
## 1         0         0.0141
## 2         1         0.0304
```

Table 3: Confidence Interval for GOTV Call by 1996 Vote

lower	upper	level
0.0106077	0.0219583	0.95

Above we see how the difference in proportion of people who did or did not receive a GOTV call is confounded by if they voted in 1996. From the table of values for the proportions, we do observe a difference. Then a confidence interval for these variables show that there is a difference with 95% confidence as 0 is not contained within the interval and there is a statistically significant difference that is between the intervals of 0.0106 to 0.0220.

```
## # A tibble: 2 x 2
##   MAJORPTY mean_GOTV_call
##      <int>          <dbl>
## 1        0         0.0178
## 2        1         0.0245
```

Table 4: Confidence Interval for GOTV Call by Affiliation with Major Party

lower	upper	level
0.0004616	0.012918	0.95

Above we see how the difference in proportion of people who did or did not receive a GOTV call is confounded by if they were affiliated with a major party. From the table of values for the proportions, we do observe a difference. Then a confidence interval for these variables show that there is a difference with 95% confidence as 0 is not contained within the interval and there is a statistically significant difference that is between the intervals of 0.0004 to 0.0129.

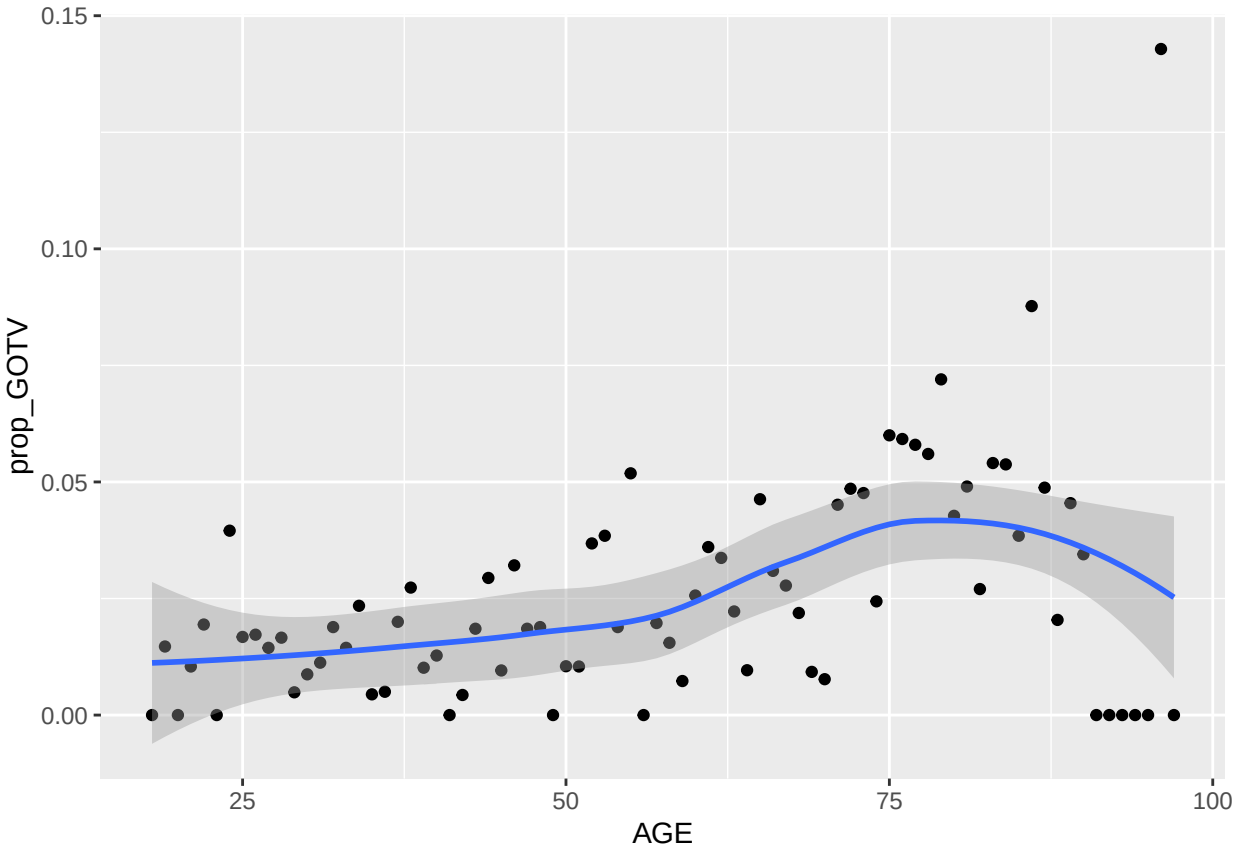

Table 5: Confidence Interval for GOTV Call by Age

lower	upper	level
-11.39505	-6.369644	0.95

The scatterplot above shows how age is correlated with the proportion of those who received a GOTV call. The points show a relationship between the variables that as age increases then the proportion of those who receive a call slightly increases. This shows that age is a confounder for if someone were to vote. The confidence interval for these variables shows that with 95% confidence, there is a difference as 0 is not contained within the interval and there is a statistically significant difference that is between the interval of around 9.82 to 11.18.

Next we wanted to prove that these variables are also confounders for if someone were to vote in 1998:

```
## # A tibble: 2 x 2
##   voted1996 prop_voted1998
##       <int>          <dbl>
## 1         0          0.229
## 2         1          0.640
```

Table 6: Confidence Interval for GOTV Call by Age

lower	upper	level
0.3932429	0.4275349	0.95

We see that there is a large observed difference between someone who voted in 1998 and whether or not

they voted in 1996 based on the proportions found in the first table above. The confidence interval for these variables also show that there is a difference with 95% confidence as 0 is not contained within the interval and there is a statistically significant difference that is between the interval of around 0.3932 to 0.4275.

```
## # A tibble: 2 x 2
##   MAJORPTY prop_voted1998
##      <int>          <dbl>
## 1        0          0.350
## 2        1          0.482
```

Table 7: Confidence Interval for GOTV Call by Age

lower	upper	level
0.1111651	0.1534422	0.95

There is also a difference between someone who voted in 1998 and whether or not they were affiliated with a major party based on the first table above. The confidence interval for these variables also show that there is a difference with 95% confidence as 0 is not contained within the interval and there is a statistically significant difference that is between the interval of around 0.1111 to 0.1534.

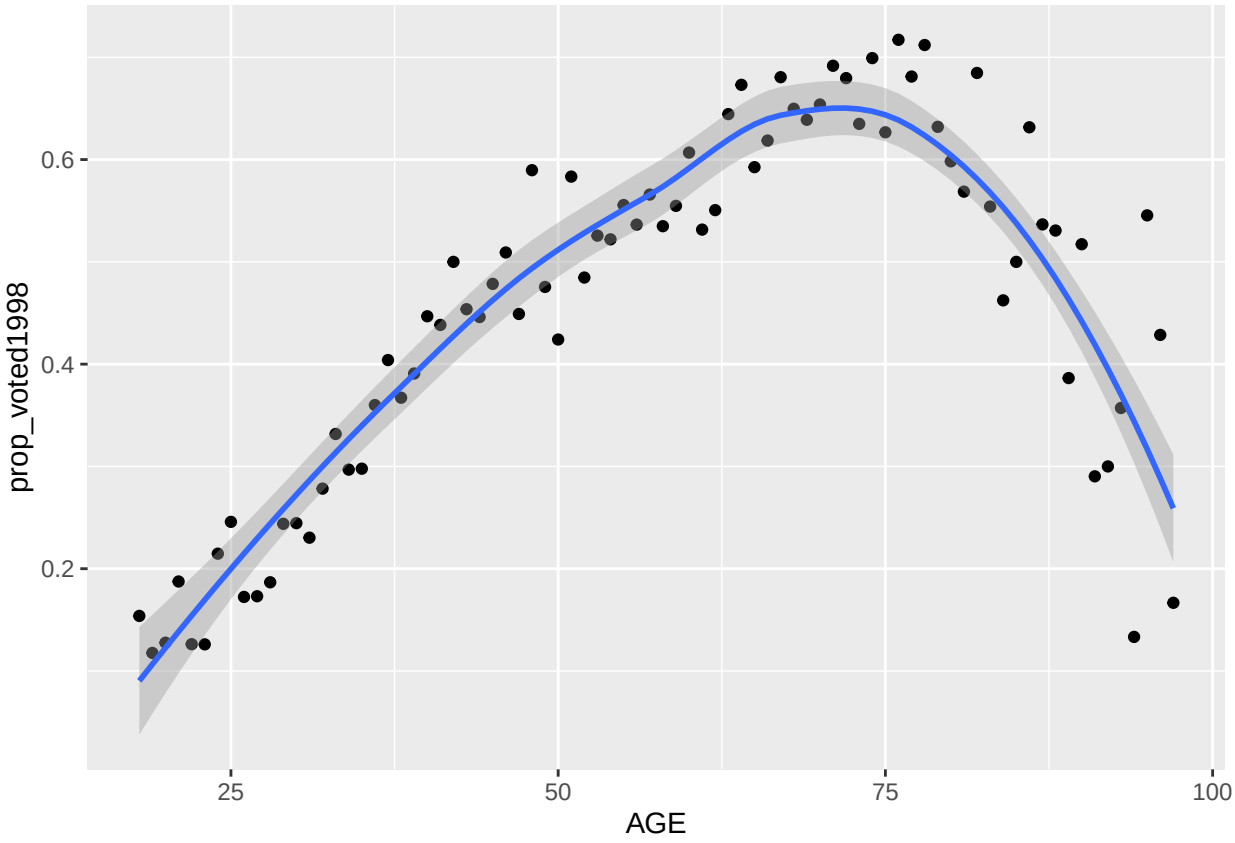

Table 8: Confidence Interval for GOTV Call by Age

lower	upper	level
-11.18201	-9.820602	0.95

The scatterplot above shows how age is correlated with the proportion who voted in 1998. The points show a

definite relationship between the variables that as age increases then the proportion of those people would vote. This shows that age is a confounder for if someone were to vote. The confidence interval for these variables shows that with 95% confidence, there is a difference as 0 is not contained within the interval and there is a statistically significant difference that is between the interval of around 9.82 to 11.18.

All of these variables were found to be confounders that affected both our variables of if they recieved a GOTV call and if they voted in 1998.

Part C

To get a better estimate on how receiving a GOTV call would affect someone's likelihood to vote, I matched or balanced the data by comparing data that had similar values for age, if they voted in 1996, and if they were affiliated with a major party.

```
## # A tibble: 2 x 2
##   voted1996 mean_GOTV_call
##       <int>          <dbl>
## 1         0          0.167
## 2         1          0.167
```

Table 9: Confidence Interval for GOTV Call by 1996 Vote

lower	upper	level
-0.0419246	0.0419246	0.95

Above is the results from after matching our data which show the same values for the proportions of those who received a GOTV call by if they voted in 1996. The confidence interval also shows that with 95% confidence there is not a statistically significant difference as 0 is contained within the interval which signifuies a difference.

```
## # A tibble: 2 x 2
##   MAJORPTY mean_GOTV_call
##      <int>          <dbl>
## 1        0          0.171
## 2        1          0.166
```

Table 10: Confidence Interval for GOTV Call by Affiliation with Major Party

lower	upper	level
-0.0555697	0.045487	0.95

Above is the results for the proportions of those who received a GOTV call by if they were affiliated with a major party or not. While there is a difference in the values, the confidence interval shows that with 95% confidence there is not a statistically significant difference as 0 is contained within the interval which signifies no difference.

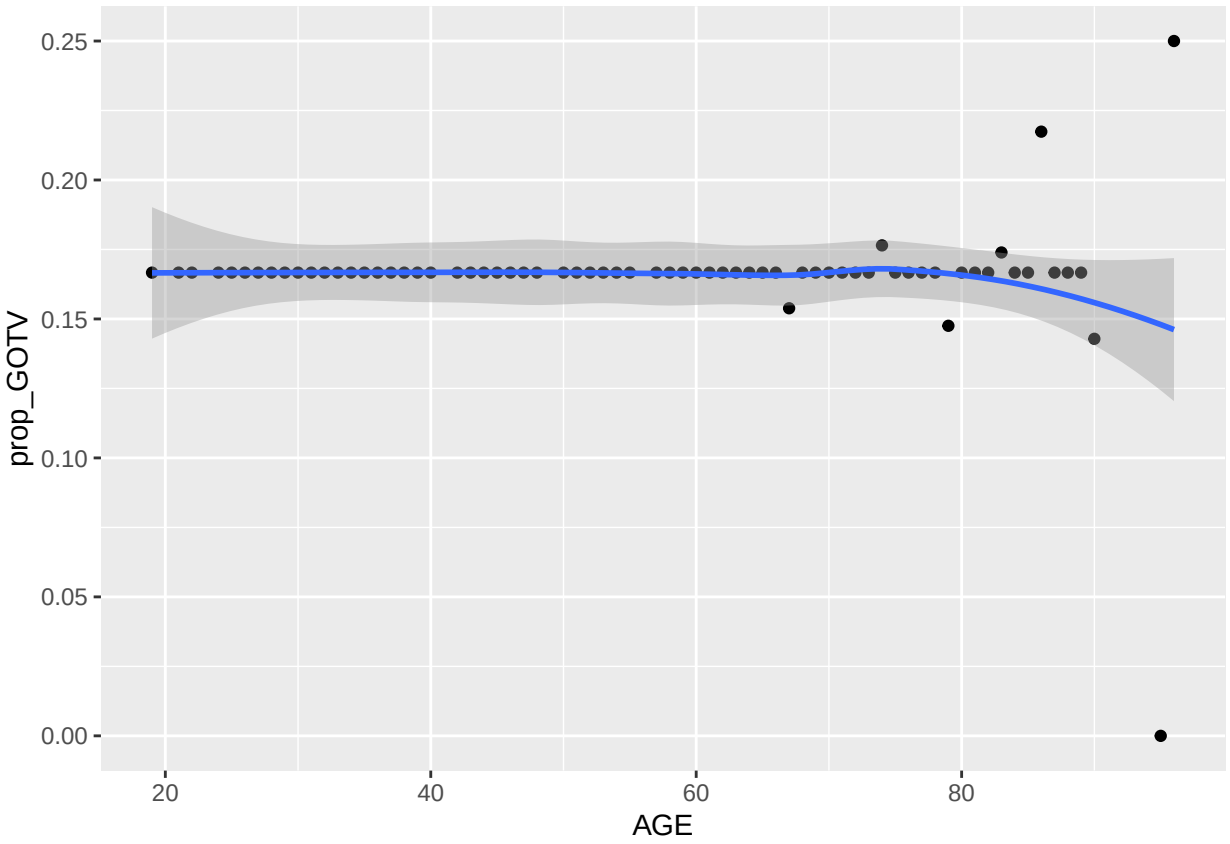

Table 11: Confidence Interval for GOTV Call by Age

lower	upper	level
-2.760375	2.677783	0.95

Above shows the plot of proportion who received a GOTV call by their age which shows that the proportion stays the same for the majority of the ages. The confidence interval for these variables also shows that with 95% confidence, there is not a statistically significant difference as 0 is contained within the interval whihc signifies no difference.

This means that out matched data balances for these confounders well and now we can find better evidence for how receiving a GOTV call would affect the likelihood of voting in the 1998 election.

```
##  prop_1.0  prop_1.1
## 0.5692308 0.6477733
```

Above are the proportions of those receiving a GOTV call and voting in 1998 and those who did not receive a GOTV call and voted in 1998. We see that the proportion who did receive a call was around 0.6478 and those who did not receive a call was around 0.5692 using the new matched data.

Table 12: Confidence Interval for GOTV Call by Age

lower	upper	level
0.0104535	0.1466315	0.95

Above is the confidence interval for difference in the two proportions which shows that with 95% confidence,

the true difference in proportions of those who received a GOTV call and voted in 1998 against those who did not receive a GOTV call and voting in 1998 lies in the range of around 0.0105 to 0.1466. This is a statistically significant difference as 0 is not contained within the interval.

From this we can conclude that a GOTV call does improve the likelihood of voting in the 1998 election as we found a statistically significant difference between those who did and did not receive a call. With 95% confidence, it will improve the likelihood of voting from a value between 1.05% to 14.66%.