HW7

Surafel Geleta ssg2775 https://github.com/surafelgeleta/HW7-SDS315

Problem 1

Part A

There are 111 female students and 106 male students in the dataset. The sample proportion of males who folded their left arm on top is 0.472, while the sample proportion of females who folded their left arm on top is 0.423.

Part B

The observed difference in the proportion of males and females who folded their left arm on top is around 0.048.

Part C

Using R's built-in "prop.test()" function, I found with 95% confidence that the true difference in the proportion of males and females who folded their left arm on top is between -0.0839 and 0.1804. Since the confidence interval captures 0, there is not enough evidence to suggest that there is a difference between the proportions of males and females who folded their left arm on top.

I found very similar results using the formula for the construction of a confidence interval, $\theta \in \hat{\theta} \pm z \times se(\hat{\theta})$, where $\hat{\theta} = \hat{p}_1 - \hat{p}_2 = 0.0483$ and $se(\hat{\theta}) = \sqrt{}((\frac{(\hat{p}_1) \times (1 - \hat{p}_1)}{N_1}) + (\frac{(\hat{p}_2) \times (1 - \hat{p}_2)}{N_2})) = \sqrt{}((\frac{(0.4717) \times (1 - 0.4717)}{106}) + (\frac{(0.4234) \times (1 - 0.4234)}{111})) \approx 0.0675$. I set $z = 1.96$ because I am constructing a 95% confidence interval, and the area across 1.96 standard errors of a normal curve covers about 95% of the curve's area. So $0.0483 + 1.96 \times 0.0675 = 0.1806$ and $0.0483 - 1.96 \times 0.0675 = -0.084$ gives us the upper and lower bound respectively of our 95% difference in proportions confidence interval, with very similar values to the confidence interval produced by the R function.

Part D

If we were to construct 100 confidence intervals on this data, then we would expect that 95 of them would capture the true difference in proportion of males and females who folded their left arm on top.

Part E

The standard error is the standard deviation of a sampling distribution of the sample estimate, that being the measured difference in proportion of males and females who folded their left arm on top.

Part F

The sampling distribution here refers to the distribution of sample estimates of the difference in proportion of males and females who folded their left arm on top across many repeated samples. What varies from sample to sample is the sample estimate of the difference in proportion, however the population difference in proportions and sample size remains the same.

Part G

The central limit theorem justifies using a normal distribution to approximate the sampling distribution of the sample difference in proportions. This is because it states that, regardless of the original population distribution, repeated sample estimates (thereby forming a sampling distribution) taken from a sufficiently large sample (which is true in this case, because the sample size is 217) will have the shape of a normal curve.

Part H

I would tell them that the confidence interval containing 0 does not necessarily mean that there is no sex difference in proportion of folding arms with the left arm on top, but rather just means that the difference is not statistically significant.

Part I

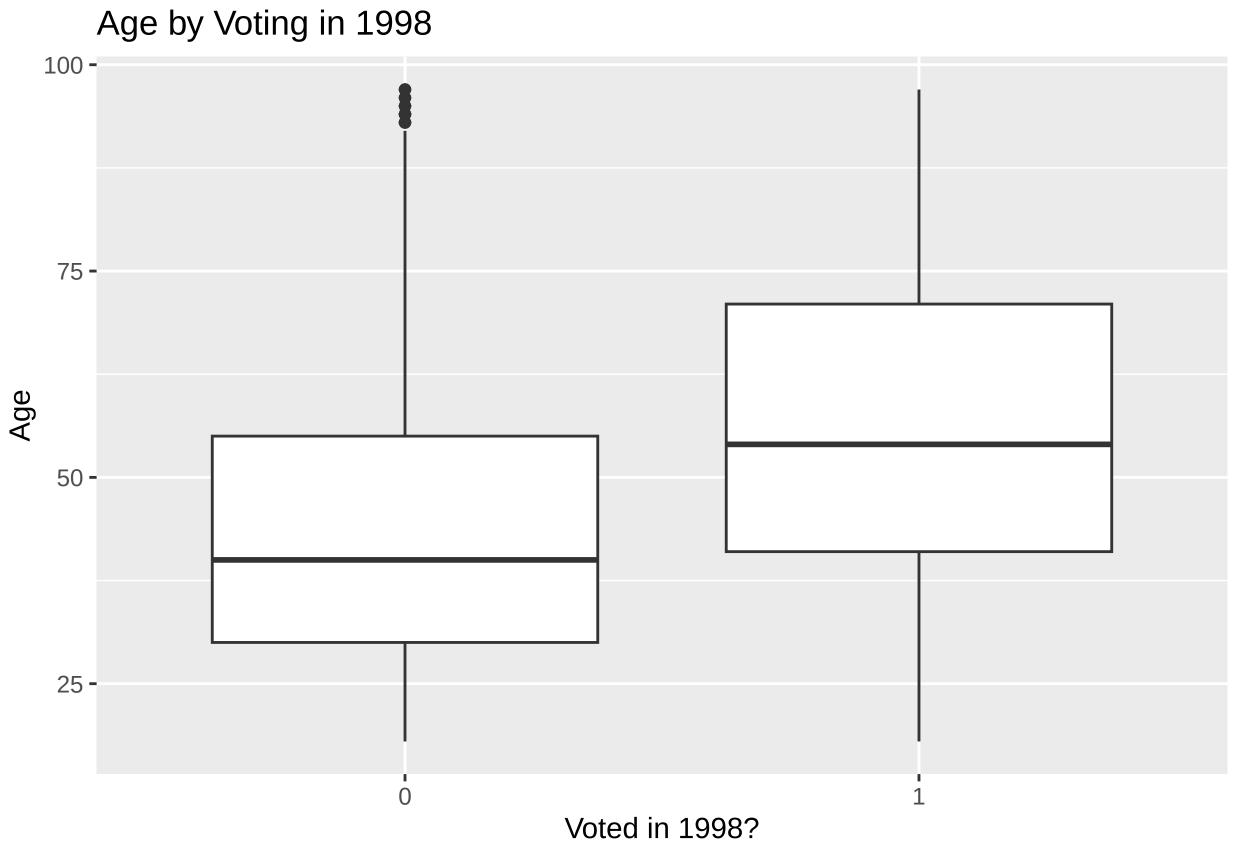
If this experiment was repeated many times with different random samples of students, the confidence interval would be different across samples because the sample difference in proportions may vary across samples. If a confidence level of 95% is used, then about 95% of the collection of these intervals should contain the true difference of proportions of males and females who fold their left arm on top.

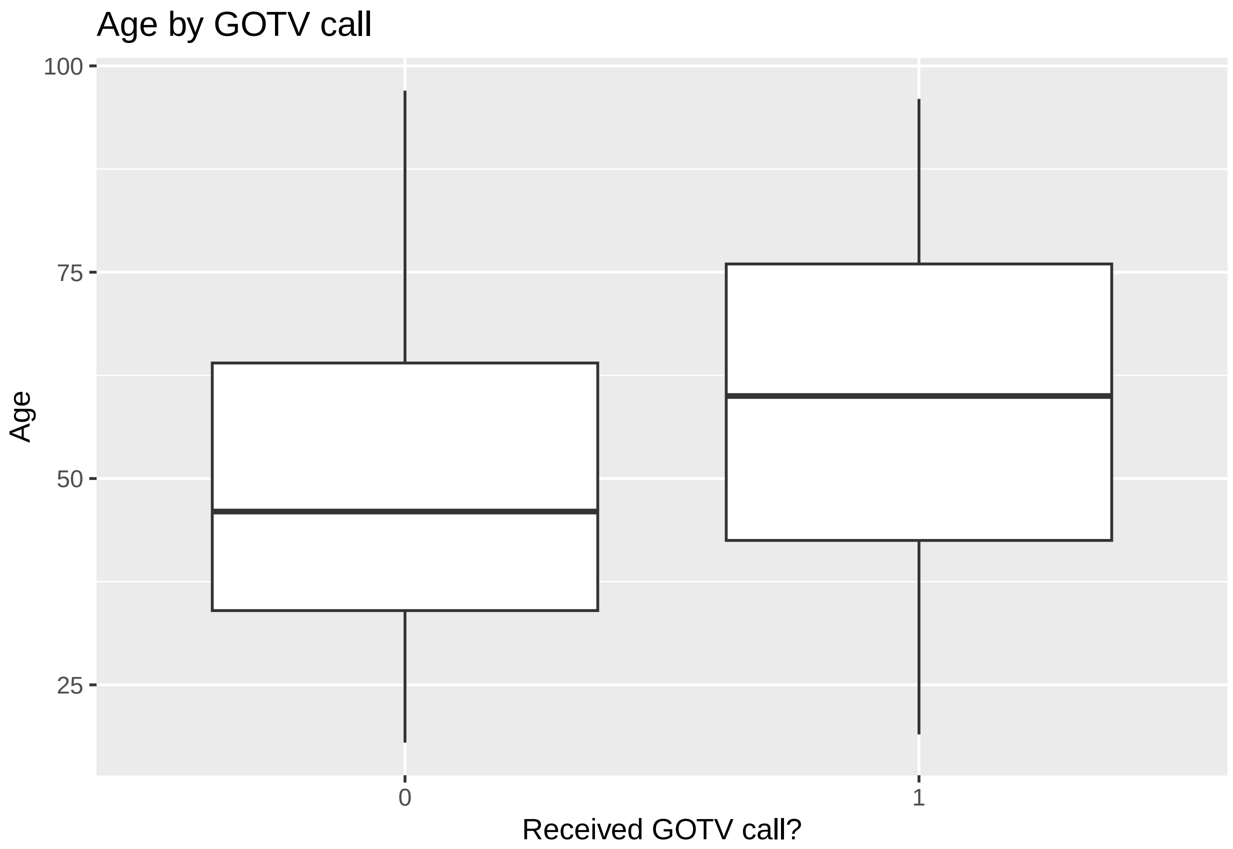
Problem 2

Part A

The proportion of those who received a GOTV call and voted in 1998 was 0.6478. The sample proportion of those who didn't receive a GOTV call who voted in 1998 was 0.4442. A 95% confidence interval shows with 95% confidence that the true difference in proportions of GOTV call receivers and nonreceivers that voted in the 1998 election lies between 0.1411 and 0.2659.

Part B


Age by Voting in 1998
100
75
50
25
Age
0
1
Voted in 1998?


Age by GOTV call
100
75
50
25
Age
0
1
Received GOTV call?

Within the sample, 1998 voters tended to be older than non-voters, and those who received GOTV calls tended to be older than non GOTV call receivers.

A 95% confidence interval demonstrates with 95% confidence that the true difference in mean age between those who did not vote in 1998 and those who did lies between -11.182 and -9.82 years. In addition, a 95% confidence interval demonstrates with 95% confidence that the true difference in mean age between those who did not receive a GOTV call and those who did lies between -11.395 and -6.369 years. Since neither of these confidence intervals contain 0, we have evidence to reject the null hypotheses that there is no difference in mean age between 1998 voters and non-voters, and that there is no difference in the mean age between GOTV call receivers and those who did not receive a call. So, since a voter's age may affect their likelihood of voting in 1998 and may also affect their likelihood of receiving a GOTV call, age is a confounder that keep us from identifying the causal effect of GOTV calls on voting propensity in 1998.

Table 1: Registration with Major U.S. Party, 1998 Voting, and GOTV Calls

Registered w/Major Party?	Prop. Voted 1998	Prop. Received GOTV Call	Size
0	0.3501818	0.0178182	2750
1	0.4824855	0.0245080	8079

Within the sample, people who were registered members of one of the two major U.S. political parties were more likely to have voted in 1998 and more likely to have received a GOTV call.

With 95% confidence, the true difference in the proportion of people who voted in the 1998 election between people registered with a major U.S. political party and people not registered with a major U.S. political party lies between 0.1114 and 0.1532. Also with 95% confidence, the true difference in the proportion of people who received a GOTV call between people registered with and not registered with a major U.S. political party lies between 0.0007053 and 0.012674. Since neither 95% confidence interval contains 0, there is evidence to reject the null hypothesis that there is no difference in the proportion of people who voted in 1998 among non-registered and registered party members, and to reject the null hypothesis that there is no difference in the proportion of people who received a GOTV call among non-registered and registered party members. So, since a person's major party registration status may affect their likelihood of voting in 1998 and of receiving a GOTV call, a person's registration status with a major U.S. political party is a confounder.

Table 2: Turnout in 1996 Election, 1998 Voting, and GOTV Calls

Voted in 1996?	Prop. Voted 1998	Prop. Received GOTV Call	Size
0	0.2293487	0.0140985	5036
1	0.6397376	0.0303815	5793

Within the sample, people who voted in 1996 had a higher likelihood of voting in 1998 and of receiving a GOTV call.

With 95% confidence, the true difference in the proportion of people who voted in 1998 between people who did and didn't vote in 1996 lies between 0.3934 and 0.4273. Also with 95% confidence, the true difference in the proportion of people who received a GOTV call between people who did and didn't vote in 1996 lies between 0.0108 and 0.0218. Since neither confidence interval captures 0, there is evidence to reject the null hypotheses that there is no difference in the proportion of people who voted in 1998 among people who did and didn't vote in 1996, and that there is no difference in the proportion of people who received a GOTV call among people who did and didn't vote in 1996. Since voting in 1996 affects one's likelihood of voting in 1998 and of receiving a GOTV call, whether of not someone voted in 1996 is a confounder.

Part C

Table 3: Summary Stats of Confounders by Treatment Value After Matching

GOTV Call Received?	Mean Age	Prop. Voted 1996	Prop. Reg. w/ Major Party
0	58.26640	0.7125506	0.8072874
1	58.30769	0.7125506	0.8016194

After matching, the three confounding variables age, voting in 1996, and registration status with a major U.S. political party appear to have similar values across the control and treatment groups.

With 95% confidence, the true difference in mean ages between people who did and did not receive GOTV calls in the matched dataset lies between -2.7604 and 2.6778 years. Since 0 is in this interval, we must fail to reject the null hypothesis that there is no difference in the mean ages between people who did and did not receive GOTV calls. This indicates that age is not a confounder in the matched dataset, since age does not appear to affect a person's likelihood to receive a GOTV call.

WIth 95% confidence, the true difference in the proportion of people who voted in the 1996 election between people who received and didn't receive a GOTV call in the matched dataset lies between -0.0618 and 0.0618. Since the confidence interval contains 0, we fail to reject the null hypothesis that there is no difference in the proportion of people who voted in 1996 between people who did and didn't receive a GOTV call in the matched dataset. This indicates that voting in 1996 is not a confounder in the matched dataset, since voting in 1996 doesn't appear to affect the likelihood of someone receiving a GOTV call.

With 95% confidence, the true difference in the proportion of people registered with a major U.S. political party between people who did and did not receive a GOTV call lies between -0.06 and 0.05. So, we fail to reject the null hypothesis that there is no difference in the proportion of people registered with a major U.S. political party between people who did and did not receive a GOTV call in the matched dataset. Since registration status with a major U.S. political party does not appear to affect the likelihood of someone receiving a GOTV call, the former variable is not a confounder in the matched dataset.

In the matched dataset, the proportion of individuals who received a GOTV call that voted in 1998 is 0.6478, while the sample proportion of individuals who did not receive a GOTV call that voted in 1998 is 0.5692.

Using a 95% confidence interval, I find with 95% confidence that the true difference in the proportions of people who voted in 1998 between people who did and did not receive a GOTV call lies between 0.01288 and 0.14420. Since this confidence interval does not contain 0, there is evidence to reject the null hypothesis that there is no difference in the proportions of people who voted in 1998 between people who did and did not receive a GOTV call.

The 95% difference in proportions confidence interval of the matched data contains smaller values than the confidence interval for the unmatched data, suggesting that the difference in the proportion of people who voted in 1998 between people who did and did not receive a GOTV call is smaller when matching for certain confounders. However, it is still not possible to identify the causal effect of GOTV calls on voter propensity in 1998. I only matched on three confounding variables available in the dataset, and there is a high risk of omitted variable bias as a result of other potential confounders not being included in the dataset. These omitted variables might include sex, education and U.S. state of residence.