Soumil Asanare (ssa2958)
Github: https://github.com/Soumil-A/HW7SDS.git

Problem 1

Part A

Number of male and female students:

There are 106 male students and 111 female students in the dataset.

Sample proportion of males who folded their left arm on top:

About 47.17% of men

Sample proportion of females who folded their left arm on top:

About 42.34% of females

Part B

The observed difference in proportions between the two groups is about **0.0483.**

Part C

```
 Female Male
  0      64   56
  1      47   50

      2-sample test for equality of proportions without continuity correction

data:  c(male_LonR, female_LonR) out of c(maleSum, femaleSum)
X-squared = 0.51118, df = 1, p-value = 0.4746
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.08393731  0.18048668
sample estimates:
   prop 1    prop 2
0.4716981 0.4234234

[1] -0.08393973  0.18048911
```

$$\sqrt{\frac{\hat{p}_1\,(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2\,(1-\hat{p}_2)}{n_2}}$$

This is the equation for the the standard error for the difference in proportions

P1 = 0.4717 (males)

P2 = 0.4234 (females)

N1 = 106 (number of males)

N2 = 111 (number of females)

I used the z* value 1.96. This is the critical value from the standard normal distribution that captures the middle 95% of the distribution. It corresponds to a two-tailed confidence level of 95%, with 2.5% in each tail.

Part D

If we were to repeatedly take random samples of students from this university's population and calculate the difference in arm-folding proportions, then we would expect that about 95% of those confidence intervals would contain the true difference in population proportions between males and females who fold their left arm on top.

Part E

The standard error is 0.0675. This means that the difference in proportions between male and female would typically vary by about 6.75% just due to chance. The standard error means value of the observed difference between the proportions of males and females who fold their left arm on top would vary if we repeatedly took different random samples from the same population of students.

Part F

The sampling distribution refers to the distribution of the difference in proportions we would see if we took infinite random samples from the same population. In this assignment, it shows how the difference between the proportions of males and females folding their left arm on top varies. The thing that varies is the observed difference in proportions. What stays the same is the true population difference, which we are trying to estimate.

Part G

The mathematical theorem that justifies a normal distribution to approximate the sampling distribution of the difference in proportions is the Central Limit Theorem. It says that with large enough sample sizes the sampling distribution will turn out approximately normal. Since both groups have enough observations and at least 10 successes and failures, the normal approximation works here.

Part H

Since the confidence interval (-0.084, 0.180) contains 0, there is no difference in arm folding between the sex of the person

Part I

Yes there would be a difference in the confidence interval because there would be different people and due to random chance. There will be a natural variation in those intervals. But if we repeat this with many samples, then about 95% of those confidence intervals would contain the true population difference in proportions.

Problem 2

Part A

```
[1] 0.6477733
[1] 0.4442449

      2-sample test for equality of proportions without continuity correction

data:  c(sum(dataGov$voted1998[dataGov$GOTV_call == 1]),
sum(dataGov$voted1998[dataGov$GOTV_call == 0])) out of c(sum(dataGov$GOTV_call
== 1), sum(dataGov$GOTV_call == 0))
```

```
X-squared = 40.416, df = 1, p-value = 2.053e-10
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1432115 0.2638452
sample estimates:
   prop 1    prop 2
0.6477733 0.4442449
```

This confidence interval suggests that those who received a GOTV call were 14.3 to 26.4 percentage points more likely to vote in 1998 than those who didn't. Since the entire interval is above 0, this provides strong evidence of a positive association between receiving a GOTV call and voting behavior.

Part B

GOTV_call <dbl>	voted1996 <dbl>	AGE <dbl>	MAJORPTY <dbl>
0	0.5308070	49.42534	0.7447552
1	0.7125506	58.30769	0.8016194

```
2-sample test for equality of proportions without continuity
correction

data:  c(voted96_gotv, voted96_nogotv) out of c(n_gotv, n_nogotv)
X-squared = 32.047, df = 1, p-value = 1.505e-08
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1245081 0.2389791
sample estimates:
   prop 1    prop 2
0.7125506 0.5308070
```

```
     2-sample test for equality of proportions without continuity
correction
```

```
data:  c(party_gotv, party_nogotv) out of c(n_gotv, n_nogotv)
X-squared = 4.1195, df = 1, p-value = 0.04239
alternative hypothesis: two.sided
95 percent confidence interval:
 0.006443461 0.107284916
sample estimates:
   prop 1    prop 2
0.8016194 0.7447552


      Welch Two Sample t-test

data:  AGE by GOTV_call
t = -6.9613, df = 256.33, p-value = 2.817e-11
alternative hypothesis: true difference in means between group 0 and
group 1 is not equal to 0
95 percent confidence interval:
 -11.395051  -6.369644
sample estimates:
mean in group 0 mean in group 1
       49.42534        58.30769
```

Since all three of these intervals do not contain 0, we can say with 95% certainty that there is a difference in the sample proportions and all these variables are confounding variables.

Part C

GOTV_call <dbl>	voted1996 <dbl>	AGE <dbl>	MAJORPTY <dbl>
0	0.7125506	58.26640	0.8072874
1	0.7125506	58.30769	0.8016194

Now The confidence Intervals at a 95% level for each confounder.

```
2-sample test for equality of proportions without continuity
correction

data:  c(voted96_gotv, voted96_nogotv) out of c(n_gotv, n_nogotv)
X-squared = 2.6633e-29, df = 1, p-value = 1
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06182709  0.06182709
sample estimates:
   prop 1    prop 2
0.7125506 0.7125506


     2-sample test for equality of proportions without continuity
correction

data:  c(party_gotv, party_nogotv) out of c(n_gotv, n_nogotv)
X-squared = 0.042347, df = 1, p-value = 0.837
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.06004775  0.04871171
sample estimates:
   prop 1    prop 2
0.8016194 0.8072874


     Welch Two Sample t-test

data:  AGE by GOTV_call
t = -0.02987, df = 350.55, p-value = 0.9762
alternative hypothesis: true difference in means between group 0 and
group 1 is not equal to 0
95 percent confidence interval:
 -2.760374  2.677783
sample estimates:
mean in group 0 mean in group 1
       58.26640        58.30769
```

Because all 3 of these confidence intervals include 0 we can say with 95% certainty that the matching successfully balanced the groups on those confounders.

```
2-sample test for equality of proportions without continuity
correction

data:  c(voted_gotv, voted_nogotv) out of c(n_gotv, n_nogotv)
X-squared = 5.2206, df = 1, p-value = 0.02232
alternative hypothesis: two.sided
95 percent confidence interval:
 0.01288268 0.14420234
sample estimates:
   prop 1    prop 2
0.6477733 0.5692308
```

We can conclude with 95% certainty that the GOTV call most likely had a real impact on voter turnout in the 1998 election because the matched data shows that the confidence interval is still positive and does not contain 0. We have a 95% confidence level that the actual variation in the population's voting rates falls between 1.29 and 14.42 percent of the actual.