Homework7

Daniel Wu

2025-04-01

To access my GitHub repository, click here: https://github.com/DanielWu3627/SDS315. Please check the file named **Homework7.Rmd**.

Problem 1

Part A

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

```
## prop_1.Female   prop_1.Male
##     0.4234234     0.4716981
```

In the dataset, there is a total of 111 female students and 106 male students. The sample proportion of females who folded their left arm on top is 0.4234234 while the sample proportion of males who folded their left arm on top is 0.4716981.

Part B

```
## [1] "The observed difference in proportions between the two sexes is 0.0483"
```

Part C

```
sd_error <- sqrt(((0.4234234 * (1-0.4234234))/111) + (0.4716981 * (1-0.4716981))/106)
up_val <- (0.4234234 - 0.4716981) + 1.96 * sd_error
up_val
```

```
## [1] 0.08393972
```

```
low_val <- (0.4234234 - 0.4716981) - 1.96 * sd_error
low_val
```

```
## [1] -0.1804891
```

```
prop.test(LonR_fold ~ Sex, data=arms, success=1)
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(LonR_fold ~ Sex)
## X-squared = 0.33454, df = 1, p-value = 0.563
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.18970817  0.09315879
```

```
## sample estimates:
##    prop 1    prop 2
## 0.4234234 0.4716981
```

According to the built-in function, the confidence interval is [-0.1897, 0.0932].

The formula I used to calculate the standard error for the difference in proportions is the square root of p1 times 1 minus p1 over n1 plus p2 times 1 minus p2 over n2. I plugged in 0.4234234 as p1 and 111 as n1 since both values represent female students. I plugged in 0.4716981 as p2 and 106 as n2 since these numbers represent male students.

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

After hand-calculating the confidence interval, I found that the 95% confidence interval is [-0.1805, 0.0839] for the difference in proportions between males and females, using males as references. The two intervals are very close to each other, despite some minor roundign differences.

Part D

Our confidence interval shows that if were were to collect many random samples and compute a confidence interval for each, then we would expect that about 95% of those intervals would contain the true difference in population proportions between males and females.

Part E

The standard error I calculated above represents the standard deviation of the sampling distribution. It is measuring the spread of the sampling distribution for the difference in proportion between males and females who folded their arms with left on top.

Part F

The sampling distribution represents the distribution of the difference in sample proportions between males and females who folded their left arm on top. From sample to sample, the difference in proportion and the confidence interval for each sample vary. The sample size and and true difference in proportion should stay the same.

Part G

The mathematical theorem that justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions is the Central Limit Theorem, which states that if the sample size is large enough, the sampling distribution of the difference in proportions will be normally distributed.

Part H

If the 95% confidence interval for the difference in proportions was [-0.01, 0.30], it means that it contains zero, indicating there is no statically significant difference between proportions of males and females who fold their left arm on top at the 95% confidence level.

Part I

If we repeat this experiment many times with different random samples of university students, the confidence intervals would indeed be different across samples due to random sampling. Each sample will yield a different difference in proportion. However, if we repeat the sampling process many times, 95% of all the confidence intervals would contain the true difference in proportion between males and females who fold their left arm on top.

Problem 2

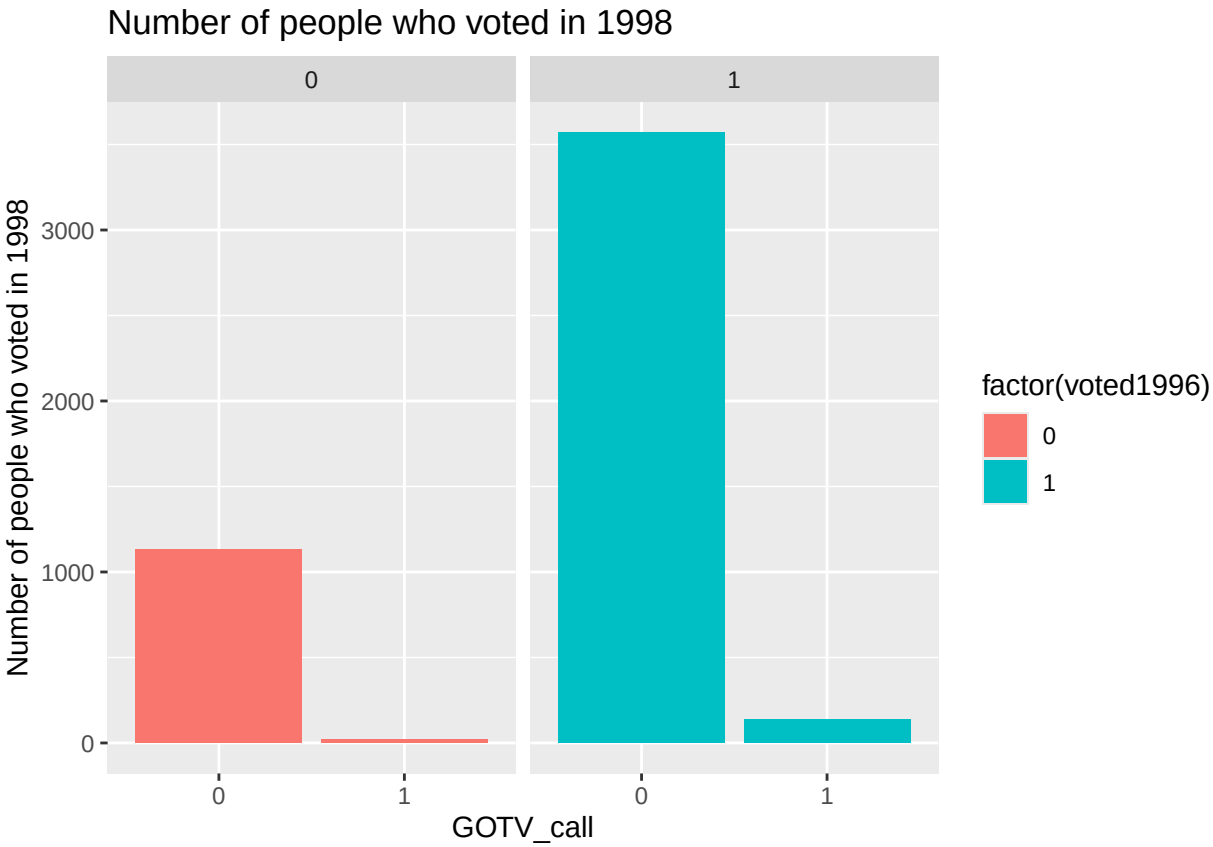
Part A

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

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(voted1998 ~ GOTV_call)
## X-squared = 39.597, df = 1, p-value = 3.122e-10
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2659167 -0.1411399
## sample estimates:
##    prop 1    prop 2
## 0.4442449 0.6477733
```

The sample proportion of those receiving a GOTV call who also voted in the 1998 Congressional election is 0.6477 while the sample proportion of those not receiving a GOTV call who voted in the 1998 Congressional Election is 0.4442. The 95% confidence interval is [-0.2659, -0.1411] in the proportion of voting in 1998 for those who received a GOTV call vs. who did not, using those receiveing a call as reference.

Part B



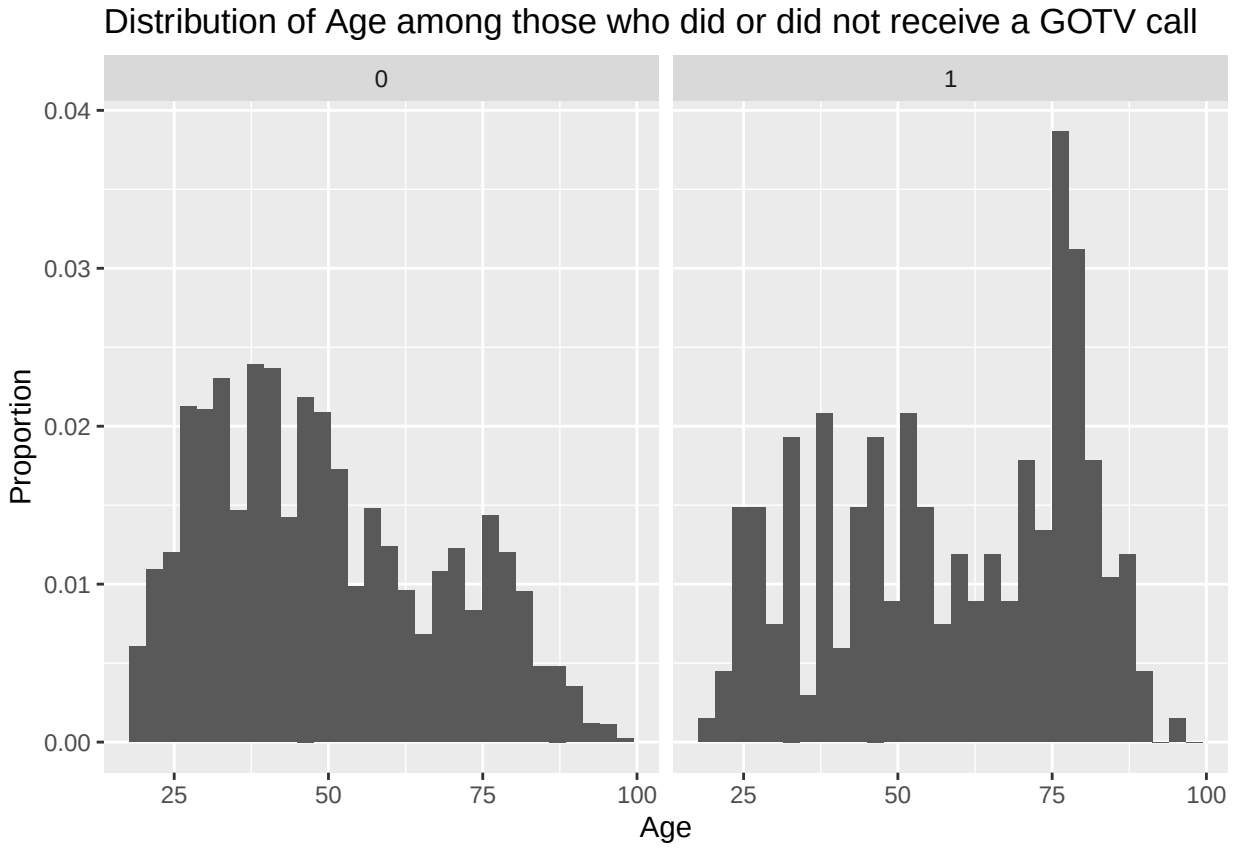

```
##   prop_1.0   prop_1.1
## 0.01409849 0.03038149
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(GOTV_call ~ voted1996)
## X-squared = 31.32, df = 1, p-value = 2.188e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.02195834 -0.01060767
## sample estimates:
##     prop 1     prop 2
## 0.01409849 0.03038149
```

According to the graphs, the proportion of receiving a call between people who voted in 1996 vs. those who did not seem different. The sample proportion of those who voted in 1996 who also received a GOTV call is 0.0304 while the sample proportion of those who did not vote in 1996 but received a GOTV call is 0.0141. The 95% confidence interval is [-0.220, -0.011] for the difference in proportion, which does not contain zero. Therefore, there is statistically significant difference between the proportions of receiving a GOTV call between those who voted in 1996 vs. those who did not.

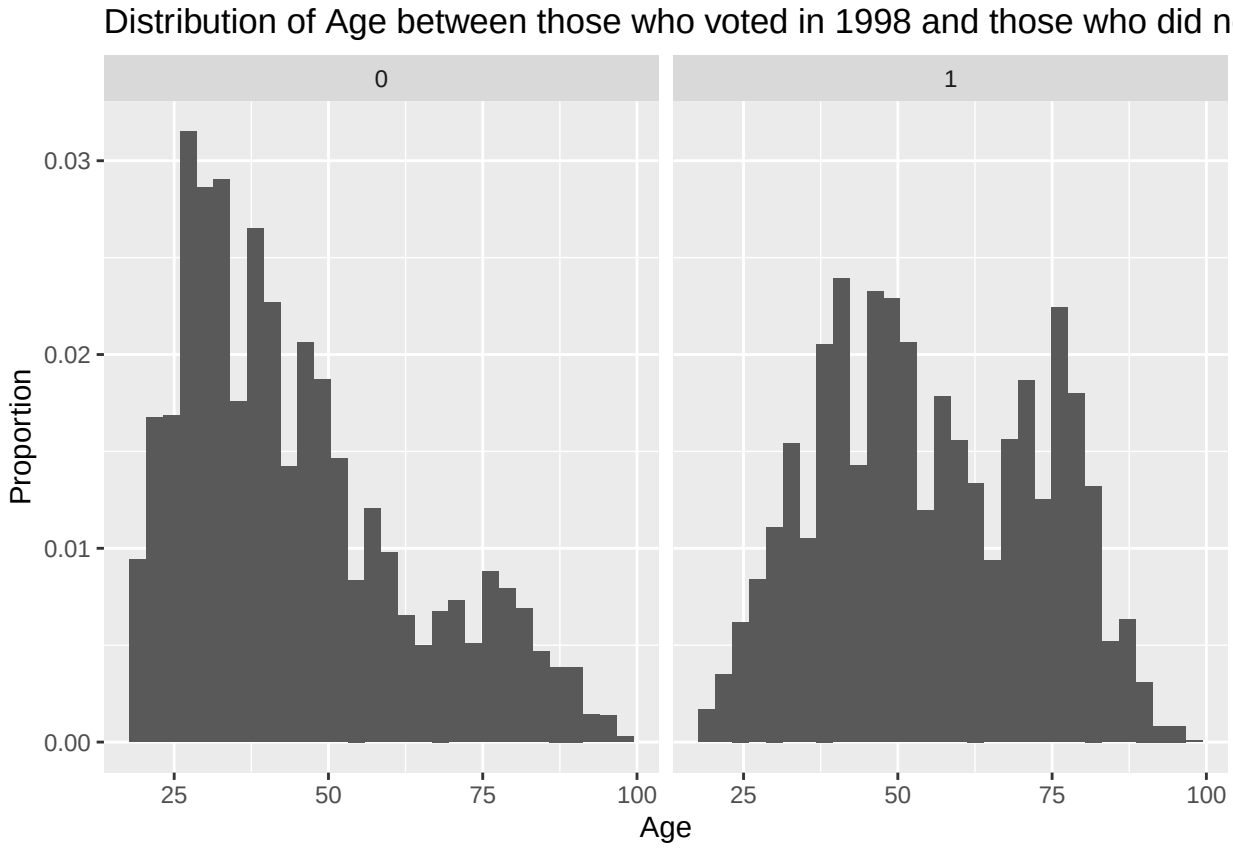
```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(voted1998 ~ voted1996)
## X-squared = 1832.4, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.4275349 -0.3932429
## sample estimates:
##    prop 1    prop 2
## 0.2293487 0.6397376
```

The sample proportion of those voting in 1996 who also voted in 1998 is 0.640 while the sample proportion of those who did not vote in 1996 but voted in 1998 is 0.229. The 95% confidence interval is [-0.428, -0.393] for the difference in proportion, which does not contain zero. Therefore, there is statistically significant difference between the proportions of voting in 1998 between those who voted in 1996 vs. those who did not. This indicates that voted1996 is a confounder.



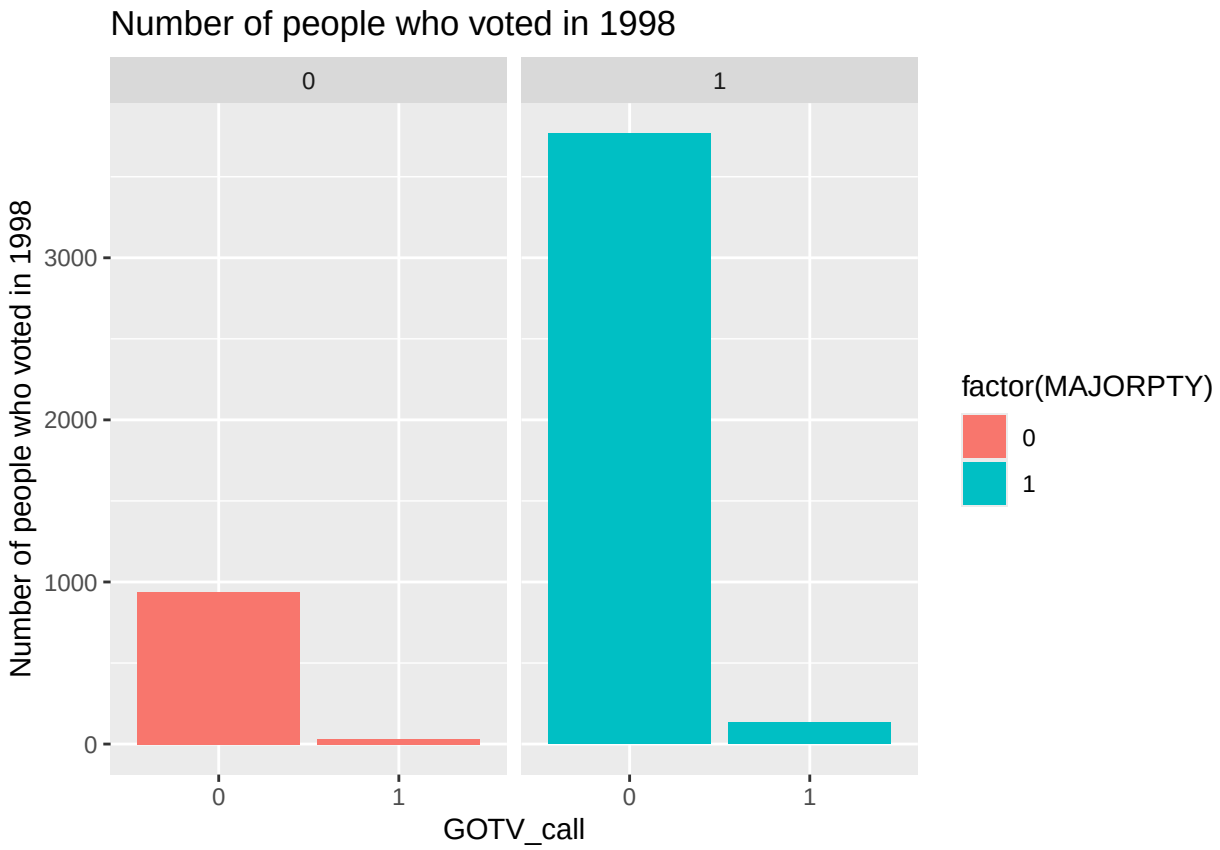

```
##
##  Welch Two Sample t-test
##
## data:  AGE by GOTV_call
## t = -6.9613, df = 256.33, p-value = 2.817e-11
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.395051  -6.369644
## sample estimates:
## mean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

Based on the graph, the age distribution between those who received a GOTV call and those who did not look different. Based on the calculations, the mean age of those who received a call is 58.3, and the mean age of those who did not is 49.4. Also, the confidence interval in difference of mean age is [-11.40, -6.37], which does not include 0 after the t-test of Age vs. GOTV_call. Therefore, there is statistically significant difference between the means of voting age between those who received a GOTV call vs. those who did not. This indicates that AGE is associated with GOTV_call.




```
##
##  Welch Two Sample t-test
##
## data:  AGE by voted1998
## t = -30.24, df = 10568, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -11.182008  -9.820602
## sample estimates:
## mean in group 0 mean in group 1
##        44.91404        55.41535
```

Based on the graph, the age distribution between those who voted in 1998 and those who did not look different. Based on the calculations, the mean age of those who voted in 1998 is 55.4, and the mean age of those who did not is 44.9. Also, the confidence interval in difference of mean age is [-11.18, -9.82], which does not include 0 after the t-test of Age vs. voted1998. Therefore, there is statistically significant difference between the difference in means of voting age between those who voted in 1998 vs. those who did not. This indicates that AGE is associated with voted1998 and AGE is a confounder.




```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(GOTV_call ~ MAJORPTY)
## X-squared = 3.8248, df = 1, p-value = 0.0505
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.0129180093 -0.0004615944
## sample estimates:
##     prop 1     prop 2
## 0.01781818 0.02450798
```

According to the graphs, the proportion of receiving a call between people who are affiliarted with a major party vs. those who are not seem different. The sample proportion of those who are affiliated with a major party who also received a GOTV call is 0.0245 while the sample proportion of those who are not but received a GOTV call is 0.0178. The 95% confidence interval is [-0.013, -0.0005] for the difference in proportion, which does not contain zero. Therefore, there is statistically significant difference between the proportions of receiving a GOTV call between those who are affiliated with a major political party vs. those who are not.

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(voted1998 ~ MAJORPTY)
## X-squared = 144.63, df = 1, p-value < 2.2e-16
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.1534422 -0.1111651
## sample estimates:
```

```
##    prop 1    prop 2
## 0.3501818 0.4824855
```

The sample proportion of those affilated with a major party who also voted in 1998 is 0.482 while the sample proportion of those who are not but voted in 1998 is 0.350. The 95% confidence interval is [-0.153, -0.111] for the difference in proportion, which does not contain zero. Therefore, there is statistically significant difference between the proportions of voting in 1998 between those who are affiliated with a major political party vs. those who are not. This indicates that MAJORPTY is a confounder.

Part C

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

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  tally(voted1998 ~ GOTV_call)
## X-squared = 4.9027, df = 1, p-value = 0.02682
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14663149 -0.01045353
## sample estimates:
##    prop 1    prop 2
## 0.5692308 0.6477733
```

After matching, the sample proportion of voting in 1998 for those who got a GOTV call is 0.6477 while the sample proportion of those who did not get a GOTV call is 0.5692. The confidence interval is [-0.1466, -0.010] for the difference in proportion, using those who got a call as reference.

In conclusion, the raw difference in sample proportion suggests a higher voting rate among those who received the GOTV call, and the confidence interval does not contain zero, so according to the matched data, those receiving GOTV call appears to have statistically significant higher likelihood of voting in the 1998 election. However, it only suggests that there is a significant association between receiving the GOTV call and voting in 1998, and we cannot infer whether there is a causal effect of the GOTV call on the likelihood of voting in 1998 election without doing a randomized control trial experiment.