Hw7

Pranita

2025-04-07

Name: Pranita Chaudhury

UT EID: pc28377

github: https://github.com/PranitaChau/Hw_7

Part 1

Part A

```
## # A tibble: 2 x 4
##   Sex    proportion standard_deviation num_rows
##   <chr>       <dbl>              <dbl>    <int>
## 1 Female      0.423              0.496      111
## 2 Male        0.472              0.502      106
```

There are 111 female students in the dataset and there are 106 male students in the dataset. The sample proportion of males who folder their left arm on top is 0.4716981 while the sample proportion of females who folded their left arm on top is 0.4234234.

Part B

```
##   diffprop
## 0.04827469
```

The observed difference in proportions between the two groups (males minus females) is 0.04827469.

Part C

```
##
##  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:
##  -0.09315879  0.18970817
## sample estimates:
##    prop 1    prop 2
## 0.5765766 0.5283019
```

```
## [1] -0.08723215
```

```
## [1] 0.1817489
```

We can be 95% confident that the true difference in proportions between males and females who folded their left arm on top is between -0.09315879 and 0.18970817. Since this confidence interval includes zero it suggests that there is no statistical significance between male and female students who cross their arms with their left arm being on top. The formula for standard error is sqrt((sd1^2 / n1) + (sd2^2 / n2)) with sd being standard deviation and n being the total amount of rows in that category (male or female). I used sd1 = 0.4963421, sd2 = 0.5015699, n1 = 106, n2 = 111. I used 2 for the z* value because I used a 95% confidence interval.

Part D

If we were to take many samples from this class where we asked a student to cross their hands and see if they put their left hand on top, then we would expect that 95% of the time the difference in proportions between males and females placing their left hand on top would be between -0.0932 and 0.1897, indicating no significant statistical difference.

Part E

The standard error is the variability of the sampling distribution. It tells us the variability between the proportions of the two groups and how off they are to the actual difference. It is measuring the difference between observed standard deviation of male and female students who folded their left arm on top compared to the reality of the true proportion when using different samples, and in this clas is 0.0052278 (male sd - female sd).

Part F

Sampling distribution is the distribution of differences in the proportions in males vs. females folding their left arm on top across different samples. In each sample the thing that would change is the difference in proportion since the proportions will change, but the sampling procedure will stay the same.

Part G

The central limit theorem shows that using a normal distribution is ok since the sample sizes are large so that the overall result will not be affected based on the skew or distribution of the original sample because the sampling distributions that are based on averages of large sample sizes which are independent. Since our sample size of both male and female students is big enough we can use normal distribution to approximate proportion differences.

Part H

Looking at the 95% confidence interval for the difference in proportions, which is -0.01 to 0.30, we cannot say for sure that there's no difference between males and females in arm folding. The fact that the interval includes 0 means it's possible there's no difference at all, but it also goes up to 0.30, which suggests that it could be more likely for females to fold their left arm on top. So, even though we can't completely rule out the idea of no difference, we also don't have enough evidence to say there's definitely a difference either.

Part I

Yes the confidence interval would be slightly difference across all samples. This is due to resampling and the fact that there may be more people of one sex who prefer their left arm. There will be slight changes based on chance. However, over many many trials of this sample being taken we can be 95% confident that around 95% of the confidence intervals will include the true population difference in proportions.

Part 2

Part A

```
## # A tibble: 2 x 4
##   GOTV_call proportion standard_deviation num_rows
##       <dbl>      <dbl>              <dbl>    <int>
## 1         0      0.444              0.497    10582
## 2         1      0.648              0.479      247
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  $ out of $new out of newvoted out of total
## 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 proportion of those receiving a GOTV call who voted in 1998 is 0.6477733 and the proportion of those not receiving a GOTV call who voted in 1998 is 0.4442449. We can be 95% confident that there is a -0.2659167 to -0.1411399 difference in the proportion of voters for those who did not receive a GOTV call compared to those who did.

Part B

```
## # A tibble: 2 x 4
##   GOTV_call proportion standard_deviation num_rows
##       <dbl>      <dbl>              <dbl>    <int>
## 1         0      0.531              0.499    10582
## 2         1      0.713              0.453      247
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  $ out of $govt_1996 out of govt_1996voted out of total
## X-squared = 31.32, df = 1, p-value = 2.188e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2410506 -0.1224366
## sample estimates:
##    prop 1    prop 2
## 0.5308070 0.7125506
```

```
## # A tibble: 2 x 4
##   voted1996 proportion standard_deviation num_rows
##       <dbl>      <dbl>              <dbl>    <int>
## 1         0      0.229              0.420     5036
## 2         1      0.640              0.480     5793
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  $ out of $both_years out of both_yearsvoted out of total
```

```
## 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
```

We can conclude that the variable voted1996 is a confounder because when we compare it to GOTV_call and voted1998, the confidence interval for the difference is between -0.2410506 to -0.1224366 for GOTV_call, and -0.4275349 to -0.3932429 for voted1998. Since the interval does not include zero this suggests a significant correlation, indicating that voted1996 (if the person voted in 1996) is a confounder variable between whether the person recieved a GOTV call and if they voted in 1998..

```
##
##  4-sample test for equality of proportions without continuity correction
##
## data:  * out of $govt_voted1998_age$proportion_voted1998 out of govt_voted1998_agegovt_voted1998_age
## X-squared = 649.5, df = 3, p-value < 2.2e-16
## alternative hypothesis: two.sided
## sample estimates:
##    prop 1    prop 2    prop 3    prop 4
## 0.1878659 0.5185185 0.4995978 0.6636364
```

```
## # A tibble: 4 x 6
## # Groups:   age_group [2]
##   age_group   GOTV_call proportion_voted1998 num_rows ci_lower ci_upper
##   <chr>           <dbl>                <dbl>    <int>    <dbl>    <dbl>
## 1 30 or Under         0                0.188     1879    0.170    0.206
## 2 30 or Under         1                0.519       27    0.330    0.707
## 3 Over 30             0                0.500     8703    0.489    0.510
## 4 Over 30             1                0.664      220    0.601    0.726
```

We can conclude that AGE is a confounder because when we compare it to GOTV_call and voted1998, the confidence interval for the difference is between 0.1702043 to 0.2055275 for those 30 or under who did not get a GOTV call, the difference for those 30 or under who did get a GOTV call is 0.3300468 to 0.7069902. The difference for those over 30 who did not get a GOTV call is 0.4890930 to 0.5101027, and for those over 30 who did get a GOTV call the difference is 0.6012033 to 0.7260694. Since none of the intervals include zero this suggests a correlation, indicating that AGE is a confounder variable between whether the person recieved a GOTV call and if they voted in 1998..

```
## # A tibble: 2 x 4
##   GOTV_call proportion standard_deviation num_rows
##       <dbl>      <dbl>              <dbl>    <int>
## 1         0      0.745              0.436    10582
## 2         1      0.802              0.400      247
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  $ out of $govt_party out of govt_partyvoted out of total
## X-squared = 3.8248, df = 1, p-value = 0.0505
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.109356458 -0.004371919
## sample estimates:
##    prop 1    prop 2
```

```
## 0.7447552 0.8016194
```

```
## # A tibble: 2 x 4
##   MAJORPTY proportion standard_deviation num_rows
##      <dbl>      <dbl>              <dbl>    <int>
## 1        0      0.350              0.477     2750
## 2        1      0.482              0.500     8079
```

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  $ out of $party_vote out of party_votevoted out of total
## 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
```

We can conclude that MAJORPTY is a confounder because when we compare it to GOTV_call and voted1998, the confidence interval for the difference is between -0.109356458 to -0.004371919 for GOTV_call, and -0.1534422 to -0.1111651 for voted1998. Since the interval does not include zero this suggests a significant correlation, indicating that MAJORPTY is a confounder variable between whether the person recieved a GOTV call and if they voted in 1998..

Therefore we can conclude that the voted1996, AGE, and MAJORPTY variables are all confounder variables since none of the confidence intervals of those variables compared to the GOTV_call and voted1998 variables included zero in their interval.

Part C

```
##
## Call:
## matchit(formula = GOTV_call ~ voted1996 + AGE + MAJORPTY, data = turnout,
##     ratio = 5)
##
## Summary of Balance for All Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.0297        0.0226          0.5130     1.3026    0.1572
## voted1996        0.7126        0.5308          0.4016          .    0.1817
## AGE             58.3077       49.4253          0.4475     1.1228    0.1114
## MAJORPTY         0.8016        0.7448          0.1426          .    0.0569
##           eCDF Max
## distance    0.2499
## voted1996   0.1817
## AGE         0.2229
## MAJORPTY    0.0569
##
## Summary of Balance for Matched Data:
##           Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean
## distance         0.0297        0.0297          0.0001      1.004    0.0000
## voted1996        0.7126        0.7126         -0.0000          .    0.0000
## AGE             58.3077       58.2664          0.0021      1.008    0.0006
## MAJORPTY         0.8016        0.8073         -0.0142          .    0.0057
##           eCDF Max Std. Pair Dist.
```

```
## distance    0.0057          0.0001
## voted1996   0.0000          0.0000
## AGE         0.0057          0.0027
## MAJORPTY    0.0057          0.0183
##
## Sample Sizes:
##           Control Treated
## All         10582     247
## Matched      1235     247
## Unmatched    9347       0
## Discarded       0       0
```

```
## # A tibble: 1,482 x 10
##    voted1998 GOTV_call voted1996 PERSONS   AGE MAJORPTY age_group   distance
##        <dbl>     <dbl>     <dbl>   <dbl> <dbl>    <dbl> <chr>          <dbl>
##  1         1         1         1       2    74        1 Over 30       0.0427
##  2         0         1         1       2    37        0 Over 30       0.0180
##  3         1         1         1       2    24        1 30 or Under   0.0160
##  4         1         1         1       2    81        1 Over 30       0.0488
##  5         0         1         0       2    34        1 Over 30       0.0113
##  6         1         1         0       2    54        1 Over 30       0.0167
##  7         1         1         1       2    65        1 Over 30       0.0359
##  8         1         1         0       2    43        0 Over 30       0.0117
##  9         1         1         1       2    52        1 Over 30       0.0278
## 10         1         1         1       2    44        1 Over 30       0.0238
## # i 1,472 more rows
## # i 2 more variables: weights <dbl>, subclass <fct>
```

```
## [1] 0.6477733
```

```
## [1] 0.5692308
```

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

The proportion of people that got a GOVT call and voted in 1998 is 0.6477733, while the proportion of people that did not get a GOVT call and voted in 1998 is 0.5692308.

We can be 95% confident that the true difference in proportions between those who got a GOVT call and voted in 1998 compared to those who did not receive a GOVT call and voted in 1998 in between 0.01045353 and 0.14663149. This shows that getting a GOTV call has a statistically significant effect on the likelihood of voting in 1998 since the voted1996, AGE, and MAJORPTY are no longer confounders and the interval does not include zero.