Homework7

Kapil Taspa

2025-04-07

UT EID: kt27955

GitHub: https://github.com/ktaspa/Homework7SDS315

Problem 1: Armfolding

A

The number of male and female students in the dataset.

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

The sample proportion of males who folded their left arm on top.

```
## [1] 0.4716981
```

The sample proportion of females who folded their left arm on top.

```
## [1] 0.4234234
```

B

What is the observed difference in proportions between the two groups (males minus females)?

```
## [1] 0.04827469
```

C

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(armfold$LonR_fold[armfold$Sex == "Male"]), sum(armfold$LonR_fold[armfold$Sex == "Female
## 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
```

The formula for standard error is sqrt ((p1 * (1 - p1)) / n1 + (p2 * (1 - p2)) / n2) p1 is the proportion of males with left hand on top (0.47) p2 is the proportion of females with left hand on top (0.42) n1 is the number of males (106) n2 is the number of females (111) With these values we get 0.067 for the SE

the value of z* for this should be 1.96 because we are using the 95% confidence level To get the lower bound we do p1 - p2 - z* x SE = -0.0839 To get the upper bound we do p1 - p2 - z* x SE = 0.1801

D

With about 95% certainty we can say that the true true difference in proportions between males and females folding their left arm on top is between -0.0839 and 0.1801.

E

The standard error represents the variance that we expect from sample to sample ie. how much the difference in sample proportions may be different in other samples.

F

The sampling distribution is the distribution of the difference in sample proportions (male minus female). The true population proportions stay fixed, while the sample proportions vary from one sample to the next.

G

The Central Limit Theorem- under specific conditions outlined by the theorem the distribution of the sample statistic will be approximately normal

H

The confidence interval has 0 in it, but we can't reject the claim that there is no difference in arm folding because it could be 0.

I

Yes, the confidence interval would likely be different when using different samples because there will be different people in the samples. There is a natural variation that occurs in those samples, but if we repeat this with many samples about 95% of the intervals we calculate would contain the true population difference in proportions.

Problem 2

A

We see that there is a difference in the proportions with 95% certainty because the confidence interval is all positive and does not contain 0. So, somewhere between 0.143 and 0.264.

```
## Rows: 10829 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## dbl (6): voted1998, GOTV_call, voted1996, PERSONS, AGE, MAJORPTY
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

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

```
## [1] 0.4442449
```

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

```
##
## data:  c(sum(turnout$voted1998[turnout$GOTV_call == 1]), sum(turnout$voted1998[turnout$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
```

B

The variables are confounders based on the table below. GOTV call recipients were more likely to have voted in 1996 (0.71 vs 0.53), were older (58.3 vs 49.4), and were more likely to be registered with a major party (0.80 vs 0.74).

```
## # A tibble: 2 x 4
##   GOTV_call voted1996   AGE MAJORPTY
##       <dbl>     <dbl> <dbl>    <dbl>
## 1         0     0.531  49.4    0.745
## 2         1     0.713  58.3    0.802
```

We can confirm this with the confidence intervals at the 95% level for each of these variables:

voted1996: does not include 0

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(turnout$voted1996[turnout$GOTV_call == 1]), sum(turnout$voted1996[turnout$GOTV_call == 0]
## 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
```

MAJORPTY: does not include 0

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(turnout$MAJORPTY[turnout$GOTV_call == 1]), sum(turnout$MAJORPTY[turnout$GOTV_call == 0]]
## 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
```

AGE: does not include 0

```
##                 2.5 %   97.5 %
## (Intercept) 49.067914 49.78278
```

```
## GOTV_call    6.515674 11.24902
```

Because all 3 of these intervals do not contain 0 we can say with 95% certainty there is a difference in the sample proportions and these variables are all confounders.

C

The first table below checks that the matched dataset is still balanced.

```
## # A tibble: 2 x 4
##   GOTV_call voted1996   AGE MAJORPTY
##       <dbl>     <dbl> <dbl>    <dbl>
## 1         0     0.713  58.3    0.807
## 2         1     0.713  58.3    0.802
```

Now we can perform the confidence intervals at 95% level for each of these variables to confirm that they are balanced on those confounders. ### voted1996: includes 0

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(matched_data$voted1996[matched_data$GOTV_call == 1]), sum(matched_data$voted1996[matche
## 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
```

MAJORPTY: includes 0

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(matched_data$MAJORPTY[matched_data$GOTV_call == 1]), sum(matched_data$MAJORPTY[matched_
## 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
```

AGE: includes 0

```
##                 2.5 %    97.5 %
## (Intercept) 57.162215 59.370579
## GOTV_call   -2.663386  2.745978
```

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. Now, we use the matched dataset and repeat the analysis from part A

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

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

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  c(sum(matched_data$voted1998[matched_data$GOTV_call == 1]), sum(matched_data$voted1998[matche
## 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 see in the matched data that the confidence interval is still positive and does not contain 0 so we can conclude wit 95% certainty that the GOTV call likely had a real impact on turnout in the 1998 election. We are 95% confident that the true difference in voting rates in the population lies between 1.29 and 14.42 percentage points.