Homework 7

Pranav Belligundu(psb898) - SDS 315 UT Austin

https://github.com/pranav-B21/SDS-315/tree/main/HW7

Problem 1: Armfolding

A. Load and examine the data

The number of male and female students in the dataset: male = 106, female = 111 The sample proportion of males who folded their left arm on top: 47.2% The sample proportion of females who folded their left arm on top: 42.3%

B. Observed Difference

The observed difference is in proportion between males and females is 4.83%

C. Compute a 95% confidence interval

```
##
##  2-sample test for equality of proportions with continuity correction
##
## data:  x out of n
## 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
## 0.4716981 0.4234234
```

The SE formula is sqrt((p1(1-p1))/n1 + (p2(1-p2))/n2).

Given the values above that we have calculated, we can input them into the equations: sqrt((0.472(1-0.472))/106 + (0.423(1-0.423))/111) = 0.0675

We are choosing z* = 1.96 because of the 95% CI.

Now to calculate the CI we need to: (diff) ± z* times SE = (0.0483) ± 1.96 * 0.0675 = 0.049 ± 0.1323 = (−0.0833,0.1813).

The prop.test from the R function gives us a CI of (-0.0932, 0.1897)

Both the CI's that have been calculated from the formula and the R function give around the same result with some small differences.

D. Interpret your confidence interval

If we were to repeat this sampling process many times, then we would expect that 95% of the resulting confidence intervals would contain the true difference in proportions of males and females who fold their left arm on top is between –8.3% and 18.1%.

The CI includes 0, which means we don't have strong evidence of a real difference between male and female arm-folding preferences in the population.

E. What does the standard error you calculated above represent? What is it measuring

The standard error is measuring the amount in variability in the difference of prop between males and females putting their left hand over their right hand completely due to random sampling and chance.

F. What does the term sampling distribution refer to in this context?

In this context, the sampling distribution refers to the distribution of differences in sample proportions (male minus female who fold their left arm on top) that we would get if we repeatedly took random samples of male and female students from the population.

G. What mathematical result or theorem justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions?

The Central Limit Theorem justifies using a normal distribution to approximate the sampling distribution of the difference in sample proportions.

H. Suppose your 95% confidence interval for the difference in proportions was [-0.01, 0.30]. Based on this, what would you say to someone who claims "there's no sex difference in arm folding"?

While the interval includes zero, which means we cannot rule out the possibility of no difference, it also includes many positive values suggesting there could be a meaningful difference, with males possibly more likely to fold their left arm on top. So, we can't confidently claim there is a sex difference, but we also can't confidently say there isn't one.

I. Imagine repeating this experiment many times with different random samples of university students. Would the confidence interval be different across samples? Why? What should be true about the collection of all those intervals?

Yes, confidence intervals would vary from sample to sample because each random sample would produce slightly different proportions, just due to natural sampling variability.

Problem 2: Get out the vote

A. How much more likely are GOTV call recipients to have voted in 1998?

The proportion of those receiving a GOTV call who voted in 1998: 64.8% The sample proportion of those not receiving a GOTV call who voted in 1998: 44.4% A large-sample 95% confidence interval for the difference in these two proportions: 20.4%

B. Prove that voted1996, AGE, and MAJORPTY are confoudners

```
##
##        0    1
##   0 4965 5617
##   1   71  176
```

```
##
##             0         1
##   0 0.4691930 0.5308070
##   1 0.2874494 0.7125506
```

```
##
##        0    1
##   0 3881 1155
##   1 2087 3706
```

```
##
##             0         1
##   0 0.7706513 0.2293487
##   1 0.3602624 0.6397376
```

Among those who did NOT receive a GOTV call (GOTV_call = 0), 53.1% had voted in 1996. Among those who DID receive a GOTV call (GOTV_call = 1), 71.3% had voted in 1996.

Among people who did NOT vote in 1996, only 22.9% voted in 1998. Among people who DID vote in 1996, 63.9% voted in 1998.

Since voting in 1996 is associated with the treatment, GOTV_call, and voting in 1996 is associated with the outcome of voting in 1998, voted1996 is a **confounder**.

```
##        0        1
## 49.42534 58.30769
```

```
##        0        1
## 18.73161 19.84806
```

```
##
##  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
```

```
##        0        1
## 44.91404 55.41535
```

```
##        0        1
## 18.45671 17.57087
```

```
##
##  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
```

Individuals who received a GOTV call were significantly older on average (mean age = 58.3) than those who did not (mean age = 49.4), with a statistically significant difference (95% CI: -11.40 to -6.37, $p < 0.00001$).

Those who voted in 1998 were also significantly older on average (mean age = 55.4) compared to those who did not vote (mean age = 44.9), with a strong statistical difference (95% CI: -11.18 to -9.82, $p < 0.00001$).

Since age is associated with both receiving a GOTV call and the likelihood of voting in 1998, it is a **confounder**

```
##
##        0    1
##   0 2701 7881
##   1   49  198
```

```
##
##             0         1
##   0 0.2552448 0.7447552
##   1 0.1983806 0.8016194
```

```
##
##        0    1
##   0 1787  963
##   1 4181 3898
```

```
##
##             0         1
##   0 0.6498182 0.3501818
##   1 0.5175145 0.4824855
```

Among those who did NOT receive a GOTV call, 74.5% were affiliated with a major party. Among those who DID receive a GOTV call, 80.2% were affiliated with a major party. This shows that major party affiliation increases the likelihood of receiving a GOTV call.

Among those not affiliated with a major party, only 35% voted in 1998. Among major party members, 48.2% voted in 1998. This shows that major party affiliation is associated with a higher likelihood of voting.

Because MAJORPTY is related to both treatment assignment and voting behavior, it is a **confounder**.

B. Matching

The proportion of those receiving a GOTV call who voted in 1998: 64.8% The sample proportion of those not receiving a GOTV call who voted in 1998: 56.9% A large-sample 95% confidence interval for the difference in these two proportions: 7.9%

After matching, there is evidence that receiving a GOTV call had a positive effect on the likelihood of voting in the 1998 Congressional election. Among individuals who received a GOTV call, 64.8% voted, compared to 56.9% of those who did not receive a call. This shows an estimated treatment effect of a 7.9 percentage point increase in voter turnout. Additionally, a large 95% confidence interval(1.3% to 14.4%), indicates that the effect is statistically significant.

Overall, these results suggest that the GOTV campaign was effective at increasing turnout, and this conclusion is more credible than the unadjusted estimate because it accounts for key confounding variables through matching.