Homework7_SDS 315

Nancy Nakyung Kwak (UT EID: nk24424)
GitHub link: https://github.com/nancy1404/sds315_hw7.git

2025-04-03

Contents

PROBLEM 1: Armfolding

a. Load and Examine the data

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

```
##
##         Female      Male
##   0 0.5765766 0.5283019
##   1 0.4234234 0.4716981
```

1-a summary

- There are 111 females and 106 males in the dataset.
- Among females, 42.3% folded their arms with left arm on top.
- Among males, 47.2% folded their arms with the left arm on top. These are out sample proportions for each group.

b. Observed Difference Between Two Groups

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

1-b summary

- The observed difference in sample proportions (male-female) is approximately 0.048, which means the proportion of males who folded their left arm on top is about 4.8% points higher than those of females who did so.

c. Confidence Interval

```
## [1] 50 47
```

```
## [1] 106 111
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## 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-c summary

- 1. R's built-in calculation
  - using prop.test() with correct = FALSE, it gave the proportion of males with left on top: 0.4717 and proportion of femals with left on top: 0.4234. Observed difference was (male-female) 0.0483.
  - The 95% confidence interval was [-0.0839, 0.1805].
- 2. Hand-calculated version

- The standard error is $SE = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$.
- Where, $\hat{p}_1 = 50/106 = 0.4717$, $n_1 = 106$ (males) and $\hat{p}_2 = 47/111 = 0.4234$, $n_2 = 111$ (females). When plugged the values directly into the SE formula, $SE = \sqrt{\frac{0.4717(1-0.4717)}{106} + \frac{0.4234(1-0.4234)}{111}} = 0.0679$.
- We use a z* value of 1.96 for a 95% confidence interval. We use z* = 1.96 because it corresponds to a 95% confidence level under the standard normal distribution. This means that about 95% of intervals calculated this way will contain the true difference in proportions. We can use the normal distribution here because the sample sizes are large, so the Central Limit Theorem applies. Using this, we get, 0.0483±1.96×0.0679 => [-0.0839, 0.1805], which matches the confidence interval from R's prop.test() function up to minor rounding.

d. Interpret Confidence Interval

- "If we were to **repeat this arm-folding survey with many new random samples of students from the sample population**, then we would expect that **about 95% of the confidence intervals calculated this way would contain the true difference in population proportions (males minus females) who fold their left arm on top**."

e. Standard Error?

- The standard error tells us how much the difference between he male and female proportions might bounce around if we kept collecting new random samples from the same group. In this case, it gives us a sense of how much the left-arm-on-top differene could vary just due to random chance. If the standard error is small, that means our estimate is pretty stable across samples; if it's bigger, there's more wiggle room because of sample-to-sample randomness.

f. Sampling Distribution?

- The sampling distribution here refers to the range of possible values we might see for the difference in sample proportions (left-on-top: males minus females) if we repeated this arm-folding survey over and over again with different random groups of students from the sample population.
- What changes from sample to sample are the actual sample proportions for males and females – because different students might be selected each time, and their folding preferences could vary a bit just by chance.
- What stays the same is the true population proportions for males and females – those are fixed (even though we don't know exactly what they are), and we're trying to learn about them using our samples. The sampling distribution is basically showing how much the difference we observe in our data might vary just due to random sampling.

g. Normal Distribution Justification

- The reason we're allowed to use the normal distribution to approximate the sampling distribution of the difference in sample proportions is because of the **Central Limit Theorem (CLT)**. This theorem tells us that when we take a large enough sample, the distribution of the sample proportion (or difference in proportions) will look roughly normal – even if the original data is not. That means the shape of all the possible sample differences (from lots of repeated random samples) will follow the bell curve.
- In this case, both male and female sample sizes are large enough, so the CLT kicks in and lets us use the normal distribution to build our CI.

h. No Sex Difference in Arm Folding?

- It is true that the confidence interval does include zero (which shows the possibility of no difference still exists), but the interval also has a range of positive values. This could suggest that there might be a real difference in the population – we just don't have strong enough evidence to be so sure.

- Therefore, I don't fully agree or disagree with the claim since there is not enough statistical evidence to confidently reject the hypothesis, but we also cannot confidently say there IS no difference. More data might help to be confident on this decision.

i. Differences in CI?

- If we repeat this experiment many times with different random samples, the confidence interval would definitely change from sample to sample. It's because each random sample will give slightly different proportions, which leads to a different estimate of the difference and a different confidence interval.
- But if we repeated this process over and over, calculating the 95% confidence interval each time, then about 95% of those intervals would contin the true population difference in proportions. That is what the 95% confidence level really shows, not what we are 95% sure this interval is correct, but that method we are using gives right answer most of time in long run.

PROBLEM 2: Get out the vote

PART A

```
##         0         1
## 0.4442449 0.6477733
```

```
##  diffmean
## 0.2035283
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  x out of n
## 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
```

2-A summary

- Among those who received a GOTV call, 64.78% voted in the 1998 election. Among those who did not receive a call, 44.42% voted. The observed difference in proportions (GOTV minus non-GOTV) was approximately 0.2035. We used a large-sample method for computing this CI, assuming the sampling distribution of the difference in proportions is approximately normal due to the fairly large sample sizes. A 95% confidence interval for the difference in proportions is [0.143, 0.264], based on prop.test() in R. Since the entire confidence interval is above 0, we have fairly strong evidence that receiving a GOTV call was associated with a higher voting rate in 1998.

PART B

```
##         0         1
## 0.5308070 0.7125506
```

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

```
##         0         1
## 0.7447552 0.8016194
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of colSumstable(turnout$voted1996, turnout$GOTV_call) out of table(turnout$voted1996, tu
## X-squared = 32.047, df = 1, p-value = 1.505e-08
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.2389791 -0.1245081
## sample estimates:
##    prop 1    prop 2
## 0.5308070 0.7125506
```

```
##
##  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
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of colSumstable(turnout$MAJORPTY, turnout$GOTV_call) out of table(turnout$MAJORPTY, turn
## X-squared = 4.1195, df = 1, p-value = 0.04239
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.107284916 -0.006443461
## sample estimates:
##    prop 1    prop 2
## 0.7447552 0.8016194
```

2-B summary

- We looked at three possible confounders: prior voting (voted1996), age (AGE) and party registration (MARJORITY).
- **voted1996**
  - for the proportion who voted in 1996, GOTV == 1 was 71.26% and GOTV ==0 was 53.08%. The 95% CI for difference in proportions was [-0.2390, -0.1245].
  - People who received a GOTV call *were a lot more likely to have voted* in 1996, and this difference is statistically significant.
- **AGE**
  - Mean age was GOTV ==1 for 58.31 years and GOTV==0 for 49.43 years. The 95% CI for difference in means was [-11.40, -6.37].
  - People who received a GOTV call *were older on average,* which is statistically clear and signficant difference in gap.
- **MAJORPTY**
  - Proportion registered with the major party was 80.16% for GOTV == 1 and 74.48% for GOTV == 0. The 95% CI for difference in proportions was [-0.1073, -0.0064].
  - People who received a GOTV call *were more likely to be registered with a major party*, and the difference is statistically significant here as well.
- **All three variables turned out to be significantly differ across treatment groups and are more likely to be related to both receiving a GOTV call and with the possibility of voting in 1998, which accordingly makes these variables confounders. Since the significant imbalance exists, the observed effect described in PART A CANNOT be interpreted as causal without adjustment.**

PART C

```
##         0         1
## 0.7125506 0.7125506
```

```
##        0        1
## 58.26640 58.30769
```

```
##         0         1
## 0.8072874 0.8016194
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of colSumstable(matched_data$voted1996, matched_data$GOTV_call) out of table(matched_dat
## 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
```

```
##
##  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
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of colSumstable(matched_data$MAJORPTY, matched_data$GOTV_call) out of table(matched_data
## X-squared = 0.042347, df = 1, p-value = 0.837
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04871171  0.06004775
## sample estimates:
##    prop 1    prop 2
## 0.8072874 0.8016194
```

```
##         0         1
## 0.5692308 0.6477733
```

```
##
##  2-sample test for equality of proportions without continuity correction
##
## data:  [ out of colSumstable(matched_data$voted1998, matched_data$GOTV_call) out of table(matched_dat
## X-squared = 5.2206, df = 1, p-value = 0.02232
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.14420234 -0.01288268
## sample estimates:
##    prop 1    prop 2
## 0.5692308 0.6477733
```

2-C summary

- Here I used matching to match each treated case (those who got GOTV call) with 5 control cases (those who didn't) to adjust these thre confounders: voted1996, AGE, and MAJORPTY. After matching, balance of the confounders between the treatment and control groups are checked. Not the groups results were similar where

- the proportion of those who voted in 1996 was 71.3% for both groups.
- the average age was 58.31 year for those who received the call and 58.27 for those who did not.
- the proportion registered with major party was 80.2% for those who received the call and 80.7% for those who did not.

- The differenes were small for all three of these cases and the 95% CI included zero which shows that matching now successfully balanced the groups.
- After that, looking at how many people voted in 1998 using ONLY the matched_data, among those who received a GOTV call, 64.8% voted, compared to 56.9% of those who did not.
- The difference in proportions was statistically significant, with a 95% CI of [-0.144, -0.013] which also does not including zero.
- Even after adjusting for confounding variables, receiving a GOTV call had a statistically significant and positive effect on the likelihood of voting in the 1998 election. Because the matching process balanced key confounders like prior voting history, age, and party registration, this provides stronger evidence that the difference in turnout was caused by the GOTV call, rather than underlying differences between groups.