Statistical Thinking HW 4

Estrella Fuentes eaf2758

2025-04-02

Contents

My GitHub Link -

Click Here

Problem 1: Armfolding

Our task for the first problem is to find if there was a difference in proportions between the proportion of men who fold their left arm on top versus the the proportion of females who fold their left arm on top. The study consisted of 217 participants, 111 being female, and 106 being male. The proportion of males who folded their left arm over their right was 47.2% and for females was 42.3%, the difference between their proportions was 4.8%. Using the prop.test built in R function, I got a 95% confidence interval of -0.094 and 0.192.

Going back in time before we had R, old statisticians used math, and in this case would use De Moivre's equation for the standard error for the difference in proportions. De Moivre's equation is the square root of the proportion of group one, multiplied by one minus the proportion of group one, divided by the sample size of group one, plus the proportion of group two, multiplied by one minus the proportion of group two, divided by the sample size of group two. The values I used in the equation were 0.4716981 which was the proportion of group 1 and 106 the sample size, and 0.4234234 the proportion of group 2 and 111 the sample size. To calculate the confidence intervals I used a z value of 1.96 because this would give me a 95% confidence interval, and through the "hand-calculated" confidence interval I got -0.084 to 0.180. If we were to conduct the study multiple times, then we would expect the true difference in proportion between the male proportion who fold their left over their right arm and the females to fall within the interval -0.094 and 0.192 in 95% of the studies. The standard error I calculated above measures how much we can expect the difference between the two sample proportions to vary from one sample population to another. In this case sampling distribution is the statistic of difference in proportions we would obtained if we sampled out population many times. From sample to sample the males and females drawn from the population would vary, resulting in different difference in proportions. The Central Limit Theorem justifies using a normal distribution because the theorem states that with a large enough sample all sampling distributions look the same and will become approximately normal.

If someone told me "there is no difference in arm folding" after I had showed them my confidence interval of -0.94 and 0.192, I would start by saying since the confidence includes 0 meaning their could be no difference between between sex difference in arm folding but we cant say it with 100% certainty their no real or consistent difference. If the experiment were to be repeated across many different random samples of university students, the results would be many different confidence intervals because each sample would have different proportions, leading to different confidence intervals. What is true about the collection of all those intervals is that 95% of those confidence intervals would contain the true population difference in arm folding behavior between males and females.

Problem 2: Get Out The Vote

Looking at Confounders

Age as a Confounder

The regression analysis shows that receiving a GOTV call significantly increases the likelihood of voting in 1998, with a 95% confidence interval for the coefficient ranging from 0.0786 to 0.1993. Additionally, age is positively associated with voting in 1998, with a 95% confidence interval for the age coefficient ranging from 0.0068 to 0.0077, indicating that older individuals are slightly more likely to vote. To determine whether age was a confounder I used a regression model to intent to make a fair comparison between whether age or GOTV impacted people who voted. The results showed that the coefficient of GOTV call is slightly higher in the model without age (0.2035), meaning that without accounting for age, the effect of GOTV call on voting is stronger. In comparison, the regression model with age has a significantly higher r-squared (of 0.0789) meaning age contributes to voting behavior. Finally we can assume age is confounder because the coefficient suggest that every one year increase in age, the likelihood in voting increases by 0.0727%, this suggest is an important factor in predicting voting behavior and is confounding the relationship between the GOTV call and voting in 1998.

Voted 1996 as a Confounder

To test whether voting in the 1996 election is a confounder, we examined whether voting in 1996 is related to receiving a GOTV call and voting in 1998. The results of the Chi-Squared test showed a p-value of 2.188e-08 between GOTV call and voting in 1996, indicating a significant association between whether a person received a GOTV call and whether they voted in 1996. In other words, people who voted in 1996 were more likely to receive a GOTV call. Additionally, the extremely small p-value of 2.2e-16 indicated a very strong and significant association between voting in 1996 and voting in 1998. The regression analysis further confirmed that voting in 1996 strongly predicts voting in 1998 95% confidence interval of 0.3912 to 0.4254, these results suggest that voting in 1996 is indeed a confounder in the relationship between receiving a GOTV call and voting in 1998.

Major Political Party as Confounder

Similar to the techniques I used above, such as regression models and the chi-squared test, I found that being part of a Major Political Party is a confounder that affects whether you get a call and whether a person is more likely to vote. First, I wrote some code to perform a chi-squared test, and the result showed a strong and significant relationship between voting and being associated with a major political party, which we can assume due to the extremely small p-value of 2.2e-16. Secondly, to check for association with the GOTV call, I noticed that the relationship between the two was very slight because of the p-value of 0.0505. When looking at the coefficient of the linear regression model of GOTV call and voted 1998, it was 0.204 and only decreased slightly to a coefficient of 0.196 when I looked at the regression model adjusted by MAJORPTY.

Confounders now Insignificant

Above we saw that AGE, voted1996, and MAJORPTY were confounders of the independent variables and dependent variables. After matching the goal is to eliminate the variables as confounding variable to GOTV call in order to determine whether the GOTV call was significantly important to people going out and voting. In order to determine whether the variables are still confounder or not, i'm going to be comparing their chi-squared p-value before and after matching.

1. Age Confounder

- PVALUE BEFORE MATCHING: 7.132e-07
- PVALUE AFTER MATCHING: 1

2. Voted 1996 Confounder

- PVALUE BEFORE MATCHING: 2.188e-08
- PVALUE AFTER MATCHING: 1

3. Major Part Association Confounder

- PVALUE BEFORE MATCHING: 0.0505
- PVALUE AFTER MATCHING: 0.9064

As seen above all off the pvalues increased significantly demonstrating their is no or little to no relationship between GOTV call and the variables.

Conclusion

Finally, the big question we have been trying to answer is whether receiving a GOTV call influenced whether people went out to vote. In the unmatched dataset, the results showed a confidence interval of 0.0123 to 0.0244, meaning the proportion of people who received a call and voted was likely to be between 1.23% and 2.44% higher than those who did not receive a call. This suggested a significant relationship between the GOTV call and voter turnout. However, after matching the data to control for confounders, the confidence interval for the difference in proportions widened to 0.0058 to 0.0839. This new confidence interval indicates that the difference in voting between those who received a call and those who did not is now between 0.58% and 8.39% higher for the treated group. While the effect is still significant in the matched data, the wider confidence interval suggests that the magnitude of the effect is less precise compared to the unmatched data. The matched data shows a larger proportion of people who received a call and voted compared to those who did not, but the uncertainty in this estimate is higher. Overall, the results imply that receiving a GOTV call has a positive effect on voting, but after controlling for confounders, the effect appears less certain and smaller in magnitude than initially suggested by the unmatched dataset.