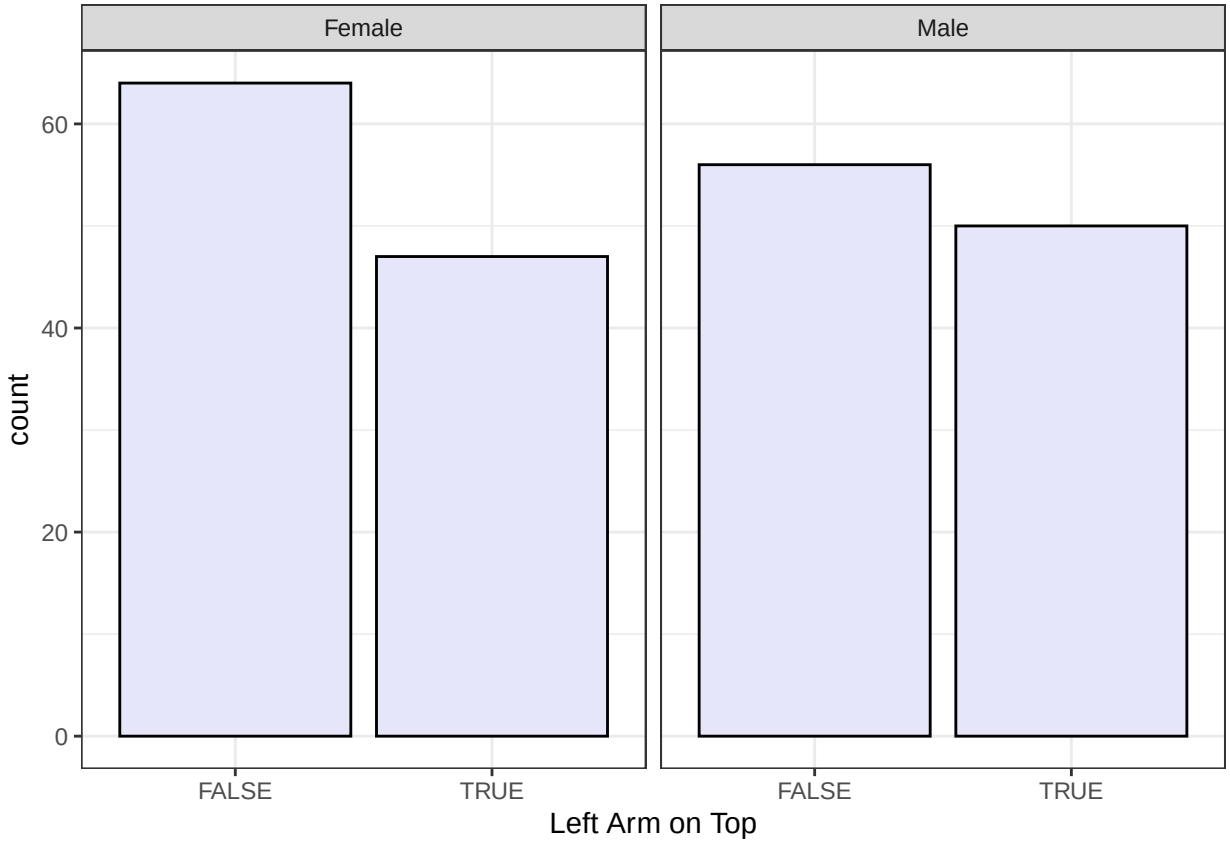
HW7 - SDS315

Rachel Chavez - rcc3342 - {GITHUB}

2025-04-02

Problem 1: Armfolding

The given dataset contains a matrix of 217 students from an Australian professor's class with their sex and whether or not their left arm was above their right arm when folding their arms. This analysis is meant to estimate the true difference in proportions (between males and females who cross their arms with their left arm on top) within the population–being the undergraduate students at this Australian university.



This sample contains 111 females and 106 males, adding up to 217 observations in total. Of these 217, 50 are *men* who place their left arm above their right, and 47 are *women* who place their left arm above their right.

Based on our sampling distribution of which arm is typically placed on top, it appears the right arm is preferred. Interestingly, it appears men are more likely to place their left arm on top than women, at least as reflected in this sample.

About 47.2% of men will place their left arm on top while only 42.3% of women will do so. Our observed proportion difference of the sample is 0.048, or 4.8%.

With this data and de Moivre's equation, we can construct a 95% confidence interval of where our true population proportion difference may lie. We begin to calculate the standard error of our difference in proportions sample by taking each sex's proportion of left hand over right multiplied by its proportion of the opposite–and then dividing that product by the total number of that sex (found above to be 111 and 106 females and males, respectively). The two sex's fractions are then added together. Once we take the square root of that sum, we have our standard error for the differences in proportion.

Since we are constructing a 95% confidence interval, we know our random variable (the proportion difference) will fall between about 1.96 standard errors to the left and right of our observed proportion difference. Thus, we calculate: $Bounds = SampleProportionDifference \pm (1.96) * \sigma$ where σ is our previously calculated standard error.

Bound.1	Bound.2
-0.09	0.19

Using this formula, we can construct a confidence interval to estimate the true difference in proportions between the two groups. If we were to repeatedly construct confidence intervals across many random samples from the population, then we would expect that the true difference in proportions between men and women who fold their left arm on top lies between -0.08 and 0.18, with 95% confidence. Because 0.00 (no difference in population proportions) falls within that interval, however, we do not have sufficient evidence to conclude the proportions of either sex significantly differ.

Because we only have one sample, the standard error (0.07) calculated here represents the uncertainty in our prediction of proportional differences between male and females who fold their arms with the left arm on top. It informs us how much the difference in proportions could fluctuate due to random sampling variation if we had more samples.

The sampling distribution in this case is the distribution of differences in proportion between men and women who fold their arms with their left arm on top due to random sampling. This proportional difference varies from sample to sample, but the true difference in proportion of the population stays fixed.

Thanks to the Central Limit Theorem and our sufficiently large sample size, we can use a normal distribution to approximate our sample mean and standard deviation using de Moivre's equation. From these ingredients, we can construct a confidence interval.

If somebody were to claim, from the confidence interval of [-0.01, 0.30], that there is no sex difference in arm folding, I would agree with 95% confidence, assuming they're referring to specifically the population of this sample (undergraduate university students). I would respond that because 0 is within our confidence interval, there is no statistically significant evidence of a difference in proportion between the two sexes. If, however, they were trying to make that statement referring to the general human population based on this data set, I would argue that that is an impossible claim to make as we lack sufficient data.

If we repeated this experiment many times across many samples, the confidence interval bounds would fluctuate due to random sampling variation. Each sample would have a slightly different proportion of men and women who fold their arms with their left hand on top because people are unpredictable and each observation is independent from the next. Across the entire collection of all of those intervals, however, approximately 95% of our intervals should capture the true population proportion difference between the two groups. The more confidence intervals we construct from samples, the narrow and subsequently more precise our 95% confidence interval will become as our sampling distribution will narrow.

Problem 2: Get Out the Vote

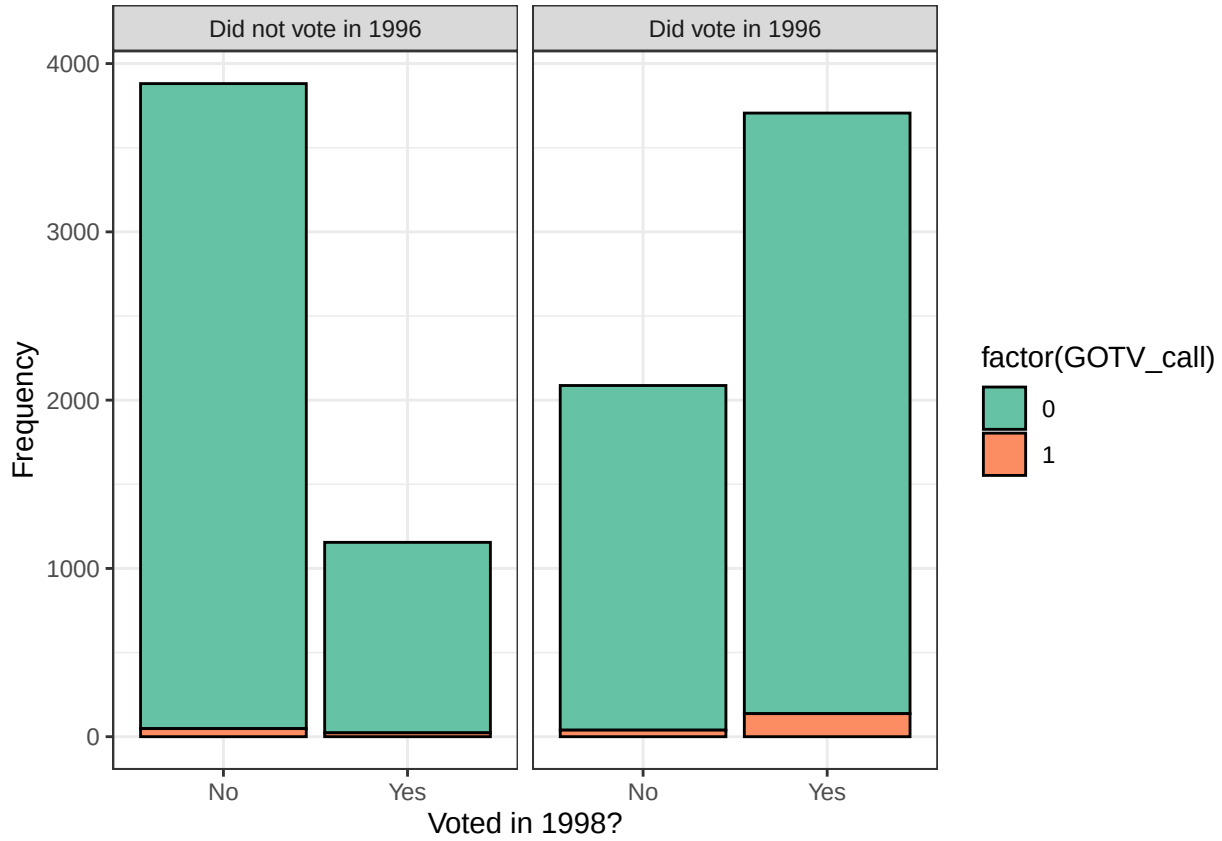
In this analysis, we examine how receiving a "Get Out the Vote" (GOTV) call influenced Congressional voter turnout in 1998. We use a sample of 10,829 registered voters from the major parties' databases. The dataset includes the following variables: 1) whether the voter participated in the 1998 election, 2) whether they received the GOTV call, 3) whether they voted in the 1996 Congressional election, 4) their age, and 5) their party affiliation.

Only 160 people both 1) voted in the 1998 election, and 2) received the GOTV call. 4,701 people voted, but did not receive the call. We find our sample proportions of those who voted in 1998 and those who did or did not receive the call to be approximately 0.65 and 0.44, respectively. This suggests a higher proportion of voters among those who received the GOTV call.

By using the same approach as above, we can construct a 95% confidence interval to estimate where the difference in proportion of the population likely lies.

Using this equation: $se(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}$, and filling in the appropriate values, where our first proportion and sample are those who voted in 1998 and received the GOTV call, and our seconds are those who voted in 1998 but did not receive the GOTV call, we can calculate the standard error (which turns out to be 0.03, approximately). By then multiplying that value by 1.96 (for 95% confidence) and adding, then subtracting it from our observed difference in proportions (being 0.20, approximately), we find that the upper bound of our interval is 0.26, and our lower bound is 0.14. Thus, we estimate with 95% confidence that the true difference in voter turnout between those who received the GOTV call and those who did not falls between 0.14 and 0.27. Since this interval does not include zero, we can confidently say that receiving the GOTV call likely had an effect on voter turnout in 1998.

There are, however, complications to this conclusion. The other listed variables could be confounders, meaning they influence both voter turnout in 1998 and receiving the GOTV independently.



Proportion-Difference-1996	lower.bound	upper.bound
Voted in 1998	0.509	0.541
Received GOTV Call	0.011	0.022

The bar plot above shows how voting behavior in 1998 is related to prior voting behavior in 1996 as well as whether a person received a GOTV call. Those who did not vote in 1996 clearly received more GOTV calls, particular the group which also voted in 1998. Both facets are very uneven, with the first indicating that the majority of those who did not vote in 1996 also did not vote in 1998. Similarly, most of those who voted in 1996 also voted in 1998.

To confirm what the graph shows us, I can construct two confidence intervals: one which contains the difference in proportion between those who voted in 1998 across those who did and did not vote in 1996, and one that contains the difference in proportion between those who received the GOTV call across those who voted in 1996.

The first confidence interval (examining the influence of voting in 1996 on voting in 1998) estimates the true difference in proportion to be within [0.51, 0.54], with 95% confidence. This result does not contain zero, indicating a significant difference. Therefore, whether or not someone voted in 1996 likely influences the likelihood of voting in 1998. Specifically, those who voted in 1996 are more likely to vote in 1998.

The second confidence interval (examining the influence of voting in 1996 on whether or not someone receives a GOTV call) estimates the true difference in proportion to fall within [0.01, 0.02], with 95% confidence, which again does not contain zero. This result indicates that whether or not someone voted in 1996 is associated with the likelihood of receiving a GOTV call. In particular, those who voted in 1996 are possible more likely to have received a GOTV call.

Based on these results, we can conclude that voting in 1996 is associated with the likelihood of receiving a GOTV call and to have voted in 1998.

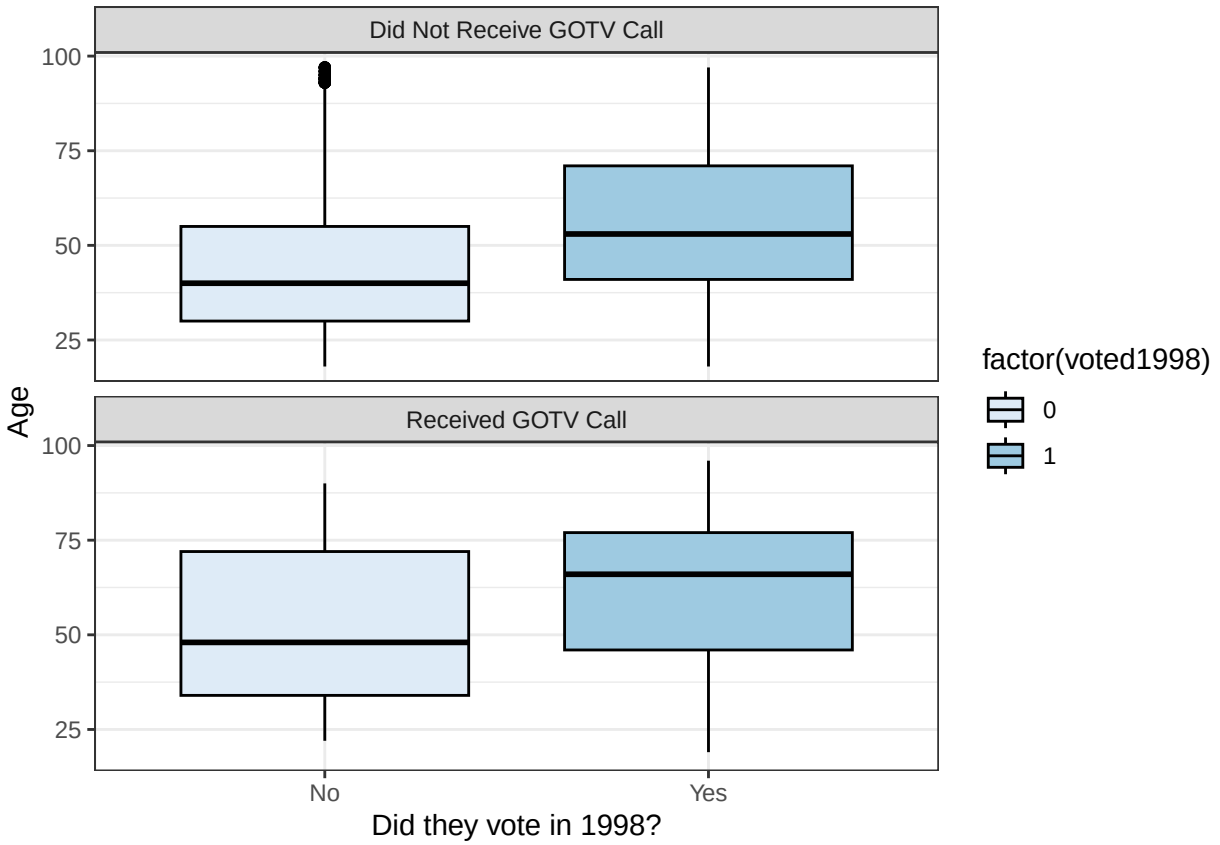
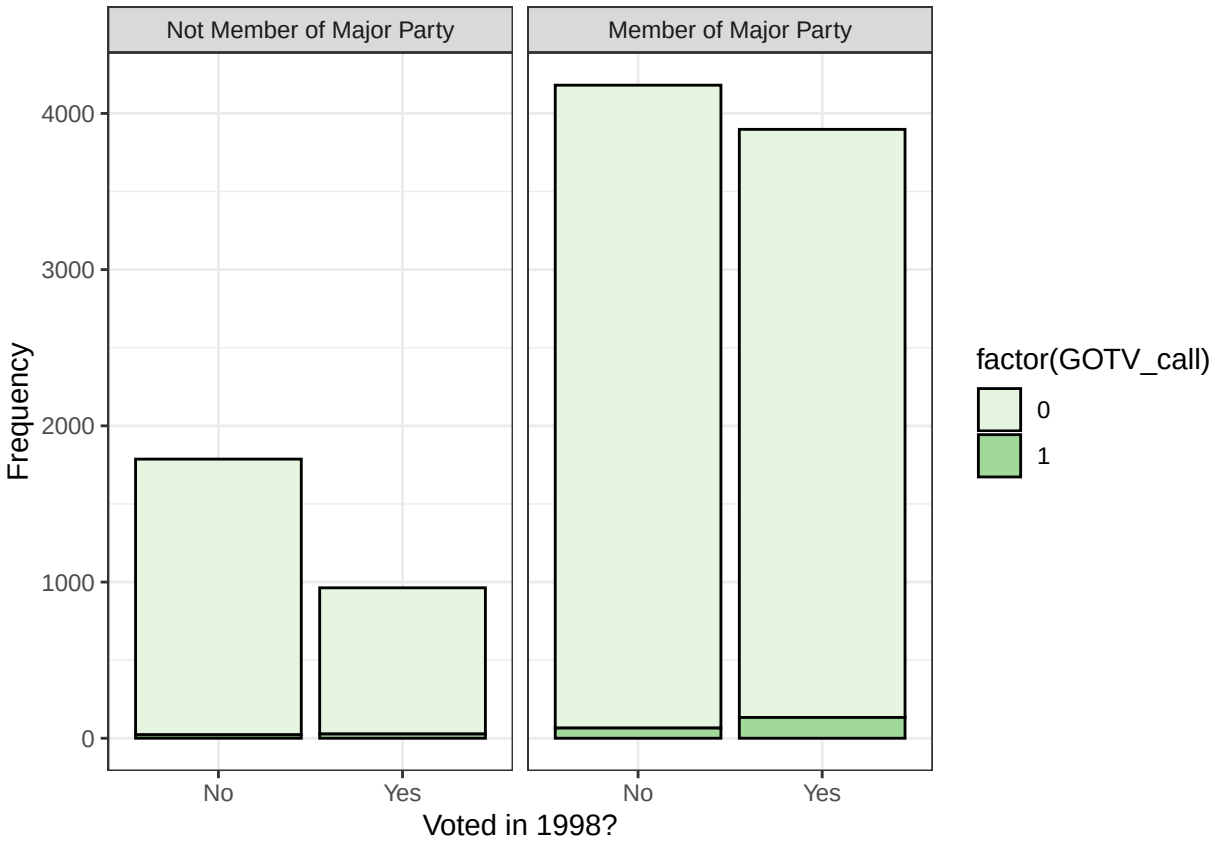


Table 3: The Effect of Age on...

Voted.1998	Received.the.GOTV.Call
-11.18 to -9.82	-11.4 to -6.37

From the box plot, we see that the median age of those who did not vote in 1998 is younger (around 38-50 years) compared to those who voted in 1998 (who tend to be older, above 50 years). Additionally, the medians for individuals who received the GOTV call appear to be higher than those who did not.

To further examine this, we repeat the process of building confidence intervals to examine the variables' influence on each other. The table shows the 95% confidence intervals from the t-tests conducted to assess the effect of age on two variables: voting in 1998 and receiving the GOTV call. For voting in 1998, the confidence interval is approximately [-11.18, -9.82], suggesting a significant difference in age between voters and non-voters. The second interval (examining the variables of the GOTV call and age) is around [-11.40, -6.37], indicating that those who received the GOTV call were significantly older than those who did not, with 95% confidence.



Proportion-Difference-1996	lower.bound	upper.bound
Voted in 1998	0.587	0.621
Received GOTV Call	0.586	0.620

Finally, we consider party affiliation as a potential confounder. The bar plot above visualizes the relationship between party affiliation, voting behavior in 1998, and receiving the GOTV call. Based on the plot, individuals affiliated with major political parties appear more likely to both vote in 1998 and receive a GOTV call.

The 95% confidence intervals for both variables (voting in 1998 and receiving the GOTV call or not) in associated with major party affiliation are [0.59, 0.62], which suggests a significant association between party affiliation and both voting in 1998 and receiving a GOTV call.

To account for the confounders, we use a matching procedure to match individuals who received the GOTV call with those who did not based on prior voting behavior (voted in 1996), age, and party affiliation. This reduces bias and provides a more accurate estimate of the causal effect of receiving the GOTV call. I used 5 control cases for each treated case in the matching process.

To confirm that the previously confounding variables are no longer confounders, we will examine each one's influence on whether or not somebody received the GOTV call and construct new confidence intervals for each.

Proportion-Difference-1996	lower.bound	upper.bound
Confidence Interval	-0.042	0.042

Now, reexamining the influence of whether someone voted in 1996 to whether they received the GOTV call with our matched data, we construct a new confidence interval, similarly to above. This new confidence interval is from -0.04 to 0.04, meaning it now includes zero. Because of this, voting in 1996 no longer has a statistically significant influence (with 95% confidence) on whether somebody received the GOTV call. Thus, it is no longer a confounder.

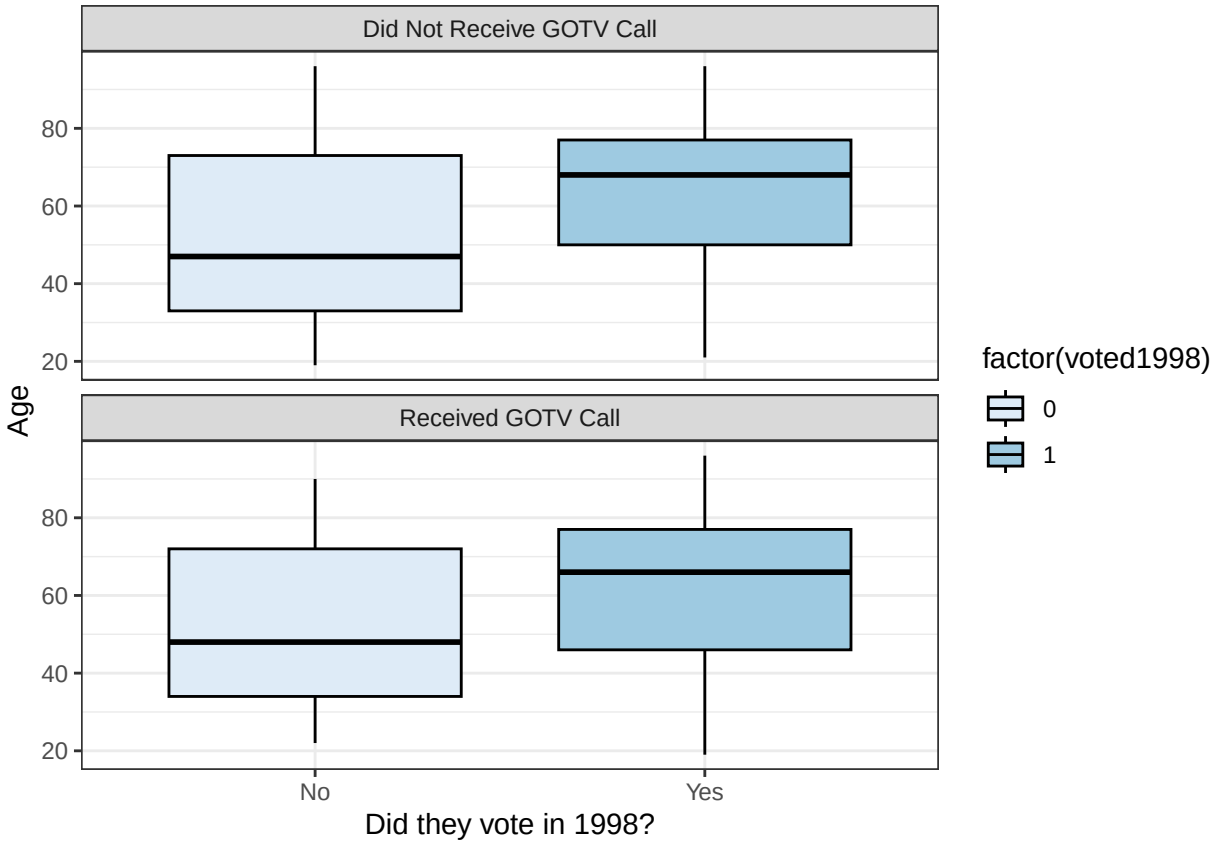


Table 6: Age Difference Confidence Interval by GOTV Call

Lower.Bound	Upper.Bound
-2.76	2.678

I've replotted the boxplot using only the matched data, and now, comparing the group who received the GOTV with those who did not, we see that their plots are almost identical. Their medians are in very similar places, unlike in the raw dataset's plot.

Now reconstructing our 95% confidence interval with the relationship between age and whether somebody received the GOTV call, I get the bounds -2.76 to 2.68. 0 years of age is within the confidence interval, so now age no longer has a statistically significant effect on the GOTV variable. Thus, age is no longer a confounder.

Proportion-Difference-Major-Party-Confidence-Interval	
Lower Bound	-0.056
Upper Bound	0.045

Finally, we make a confidence interval with the matched data, estimating the proportion difference between

those within major parties and those who received the GOTV call. The difference in proportions measures how much the proportions of people from each party who received a GOTV call differ between the two groups (treated vs. control). From this new confidence interval, I conclude with 95% confidence that the true difference in proportions between the two groups falls between -0.06 and 0.05. Because zero is also included in this confidence interval, we can conclude that whether somebody is or is not in a major party does not have a statistically significant effect on whether received the GOTV call. Thus, it is no longer a confounder.

Using these matched pairs, we can reexamine the influence on voting in 1998 by receiving the GOTV call.

Table 8: Difference in 1998 Turnout: GOTV Call vs. No Call

lower.bound	upper.bound
0.013	0.144

The final mean difference in proportion is approixmately 0.08. Thus, the true difference in proportion between voters in 1998 who did receive the GOTV as opposed to those who did not falls within [0.01, 0.15], with 95% confidence. Since this confidence interval does not contain 0, it suggests that the difference in the proportion of voters who participated in the 1998 election between those who received the GOTV call and those who did not is statistically significant.

This analysis indicates that receiving a GOTV call increased the likelihood of voting in the 1998 election. The bounds of the confidence interval suggest that the effect of the GOTV call on voting behavior could range from a 1% to 15% increase in the probability of voting in the 1998 Congressional election.