Difficulty as Motivators for Voter Turnout

W203 Lab 1: Statistical Analysis

Jessica, Josie and Heaven

Contents

1 Importance and Context

Just like the COVID-19, the US 2020 elections raised unprecedented challenges for its citizens. While the nation was still recovering from the pandemic, the Americans, amongst the turmoil, stood up to cast their vote and fulfill the needs to democracy. Access to voting, a discussion rooted in our countries' racist history, was once again a heated political topic. In particular, Democrat commentators argued that Republicans used a variety of tactics to make it more difficult for likely Democrat voters to actually submit their ballots. For example, there are allegations of limited voting hours, fewer voting locations, staffing issues that lead to long lines, and sparse drop-boxes for absentee ballot drops. If such speculation is true, we should be concerned about the integrity of our elections. As a first step, this statistical analysis looks to nationally representative survey data to gain insight from voters themselves on the situation on the ground in the 2020 election to answer the question of interest:

> *Did Democrat-leaning or Republican-leaning voters claim to have more difficulty voting in the 2020 election?*

We utilize data from the 2020 American National Election Studies (ANES), which surveyed a nationally-representative set of respondents before (Aug 18, 2020 - Nov 3, 2020) and after the 2020 national election that occurred on Nov 8, 2020 (Nov 8, 2020 - Jan 4, 2021). It was a survey of non-institutionalized U.S. citizens age 18 and older.

2 Data and Methodology

Conceptually, we think of a *voter* as broadly someone is eligible to vote and has some intention of participating in the election process, regardless on if they actually cast a vote. We believe registering to vote signals an intention of voting. Based on this definition, we believe individuals who registered to vote (or said that they plan to register to vote) should be classified as voters. Our ideal measure would include individuals who voted, registered individuals who did not vote, and individuals who planned to register but did not manage to do so. By including registered voters (or those who planned to register), we can more fully understand how difficulties in the voting process may have deterred individuals from voting.

We think of a *Republican* or *Democrat party-affiliation* as someone who leans toward either the Republican or Democrat party in their voting preferences. Since we think of voters broadly of those who had an intention of voting, rather than just those who managed to vote, we do not want to restrict our definition to actual votes for one party or the other on the ballot.

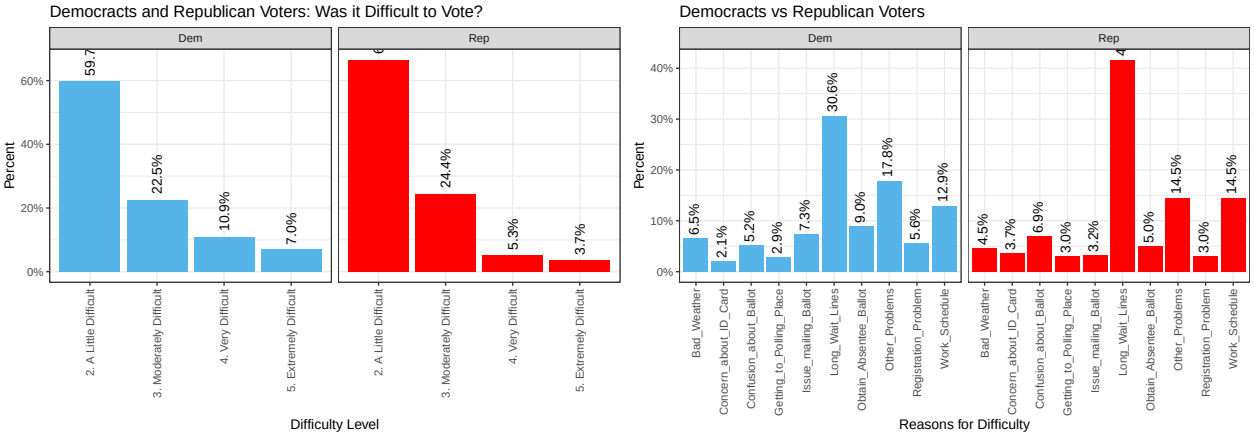
Last, we think of *difficulties* in voting quite broadly. These can vary from personal circumstances such as work schedules, care for small children, illness, or bad weather to more structural issues around the voting infrastructure, such as confusion about how to request an absentee ballot, not having the proper documentation to register, not meeting registration deadlines, a long commute to a polling site from your house, and limited hours of a polling site. Each of these individual items would require a separate measurement. We would want to know "how" big a barrier each of these things actually is, measuring them quantitatively. For example, we could measure distance to a polling site in terms of minutes of commute time; total hours or days voting booths in a county are open; or compare how voting registration deadlines are across states.

Our survey dataset is not sufficient to measure the ideal version of these concepts, however there are useful proxies. To measure "difficulty of voting", we are using the post-election survey question: "How difficult was it for you to vote?" (V202119) from ANES that had 5 categories: Not difficult at all, a little difficult, moderately difficult, very difficult, and extremely difficult. We specify our "difficulty" measure as an ordinal variable with the values of 1 through 5 that map back to the survey question categories, where 1 = Not difficult at all and 5 = extremely difficult. This question is a good proxy because it broadly encompasses any difficulties an individual may have had, personal or voting-infrastructure related. The disadvantage of this survey question is that it is only asked of respondents who actually voted in the Nov 2020 election. Individuals who may have planned to vote but faced substantial enough barriers doing so, either while trying to register or trying to vote, are excluded. Self-reports can also be subject to recall or social-desirability bias.

Since our "difficulty" survey measure is constrained to those who voted, our operational definition of voters is also narrowed to just those who voted. This includes both the pre-survey (V201023) and post-survey (V202066) questions with respondents who said "Yes, voted" or "I am sure I voted". This of course excludes respondents who may have intended to vote but had difficulty doing so. Furthermore, the survey has a sequence of questions we used categorize respondents who have a preference for the Democrat or Republican parties in the pre-survey (V201228, V201230). We labeled respondents as Democrats who said they "think of themselves as a Democrats" or "think of themselves closer to the Democrat party." We labeled respondents as Republicans who said they "think of themselves as a Republican" or "think of themselves closer to the Republican party. Our party variable is a categorical variable where 1 = Republicans and 2 = Democrats.

From the total survey sample of 8,280 respondents, there are 7,282 that classified themselves as either a Republican or Democrat. Once we filter the sample down to those who voted in the 2020 election, the sample is 5,874. Then, our final sample includes voters who answered the difficulty question, which dropped down to 5,831 respondents (2,700 Republicans and 3,131 Democrats).

We show two charts to provide an understanding of the data. On the left, we showed how Democrats and Republicans rated themselves on the voting difficulty score. While most voters had no difficulty at all, more Republican voters said that it was a little to moderately difficult, while more Democrats answered it was very to extremely difficult. The chart on the right shows different problems voters mentioned when voting. We see again similar distributions for Republican and Democrats and the top problem for both groups is long wait times and second being other problems. Since this survey does not delve into those problems, we are unsure what other problems could be causing the difficulty. We will move next to a statistical test to discern if there are meaningful differences by party.



We are using the non-parametric Wilcox Rank Test (hypothesis of comparisons version) to do our statistical test. This test is appropriate because we are comparing two groups (Democrats and Republicans) on a ordinal outcome variable (difficulty score). Other standard 2-sample tests require metric data.

There are two key assumptions for the Wilcox Rank Test test to provide valid inferences. The first is that we are measuring our outcome on an ordinal scale. Our outcome variable (difficulty voting) is indeed on an ordinal scale (values have a meaningful order ranging from 1 to 5). Second, our data must be independent and identically distributed (iid). It is reasonable to assume the sample is iid because the respondents were randomly-selected across the country by a professional survey firm. With a true random sample across the whole country, we can believe that a given respondent does not influence a subsequent respondent's answer to the survey. Our null hypothesis is as follows:

> **Null Hypothesis**: *The probability that a difficulty score for a Republican is less than the difficulty score for a Democrat is equal to the probability that a difficulty score for a Republican is greater than a difficulty score for a Republican.*

3 Results

```
wilcox.test(anes$difficult ~ anes$party,
            alternative = "two.sided", paired=FALSE, exact=FALSE, use = "complete")
```

The test shows that there is a statistically significant difference in the difficulty of voting between Republicans and Democrats (p-value = 0.003687). We can reject the null hypothesis at a 99% confidence level (i.e., there is only a 1% chance that we are rejecting the null hypothesis when we should not).

Some limitations of our test affect the conclusions drawn from the analysis. Foremost, this test only indicates correlation (not causation) between party and difficulty voting. It does also not hint at why we might be seeing a difference. It is possible that the driver is actually a third factor (e.g., socioeconomic) that is correlated with party affiliation. From here, we can analyze more deeply which party experienced more difficulty while voting. In addition, we are only seeing a subset of all voters. This measure only captures those who responded to this survey (only roughly a 30 percent response rate), voted, and answered the difficulty question on the post-survey. Each step reduces the sample farther away from our ideal comprehensive conception of a "voter" we described above. Furthermore, because we did not apply the survey weights, the result is also not a generalization of the full U.S. population.

4 Discussion

This study found that party affiliation is associated with voting difficulty. From a pragmatic lens, we are interested in both the direction and magnitude of this trend. Using Spearman's Correlation, we found that in possible pairings of a Democrat and Republican, only 3.8% of the time, Democrats had more difficulty voting than Republicans. Taken together with the Wilcox Rank Test, we find that while there is a statistically significant difference in difficult scores between Democrats and Republicans, the magnitude of this difference is quite small. It is of little practical explanatory value. This could be due to either the limitations of the survey data or that another factor (e.g., socioeconomic) actually drives this difference.

Our descriptive analysis looked at what proportion of each party reported difficulty in voting and why did voters find voting difficult. As mentioned above, a higher proportion of Republicans voted on the lower side of difficulty while the opposite is true on the higher side. The difficulty reasons gives another important layer to understanding the difficulty distribution but lacks two key insights: those who found it so difficult to even register they did not vote, and other reasons why it was difficult to vote outside the ones (since the second highest amount of responses was "other reasons"). To continue this research, we would need to include a similar likert scale for those who found it difficult to register and see if the barrier to entry started there, which may change the difference in Republicans and Democrats difficulty. Non-likert scale data should also be used since data such as where a voter lives determines much of their difficulty (i.e., voting regulation is at the state level).

In future steps, we hope to do a further analysis of party affiliation and other factors that may better explain why certain groups of voters have more trouble voting than others. We wish to study the party affiliation driver with better data, such as with a more comprehensive definition of voters (i.e., include registered voters and those who intended to register). More feature engineering can be done to capture voting behavior of lean Democrats or lean Republicans. Using further data outside of the ANES report, we could more accurately look into causes for the difference in voting.