Analyzing Voting Difficulty

Lab 1: Datasci 203

Sonia Song, Kenneth Hahn, Mei Qu

Contents

Importance and Context

Voter turnouts in U.S. elections have historically been below two thirds of the eligible voting population. With about 66% of the eligible voting population turnout, 2020 presidential election saw one of the highest rate for any national election since 1900. [1]

It is important to note that voters don't vote consistently along the party lines over time or across issues, which increases the unpredictability of election outcomes. Given the rising political divide in the U.S., any voting irregularities can potentially create an outsized effect on election outcomes. One of those variables is difficulty of voting.

Our analysis seeks to answer the below research question using statistical methods:

> *Do Democratic voters or Republican voters experience more difficulty voting?*

It is critical to understand whether there is any systematic difference in how difficult it is to vote among Democrat and Republican voters for fair report and analysis of election results. The additional transparency can increase public confidence in elections. Moreover, further investigation of the underlying drivers of voting difficulty can provide valuable insights for improving political and civic engagement.

Data and Methodology

Our analysis uses the American National Election Studies (ANES) 2022 Pilot Study dataset. This is an observational dataset based on sample respondents collected from YouGov to test questions for potential inclusion in the 2024 time series study and to understand public opinion after the 2022 midterm elections. There are a total of 1585 cases in the study. We removed 85 unweighted cases not selected by the sample matching procedure, which the study recommends excluding for making any inferences about the general population.

We define voters as those who are registered to vote (responded 1 or 2 to `reg`) or those that answered the "how difficult was it to vote" question (response to `votehard` was not equal to -1, or a skipped answer). We used an "or" statement because there could be voters who are not currently registered to vote but voted in the November 8th election and they should be accounted for in our analysis. We note from the documentation that only respondents who definitely voted or probably voted received the `votehard` question. By defining voters as such, we realize we may miss responders who say they didn't register to vote, answered that they voted or likely voted in the November 8th election, but then skipped the `votehard` question. However, we observed that there were no respondents in this category, making our definition holistic.

To differentiate a Democrat from a Republican voter, we recognize that the survey generates a `rand_pid` (a random integer value) from 1-3 for each respondent. Respondents who get assigned a `rand_pid` of 1 or 3 receive question `pid1d` and respondents who get assigned a `rand_pid` of 2 receive `pid1r`. Both ask the question "Generally speaking, do you usually think of yourself as a Democrat, a Republican, an independent, or what?" but with different phrasing of the questions and answers. We also see a question `pidlean` asking respondents which party they are closer to if they stated they were "Independent", "Something else", or skipped the earlier question. Therefore, we categorized Democrats as those who responded "Democrat" to question `pid1d`, `pid1r` or `pidlean` and Republicans as those who responded "Republican" to those same questions, resulting in 1317 responses. We considered using variable `pid_x` instead in our definition which is on a scale from 1-7 from "Strong Democrat" to "Strong Republican", but ultimately chose not to use this variable as we didn't see any documentation of where this variable came from in the Questionnaire Specifications. We also decided not to try to further categorize those who hadn't been categorized as we shouldn't make the unnecessary assumption that their political views or who they elected and would elect for president reflects their political affiliation.

[1] https://www.pewresearch.org/politics/2023/07/12/voter-turnout-2018-2022/

The survey asks the respondents “How difficult was it for you to vote?” (also known as `votehard`) with 5 choices ranging from “Not difficult at all” (1) to “Extremely difficult” (5). Once we filtered the data for Democratic and Republican voters, we used the `votehard` ranking to conduct our statistical test to test the null hypothesis, as this variable is directly applicable to the research question at hand. After removing anyone who is not a Democrat or Republican and removing those who did not answer the `votehard` question (a response of -1), we remain with 976 rows.

The sample test we choose is the Wilcoxon rank-sum test. The data is not paired as each surveyee has their own individual score. We believe that this is the most appropriate test as the dataset satisfies the following assumptions for the Wilcoxon rank-sum test: 1. The data is I.I.D. because one surveyor’s answer does not depend on the other and both groups are pulled from the same distribution of surveyors 2. Data is ordinal and not metric because intervals are different from person to person, e.g. the difference between a 4 (very difficult) to a 5 (extremely difficult) will be different from one surveyee to the next. Because the data is ordinal, then it must also be non-parametric.

Given that the data is at least ordinal and I.I.D., we must use the Hypothesis of Comparisons to define our null hypothesis:

> **Null Hypothesis:** *The probability of Democrat voters experiencing voting difficulty being greater than Republican voters is the same as the probability of Democrat voters experiencing voting difficulty being less than Republican voters.*

We will be utilizing a two-tailed Wilcoxon Rank-Sum test because a one tailed test will not only make it easier to reject the null hypothesis but a one tailed test also assumes that the opposite case cannot occur, which in this scenario either probability can be just as likely to occur.

As a result our alternative hypothesis is as follows:

> **Alternate Hypothesis:** *The probability of Democrat voters experiencing voting difficulty being greater than Republican voters is not equal to the probability of Democrat voters experiencing voting difficulty being less than Republican voters.*

Results

Table 1: Votehard Summary Table by Party

party	Count_votehard	Mean_votehard	Standard_Deviation_votehard
Democrat	525	1.281905	0.6629193
Republican	451	1.124168	0.4691519

Table 1 above shows the results of filtering our dataset for only Democratic and Republican voters. The table portrays that there are 74 more Democrat voters than Republican voters and that Democrats have an average `votehard` score that is 15.77% higher than Republicans. Because this data is ordinal, we cannot directly compare the means of the two groups and we conducted the Wilcoxon Rank-Sum test with the code below.

```r
results <- wilcox.test(dem_data_votehard$votehard,
                       rep_data_votehard$votehard, alternative = "two.sided")
```

After performing the Wilcoxon rank-sum test, we concluded that the probabilities are not equal (W = 1.306215×10^5, p = 3.6163904×10^{-6}). This result indicates that the probability that a Democrat has more difficulty voting than a Republican is not equal to the probability that a Democrat has less difficulty voting than a Republican. Figure 1 delineates the differences between the two survey groups and how they answered the question of how difficult voting was for them. We calculated by the percentage of the respective party that answered each of the categories for votehard, since there are more Democrat voters who answered the question than Republican voters (525 Democrat voters and 451 Republican voters).

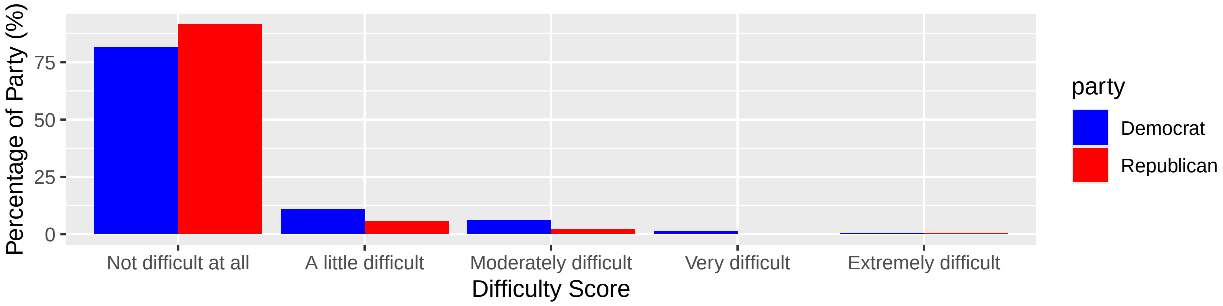

Figure 1: Percentage of Respective Party that Answered "votehard" from 1-5

Figure 1 shows a heavy tail where a vast majority of both Democrats and Republicans did not think that voting was difficult at all (81% and 91%, respectively). We also do observe that a higher percent of Democrats responded to the "more difficult" option than Republicans did. Ultimately, the Wilcoxon Rank-Sum test concludes that we can reject our null hypothesis that the probabilities are not equal; however, the limitations of the test cannot conclude which probability is more likely to occur and the result is restricted to the sample set of surveyees, meaning that we cannot apply these results directly to the U.S. population.

Discussion

The study concludes that we can reject the null hypothesis: the probability that Democratic voters find voting more difficult than Republican voters is not equal to the probability that Republican voters find voting more difficult than Democratic voters. Again, this analysis does not delineate which party has a higher probability and does not represent the population as a whole. Understanding these two restrictions can allow us to develop future studies to clarify the answer further.

For example, we can utilize the 10 different `vharder` categories, which is a series of questions where surveyees can select whether they found a specific reason that made it difficult to vote or not. Upon initial investigation with Figure 2, we can conclude that for each of the categories, a higher percentage of Democratic voters responded than Republican voters. There is opportunity to conduct more statistical tests (e.g. a two-proportion test) to see if there are any significant differences between each of the groups of data.

Furthermore, we can get a better understanding of the population at large by conducting the same Wilcoxon Rank-Sum test but utilizing the weights that ANES provides to help gain a better inference on the population's differences to voting difficulty. Both of these studies are outside of the scope of our analysis, but provide a framework to further understand voter turnout.

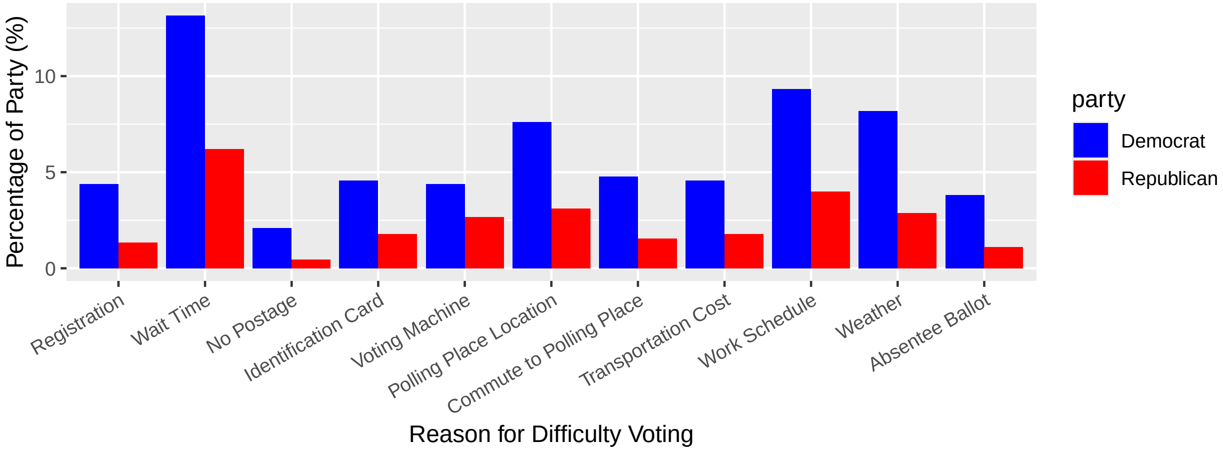

Figure 2: Reason for Difficulty Voting by Political Party