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]

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?*

The answer to this question can provide additional transparency, increasing 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. 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 voted in previous elections if they did not register to vote and 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?" 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 the variable `pid_x` which is a ranking from 1-7 from "Strong Democrat" to "Strong Republican", but we chose not to use this variable as it was not mentioned in the Questionnaire Specifications. We also decided not to further categorize respondents who had not been categorized to avoid false assumptions on their political views.

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 (`votehard` = -1), we remain with 976 rows. We also considered using the question "How difficult was it

[1] Hartig, H., Daniller, A., Keeter, S., & Green, T. V. (2023, July 12). 1. Voter turnout, 2018-2022. Pew Research Center. https://www.pewresearch.org/politics/2023/07/12/voter-turnout-2018-2022/

for you to register to vote?" for inclusion of voters with difficulty registering but decided to leave it out because only 12% in the eligible sample did not skip the question and based on the questionnaire those who responded were voters who successfully registered to vote, so the responses are not sufficiently informative.

The sample test we choose is the Wilcoxon rank-sum test. The data is not paired as each respondent 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 respondent's answer does not depend on the other and both groups are pulled from the same distribution 2. Data is ordinal, not metric, because the difference between intervals, such as a 4 (very difficult) to a 5 (extremely difficult), is not universally agreed upon. Because the data is ordinal, it must also be non-parametric. Given that the data is at ordinal and I.I.D., we must use the Hypothesis of Comparisons to define our null hypothesis, where P is the probability, and we are comparing the `votehard` scores for Democrats and Republicans, respectively:

H_0: $P(votehard_{Democrat} > votehard_{Republican}) = P(votehard_{Democrate} < votehard_{Republican})$

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 also assumes that the opposite case cannot occur, which in this scenario either probability can be just as likely to occur. We will also be defining our type I error as $\alpha = 0.05$ as our threshold for rejecting the null hypothesis. As a result our alternative hypothesis is as follows:

H_a: $P(votehard_{Democrat} > votehard_{Republican}) \neq P(votehard_{Democrat} < votehard_{Republican})$

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.

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

After performing the Wilcoxon rank-sum test, we observed that our p-value is less than 0.05. This means 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 ($W = 1.306215 \times 10^5$, $p = 3.6163904 \times 10^{-6}$). Figure 1 delineates the differences in responses between the two survey groups at each difficulty level. 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 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 respondents, meaning that we cannot apply these results directly to the U.S. population.

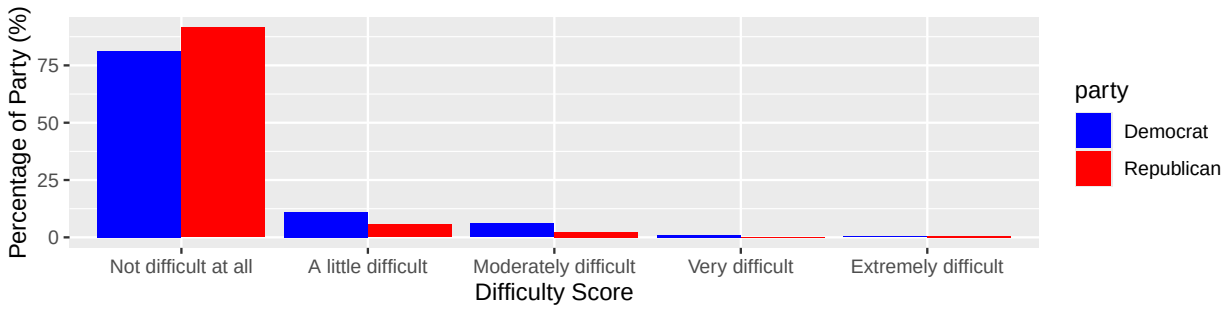

Figure 1: Percentage of Respective Party that Answered "votehard" from 1-5

Discussion

The study found evidence that 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. Our results may be of key interest to future political campaigns, who may have the goal of increasing voter turnout for their party. While this study addresses overall voting difficulty, future studies may build on this framework to further understand the implications behind voter:

1. Conduct more statistical tests (e.g. a two-proportion test) at the category level. For example, we can utilize the 10 different `vharder` categories, which is a series of questions where respondents can select whether they found a specific reason that made it difficult to vote or not. Upon initial investigation with Figure 2, we observed that for each of the categories, a higher percentage of Democratic voters responded than Republican voters. Make better inferences about the population by utilizing the weights that ANES provides while conducting the sample test.

2. Make better inferences about the population utilizing the weights that ANES provides while conducting the sample test

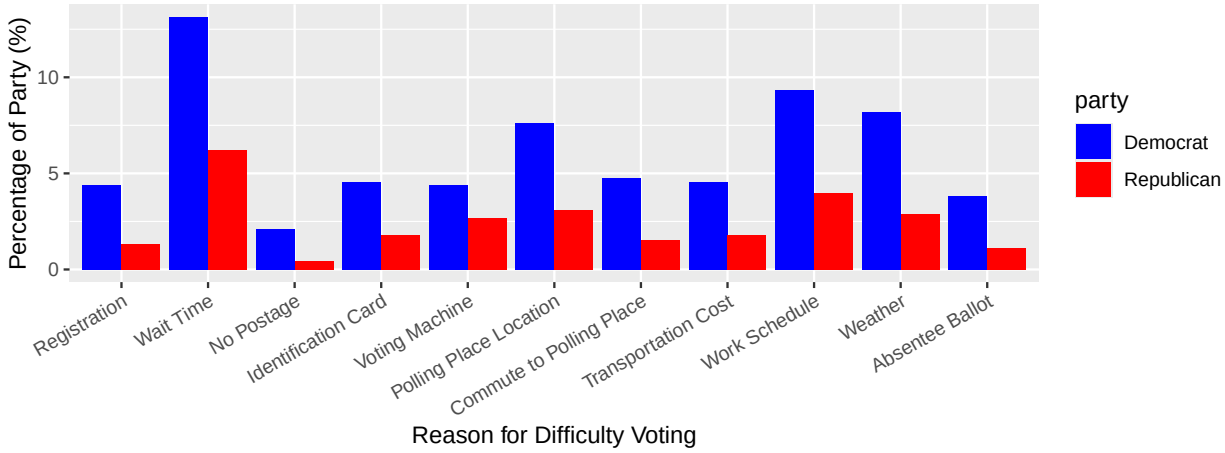

Figure 2: Reason for Difficulty Voting by Political Party