Lab 1: Analyzing Voting Difficulty

Datasci 203

Nicolas Aragon, Bella Davies, Andrew Main

Contents

1 Introduction

The analysis of the motivations of American voters has become increasingly complex and yet voter behavior seems consistent for the past 50 years. Notably, race has become much more visible as a factor that denotes social difference. The victories of Carter and Clinton and the defeat of Gore were marked by black voters creating marginal majorities in southern states. Whites support republican candidates by 47-43% while Blacks preference for democrats has increased from 57-19% to 84-7% (See Petrocik 2009, University of Missouri-Columbia). Ensuring that people are accurately represented by elected public officials has become increasingly relevant in today's political climate, where increased partisanship has resulted in hostility towards members of the other party and an unwillingness to compromise on important matters (see Brooker and Beitsch). As a first step in this analysis, we address the question: *Do Democratic voters or Republican voters experience more difficulty voting?* We approach the question via non parametric hypothesis tests over a ranking for the perceived difficulty to vote. The answer to this question could help validate concerns over equal access to voting in America and help legislators define policies to remedy the situation going forward.

2 Conceptualization and Operationalization:

To further clarify this question so that it can be studied, the terms must be conceptualized and defined. Voters are U.S. citizens over the age of 18 who responded to a 2022 survey from ANES. The groups of "Republican" and "Democrat" consist of voters who support either political party. There are a few variables in the ANES dataset that denote support for a party such as "pid1d", "pid1r", "pidstr", and "pidlean". These variables ask respondents which party they "think of themselves as closer to" and how strongly they lean in either direction. Among these alternatives, we chose the "pid_x" variable which represents "Party ID" because it is considered the most representative of which party was supported by the voters on the ballot rather than what voters "think of themselves as". We made this decision to account for Americans who think of themselves as independent leaners, so that we can measure their support for either the Democrat (pid_x $\in [1:3]$) or Reuplican (pid_x $\in [5:7]$) party. Notably, we exclude "Independent voters" because they don't support either party (pid_x $= 4$).

To conceptualize or define the term "difficulty voting" in the research question, variables such as 'regdiff', 'waittime', and 'triptime' were considered as these are all factors which can be correlated with difficulty voting. Additionally, 'vharder 1-12' includes further specific difficulties and complications that voters may experience that could be considered as they impact overall difficulty, such as location, transportation, time, weather, postage, etc. However, since the 'votehard' variable most directly represents the answer to the question "How difficult was it for you to vote?", this variable will best conceptualize and define the phrase "voting difficulty" from the research question for this dataset. We remove 538 records to exclude independent voters or voters for which there is no data available for votehard.

3 Data Wrangling & Data Understanding

The data for this project came from the American National Election Studies (ANES). The ANES conducts surveys that are typically administered as in-person interviews in order to produce data on voting, public opinion, and political participation during most years of national elections. This dataset is the 2022 Pilot Study, which covered the 2022 midterm elections. Our study is focused on two key variables from the dataset: "votehard" and "pid_x". "votehard" is an ordinal variable that asks the respondents how difficult it was for them to vote, with six unique values ranging from "not difficult at all" to "extremely difficult". This question was only asked of respondents who actually voted, excluding those that did not vote for one reason or another. Therefore, our analysis will exclude those who found it so difficult registering to vote in the first place that they were not able to vote. We believe a strong supporting analysis of registration difficulty by party would be an important factor to consider when thinking about the overall difficulty of voting under the current system.

For our analysis, we have grouped all strong and lean voters into their corresponding party. This is because literature has shown that even party "leaners" are partisan and treating them as independent is a misconception in the public discourse (Petrocik, 2009). "pid_x" is an ordinal variable that has respondents self-categorize into either "strong", core (e.g, "Democrat"), or "lean" for both Democrats and Republicans. Since we are looking to compare only Democrats and Republicans as two different groups, we can group the leaners into the two categories without losing relevant information for our question of interest. To do this, we have grouped "1 = Strong Democrat", "2 = Democrat", and "3 = Lean Democrat" to be our full group of Democrats. We also grouped "5 = Lean Republican", "6 = Republican", and "7 = Strong Republican" to be our full group of Republicans. The survey also allows respondents to identify themselves as "4 = Independent". We remove voters who identify as independent and those who did not identify with any category, as they don't offer information around the differences in voting difficulty between the two major parties. We also remove voters who selected "-1 = inapplicable, legitimate skip" for the variable "votehard" or voting difficulty, as this gives us no information about the relative ease or difficulty of voting for the respondent. After dropping these irrelevant observations, there are 1047 observations left out of 1585 originally as the chosen variables represent the majority of the survey dataset.

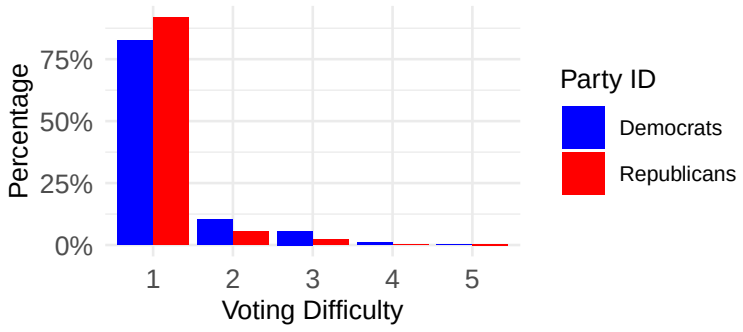
As shown in Table 1, there are 565 self-identified Democrats and 482 self-identified Republicans, resulting in a 53.96%-46.04% split. Using this information, we were able to conduct an initial analysis to show what proportion of each group felt that voting was or was not difficult, as seen in Figure 2. Percentages are plotted along the difficulty scale instead of counts because the sizes of the two groups are unequal, and proportions will provide a more representative visualization. From the histogram, we see that Democrats appear to experience more difficulty voting than Republicans. The hypotheses which follow will be based on this initial analysis, and further statistical testing will then be done to demonstrate whether or not one party has more difficulty voting than the other.

4 Visuals

Table 1: Counts and Percentages of Democrats and Republicans

Party Affiliation	Count	Percentage
Democrats	565	53.96371
Republicans	482	46.03629
Total	1047	100.00000

Figure 2: Voting Difficulty by Percentage of Total Party



5 Hypothesis

H_0: Democrats and Republicans experience the same amount of difficulty voting.

- $P(D < R) = P(D > R)$. The probability that a draw from Republicans will have more difficulty voting than a draw from Democrats is the same or equal to the probability that a draw from Democrats will have more difficulty voting than a draw from Republicans.

H_a: Democrats and Republicans experience unequal amount of difficulty voting.

- Two tailed: $P(D < R) \neq P(D > R)$. The probability that a draw from Republicans will have more difficulty voting than a draw from Democrats is not the same or unequal to the probability that a draw from Democrats will have more difficulty voting than a draw from Republicans.
- One tailed: $P(D < R) < P(D > R)$. The probability that a draw from Republicans will have more difficulty voting than a draw from Democrats is less than the probability that a draw from Democrats will have more difficulty voting than a draw from Republicans.

6 Test Selection and Evaluation of Assumptions

We chose to use the Wilcoxon Rank-Sum test with the Hypothesis of Comparisons Version to evaluate our Hypothesis. Our data meets the key assumptions required for the test, namely that the data is ordinal and that it is Independent and Identically Distributed Data. Ordinal data means that the data has an order, and the variable of interest here, "votehard", is measured in ordered categories, "1 = Not difficult at all", "2 = A little difficult", "3 = Moderately difficult", "4 = Very difficult", and "5 = Extremely difficult". This scale represents an ordinal variable of the Likert scale. Data samples are also independent and identically distributed, since individual survey respondents are independent from each other, and both samples are drawn from the same distribution of U.S. citizens over the age of 18.

7 Test Results and Interpretation

Voting difficulty is different across the two groups, with high statistical significance. We ran the Wilcoxon Rank Sum Test with the Hypothesis of Comparisons Version and found that the test results in a p-value of 6.01×10^{-6}. Because this value is less than 0.05, we reject our null hypothesis that "it's equally likely for a draw in one group to rank difficulty higher than the other group". This has high statistical significance according to the p-value.

The data shows that one of the two groups is more likely to rank difficulty higher than the other, but we're not yet declaring which group perceives higher difficulty. Notably, the effect size is -0.14 which is considered "small" in the context of Pearson's r for a sample size of 1047 voters. The negative sign in the effect size indicates that the second group in the comparison tends to have larger values than the first group. The practical interpretation for a small effect size is that the difference may be subtle.

Since the initial two-sided test shows that the two groups are not equal with high statistical significance, we wanted to then further understand which of the parties experienced more increased difficulty. We hypothesize that democrats have a higher ranking of difficulty than republicans and proceed to use the one-tailed alternate hypothesis stated above. We observe a p-value of 3.01×10^{-6}, which is less than 0.05. It seems there is a higher probability that Democrats have more difficulty voting than Republicans, with high statistical significance. This answers our research question.