Who has it harder? An analysis of aggregate voting difficulty

Alice Drozd, Rex Gao, Brian Lin, Daniel Kim

Introduction

The study of political science studies with a plethora of fundamental problems pertaining to the partisan division of the population of the United States. In recent years, the question of voter suppression has come up across the scientific and non-scientific communities alike. In particular, Democrats and Republicans both have claimed that their votes have been suppressed, or that there have been boundaries that disable them to vote. In this report, made possible by the American National Election Studies 2022 Pilot Study, we investigate one question: Do Democratic voters or Republican voters experience more difficulty voting? We perform a Wilcoxon Rank Sum Test on a summative variable including responses about difficulty related to different aspects of voting.

Conceptualization and Operation

To define a few key terms for the purposes of this report: a **voter** is defined as any individual who was eligible to vote in the 2022 election. **party** will be defined through identification on the ballot. If a voter is not registered as either Democrat or Republican, we will defer to their "lean", which they can self-identify.

For the purposes of this experiment, we defined **"experiencing difficulty voting"** as a aggregate sum of the "vharder_i" variables, where i=0,1,2,. . . ,11. These questions were of the form "Please indicate if any of the following made it harder for you to vote in the 2022 general election," with mulitple items pertaining to difficulty listed.

The summative vharder metric was chosen over options such as "How difficult was it for you to vote?" with the corresponding 1-5 Likert Scale variable "votehard,". This mitigates potential biases to using only one metric. As voting difficulty in reality is a combination of multiple factors, not just one. This also allowed a closer look at the true difficulty voters experienced, as opposed to perceived difficulty. Perceived difficulty is an important metric to capture as well; however, this experiment is focused on capturing difficulty as close to objectively as possible.

Data Wrangling

To perform our analysis: We transformed each of the "vharder" columns into a binary variable, where 1 indicated that the individual had experienced the factor contributing to difficulty voting, and 0 indicates that they had not. Then, we added a new column that takes the sum of the 12 transformed "vharder" columns to get our target variable. To determine whether the individual was a Democrat or Republican, we used the "pid_x" column. The column is on a scale of 1-7, 1 being "Strong Democratic" and 7 being "Strong Republican". Independents (pid_x =4) were dropped as it was not relvant to the research question. This only dropped 273 out of 1583 responses, still leaving a high sample size for analysis. Figure 1 shows a summary of this data. We did not drop n=223 individuals who responded, "definitely did not vote." We want to include individuals who had an extremely difficult time attempting to vote, to the point where they were unable to cast a vote in the end. These individuals would be at the extreme end of our target variable, meaning that excluding them could drastically impact our analysis. We acknowledge that this design choice could introduce confounding variables. By including individuals that did not have the intention to vote, we get a lot more scores of 0 in our summed "vharder" variable. If one party has more individuals that were not planning to vote than the other party, it could skew the result toward showing that party had an easier time voting. However, we weighed the trade-offs of not including this important data, and

decided to keep the records of those who answered "definitely did not vote.". Since higher sum = more difficult, democrats seem to observe a higher difficulty on average (Table 1). The statistical significance of this difference will be explored in the subsequent sections".

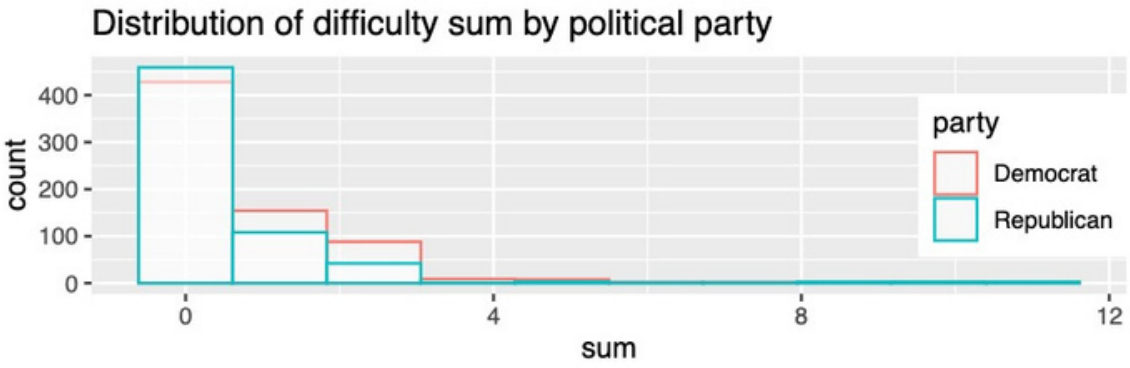

Figure 1: Distribution of difficulty sum by political party

Table 1: Means of difficulty sums.

party	mean_of_sum
Democrat	0.7208633
Republican	0.3902439

Hypothesis Testing

In our statistical testing, we used the aforementioned summative variable combining 12 binary values measuring voting difficulty. We wanted to compare two samples from two independent groups (Democratic and Republican voters) so the most appropriate statistical test for the testing of our hypothesis was the Wilcoxon Rank Sum test. The assumptions for this test are that the data must be ordinal, and independent and identically distributed. The summative vharder variable is non-continuous and has order (a score of 11 means the voter experienced more difficulty than one with a score of 1). We also can not meaningfully measure differences in difficulty between two scores (say, a voter who had a long wait time and a work schedule conflict vs. a voter with just a work schedule conflict). As such, the first assumption of an ordinal scale variable is met. Additionally, the distribution is not normal as can be observed in figure 2.

As the two groups we are interested in constitute the two major opposing political parties in the United States, there could be socio-political interactions within and between them that may cause some dependence among our two samples. However, each observation represents a unique voter with their own thoughts and experiences, and each voter had freedom to vote in a ballot and identify with either political party. Therefore the assumption of independence is met within our bounds of expectation. We used the responses of the participants of the American National Election Studies 2022 Pilot Study for our data. The survey for the study was conducted online; as each voter responded to the same survey, we expect each participant to be identically distributed from the online population. Although the online population may differ from the overall population of the United States, this would be true for all datapoints, therefore this does not affect the identically distributed assumption. Furthermore, the sampling from the ANES Study has been balanced according to the U.S. demographics. We conclude that the identical distribution assumption has been met. Therefore, we have met all the assumptions of the Wilcoxon rank sum test and we can proceed with our hypotheses. Our null hypothesis, in the context of our summative variable, can be stated as follows: The probability that a draw from the Democratic voters group ranks higher in the summative difficulty variable than a draw from the Republican voter group is not lower than the probability that a draw from the Republican voters group ranks higher than a draw from the Democratic voters group. Our alternative hypothesis is therefore that the probability that a draw from the Democratic voters group ranks higher in the summative difficulty variable than a draw from the Republican voter group is lower than the probability that a draw from the Republican voters group ranks higher than a draw from the Democratic voters group.

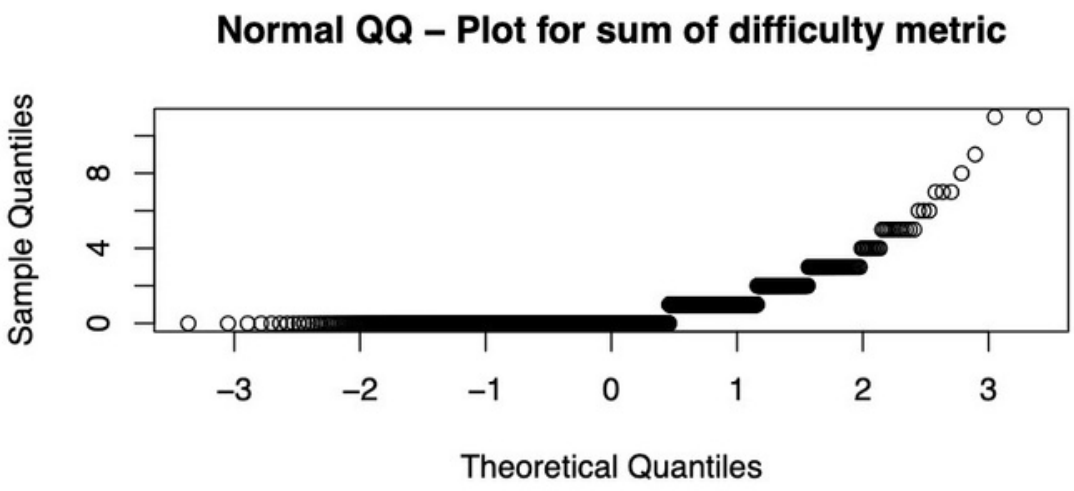

Figure 2: qqnorm plot to check for normality in data set. Due to the long tail it can be seen that the dataset is non-normal for the sampled population

We execute the One Tailed Wilcoxon ranked-sum test as follows:

```
wilcox.test(sum ~ party, data = voting_d_r, alternative = c("greater"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sum by party
## W = 244192, p-value = 3.263e-08
## alternative hypothesis: true location shift is greater than 0
```

Test, Results and Interpretation

For this analysis, H0 is that Democrats and Republicans do not have a different difficulty when voting.

It can be observed in the one tailed Wilcoxon rank sum test, with a p-value «< alpha of 0.01 (99% Confidence interval), that the populations have a statistically significant difference between them and H0 can be rejected. This leads us to the conclusion that Democrats had self-assessed more barriers to voting than Republicans in this election. The significance of this result indicate that more focus may need to be placed on projects surrounding ease of voting for Democrats in the short term, as higher perceived difficulty may result in lower voter turnout. However, this action must be carefully deliberated as unintended consequences may appear such as skewing election results. Future research on this topic may have to approach Democrats and Republicans as unique parties as they have different perceptions in difficulties in voting. For example, some additional research questions that have arose from this analysis are: how do perceived and actual difficulty differ between the two parties, and a granular analysis on each of the items in the aggregate.