Voter Difficulty between Republicans vs Democrats during 2020 Election

Kisha Kim, Sean Koval, Nicholas Brown, and Michael Townsend

Contents

1 Overview

Every election cycle, accessibility and ease of voting is on top of our minds in order to encourage all voters to participate in the election process. Various campaigns focus on voter turnout, as the recent elections have been decided by very small margins. In order to promote voter turnout, we first need to understand the 2020 presidential election. We, as a non-profit data science organization specializing in political science, have received 2020 election data sets to further this research. American National Election Studies (ANES) is a foundation that provides high quality data from its surveys on voting, public opinion, and political participation in order to support research needs. Restricted time series data from ANES was used for our analysis.

The analysis focuses on identifying whether voters from each party faced the same amount of difficulty in the 2020 presidential election. Our goal is to use data sets to compare difficulties that were faced in the 2020 presidential election, to help better prepare for the upcoming 2024 presidential election. Below key research question:

Did Democratic voters or Republican voters experience more difficulty voting in the 2020 election?

Results of this study could provide more in depth insight for the federal government election department, when creating policies on campaigns that impact voter turnout and ensuring that Democrats and Republicans voting difficulty is not biased. Furthermore, this could increase the equality in the voting process.

2 Key Terms & Limitations

Key Terms:

1. **Voter** - Defined who exercised the right to pick one of either Democrat or Republican candidates in the 2020 presidential election, pre and post-election
2. **Difficulty in Voting** - This is a ordinal measure (1-5 scale) that has been identified through individual voter experience, faced in the 2020 election process.
3. **Democrat vs. Republican** - We've classified Republican vs Democrat, by identifying voters who 'lean' in one direction or another as members of that party. This means that someone who "Leans Democratic" should be classified as a Democrat; and someone who "Leans Republican" should be classified as a Republican.

Understanding the ANES Time Series Study, the data type is in an ordinal scale. Given that difficulty was measured in a 1-5 scale, the difference between each scale (e.g. Very difficult vs Extremely difficult) is not really known. Sample size is about ~8000 (Total number of interviews: 8,280 pre-election; 7,449 post-election).

Limitations:

- Due to the nature of the test, this research will not be able to prove that the difficulty faced by Democrats is higher or lower than the difficulty of the Republicans. The analysis is limited to finding out whether either the difficulty is the same or not.
- The difficulty was measured by people who have participated in the interview. There might be difficulty faced by non-interviewers that this data set is limited to.
- High percentage in the data set has no difficulty in the election process (1 difficulty measure in the Interview result) amongst the Democrat and Republican. This leads to the high percentage resulting in the tie when comparing two measures, when performing the Wilcoxon Rank-Sum - comparison test.
- More than 50% of data collected fell into the category of "Unknown" political party subset, limiting the potential sample size of the data studied.

3 Data Wrangling/Analysis and Plots

To answer this question, we will utilize the data from the 2020 American National Election Study (ANES). The data set is observational and is based on a sample of survey respondents from the YouGov platform.

As a part of our analysis, we reduced the original data set to a smaller sample through data wrangling to eliminate unnecessary or incomplete information. The data set was reduced to four columns: "2020 Case ID", "Party of Registration (Pre-election)", "Voter Turnout 2020", "How Difficult it was to Vote". The most important variables in our analysis were difficulty voting (response variable) and the Party of Registration (explanatory variable). The variables for 2020 Case ID and Voter Turnout 2020 served to ensure that each voter had a unique ID in the data set and to validate that each respondent voted in the 2020 election.

The data set was further reduced by eliminating values that were associated with responses that would not provide useful information for our hypothesis test. Starting with the ordinal variable for difficulty voting, which is measured on a likert scale from -9 (refused to answer) to 5 (extremely difficult), information needed to be filtered out that was not useful to our analysis. We filtered out values from the Difficulty Voting variable < 0 since those values were indicative of a lack of complete data. The data was then separated into two distinct subsets by party affiliation.

Looking at Figure 4.2 the pie chart we noticed that there are a roughly similar number of Democrat and Republican respondents with a significantly larger set that fell into the "Unknown" political party. Roughly 57% of the trimmed data set contained respondents that belonged to neither the Republican nor Democrat subset. It is evident from the pie chart that much of the data collected regarding difficulty voting could not be used when comparing the subsets differentiated by party affiliation between Republican and Democrat. There was also a severe skew of the responses towards a voting difficulty of 1 (not difficult at all) as seen in the bar graph. When comparing both parties' responses we can see that for both the Republican and Democratic, a majority of the responses for voting difficulty were a 1 value. However, what we can not conclude from this observation is that those values are the same across parties since it is measured on a likert scale.Figure 4.1 depicts the difference in count between difficulty measured amongst Democrat vs Republican voters.

Below are key parameters used in the analysis:

1. **ID_2020(V200001)** - This is a case id.
2. **political_party(V201018)** - This field helps understand whether the voter is registered with any political party.
3. **voted_2020(V202109x)** - This field indicates whether the interviewer voted through a 3 number scale - 2. Not reported 0. Did not vote 1. Voted.
4. **Difficulty in voting(V202119)** - This field indicates how difficult was it for the interviewer to vote in this election.

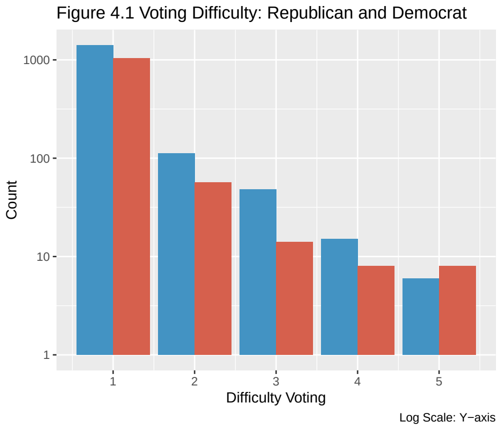

Figure 4.1 Voting Difficulty: Republican and Democrat

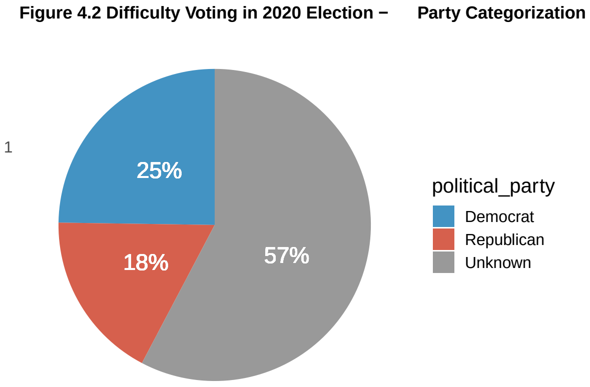

Figure 4.2 Difficulty Voting in 2020 Election − Party Categorization

4 Test Results and Interpretation

In order to confirm the test for this data set, we've evaluated a few steps. First, this is not two measures on the same sample given that one interview can be only either Democrat or Republican, not both. This allowed us to follow the unpaired test decision tree. Next, given the ordinal dataset, the data gathered is not in an interval or ratio scale. This has led us to choose to conduct the Wilcoxon Rank-Sum Test - Comparison version.

The Wilcoxon Rank-Sum - Comparison Test requires below assumptions to be true:

- **Ordinal Scale** – Each person answers the survey with a 1-5 rating describing how difficult it was to vote, representing ordinal data, so this assumption is valid.
- **I.I.D.** – Each draw from Republicans and Democrats will be from the same distributions and Republicans and Democrats here are mutually independent.

In this research, the null hypothesis that we are testing is below:

Null Hypothesis: The probability that a draw from difficulty faced by Democrats (D) ranks higher than a draw from difficulty faced by Republicans (R) is the same as the probability that a draw from difficulty faced by Republicans ranks higher than a draw from difficulty faced by Democrats.

$$(P(R > D) = P(D > R))$$

Alternative Hypothesis: The probability that a draw from difficulty faced by Democrats (D) ranks higher than a draw from difficulty faced by Republicans (R) is different from the probability that a draw from difficulty faced by Republicans ranks higher than a draw from difficulty faced by Democrats.

$$(P(R > D)! = P(D > R))$$

5 Results - Wilcoxon Rank-Sum Comparison Test

```
wilcox_test <- wilcox.test(Democrats$difficulty_voting, Republicans$difficulty_voting,
            data = voting_diff,
            paired=FALSE,
            exact = FALSE,
            conf.int = TRUE,
            conf.level = 0.95)
```

The test above ran on our data yields evidence to suggest that we should reject our null hypothesis since our p value is 0.0016 compared with our alpha assumption of 0.05. However, when looking deeper into the results the confidence interval of the difference between these two groups is extremely small (between -3e-05 and 3e-05). Since a high percentage (~90%) of the data answered a "1" for difficulty voting, we anticipate the Wilcoxon had a number of ties in it's Rank-Sum, leading to a p value that seems significant, but has small effect size in practicality. So, from a statistical perspective we are able to reject the null hypothesis, but in practice this effect size is far too small and the study needs to be improved in the future to yield actionable results.

6 Summary

Our study provided evidence that from a statistical perspective, both groups did not experience the same difficulty when voting in the 2020 presidential election. However, due to the type of test that we performed, the ordinal nature of the data, and the number of ties present in the question studied, the effect size we were measuring was far too small to have practical significance.

Our next step is to design a study that addresses some of the limitations that we encountered in this study to hopefully get a practical result in the future. To address these limitations, we hope to be involved in the complete data collection process, with the goal of collecting rich data that will have less ties to quantify the difficulty voting and be able to compare two groups more effectively. We also hope that in the next iteration of this study we're able to categorize a higher percentage of the population into the two groups and study more parameters / sub-populations to determine the optimal corrective actions to improve voter turnout.