

Part Two

Ken Trinh, Kesha Julien, David Corrales Garcia

2/14/2022

Contents

1	Introduction	1
2	Conceptualization and Operationalization	1
3	Data wrangling	2
4	Hypothesis testing	2
5	Test Results & Interpretation	3

1 Introduction

According to the U.S. Constitution, voting is a right. However, there have been recent allegations of voter irregularities and suppression during the 2020 presidential elections that made voting difficult for some Americans. The BBC report done in Georgia in October 2020¹ highlighted that restrictive voting laws and limited access to polling stations have led to some voters experiencing difficulty such as long wait times. During the 2020 presidential election, there was a historic number of popular votes (the total number of votes by citizens) with 155 million Americans or 66.8% of the citizens 18 years and older voting². Of this 66.8%, Biden received 81.2 million votes while Trump received 74.2 million votes. In spite of the historical voting turnout, still 33.2% of the voting population did not vote. Such statistics warrant an investigation on the question:

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

An answer to this question could provide insights into which party is experiencing more difficulty. It could also help shed some light on what is preventing those voters from voting. Addressing these issues could further increase voter turnout in future elections.

2 Conceptualization and Operationalization

In order to tackle the research question, we first need to clarify a few questions: 1) Who or what is a voter? 2) Who is a “Republican” and who is a “Democrat”? 3) What does it mean to experience difficulty in voting?

In this study, we conceptualize voters as American citizens who are eligible to vote or have voted in the 2020 election. Second, we define Republicans as individuals who identify themselves as strong, weak or leaning republicans, likewise for Democrats voters. Finally, we define difficulty in voting as any challenge that hinders a voter ability to vote for the 2020 presidential election.

Our analysis leverage data from the 2020 American National Election Studies (ANES)³. We planned to use some of the observational data collected as follows:

- First, to create a measurement for voters, we will use the data pertaining to individuals who have registered to vote and post election data on those that voted. We would merge these two columns with voters who have experienced difficulties in registering to vote. Treating the data this way, we would capture those that registered to vote or those that have experienced difficulties in voting.
- Second, to classify partisanship and based on a study done by Petrocik 2009, we will collect data on those who identified themselves as having strong partisanship, weak partisanship, or leaning toward a particular party. Here we will be grouping weak partisanship and those who lean toward a particular party in the same group. This is the best method for treating this part of the data because learners are partisans. Characterizing them as independents underestimates the partisanship of Americans and leads to inaccurate estimates of party effects and the responsiveness of the electorate to short-term electoral forces (Petrocik, 2009).
- Finally, to identify those who experience difficulties in voting, we will collect data based on a survey report of individuals experiencing difficulties to vote on a Likert scale from 1 to 5 with 1 being not difficult at all and 5 being extremely difficult. This information would enable us to measure the level of difficulties among individuals of each party.

¹Levinson-King, R. (2020, October 20). US election 2020: Why it can be hard to vote in the US. BBC News. Retrieved February 26, 2022, from <https://www.bbc.com/news/election-us-2020-54240651>

²Bureau, U. S. C. (2021, October 8). 2020 presidential election voting and registration tables now available. Census.gov. Retrieved February 26, 2022

³Home - anes: American National Election Studies. ANES | American National Election Studies. (2022, February 10). Retrieved February 26, 2022, from <https://electionstudies.org/>

3 Data wrangling

First, we isolated the following columns from the main dataset and renamed them as seen in Table @ref(tab:data-columns-identification).

Table 1: Data Columns Identification

Code	Name
V200001	Survey id
V202120a	Difficulty in registering to vote
V202119	Difficulty to vote
V201228	Party identification
V201229	Strong party identification
V201230	Leaning party identification
V202066	Whether or not they voted
V202051	Registered to vote

Second, we filtered out any data points that had negative values and labeled them as NA because according to the ANES dataset dictionary, negative integers mean that there was either no answer for this individual or the surveyor had experienced some form of technical difficulties when interacting with this individual.

According to our conceptualization of voters, we identified three columns that contain information on voters. These columns are people who experienced difficulty in voting, people who registered to vote, and people who have voted. Using these columns, we created the following conditional statements: Voters are people who have voted or have not voted and have registered or have not registered but have experience difficulty in registering. We stored this column as a Boolean statement named isVoter

To identify voter partisanship, we used the columns party identification, strong party identification, and leaning party identification to create we created a conditional statement to generate two dataset, republicans identified voters and democrats identified voters. We summarized the result as percentage of voters in Table @ref(tab:summary-table).

Table 2: Percentage of Voter by Partisan in 2020

	Democrats	Indepedents	Republicans
Eligible Voters	0.48	0.09	0.43

4 Hypothesis testing

To begin hypothesis testing, we formulate a null hypothesis and an alternate hypothesis. We decided to go with the alternative hypothesis that Democratic voters experience more difficulty than Republican voters based on our analysis of the data is represented in Figure @ref(fig:voter-difficulties). The barplot showed that the percentage of Democrats experience no difficulty is less compared to Republicans voters.

Null Hypothesis, (H_o): The probability that a Republican voter's level of difficulty is higher than that of a Democratic voter is the same as the probability that a Democratic voter's level of difficulty is higher than that of a Republican voter:

$$P(R \text{ level of difficulty} > D \text{ level of difficulty}) = P(R \text{ level of difficulty} < D \text{ level of difficulty})$$

Alternative Hypothesis, (H_a): The probability that a Democratic voter's level of difficulty is higher than that of a Republican voter:

$$P(R \text{ level of difficulty} < D \text{ level of difficulty})$$

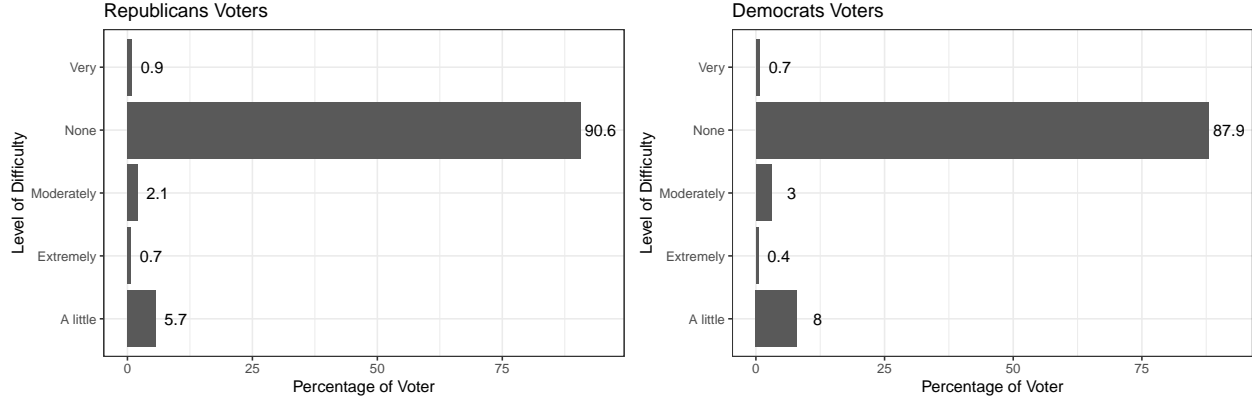


Figure 1: Percentage of Voters Experience Difficulties for Republicans vs Democrats

To test our hypothesis, we used the Wilcoxon Rank Sum Test (Hypothesis of Comparison). For the test to produce reliable results, the following assumptions must hold: 1. Ordinal Scale: - The collected data is based on a survey report of individuals experiencing difficulties to vote on a Likert scale from 1 to 5. At each rating, there is no continuous relationship between it and the next. This means that a rating of experiencing no difficulty at all is distinct from a rating of experiencing the most difficulty. Therefore, the condition of being ordinal scale is met

2. IID: Each X_i is drawn from the same distribution, each Y_i is drawn from the same distribution, and all X_i and Y_i are mutually independent.
- Let X be the Random variable for Republicans voters and let Y be the Random Variable for Democrats voters. Here, each draw of either X_i or Y_i is drawn from the same pool of people who are identified as voters in our dataset. So they all have the same underlying distribution. For each draw of X_i , there is no information on the next draw of X_i , the same case goes for each draw of Y_i . Therefore, the two samples are considered Independent and Identically Distributed (IID).

5 Test Results & Interpretation

```
wilcox_test <- wilcox.test(democrats_voters$difficult_to_vote,
                           republicans_voters$difficult_to_vote,
                           paired = FALSE,
                           alternative = "greater")
```

For our Wilcoxon rank sum test, we set X parameter to Democratic voters, the Y parameter to Republican Voters, and the alternative hypothesis parameter to greater, indicating that we want to test the alternative hypothesis that Democratic voters experienced a greater level of difficulty than Republican voters.

The Wilcoxon rank sum test result showed significant evidence that Democratic voters had a higher probability of experiencing difficulty in voting than Republican voters in the 2020 presidential election ($W = 3840387$, $p = 0.000855$). With our p -value being less than 0.05, we can reject the null hypothesis. This means that Democratic voters experienced higher levels of difficulty in voting than that of their Republican peers. Which seems to be in alignment with the allegations of voter suppression and irregularities noted in the 2020 presidential elections. By understanding the patterns of voting difficulties, law makers can enact initiatives to ensure that all voters are able to freely and fairly vote, potentially increase voter turnout in future elections.