W203 Lab 1 Analysis

Angela Guan, Meer Wu, Shanie Hsieh

Importance and Context

In the US election, we have come to face a recurring problem every voting cycle which is the difficulty voters face come election day. Taking a look into voter difficulty could be imperative to the outcome of elections. For example, if a single party faces more difficulty voting than another party, this could cause low voter turnout and consequently sway the votes to one party. We even see many articles now after the 2020 US elections talking about the bias and sway resulting from the voting process due to COVID-19. One article titled, "After Record 2020 Turnout, State Republicans Weigh Making It Harder To Vote [1]," talks about bringing more difficulty into the process in order to hopefully see the scale tip in favor of a specific party. This is especially significant in this past 2020 Election as we saw record high turnout despite the pandemic challenges [2] we faced. We want to use the ANES 2020 Time Series Study to see the impact of difficulty on voters. As the 2020 US Elections showed an intense debate with strong favors for both presidential parties, we want to take a closer look at the voting difficulty for Democrats and Republicans. Thus, our research question hopes to analyze this by asking:

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

To begin, we take a look into the key terms to look at in our research question to better apply these definitions to our data. First, we define voters as survey respondents who voted in the 2020 US Elections. In the ANES 2020 Time Series Study, we looked at variables V201028 which asks "Did R vote for President" for pre-election surveys and V202066 which asks "Did R vote in the November 2020 election" for post-election surveys. Next, we define party affiliation as voters who indicated they were Democratic OR identified closer to the Democratic Party were analyzed as Democrats and Republicans OR those who identified closer to the Republican Party were analyzed as Republicans. We looked at variables V201228 which asks "Does R think of self as Democrat, Republican, or Independent" and V201230 which asks "No Party Identification - closer to Democratic Party or Republican Party." Lastly, to identify "difficulty", we decided to use the question from the voter survey that directly asks a participant to rate from 1-5 how difficult it was to vote. This last variable was V202119 on the Time Series Study which asks "How difficult was it for R to vote."

Data and Methodology

Our data comes from the 2018 American National Election Studies (ANES). The ANES 2020 Time Series Study is a continuation of the series of election studies conducted since 1948 to support analysis of public opinion and voting behavior in U.S. presidential elections.

All respondents were assigned to interview by one of three mode groups—by web, video, or telephone. The study has a total of 8,280 pre-election interviews and 7,449 post-election re-interviews.

Most of our data wrangling involved filtering out respondents who did not fit into our operational definition of voters, party affiliation, and voting difficulty. First, we selected and renamed the 5 columns we needed

---

[1]National Public Radio. "After Record 2020 Turnout, State Republicans Weigh Making It Harder To Vote." (2021).

[2]United States Census Bureau. "Despite Pandemic Challenges, 2020 Election Had Largest Increase in Voting Between Presidential Elections on Record." (2021).

from the entire ANES dataset. Second, we filtered for those who indicated they had voted in either the pre-survey (value 1 = "Yes, voted for President") or the post-survey (value 4 = "I am sure I voted"). Third, we created a binary column "is_democrat" by first filtering for everyone with a party affiliation then assigning a 1 to those who identified as Democrats and 0 to those who identified as Republicans. Any person who identified as a Democrat/Republican or anyone without a specific party identification (eg. indicated they had no party preference or identified as an Independent) but felt closer to being a Democrat/Republican was classified into the group affiliated with the Democratic party. The same logic applied to those affiliated with the Republican party. Finally, we filtered out the respondents who didn't have a difficulty score between 1 and 5 in the question "How difficult was it for you to vote in this election?" since we were only interested in learning about the difficulty in voting experience of this demographic. We then dropped any columns that didn't add new information to the columns we kept.

The cleaned data contains 5,821 survey respondents who indicated that they have already voted- either in pre- or post- survey, expressed their party affiliation (including those leaning closer to either Democrats/Republicans), and gave a difficulty score from 1 to 5 on the post-survey question, "How difficult was it for you to vote?" As seen in figure 1, 54% (n = 3,126) of those respondents identified as a Democrat, while 46% (n = 2,695) identified as a Republican. In order to better understand how scores distribute across the two groups, we plotted two histograms showing the distribution of difficulty scores for each of the two groups- Democrat voters and Republican voters. Distributions from both groups are very similar. Both have a heavy left tail, indicating that most Democrats or Republicans experienced no difficulty voting. At first glance, it is hard to tell whether there is a difference in voting experience difficulty for the two groups. This motivated us to dive deeper into this topic using a statistical test.
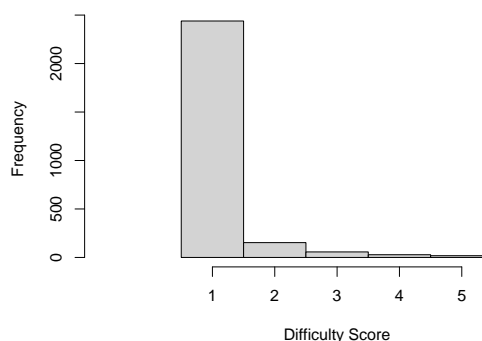
Visualizations

```
##    Party Affiliation Count
## 1          Democrat  3126
## 2        Republican  2695
```


[Figure 2] Distribution of Difficulty Scores for Democrat voters
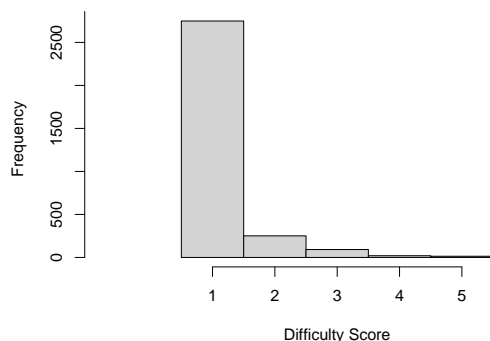
[Figure 3] Distribution of Difficulty Scores for Republican voters

From Figure 2 and Figure 3 above, we observe that the distribution of difficulty scores for Democratic voters and Republican voters appear very similar. Thus, we designed our hypothesis test to explore whether there is a difference in voting difficulty between the two groups, using a two-tailed test. We proceed our analysis with a two-sample unpaired hypothesis test: the Wilcoxon Rank-Sum test. The null hypothesis for the Wilcoxon Rank-Sum can be expressed as follows:

> **Null Hypothesis:** *In the 2020 election, there was no difference in difficulty voting experience between Democratic voters and Republican voters.*

The Wilcoxon Rank-Sum test requires the following assumptions to be true:

1. **Ordinal variable** Our variable of interest is voting difficulty, which is a categorical variable on the ordinal scale. The variable of interest ranges from 1 to 5, inclusive, and we represent voting difficulty by a score that answers the question: "How difficult was it for you to vote in this election?" The ordinal variables are represented by 1 being "Not difficult at all" to 5 being "Extremely difficult." The difference between "very difficult" and "extremely difficult" versus the difference between "not difficult at all" and "a little difficult" is not necessarily equal. In addition, a score of 3.5 is not interpretable since it lies between 3: "Moderately difficult" and 4: "Very difficult."

2. **i.i.d data** The data must be generated via an iid, or independent and identically distributed, process. The ANES 2020 Times Series data uses a panel of individuals from the YouGov platform. There could be potential dependencies or clusters in which individuals produce similar responses to the survey since a participant might share YouGov with their friends and family members with similar behavior or affiliations. Nevertheless, YouGov claims that they have millions of users, suggesting relations between participants should be rare.

Results

```
test <- wilcox.test(voters$difficulty, voters$is_democrat, paired = TRUE, alternative = "two.sided")
```

The test yields that there is no evidence that there was no difference in difficulty voting experience between Democratic voters and Republican voters in the 2020 US election (p-value < 2.2e-16). From the hypothesis test, we have a p-value well below zero, and so, we reject the null hypothesis.

Some limitations of our test affect the conclusions drawn from the analysis. Through this hypothesis test, we can only measure associations of the level of difficulty between Democratic voters and Republican voters, not causations. Additionally, our analysis examines data collected from pre-election and post-election surveys, suggesting that our definition of a voter is a proxy for individuals who voted. Our variable of interest, which captures how difficult it is for an individual in our data to vote, also contained options where the individual refused to answer, had no post-election data or had an incomplete interview, no post-election interview, an interview breakoff, or was inapplicable. We narrowed the scope of our analysis to focus on our audience of interest: voters who indicated a level of difficulty in voting.

Discussion

From the 2020 US ANES election data, we analyzed voter difficulty between Democratic voters and Republican voters. Through our analysis and statistical test, we conclude that there is no evidence that there was a difference in voting difficulty between the two groups of voters.

In future steps, we hope to analyze which party has more voting difficulty. In addition, we hope to have a more robust definition of voters and party affiliation to accurately capture the voter's characteristics. For instance, more feature engineering can be done to capture voting behavior of lean Democrats or lean Republicans. Taking this one step further, we can tap into what specific difficulties individuals faced in voting.

Voting difficulty has been a challenge to overcome that may be significant to the outcome of the election. Our analysis takes a deeper look at voter difficulty in the 2020 US Election to analyze how this factor was seen by Democrats and Republicans. Our results can be used to take further action in analyzing the 2020 Election results as well as aid in the future US elections.