Vote of Confidence: Assessing Statistical Significance of Trends in Voting Difficulty

DATASCI 203: Lab 1

Aman Kumar, Ayodele Oyewole, Assaph Aharoni, Gia Nguyen

2024-03-04

Contents

1 Importance and Context

Presidential elections in the United States have become increasingly close contests over the last 35 years. In the first ten presidential elections after World War II, the winning candidate won the popular vote by more than nine percentage points six times; in the nine presidential elections since then, no candidate has won by that margin. Had 43,000 votes in certain states in 2020 switched candidates, three of the last six elections would have been decided in favor of the candidate who lost the popular vote [1].

In the context of such narrow margins, ease of voting has become a crucial issue. In recent years, states have enacted laws that some analysts believe are intended to reduce access to voting. Since January 2021, 18 states have enacted over 30 separate laws to restrict voting by mail, limit early voting, and provide broader authority for vote purging. Over 400 similar bills are currently under consideration. Although an often-cited reason for these bills is voter fraud, some commentators believe these bills are intended to reduce votes for Democrats, particularly from people of color [2].

In this analysis, we attempt to contribute to discourse surrounding voting difficulty. In particular, we endeavor to assess whether a statistically significant difference in perceived voting difficulty exists between Democrats and Republicans through the following research question:

Do Democratic voters or Republican voters experience more difficulty voting?

The answer to this question may be useful in assessing recent measures to limit voting access. Better understanding the difficulty in voting could lead to fairer laws, higher turnout, and more public confidence in election results.

2 Data and Methodology

Our analysis relies on data from the 2022 American National Election Studies (ANES) Pilot Study. This study is an observational cross-sectional survey of US citizens aged 18 years or older conducted between November 14, 2022 and November 22, 2022. The survey data contain responses from 1,585 individuals with survey weights to generalize to the US population [3].

To operationalize our research question, we need to define certain key terms in our research question. First, we identify a "voter" as anyone eligible to vote (i.e., any individual observed in the ANES dataset). An alternative definition would include only those who actually voted in the relevant election (in this case, 2022). However, our research pertains to difficulty in voting, and those who had the most difficulty voting may have ended up not voting at all. Therefore, they should not be excluded.

Second, we identify "Democrats" and "Republicans" based on their responses to questions regarding party affiliation and party "lean," as captured in ANES field pid_x. We rely on party lean as opposed to only explicit identification because academic research suggests that, although up to 40% of Americans identify themselves as independents, very few lack a party preference [4]. We drop from our analysis the 273 observations associated with individuals labeled as "independents." We also drop 2 observations with missing political party affiliation data. We mark the remaining 1,310 observations as "Democrat" if they at least lean Democratic and "Republican" if they at least lean Republican. Table 1 shows a summary.

Table 1: Observations Analysis

Cause	Number of Samples Available for Analysis (after removal for cause)	Remove number samples for cause
Start	1,585	N/A
Missing values on control variables (political parties)	1,583	2
Independents (non-partisans)	1,310	273

Finally, we need to define a measure of "difficulty voting." There are two sets of directly relevant questions in the ANES survey regarding ease of voting. The first set consists of a question that asks participants to rank their voting difficulty between "Not difficult at all" and "Extremely difficult" on a Likert scale (as captured in the field votehard). However, this question was asked only of those who voted, and as discussed previously, it is essential for our analysis to include those who did not end up voting. The second set of questions consists of a series of binary multipunch questions regarding issues that might make voting harder, such as difficulty with registration, long wait times, and bad weather (as captured in the fields vharder0 through vharder11). To operationalize our analysis, we therefore define "difficulty voting" as the sum of the number of difficulty areas selected by respondents. For example, a respondent selecting only their work schedule as a difficulty area would receive a difficulty score of 1, whereas a respondent selecting difficulties with registration, wait time, and bad weather would receive a difficulty score of 3. A summary of how we define these measures are shown in Table 2.

Table 2: Measure Definitions

Term	Features	Reason	Codebook Definitions	Data Manipulations
Political Party Affiliation	pid_x	This was the summary of party ID given in the codebook that shows whether a participant identified as Democrat or Republican as well as their lean. This feature was the most direct indicator of party preference.	1: Strong Democrat, 2: Democrat, 3: Lean Democrat, 4: Independent, 5: Lean Republican, 6: Republican, 7: Strong Republican	We map the values 1, 2, 3 to a value of 0 for Democrat, and map the values 5, 6, 7 to a value of 1 for Republican. We drop all rows with value 4.
Difficulty Voting	vharder_0 to vharder_11	These fields are binary variables, qualifying as metric data and capturing the breadth of difficulties experienced by voters.	1: Selected, 2: Not Selected	To generate binary variables, we map Not Selected to 0 and Selected to 1 for each field. We generate the count of issue experienced by each person by summing the binary variables.

Before generating the discussed difficulty scoring, we test whether any of the difficulty areas captured by vharder0 through vharder11 are correlated with one another. For example, suppose that everyone who checks "takes too long to get to the polling place from where I live" also selects "costs too much for transportation to my polling place" because those issues are directly related. In that circumstance, we would not want to double count these responses in generating the difficulty score. Therefore, we begin our analysis by assessing the correlation between answers to vharder0 through vharder11 in a correlation matrix. We determined to take the maximum response across fields with a pairwise correlation greater than 0.75. However, as Figure 1 shows, no pair of response categories yields a correlation this high.

Therefore, we generate a difficulty score for each individual using the sum of all fields vharder0 through vharder11. Finally, we run a two-tailed two-sample t-test to assess the difference in average difficulty score for Democrats versus Republicans. The null hypothesis of our t-test can be phrased as follows:

> *The difference between the mean difficulty score for Democrats and the mean difficulty score for Republicans is equal to zero.*

To ensure relying on the t-test is appropriate, we confirm the assumptions for the test are met. Namely, the following must be true: the data must be IID, measured on a metric scale, and be drawn from a normal distribution.

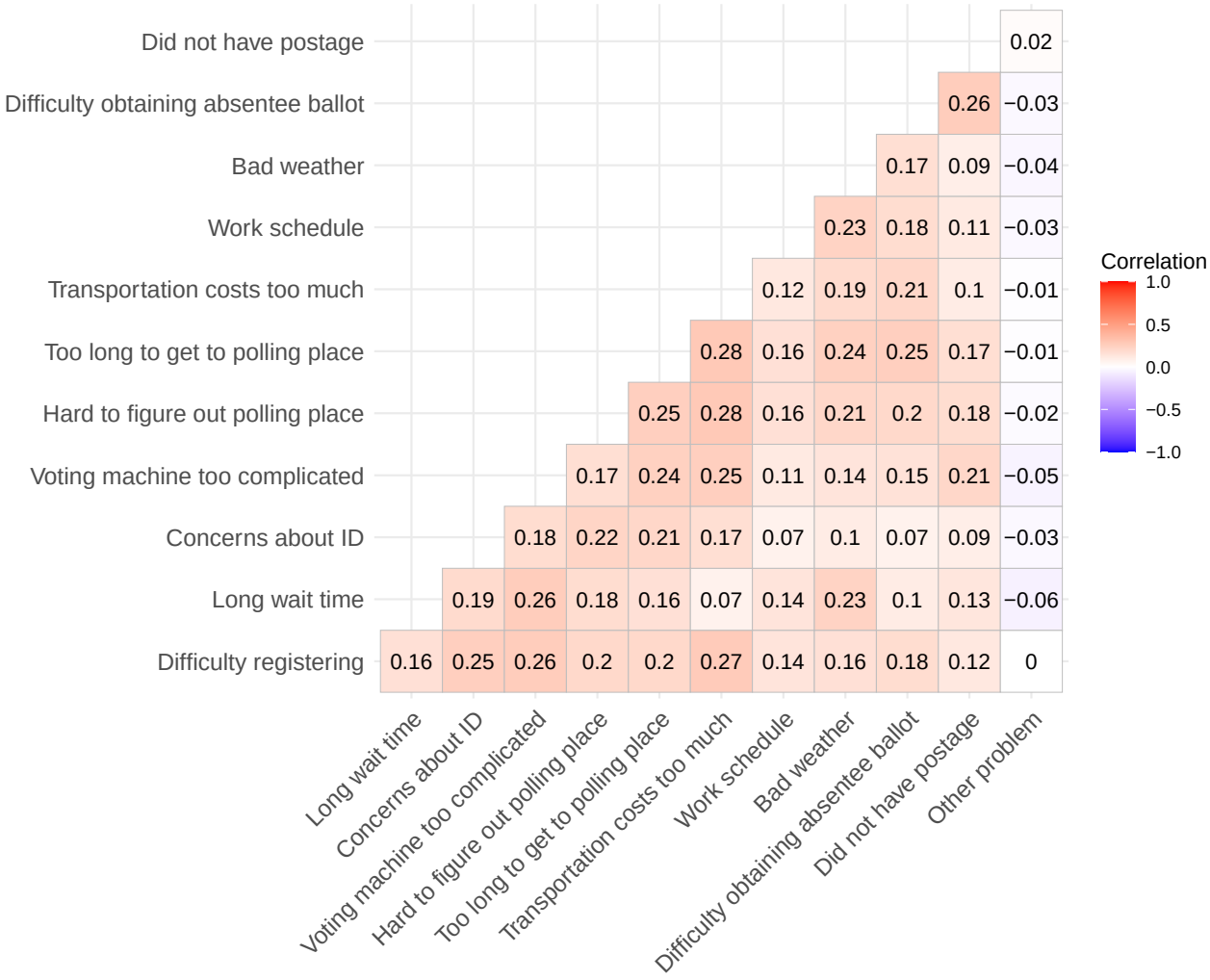

Figure 1: Correlation Matrix Illustrating Associations Between the Difficulty Variables

IID: The 2022 ANES study uses the YouGov platform, which rewards participation. It's possible that participants share information about the survey with their friends and family, leading to unexpected dependency clusters. But given the platform has millions of users, links between individuals should be rare.

Metric scale: The difficulty score we constructed is a ratio metric. In other words, it contains a true zero value (no difficulty in any category), and there is an equal distance between adjacent values.

Normality: With a sample size of 1,310, based on the Central Limit Theorem, the sampling distribution of the sample mean will be approximately normally distributed, even if the underlying data is not (skewness = 3.55). Therefore, the sampling data meets the normality assumption.

3 Results

```
t.test(difficulty_score ~ party, anes_selected)
```

The t-test provides evidence that Democrats experienced more difficulty voting than Republicans ($t = 5.48$, $p = 5 \times 10^{-8}$). At an alpha level of 0.05, we therefore reject the null hypothesis that the difference in voting difficulty between Democrats and Republicans is zero.

In terms of practical significance, Democrats had an average difficulty score of 0.72, about 85% higher than Republicans' average score of 0.39. Additionally, Cohen's d shows that Democrats' difficulty score is 0.296 standard deviations higher than Republicans'. Although the effect may appear small, with razor-thin election margins, even minor differences in voting difficulty could impact voter turnout and affect election results.

Our test does contain limitations that should be considered. We measure the difficulty score as the number of difficulty categories faced by a voter. However, it is possible that one category of difficulty poses greater difficulty than another category. Additionally, one of the difficulty choices is "other problem" (vharder11). The "other problem" may contain multiple categories that are not captured in our analysis.

4 Discussion

Studies of voting difficulty have become critical in US political discourse that is more divided and closely-contested than ever. Our study found evidence suggesting that difficulties may be affecting Democrats disproportionately relative to Republicans. However, further research into this topic is necessary.

Future studies would benefit from integrating a measure of the specific difficulty level associated with each difficulty category, rather than solely relying on its presence or absence. Additionally, we suggest that more information be collected on non-voters' reasons for abstaining from voting. This data dimension would help measure whether difficulty in voting actually led to lower turnout, as opposed to other considerations (e.g., not having an attractive candidate for whom to vote). Finally, this study could be further refined to more directly account for the sampling weights provided in the ANES dataset.

5 References

[1] Lindsay, J. M. (2024, February 23). CAMPAIGN roundup: Close presidential elections have become the norm. Council on Foreign Relations. https://www.cfr.org/blog/campaign-roundup-close-presidential-elections-have-become-norm

[2] The importance of protecting voting rights for voter turnout and economic Well-Being. (2021, November 30). The White House. https://www.whitehouse.gov/cea/written-materials/2021/08/16/the-importance-of-protecting-voting-rights-for-voter-turnout-and-economic-well-being/

[3] For purposes of our analysis, we did not rely on the sampling weights (as instructed in the lab assignment).

[4] Petrocik, J. R. (2009). Measuring party support: Leaners are not independents. Electoral Studies, 28, 562-572.