Lab 1: Question 1: Are Democrats voters older or younger than Republican voters in 2020?

Elaine Chang, Dom Dillingham, Jesse Miller, Michael Wang

1. Importance and Context

Politics often use basic demographics as a lens to better understand and plan around voters. Political operatives - e.g., current / aspiring politicians, political consultants, third-party watchdogs to name a few - may use the variable "age" as a specific demographic field to better understand their voters and highlight differences between the two major parties. This can impact strategic functions such as policy formation or tactical initiatives such as marketing outreach so as to better engage voters. For example, Democrats have taken up **canceling student debt** as part of their policy mantle with institutional support very recently, though historically it sat as a fringe idea, popular amongst Millennials. Thus understanding if there is an age difference, and further what the age difference is, between Democratic and Republican voters, can help political operatives better navigate their party towards a path to power.

2. Description of Data

The data used in this analysis is from the preliminary 2020 American National Election Studies (ANES) Time Series Study, released February 11, 2021. The ANES 2020 Time Series Study reflects a total of 8,280 pre-election interviews conducted by one of the three mode groups - by web, video or telephone. Readers are encouraged to refer to the **ANES full codebook** for additional queries on the original data that is available online.

In operationalizating our research question ("Are Democrats voters older or younger than Republican voters in 2020?"), we sought to clearly define the variables we were most interested in. While the ANES study encompasses a dataset of over 700 different variables, we identified three variables in ANES that is most pertinent to investigating our research question.

a) Age. The survey asked respondents for their age, which will be used as the primary response variable. We identified variable "V201507x" as most suitable. Any responses of "-9" which maps to "Refused" were filtered out of the dataset due to the core consideration factor that age factored into this research question. Ages were reported in the survey as their true value from ages 18, the youngest eligible age to vote, to age 79. Respondents over the age of 80 were grouped together and reported as one value. While this may introduce some bias in the analysis, we expect their contribution to the overall age distribution to be relatively small because the total number of Democrats and Republicans respondents over 80 years old are similar in amount (178 and 195, respectively).

b) Party Affiliation. In defining "Democrat" and "Republican" groups, we relied on the respondent's preference for presidential candidate (captured in V201075x). There were several contending interpretations to define party affiliation such as party registration (V201018 "Party of registration"), self-identification (V201228 "Does R think of self as Democrat, Republican, or Independent?"), or preference for a selected office (e.g., V201070 "For whom does R intend to vote for governor?"). However our qualitative research found that party registration may be an unreliable indicator. For example, an **analysis by DC Report** of Mitch McConnell's recent re-election in 2020 discusses that conventional political wisdom warns that "analysts shouldn't correlate party registration with voting patterns." Similarly, we ruled out the party self-identification variables as being too subjective on voter interpretation, without the specificity of action weighed with it. Finally, we chose a variable that tied action and preference to the tangible outcome of the presidential office as that reflects the national leaders of their respective major parties. The presidential candidate for each party increasingly define party identity boundaries a la Donald Trump and the GOP. Down ballot (e.g., Congressoinal) and local offices could introduce complications on state or city specific issues.

c) Voters. For the purposes of this survey, our team wanted to pointedly analyze the fullest scope of voters with an emphasis on respondents who will cast a ballot. However because this data is pre-election, it is unlikely to be accurate in its capture of respondents as voters at the time of data collection. When the full data set for 2020 is released by ANES in a future date, we could re-analyze the data. In the meantime, we defined voters by capturing respondents who voted, intended to vote or preferred a presidential candidate (V201075x) with an additional filter of those who also intended to register to vote (V201019). Together, this implied enough knowledge and action on behalf of the respondent to create our definition of "voter."

We plot the ages of voters in the Democrat and Republican groups in Figure 1 to generate basic observations of the dataset. We observe a relatively greater population of younger voters in the Democrat group, although tests need to be performed to determine statistical significance. We also note that the data in both groups do not appear skewed, and when combined with the large data count, assures general normality of distributions.

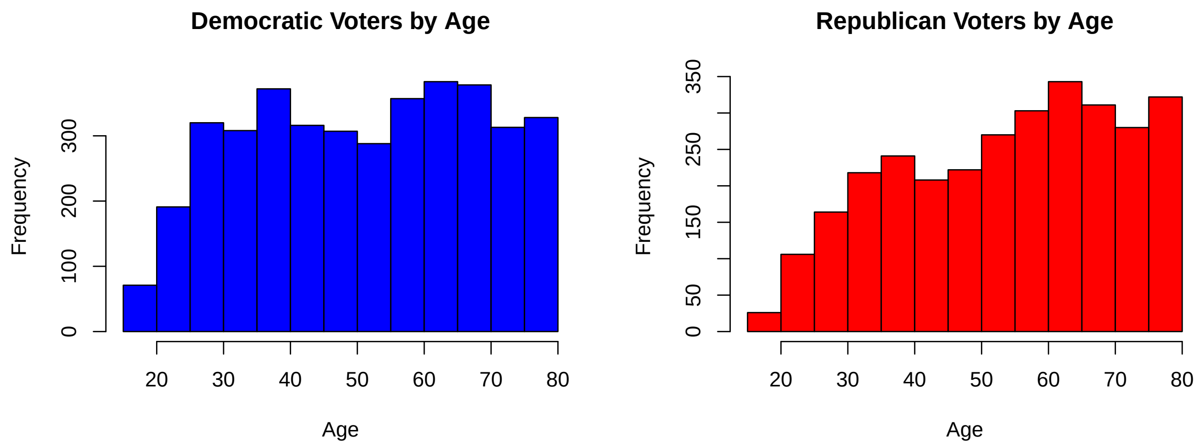

Figure 1: Age Distributions of Democratic and Republican Voters

3. Most appropriate test

We are interested in whether the average age of a Democratic voter differs from the average age of a Republican voter. As such, we determined that the most appropriate statistical test is the Welch two-sample two-sided t-test. This will be an unpaired t-test as the variables are independent with no known systemic dependencies. It satisfactorily fulfills the following core requirements of the Welch two-sample two-sided t-test:

a) Metric data. The variable measuring age (V201507x) is on a metric scale.

b) Approximately normal distribution. The histograms constructed in Figure 1, which voter age distributions across the two parties, do not show any obvious skewness. While the distribution of age is not a definitively normal distribution, the sample size is large enough such that the Central Limit Theorem will approximate a normal distribution and compensate for any skewness.

c) Data drawn IID. It is unlikely that the sample of survey respondents drawn and reported on fully satisfy the requirement of IID. Per the ANES userguide / codebook, weights are to be used to mitigate sampling bias identified by the study's researchers. The large sample size drawn across the 50 states provide some assurance in some partial fulfillment of IID. Additionally, per instructions for this research assignment, weights will be ignored. Therefore, we will proceed with our analysis assuming the data fulfills the IID requirement.

In addition, it is worth noting that we explicitly chose to continue with Welch's t-test as opposed to the Student's t-test. Similar to many different statistical programs, R's default t-test is set to the Welch t-test. In designing our test, there was no obvious indication to alter the default Welch t-test in R the Student's t-test. We would rather withhold applying assumptions on the sample variances. Cursory reports by others interested in the differences between these two t-tests show that the Welch t-Test performs better than the Student's t-Test when sample sizes and variances are unequal between groups but gives identical results when variances are equal. Thus there does not appear to be an advantage in making upfront assumptions regarding variances and move away from the Welch t-test.

Our hypotheses for this test will be:

- H0: The average age of Democratic voters and the average age of Republican voters are the same
- Ha: The average age of Democratic voters and the average age of Republican voters are not the same

```
d = dat[dat$Voting_Party == 'Democrat' ,]$Age
r = dat[dat$Voting_Party == 'Republican' ,]$Age
t.test(d,r)
```

```
##
##  Welch Two Sample t-test
##
## data:  d and r
## t = -7.3618, df = 6595.7, p-value = 2.036e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.805399 -2.204937
## sample estimates:
## mean of x mean of y
##  51.39496  54.40013
```

4. Test, results and interpretation

The results of our Welch two-sample t-test show that the p-value is 2.036e-13. Comparing this p-value to the industry-accepted alpha level of 0.05 on a two-sided test, this test suggests that there is a "very highly significant" statistical finding to reject the null hypothesis that Democrats and Republicans have the same average age. Statistically, out of all confidence intervals constructed according to our procedure outlined, we are 95% confident that Democrats are 2.2 to 3.8 years younger than their Republican counterparts.

While statistically there is strong evidence to suggest a difference in average age between the two party voters, there seems to be limited practical application of this procedure and result. The null hypothesis we rejected was that Democrat voters and Republican voters do not have the same average age. We can infer based on this that Democrat voters are indeed younger than Republican voters. However the test also shows that the actual difference in age is a relatively narrow margin of approximately three years on average. Thus on a practical level, the application of this knowledge is limited. Policy direction, candidate messaging, marketing outreach and the like will not be swayed using a strictly age-related lens if the basis of that difference is three years between the two political parties.

5. Limitations

There were several limitations to the data and analysis presented that should be noted.

- It should be emphasized that this dataset is both a recent and preliminary release of the 2020 ANES Time Series Study. As the data is updated, our results may be updated.
- As mentioned earlier, weights were encouraged by the ANES user guide and those were not applied, per assignment instructions and time limitations.
- There are two macro events to highlight in 2020 that could cause irregularities in the data. First, the spread of COVID-19 throughout the 2020 year. This would impact nearly all stages of the data collection and preparation as the study researchers modified the design of the survey and portions of its content to adapt to the COVID-19 pandemic. Second, the 2020 election was highly unusual in its vote-by-mail process as well as then-President Trump's false accusations in the untrustworthiness of that process. This repeated claim may have caused respondents to opt-out of voting and this would impact our "voters" group in the analysis.