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

Tilman Bayer, Kieran Berton, Ruilan Liu, Toby Petty

Importance and Context

The notion of there being a relationship between age and political ideology is a persistent belief, often held by older generations lamenting the changing political opinions of younger generations. Many cite evidence for there being a connection between the two by pointing to young community leaders spearheading generation-defining social movements, which inevitably come into conflict with older, more established systems. But these anecdotal examples do not provide solid evidence for there being a connection between one's political leanings and one's age.

Using data from the American National Election Study (ANES) 2020 Pre-Election Survey, we will try to answer the question of whether there is a relationship between one's party affiliation and their age. We will compare populations of self-reported Democrats and Republicans to understand if there is a statistically significant difference in the average age of the two parties' voters.

Description of Data

To answer our research question, we will use data from the ANES survey of voters prior to the 2020 General Election. This dataset[1] comprises responses from 8280 respondents to approximately 700 questions concerning their political opinions, voting behaviors, as well as some qualitative questions about their opinions on social issues[2].

We interpret the question as referring to anyone eligible to vote in the 2020 elections, regardless of whether or not they actually voted in that election or made steps to do so.

The survey's target population is defined as "the 231 million non-institutional U.S. citizens age 18 or older living in the 50 US states or the District of Columbia" (for the "fresh cross-section" part, which is combined with respondents drawn from earlier ANES surveys which had similar target populations). This matches the definition of "eligible voter" rather well - unsurprisingly, considering that this a survey conducted with the specific purpose of providing data for election studies. The "non-institutional" condition excludes a large group of ineligible voters, namely institutionalized (incarcerated) citizens who are disenfranchised as felons. Limitations include the non-inclusion of military personnel on active duty (which count as institutional too[3]) and US citizens living overseas.

Our decision not to consider whether someone actually voted in this 2020 election is partly a pragmatic choice due to various limitations of the dataset: since the survey took place before the actual election day, many or most actual voters will not have voted yet at the time of the survey (despite voting by mail having become more frequent in 2020). As a proxy, we considered using the information as to whether a respondent has registered to vote - the survey asked several question about their registration status. However, we concluded

[1] Available at https://electionstudies.org/data-center/2020-time-series-study/ . As of February 11, 2020, ANES describes this as a preliminary release, but we assume that they have already done considerable work on vetting the data and have thus not conducted a systematic search for e.g. fake responses ourselves.

[2] As advised by ANES, an accurate analysis based on this data would have to account for the survey weights included in the dataset. Per the instructors, we are ignoring them for the purposes of this labs project.

[3] See e.g. https://en.wikipedia.org/wiki/Civilian_noninstitutional_population

that the risk was too great that this would distort the data, as no less than 21 states offered same-day registration in the 2020 elections[4].

Now, let's operationalize **"Democratic voters"** and **"Republican voters"**. The dataset contains several variables providing information about the respondent's support for either the **"Democratic"** or **"Republican"** party, with varying levels of response rate. These range from broad self-categorizations of party identification, to specific questions about how respondents voted in House, Senate, and presidential elections. For example:

- **V201228** PRE: PARTY ID: DOES R THINK OF SELF AS DEMOCRAT, REPUBLICAN, OR INDEPENDENT (2448 Republican, 2786 Democrat)
- **V201029** PRE: FOR WHOM DID R VOTE FOR PRESIDENT (109 Republican, 260 Democrat)

We decided that V201228 (the respondent's self-identification as Democrat or Republican) is the most suitable data for the purposes of this question, as it provides the broadest definition of what it means to be either a 'Republican' or 'Democrat'. Preference for specific candidates or voting behavior in a specific election may differ from general party affinity (e.g. think of "Never-Trumper" Republicans).

To assess which of the two groups is younger or older, we aim to compare their mean age using standard statistical tests (see below). The survey data provides the respondent's age (as of 2020) in integer years. It has some limitations that we discuss in more detail below. For privacy reasons, the precise dates of birth have been removed from the survey data available to us. More importantly, for respondents aged 80 and over, the data is binned to a single "Age 80 or older" value.

We first dropped 354 rows from the dataset containing responses from survey participants who refused to provide their age. This leaves a total of 7926 rows of data from respondents who provided their age.

Having decided to use metric V201228 ("Party ID") for our analysis, we also drop an additional 2692 rows from the dataset from participants who either did not respond to this question for technical reasons or refused, or claimed to be independent/have no preference, leaving us with 5234 rows (from an original total of 8280).

To get a sense of the overall distributions of ages of Republican and Democratic voters in our sample, we plot a histogram of the number of voters of each unique age in the dataset. We set the "80 or older" values to 80 for the time being, resulting an a spike at the right end of each histogram:

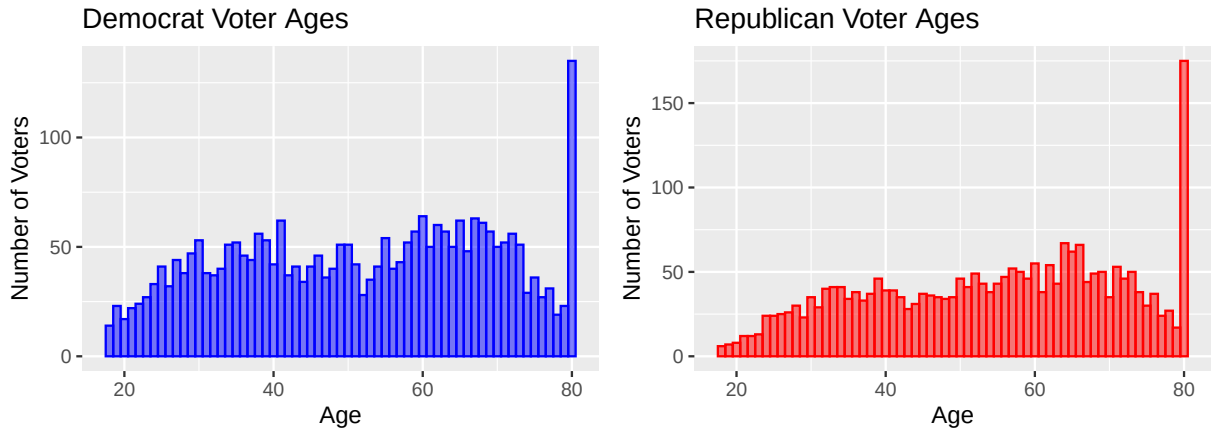

Figure 1: Ages of Respondents

The respondent's age in (integer) years appears to be precise enough for our purposes. But the binning of the age 80 and over answers presents a challenge for our plan (justified in more detail below) to use a t-test to answer the question: its test statistic is calculated from the sample means and variances of this data, which

[4]https://www.ncsl.org/research/elections-and-campaigns/same-day-registration.aspx

might differ considerably depending from the (unknown, to us) actual age distributions among those aged 80 and over. After performing a search of relevant literature, we learned that there are several generic statistical techniques available to handle such binned data, e.g. in the context of regression analysis (Tobit models[5]). But since this was a one-week project confined to exercising what we learned in W203 so far, we resorted to simply setting the "80 or older" values to 80 for the calculation, so as make us of all available information, even at the cost of some statistical accuracy

Most appropriate test

We are going to use a two-tailed unpaired two-sample t-test to assess the following null and alternative hypotheses:

- **Null Hypothesis**: There is no statistically significant difference between the mean ages of Republican and Democratic voters.
- **Alternative Hypothesis**: There is a statistically significant difference between the mean ages of Republican and Democratic voters; Democratic voters are either younger or older on average than Republicans.

We will assess the statistical significance of our test using an alpha value of $\alpha = 0.05$, a standard value used in many political science analyses.

The validity of the test relies on the following assumptions: * We are applying it to a metric random variable: Age (derived from physical time). * The samples are independent and identically distributed: Due to the survey setup, we can assume that one respondent's age doesn't influence another one's (ANES did not use snowball sampling, for example). * Our sample size is large, and while the above histograms indicate that the distribution is unlikely to be perfectly normal, it is also not very highly skewed (being confined within the ages of 18 and 116, the age of the oldest known American as of 2020[6]). As discussed above, the binning of ages ≥ 80 years means that the sample means and test statistic we are actually calculating deviate somewhat from the true result, but we are hopeful that this doesn't impact the validity of our result too much.

We think this is the most appropriate test to use in this situation. A non-parametric test would have the disadvantage of lower statistical power (i.e. being less likely to reject the null hypothesis if it is not true). And since there is no natural correspondence between individual Democrat and Republican voters, a paired test would not make sense.

Test, results and interpretation

```
ttestresult <- t.test(dems$V201507x, reps$V201507x,
                      paired = FALSE, alternative = "two.sided", var.equal = FALSE)
```

Calculated as described above, the mean ages of the two groups in our sample are 51.6 years for Democrats and 54.3 years for Republications.

The p-value of the test statistic is extremely small (9.3659568×10^{-9}, i.e. < 0.00000001). So we reject the null hypothesis and judge this age difference to be statistically significant.

This result is also practically significant: The estimated age difference (Republicans being 2.7 years older than Democrats on average) is politically meaningful information, considering the potential impact of both generational differences (i.e. a person's birth year is related to various formative cultural and political events they may have experienced) and individual life changes (e.g. reaching retirement age) on voting decisions. Cohen's D statistic for the effect size of the age difference observed is 0.159, which using conventional interpretation would rank as slightly less than a "Small" effect ($d = 0.20$)[7]. We note that the direction of the observed effect agrees with the conventional wisdom outlined in the introduction, that older voters tend to skew more conservative.

[5] See e.g. https://en.wikipedia.org/wiki/Tobit_model
[6] https://www.cbsnews.com/news/hester-ford-oldest-living-american-celebrates-birthday/
[7] https://en.wikipedia.org/wiki/Effect_size#Cohen's_d