Title: Lab 1: Question 1 | Author: "Jot, Leena, Jessica" |

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

1.1 Importance and Context

In Democracy each vote matters. In such a backdrop, political parties look for strategies to increase voter turnout. The parties rely on an understanding of their demographics to increase the voter turnout for their parties. Many different factors influence voter turnout. In the aggregate, voters tend to be older, wealthier, more educated than non-voters. This difference in turnout affects public policy.

The political parties need an understanding of their base, both in absolute terms and comparative terms. One of the important metrics is the age of the voters. There are two certainties in life: Death and taxes. The parties worry about the former as it affects their votes. The parties also rely on the age to roll out the policies that are important to their voter base. There is a general consensus that Democratic voters are younger as compared to Republican voters. Maybe, that is why Democrats love to talk about canceling student loans, while Republicans pushed to give tax breaks on the passive income sources.

Description of Data

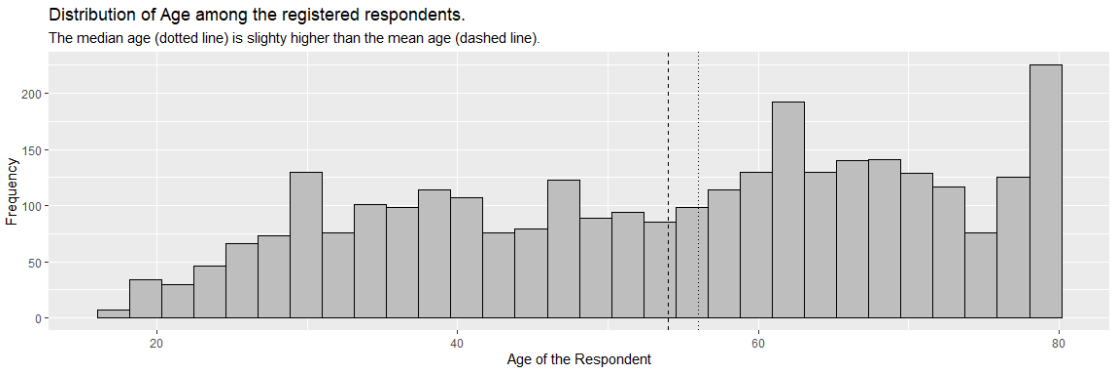
We will address this question using data from the 2020 American National Election Studies (ANES).

We want to know if Democratic voters are older or younger than Republican voters in 2020. This question has three important pieces. We first need to understand who a voter is in this context. Next, how would a voter classify as a Democrat or Republican? Finally how to calculate the age of the respondents.

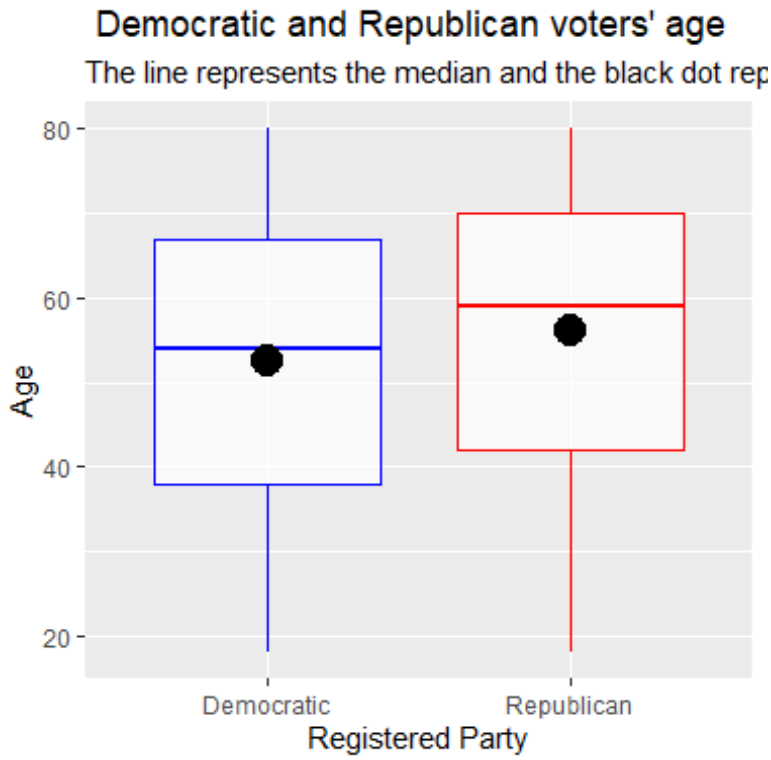
To classify a respondent as a voter, we look at their likelihood of voting in November. 5.25% of the respondents know that there are not likely at all to vote in the 2020 general election. In this study, We classified the voter as anyone in the remaining categories. The remaining of the respondents are either likely to vote, or they have already voted or don't currently know their preference yet, or finally refused to answer the question.

To classify a voter as a Democrat or Republican, we look at their party affiliation. Of the respondents, 48.42% have not mentioned their affiliation, 22.47% have registered as Democrats and 16.14% have registered as Republicans. For the purposes of this question, we focused on respondents who have registered as Democrats and Republicans.

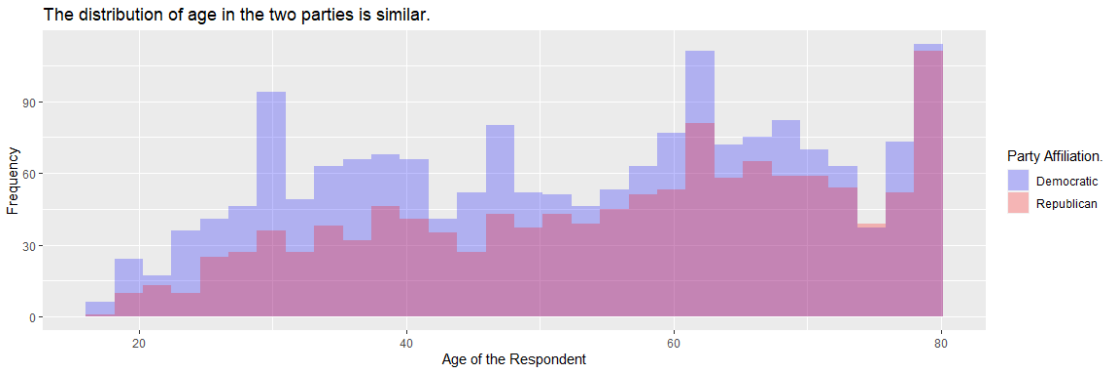
Many of the respondents have shared their age in the survey, however, 4.275% refused to share their age. We have removed those data points in the survey. The distribution of the age of the overall respondents is as follows:



We also take a look at how five-number summary of the age of respondents grouped by their party affiliations looks like. From the box plot, we can decipher that the mean and median for Democratic and Republican party voters is between 50 and 60. I also notice that the mean and median for the Democratic party are close to each other, while the median of the Republican party is greater than its mean.



Next, we look at the histogram of the two parties by age groups. We know that we have more data points for Democrats than Republicans, we concentrate on the shape of the histogram, rather than the frequency. Both the histograms are similar in shape.



To run the test, I have created two groups from the data based on the party affiliation, filtered for voters, and availability of the age data.

Most appropriate test

We then test whether the age of the two groups is the same or not. Because we would not be working with two parameters on the same sample, the unpaired test is appropriate. Since the data is interval or ratio, sample size greater than 30, and OK skew, we look at the variances of the samples. Since the variances are similar we would like to run two-sample t-test.

For a t-test, the following assumptions need to be true:

I.I.D - The ANES 2020 pilot uses a panel of individuals that use YouGov. This is an online system that rewards individuals for filling out surveys. There is a possibility that this introduces dependencies. For example, participants may tell friends or family members about YouGov, resulting in a cluster of individuals that give similar responses. Nevertheless, YouGov claims to have millions of users, which suggests that links between individuals should be rare.

Metric Scale - The assumption for a t-test is that the scale of measurement applied to the data collected follows a metric scale, such as the scores for an IQ test. The ages of the respondent is of metric scale.

Normally Distributed - The third assumption is the data when plotted, results in a normal distribution, bell-shaped distribution curve. The age distribution does not follow the normal distribution, especially since it has a spike at age 80. However since our sample size is large, the distribution of the means would approach normal distribution due to CLT.

Test, results, and interpretation

The null hypothesis of the test is that mean of the age of a Democratic voter is the same as the age of a Republican voter. If this test were to reject the null hypothesis, we would conclude that there is a difference in the age of the Democratic and Republican voters. If the test were to fail to reject the null hypothesis then we would conclude that either there is not enough data, there is no difference in age, or the test was inappropriate to conduct against data collected in this manner.

```
Q1test = t.test(subsetAnes$Age ~ subsetAnes$RegisteredParty)
```

We reject the null hypothesis that mean of the age of a Democratic voter is the same as the age of a Republican voter. The p-value for the test is 4.976e-08, which is well inside the rejection region. In many ways, this is similar to what we expected when we noticed that all the age and median were close to each other in the boxplot.

For a measure of practical significance, the simple difference in means, 56.02387 - 52.62192 = 3.40195 is probably the best measure. Republican voters are an average of 3.4 years older than Democratic voters. There is no need to standardize the units here because years are easy for us to understand. Though 3.4 years is a long time, as compared to a lifetime, it is not that much. The voters fall in the same demographic and would be difficult to differentiate based on age for targeting, marketing, and mobilizing purposes.

Test Limitations

We have conducted this test based on the data available in the ANES. The data is limited in the way the age is calculated. All the people above 80 are bucketed into one bin 80+ which makes it impossible to know the real mean/var and true distribution shape. We could also use only 50% of the respondents' data as party affiliation was missing or inapplicable.