

Lab 1: Question 1

Yao Chen, Jenny Conde, Satheesh Joseph, Paco Valdez, Yi Zhang

Importance and Context

The 2020 general election was very different from the ones that came before. It happened in the middle of a pandemic. It elected the first ever female VP. And it might have been the most polarized election in recent American history. What drove people apart, among other things, is their age.

Many people suspected that the Republicans have an older supporter base than the Democrats.

Is it just a myth or does it have some truth to it? That is what we're going to find out in this section, using the comprehensive 2020 Time Series Study from ANES (American National Election Studies).

Description of Data

From the ANES data set gathered before the election, there are 2 variables that are particularly relevant for us to answer this question, they are:

- V201018: PARTY OF REGISTRATION
- V201507x: SUMMARY: RESPONDENT AGE

We noticed for both variables, there are irrelevant answers in the data set. For PARTY OF REGISTRATION, we'll only keep Democrats and Republicans, and remove other parties as well as other non-answers, because we're only interested in the supporters of these two parties.

Similarly, for SUMMARY: RESPONDENT AGE, we will remove people who refused to answer.

After these cleanup operations, we are left with only 3074 observations to work with.

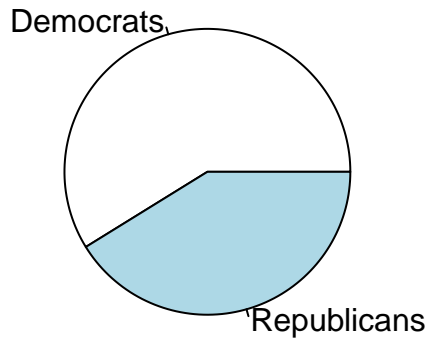
Looking at their summaries, it looks like the variables are now all in the correct range.

##	party	age
##	Min. :1.000	Min. :18.00
##	1st Qu.:1.000	1st Qu.:39.00
##	Median :1.000	Median :56.00
##	Mean :1.412	Mean :53.91
##	3rd Qu.:2.000	3rd Qu.:68.00
##	Max. :2.000	Max. :80.00

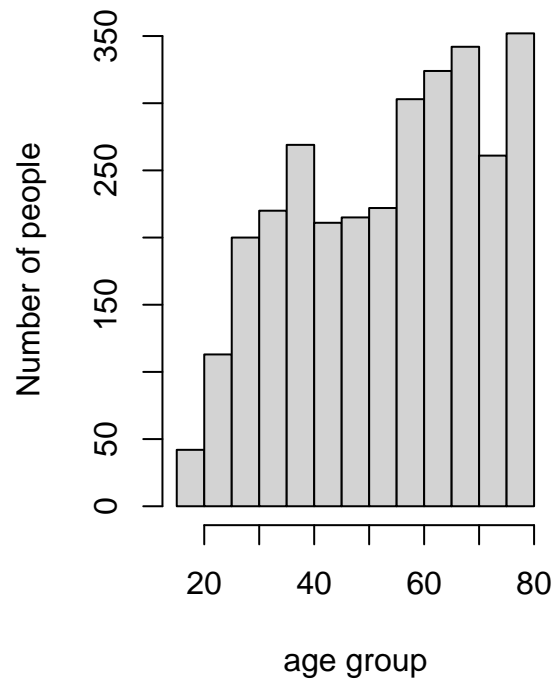
From the following graphs, we can see that the number of Democrats and Republicans are not too disparate and the Age distribution is not very skewed.

Notice the age number has a hard cutoff at 80 due to the way the survey is constructed, so everyone above age 80 simply gets grouped into "80 or older" group, so actually the mean age is somewhat under-representing the true average age of the participants.

Participants Party Affiliation



Participants Age distribution



Most appropriate test

The unpaired t-test seems to be the most appropriate to answer this question.

1. Even though there are only 3074 valid samples, it is still large enough for the CTE to apply on the sample average, so it satisfies the normality condition
2. Given the sampling frame based on a cross-section of registered addresses across 50 states and the District of Columbia, we feel the data are sufficiently close to be i.i.d.
3. And of course **age** is a metric scale variable.

Furthermore, because we're comparing two distinct groups of people with no natural pairing between them, this directs us to the unpaired t-test.

Test, results and interpretation

For the test itself, we establish the *null hypothesis* to be that the average age of Democrats and Republicans are the same.

And *the alternative hypothesis* is that they're not. Given there is no bias towards either side, this should be a two tailed test.

We'll be using the standard 5 significance level.

```
t.test(df_clean$age ~ df_clean$party)
```

```
##
## Welch Two Sample t-test
##
## data: df_clean$age by df_clean$party
## t = -5.3376, df = 2781.1, p-value = 1.017e-07
## alternative hypothesis: true difference in means is not equal to 0
```

```
## 95 percent confidence interval:
## -4.531263 -2.096511
## sample estimates:
## mean in group 1 mean in group 2
##      52.54867      55.86256
```

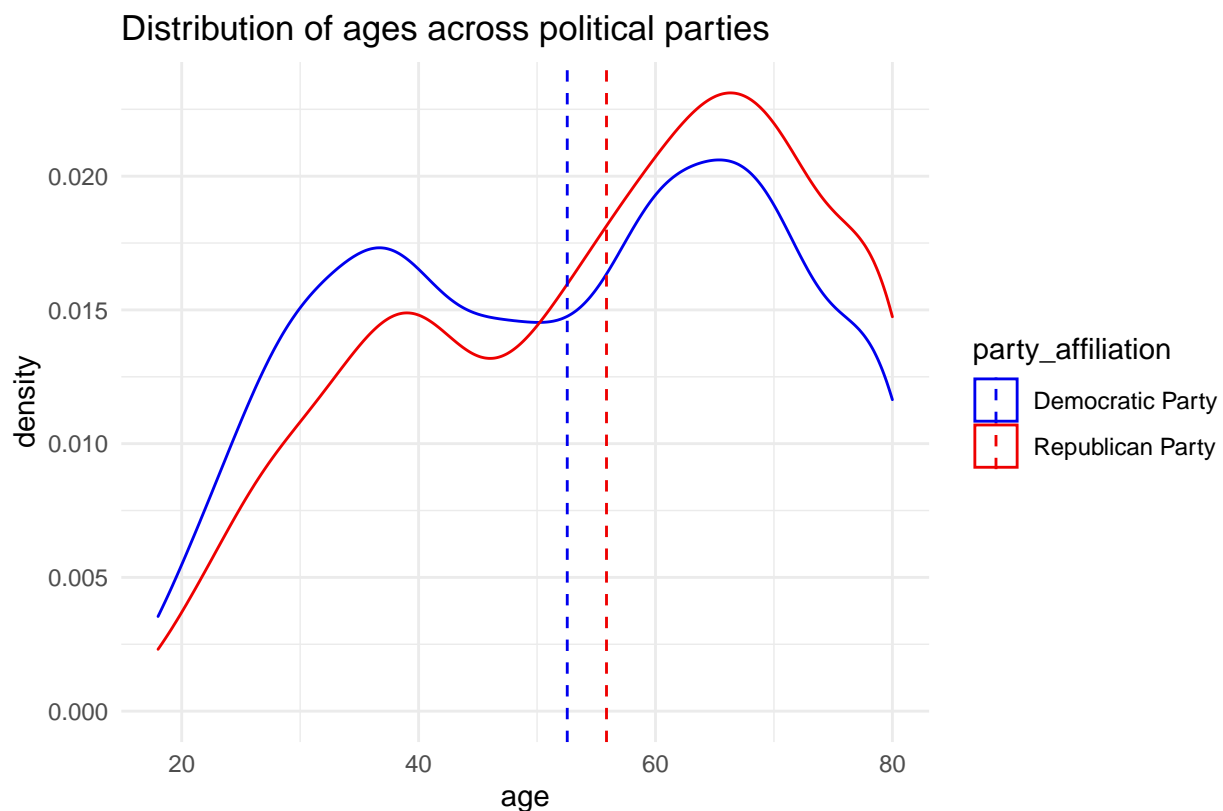
From the test, it looks like we have a very very small p-value, representing a highly significant result.

This gives us evidence to reject the null hypothesis in favor of the alternative and believe that the average age of the Democrats and the Republicans are indeed different, given the data and a 5% significance level.

Practically, as can be seen above, the average age of a Republican is more than 3 years older than a Democrat.

Plotting the distribution of ages within each group, it's clear that Democrats participants are more evenly distributed between “young” and “old” whereas Republicans are much more skewed towards people above 60.

At the same time, there are larger proportion of Democrats in the entire age group between 18 and 50 than Republicans and vice versa for over 50.



The dotted lines represent the average age in each respective political party.

More quantitatively, if we look at the correlation between the two groups, we see that there is a mild correlation, meaning that there is mildly strong (linear) relationship between the political party affiliation and age.

```
cor(df_clean$party, df_clean$age)
```

```
## [1] 0.09527008
```