Lab 1 Question 1: Are Democratic Voters Older or Younger Than Republican Voters in 2020?

GitHub Repo Link: https://github.com/yizhang7210/mids-w203-labs

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

```
theme_set(theme_minimal())
```

```
## # A tibble: 3,074 x 2
##      age party
##    <dbl> <chr>
##  1    46 Republican Party
##  2    41 Republican Party
##  3    37 Democratic Party
##  4    55 Democratic Party
##  5    31 Republican Party
##  6    80 Democratic Party
##  7    24 Democratic Party
##  8    72 Democratic Party
##  9    66 Republican Party
## 10    41 Democratic Party
## # ... with 3,064 more rows
```

Importance and Context

Like many events in the year 2020, the 2020 United States general election was unprecedented. Occurring during a global pandemic with political polarization at an all-time high and the most diverse candidate pool in U.S. history, the 2020 election posed new challenges and opportunities for American citizens and politicians. One key component that both major political parties utilized was appealing to voters and encouraging voter turnout. In order to appeal to the correct demographic base, it is helpful to understand who comprises each political party. One distinguishing factor could be age. This report uses comprehensive data from the American National Election Studies (ANES) 2020 Time Series Study to analyze the relative ages of voters registered as either Republican or Democratic using an unpaired two-sample t-test. Understanding age distributions could help politicians target their campaigns to appropriate demographics and reach audiences with whom their messages will resonate.

Description of Data

The ANES data set contains information from 8,280 pre-election interviews with U.S. citizens of voting age. Two variables are particularly relevant for us to operationalize the research question:

- `V201018: PARTY OF REGISTRATION`

- V201507x: `SUMMARY: RESPONDENT AGE`

Each variable takes on both relevant and irrelevant values for our study. For `PARTY OF REGISTRATION`, we are only interested in registered Democrats and Republicans, so we remove individuals who identify as independents, other parties, as well as other non-answers, i.e. "inapplicable," "don't know," and "refused" to respond. Similarly, for `SUMMARY: RESPONDENT AGE`, we remove people who refused to answer. After these cleanup operations, we are left with only 3,074 observations to work with. Looking at summaries for each variable, the variables are now all in the correct range.

```
##       age           party          
##  Min.   :18.00   Length:3074       
##  1st Qu.:39.00   Class :character  
##  Median :56.00   Mode  :character  
##  Mean   :53.91                     
##  3rd Qu.:68.00                     
##  Max.   :80.00                     
```

From Figure 1, we can see that the ANES data contains a larger number of Democrats than Republicans, and the age distribution is not very skewed. We also note that ANES cuts off the values for age at 80, so everyone above age 80 is grouped into the "80 or older" group. This means the average age shown in the ANES data will underestimate the true average age of the participants.

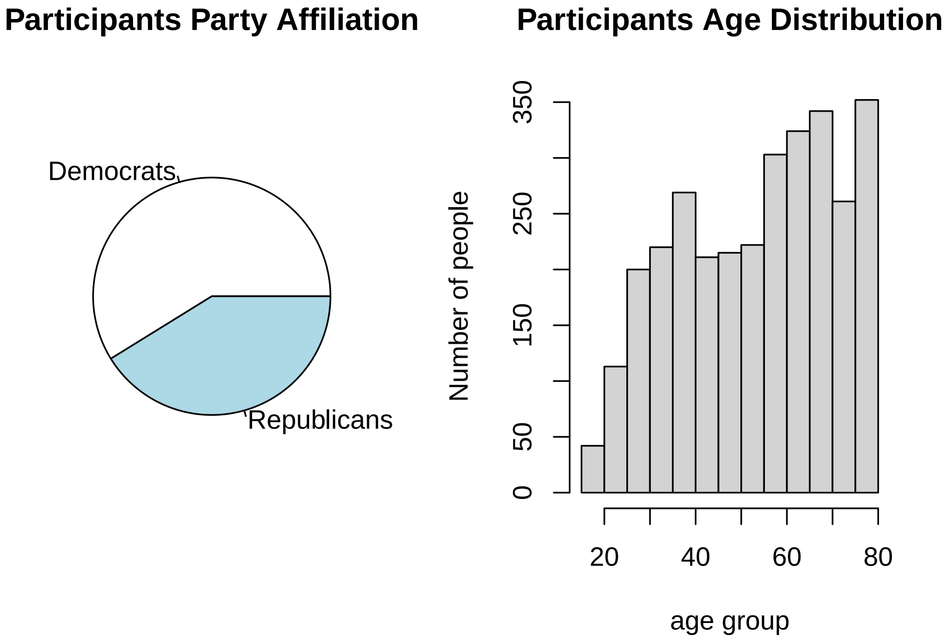

Figure 1: Distributions of party affiliation and age within the ANES data set

Most appropriate test

The unpaired two-sample t-test is most appropriate to answer this question. We are comparing a quantitative, interval variable for two distinct groups of people with no natural pairing between them. This directs us to an unpaired t-test, and we evaluate the validity of the assumptions under the two-sample unpaired t-test:

1. Although the age data does not quite follow a normal distribution as shown in the histogram in Figure 1, the data set is large enough with 3,074 valid observations for the Central Limit Theorem to apply. Therefore, this data satisfies the normality assumption of the unpaired t-test.

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. `SUMMARY: RESPONDENT AGE` is a metric scale variable.

Test, results and interpretation

For the test itself, we establish the *null hypothesis* to be that the average age of Democrats (μ_D) and average age of Republicans (μ_R) are the same. The *alternative hypothesis* is that they are not. Given we have no strong initial inclination in either direction, this should be a two tailed test.

We'll be using the standard 5% significance level.

$H_0 : \mu_D = \mu_R$ $\quad$ $H_a : \mu_D \neq \mu_R$ $\quad$ $\alpha = 0.05$

```
t.test(age ~ party, data = df)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by 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 Democratic Party mean in group Republican Party
##                       52.54867                       55.86256
```

From the test, we have a very small p-value that is much less than our significance level α, 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. In addition, a 95% confidence interval for the difference of mean between Democrats and Republicans $\mu_D - \mu_R$ is (-4.5312, -2.0965). Meaning we are 95% confident that the true difference lies in this range, implying that we cannot claim there is no difference in age between the two groups since the value 0 is not in the confidence interval.

Practically, as can be seen in Figure 2, the average age of a Republican is more than 3 years older than a Democrat. Plotting the distribution of ages within each group, we observe that Democratic 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. Knowing this information could help the Democratic and Republican parties target their political campaigns. For instance, because Republican registered voters are likely to be older than Democratic registered voters, the Republican Party could implement new strategies encouraging older citizens to vote in hopes of increasing their chance at winning the election. Conversely, the Democratic Party could leverage online campaign or social media to increase turnout rate for younger voters.

As a quantitative measure of practical significance, we analyze the correlation. Although political party is a nominal variable, the ANES data set records Democrats as the number 1 and Republicans as 2. Therefore,

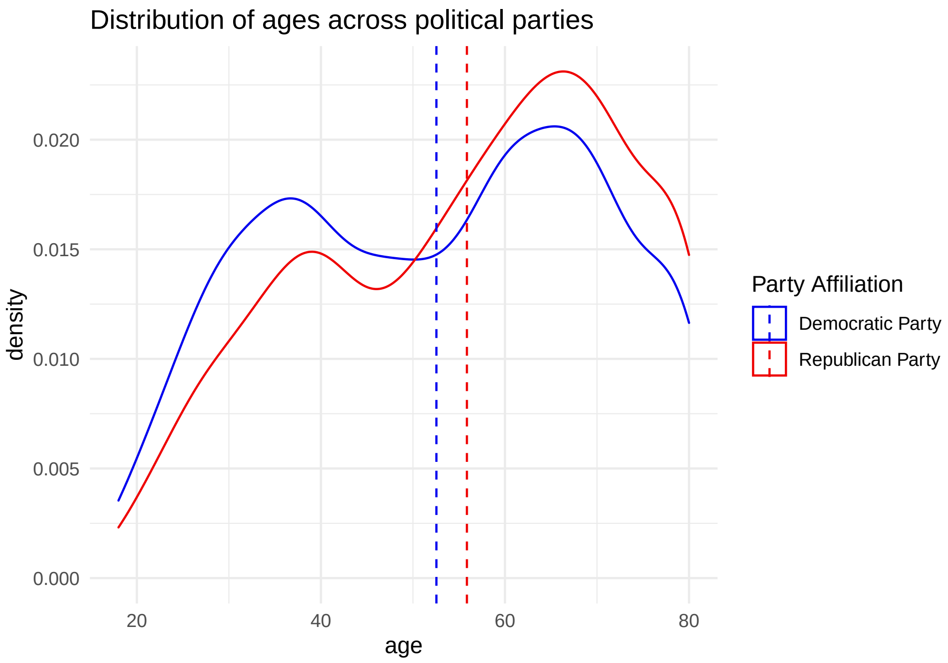

Figure 2: The distribution of age differs for each political party we analyze. The dotted lines represent the average age in each respective political party. Average age of Democrats surveyed is 52.5. Average age of Republicans surveyed is 55.9.

we can calculate the correlation since these qualitative values have been converted to numeric values. We find the correlation to equal 0.0953. This number shows a mild correlation between the two variables, indicating that higher values of age tend to correspond to "higher values" of political party, or the Republican party as ANES has recorded the data.