Lab 1: Question 1

Jun Qian, Lucas Schroyer, Ryan Mitchell, Oliver Chang

Contents

Are Democratic voters older or younger than Republican voters in 2020?

Importance and Context

Are Democratic voters older or younger than Republican voters in 2020?

The age is always one of the most important factors in forecasting and understanding American election outcomes. People in different generations are usually associated with different political ideologies. Moreover, the turnout rate for different age groups vary significantly. For example, in 2018 midterm elections, the turnout rate of people age 65 and above is 64%. In contrast, only 44% of people age 35-45 and 30% of people age 18-24 cast a ballot. As we know, since the American population is aging, older people will have even bigger impacts on the election.

It has been widely believed that older American electorates are more likely to vote for Republicans. However, is this stereotype still grounded in fact? We analyzed the 2020 pre-election data to confirm whether nowadays there is an age difference between Democratic voters and Republic voters.

Description of Data

We used the pre-election data from the 2020 American National Election Studies (ANES) site to address this question. The following variables are extracted for further processing and exploration this topic:

Variables to define "voters": registered_to_vote_status, voted_early_in_gen_election, plan_to_vote_in_gen_election, Variables to define "party": party_of_registration_str, party_of_registration, democrat_party_rating, republican_party_rating, Variables to define "age": age

In this study, we defined "voters" as who 1)registered to vote and either 2) voted early in general election or 3) plan to vote. This meant, to be eligible as voters for our analysis, the respondents should either voted already or already register to vote and also plan to vote for president. We then looked into the registered party of the voters. If there is registered party information, we assigned it as party affiliation for the respondent.

However, we noticed that around 48% of the voters do not have applicable registered party information. Since it consists a significant portion of the sample, we tried to infer their partisan leanings for these respondents based on if they have a strong preference for one party. To determine such preference, we used the difference in party rating, a "feeling thermometer" range from 0-100. If one party is rated 50 higher than the other party, we assigned a party to that respondent since it indicates a clear partisan leaning. To illustrate this party assignment methodology, Figure 1 shows the preference of Democratic party over Republican. The right tail with value greater than 50 represents all the respondents we assigned as Democratic voters. Respondents on the left tail with less than -50 were assigned as Republic voters

In this way, we assigned party information to 2105 out of 3929 respondents whose registered party information is inapplicable, bringing the total sample size to 7624.

Then, we examined the age of the voters. The age is by year expect for all the people above 80 who are bucketed into the "80+" bin. In addition, people refused to answer the age question were assigned as -9. Figure 2 shows there are more Democratic voters (purple) in all the age buckets below 50 and more Republican (orange) voters in age buckets above 50, including the "80+" bin, as well as "refuse to answer" bucket.

Since there is little information available to impute the age of the "refuse to answer" group. We excluded them from the samples for statistics tests calculation and hypothesis test. For the samples in "80+" bin, we assumed them as age of 80 in the analysis. But more details of the impacts will be discussed in the last section - "test limitation" section.

Table 1 shows the mean and median of Democratic voters v.s. Republican voters and the trimmed mean calculated without the "80+" bin. In both scenarios, the mean and median of the Democratic voters are smaller the the Republican voters.

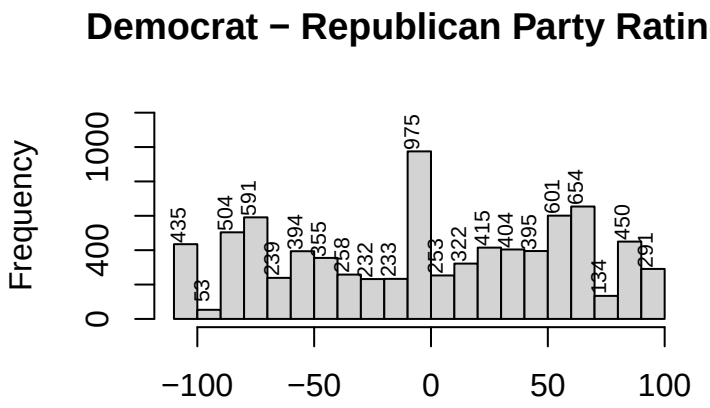
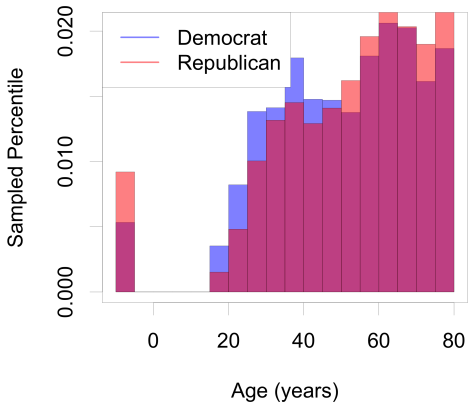


Figure 1: If a survey response's abs(Democratic Rating - Republican Rating) > 50 and they are listed as NA for their party affiliation, remap that respondent to either a Democrat/Republican from NA.

Intuitively, we expect the democratic voters cohort are younger than the republic voter group. However, is the difference statistically and practically significant?

```
## pdf
##   2
```



Political Party	Median Voter Age	Mean Voter Age
Democrat	55	53
Republican	58	55
Democrat_filtered	53	51
Republican_filtered	56	54

Figure 2: Age of Survey Respondents by Political Party and Table 1: Age by Party for the Unfiltered Dataset and for with the 80+ Age Bin Removed

Most appropriate test

To determine whether statistically there is a age difference between democratic voters and republic voters, we compared the mean of the age of two groups using the unpaired welch t two sample test. The reasons we selected this test are as follows:

(1) The two sample groups consists sample data points unrelated to each other.

(2) The data is interval except the 80+ bin. For the 80+ bin, as in the mean/median calculation, we just assumed it is 80 for testing. It is a simplified treatment of the data. However, the possible impacts will be discussed in the test limitation section.

(3) The data extracted from the population is I.I.D.. Clustering effect could skew the data during sampling

process. But the procedures how the survey was done are likely to keep the clustering effect should be minimal.

(4) The sample size is very large. The sample sizes of Democratic voters and Republic voters are 2701 and 2260 respectively. As a result, although the samples are not perfectly normally distributed, we decided the large sample size overweight the skewness.

(5) The sample variance actually is unknown due to the 80+ bin. Under such circumstance, the Welch t test is generally regarded as a more appropriate t test.

Therefore, we decided to use the Welch t two sample test.The Null Hypothesis for the test is the means of the age of two voters groups are equal. We used a two-tailed test in this study because we didn't want to assume one group is older or younger than the other group beforehand.

In addition to the statistical difference, we also would like to examine whether the difference is practically significant enough for us to act on. Therefore, we also calculated the effect size to reach a final conclusion.

Test, results and interpretation

According to the results, we can not simply assume the age of Democratic voters is different from the Republican voters.

From the first glance, the t-test results indicated that the Null Hypothesis was rejected at 95% confidence level. The reported p-value was 0 , almost zero. So statistically, we can conclude the mean age of two groups are different. With a t-value of -5.89 , we anticipated the mean of Democratic voters is smaller than Republican voters. This also seems to be consistent with our observations earlier the the Democratic voters are younger.

However, the effect size of the test is c(`Cohen's d` = -0.17). This meant the practical significance in the difference in the age is fairly small, almost negligible. The possible reason for the conflicting results might be our large sample size. As the sample size increases, the hypothesis test possesses increasing statistic power to detect a small difference in two sample groups. But the low effect size indicates the actual difference in practice is fairly minor.

With a relatively insignificant effect size,we can not simply conclude the whole Democratic voters group are younger or older than the Republican voters in reality. The situation might be much more complex.

Test limitation

In our analysis, we simply assume the age "80+" bin as 80 years old since there is no way for us to know the exact distribution of data in this bin. However, this bin consists 5.04% of the Democratic voters' group and 5.81% of Republic voters' group. Different age patterns in this bin can vary the result significantly. For example, if the distributions of two groups are heavily skewed toward different directions, obviously assuming it as 80 can distort the test results. If we can get the birth year of the respondents, it will largely reduce the uncertainty.