Lab 1: Question 1

RBG (Ruby Bajaj Gerrit) Team

Contents

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

1.1 Importance and Context

Much is known in the political science realm about sociodemographic factors such as race and household incomes on political affiliations. However, age has always been ignored despite the fact that age is one of the strongest predictors in terms of voter turnouts in presidential elections (Holland 2013). In both the 2008 and 2012 presidential elections, younger crowds overwhelmingly voted for then Democrat Barack Obama (Holland 2013). Does that mean that Democratic voters are younger than their Republican counterparts?

We will further explore this research question using the ANES 2020 Time Series Study which is a continuation of series of election studies conducted since 1948 to support analysis of public opinion and voting behavior in US presidential elections. The result could potentially help provide additional insights into the complexity of what drives voter turnout and choice within the American electorate in which a special emphasis will be placed on the role of age in spurring voter choice.

1.2 Description of Data

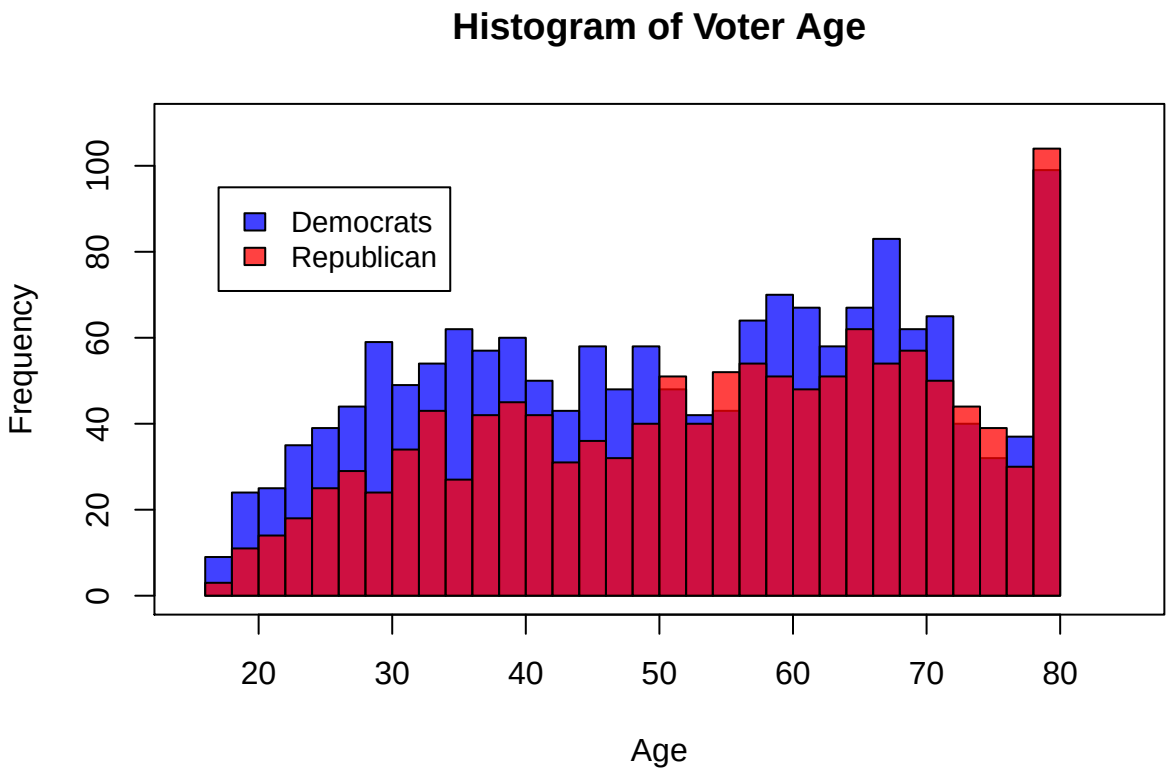

Figure 1: Voter Age by Party

We have two main variables that we need to operationalize in order to test the research question. We used ANES variable V201507x to determine voter age. There were 354 instances of respondents who refused to disclose their ages and were omitted from the dataset. The age variable is bucketized from 18 to 79 as actual age values and any age 80 or older responses is grouped into the 80 bucket. This accounted for 184 instances and there is a risk of skewing the end result due to data obfuscation.

Table 1: Democrat and Republican Distribution

Party	Frequency
Democrats	1651
Republicans	1283

Subsequently, we made use of ANES variables V201018 (`party of registration`), V201228 (`party id`) and V201019 (`intent to register to vote`) to represent our `voter by party` (Democrat or Republican) variable. We made an important distinction between `party of registration` and `party id` as the responses do not match in some cases. As a registered voter is not required to affiliate themselves with a party prior to voting in a party's primary election, we made the assumption that the `inapplicable` category within `party of registration` variable is still a valid category to use provided that respondents who answered `inapplicable` also answered `yes` in their `intent to register to vote` question. All other categories were considered to be erroneous and unknown data and thus will be omitted. We then filtered and cross-validated the data subset with variable `party id` to aggregate Democratic and Republican voters.

As reported in Table 1, we determined 1651 Democratic and 1283 Republican voters based on our methodology of defining `voter by party` variable.

In Figure 1, distributions of age for both Democratic and Republican voters share similar traits with a sharp spike on the right. This is probably attributed to the age cap of 80 for all 80 or older voters as mentioned previously.

1.3 Most appropriate test

We will then need to test and compare the age across this subgroup. As these variables are metric and continuous, a parametric test is appropriate. In addition, the data is unpaired since we are comparing a group of voters by age and party choice. Based on these initial assumptions, we will use an unpaired t-test, implemented in R using `t.test`.

The unpaired t-test requires the following validations to be true:

- **Approximately normal:**

Based on Figure 1, although the distribution isn't clearly normally distributed, it really isn't that poorly distributed. There isn't a clear central tendency but the dispersion is reasonably contained. We have 2934 and this satisfies the Central Limit Theorem (CLT) requirement of more than 30 observations to achieve a normal sampling distribution as shown in the Appendix section, Figure 2. However, as noted in the previous section, there appears to be a sharp skew on the right of the histogram. Once again, this is due to the age 80 or older data points being truncated at 80 resulting in the skew. Nevertheless, this skew is negated by 2934 data points.

- **Metric scale:**

Both age and number of voters by party are on a metric scale in which the frequency or count represent a meaningful measurement.

- **I.i.d data:**

There is some hesitation in saying that the dataset is completely i.i.d. as respondents who complete the 2020 survey online are rewarded and thus incentivized to complete the survey. There is a possibility of introducing dependencies. For example, respondents may refer friends or family members who receive similar invitations to complete the survey resulting in a cluster of individuals giving similar responses. Nevertheless, the data is assumed to collected from a representative, randomly selected portion of the population as this sample consists of the full set of sample members from the ANES 2016 Time Series Study who have completed the post-election interview. Thus, it is i.i.d. enough to justify using the unpaired t-test.

In addition, we chose to select a two-tailed t-test to cover all our bases and eliminate any bias in our test. In short, even though not all requirements are met, they are close enough given the large sample size to validate using the unpaired two-tailed t-test.

1.4 Test, results and interpretation

```
t.test(age ~ party, data = q1, paired = FALSE, alternative = "two.sided")
```

```
##
##  Welch Two Sample t-test
##
## data:  age by party
## t = -4.6259, df = 2794.5, p-value = 3.9e-06
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.171166 -1.687700
## sample estimates:
##   mean in group Democrats mean in group Republicans
##                  51.81647                  54.74591
```

The unpaired t-test for difference in mean between the age of Democratic and Republican voters yields a p-value of 0.0000039 which is lower than $\alpha = 0.05$. Since the p-value computed is so much smaller relative to the significance level, α, this suggests a strong evidence in favor of the alternative hypothesis in which there is a statistical significance of observed differences between the mean ages.

This leads us to reject the null hypothesis that the true difference in means is equal to 0. The practical significance of the result is that we have sufficient evidence to say that the age difference between Democratic voters and Republican voters is not 0 and the hypothesized effect exists. Among the Democratic group, the mean age is approximately three years younger on average compared to its counterpart. This difference of about three years might typically be considered a small effect but this effect is large enough to produce a statistically significant result due to a polarized electorate. Interestingly enough, the means for both group lie in the middle age range and this leads to the question on why the younger age range group (18-30) is less likely to vote and further exploration will be needed to provide guidance on how to increase voter turnout amongst this group.

1.5 Test limitations

Even though the unpaired t-test produced a statistically significant result, a different outcome altogether could be obtained due to the below test limitations. First, respondents 80 or above are bucketed into one bin 80+ which limits the test and creates inaccuracy in terms of not knowing the true mean or variance and distribution shape. Second, we could not pinpoint 'voters' perfectly. Hence, the sample may be misrepresented if our define group opts to not vote or if our exclusion technique removed those that have the intent to vote. Third, respondents are incentivized to complete the survey and thus introducing the possibility of dependencies. Fourth, the political affiliations for all observations may not be accurate as some respondents chose to not disclose their answer or others experienced technical difficulties. Nonetheless, these realizations should not discredit our test utilization and conclusion. However, these test limitations must be kept in mind when interpreting the test result.

2 Appendix

2.1 References

American National Election Studies (2021). 2020 Time Series Study (February 11, 2021 Version) [.dta]. Retrieved from https://electionstudies.org/data-center/2020-time-series-study/

Holland, Jenny Lynn (2013). Age Gap? The Influence of Age on Voting Behavior and Political Preferences in the American Electorate. Retrieved February 28, 2021 from http://hdl.handle.net/2376/4982

2.2 Plot

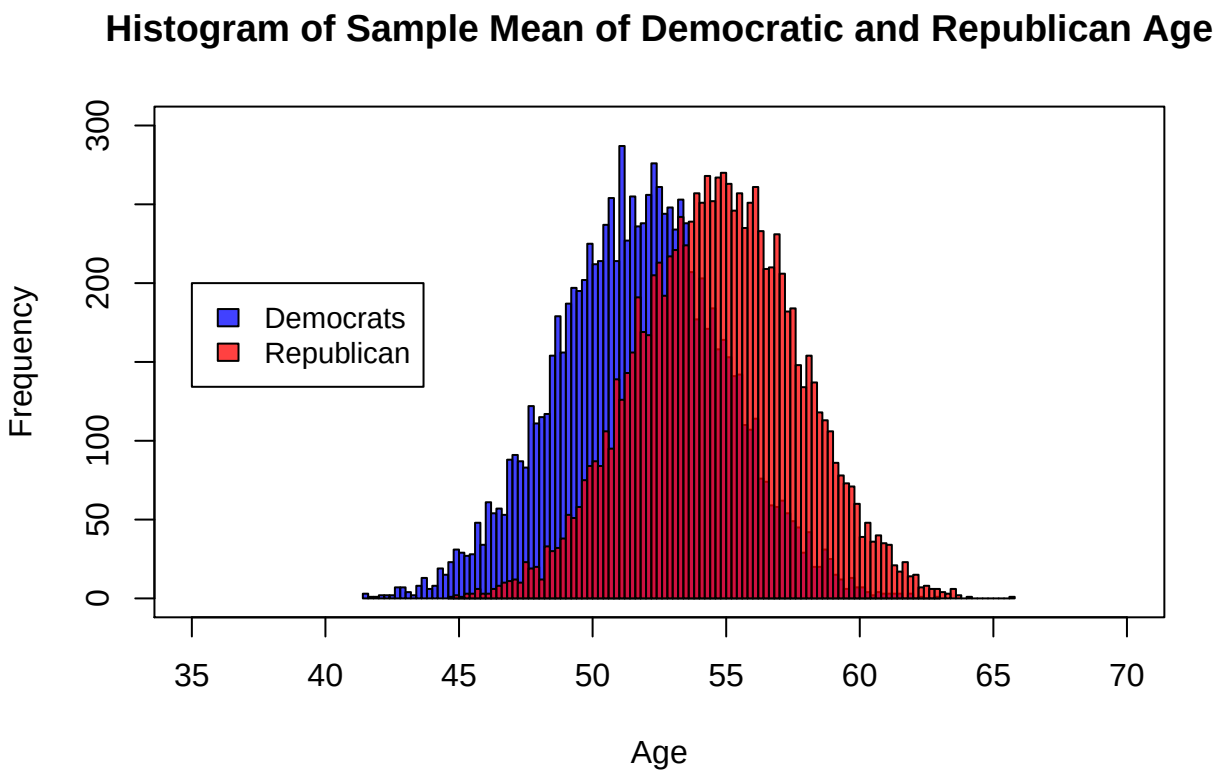

Figure 2: Sample Mean