Lab 1: Question 1

Devashish Kulkarni, John-Michael Stilb, March Saper

Contents

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

1.1 Importance and Context

The 2020 US presidential election has been a uniquely historic election. Falling squarely in the midst of a once in a century pandemic, a highly polarized electorate voted to choose a President between two candidates, both representing incredibly different paths for the future of the country. A wide range of strategies were used by the two major political parties in order to turn out their respective voters to vote. Among other demographic traits, voter's age is one of the key traits that drives political messaging and policy decisions for a party. Questions about the age of Democratic and Republican voters can help understand the effectiveness of this messaging in turning out the vote. This would also help the parties to tailor their political agenda towards specific age groups in order to increase support. In addition, since young people are the future of politics in the country, understanding whether voters of one party are younger or older than voters of the other party would be crucial in mapping out the future direction of the party itself.

1.2 Description of Data

We will address this question using the data from the 2020 American national Election Studies (ANES). This data set is collected as part of a series of election studies to support analysis of public opinion and voting behavior in US presidential elections. A total of 8280 participant responses are included in this data set on a wide variety of survey questions. For the purpose of answering this question, we will consider a respondent a 'voter' if they are registered to vote. It can not be assumed that a respondent not registered to vote did end up voting in this election. Also, a respondent who is registered to vote has shown an intention for voting. After subsetting, 7551 respondents remain in the data set.

To group the remaining respondents into Democratic or Republican voters, we will use the respondents' answer to a question asking to self identify as a Democrat, Republican, Independent or Other. We choose this definition, instead of say, party registration or prior voting record because of the unique nature of this election. Based on the political scenario in the US, the 2020 election was largely a referendum for the incumbent president. Hence, it is likely that voters registered for one party or having previous voting records for one party might vote for the candidate of the other party. Voters might also register as independents because of the rapidly changing politics of their party.

Finally, the age of respondents is gathered as an answer to a demographic question on the survey. We exclude 354 responses from the data set without a valid age to be able to perform a hypothesis test. In addition, the age of respondents older than 80 is set at 80 as part of the data gathering process.

Subsetting the data using the above criteria leaves us with two groups of voters, namely, 2727 Democratic voters and 2370 Republican voters.

To explore the distribution of age across the two groups, we look at a comparative histogram (Figure 1). We have a distribution starting just below 20 years to 80 years of age, with the age of voters older than 80 years of age set to 80. The minimum age is constrained by the legal voting age. We see that the proportion of young voters is lower than the proportion of older voters. In particular, it seems like there is a gradual increase in the number of voters through the ages of 20 to 40. There is an approximately uniform distribution of counts from ages approximately 40 to 60. There is a peak around the age of 70 and then a sharp decline towards the age of 80, ending with the spike. The comparative distribution of age with party has some interesting characteristics. There is a clear Democratic leaning for voters younger than approximately 40 years of age. The split across party evens out as we look at older voters. The spike at 80 years old is Republican leaning.

1.3 Most appropriate test

The hypothesis test used to answer this question requires a comparison of means between two unpaired groups of metric data. The appropriate test in this case is Welch's two sample t-test. The assumptions required for the validity of this test are:

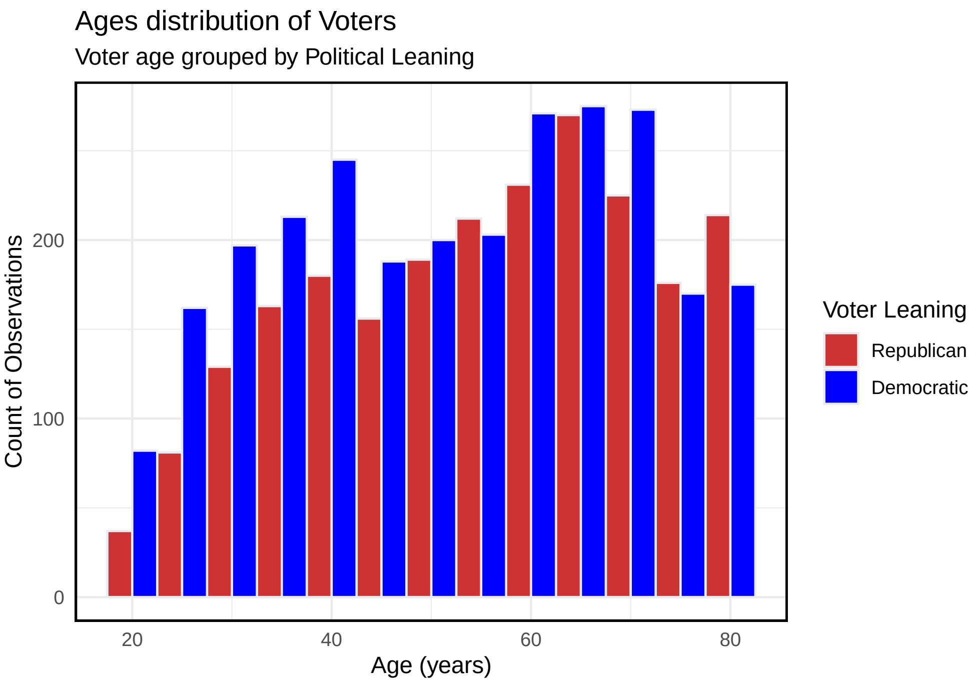

Figure 1: Histogram for Voter Age

1. Metric variable - Age, as measured in this data set is sufficiently metric. It is measured as integers and ranges from 20 to 80. The value of 80 is set for voters aged over 80. This trait of the data set is expected to affect the results of this test, however, since the proportion of observations with age >80 is not a large majority and the fact that observations in both groups are treated identically, we do not expect this to invalidate the results of the test.
2. IID Sample - The respondents for this survey are randomly selected using information from the USPS. Any person in the US with a registered address with USPS had an equal probability of being included in the survey, hence the assumption of IID for this sample is sufficiently met.
3. Sufficient normality of the data - There is little skew in the data as seen from Figure 1. In addition, the number of samples for each group is sufficiently large for the applicability of the Central Limit Theorem. The normality of the data is sufficient to establish validity of the test.

A two tailed test is appropriate for this case because we have no prior intuition or reasons to believe if one group has a lower or higher mean age than the other. We would reject the null hypothesis if there is a statistically significant difference between the mean ages of the two groups in either direction.

We use the following null and alternative hypothesis for this test.

Null hypothesis: There is NO difference between the mean age of Democratic voters and Republican voters.

Alternative hypothesis: There IS a difference between the mean age of Democratic and Republican voters.

The conventional significance level of 0.95 is appropriate for this test as there is no reason to use a more strict test. We will use a rejection criteria of p-value < 0.05 for this test.

1.4 Test, results and interpretation

```
t.test(V201507x ~ is_democrat, data = data_registered_voters)


cohen.d(V201507x ~ is_democrat, data = data_registered_voters)
```

The test produces a very small p-value (<1 e-10), indicating that the test is highly statistically significant. We reject the Null hypothesis that there is no difference between the mean ages of Democratic and Republican voters. Indeed, the observation made from looking at the distribution of the data in the above histogram hinted towards the fact that Democratic voters are younger than Republican voters. The sample mean for Democratic voters was 52.13 years while the sample mean for Republican voters was 55.30 years. The difference in sample means of the two groups is ~3 years and a calculation of the effect size using Cohen's D (0.188) indicates a small effect. This difference in means may not seem like large compared to the range of ages in the data set, however, it may be practically significant, especially looking distribution of ages across the two groups that shows a Democratic leaning for voters less than 40 years old as seen in Figure 1.

The conclusions of this report could provide important insights for the two major political parties in the US. The Republican party may choose to look into why younger people favor the Democratic party. On the other hand, the Democratic party may reassess their messaging to focus more towards older Americans. This insight could be used to affect the future of both the political parties, and in turn the political landscape in the US.