Lab 1: Question 1

Peter Morgan, Bruce Lam, Mia Yin

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?

In American politics, understanding voter demographics is very important. A good understanding of voter demographics allows political parties and individual candidates to portray a message that can better connect with its intended audience.

Our research question scratches the surface of achieving this goal. By understanding basic demographics like age, we can then conduct follow up research on what messages are more effective on certain age groups. Research like this can lead to more informed decisions for campaigns and individual candidates.

Description of Data

The dataset used to address this question will be from the 2020 American National Election Studies (ANES). This dataset is a random sample of around 8,000 participants meant to represent the population of 231 million non-institutional U.S. citizens aged 18 or older living in the 50 US states or the District of Columbia.

The first column of interest in this dataset is our grouping variable, Party of Registration. This is a simple and elegant way to filter between Republicans and Democrats which are the two groups of interest in our study. This data was collected by asking a participant “What political party are you registered with if any?” The answers available to the participant are refused, don’t know, inapplicable, democratic party, republican party, none or independent, or other. Table 1 below shows how this variable is distributed in this sample of data.

Table 1: Party Affiliation Distribution, Unfiltered

Refused	Don’t know	Inapplicable	Democrat	Republican	None/Independent	Other
9	2	4010	1861	1336	1029	33

As you can see, a lot of this data is irrelevant to our question as we are only interested in comparing the Republican and Democrat groups. Because of this, we will filter out all data where the Party Affiliation is not democrat or republican. The next column of interest for our research question is the age of the registered

voters. This is for the most part straightforward metric data. The exceptions are that there are some people that refused to specify their age, and all people aged 80 or older are reported as age 80. The minimum age value is 18 because that is how old you are when you are eligible to vote in the United States. The distribution of age data is shown below in table 2.

Table 2: Age Distribution, Unfiltered

Refused to Answer	18-19	20-29	30-39	40-49	50-59	60-69	70-79	80+
354	83	918	1378	1208	1329	1561	1048	401

To handle these discrepancies, we will filter out any data where the subject refused to give their age. We will still include all data labeled age 80 because there is no way to determine who is exactly 80 years old and who is older, and we do not want to throw away the data for older voters. Figures 1 and 2 below show density and frequency histograms for our filtered data.

Fig. 1: Density Histogram of Age

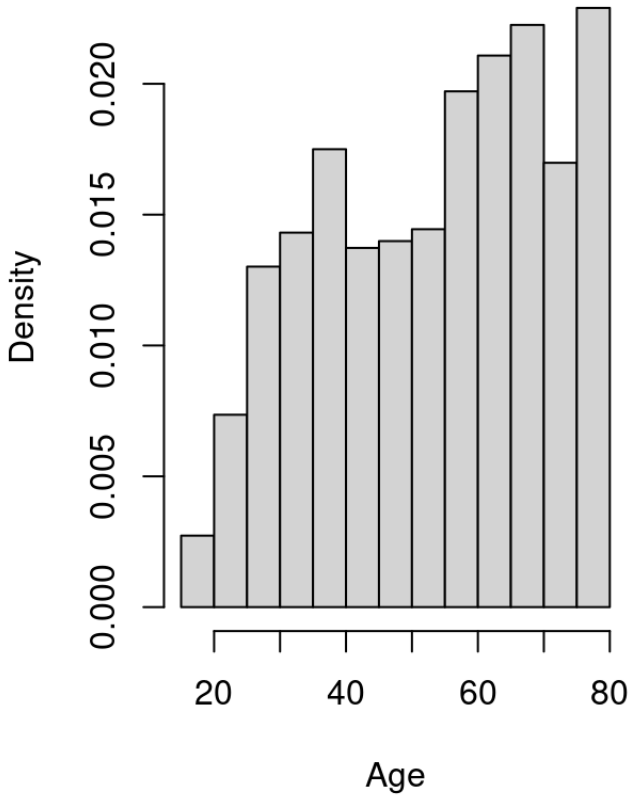
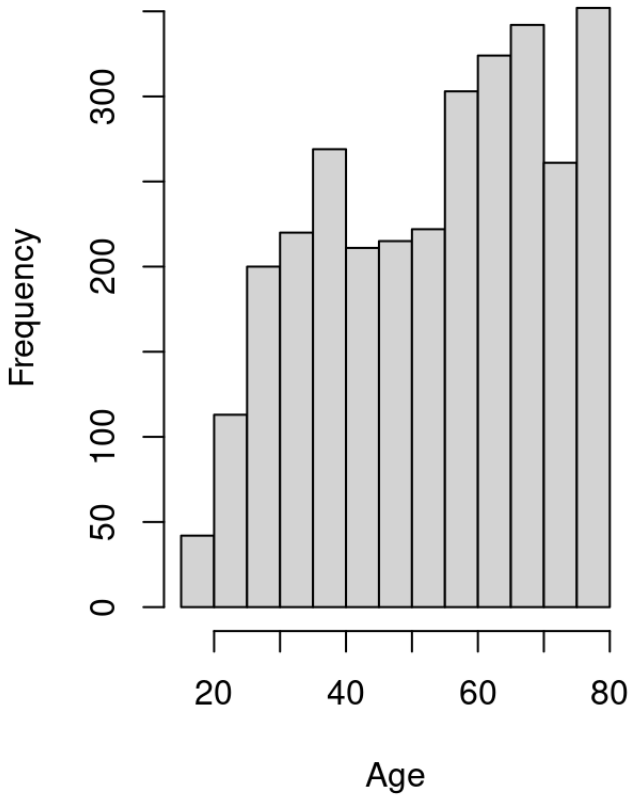


Fig. 2: Counts Histogram of Age



Figures 1 and 2 above provide insight on the overall subject age. You can see that participants in this study tend to be relatively uniformly distributed until you get to ages 30 and under. At the younger ages, there is a lower frequency of participants in this study. To gain insight relative to our research question, we will filter the ages based on political party affiliation. This distribution is shown in figure 3 below.

Figure 3: Boxplot of Age and Party Affiliation

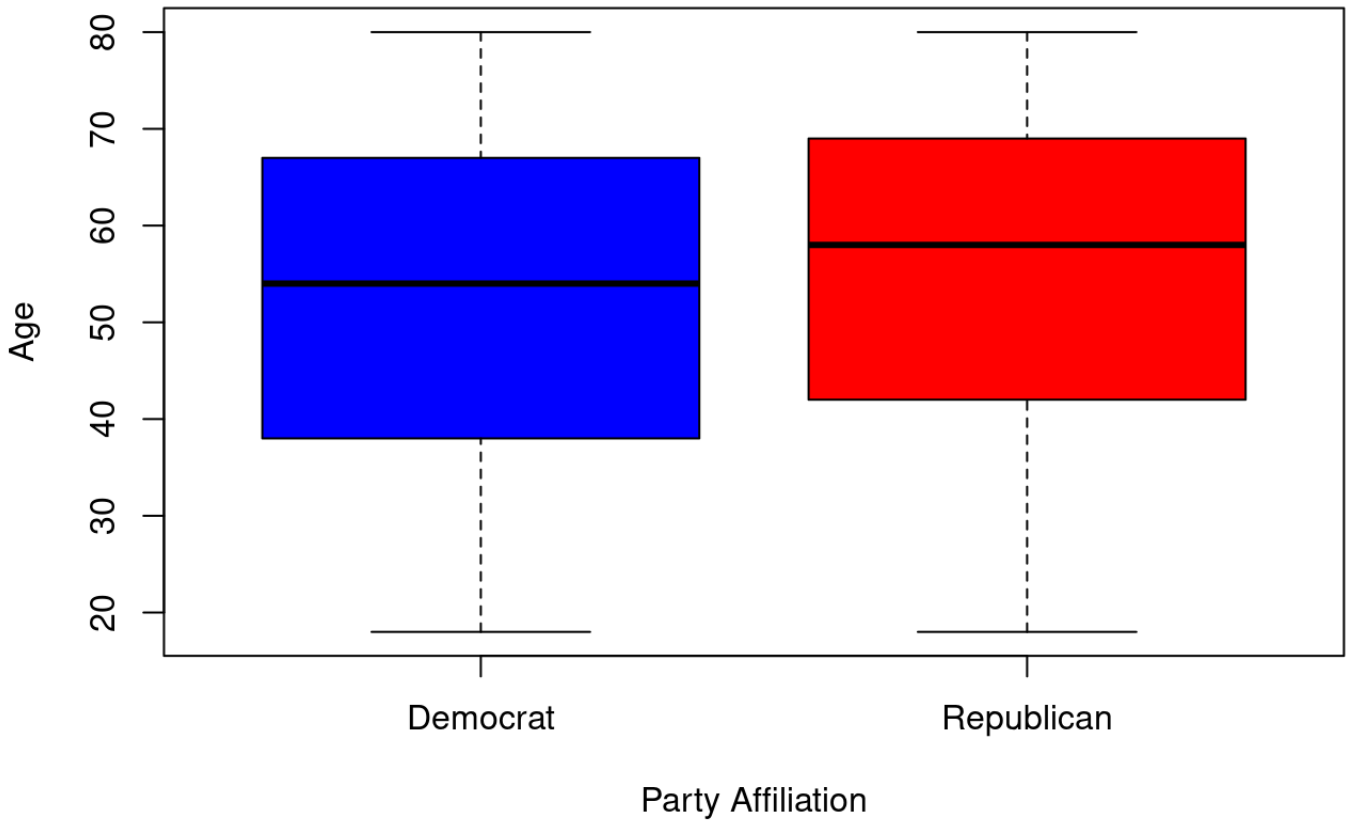


Figure 3 shows that the sample median age of democrats is slightly smaller than the sample median age of republicans. The spread of the distributions looks relatively similar between the two parties, with the maximum and minimum values being the same. This makes sense because our data has a lower bound of 18 and an upper bound of 80. Now that we have explored and cleaned our data, it is time to determine what is the most appropriate test for our research question.

Most Appropriate Test

The first step in selecting a test is determining if we will use a paired or unpaired test. For our data and research question, we will use an unpaired test because there is only one age measurement per subject. The next step in determining the best test is to decide if we have metric or ordinal data. In this case, we have metric data because we are doing our test on the age variable. The next decision is evaluating normality. Figures 1 and 2 show that this distribution looks somewhat normal for the left tail, but as the ages get higher,

the distribution begins to look uniform. Because of this, we are not able to assume normality. Since we cannot assume normality, a t-test can only be used if we can invoke the central limit theorem. In this case, the central limit theorem can be used due to the extremely high sample size and the skew not being too bad. We will proceed with a two-sample t-test.

When conducting a two-sample t-test, there are three assumptions that must be met. The first is that we have a metric scale. This assumption is met because age is a metric variable. The second assumption for a two-sample t-test is IID data. This assumption is valid according to the sample design section of the ANES user guide. The final assumption is that there are no major deviations from normality given the sample size. While our distribution is certainly not normal, we are able to invoke the central limit theorem to make this assumption valid. Now it is time to state our null hypothesis.

Even though figure 3 shows the median age of democrats is lower than the median age of republicans, we will still run a two-tailed t-test at 95% confidence to be slightly more conservative in our analysis. The null hypothesis is the mean age of democratic voters (μ_D) equals the mean age of republican voters (μ_R). These hypotheses are stated mathematically below.

$$H_0 : \mu_D = \mu_R$$

$$H_a : \mu_D \neq \mu_R$$

Test Results and Interpretation

To evaluate our null hypothesis, we will perform our two-sample two-tailed t-test with a 95% confidence level. Performing this test yields the following results.

$$\mu_D = 52.55, \mu_R = 55.86$$

$$p = 1.0171\text{e}{-7}$$

$$95\%\ T\ statistic\ confidence\ interval : (-4.53, -2.10)$$

Based on the results of our t-test, we will reject the null hypothesis. This is because the p-value is significantly less than 0.05. Another way to interpret the results is using the confidence interval. It says that we have 95% confidence that the t-statistic lies between -4.53 and -2.10. This tells us to reject the null hypothesis because 0 is not in this confidence interval. Now that we have determined that there is a statistically significant difference in the mean age of republicans and democrats, we can move on to practical significance. Since we are comparing the difference between two means, we will use Cohen's d to evaluate practical significance.

Performing the Cohen's d calculation gives us a d value of 0.19 meaning a practical significance of little or small effect size. This is reflected in the fact that the raw difference of means is only 3.31 years. This test is statistically significant because the large sample size makes the difference more statistically meaningful, but the difference is still relatively small which means a low practical significance.

In this research we have determined that democratic voters mean age is not the same as republican voters mean age at a 95% confidence level. However, the true difference is not very large which means a low practical significance and results that do not help us too much. If anything, we did learn that to understand political party demographics in America, we must look beyond age.