Lab 1: Question 1

Patrick Old, Xander Snyder, Tal Segal

Importance and Context

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

An old saying claims that if you're young and not a liberal, you have no heart, but if you are old and not a conservative, you have no brain. That is - there is a long held belief that the younger people lean left whereas older people lean right. Since older people tend to turnout to vote at much higher rates, for reasons such as habit formation and time availability[1], understanding the extent of that tendency could greatly affect election results.

The differences in party age could affect the candidate selection as it is possible that individuals will prefer candidates closer to them in age. According to a study by Washington State[2], the mainstream-insurgent hypothesis states that younger people prefer political candidates that are viewed as ideologically extreme and iconoclastic to the traditional party establishment, while older people prefer more moderate and traditional mainstream candidates. A better understanding of whether Democratic voters are truly older or younger than Republican voters could influence decisions to improve the appeal of political campaigns to their party members.

Description of Data

The data that we will be using to answer this question is drawn from the 2020 American National Election Studies ANES[3]. According to the codebook, the target population for the study "was the 231 million non-institutional U.S. citizens age 18 or older living in the 50 US states or the District of Columbia". The samples were randomly drawn, with all included residential addresses having equal probability of selection.

We will be using this dataset to extract the subset of respondents who are A) Willing to reveal their age B) Voters C) Either democratic or republican. After all filtering, our subset contained 61.5% of the total dataset.

A) Ages were extracted using the variable asking for the respondent's age. If a respondent refused to give their age, they were excluded from our subset. Approximately 95.7% of responders were willing to give their age.

B) We defined "voter" as the subset of respondents that either claimed to be registered voters or indicated that they intended to register to vote in 2020. Specifically, we filtered on three variables to arrive at our subset of voters, which accounted for approximately 91.3% of respondents.

The first variable asked respondents if they intend to register to vote. If they replied that they did intend to, we included them in our sample. The second variable asked respondents the address at which they are registered to vote. As an address is required to register, if they replied that they were registered to vote at either their mailed survey address or at a different address we included them in our subset. Since the survey had mixed method sampling, some respondents did not receive the survey through the mail. Thus, the third variable asked respondents who were answering the survey online if they are registered to vote. If they replied that they are registered to vote without an address we included them in our subset.

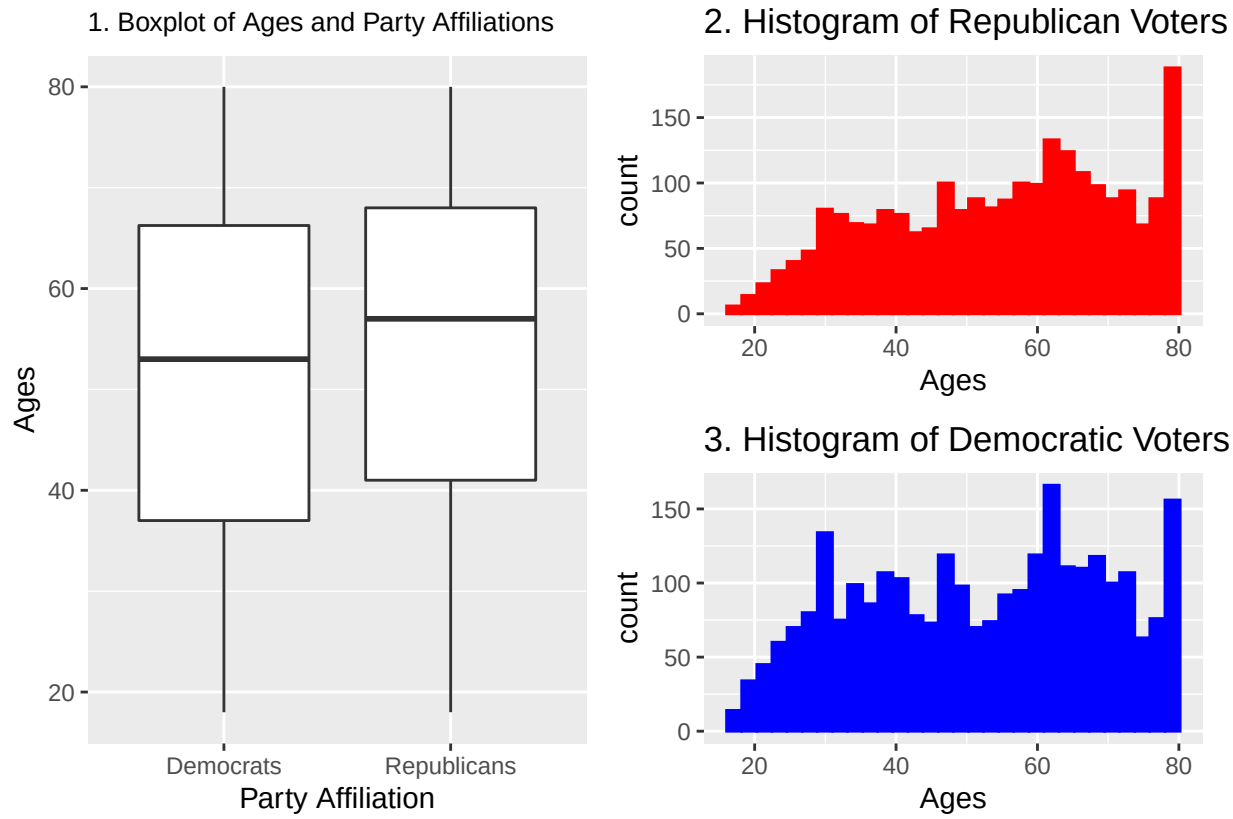
[1] American Sociological Review Vol. 33, No. 4 (Aug., 1968), pp. 563-575 (13 pages)

[2] Symonds, Alexandria. "Why Don't Young People Vote, and What Can Be Done About It?" The New York Times, The New York Times, 8 Oct. 2020, www.nytimes.com/2020/10/08/upshot/youth-voting-2020-election.html.

[3] American National Election Studies. 2021. ANES 2020 Time Series Study Preliminary Release: Pre-Election Data [dataset and documentation]. February 11, 2021 version. (www.electionstudies.org)

C) Party affiliation was filtered using the variable which asked if respondents consider themselves as a Democrat, a Republican, or of another party. 67.4% of respondents self-identified as either Democratic or Republican, with 36.2% identifying themselves as Democrat and 31.3% as Republican.

The mean age for a Democratic voter is 52 and the mean age for a Republican voter is 55. Distributions of age for both Republican and Democratic voters had a left skew (figures 2,3 below), indicating that voters of both parties tend to be on the older side. These plots seem to support the notion that older folks are more likely to vote than younger ones. Comparing the box plots of both parties (figure 1 below), Democrats appear to have a younger median value compared to the Republican party. The 25% and 75% percentiles of the Democrats are wider apart, but both are lower in value than the 25% and 75% percentiles of the Republican party, indicating a potential difference in ages.



Figures 1,2,3: Ages of Democratic and Republican Voters through a Box Plot and Histograms

Limitations to our dataset:

1) By definition, a preliminary release is an early version of a dataset, and is subject to change in the full release.

2) Although the dataset came with weights, which are meant to more accurately represent population parameters, those weights were not applied due to time constraints.

3) All people 80 and over were clustered into the same age category in the survey ("Age 80 or older"), which means we didn't have the exact age of older voters. This contributed to the left skew mentioned above. Approximately 6.0% of our subset fell into this category.

4) Respondents who were unable to answer questions regarding their registration status, age or party affiliation were filtered out.

Most appropriate test

We used the two-sided Welch's t-test at a 95% confidence level to evaluate whether the mean age of Democrat and Republican voters is different. While our data could qualify for testing under Wilcoxon-Ranked Sum Test, which does not have as many assumption requirements, we chose the t-test as a more robust option over a nonparametric test.

Our null hypothesis H_0 is that mean Republican age is identical to mean Democrat age. $H_0 : \mu_D = \mu_R$ Our alternative hypothesis H_a is that there is a difference between the means $H_A : \mu_D \neq \mu_R$

Assumptions of the Welch's t-test:

1) Normally distributed data. While our data is not normally distributed, the sample size of 5094 is large enough for the Central Limit Theorem to kick in.
2) Sample is i.i.d. The response rate overall was 36.7% thus there is a potential for nonresponse bias. The pre-election study was carried out from August 2020 up to election day on November 3rd, thus depending on the timing in which the responder was surveyed they may have different answers. Nonetheless, the data is relatively independent and identical, cross-sectional samples were randomly drawn with all included residential addresses across the all states having an equal probability of selection.
3) Metric data. Since age is quantitative with a set zero point, it is metric ratio data.
4) Unpaired data. We consider democrats and republicans to be independent groups.
5) As the variance of the subsets is unknown, and the Levene test rejected the hypothesis that the variances are equal, we used the Welch's t-test, which does not require sample variances to be the same.

We also tested the effective size of our mean difference using the Cohen's D test. The test assumes normally distributed, continuous metric data, which we address in the above bullets.

```
#conducting Levene's test
vareq = t.test((repub-mean(repub))^2,(dem-mean(dem))^2,alternative="two.sided", var.equal=TRUE)
#since the variability is still different (p<0.05), we will need to use Welch's test
t <- t.test(repub, dem,alternative="two.sided", var.equal=FALSE)
cat('Welch\'s unpaired t-test p-value:',t$p.value)
```

```
## Welch's unpaired t-test p-value: 3.352032e-10
```

```
#using Cohen's D to calculate effect size
cohens_d <- function(x, y) {
    csd <- ((length(x)-1)*var(x)+(length(y)-1)*var(y))/((length(x)-1)+(length(y)-1))
    cd  <- (abs(mean(x) - mean(y)))/(sqrt(csd))}
cd <- cohens_d(repub,dem)
cat('Cohen\'s D Value: ',cd)
```

```
## Cohen's D Value:  0.1763146
```

Test, results and interpretation

Since the p-value of the Welch's t-test was 3.35^{-10}, we reject the null hypothesis that the mean voter age of the two parties are equal. Democrats, with a mean age of 52, on average are younger than Republicans, which had a mean age of 55. These results align with the lower value distributions for Democrats seen in the boxplot (Fig 1 above).

Next, we evaluated the practical significance of this difference using Cohen's D. In this, our Cohen's D value of the magnitude of the difference was only 0.176, which corresponds to a little to no effective size. The practical significance of the difference in means is therefore rather minor. Intuitively this makes sense, since we observed that the mean ages in the two parties are only 3 years apart. The significance of the p-value in our Welch's t-test could be indicative of our relatively large sample size of 5094 test points. From both tests we conclude that Democratic voters in the 2020 election are, on average, younger than Republican voters, even though the size of this difference is minor.