Lab 1: Question 1

Jeff Adams, Brittany Dougall, Li Jin, Jerico Johns

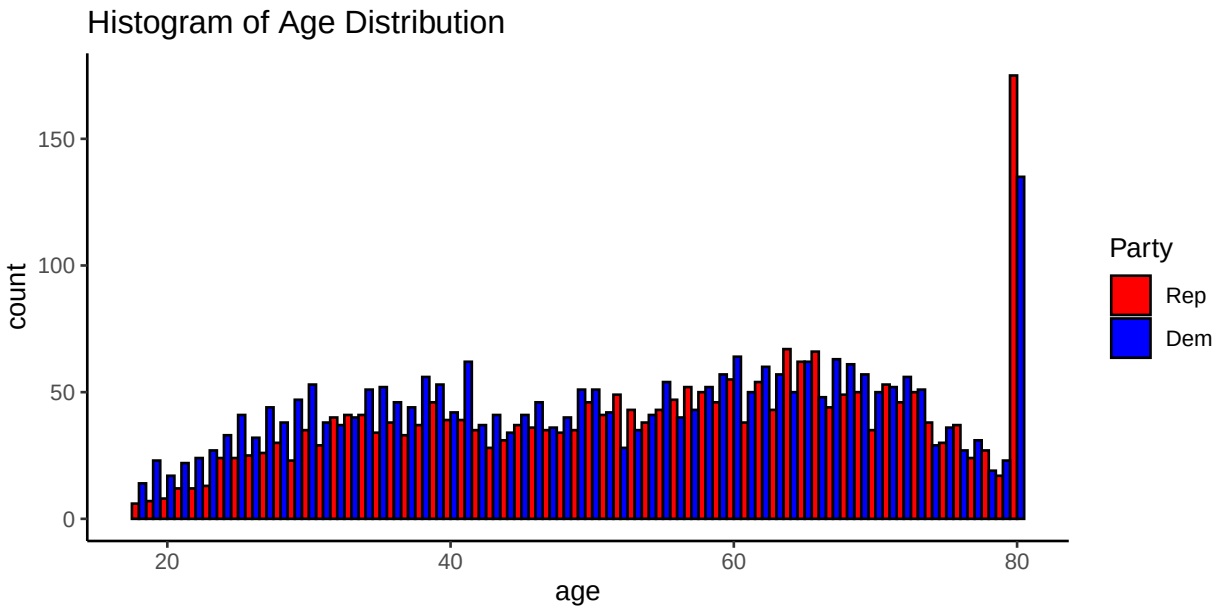
```
install.packages("effsize", repos = " https://CRAN.R-project.org/")
library(dplyr)
library(ggplot2)
library(tidyverse)
library(haven)
library(effsize)
```

Importance and Context

The United States has been on a two party political system since the 1850's. These two parties have been the republican and democratic party. Each party has had its ups and downs throughout the years. They have also appealed to different demographics over the years. It has long been the goal of each party to gain the vote of the younger generation. The younger generation has consistently had the lowest voter turn out (http://www.electproject.org/). If a party could capture this vote they would be gaining a large percentage of brand new voters. As well these voters have many years of voting ahead of them. So the young vote helps the longevity of the party. We would like to look at a recent survey to see if the age demographic between the two parties is different by comparing the mean age for Republican respondents to the mean age of Democratic respondents using a two sample t-test.

Description of Data

Data is collected from the 2020 America National Election Survey (ANES), that uses cross-sectional random sampling on USPS household records to survey a single individual from a randomly selected household regarding their political opinions and voting behavior in the upcoming U.S. presidential election (data collected between August 18, 2020 and November 3rd, 2020). The two main variables for this analysis are party affiliation and age. We used code V201228 to determine what party each respondent affiliated with. The question on the survey asked the respondent what party they self identified with. We also used code V201507x to determine the respondents age. The respondent could refuse to submit an age, so they were excluded from the data set. This resulted in the removal of 78 Democrats and 116 Republicans from the data. This represents a small portion of the total sample size. So we were not concerned with the removal of these data points. After the data clean up we had 2,786 Democratic respondents and 2,448 Republican respondents. You can see the distribution of ages for each party below. As you can see both parties have similar distributions, but one thing to note is that both parties have a large number of respondents that are marked as 80 years old. This is because the survey had a category of 80+ and counted that as 80 years old in the table. This means that the respondents could actually be much older than we are showing, but because of the nature of the survey we are missing that data and truncating it at 80. Republicans have a slightly large portion than Democrats at 80 years old (7% vs 5%). This will be important to account for when looking at the practical significance of the results.



Most appropriate test

We believe that the most appropriate test for our question is Welch's Two Sample t-test (for independent samples). The two samples that we are comparing are completely independent of each other, because one person can either be a Democrat or a Republican, these two are not linked. We also don't have a very strong assumption of the data leaning one way or another so we need to use the two tailed test.

There are three main assumptions needed for this test:

- The data needs to be numeric. The age data that we are using is numeric between 18 and 80.
- The sample needs to be independent and identically distributed (iid). The data is being pulled from only one individual per household. So there should not be respondents that are impacting each others results. So the data is independent. Also the data is being collected in a very short period of time period within the election, so the population is not changing. Therefore the data is also identical.
- The data should have no major deviations in normality, considering the size of the sample. As seen in the charts above the data is not normally distributed, but our sample sizes are 2,500 or greater. This is well above 30, so the central limit theorem will kick in and we will have no problem meeting this assumption.

Given these assumptions being met we will evaluate the null hypothesis with an α of .05:

- H_0: There is no difference in the mean ages between Democrats and Republicans. $\mu_1 = \mu_2$
- H_a: There is a difference in the mean ages between Democrats and Republicans. $\mu_1 <> \mu_2$

Test, results and interpretation

Below you can see the results of our Welch's Two Sample t-test. The t value for this test is -5.75, this is well outside of our 95% confidence interval [-3.63 to -1.78]. This results in a $p = 9.37e-09 < \alpha = 0.05$. Therefore we have enough statistical significance to reject the null hypothesis. It is important to look at the practical significance as well. If you look at the two means they are only showing a difference of 3 years (Effect Size) for the two groups. This isn't a lot of practical significance. As well the effect size from Cohen's D is |.16| which is very small, showing little to no practical significance using this metric as well. So although we can reject the null hypothesis the actual age difference might not be very significant, but it does show that Democrats lean younger than Republicans. Another important note is that Republicans have a higher

percentage of people in the 80+ group, this means that we could possibly be truncating older people from the republican group. This could cause a larger spread in our means, giving us more practical significance.

```
##
##  Welch Two Sample t-test
##
## data:  df_dem$age and df_rep$age
## t = -5.7512, df = 5185, p-value = 9.366e-09
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.630179 -1.784487
## sample estimates:
## mean of x mean of y
##  51.61701  54.32435
```

```
##
## Cohen's d
##
## d estimate: -0.1589722 (negligible)
## 95 percent confidence interval:
##      lower      upper
## -0.2133662 -0.1045781
```