# Lab 2

W203 Statistics for Data Science

*Section 6 - Team 1*

## Before Starting: General Data Considerations

To ensure the quality of the dataset, we performed analysis only on individuals whose responses were mostly trustworthy and sincere. We pre-filtered our participants based on their "RESPONSE QUALITY" and considered only those who were "Never" or "Some of the time" insincere in their responses, and those who answered questions honestly "Most of the time" or "Always". For this subsample, we believe that we can expect reliable answers more often than not. This reduced the sample size from 2500 to 2064, which is not a significant drop for integrity of analysis.

```r
setwd("~/Desktop/Stone/Berkeley_MIDS/Statistics/Labs/Lab_2")
full_data <- read.csv('anes_pilot_2018.csv')

library(ggplot2)
suppressMessages(library(dplyr))

A <- full_data %>% filter(honest >= 4 & nonserious <= 2) #subsample stored in A

print(dim(full_data)[1])
```

```
## [1] 2500
```

```r
print(dim(A)[1])
```

```
## [1] 2064
```

# Question 1: Do US voters have more respect for the police or for journalists?

## Introduction and Operationalization

The survey asks participants to rate journalists (ftjournal) and police (ftpolice). Responses range from 1 to 100, but the survey was presented as a thermometer with 9 labels from very cold (0) to very warm (100). We believe that these rankings are similar to a Likert scale. Namely, we do not believe that giving a ranking of 75 is different from a ranking of 80 because both numbers are between the "Fairly Warm" and "Quite Warm categories. For this reason, we binned respondents into 1 of 9 categories as defined by the survey scale. 0-15 was"Very Cold", 15-30 was"Quite Cold", etc. Those who responded 100 received their own bin for two reasons: 1. The scale was capped at 100, so we do not know whether these individuals would have gone higher if allowed; and 2. These are the only individuals who can be classified as"Very Warm" (even 99 does pass the bar on the thermometer into "Very Warm").

One deficiency here is that participants are asked to "rate" a group, which does not necessarily translate to respect for the group. For example, one can rate the police highly because they believe the group is effective at responding to emergencies, but one may lack respect if one believes the police sometimes perform their jobs unethically.

We also assume that a single person's judgement for these bins do not change between police and journalists. Namely, if individuals rate the police and journalists both in the "Fairly cold" bin, we assume that their opinions for both groups are roughly equal.

We wanted to compare the rating given to police and journalists for each individual voter. Because of this, we required that all respondents answered both questions. Also, to focus the sample on US voters, we only looked at individuals who were registered to vote. These are many reasons why someone might not vote, but if they are registered, they should be classified as eligible US voters. There were only two individuals who did not respond to ftjournal, and all responded to ftpolice. In addition, 316 were not registered to vote. This reduces our dataset to 1746 participants.

```
dim(A %>% filter(ftjournal >=0 & ftpolice >=0) %>% filter(reg<=2, reg>0))[1]
```

```
## [1] 1746
```

```
JR_PL_data <- A %>%
  filter(ftjournal >=0 & ftpolice >=0) %>%
  filter(reg<=2) %>% select(caseid, ftpolice, ftjournal)

#Ensures we get the same bins as in the study
ftbins <- c(-0.01,15,30,40,50,60,70,85,99.99,100)

JR_PL_ordinal_data <- data.frame(
  cut(JR_PL_data$ftpolice, breaks=ftbins) %>% as.numeric(),
  cut(JR_PL_data$ftjournal, breaks=ftbins) %>% as.numeric()
)
colnames(JR_PL_ordinal_data) <- c("Police Rating", 'Journalist Rating')

JR_PL_ordinal_data$Police_minus_Journalist <-
  JR_PL_ordinal_data$`Police Rating` -
  JR_PL_ordinal_data$`Journalist Rating`
```

## Exploratory Data Analysis (EDA)

First, we looked at summary statistics Police and Journalist ratings

```r
summary(JR_PL_ordinal_data$`Police Rating`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   7.000   6.036   8.000   9.000
```

```r
summary(JR_PL_ordinal_data$`Journalist Rating`)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   5.000   4.883   8.000   9.000
```

We see that these take anticipated min/max bin values of 1 and 9.

Second, we looked at summary statistics for the difference of Police rating minus Journalist rating for each voter.

```r
summary(JR_PL_ordinal_data$Police_minus_Journalist)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -8.000  -2.000   0.000   1.152   5.000   8.000
```

The difference takes on both negative and positive values with the same minimum and maximum difference in ranks. There are individuals who rated Police in bin 1 and Journalists in bin 9, and those who rated Police 9 and Journalists 1 (-8 and 8 respectively). Our 3rd Quartile is larger in magnitude than our 2nd Quartile, with the mean shifted upwards, indicating that there is skew toward favoring the police more.

To visualize this, we plot a histogram of the difference in ranks.

```r
ggplot(data = JR_PL_ordinal_data, aes(x=Police_minus_Journalist)) +
  geom_histogram(alpha=0.5,
                 binwidth = 1) +
  geom_density(alpha=0.5)+
  labs(title='Bin difference between Police Rankings and Journalist Rankings for U.S. Voters',
       x='Difference per Respondent (Police Ranking - Journalist Ranking)',
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  ylim(0,275)+
  scale_x_continuous(breaks=seq(-10, 10, by=1))+
  stat_bin(aes(y=..count.., label=ifelse(..count.. > 0, ..count.., "")),
           geom="text",
           vjust=-.5,
           binwidth=1
           )
```

## Bin difference between Police Rankings and Journalist Rankings for U.S. Vote
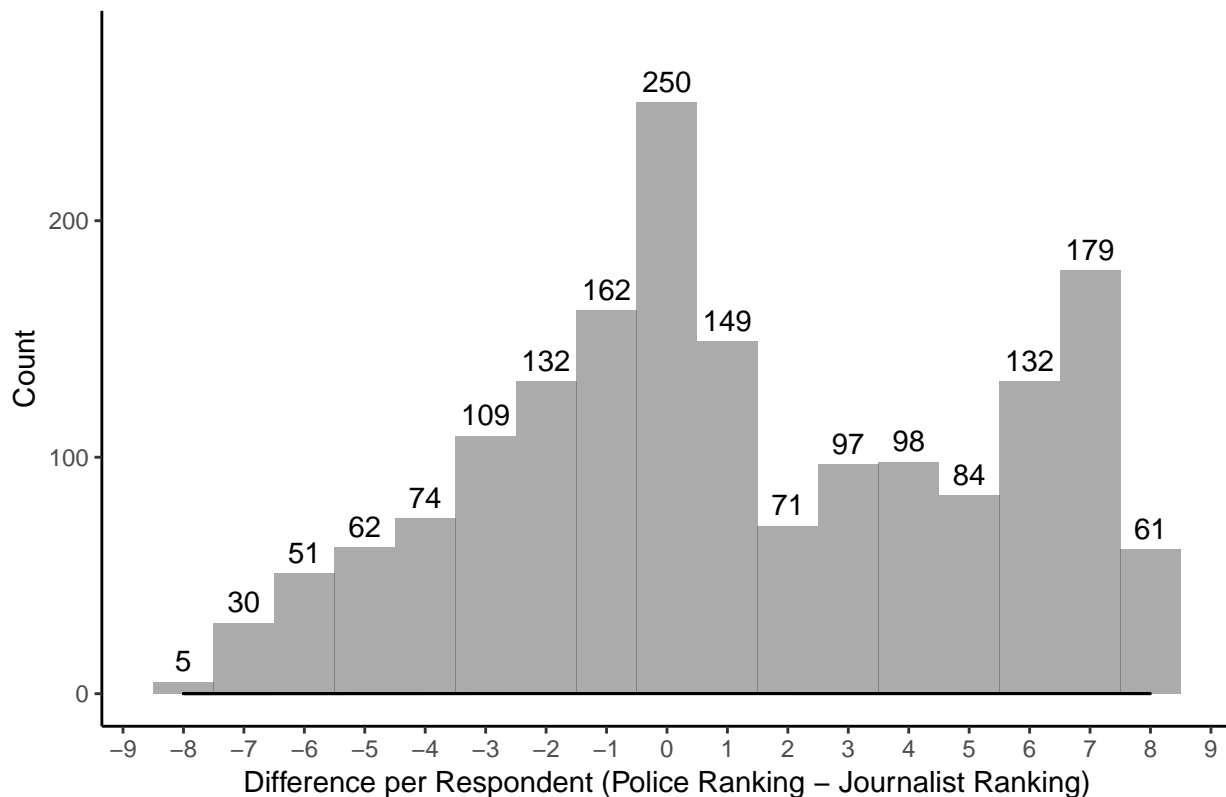


While the distribution mode is 0, there are sharp peaks in the positive region skewing the data to the right (the police was rated higher than the journalists).

## Hypothesis Test Selection

The most appropriate test here is the Wilcoxon signed-rank test, which modern research shows is valid for ordinal data. Consider the **assumptions** this test makes.

1. **Data is paired, but each pair is iid.**

The data is paired since it is the same respondent on two different questions. There is no reason to believe the answers of a participant will affect the answers of others since the survey was administered online, so we assume independence. In additions, even though the sample does not represent all eligible US voters, the samples are derived from the same population, which is the pool of US voters who responded to the survey.

2. **The differences are measured on at least an ordinal scale.**

Since the variables are on an ordinal scale, we cannot use the paired t-test regardless of the sample size. But we can subtract the ranks we structured (scale 1-9) and look at the difference between ranking between police and journalist rating for each respondent.

3. **The distribution is symmetric around its mean and median.**

This is true under our null hypothesis, which is that the difference between the ranks of the two groups is 0. The ranking difference takes integer values from -8 to +8 and is symmetric around 0.

## Hypothesis Testing

Let D represent the r.v. of the difference between ratings of the police and journalists. At a significance level of $\alpha = 0.05$, we state our hypotheses as:

$$H_0 : D = 0$$
$$H_a : D \neq 0$$

We used a two-tailed test since we did not have prior evidence that the rating of one group should be greater than the other.

After conducting the test:

```r
wilcox.test(JR_PL_ordinal_data$`Police Rating`,
            JR_PL_ordinal_data$`Journalist Rating`,
            paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  JR_PL_ordinal_data$`Police Rating` and JR_PL_ordinal_data$`Journalist Rating`
## V = 748390, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

We observed that there is a large statistical difference between the rating of the two groups, with $p-value << 0.001 << \alpha$. This means that if $H_0$ was true, we would have about 2.2e-16 % chance of seeing data as extreme as what was observed. Therefore, we rejected our null hypothesis in favor of the alternative that the difference in rating between the two groups is not 0.

We can also see that there are more observations for which the paired difference was positive for $police - journalists$:

```r
sum(JR_PL_ordinal_data$Police_minus_Journalist > 0)
```

```
## [1] 871
```

```r
sum(JR_PL_ordinal_data$Police_minus_Journalist < 0)
```

```
## [1] 625
```

Since more individuals rated the police higher than journalists, our best conclusions is that **US voter have more respect for the police than they do for journalists.**

In order to compute the effect size, we computed a correlation-like value by dividing the z-score of the p-value by sqrt(n). We could do this because our sample size is very large (n=1746).

$$r = \frac{z}{\sqrt{n}}$$

This approach produces an effect size

```r
p.value <- 1.169115e-29 #from wilcox.test
z.score <- -qnorm(p.value/2)
effect.correlation <- z.score / sqrt(length(JR_PL_ordinal_data$Police_minus_Journalist))
print(effect.correlation)
```

```
## [1] 0.2706734
```

The effect correlation r is between small (0.1) and medium (0.3) effect size. While the test is highly statistically significant, the effect size is considered weak. Consequently, there is not a large practical significance. This agrees with intuition: the median individual had a 0 difference in rank for both police versus journalist, while the average individual in this sample had a slightly greater than 1 bin difference, which does not feel like a major difference.

## Question 2: Are Republican voters older or younger than Democratic voters?

### Introduction and Operationalization

In order to determine age in years, the variable birthyr (year of birth) was subtracted from 2018, when survey was conducted. This is the best guess for a person age since there is no further information on birth months. Because of this, we may be -1 year off if the individual hasn't had their birthday in 2018.

To determine voter status and party affliation, we used self-classified categories from pid7x. We chose this over how participants actually voted in 2016 and 2018. It can often be the case that self-identified Democrats will vote Republican and vice-versa if for a particular election they happen to favor an opposing candidaten. This behavior does not change their true party affliation, and the best guess we have for that is what they consider themselves to be. The age and party information was collected below.

```
AG_PY_data <- A %>% mutate(party =
                              ifelse(pid7x == 1 | pid7x == 2, 'D',
                                     ifelse(pid7x == 6 | pid7x == 7, 'R', 'I'))) %>%
  mutate(age = 2018 - birthyr) %>%
  filter(party != 'I', reg == 1 | reg == 2) %>%
  select(age, party)
```

### Exploratory Data Analysis (EDA)

We first examined the age variable.

```
summary(AG_PY_data$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   42.50   57.00   54.18   65.50   91.00
```

Age extend over a large range, from the minimum voting age 18 up to age 91. There are no missing values in the age. The median is 57, while mean is a little over 54. We can also get a sense of the ages of Democratic versus Republican voters.

```
AG_PY_data %>% filter(party == 'D') %>% select(age) %>% summary
```

```
##       age
##  Min.   :18.00
##  1st Qu.:39.00
##  Median :55.00
##  Mean   :52.57
##  3rd Qu.:64.00
##  Max.   :91.00
```

```
AG_PY_data %>% filter(party == 'R') %>% select(age) %>% summary
```

```
##       age
##  Min.   :18.00
##  1st Qu.:47.00
##  Median :59.00
##  Mean   :56.49
##  3rd Qu.:67.00
##  Max.   :90.00
```

The min/max ages for both groups are very similar, but the median/mean seems to indicate that Republicans seem to be on average a bit older. We checked whether there is an equal number of Democrats and Republicans in our sample.

```
dim(AG_PY_data %>% filter(party == 'D'))[1]
```

## [1] 669

```
dim(AG_PY_data %>% filter(party == 'R'))[1]
```

## [1] 466

Our total sample size is 1135. There are 203 more Democrats than Republicans in our sample. Therefore, to visualize any differences, it would make more sense to plot a density histogram. If we were to plot the count, we would expect Democrats to dominate the bins simply because that there are more of them.

```
ggplot(data = AG_PY_data) +
  geom_histogram(aes(x=age, y=(..density..), fill=party),
                 alpha=0.3,
                 breaks=seq(15,100,5),
                 position='identity') +
  labs(title='Frequency of Age Grouped by Political Leaning',
       x='Age',
       y = "Frequency of Respondents",
       fill = "Party") +
  scale_fill_manual(labels = c("Democrats", "Republicans"),
                    values=c("blue1", "red1"))+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  scale_x_continuous(breaks=seq(15,100, by=5))
```
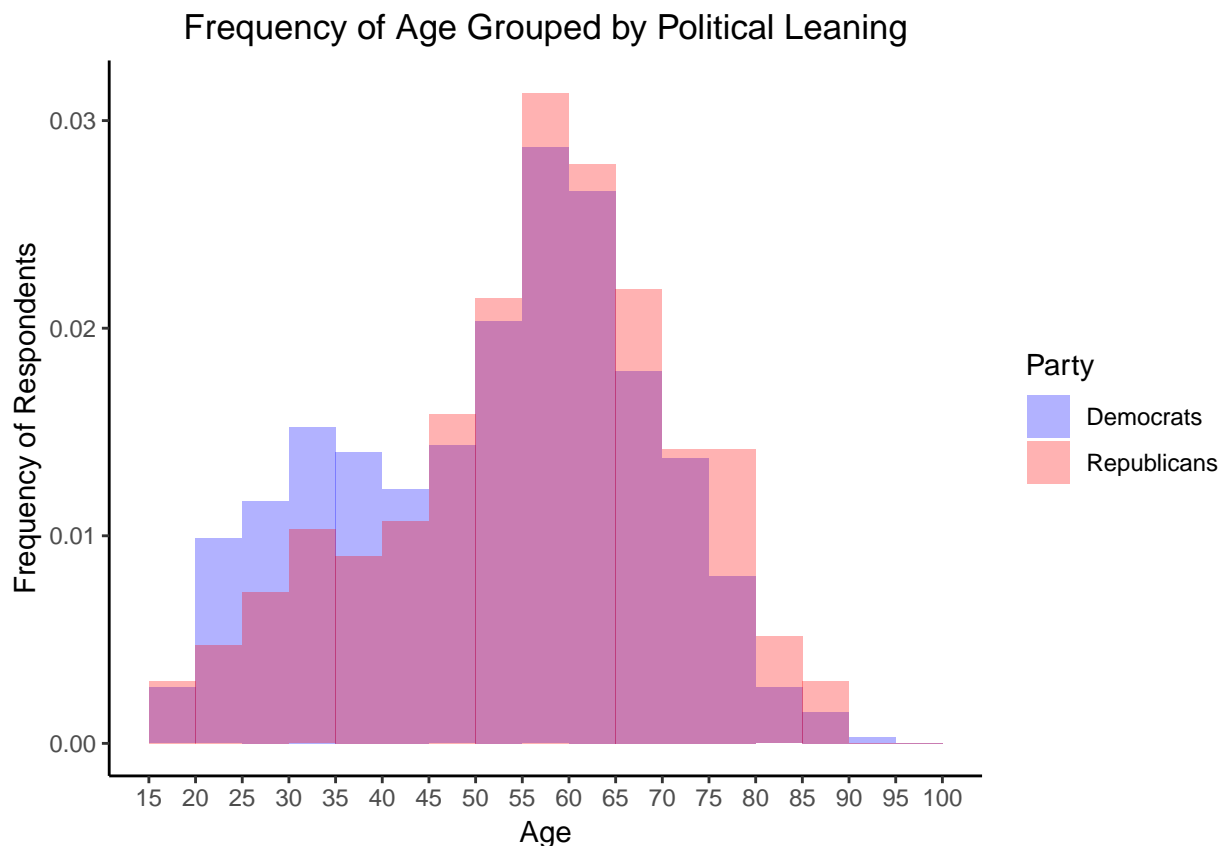


While the overlap (magenta color) is large with mode for both parties around 55-60, there appears to be a bigger population from 20 to 40 for Democrats. From ages 45 and up, there are at least as many Republicans

7

as there are Democrats, with the exception of the very last bin of n=1, where the oldest candidate (91 years) was Democratic. Both ages are skewed left, with Democrats having a greater skew left.

## Hypothesis Test Selection

Since we have a metric variable, and we do not know the population variance, and there is no natural pairing between the participants across parties, we performed an independent two sample t-test.

Consider the **assumptions** this test makes and the data used in this section.

1. **Ratio data**

Since we are on a metric scale, and there is an absolute zero (age = 0), we have ratio data (age). We can add/subtract/take ratios of age and get sensible statistics.

2. **iid observations**

As in Question 1, the participants ages do not affect one another and are taken from the same survey.

3. **Normal population or large sample**

Since our sample sizes are extremely large for both samples (n=669 for Democrat; n=446 for Republican; both>>30), according to the Central Limit Theorem, the sampling distribution of our mean for both samples will be close to normal. This makes the t-test valid even if the population distribution is skewed left as demonstrated above.

4. **No outliers**

Outliers negatively affect the quality of a t-test, and are represented by boxplots as data point lying outside of 1.5 times the interquartile range above the upper quartile or below the lower quartile. We could see from the plots that there were no outliers (which would be displayed as RED).

```
ggplot(data = AG_PY_data) +
  geom_boxplot(aes(x=party, y=age),
  position = "dodge2",
  outlier.color = 'red')+
  labs(title='No outliers in the boxplot of ages of Democratic and Republican voters',
       x='Party',
       y='Age') +
  scale_x_discrete(
        labels=c("Democrats","Republican")) +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5))
```

## No outliers in the boxplot of ages of Democratic and Republican voters
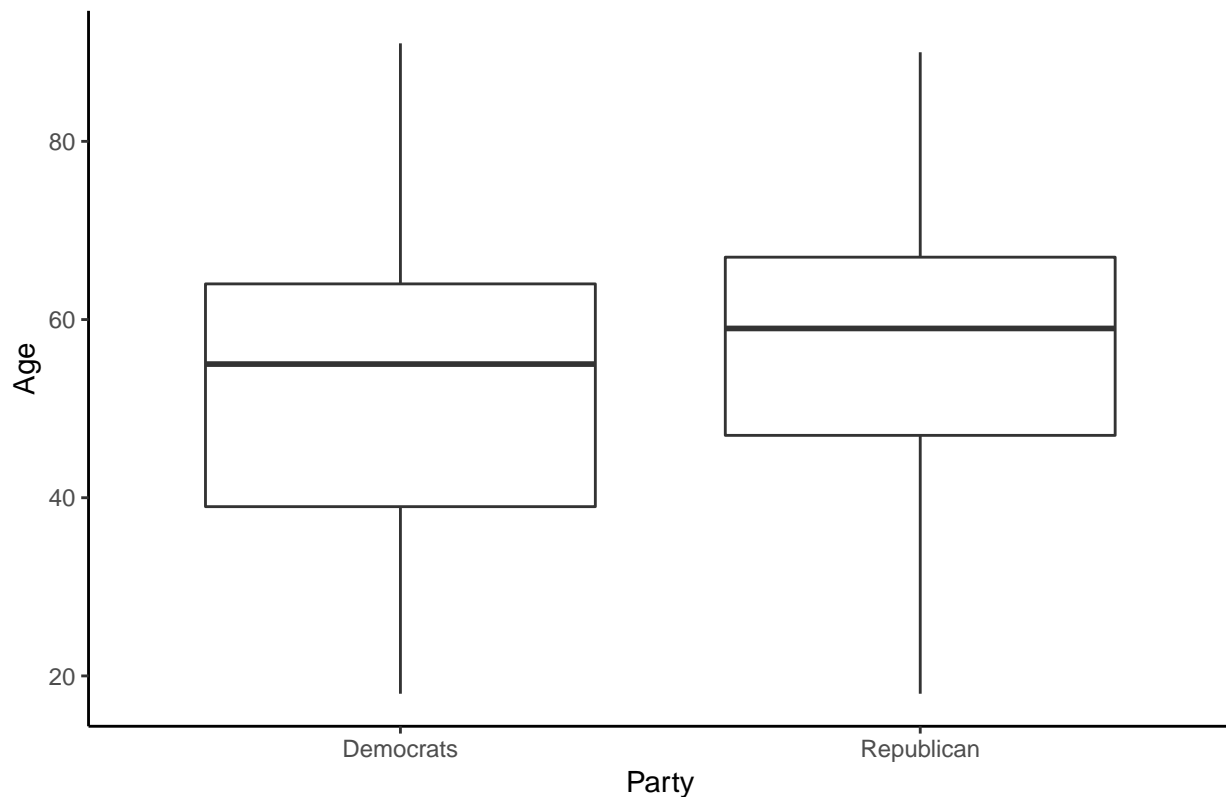


5. **Homogeneity of variance**

To determine whether the standard deviation of our two independent samples are equivalent or different we applied Levene's test for Equality Of Variances at a significance level of 0.05. We state our hypotheses as follows:

$$H_0 : \sigma_D = \sigma_R$$
$$H_a : \sigma_D \neq \sigma_R$$

```
library(inferr)

age.levene <- infer_levene_test(AG_PY_data,age, group_var = party)
print(age.levene$p_lev)
```

```
## [1] 0.0586
```

Since $p - value > 0.05$, we fail to reject the null hypothesis that the variance of the two groups are equal. As a result, our t-test will assume equal variances in the two populations' standard deviation. We do note that the p-value is very close to 0.05; however, we cannot ignore the fact that under H0, we had a $>5\%$ chance of seeing data this extreme.

## Hypothesis Testing

We setup our t-test as follows with significance $\alpha = 0.05$, and subscripts on the mean representing party:

$$H_0 : \mu_D = \mu_R$$
$$H_a : \mu_D \neq \mu_R$$

We used a two-tailed test since we did not have prior evidence that the age of one group should be greater than the other.

```r
t.test(AG_PY_data[AG_PY_data$party == 'D',]$age,
       AG_PY_data[AG_PY_data$party == 'R',]$age,
       paired = FALSE,
       var.equal = TRUE,
       alternative = 't')
```

```
##
##  Two Sample t-test
##
## data:  AG_PY_data[AG_PY_data$party == "D", ]$age and AG_PY_data[AG_PY_data$party == "R", ]$age
## t = -4.0837, df = 1133, p-value = 4.744e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.809705 -2.038791
## sample estimates:
## mean of x mean of y
##  52.56502  56.48927
```

Since $p\text{-value} << 0.001$, we reject our null hypothesis in favor of the alterative hypothesis that the difference in age between Democratic voters and Republican voters is not 0. Since the mean for Democratic voters are younger from our EDA, **we concluded that Democratic voters are younger on average**.

To evaluate effect size, we used the the Cohen's D statistic.

```r
library(effsize)
cohen.d(AG_PY_data[AG_PY_data$party == 'R',]$age,
        AG_PY_data[AG_PY_data$party == 'D',]$age)
```

```
##
## Cohen's d
##
## d estimate: 0.2464019 (small)
## 95 percent confidence interval:
##     lower     upper
## 0.1275807 0.3652230
```

Since our Cohen's D value is 0.246, it is between what is typically considered small (0.2) and medium (0.5), so we have a reasonably small effect size. This is also in-line with intuition. From the EDA above, the mean difference in age is 3.92 years, which is non-zero, but fairly small given that the range of ages goes from 18 to 91, a difference of 73 years.

## Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?

### Introduction and Operationalization

We considered only respondents who were registered to vote (as in question 2), and self-identified as Independent. To determine who believed the investigation was baseless, we selected individuals who believed that there wasn't Russian interference in the 2016 presidential elections to try to help Donald Trump win. One caveat that we needed to make an assumption: for participants who believed Russian interference had occurred, they also believed the interference was in favor of Donald Trump. For example, if survey takers had believed the interference occurred but favored neither candidate, due to the inclusion of Trump in the survey question, these individuals would have selected "This probably did not happen." However, we feel this is a fair assumption since the general belief is that any interference would have been in favor of Trump.

Another relevant variable is the approval of the Mueller investigations (muellerinv). We feel this survey question addresses more than whether the investigation was baseless. Namely, those who "disapproved"" could have found that the investigation was indeed warranted (not baseless), but instead poorly executed, or poorly communicated to the public. Due to potential irrelevant effects, this variable was not taken into account.

```
FBI_IN_data <- A %>% filter(pid7x <= 5 & pid7x >=3) %>%
  filter(reg <= 2, reg>0) %>%
  select(russia16)
```

### Exploratory Data Analysis (EDA)

First, we generated a summary of the russian16 variable for Independent voters.

```
min(FBI_IN_data$russia16)
```

```
## [1] 1
```

```
max(FBI_IN_data$russia16)
```

```
## [1] 2
```

```
FBI_IN_data <- FBI_IN_data %>% filter(russia16 > 0)
sum(FBI_IN_data$russia16 == 1)
```

```
## [1] 340
```

```
sum(FBI_IN_data$russia16 == 2)
```

```
## [1] 263
```

The minimum value is 1 and maximum is 2, i.e. everyone in our sample answered the relevant question. There are a total of 340 individuals who believed that Russia probably interfered (Not Baseless), and 263 that believed that Russia probably did not (Baseless). We can also visualize this with a bar plot of this binary variable.

```
Count <- c(sum(FBI_IN_data==1), sum(FBI_IN_data==2))
Group <- c('Not Baseless', 'Baseless')
df_counts <- data.frame(Count, Group)

p <- ggplot(df_counts)
p + geom_bar(aes(x=Group, y=Count), stat='identity', position=position_dodge())+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(x=Group, y=Count, label=Count), vjust=1.6, color="white",
```

```
            position = position_dodge(0.9), size=3.5)+
    labs(title="Opinion on Base for FBI Investigation among Independent Voters")+
    scale_fill_brewer(palette="Paired")
```
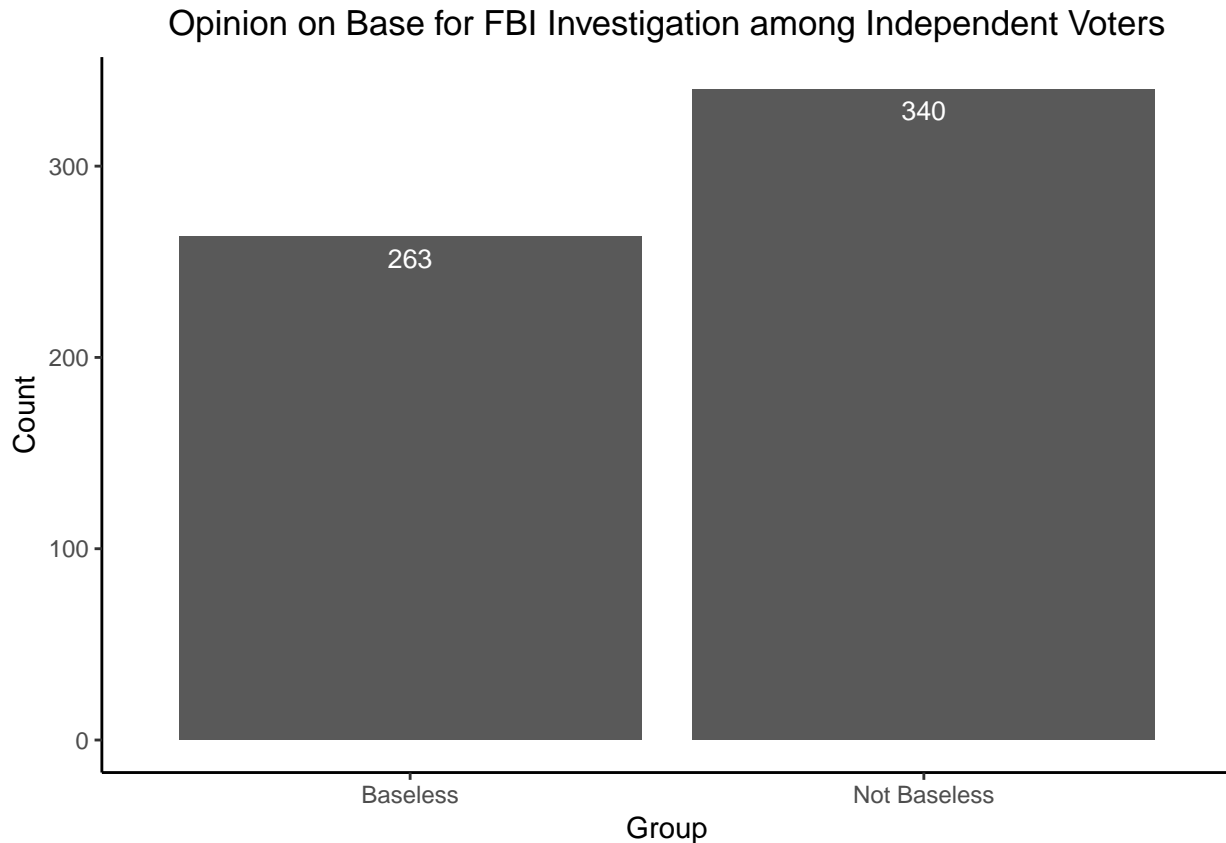


While more responded as "not baseless," it is not by a blowout rate (~56%).

## Hypothesis Test Selection

For this question, we used a binomial test for population proportion. This test compares the number of successes to the hypothesized number of successes for a sample of particular size. In this case, we call success as individuals who believe the investigation was baseless.

Consider the **assumptions** this test makes and the data used in this section.

1. **Samples are dichotomous and nominal**

The subjects are classified as either believing the investigation was baseless or not, with no third category, and no numerical value associated with each category.

2. **iid samples**

Independence and identical distribution are assumed because respondents do not influence one another and all respondents were drawn from the same population of survey takers.

## Hypothesis Testing

With $p$ representing the successful population proportion, our null hypothesis was that the same proportion of individuals believe the investigation was baseless as those who did not. We performed a 2-tailed test because we do not have a strong prior that the population was strongly leaning in either direction, and set our confidence level at 0.05.

$$H_0 : p = 0.5$$
$$H_a : p \neq 0.5$$

For testing, we used the binom.test function in R. This takes a parameter for number of successes (number of people who believe the investigation was baseless), number of trials (sample size), and 't' for two-sided test:

```
binom.test(sum(FBI_IN_data$russia16 == 2),
           length(FBI_IN_data$russia16),
           p=0.5,
           alternative = 't')
```

```
##
##  Exact binomial test
##
## data:  sum(FBI_IN_data$russia16 == 2) and length(FBI_IN_data$russia16)
## number of successes = 263, number of trials = 603, p-value =
## 0.001944
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3961389 0.4767945
## sample estimates:
## probability of success
##              0.4361526
```

Since $p - value < 0.05$, we rejected the null hypothesis and state that the proportion of individuals who think the investigation was baseless was not 0.5. Since the fraction of individuals who thought the investigation was baseless was 0.456, **we concluded that a statistically significant majority of Independent voters did not believe that the investigation was baseless**.

In order to calculate the practical significance, we will use the metric of Cohen's g, which is valid for the one-sample binomial test where the null hypothesis states that the proportion is 0.5. Cohen's g is simply the difference of sample proportion of individuals who believed the investigation was baseless with 0.5, and then taking the absolute value. The closer the sample proportion is to 0.5, the smaller the effect size.

```
cohen.g <- abs(sum(FBI_IN_data$russia16 ==  2) / length(FBI_IN_data$russia16) - 0.5)
print(cohen.g)
```

```
## [1] 0.06384743
```

The effect size is $\approx 0.064$, which is small. Typically, the ranges are <0.15 (small), 0.15-0.25 (medium), and >=0.25 (large). Therefore, our study was statistically significant with low level of practical significance. This makes sense since almost 46% of the sample believe the investigation was baseless. This is not a large distance from 50%.

# Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?

## Introduction and Operationalization

The voter turnout categories are stored in turnout16 and turnout18. The overall turnout for 2018 was actually lower than for 2016. However, they were different races: 2016 was presidential and 2018 was for House and Senate. This was likely an important reason for the overall decline. As a result, the data sample for this question was limited to the individuals that did not vote in 2016, but did vote in 2018. Also, we eliminated anyone who was 19 years or younger at the time of the survey, because they were below the legal voting age in 2016.

There are three questions on the survey for which anger and fear were measured. The first is on Donald Trump, second on immigration, and third on general feeling about conditions in the country. However, only the third (stored in the geafraid and geangry variables) was asked for all individuals. We do not believe that immigration feelings are comparable to feelings about Trump; as a result, we only used the general feeling variables.

We assumed that for each person, their perceived scales for anger and fear were comparable. Namely, if they choose 4 for anger and 3 for fear, anger was a more important factor for that individual. We also eliminated individuals that gave both feelings the same ranking, so neither was more important to them. As a result, we could classify individuals into more angry or more fearful categories by looking at the sign of the difference in their ranked values.

One deficiency here is that we could not measure casual relationships, which is what the question seeks with the term "driving". The question we could answer is: "for the group of individuals that were eligible to vote in 2018 but did not vote in 2016, was fear or anger levels higher amongst that group?". To determine casuality would require careful controlled experimentation.

## Exploratory Data Analysis (EDA)

The following gets the data as described above.

```
AG_FR_turnout <- A %>%
                  filter(turnout16 == 2 & turnout18 == 1) %>%
                  filter(2018 - birthyr > 19) %>%
                  filter(reg > 0 & reg <=2) %>%
                  select(turnout16, turnout18, geangry, geafraid, birthyr) %>%
                  mutate(age = 2018 - birthyr) %>%
                  mutate(angry_minus_fear = geangry - geafraid)
```

We did a sanity check on age, for them to be equal or greater than 20 in 2018:

```
summary(AG_FR_turnout$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    22.0    28.5    43.0    42.6    55.5    73.0
```

We could also visualize the difference between angry and fear levels with a histogram:

```
ggplot(data = AG_FR_turnout, aes(x=angry_minus_fear)) +
  geom_histogram(alpha=0.5,
                 binwidth = 1) +
  labs(title='Difference Between Anger & Fear Levels for 2018 Voters that did not vote in 2016',
       x='Difference per Respondent',
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
```

```
ylim(0,20)+
scale_x_continuous(breaks=seq(-2, 3, by=1))+
stat_bin(aes(y=..count.., label=ifelse(..count.. > 0, ..count.., "")),
         geom="text",
         vjust=-.5,
         binwidth=1
         )
```

## Difference Between Anger & Fear Levels for 2018 Voters that did not vote in 20
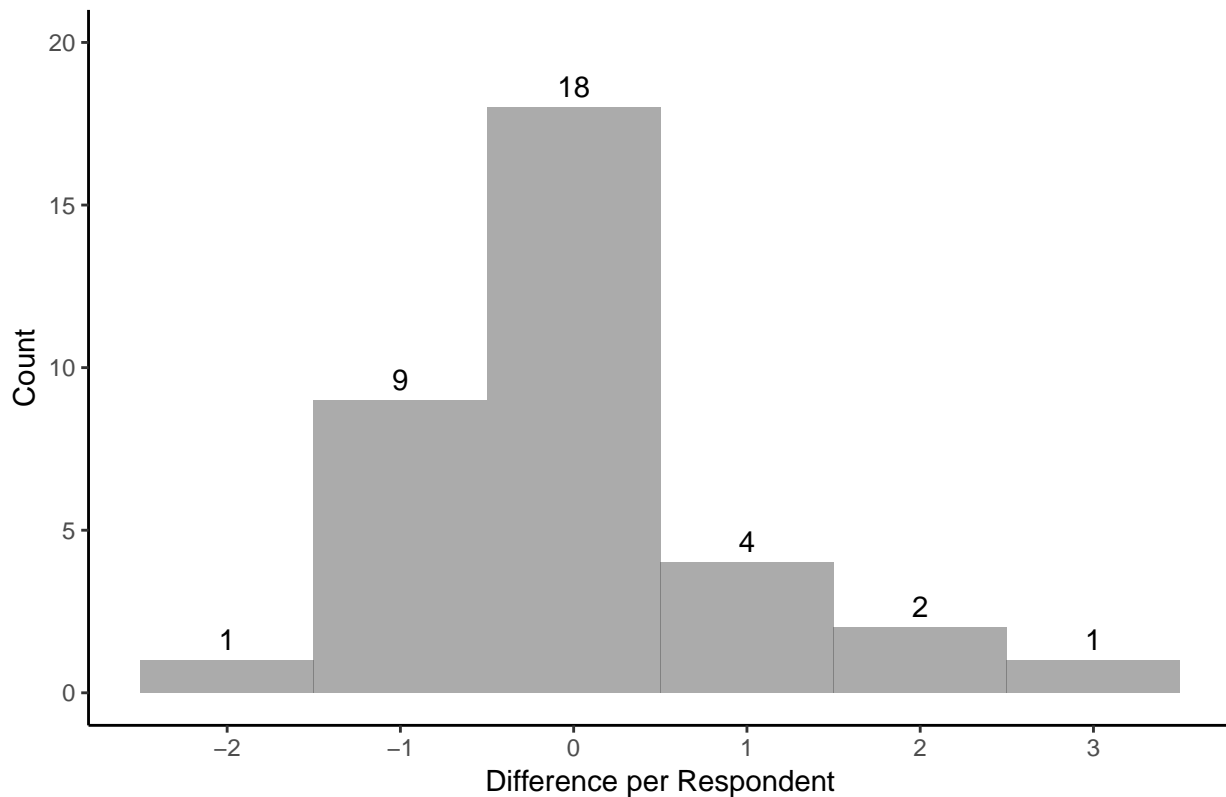


The minimum is -2, so fear was more relevant for some individuals, and maximum is +3, so anger was more relevant for other individuals. We see from this histogram that for the majority, the difference was 0. We have a small sample size of 17 (excluding those who had no preference), and the data appears slightly skewed to the right. The mode for those who had a preference was -1 (anger levels were 1 higher than anger).

### Hypothesis Test Selection

The most appropriate test is the exact binomial test for proportion due to small sample size. Consider the **assumptions** this test makes and the data used in this section.

1. **Samples are dichotomous and nominal.**

After our initial treatment of the data, the subjects were classified as either more angry or more afraid, with no possible third category. There is also not a numerical value associated with either category.

2. **iid samples**

As before, we assume respondents do not influence one another, and all respondents were drawn from the same population of survey takers.

## Hypothesis Testing

Let $L$ represent the proportion of individuals in which anger was more important for driving them to the voting polls. We tested at a significance level of $\alpha = 0.05$ to see if $L = 0.5$. We used a two-tailed test since we did not have prior evidence that either fear or anger were more relevant.

$$H_0 : L = 0.5$$
$$H_a : L \neq 0.5$$

To conduct the test, we transform the variables and use the following:

```
anger_more_important <- sum(AG_FR_turnout$angry_minus_fear > 0)
trials <- sum(AG_FR_turnout$angry_minus_fear != 0)
binom.test(anger_more_important,
           trials,
           p = 0.5,
           alternative = 't')
```

```
##
##  Exact binomial test
##
## data:  anger_more_important and trials
## number of successes = 7, number of trials = 17, p-value = 0.6291
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.1844370 0.6707528
## sample estimates:
## probability of success
##              0.4117647
```

Here $p-value >> 0.05$. This means that we fail to reject $H_0$ as we have a high chance under $H_0$ to see data as extreme as our sample. Therefore neither anger nor fear levels were greater than the other within this group of individuals, and so we can postulate that neither fear nor anger was more effective at driving voter turnout.

As in 3, we again used Cohen's g to calculate the practical significance:

```
cohen.g <- abs(sum(AG_FR_turnout$angry_minus_fear > 0) / trials - 0.5)
print(cohen.g)
```

```
## [1] 0.08823529
```

The effect correlation $r < 0.1$, which is very small. This is consistent with the fact that for 7/17 individuals, or ~41%, anger was important than fear, which is not too far from 50%. Consequently, there is neither a statistical or practical difference between anger and fear for 2018 voters that did not vote in 2016.

## Question 5: For voters that did not vote in the 2016 Presidential Election, but plan to vote in the 2020 Presidential Election, did more of these individuals change their behavior because they have been satisfied with Trump's performance as president, or dissatisfied with Trump's performance?

We believe it is important to know whether voters are more likely to show up to the polls to vote for a candidate they favor, or to show in order to vote against a candidate they dislike. During a time when political emotions are high, we believe it is fundamental to understand whether positive or negative emotions are more important for predicting voter turnout.

Understanding this question could help campaigns create more targeted ads. If individuals are more likely to show up to support someone they favor, then campaign ads should focus on promoting individuals rather than attacking the opposition. This question could also help advocacy groups devoted to increasing voter turnout. If individuals are more likely to show up to vote against someone they dislike, advocacy groups might focus on reinforcing how not voting could result in a disliked candidate entering office.

To address this question, we will look at individuals who were of voting age in 2016, and select the proportion of these individuals who did not vote in 2016 but anticipate to vote in 2020. The first variable is turnout16, and the second is percent20. Since percent20 is a percentage from 1 to 100, we classify individuals with percent20 > 50 to be ones who will likely vote. percent20 > 50 is the best decision boundary between which we can anticipate whether or not the individual will vote. To evaluate whether or not the voter was satisfied with Trump, we will look at the apppres variable. The variable is categorical and ranges from 1 to 7, from Approve extremely strongly to Disapprove extremely strongly. All individuals that selected "Approve" to some level (1-3) will be classified as favoring Trump, and all individuals that selected "Disapprove" to some level (5-7) will be classified as disfavoring Trump. Those who were neutral will not be considered relevant to the question.

One caveat is that those who say they will vote (especially close to 50%) may not end up voting on Election Day. However, these individuals still have the intention to vote (regardless of perhaps trivial reasons such as a busy work day preventing them from doing so). The same is true for self-reported non-voters. In addition, as in Question 4, the outcome cannot be interpreted as causal. We must stick to the question: among this sample, did more or less individuals approve of Trump's performance?

### Exploratory Data Analysis (EDA) and Hypothesis Test Selection

We filter our set according to the specifications above:

```r
non_voters_2020 <- A %>% filter(turnout16==2, percent20>50, birthyr<=1998) %>% #turnout/age
                      mutate(party = ifelse(pid7x == 1 | pid7x == 2, 'D', #parties
                             ifelse(pid7x == 6 | pid7x == 7, 'R',
                             ifelse(pid7x == 3 | pid7x == 4 | pid7x == 5, 'I', 'N')))) %>%
                      mutate(age = 2018 - birthyr) %>%
                      mutate(approve_trump = ifelse(apppres >= 5, 'N', #approval
                                           ifelse(apppres <=3, 'Y', 'Neither'))) %>%
                      filter(approve_trump != 'Neither') %>% #remove neutral approvals
                      select(age, party, approve_trump)
```

First, we checked the integrity of our data:

```r
summary(non_voters_2020)
```

```
##       age              party            approve_trump
##  Min.   :20.00    Length:122          Length:122
##  1st Qu.:26.00    Class :character    Class :character
```

17

```
##  Median :36.50    Mode  :character    Mode  :character
##  Mean    :38.48
##  3rd Qu.:49.75
##  Max.    :76.00
```

Our age varible has a minimum of 20, meaning all individuals were of voting age in 2016. Our sample size is decent at 122.

Then, we checked whether or not our sample was biased in favour of either Democrats or Republicans:

```r
sum(non_voters_2020$party == 'I')
```

```
## [1] 63
```

```r
sum(non_voters_2020$party == 'R')
```

```
## [1] 32
```

```r
sum(non_voters_2020$party == 'D')
```

```
## [1] 26
```

```r
sum(non_voters_2020$party == 'N')
```

```
## [1] 1
```

The sample had similar amounts of Democratic and Republican. Interestingly, more than half of the sample were Independent voters, and only 1 non-disclosed. Our sample is not biased to either Trump's party, or the party that's the biggest opponent of Trump.

Then, we visualized the opinion distribution of our sample:

```r
Count <- c(sum(non_voters_2020$approve_trump == 'Y'),
           sum(non_voters_2020$approve_trump == 'N'))
Group <- c('Approve Trump', 'Disapprove Trump')
df_counts <- data.frame(Count, Group)

p <- ggplot(df_counts)
p + geom_bar(aes(x=Group, y=Count), stat='identity', position=position_dodge())+
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(x=Group, y=Count, label=Count), vjust=1.6, color="white",
            position = position_dodge(0.9), size=3.5)+
  labs(title="Opinion on Trump for 2020 Anticipated Voters that did not vote in 2016")+
  scale_fill_brewer(palette="Paired")
```

## Opinion on Trump for 2020 Anticipated Voters that did not vote in 2016
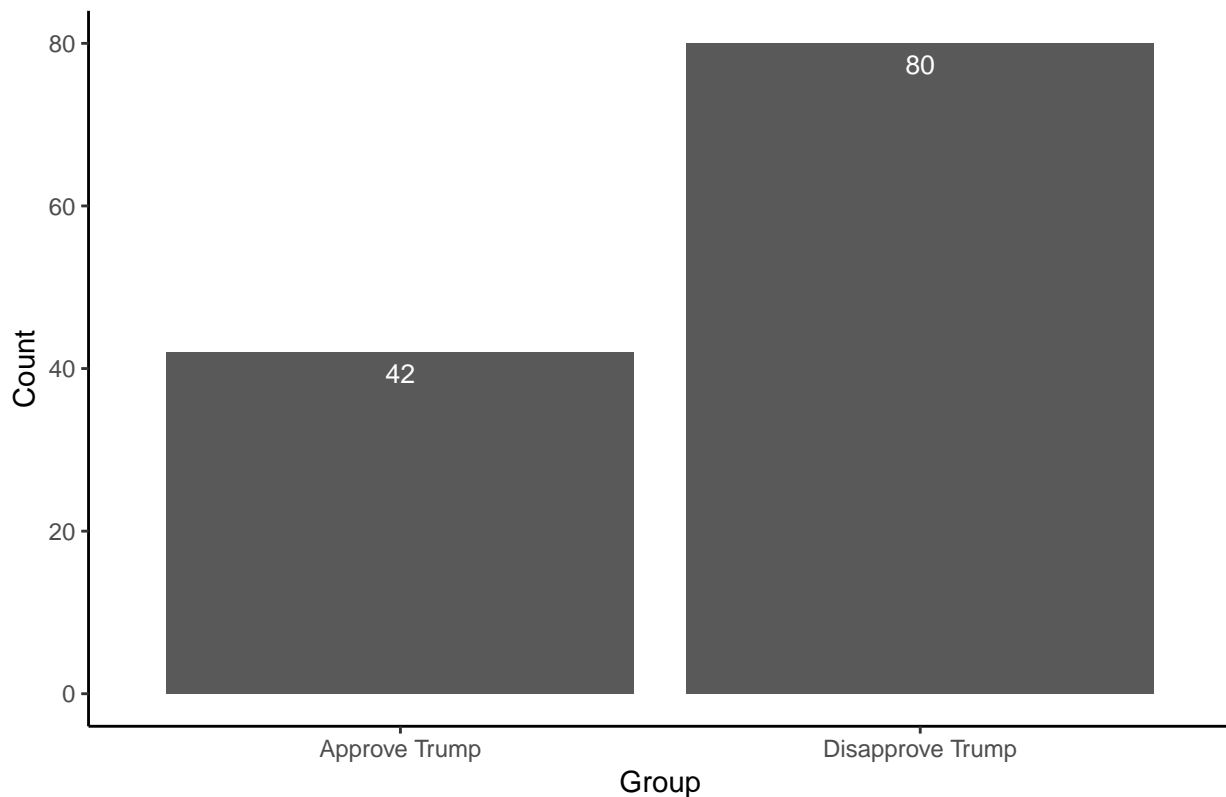


Almost double of these individuals disapprove of Trump to some degree.

We can also look at the opinion distribution of Republicans:

```
print.data.frame(non_voters_2020 %>%
                 filter(party == 'R') %>%
                 group_by(approve_trump) %>% count)
```

```
##    approve_trump  n
## 1              N  7
## 2              Y 25
```

And of Democrats:

```
print.data.frame(non_voters_2020 %>%
                 filter(party == 'D') %>%
                 group_by(approve_trump) %>% count)
```

```
##    approve_trump  n
## 1              N 25
## 2              Y  1
```

Interestingly, the number of Democrats that disapproves of Trump cancels out the number of Republicans that approve of Trump. However, more Republicans disapprove of Trump than Democrats that approve of Trump.

For Independents and non-identified:

```
print.data.frame(non_voters_2020 %>%
                 filter(party == 'I' | party == 'N') %>%
                 group_by(approve_trump) %>% count)
```

```
##   approve_trump  n
## 1             N 48
## 2             Y 16
```

Many more disapprove of Trump.

The most appropriate test here is the binomial test for proportion. Consider the **assumptions** this test makes and the data used in this section.

 1. **Samples are dichotomous and nominal**

After our data treatment, the subjects were classified as either approve or disapprove of Trump with no third category.

 2. **iid sample**

Independence and identical distribution are assumed because respondents do not influence one another and all respondents were drawn from the same population of survey takers.

## Hypothesis Testing

Let R represent the proportion of individuals that did not vote in 2016 but plan to in 2020, and approve of Trump. We will test at a significance level of $\alpha = 0.05$ to see if $R = 0.5$. We used a two-tailed test since we don't have independent, prior evidence, outside of the sample itself that the majority is satisfied or dissatisfied. We do not want, for example, to subject the study to our personal experiences with the media, so we err on the conservative 2-tailed test.

$$H_0 : R = 0.5$$
$$H_a : R \neq 0.5$$

To conduct the test, we transform the variables and use the following:

```
approve_trump <- sum(non_voters_2020$approve_trump == 'Y')
trials <- length(non_voters_2020$approve_trump)
binom.test(approve_trump,
           trials,
           p = 0.5,
           alternative = 't')
```

```
##
##  Exact binomial test
##
## data:  approve_trump and trials
## number of successes = 42, number of trials = 122, p-value =
## 0.0007408
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.2606129 0.4356696
## sample estimates:
## probability of success
##              0.3442623
```

Here, $p-value << 0.05$; therefore we reject $H_0$. We concluded that the proportion of individuals that did not vote in 2016 but most likely will vote in 2020 who favored Trump is not 0.5. Since more individuals disapprove of Trump, we conclude one step further than more individuals statistically disapprove of Trump than those that approve of Trump in this sample.

In order to calculate the practical significance, we will use the metric of Cohen's g as before.

```
cohen.g <- abs(approve_trump / trials - 0.5)
print(cohen.g)
```

## [1] 0.1557377

We get an effect size slightly greater than 0.15, which is medium effect size. Our study was highly statistically significant with medium level of practical significance. This makes sense since almost more than double the number of individuals disapprove of Trump than approve of Trump, equating to about 65.6% of the sample, which appears to be a large majority. This warrants that perhaps further studies on whether positive or negative emotions are more effective at driving voter turnout could be fruitful.