w203: Statistics for Data Science, Lab 1: Comparing Means

Blair Jones, Sean Norris, Mikayla Pugel

Purpose of the Lab and Research Questions

The following document addresses five questions regarding voter attitudes and behavior. Question 5 was chosen by the team.

- Question 1: Do US voters have more respect for the police or for journalists?
- Question 2: Are Republican voters older or younger than Democratic voters??
- Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?
- Question 4: Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?
- Question 5: "Do women from wealthy communities perceive more gender discrimination than women from poor communities?"

Each question is explored in its own section, following the same analysis approach:

- Question Introduction
- Exploratory Data Analysis (EDA)
- Hypothesis Test Definition
- Test Results and Analysis

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

Question Introduction

For the purpose of this estimate we are going to study people who said they voted for a presidential candidate in 2016 as "voters." This is because the question asks about voters and not potential voters. I'm taking this to mean people that have voted before. In general, when we talk about whether or not people are voting we generally are referring to the general, presidential election unless otherwise specified. The column we are going to use for this is the presvote16post column. We chose this variable because it is closer to observed behavior than the other variables in the dataset, including one where respondents were asked "if" they will vote in a future election. We omit this variable because people are generally bad at predicting what they will do in the future.

"Respect" for police and journalists is being measured by the "feeling thermometer" metric in this survey. This is where respondents were asked how they feel on a scale of 0 - 100, with 100 representing the most positive feelings, toward both groups. Respect and feeling may not approximate to the exact same thing. However, if we see large differences between our population's feelings towards both groups it may suggest that a similar gap exists in how respected both groups are by the voting public.

Exploratory Data Analysis (EDA)

Based on the information available, we feel these data meet the criteria of being I.I.D. We specifically refer to YouGov's attempts to match respondents to the American Community Survey from the U.S. Census Bureau (a highly reputable data source) based on various demographics and this passage from the ANES 2018 Pilot Study Codebook, "survey participation is independent of variables measured in the survey." We are aware that YouGov panel pays respondents but a desk review of their practices did not reveal any red flags.

There were no NA's in these data but there were a significant number of "0" responses within the feeling thermometer sample. We believe these to be legitimate responses. This is because the "Questionnaire Specifications" document specifies 0 as a legitimate response and there is a code for those who have skipped the question.

We created a new dataframe that included the case ids, survey weights, 2016 voting data, and police and journalist feeling thermometer data. We then stripped out the people who did not answer the voting question, or were skipped legitimately, because we are only concerned about voters for this question. We then stripped out people who did not respond to the journalist and police feeling questions because we considered that to be a non-response. The filtering left us with 1,707 rows which we felt is still a large enough sample size (e.g., < 30) to generalize to the population. Finally, we multiplied the feeling thermometer data by the weights so all the answers are weighted.

Figure 1.1 Naive Means of Sample Feelings Towards Police and Journalists

```r
#Running summary statistics on each of the important variables
summary(jpdata$police_weighted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   30.31   48.66   61.19   75.92  700.78 
```

```r
summary(jpdata$journos_weighted)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   15.03   37.85   47.04   60.55  578.29 
```

```r
summary(jpdata$diff)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
## -385.429  -13.073    3.005   14.143   40.830  658.733 
```

The summary data presented in Figure 1.1 suggests there is a difference in the mean between who is trusted more: police or journalists. We will test this in the next section.

Figure 1.2 Sample Feelings By Race (Black and White Only)

```
##       n (Weighted) Journalists (Avg. Feeling) Police (Avg. Feeling))
## Total 1707         54.6                       69.18                 
## White 1366         52.75                      71.62                 
## Black 137          71.11                      54.2                  
```

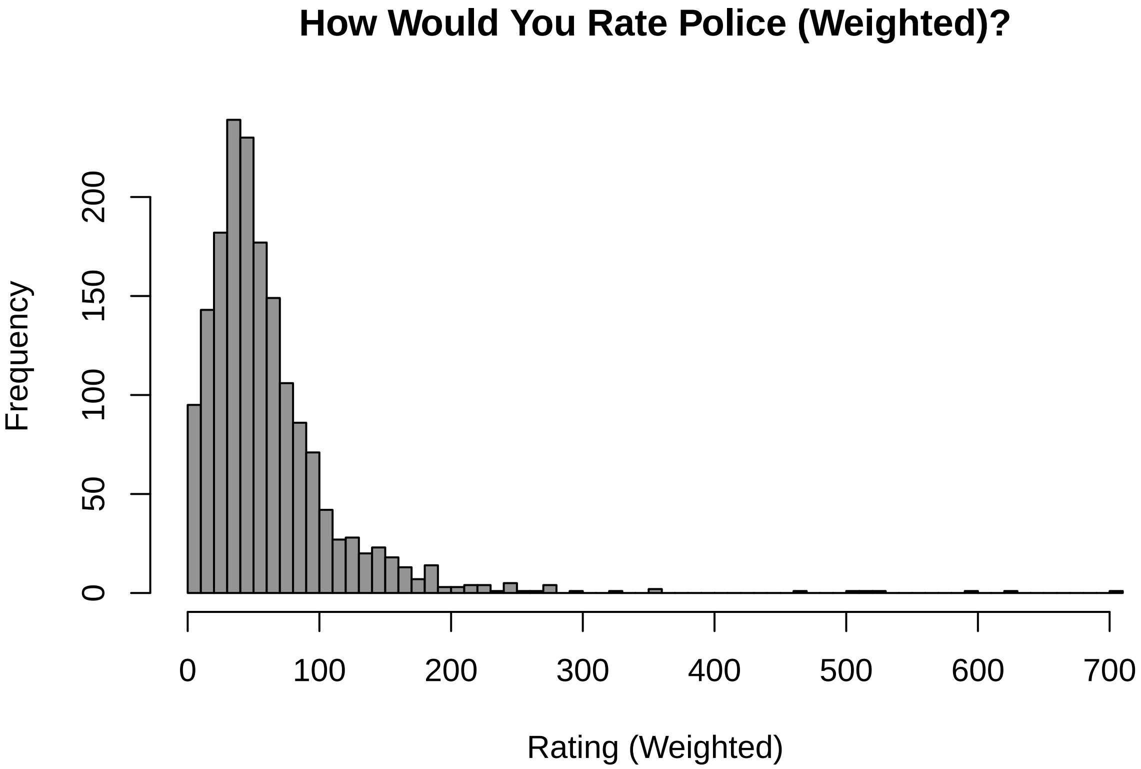
Before we run our test, we first wanted to perform a sanity check on these data as can be seen in Figure 1.2. If the feeling thermometer ratings provided by white people and black people when asked about police were the same or similar, we believe it would call into question the validity of these data. This is because police policy has affected these populations differently. For instance, Pew in 2019 reported that black people make up 12% of the general population and 33% of the prison population.

As you can see in the table above, white people reported an average feeling thermometer score of 71.6 when asked about police and black people recorded an average feeling of 54.2. While informal, we feel this reinforces the validity of these data.

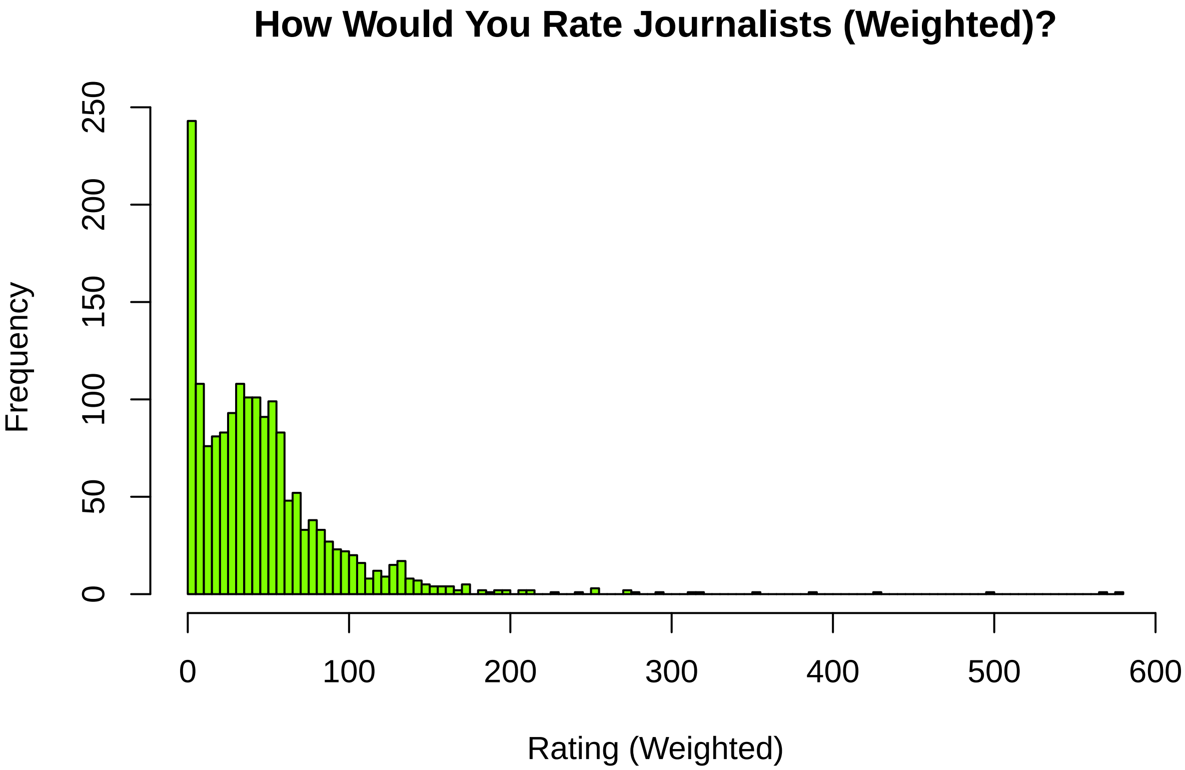
Hypothesis Test Definition

Figure 1.3 Sample Feeling Distribution (Police, Journalists and Difference Between)

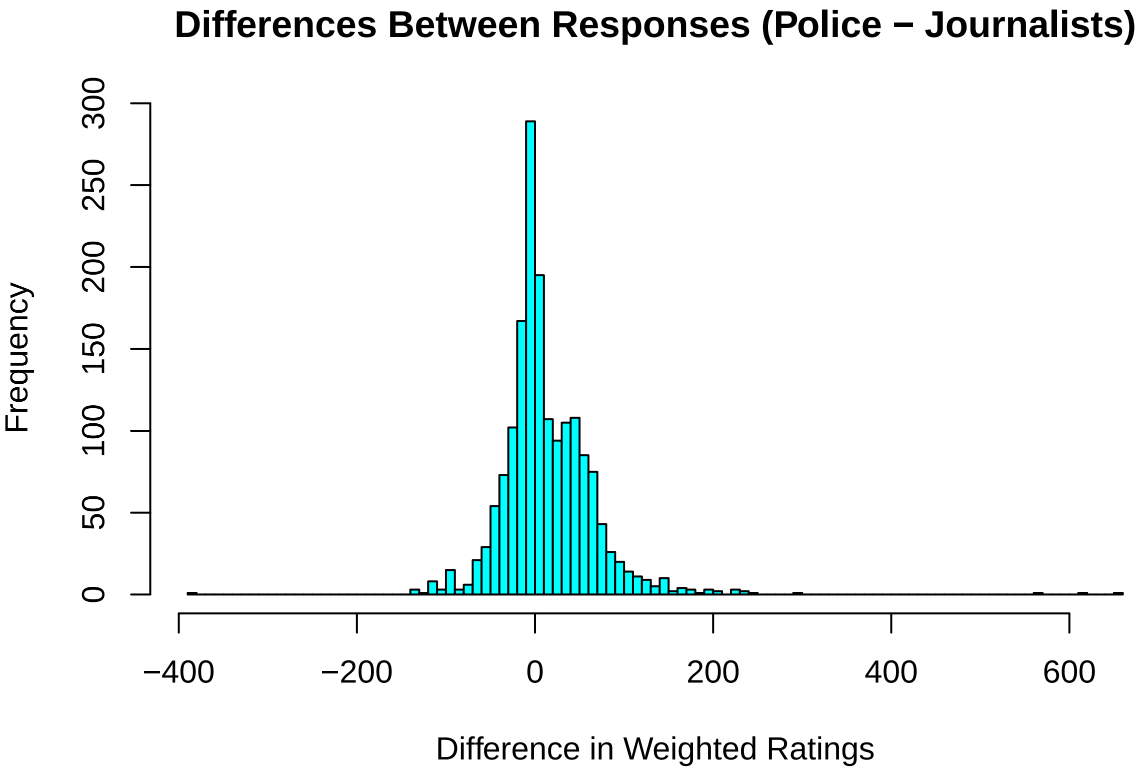
```
hgram_p <- hist(jpdata$police_weighted, main = "How Would You Rate Police (Weighted)?",
                xlab = "Rating (Weighted)", col = 'gray58', breaks = 100)
```



```
hgram_j <- hist(jpdata$journos_weighted, main = "How Would You Rate Journalists (Weighted)?",
                xlab = "Rating (Weighted)", col = 'chartreuse', breaks = 100)
```



```
hgram_d <- hist(jpdata$diff, main = "Differences Between Responses (Police - Journalists)",
                xlab = "Difference in Weighted Ratings", col = ’cyan’, breaks = 100)
```



Once graphed, as in Figure 1.3, you can see the differences between the two ratings is roughly normal but slightly skewed right. However, because our sample is large we can assume the Central Limit Theorem (CLT) which states the distribution of a sample approximates a normal distribution as the sample size approaches infinity.

We cannot use a parametric test because we do not have a metric scale. The feelings people have toward police and journalists is subjective and though these data look parametric they are more like a likert scale. We do have paired data because one person produces an observation for each variable (e.g., police and journalists). Additionally, we have a large (e.g., > 30 observations) and I.I.D sample. This means we can use the Sign Test.

Our null hypothesis will be the probability of a decreasing pair is the same as the probability of an increasing pair. The alternative hypothesis will be these probabilities are not equal.

Test Results and Analysis

Figure 1.5 Paired T-Test Results

```
##
##  One-sample Sign-Test
##
## data:  jpdata$diff
## s = 932, p-value = 7.967e-07
## alternative hypothesis: true median is not equal to 0
## 95 percent confidence interval:
##  1.568217 5.089367
## sample estimates:
## median of x
##    3.005393
##
## Achieved and Interpolated Confidence Intervals:
##
##                   Conf.Level L.E.pt U.E.pt
## Lower Achieved CI     0.9472 1.6365 5.0659
## Interpolated CI       0.9500 1.5682 5.0894
## Upper Achieved CI     0.9529 1.4985 5.1133
```

Our test suggests the difference between the means has high statistical significance because our p-value is very low.

From a practical standpoint, the sample suggests the mean "feeling" toward police is higher than for journalists in this sample of voters. Considering what we've discussed above, that would lead us to believe overall, voters have more respect for police than journalists.

However, given some of the different views between races we've examined above, there may be subgroups within this sample that have very different feelings towards police and journalists.

One last caveat, these data were collected in 2018 and since then, police and journalists have been featured prominently in the news. Should these data be collected again, it will be interesting to see how and if these attitudes may have shifted.

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

Question Introduction

All survey respondents will be considered for this question. We will not filter out candidates who have not recently voted, on the basis that the question is asking about voters who would generally vote for either Republicans or Democrats.

Voter Age

One field, birthyr, captures the age of the voter. The voter's exact birth date is not captured, so it is not possible to calculate a more precise voter age. Birth year (*birthyr*) is subtracted from 2018, the year of the survey, to calculate voter age.

Political Party

The dataset contains a field for political party, *pid7.* The associated document mentions that field *pid7* is from YouGov profile survey data, which were collected on previously-completed questionnaires. *pid7* is not available in this dataset due to a processing error but may be available on a future release. Note that the party ID variable, *pid7x*, is provided.

There are other party-related fields in the dataset (example: *pidlean* and *ideo5*). However these were examined and determined to not be relevant because they are indications of political philosophy but not specific to actual voting.

The vector *pid7x* contains voter responses for these options in the spectrum ranging from Democrat to Independent to Republican

```
  1 Strong Dem
  2 Not very strong Dem
  3 Ind, closer to Dem
  4 Independent
  5 Ind, closer to Rep
  6 Not very strong Rep
  7 Strong Rep
```

We define:

- Democrat: response of 1 or 2
- Republican: response of 6 or 7

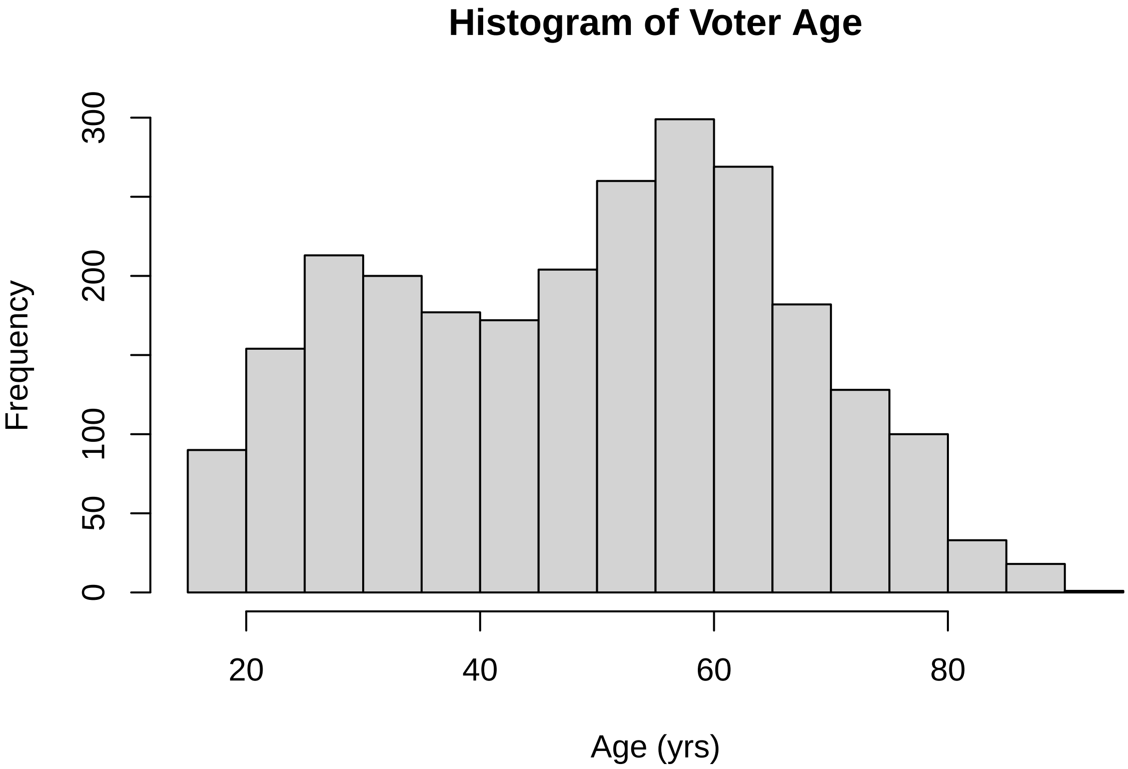
Independents are excluded from this analysis.

Exploratory Data Analysis (EDA)

Voter Age

The *birthyr* vector does not contain any NA values. When calculating age, it does not generate any invalid or unlikely values, such as voter age less than 18 years or greater than 120 years.

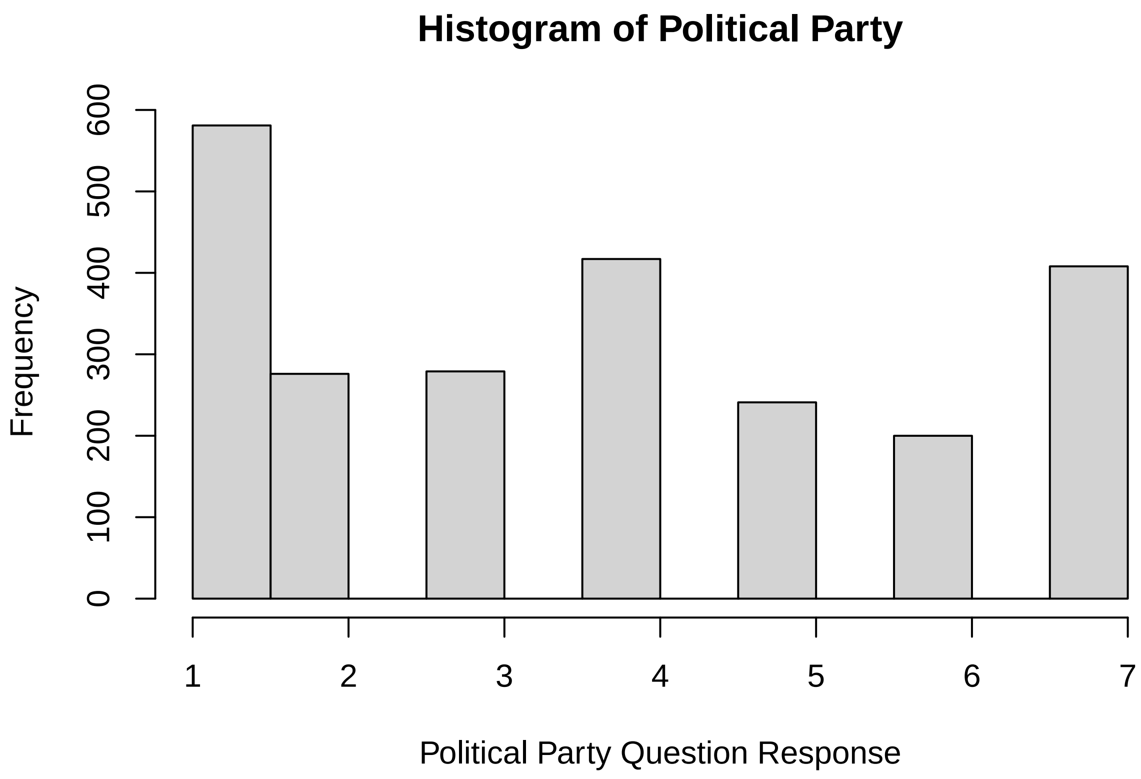
Figure 2.1 Distribution of Voter Age (Democrats and Republicans)



Political Party

The *pid7x* vector does not contain any NA values. It contains the expected range of response values of 1, 2, 3, 4, 5, 6, 7 and a non-response value of -7. There are 98 non-response entries out of the total of 2,500 records. The non-response records will be excluded from the analysis. There are almost twice as many Democrats (i.e., 1,136) as Republicans (i.e., 608) in our sample.

Figure 2.2 Distribution of Political Party Affiliation



Voter age grouped by party

Figure 2.3 Distribution of Age (Democrats Only)

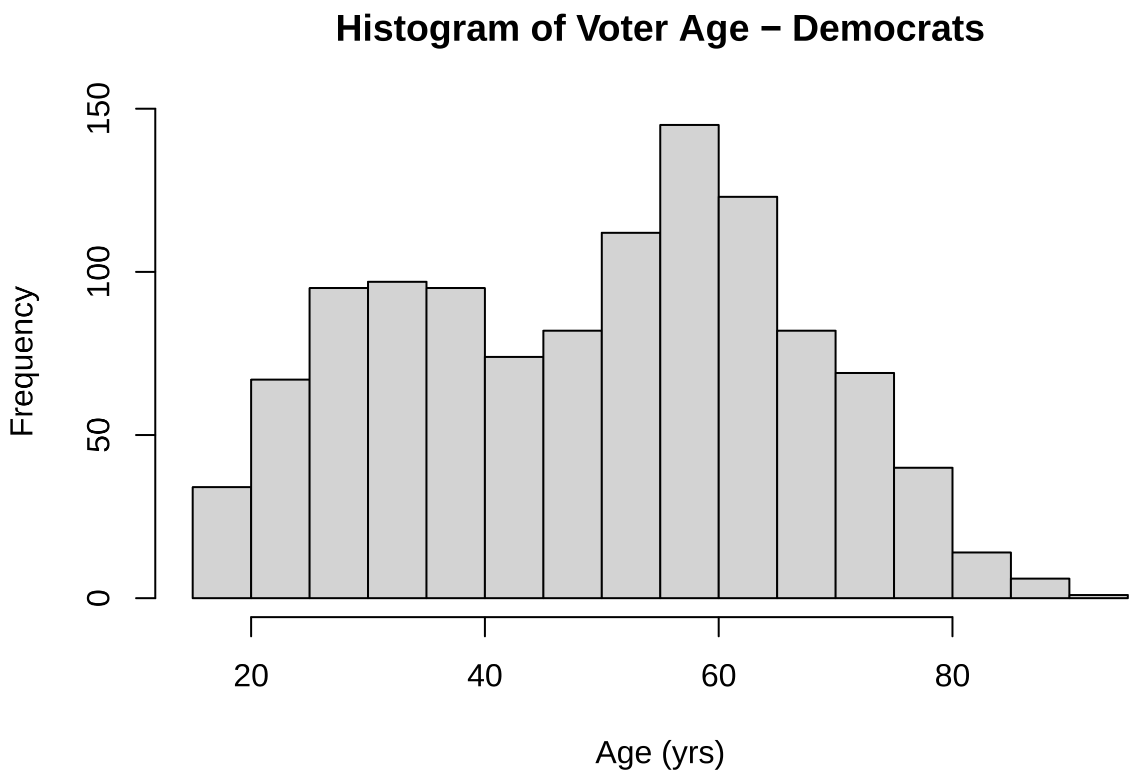
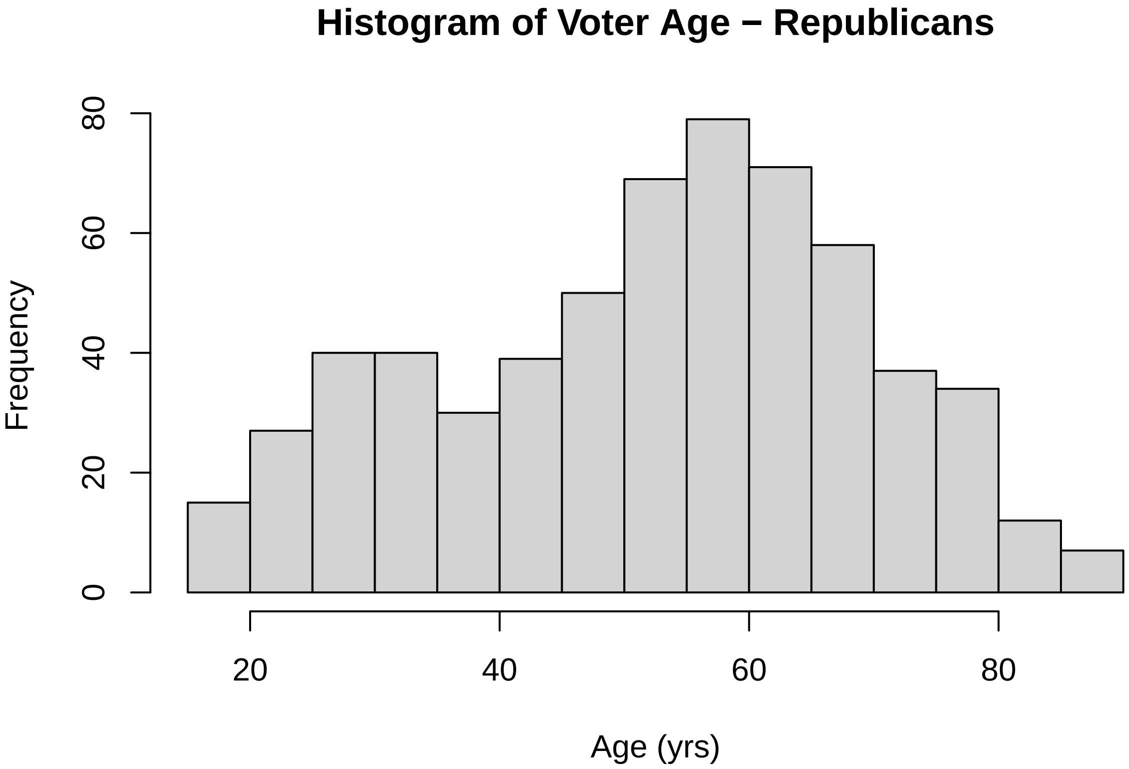


Figure 2.4 Distribution of Age (Republicans Only)



Hypothesis Test Definition

We accept these data are I.I.D. for the reasons outlined in question 1 because these variables are drawn from the same sample.

Voter age is captured once per respondent and is a metric variable. Similarly, party affiliation is captured once per respondent. This leads us to consider an unpaired test.

Voter age does not demonstrate a normal distribution for either the Democrats or Republicans. However, because of the I.I.D. nature of these data, the large sample sizes (e.g., > 500 for each group), we will assume the Central Limit Theorem applies.

Given the analysis above we will proceed with a parametric test. Specifically, we will use an unpaired t-test to determine if there is a significant difference in age in our sample.

Null hypothesis

There is no difference in mean age between Democrats and Republicans.

Test Results and Analysis

Figure 2.5 T-Test Results

```
##
##  Welch Two Sample t-test
##
## data:  republicans$age and democrats$age
## t = 3.851, df = 1244.9, p-value = 0.0001236
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.598264 4.917860
## sample estimates:
## mean of x mean of y
##  52.89803  49.63996
```

Statistical Significance

The test shows that within the sample, there is a statistically significant difference between the mean ages of the two groups. A p-value less than 0.01 is a very strong indication to reject the null hypothesis

Practical Significance

Our sample suggests the mean age for Republicans is 3 years older than for Democrats. This difference, while statistically significant, is not practically significant. This does not demonstrate a generational gap as has been discussed in the media. A more interesting analysis may be to examine mean age over time, seeking to determine if one party is getting older faster than another.

Test conducted on weighted values

This section will re-run the analysis by apply the survey weights.

Voter age grouped by party

Figure 2.6 Voter Age - Democrats (Weighted)

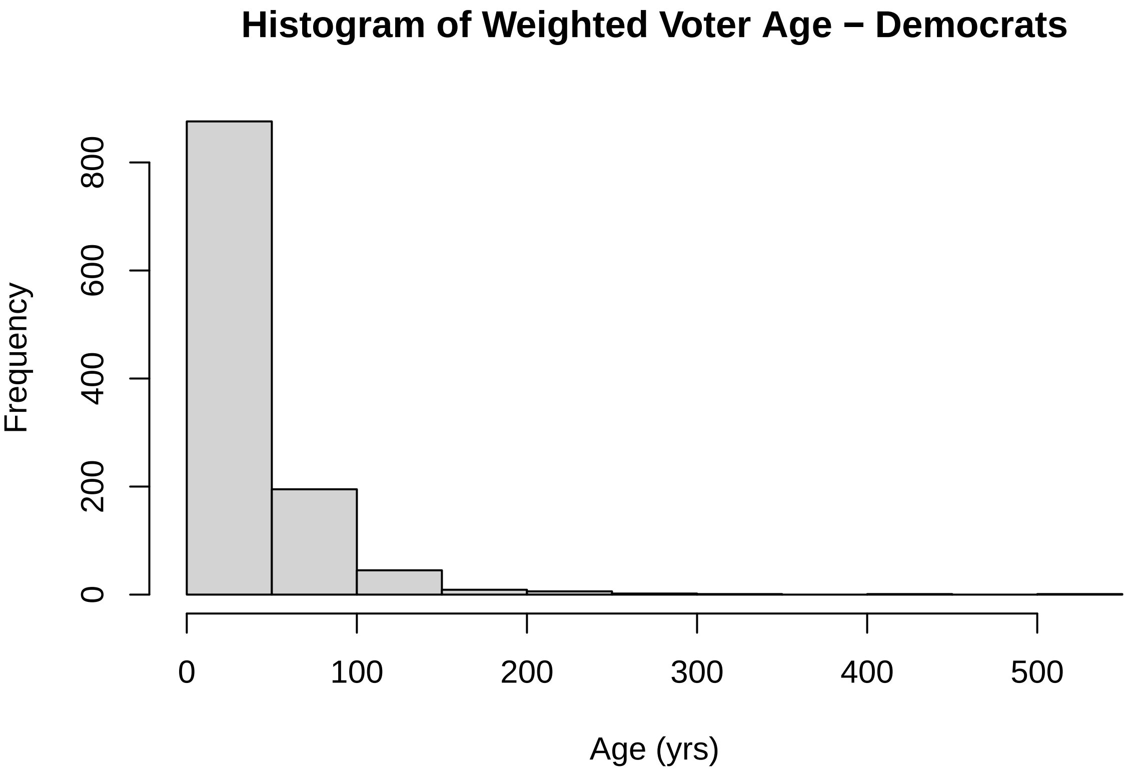


Figure 2.7 Voter Age - Republicans (Weighted)

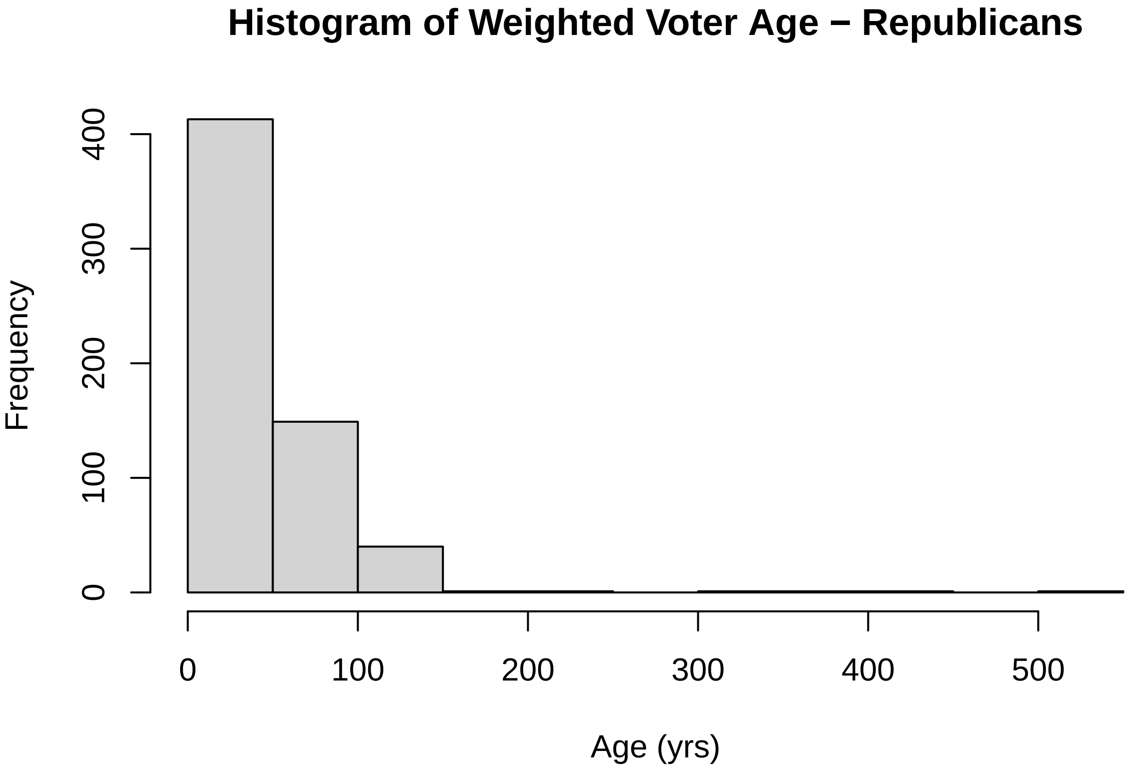


Figure 2.8 T-Test Results (Weighted)

```
##
##  Welch Two Sample t-test
##
## data:  republicans$age_weighted and democrats$age_weighted
## t = 3.1613, df = 1156.6, p-value = 0.001612
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   2.358716 10.076500
## sample estimates:
## mean of x mean of y
##  49.92250  43.70489
```

Once again, we find a statistically significant difference between the 2 groups, and can reject the null hypothesis. The mean weighted age of Republicans is higher than for Democrats.

The weighted difference is larger than the unweighted difference in absolute terms. With a mean weighted difference of approximately 6 years, this could be practically important to users of this analysis. For example, a difference of this magnitude could imply a need to employ different communication strategies in campaigns and other political activities..

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

Question Introduction

We will examine three elements to study this question: party affiliation, voter results from the 2016 election and responses regarding whether the respondent believes federal investigation of Russian election interference is baseless.

For the reasons stated in Question 2, we again used the *pid7x* variable to estimate party affiliation. This time we used the grouped variable indicating "Independent" party affiliation and excluded Republicans and Democrats. As in Question 2, non-responses were removed from this analysis. As in Question 1, respondents who said they did not vote in the 2016 election were also removed. We believe this combination of variables captures the intent behind the "independent voters" portion of the stated question. Meaning, this would enable us to capture the attitudes of people who identify as "Independent" and voted in the 2016 election.

Feelings on baselessness were measured using the Russia16 variable. This variable contains responses to the question, "Do you think the Russian government probably interfered in the 2016 presidential election to try to help Donald Trump win, or do you think this probably did not happen?" Our thinking in using this variable is as follows: if an independent voter indicates they do believe Russia interfered in the election then they would also believe the federal investigation had merit. This assumption will help us relate the question of interest to the null and alternative hypothesis formulated from this survey question answered. As in previous questions, non-responses and skips were removed. The surveyor had the options presented below:

```
[1] Russia probably interfered
[2] This probably did not happen
```

Exploratory Data Analysis (EDA)

Voter Party Data These data were cleaned to remove any non-responses, and to remove all the party affiliations other than the Independent party. Also, any voter who did not vote in the 2016 election was removed for the study. This was completed in order to preserve the true party affiliation, as anyone who voted in the 2016 election was to be sure of their party affiliation, and any non active voter would be less sure.

Russia Interference Data These data were cleaned to remove any non-responses and to remove any skips.

```
ids <- uGov$caseid
weights <- uGov$weight
voter_party <- uGov$pid7x
russia16 <- uGov$russia16
russia16_skp <- uGov$russia16_skp
turnout_16 <- uGov$turnout16

dataq3 <- cbind(ids,weights,voter_party, russia16, russia16_skp, turnout_16)
dataq3 <- data.frame(dataq3)
dataq3 <- filter(dataq3, voter_party != -7)
dataq3 <- filter(dataq3, voter_party != 7)
dataq3 <- filter(dataq3, voter_party != 6)
dataq3 <- filter(dataq3, voter_party != 1)
dataq3 <- filter(dataq3, voter_party != 2)
```

```
dataq3 <- filter(dataq3, turnout_16 != 2)
dataq3 <- filter(dataq3, turnout_16 != 3)
dataq3 <- filter(dataq3, russia16 != -7)
dataq3 <- filter(dataq3, russia16_skp != 1)
dataq3 <- filter(dataq3, russia16_skp != 2)

nrow(dataq3)
```

```
## [1] 595
```

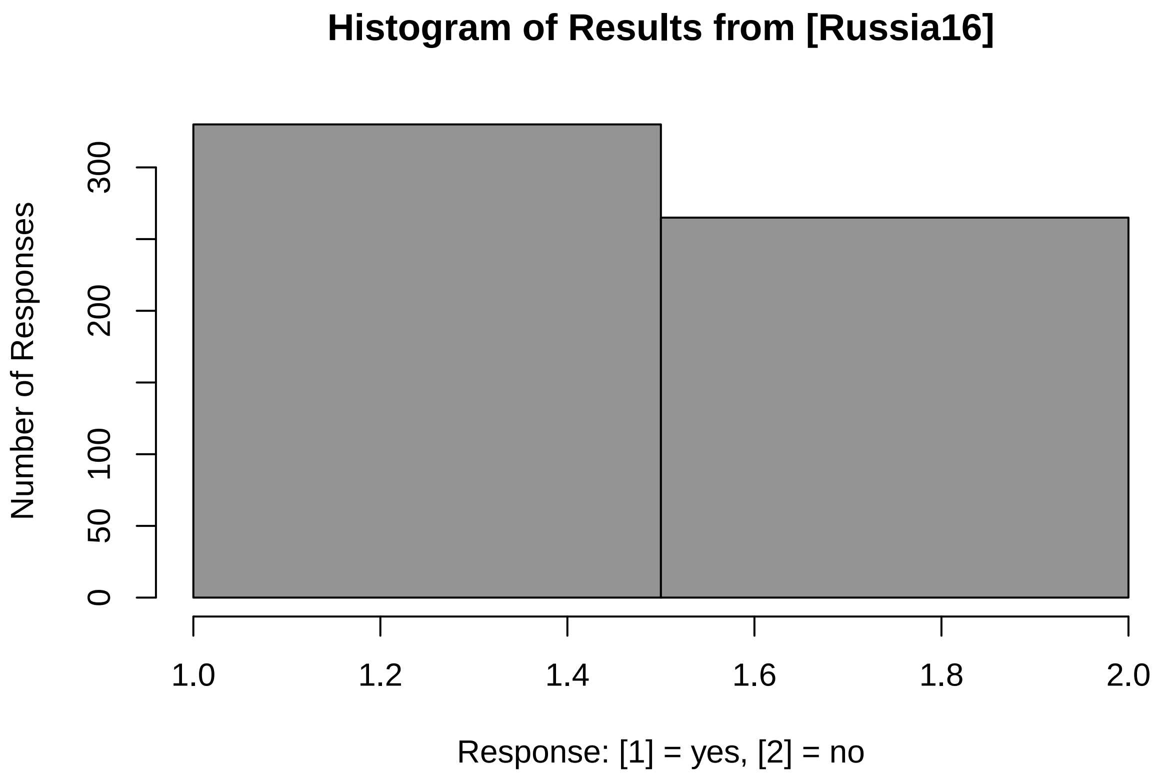
```
sum(dataq3$weights)
```

```
## [1] 487.9974
```

Data Visualization and Exploration These data were explored by observing the summary and histogram of the results from the vector [russia16]. Based on the histogram, it can easily be seen that more independent voters believe that Russia probably interfered. The mean from the summary data also confirms this belief. However, it is only a slight majority, and therefore may not be significant to declare the research question true.

Figure 3.1 Distribution of Russian Interference Results

```
hgram <- hist(dataq3$russia16, col = ’gray58’, breaks = 2,
              main = ("Histogram of Results from [Russia16]"),
              xlab = ("Response: [1] = yes, [2] = no"), ylab = ("Number of Responses"))
```



```r
summary(dataq3$russia16)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   1.000   1.000   1.445   2.000   2.000
```

Hypothesis Test Definition

Based on the exploratory data analysis, the study to best fit these data is a binomial test. This is because the survey question being studied is a yes or no question, which is assumed to be nominal as either "success" or "failure". There is also only one group in question, which is the independent voters. Based on this criteria, the binomial test is the correct option for this hypothesis testing. This test conducted will be a two-tailed test and based on the p-value from the results, the null hypothesis will either be rejected or failed to be rejected. This test will be calculated with a 95% confidence level.

Null Hypothesis (Ho): The relative frequency independent voters believed Russia interfered = 0.5

Alternative Hypothesis (Ha): The relative frequency independent voters believed Russia interfered is not 0.5

Assumptions for this test: 1. Items are dichotomous and nominal; in this case we assume that yes and no are equivalent to 0 and 1. 2. The sample size is significantly less than the population size 3. The sample is a fair representation of the population - this is true with the use of weights to better model the population 4. Sample items are independent

Test Results and Analysis

For the binomial test, the values used in the test are, the sum of the weighted observations that independent voters believe that Russia interfered, the sum of all the weighted values, and the null hypothesis probability. The sum of the weighted values of the vector [russia16] were used in order to model the sample more like the population. These variables and the binomial test are calculated below.

```r
sum_weights_yes <- round(sum(dataq3$weights[which(dataq3$russia16 == 1)]))
total_weights <- round(sum(dataq3$weights))

binom.test(sum_weights_yes, total_weights, 0.5)
```

```
## 
##  Exact binomial test
## 
## data:  sum_weights_yes and total_weights
## number of successes = 255, number of trials = 488, p-value = 0.3418
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4771781 0.5676292
## sample estimates:
## probability of success 
##               0.522541
```

Based on the results from the binomial test, we would fail to reject the null hypothesis. This is because the p-value from the test is sufficiently large. The p-value is the probability of obtaining results at least as extreme as the observed results in the test. Since the value is large, our results did not deviate from the null hypothesis by a significant amount. This test was completed with a 95% confidence level.

Based on the results from the study, we cannot say with statistical confidence that the majority of independent voters believe that the federal investigations of Russian election interference are baseless.

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

Question Introduction

Overall, voter turnout decreased from 2016 to 2018. Therefore our interpretation of this question will be to consider: which was more effective, fear or anger, in getting people who did not vote in 16 to vote in 18. This would make the overall net increase in the population less relevant.

To do this, we will study the subset of people who voted in 2018 and did not vote in 2016. We will then use the data available to determine if these people were more scared or afraid in 2018.

Voter Turnout

There are vectors in this dataset to represent whether a survey respondent voted in 2016 and again in 2018. The data for 2018 are more granular to examine the method by which the vote was cast, but otherwise the two vectors are comparable across 3 values:

- Did vote
- Did not vote
- Not sure

The two vectors of interest are *turnout16* and *turnout18*.

Anger / Fear

There are several questions in the dataset which address anger and fear, grouped into 3 topics:

- General
- President
- Immigration

The question posed is not specific to the president, nor to immigration. Therefore we will focus our analysis on the general topic.

The two vectors of interest are *geangry* and *geafraid*.

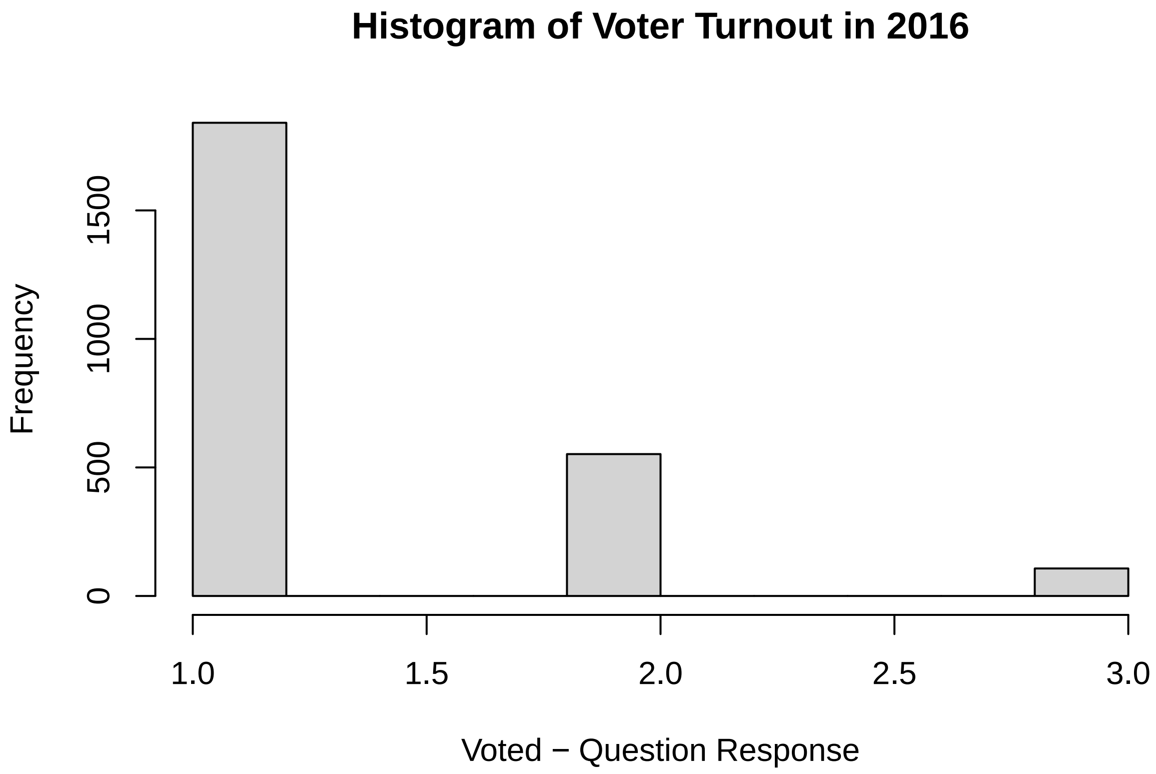
The response values for both these vectors are:

- 1 Not at all
- 2 A little
- 3 Somewhat
- 4 Very
- 5 Extremely

Exploratory Data Analysis (EDA)

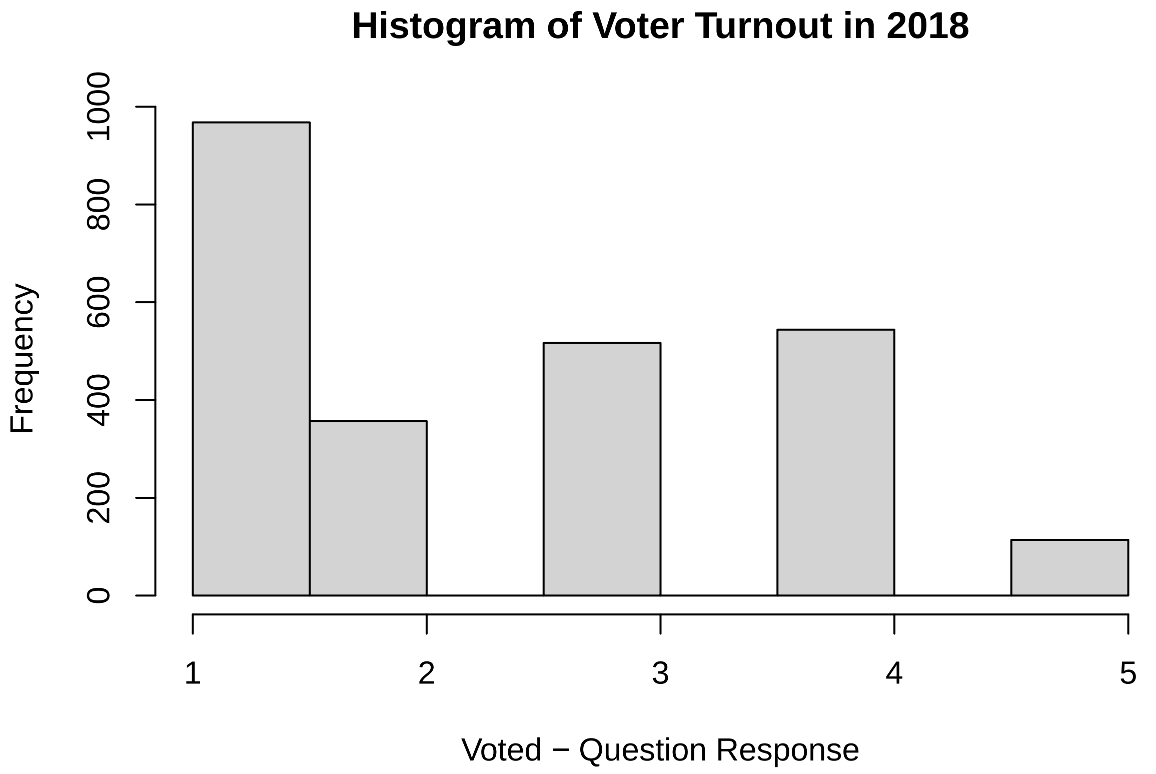
Turnout

Figure 4.1 - Voter Turnout 2016



Values equal: 1 - Did vote 2 - Did not vote 3 - Unsure

Figure 4.2 - Voter Turnout 2018



Values equal: 1, 2, 3 - Did vote 4 - Did not vote 5 - Unsure

Anger / Fear

There are no NA values in any of the records for the *geangry* and *geafraid* vectors. The number of skips/non-responses is very low at 3 and 6 respectively.

Figure 4.3 - Distribution of General Voter Anger 2018

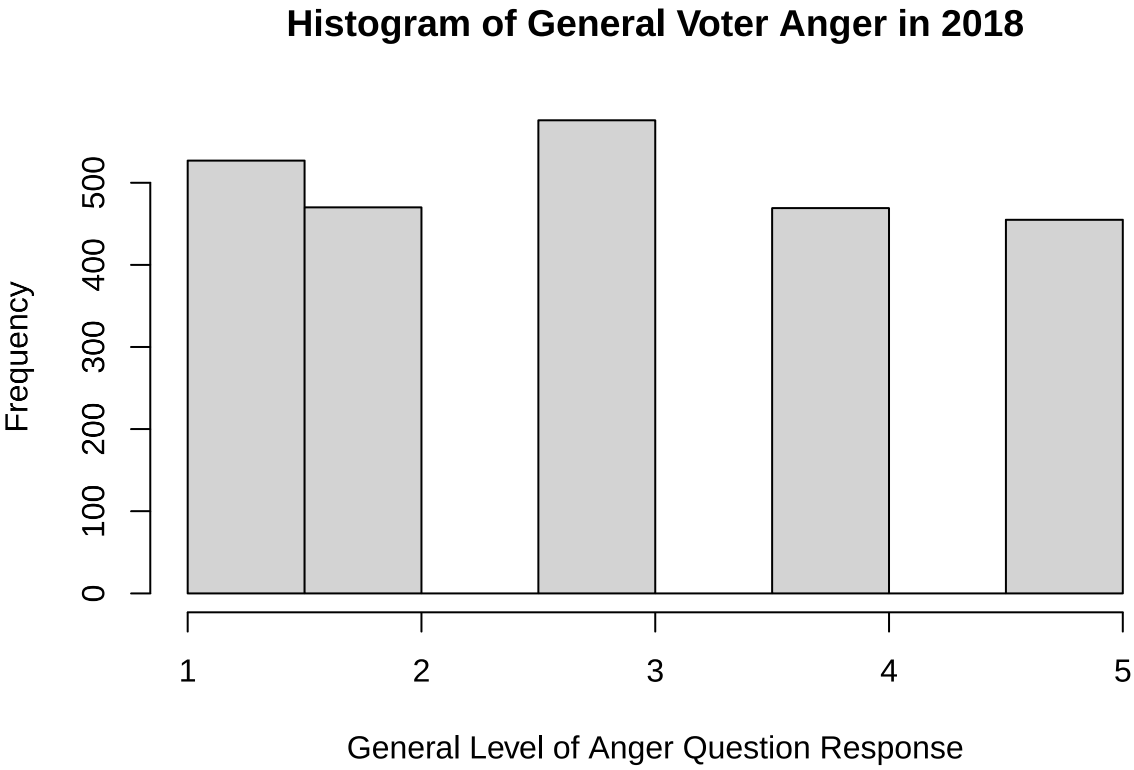
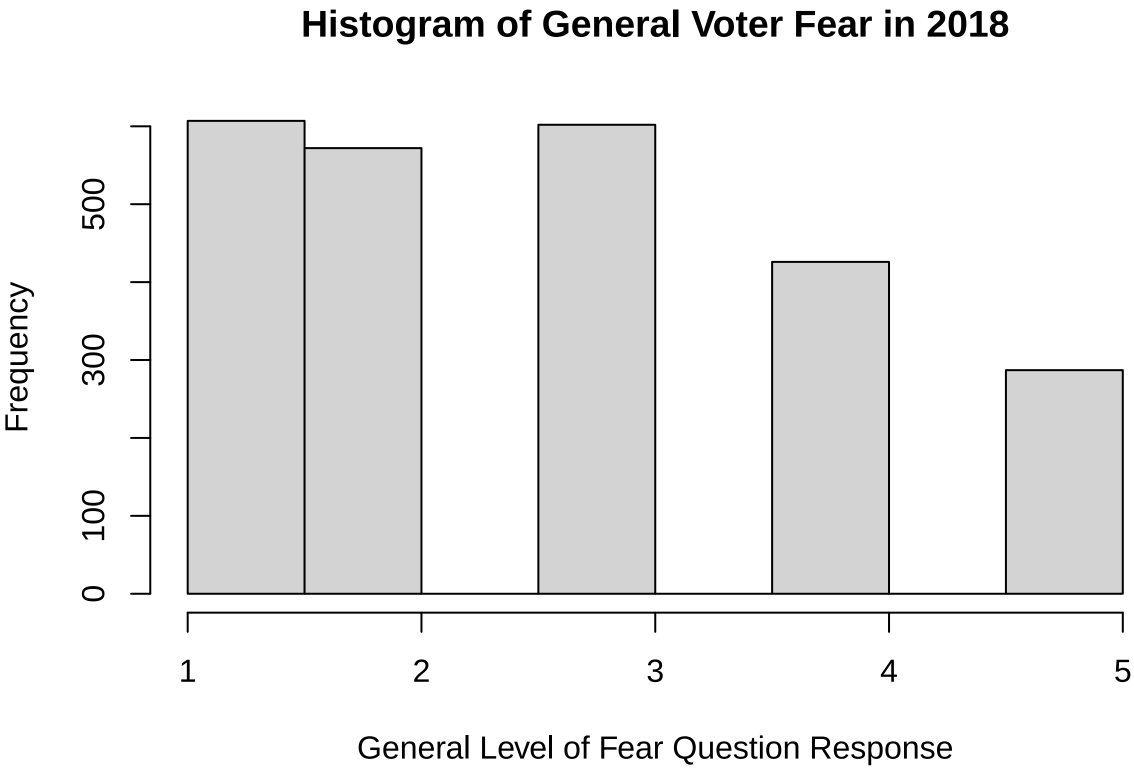


Figure 4.4 - Distribution of General Voter Fear 2018



Hypothesis Test Definition

The turnout response for 2016 and 2018 are two measurements from the same person, therefore we will apply a paired test approach.

The angry/afraid responses are ordinal in nature. The differences between angry and afraid responses are not symmetric.

None of these data exhibit a normal distribution, however given the large sample size the CLT applies.

Therefore, given these data are paired, ordinal and asymmetric, a Sign Test will be used to test the null hypothesis.

Null Hypothesis

There is no difference between the level of Anger and Fear amongst voters who voted in 2018 who did not vote in 2016.

Test Approach

For the test we will retain the voters who voted in 2018 and did not vote in 2016. There are 552 records in the resulting dataset.

We then performed a comparison of the vectors *geangry* and *geafraid*, which are likert scale variables. This approach compares responses for the same person across the two questions so that we have paired data. This addressesed the possibility that two people may not rate the levels in the scale the same way. This would make it difficult to compare one person's level of anger to another person's level of fear. Therefore, we compared each individual's responses to the two questions to determine the relative outcome for that person.

This comparison is stored as a new vector for the result of the comparison:

- $geangry > geafraid = \text{A}$ (more angry)

- $\mathit{geangry} < \mathit{geafraid} = \mathrm{F}$ (more afraid)
- $\mathit{geangry} = \mathit{geafraid} = \mathrm{N}$ (equally angry and afraid)

This new vector now gives us an indication, for each person, whether they felt more anger than fear, more fear than anger, or felt equally about both emotions. From that information, we can conduct a Bernoulli test, excluding those who felt no difference. This will provide us with a view of whether there is a statistically significant difference in the expressed level of anger versus fear for additional voters in 2018.

Test Results and Analysis

Figure 4.5 - Sign Test on Anger versus Fear

```
##
##  Exact binomial test
##
## data:  c(more_angry, more_afraid)
## number of successes = 164, number of trials = 319, p-value = 0.6543
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4577821 0.5701668
## sample estimates:
## probability of success
##              0.5141066
```

The high p-value here suggests we should fail to reject the null hypothesis that there is no difference between fear and anger in the new voter population. In figure 4.7 we examine a raw summary of these data.

Figure 4.6 - Summary of Anger Vs. Fear Results

```r
cat('Voters more angry than afraid:', more_angry, '\nVoters equally angry and afraid:',
    neutral, '\nVoters less angry than afraid:', more_afraid, '\n')
```

```
## Voters more angry than afraid: 164
## Voters equally angry and afraid: 233
## Voters less angry than afraid: 155
```

The sample shows slightly more new voters in 2018 signaled a stronger feeling of anger than of fear (geangry > geafraid). However our test showed this is not statistically significant.In addition, we believe these results are also not practically significant because the sums are too close and a plurality of people fall into the middle group of being equally afraid and angry.

Question 5: "Do women from wealthy communities perceive more gender discrimination than women from poor communities?"

Question Introduction

The question of study for this experimental design is "Do women from wealthy communities perceive more gender discrimination than women from poor communities?"

Wealthy and Poor Women This variable is conceptualized from the vectors [gender] and [faminc_new]. The variable will be split into two groups. The women will be grouped based on their response to the survey question “Thinking back over the last year, what was your family’s annual income?” The responses from these data are comprised of 97 different ranges in income. The wealthy women will comprise the upper half of responses and the poor women will comprise the lower half of the responses. For this analysis, we did not consider the poverty line, but rather we are assuming relative wealth. Therefore, poor women are considered the lower half of the family income and the wealthy are the upper half. Also, we have assumed that if a family has a high income, that they also reside in a wealthy community and vice versa for the poor communities. This assumption will help us answer the research question based on the survey results.

Gender Discrimination Gender Discrimination will be measured from the vector [disc_selfsex]. For pusposes of this study we will assume women who answered this question will be truthful in their experiences with gender discrimination. This will allow us to use the responses to the question to calculate an answer to our research question. This question asked, “How much discrimination have you personally expeierenced because of your sex or gender?” The possible responses are shown below:

```
[1] None
[2] A little
[3] A moderate amount
[4] A lot
[5] A great deal
```

Exploratory Data Analysis (EDA)

We isolated the variables needed to complete our analysis. We believe using the weighted responses will enable us to more accurately measure the sentiment in the broader population.

```r
ids <- uGov$caseid
weights <- uGov$weight
gender <- uGov$gender
income <- uGov$faminc_new
gender_disc <- uGov$disc_selfsex
gender_disc_skp <- uGov$disc_selfsex_skp
```

Gender and Income Data

The gender data does not contain any NA values or non-response values. The men were removed from the dataset by filtering using the gender variable.

Gender Discrimination data

The skips from the [disc_selfsex] question as well as the non-responses were removed from these data.

Data Visualization and Exploration A summary of these data is provided below. Based on these data, the median income will be used to separate the poor women and wealthy women into equal groups for the statistical test. The histrogram of gender discrimination in women is shown to be skewed right.

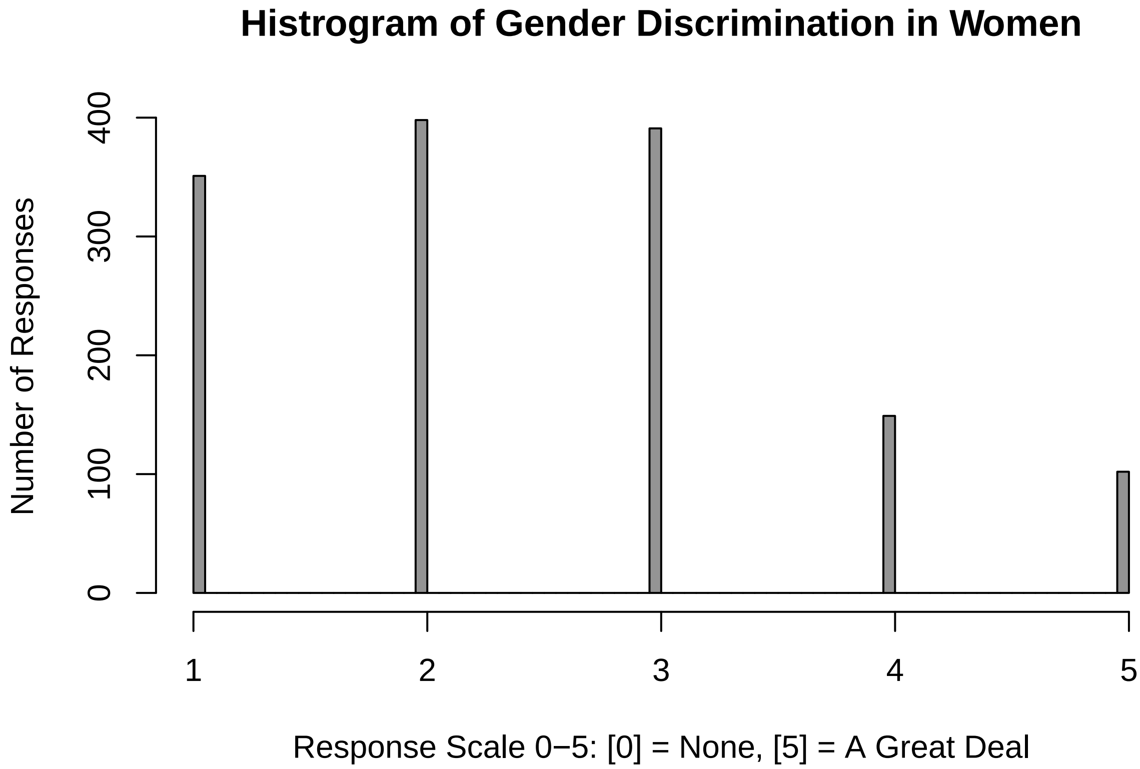
```r
dataq5 <- cbind(ids,weights, gender, gender_disc, gender_disc_skp, income)
dataq5<- data.frame(dataq5)
dataq5 <- filter(dataq5, gender != 1)
dataq5 <- filter(dataq5, gender_disc != -7)
dataq5 <- filter(dataq5, gender_disc_skp != 1)
dataq5 <- filter(dataq5, gender_disc_skp != 2)
summary(income)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    6.00   17.44   10.00   97.00
```

```
summary(gender_disc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -7.000   1.000   2.000   2.143   3.000   5.000
```

Figure 5.1 - Histogram of Gender Discrimination



Hypothesis Test Definition

Based on the exploratory data analysis, the best test for these data is a wilcoxon rank-sum test. In this case, we have assumed these data are unpaired and ordinal. The test will a two sided because we want to see whether wealthy women experience more or less discrimination when compared to poor women. Also, for a wilcoxon rank-sum test the data can be non-normal or skewed. This test will be calculated with a 95% confidence level.

Null Hypothesis: (Ho): The probability a woman from a poor community experiencing discrimination because of her gender is equal to the probability of a woman from a wealthy community experiencing discrimination because of her gender.

Alternative Hypothesis (Ha): The probability that a woman from a poor community has experienced discrimination because of her gender is not equal to the probability of a woman from a wealthy community experiencing discrimination because of her gender.

The assumptions for the test: 1. The variable is ordinal - this is true based on the selection options in the question 2. Sample points are IID - this sample was completed in an independent and identically distributed survey 3. Independent groups - poor women and wealthy women are assumed to be independent as discrimination towards a poor woman would not cause or not cause a wealthy woman to experience discrimination.

Test Results and Analysis

The variables for gender discrimination of poor woman and gender discrimination of wealthy women is calculated below based on the family income they reported. Based on the median reported income the poor women are defined as having a family income reported of $0-$59,000 and wealthy women are defined as having a family income reported of $60,000+. From these variables the wilcoxon rank sum test was conducted.

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  poor_gen_disc and wealthy_gen_disc
## W = 238893, p-value = 0.9874
## alternative hypothesis: true location shift is not equal to 0
```

Based on the results from the wilcoxon rank sum test, we would fail to reject the null hypothesis. This is because the p-value is significantly large. This test was conducted with a significance level of 95%.

Since we fail to reject the null hypothesis, we do not accept that the alternative hypothesis to be true, in that a woman from poor community has experienced more or less gender discrimination than a woman from a wealthy community. Additional testing is needed to support this but from a practical perspective this could be used as evidence the discrimination faced by these women is gender based and not income based.