Lab 1: Comparing Means

Sristhi Mehra, David Djambazov, and Andi Morey Peterson

10/17/2020

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States. While its flagship survey occurs every four years at the time of each presidential election, ANES also conducts pilot studies midway between these elections. We will be using this data to ask five (5) questions about the respondents:

1. Do US voters have more respect for the police or for journalists?
2. Are Republican voters older or younger than Democratic voters?
3. Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?
4. Was anger or fear more effective at driving increases in voter turnout from 2016 to 2018?
5. (Student Choice) Do Hillary Clinton voters and Trump voters view the US income gap differently?

General Study Comments (applicable for all questions)

Since all questions draw from the same study sample, it is useful to state comments here that we can refer to for all questions. For almost any test we run, we will need to determine if the data is i.i.d. and if the respondents to the survey accurately represent the average U.S. voter that we can generalize accordingly.

Independence - Unless multiple people in the same locale, same family, or same household, for example, are used in the way the survey was conducted, we can safely assume each respondent is independent from another.

Identically Distributed - Once on person has taken the survey, they cannot take the survey again; so the distribution for the next "draw" is changed. But the change in the population distribution for the next raw is so small, we can safely ignore this effect.

Generalizability - Because this is a modern, paid, opt-in survey, the sample data will only include individuals who have the propensity or financial motivation to complete the survey. However, the financial impact is small, 21-50 cents for this 30 minute survey (see the ANES User Guide Code Book). In addition, the survey provided weights in which the survey recommends to use when making inferences to the target population of U.S. adult citizens.

Given these, we can assume the iid assumption is valid and for results we worry about generalizability, we can use the weights to help us on questions in which we are concerned about generalizing to the population. (One concern/gap worth mentioning – the data did not account for people who are ineligible to vote due to a felony. Perhaps the survey rid of these samples prior to publishing but it is not mentioned in the materials. These people didn't vote not because of anger or fear, but because they were unable to do so).

Confidence Interval - All tests will use a 95% confidence as standard practice.

Voting Population - Nearly all the question asks about voters. The survey provides a few sample questions about the respondent to determine if that respondent was actually a voter. Variables *turnout*18 and *turnout*18*ns* can be used to determine if they were a voter. For our analysis, we will only consider the population that have value 1, 2, or 3 for the variable *turnout*18 to take the conservative approach in considering US voters.

Let's look at how many observations are in those categories:

```r
paste("Number of NA in turnout18: ", length(which(is.na(A$turnout18))))
```

```
## [1] "Number of NA in turnout18:  0"
```

```r
paste("Number of NA in turnout18ns: ", length(which(is.na(A$turnout18ns))))
```

```
## [1] "Number of NA in turnout18ns:  0"
```

```r
paste("Definitely Voted: ", sum(A$turnout18 <= 3))
```

```
## [1] "Definitely Voted:  1842"
```

```r
paste("Not completely sure or Probably did vote: ", sum(A$turnout18 == 5 & A$turnout18ns == 1))
```

```
## [1] "Not completely sure or Probably did vote:  18"
```

```r
paste("Number of Duplicate Voters for the main dataset: ", length(which(duplicated(A))))
```

```
## [1] "Number of Duplicate Voters for the main dataset:  0"
```

```r
A$voted2018 <- ifelse((A$turnout18<=3), 1, 0)
```

About 1% not sure or "probably" voted. We argue that because is a fairly small number and being uncertain about having voted in an election that has just taken place can be reasonably viewed as grounds for exclusion from the population of US 2018 voters. We will use this population for all questions in this lab because all questions ask specifically about "voters".

We also did a check to confirm there are no duplicate rows in the set of voted18. If there are duplicates, we will need to filter them out, in this case where were none so we will not need to worry about this for the rest of the lab.

Research Questions

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

Introducing the topic

Concept: To understand if US voters have different amount of respect police and jounralists. What we are trying to measure here, is a subjective variable.

Operationalization: From a list of data points collected in the ANES 2018 Pilot Study, two are people's ratings of the police and of journalists in variables 'ftpolice' and 'ftjournal' respectively. The questions to collect these are "How would you rate the police?" and "How would you rate journalists?". These are the closest data points available to our question since they describe to what the sample population feels about the police and journalists respectively. The answers to these questions are collected on a rating scale between 0 to 100.

To get the subset of voters, we will use the voted2018 as descibed in the introduction.

Concerns or Gaps:

1. The data collected is for rating police and journalists in general, not particularly based on respect. We have to be cautious while interpreting our conclusions since we are inferring respect from a total rating.

2. One person's interpretation of the rating scale can be very different from another's, so if one person could think of brilliant as 90, another could think of brilliant as 94. This must also be kept in mind as caution while reading the conclusion.

Exploratory data analysis (EDA) of the relevant variables

Variables: ftpolice, ftjournal

Types: Both ftpolice and ftjournal are Ordinal type variables. Since the variables are measured on a rating scale, they have defined categories that have an order. So can apply operators like <, >, as well as =. We cannot however, apply any arithmetic operaters (like +, -, /) to such variables.

Number of Entries

```r
library(tidyverse)
usVoters <- A  %>% filter(A$voted2018  == 1)
paste("Number of Entries for ftpolice:", length(usVoters$ftpolice))
```

```
## [1] "Number of Entries for ftpolice: 1842"
```

```r
paste("Summary for ftpolice: ")
```

```
## [1] "Summary for ftpolice: "
```

```r
summary(usVoters$ftpolice)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   50.00   74.50   68.49   91.00  100.00
```

```r
paste("Number of Entries for ftjournal:", length(usVoters$ftjournal))
```

```
## [1] "Number of Entries for ftjournal: 1842"
```

```r
paste("Summary for ftjournal: ")
```

```
## [1] "Summary for ftjournal: "
```

```r
summary(usVoters$ftjournal)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   -7.00   20.00   59.00   54.42   88.00  100.00
```

There are no Null Values for ftpolice and ftjournal

```r
paste("Number of Null Values for ftpolice:", length(which(is.na(usVoters$ftpolice))))
```

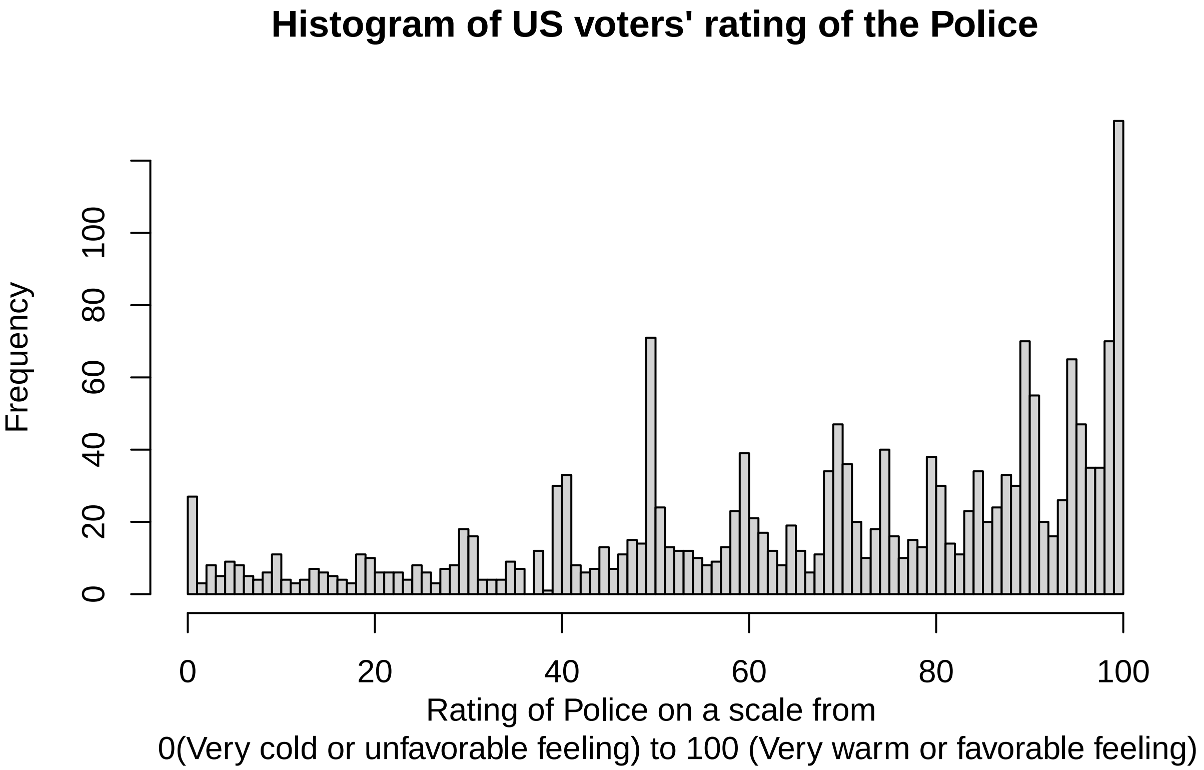
```
## [1] "Number of Null Values for ftpolice: 0"
```

```r
paste("Number of Null Values for ftjournal:", length(which(is.na(usVoters$ftjournal))))
```

```
## [1] "Number of Null Values for ftjournal: 0"
```

We will draw a histogram of both ftpolice and ftjournal to observe the distribution and any outliers

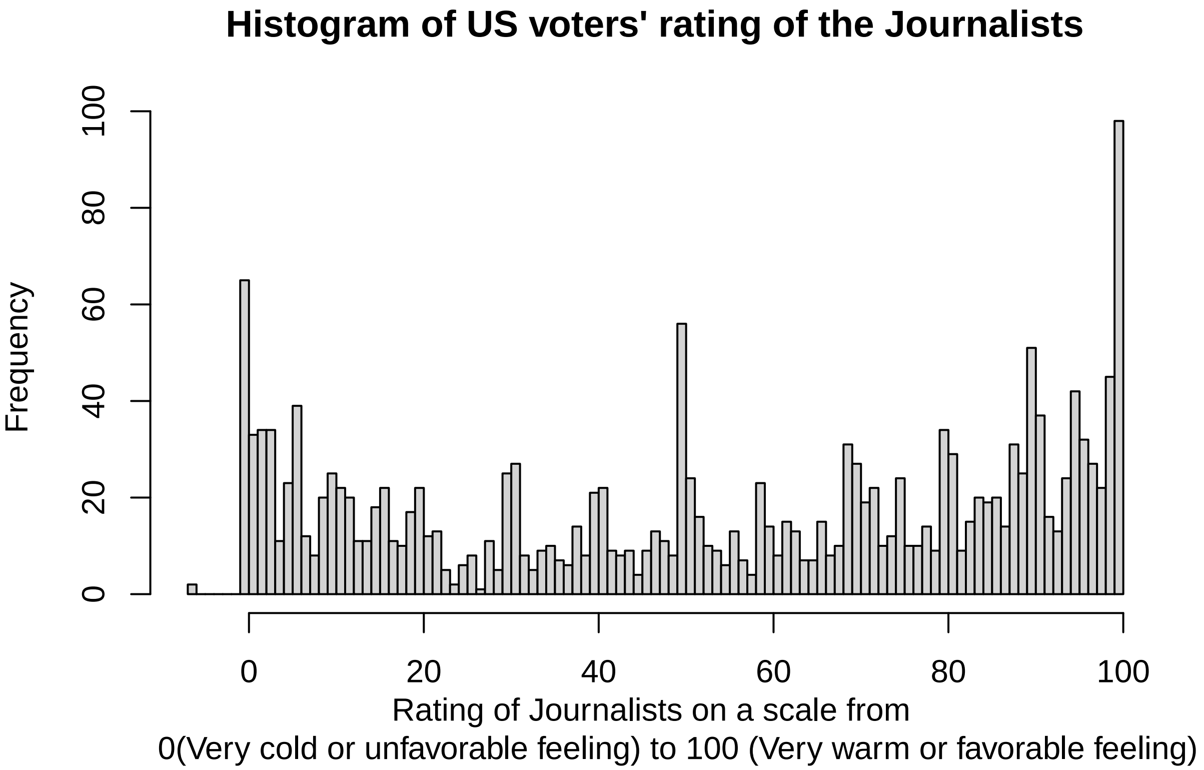
```r
hist(usVoters$ftpolice, breaks = 100,
     main = "Histogram of US voters' rating of the Police",
     xlab = "Rating of Police on a scale from
     0(Very cold or unfavorable feeling) to 100 (Very warm or favorable feeling)")
```



>

The distribution for ftpolice seems a little skewed towards the higher end, with upticks in the center.

```
hist(usVoters$ftjournal, breaks = 100,
     main = "Histogram of US voters’ rating of the Journalists",
     xlab = "Rating of Journalists on a scale from
     0(Very cold or unfavorable feeling) to 100 (Very warm or favorable feeling)")
```



The distribution for ftjournal is lesser skewed than that of ftpolice.

With the number of observations in both samples being well above the thumb rule (n=30), we can use the Central Limit Theorem to deduce that the distribution of the means approaches a normal distribution.

Based on our EDA, selecting an appropriate hypothesis test

1. Since we are comparing two samples, it will be a two-sample test.
2. Since the variables are of ordinal scale, we will use a non-parametric test.
3. Since the two variables we are looking at are answered by the same person, there is enough dependence that we can take advantage of if we do a paired test. With all these above reasons, we narrow our choice of test down to Sign Test.

Assumptions to be able to use the Sign Test are: 1. Variables represented are of ordnial Scale 2. Data is paired and drawn from i.i.d samples

Since both the assumptions are satisfied, we will go ahead with using the sign test.

The Null Hypothesis we will be testing is: US voters have equal respect for the police and journalists

We will do a two-tailed test since we are only interested in the alternative hypothesis: US voters do not have equal respect for the police and journalists

Conducting the chosen test

```
more_police = sum( usVoters$ftjournal < usVoters$ftpolice)
trials = sum( usVoters$ftjournal < usVoters$ftpolice | usVoters$ftjournal > usVoters$ftpolice)
binom.test(more_police , trials)
```

```
##
##  Exact binomial test
##
## data:  more_police and trials
## number of successes = 1008, number of trials = 1795, p-value =
## 2.004e-07
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5382418 0.5846764
## sample estimates:
## probability of success
##              0.5615599
```

> The p-value being well under 0.05 (statistical significance we decided to consider) indicates that we can reject the null hypothesis that US voters have equal respect for the police and journalists.
>
> Practical significance: The number of US voters in this sample who respect police more divided by the total number pairs of respecting either police or journalists more is 0.56. This is instead of the 0.5 we assumed for our null hypothesis, which would have reflected that US voters have equal respect for the police and journalists.

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

Introducing the topic

Conceptualize

Here we are trying to understand if there is an age component in the differences in political parties.

Operationalize

The survey itself asks its respondents many questions surrounding their political identity. We have several variables in which we can consider to use to define political identity. First, there is *pid1d* / *pid1r* which asks “Generally speaking, do you usually think of yourself as a Democrat, a Republican, an Independent, or what?”. Each of these reverse the order of which is asked first. There are follow up questions with these to try and figure out which way the respondent leans or how strongly *pidstr* and *pidlean*. There is also *pid7x* that also asks on a Likert-type scale. Since we are asking just ages of Republican vs Democrat and not HOW far they lean or how STRONGLY they feel that way, we can use the data from *pid1d*/*pid2d* (jointly, since 50% were asked one way and 50% the other). We will disregard WHO they voted for, as the person they voted for doesn’t represent which party they are affiliated with. For age, we can simply use *birthyr* and know that we will only have a year granularity for our data (rather than month and day).

Exploratory Data Analysis

Variables: *pid1d*, *pid1r*, *birthyr*

Types: Both *pid1d* and *pid1r* are ordinal type variables. However, since we are only limiting the sample to Republicans(0) and Democrats(1) which will make our new variable “Party” as nominal or metric. First, lets determine if we have enough Democrats and Republicans. We will have to acknowledge that using only responses 1 or 2 for *pid1d* and *pid2d* will remove non-responses, Independents, and “others” from the analysis. We will also have to confirm these individuals are voters in 2018 using *voted2018*.

Number of Entries:

```
Age_and_Party <-select(A, c("pid1d", "pid1r", "voted2018", "birthyr", "weight", "weight_spss", "caseid"
Age_and_Party$Party <-
  ifelse((Age_and_Party$pid1d==2 | Age_and_Party$pid1r == 2), 0,
  ifelse((Age_and_Party$pid1d==1 | Age_and_Party$pid1r == 1), 1, -1))
Age_and_Party<-filter(Age_and_Party, Party >= 0)
Age_and_Party<-filter(Age_and_Party, voted2018 >= 1)
paste("Number of Republicans: ", sum(Age_and_Party$Party==0))
```

```
## [1] "Number of Republicans:  503"
```

```
paste("Number of Democrats:   ", sum(Age_and_Party$Party==1))
```

```
## [1] "Number of Democrats:    725"
```

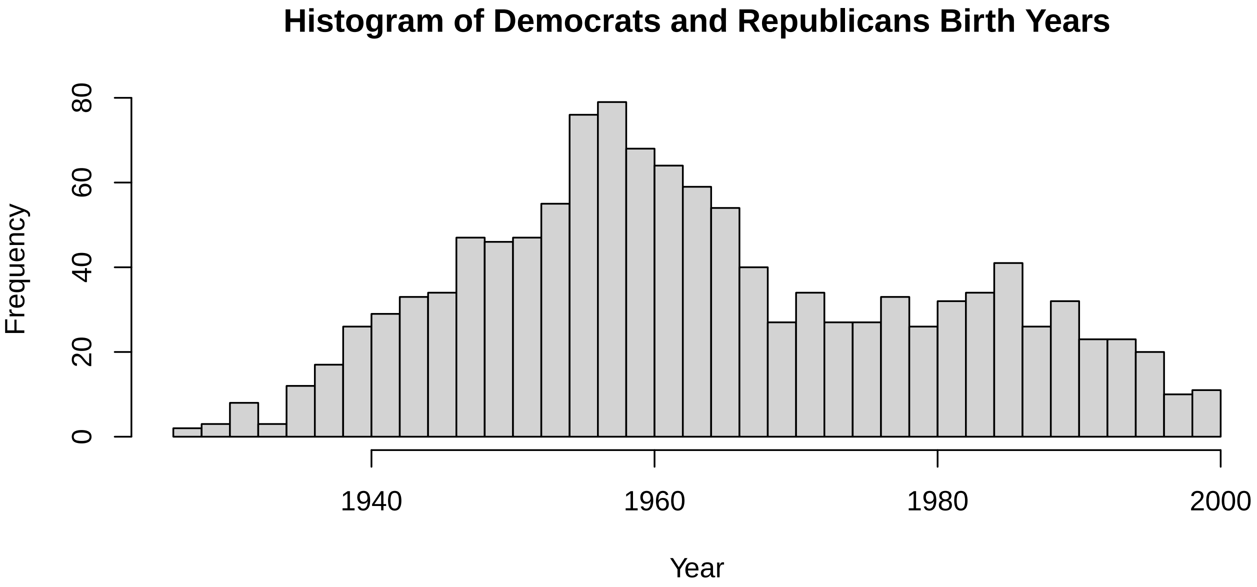
There are no Null Values for birthyr

```
paste("Number of Null Values for birthyr:", length(which(is.na(Age_and_Party$birthyr))))
```

```
## [1] "Number of Null Values for birthyr: 0"
```

We have 1228 people who claimed they are part of either party, Democrats and Republicans. Now we need to see if the sample distribution of their ages are normal enough to use CLT. We will draw a histogram of *birthyr* to observe the distribution and any outliers.

```r
hist(Age_and_Party$birthyr, breaks = 50,
     main = "Histogram of Democrats and Republicans Birth Years",
     xlab = "Year")
```



The histogram of ages has a somewhat bimodal distribution however, with 1228 responses, we can rely on CLT.

The Test

Since our sample's political id variable is metric (0 or 1) and *birthyr* is metric, i.i.d (see general assumptions on page 1) and the data cannot be paired, we would argue that using an unpaired t-test. The assumptions using this test is that the variables represented are metric and the data is drawn i.i.d. (See introduction). The *Null Hypothesis* is that the difference ages between Democrats and Republicans is equal to 0.

$$\mu_{D\_age} - \mu_{R\_age} = 0$$

```r
t.test(Age_and_Party$birthyr~Age_and_Party$Party)
```

```
##
##  Welch Two Sample t-test
##
## data:  Age_and_Party$birthyr by Age_and_Party$Party
## t = -3.6785, df = 1104.4, p-value = 0.0002459
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.235344 -1.593043
## sample estimates:
## mean in group 0 mean in group 1
##        1962.044        1965.458
```

Here we have a statistically significant test is the p-value being less than 0.05. We can reject the null hypothesis that there is an no age difference between Democrats and Republicans voters. Republican voters are, on average, 3 years older than Democrat voters since they have a mean of a birth year of 1962 and the Democrat voters have a birth year mean of 1965.

The Weighted Test

```r
library(survey)
```

```
## Loading required package: grid
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loading required package: survival
```

```
##
## Attaching package: 'survey'
```

```
## The following object is masked from 'package:graphics':
##
##     dotchart
```

```r
Age_and_Party_Weighted <-
  svydesign(id      = ~caseid,
            weights = ~weight,
            data    = Age_and_Party)

paste("Does Weighting make a difference in the mean birth year?")
```

```
## [1] "Does Weighting make a difference in the mean birth year?"
```

```r
paste("Average Birth Year before Weights: ", mean(Age_and_Party$birthyr))
```

```
## [1] "Average Birth Year before Weights:  1964.05944625407"
```

```r
paste("Average Birth Year after Weights: ", svymean(~birthyr, Age_and_Party_Weighted))
```

```
## [1] "Average Birth Year after Weights:  1967.45390343812"
```

```
svyttest(birthyr~Party, Age_and_Party_Weighted)
```

```
##
##  Design-based t-test
##
## data:  birthyr ~ Party
## t = 2.8754, df = 1226, p-value = 0.004104
## alternative hypothesis: true difference in mean is not equal to 0
## 95 percent confidence interval:
##  1.243632 6.568711
## sample estimates:
## difference in mean
##           3.906171
```

When applying the weights, the mean birth year of the respondants increased from 1964 to 1967, making the average population of 2018 Democrate and Republican voters younger by nearly 3 years. Even weighted, we still have a statistically significant test is the p-value being less than 0.05. We can again reject the null hypothesis that there is an no age difference between Democrats and Republicans voters. Using weights, Republican voters are, on average, 3.9 (not 3.4) years older than Democrat voters. It didn't change our outcome of the test, but it slightly modified our practical significance, in which the difference in mean was closer to 4 years than to 3 years.

> Why did weights make a difference? The survey designers added in the weights to make sure the sample is a full representation of the US voting population. If they couldn't fully complete their stratified random sampling, or there was a lot of non-respondants, groups could be underrepresnted or overrepresented in the sample. Weights are then calculated to give respondants from these groups more or less "power" to the data represented so that one can infer more from the data to the population … or in other words calibrate the sample to the population.

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

Introduce your topic briefly. (5 points)

Operationalization Te three most relevant fields look to be `russia16` (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?), `muellerinv` (Do you approve, disapprove, or neither approve nor disapprove of Robert Mueller's investigation of Russian interference in the 2016 election?) and `coord16` (Do you think Donald Trump's 2016 campaign probably coordinated with the Russians, or do you think his campaign probably did not do this?). These variables are likely to be highly correlated, so it will be important to operationalize them wisely as to try to answer the question in the most valid manner.

The key question is what does it mean "to believe that the federal investigations of Russian election interference are baseless"? Our interpretation is that this phrasing is equivalent to "to believe that no Russian election interference happened". Using a measure of approval of the specific Mueller investigation is not an appropriate answer to this question as it is completely credible to "not believe that the investigation is baseless", but disapprove of it for some other reason.

The other possible variable, measuring opinions whether the Trump campaign coordinated with the Russians, is also tangential to the main question, which refers to the investigation as one of the Russian interference, not as an investigation of the Trump campaign's coordination.

With the exception of possible "No Answers", the variable `russia16` is a binary variable (yes/no) and therefore metric. If we get some "No Answers" we'll have to consider eliminating them.

Population: We can take the self-defined Independent voters for whom `pid7x` is 3, 4, or 5.

Gaps: The main issue here is about using disbelief in Russian interference as an indicator of belief in the baselessness of the investigation. At the same time, it seems reasonable that if independent voters believed in the existence of Russian interferece, they would not find an investigation of such interference baseless. So the variable seems like a good measure.

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

Let's pull out our sample.

```r
ind_voters2 <- A %>%
  filter(pid7x == 3 | pid7x == 4 | pid7x == 5)
ind_voters2 <- filter(ind_voters2, voted2018 >= 1)
table(ind_voters2$russia16)
```

```
##
##   1   2
## 336 265
```

So the subset of independent voters has no -7 (No Answer) values and is properly binary.

Based on your EDA, select an appropriate hypothesis test. (5 points)

The variable is metric and its distribution is bimodal, but that is not a major problem for the CLT. In addition, there is a large sample size. We can run a one-sample t.test.

Our null hypothesis is that exactly 50% of independent voters find the investigation baseless. That translates to a mean of the variable of 1.5, but for more clarity we can just recode the negative responses to 0 and the positive responses to 1.

Since we don't have a clear view to the directionality of a possible rejection of the null hypothesis and since we would be interested in a statistically significant result in wither direction, we'll run a two-tailed test.

Conduct your test. (5 points)

```r
test_var <- ifelse(ind_voters2$russia16 == 1, 1, 0)
t.test(test_var, mu = 0.5)
```

```
##
##  One Sample t-test
##
## data:  test_var
## t = 2.9141, df = 600, p-value = 0.0037
## alternative hypothesis: true mean is not equal to 0.5
## 95 percent confidence interval:
##  0.5192605 0.5988760
## sample estimates:
## mean of x
## 0.5590682
```

Statistically significant result. We reject the null hypothesis. There is evidence in the data in support of the alternative hypothesis. Since we ran a two-tailed test we are able to report that the calculated t-statistic of 2.9141 is in the upper tail, that is pointing to the conclusion that a majority of independent voters believe that Russian interference did in fact occur, and hence by the logic laid out in our introduction, that only a minority of independent voters find the federal investigation baseless, the null hypothesis can be rejected.

In terms of practical significance, given the polarization and relative parity of the Democratic and Republican voting blocks, if indeed in the population 56% of independent voters believe that the federal investigation is warranted and that motivates them to vote against the party that benefited from the alleged interference, that could be a very significant effect. In close elections, which are usually zero-sum games, a 6% percent swing from parity to one side is actually a 12% swing in the overall result. In the political campaign world that is massive, even when it is only among a subsection of the electorate. And in this case that happens to be the most sought after group of voters.

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

Introduction: This question is complex and touches on several different ideas than need to be unpacked. First we need to find in the dataset some measure of voter fear and anger. Then we need to address if it is at all feasible to reason about "effectiveness in driving voter increases". Finally, we need to somehow define what voter turnout increases between 2016 and 2018 might be and what sample from the dataset might reflect that.

Variables The leading candidates for looking at voter anger and fear and `geangry` and `geafraid` - two ordinal variables measuring the reaction to possible responses to the question "Generally speaking, how do you feel about the way things are going in the country these days?" The scale ranges from 1 (Not at all) to 5 (Extremely).

There are two other candidate survey questions featuring fear and anger, but they address some narrower issues (Donald Trump's behavior and immigration to the US) and were also randomized (one half of respondents saw one question, the rest saw the other), they would require a lot more assumptions to operationalize.

Gaps The crucial gap here is that the question is implicitly one about causality. Further more, it presumes that anger and fear drive (cause) higher turnout and asks which of the causal effects is stronger. We cannot answer that question with this data. What we can look into is whether there's a difference between the anger/fear responses of those who voted in 2018 after sitting out the 2016 election and the responses of those who didn't vote in either.

The second gap is closely connected to choosing our sample. The 2018 election did indeed see a historically high turnout, but still that was lower than the turnout of the 2016 election. The reason is that 2016 was a presidential election year, while 2018 was only a midterm election and those typically have lower turnouts than presidential elections.

Operationalization and population In order to address the question, while at the same time acknowledging the two significant gaps, we are going to make the case that by comparing a sample of people who voted in 2018, but not in 2016 to a sample of people who didn't vote in either election, we are indeed looking at "an increase" in turnout. To put it slightly differently, it can be reasonably expected that close to an entirety of the population of people who care about elections and do vote would definitely turn up for the most consequential of elections - that for President. Under that scenario, in your run-of-the-mill regular midterm election, midterm voters can reasonably be expected to be a subset of those who usually vote for President. So if we see a substantial group of voters voting in a midterm election after not voting for President, we can reason that those voters represent an increase in turnout.

For extracting the two samples, we first start with our `voter_sample` dataframe and use the field `turnout16` ("In 2016, the major candidates for president were Donald Trump for the Republicans and Hillary Clinton for the Democrats. In that election, did you definitely vote, definitely not vote, or are you not completely sure whether you voted?"), `turnout16b` ([IF turnout16=3] "Do you think you probably voted or probably did not vote?"), and `birthyr` to determine our sample of respondents who were eligible in 2016, didn't vote in 2016 and voted in 2018.

```
voter_sample = A[ which(A$turnout18 < 4),]
to_increase <- voter_sample[which((voter_sample$turnout16 == 2 | voter_sample$turnout16b == 2) & voter_s
```

Then for the sample of non-voters we go back to the entire dataset, filter out our 2018 voters and then grab just those who were eligible in 2016, but didn't vote.

```
non_vote <- A[which((A$turnout16 == 2 | A$turnout16b == 2) & A$turnout18 > 3 & A$birthyr < 1999),]
```

Perform an exploratory data analysis (EDA) of the relevant variables. (5 points)

This should include a treatment of non-response and other special codes, basic sanity checks, and a justification for any values that are removed. Use visual tools to assess the relationship among your variables and comment on any features you find.

Let's look at the sizes of our samples.

```
paste(count(to_increase), count(non_vote))
```

```
## [1] "88 489"
```

Let's look at missing answers for `geangry` and `geafraid`

```
table(to_increase$geangry)
```

```
##
## -7  1  2  3  4  5
##  1 16 18 28  9 16
```

```
table(to_increase$geafraid)
```

```
##
## -7  1  2  3  4  5
##  3 18 18 26  6 17
```

```
table(non_vote$geangry)
```

```
##
##  -7   1   2   3   4   5
##   1 146  91 129  67  55
```

```
table(non_vote$geafraid)
```

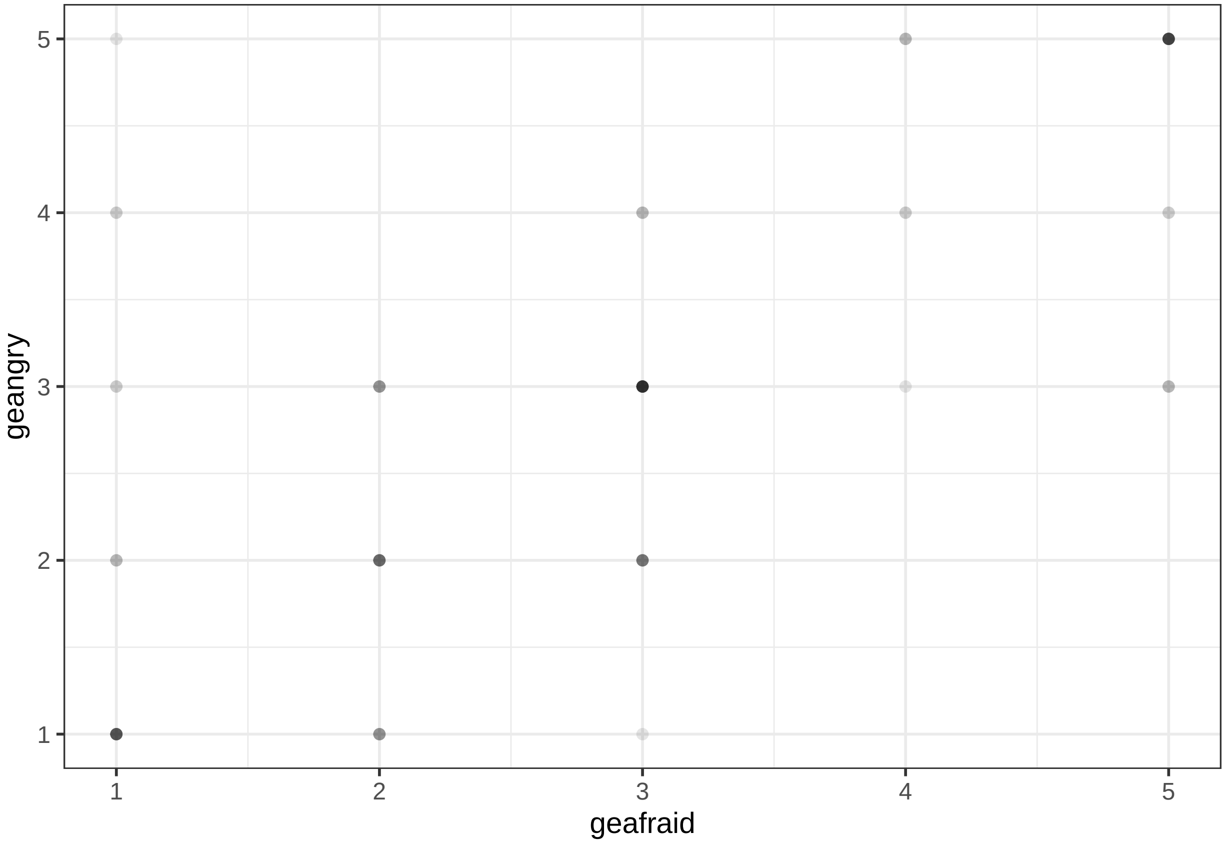
```
##
##  -7   1   2   3   4   5
##   2 138 113 129  60  47
```

So we have 88 observations in one sample and 489 in the other. Of those, only a handful are have value of -7 (No Answer) for `geafraid` or `geangry`. It seems reasonable to omit them.

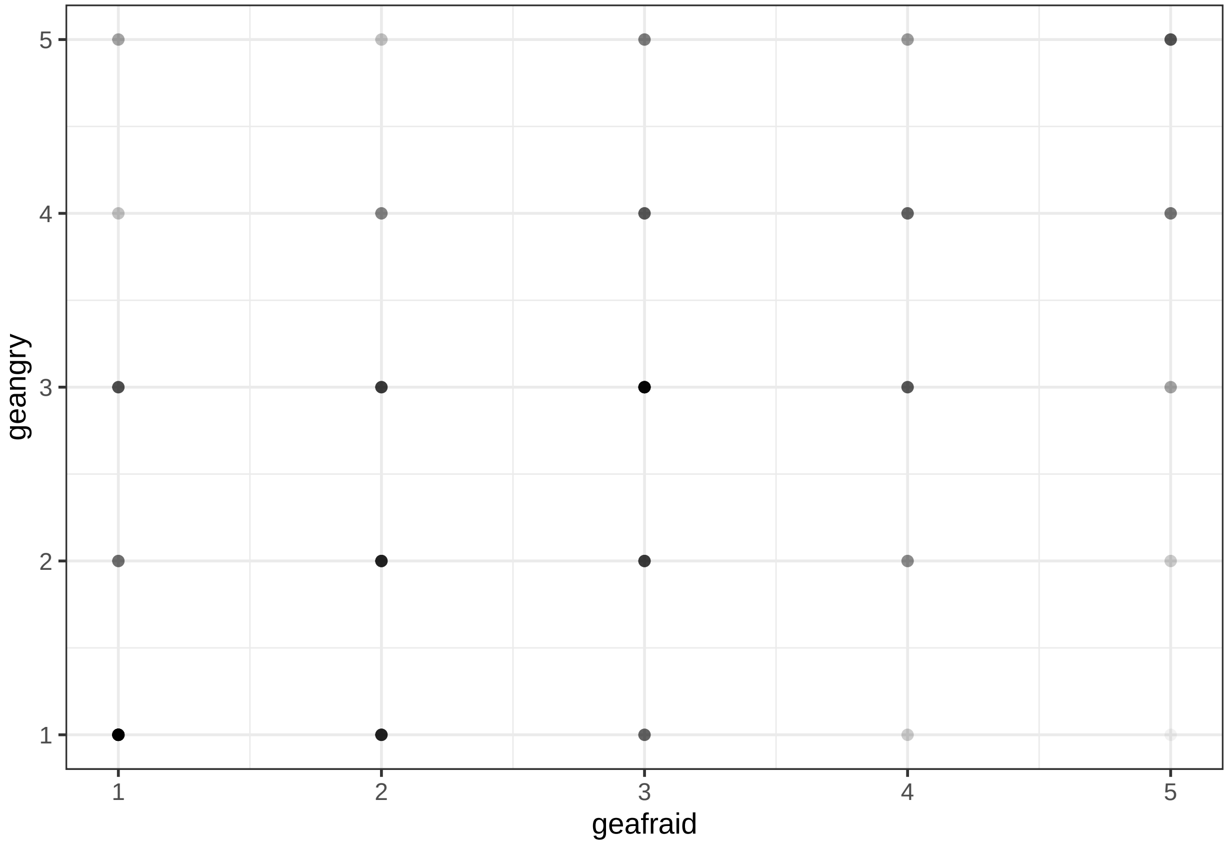
```
to_increase <- to_increase %>%
  filter(geafraid > 0) %>%
  filter(geangry > 0)

non_vote <- non_vote %>%
  filter(geafraid > 0) %>%
  filter(geangry > 0)
```

Now we can create two scatter plots to see where the values for `geafraid` and `geangry` fall for the two samples.

```
ggplot(to_increase, aes(x=geafraid, y=geangry) ) +
  geom_point(alpha = 0.1) +
  theme_bw()
```



```
ggplot(non_vote, aes(x=geafraid, y=geangry) ) +
  geom_point(alpha = 0.05) +
  theme_bw()
```



Based on your EDA, select an appropriate hypothesis test. (5 points)

These are paired ordinal variables, so we have to use sign tests. Since we have two different samples and we're interested in the difference between `geafraid` and `geangry` responses for both of those, we propose running seperate tests on each sample.

Both null hypotheses are that in each relevant sample there is no difference between `geafraid` and `geangry` responses.

Newly turned out voters:

```
more_afraid <- sum( to_increase$geangry < to_increase$geafraid)
trials <- sum( to_increase$geangry < to_increase$geafraid | to_increase$geangry > to_increase$geafraid)
binom.test(more_afraid, trials)
```

```
##
##  Exact binomial test
##
## data:  more_afraid and trials
## number of successes = 19, number of trials = 38, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.333789 0.666211
## sample estimates:
## probability of success
##                    0.5
```

Non-voters:

```
more_afraid <- sum( non_vote$geangry < non_vote$geafraid)
trials <- sum( non_vote$geangry < non_vote$geafraid | non_vote$geangry > non_vote$geafraid)
binom.test(more_afraid, trials)
```

```
##
##  Exact binomial test
##
## data:  more_afraid and trials
## number of successes = 139, number of trials = 272, p-value = 0.7618
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4499470 0.5718689
## sample estimates:
## probability of success
##              0.5110294
```

We fail to reject the null hypothesis in both samples. As Yoda said, fear leads to anger and anger leads to suffering, so perhaps fear and anger are closely related, but it's unclear which, if any, drives turnout more efficiently.

Question 5: Do voters for Hillary Clinton perceive the chane in income inequality over the last 20 years differently from how voters for Donald Trump perceive it.

Introducing our Research Question

There are a variety of issues people care about and consider significantly while voting and while choosing a candidate to vote for. Income inequality is one such issue. Income inequality has fluctuated over the years, both increasing and decreasing in portions of the last 20 years.

The question we are interested in is how Hillary voters view the change in income inequality over the last 20 years versus how Trump voters view the change in income inequality over the last 20 years.

It has been argued that the extent and nature of participation in choosing leaders is closely associated with the distribution of resources in society (https://nathanjkelly.utk.edu/wp-content/uploads/2017/10/Franko-et-al-2016.pdf). It is plausible that voters' perception of change in income inequality could affect their voting behavior. Therefore, we feel this is an important question in understanding voting behavior.

We will evaluate this by building the appropriate hypothesis and using appropriate statistical tests under the assumptions that make them valid.

Concept: To evaluate if Hillary voters look at the change in income inequality over the last 20 years differently from the way Trump voters look at the same. We are interested in seeing what people's perception of change in income inequality is in 2018 based on what candidate they preferred in 2016.

Operationlization: The variable 'richpoor' captures the answer to the question "Do you think the difference in incomes between rich people and poor people in the United States today is larger, smaller, or the same as it was 20 years ago?". This question was presented to all the participants of the survey. It can take one of the following seven values (along with the representing choices that the participants given to choose from): 1 - A lot larger 2 - A moderate amount larger 3 - A little larger 4 - The same 5 - A little smaller 6 - A moderate amount smaller 7 - A lot smaller We will use this variable to understand how the voters viewed the change in income inequality in the last 20 years.

The variable 'vote16' captures the answer to the questions "In the 2016 presidential election, who did you vote for? Donald Trump, Hillary Clinton, or someone else?". It takes one of the following three values (along with the representing choices that the participants given to choose from): 1 - Donald Trump 2 - Hillary Clinton 3 - Someone else We will use this to distinguish between Hillary voters and Trump voters.

Exploratory data analysis (EDA) of the relevant variables

Variables: richpoor, vote16

Types: richpoor is an ordinal variable since it has categories that have an order. Operataions like >,<,= are valid for the 'richpoor' variable. vote16 is a nominal variable since it has categories that do not have an order. Operataions like >,<,= do not make sense for the vote16 variable. We are only using the vote16 variable to separate out the two samples we are interested in comparing.

Number of Entries

```
hillary_voters <- A %>% filter(A$vote16==2)
trump_voters <- A %>% filter(A$vote16==1)

paste("Number of voters who voted for Hillary Clinton in 2016:", length(hillary_voters$vote16))
```

```
## [1] "Number of voters who voted for Hillary Clinton in 2016: 898"
```

```
paste("Number of voters who voted for Donald Trump in 2016:", length(trump_voters$vote16))
```

```
## [1] "Number of voters who voted for Donald Trump in 2016: 770"
```

There are no Null Values for the variable richpoor for Hillary or Trump voters

```
paste("Number of Null Values for richpoor for Hillary voters:", length(which(is.na(hillary_voters$richp
```

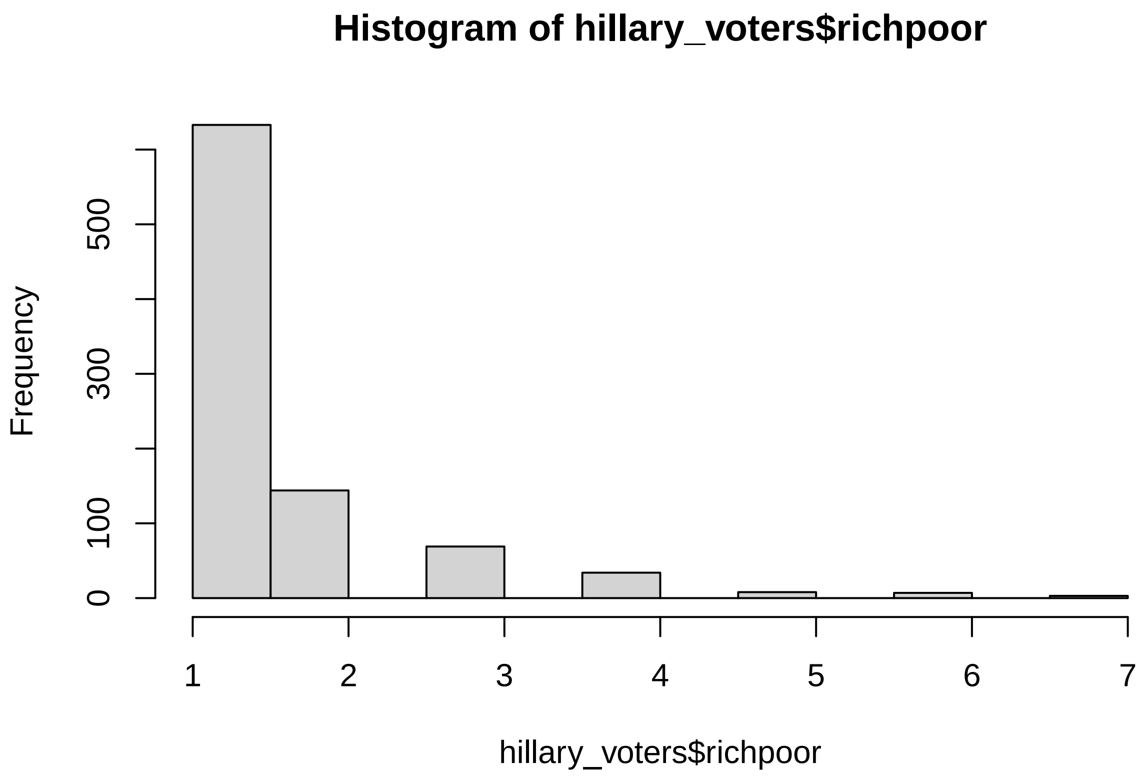
```
## [1] "Number of Null Values for richpoor for Hillary voters: 0"
```

```
paste("Number of Null Values for richpoor for Trump voters:", length(which(is.na(trump_voters$richpoor)
```

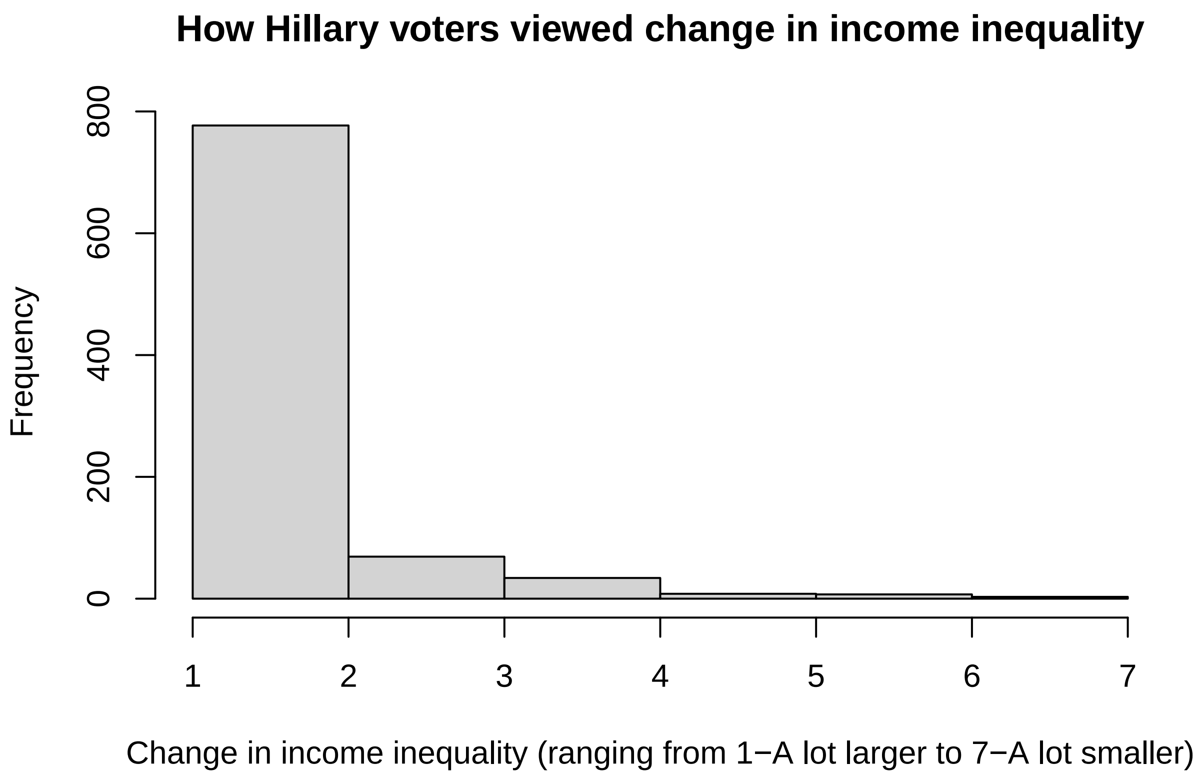
```
## [1] "Number of Null Values for richpoor for Trump voters: 0"
```

Drawing out the histograms for the richpoor variable for both Hillary and Trump voters to observe sampling distribution

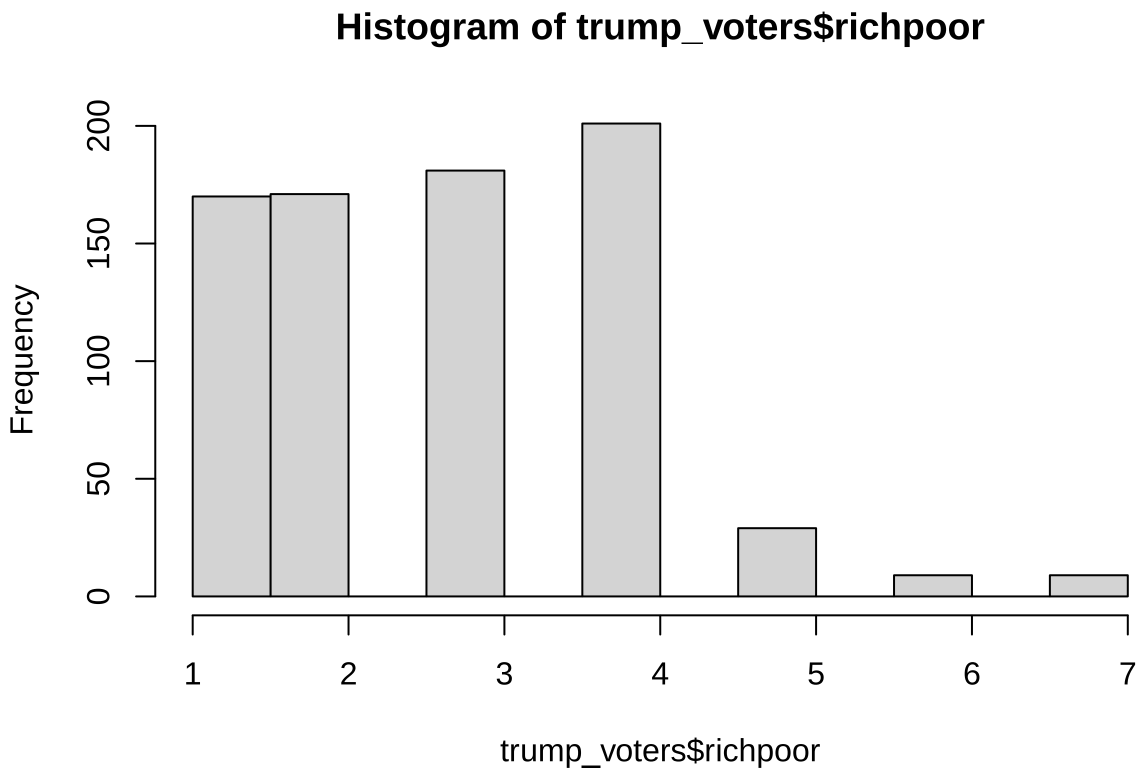
```
hist(hillary_voters$richpoor)
```



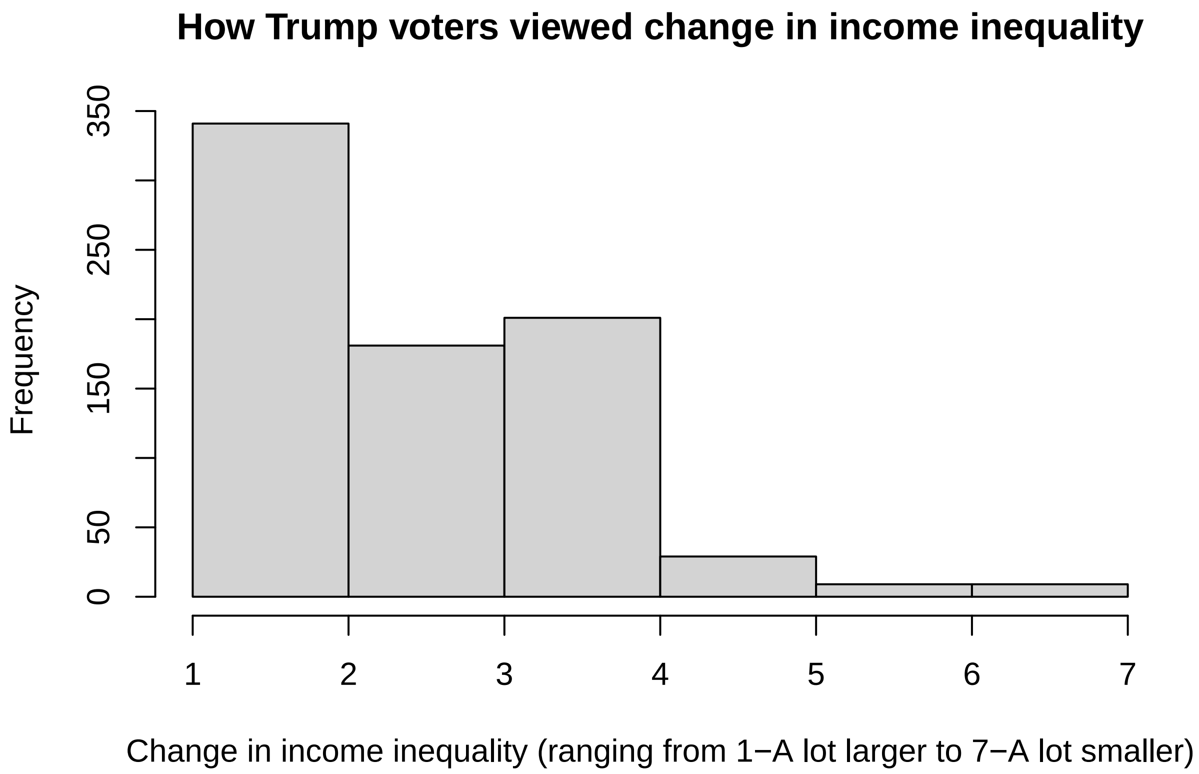
```
hist(hillary_voters$richpoor, breaks = 7,
     main = "How Hillary voters viewed change in income inequality",
     xlab = "Change in income inequality (ranging from 1-A lot larger to 7-A lot smaller)")
```



```
hist(trump_voters$richpoor)
```



```r
hist(trump_voters$richpoor, breaks = 7,
     main = "How Trump voters viewed change in income inequality",
     xlab = "Change in income inequality (ranging from 1-A lot larger to 7-A lot smaller)")
```



Based on our EDA, selecting an appropriate hypothesis test

1. Since we are comparing two samples, it will be a two-sample test.
2. Since the variables are of ordinal scale, parametric tests cannot be used.
3. There is no dependence to rely on and therefore we will use an unpaired test. The distributions from the histograms look slightly skewed towards the left however, since we are intending to use the Wilcoxon Rank-Sum test, the skewness does not affect our choice of test.

With all these above reasons, we narrow our choice of test down to Wicoxon rank-sum test.

Assumptions to be able to use the Sign Test are: 1. Variables represented are of ordnial Scale 2. Data is paired and drawn from i.i.d samples

Since both the assumptions are satisfied, we will go ahead with using the Wicoxon rank-sum test.

The Null Hypothesis we will be testing is: Hillary voters do not view change in income inequality over the last 20 years any differently than Trump voters do.

We will do a two-tailed test since we are only interested in the alternative hypothesis: There is a difference in the way Hillary voters and Trump voters view change in income inequality over the last 20 years.

Conducting the chosen test

```
wilcox.test(trump_voters$richpoor,hillary_voters$richpoor)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  trump_voters$richpoor and hillary_voters$richpoor
## W = 536432, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

The p-value is less than 0.05 indicating that this test is statistically significant to reject the null hypothesis that Hillary voters do not view change in income inequality over the last 20 years any differently than Trump voters do.

Conclusion

From this test we see that Hillary voters view the change income inequality over the past 20 years differently from how Trump voters view this change.

We have noticed that the histograms we saw for the sample distributions reflected differences in how Hillary voters perceived change in income inequality over the past 20 years versus how Trump voters did. More Hillary voters thought of the inequality to have become a lot larger whereas Trump voters had opinions more spread on the scale.

From the wilcox.test documentation, “R’s value can also be computed as the number of all pairs (trump_voters$richpoor, hillary_voters$richpoor) for which hillary_voters$richpoorisnotgreaterthantrump_voters$richpoor the most common definition of the Mann-Whitney test.”

The total number of pairs is,

```
(sum(A$vote16 == 1 & !is.na(A$richpoor)) * sum(A$vote16 == 2 & !is.na(A$richpoor)))
```

```
## [1] 691460
```

536432 (value of W obtained from our test) out 691460 pairs showed that there was a shift in the percieved rating for change in income inequality between Trump and Hillary voters. The percentage that gets us is

```
paste(536432*100/691460, "%")
```

```
## [1] "77.5796141497701 %"
```

77% is a percent we feel comfortable to use and say that this shows that the test is practically significant along with being statistically significant.

Common wisdom is that Democrat voters are more concerned about issues of inequality as compared to Republican voters and test results are consistent with this view.