Lab 1: Comparing Means

Srishti Mehra

```
A = read.csv("anes_pilot_2018.csv")
```

Research Questions

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

Introducing the topic

Concept: To measure the amount of respect US voters have for police and jounralists, since it is subjective, we cannot use a metric variable. We have to use a scale like the Likert scale to possibly capture this data and therefore it can be best defined with an ordinal 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 are the only ones related to what the sample population feels about the police and journalists respectively. The answers to these questions are collected on a Likert scale. We are also going to take a subset of the total sample available from the ANES 2018 Pilot study of those people who have reported that they voted. To do so we will consider the variable 'turnout18' which contains the answer to the question "In the election held on November 6, did you definitely vote in person on election?". The variable holds one of the following five values:

1 - Definitely voted in person on Nov 6 2 - Definitely voted in person, before Nov 6 3 - Definitely voted by mail 4 - Definitely did not vote 5 - Not completely sure

Another variable that could inform a portion of the population who have the 'turnout18' variable value as 5 who could have possibly voted can be covered from the variable 'turnout18ns'. This variable captures answer to the question "If you had to guess, would you say that you probably did vote in the election?". This variable takes one of the following three values:

1 - inapplicable, legitimate skip 2 - Probably did vote 3 - Probably did not vote

For our analysis, we will only consider the population that have value 1, 2, or 3 for the variable 'turnout18' to take the conserative approach in considering US voters.

Concerns or Gaps:

1. The data collected is for rating police and journalists in general, not particularly based on respect. Inferring respect from a total rating will not give us accurate conclusions to draw.
2. One person's interpretation of the Likert 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 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 Likert 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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.4     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
usVoters <- A  %>% filter(A$turnout18 > 0 & A$turnout18 < 4)
paste("Number of Entries for ftpolice:", length(usVoters$ftpolice))
```

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

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

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

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

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

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

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

```
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"
```

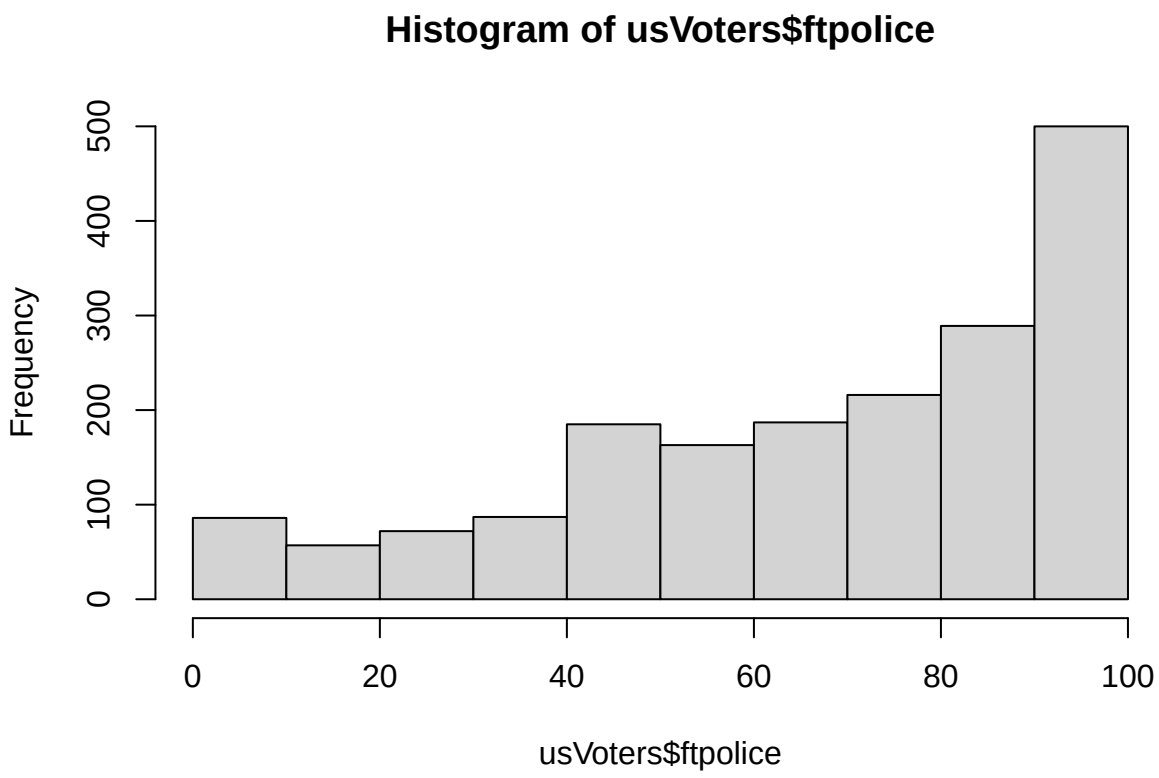
There are no duplicate rows in the dataset of US Voters

```r
paste("Number of Duplicate Values for the dataset usVoters:", length(which(duplicated(usVoters))))
```

```
## [1] "Number of Duplicate Values for the dataset usVoters: 0"
```

[ATTENTION!!] Will draw a histogram of both ftpolice and ftjournal to observe the distribution and any outliers

```r
hist(usVoters$ftpolice)
```



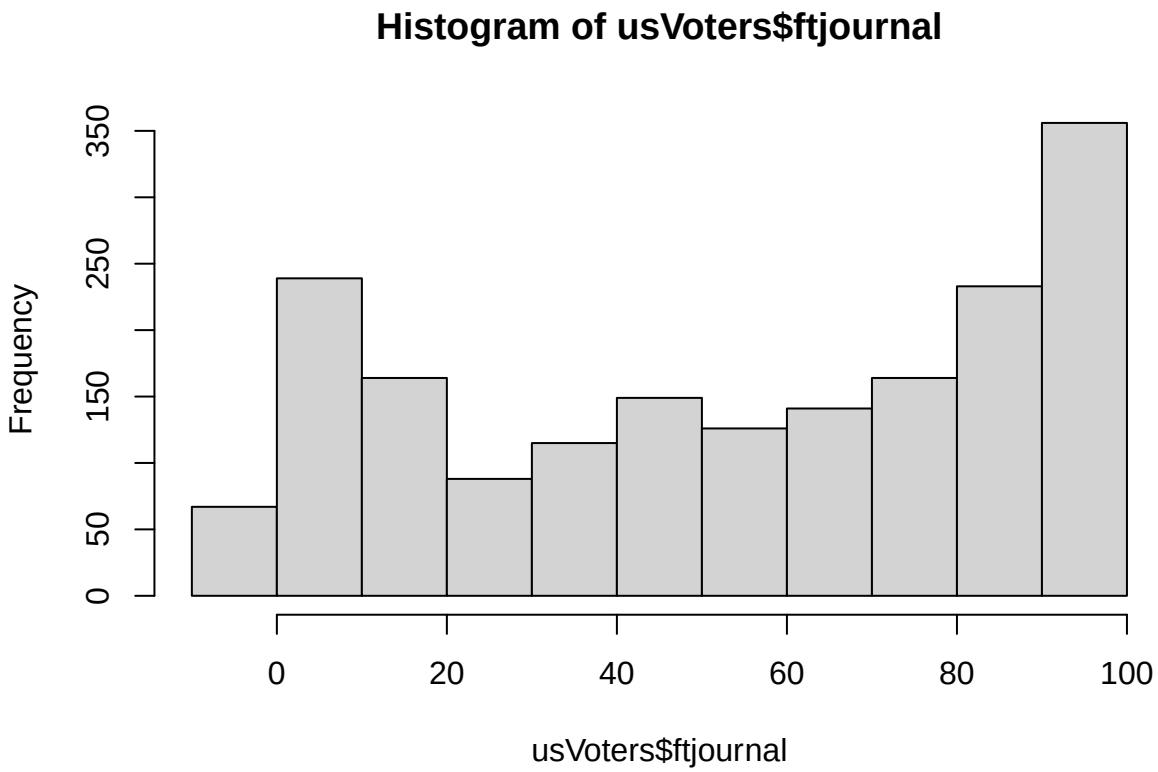
We will now observe the skewness of ftpolice variable

```
library(e1071)
skewness(usVoters$ftpolice)
```

```
## [1] -0.7689949
```

The distribution for ftpolice seems a little skewed towards the higher end, but the skewness level is moderate (skewness is between -1 and -0.5 or between 0.5 and 1).

```
hist(usVoters$ftjournal)
```



We will observe the skewness of ftpolice variable

```
skewness(usVoters$ftjournal)
```

```
## [1] -0.2235562
```

The distribution for ftjournal is lesser skewed than ftpolice and the skewness level is at a low (skewness is between -0.5 and 0.5,).

We will now observe the variable to see if it is identically and independently distributed (iid)

As a modern survey, the ANES 2018 Pilot Study we can expect the selection to be from iid drawn. You might worry that some people are more likely to respond than others (some more motivated for the money since this is a survey they get paid for). But this isn't a problem for iid, it just means that the population of respondents is not the same as the population representing all Americans. It's a problem for generalizability. Since there is only a finite number of Americans, it is impossible to have perfect independence. Once a person is drawn, they cannot be drawn again, so the distribution for the next draw has changed. However, the change is likely to be very small, so we can safely ignore these finite-sample effects. On the whole, there are issues to worry about, but modern surveys like the ANES 2018 Pilot Study do a pretty good job of making the iid assumption valid. Codebook and User's Guide to the ANES 2018 Pilot Study mentiones that the survey participation is independent of variables measured in the survey.

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. 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

We will consider the significance level of 0.01 since we are looking at a large number of observations. We understand that with greater number of observations, statistical significance gets easier to achieve for small effects, so we are considering 0.01 and not 0.05 as our significance level for this test.

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 0.01 (statistical significance we decided to consider) indicating that we can reject the null hypothesis that US voters have equal respect for the police and journalists.

Practical significance: We see that the probability of US voters have respecting police more comes up as 0.56 from the Sign Test. In our hypothesis we assumed it would be 0.5 reflecting that US voters have equal respect for the police and journalists. Thus, we see that the effect is not a very large one even though the p-value was very low to reflect a statistically significant result.

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

Introducing the topic

Concept: To measure the difference in ages of voters who identified as Republican or Democratic, we can use a metric variable because age is (just!) a number.

Operationalization: We have to find a variable that shows if the voter identifies as Republican or Democratic. The variable 'pid1d' captures the result of the question "Generally speaking, do you usually think of yourself as a Democrat, a Republican" in the survey matches closely to what we want and thus, we will use it. This variable takes one of the following six values: -7 No Answer -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3 independent 4 something else

There is another variable 'pid7x' which is a part of YouGov profile survey data that contains this affiliation. However, since these data were collected on previously-completed questionnaires, we identify the value in 'pid1d' being more recent and therefore will use that for the analysis our question.

For the age, we can use the 'birthyr' variable that is captured as a part of the survey, subtract that from the current year and get age.

Exploratory data analysis (EDA) of the relevant variables

Variables: birthyr, pid1d

Types: pid1d is a nominal level variable, however, we are only using that to separate our two samples that we are going to compare the ages for. Age is a ratio level variable.

Number of Entries

```
currentYr <- lubridate::isoyear(Sys.Date())
usVoters$age <- currentYr-usVoters$birthyr
democrats <- usVoters  %>% filter(usVoters$pid1d==1)
republicans <- usVoters  %>% filter(usVoters$pid1d==2)

paste("Number of Entries for age of democrats:", length(democrats$age))
```

```
## [1] "Number of Entries for age of democrats: 375"
```

```r
paste("Summary for age of democrats: ")
```

```
## [1] "Summary for age of democrats: "
```

```r
summary(democrats$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   20.00   41.00   57.00   54.96   67.00   93.00
```

```r
paste("Number of Entries for age of republicans:", length(republicans$age))
```

```
## [1] "Number of Entries for age of republicans: 267"
```

```r
paste("Summary for age of republicans: ")
```

```
## [1] "Summary for age of republicans: "
```

```r
summary(republicans$age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.00   47.00   60.00   57.92   69.50   92.00
```

There are no Null Values for ages of democrats or republicans

```r
paste("Number of Null Values for ages of democrats:", length(which(is.na(democrats$age))))
```

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

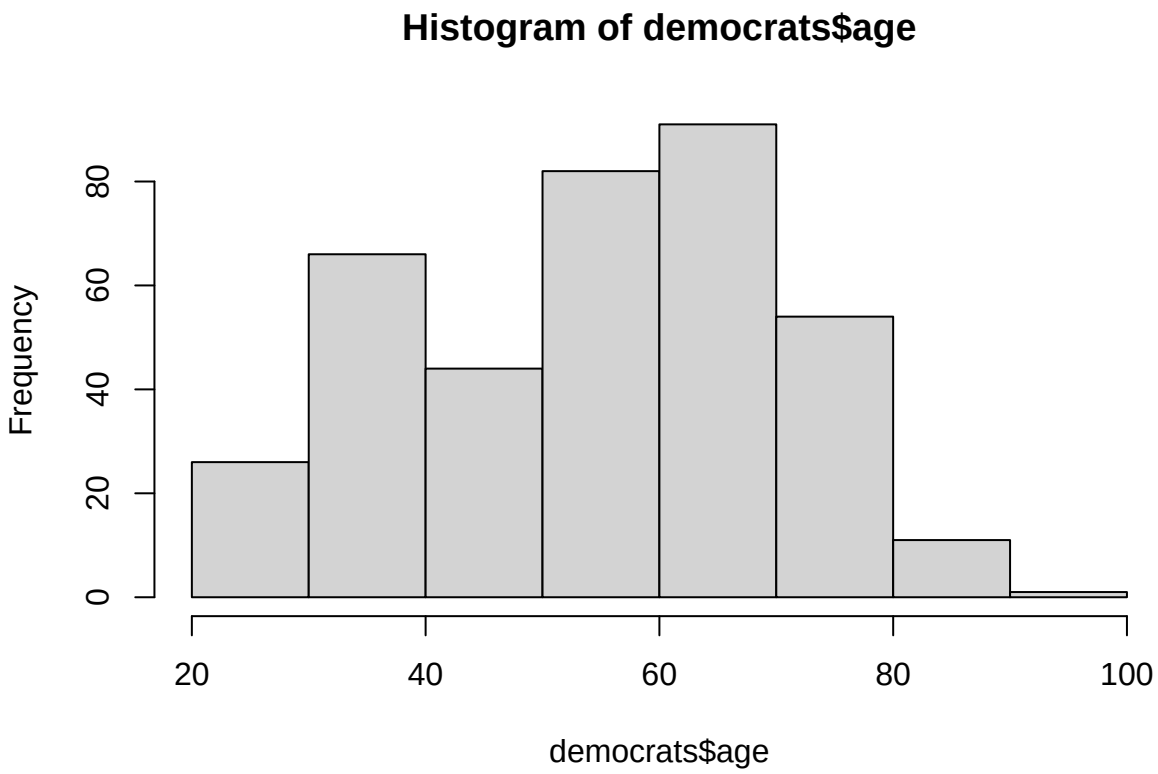
```r
paste("Number of Null Values for ages of republicans:", length(which(is.na(democrats$age))))
```

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

There are no duplicate rows in the dataset of US Voters as we had already established while answering research question 1

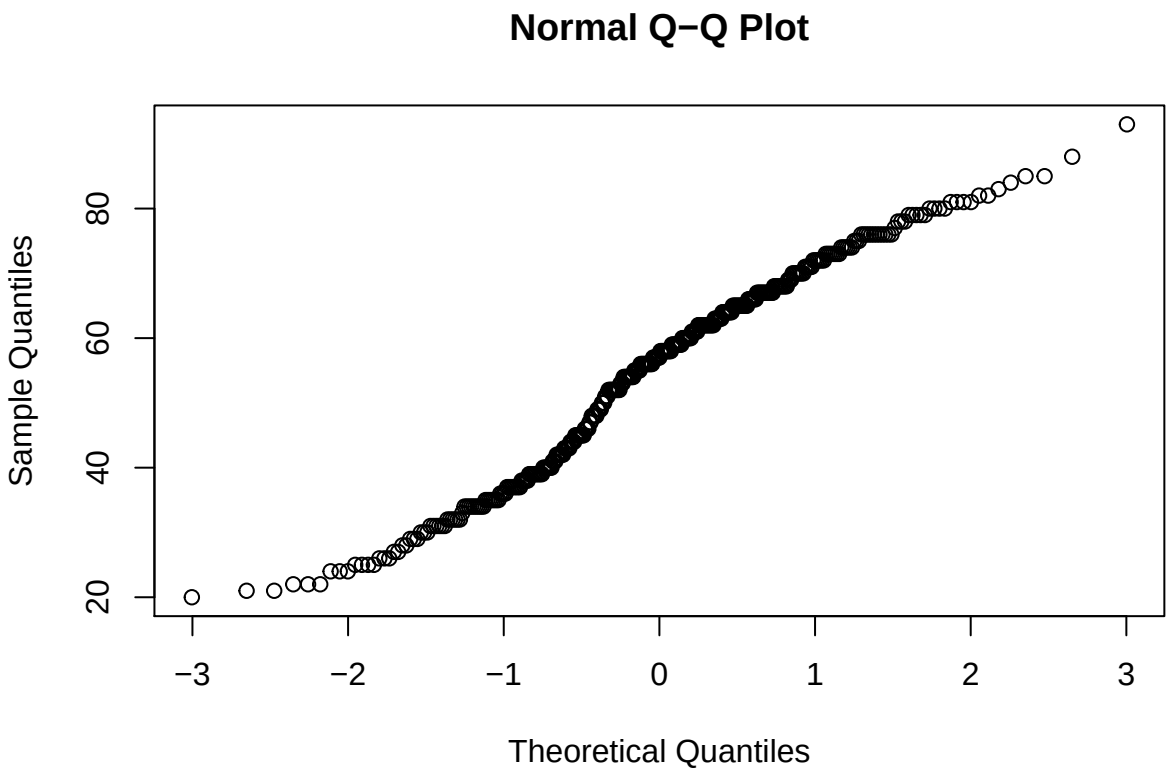
We will draw a histogram of ages of democrats to observe the distribution and any outliers

```r
hist(democrats$age)
```



To see the deviation from normality, we will plot a qqnorm plot

```
qqnorm(democrats$age)
```



We will observe the skewness of age of democrats

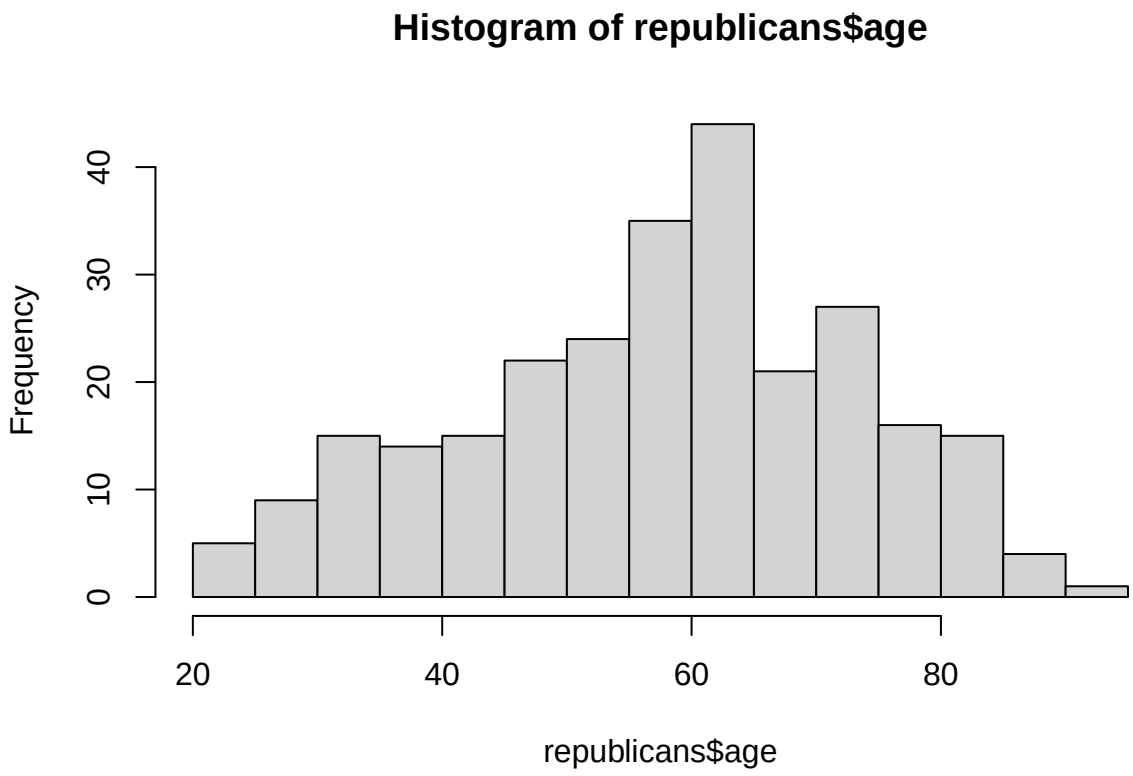
```
skewness(democrats$age)
```

```
## [1] -0.2216287
```

The distribution for age of democrats is not far from normal, and with the number of observations being well above the thumb rule and the skewness being at at a low level (skewness is between -0.5 and 0.5), we can use the Central Limit Theorem to deduce that the distribution of the means approaches a normal distribution.

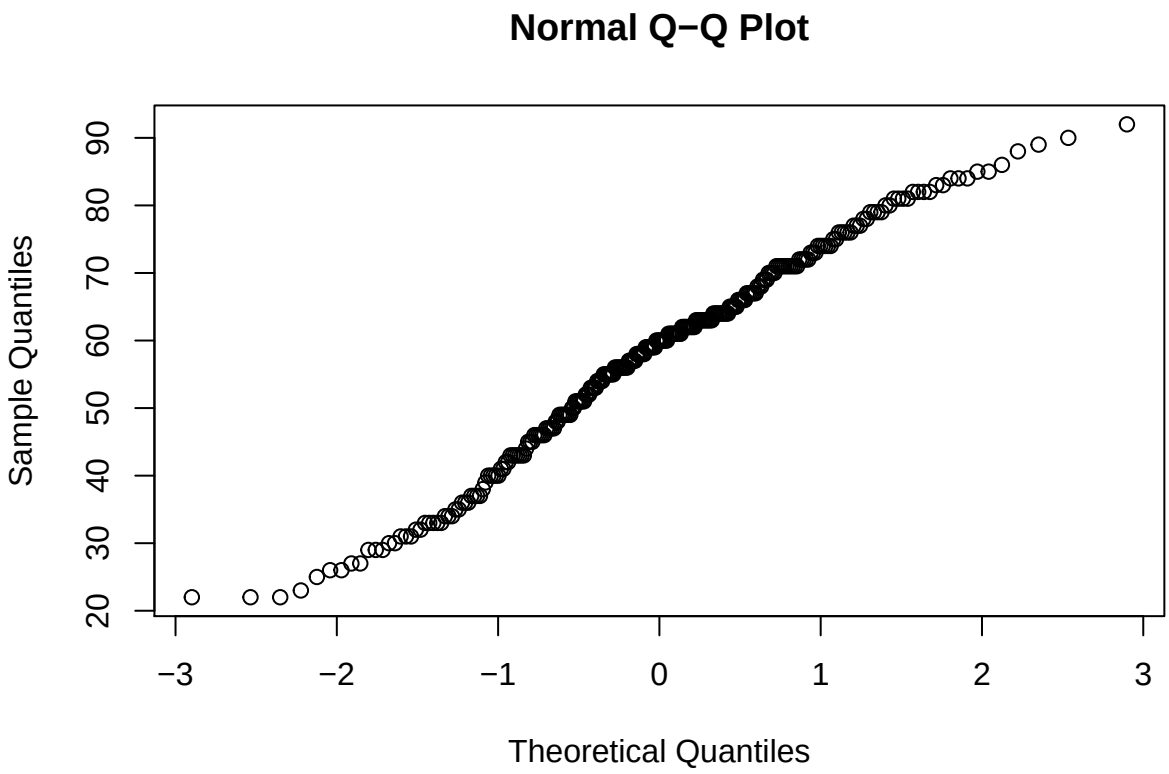
We will draw a histogram of ages of republicans to observe the distribution and any outliers

```r
hist(republicans$age)
```



To see the deviation from normality, we will plot a qqnorm plot

```r
qqnorm(republicans$age)
```



We will observe the skewness of age of republicans

```
skewness(republicans$age)
```

```
## [1] -0.2614682
```

The distribution for age of republicans is not far from normal either, and with the number of observations being well above the thumb rule and the skewness being at a low level (skewness is between -0.5 and 0.5), we can use the Central Limit Theorem to deduce that the distribution of the means approaches a normal distribution.

As discussed earlier the sample is iid drawn. However, we must be cautious in this set with the age as the respondents of lower ages might be more motivated (because of the money they can earn through it) to respond than those of higher ages. It will apply to both the sets - democrats and republicans.

Based on our EDA, selecting an appropriate hypothesis test

We will compare the means for the two samples. However, we should know that the mean alone cannot represent the enture distribution and answer our question. The conclusion thus, must be read with that in mind.

1. Since we are comparing means of two samples, it will be a two-sample test.

2. Since the variable we are comparing is a metric variable, parametric tests can be used.
3. Since the two variables we are looking at are answered by the different people and there is no dependency we can take advantage of, we will do a unpaired test. With all these above reasons, we narrow our choice of test down to an unpaired t-test.

Assumptions to be able to use the unpaired t-test are: 1. Metric variable 2. Data is paired and drawn from i.i.d samples 3. Distribution of the variable is (not too un-)normal

Since all the assumptions are satisfied, we will go ahead with using the unpaired t-test.

The Null Hypothesis we will be testing is: Republican voters are no more older or younger than Democratic voters

We will do a two-tailed tests since we are only interested in the alternative hypothesis: Republican voters are either older or younger than Democratic voters

We will consider the significance level of 0.05 since we are looking at a only a few hundred observations. Since the number of observations is not very large here, we will consider rejecting the null hypothesis if the p-value is <0.05.

Conducting the chosen test

```
t.test(democrats$age, republicans$age, paired=F)
```

```
##
##  Welch Two Sample t-test
##
## data:  democrats$age and republicans$age
## t = -2.3494, df = 580.23, p-value = 0.01914
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.4369362 -0.4857605
## sample estimates:
## mean of x mean of y
##  54.96000  57.92135
```

Since the p-value is less than 0.05, the test is statistically significant to reject the null hypothesis per the significance level we considered before doing the test.

Practical significance:

```
meanage_dem = mean(democrats$age)
sdage_dem = sd(democrats$age)
paste("Mean of Age of Democrats: ", meanage_dem, " Sampling Standard Deviation in Age Democates: ", sda
```

```
## [1] "Mean of Age of Democrats:  54.96  Sampling Standard Deviation in Age Democates:  15.934894276868
```

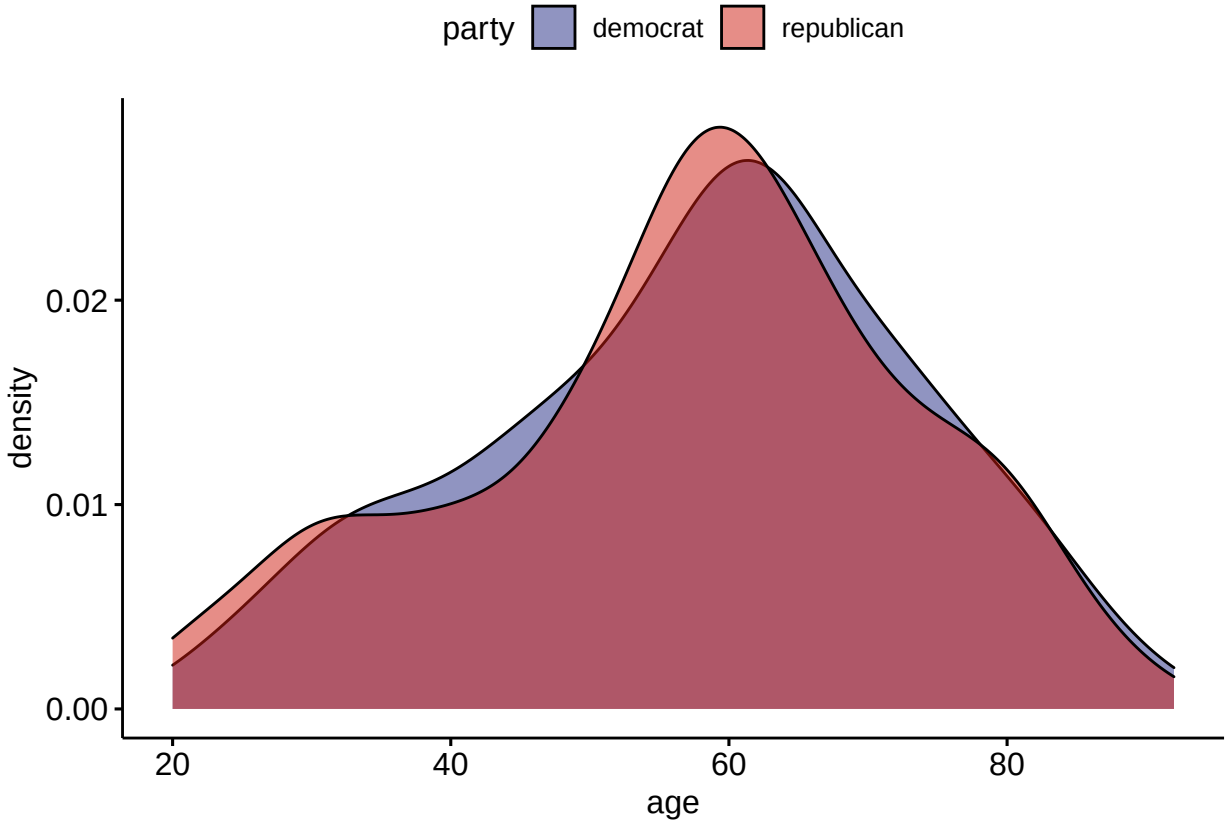
```r
meanage_rep = mean(republicans$age)
sdage_rep = sd(republicans$age)
paste("Mean of Age of Republicans: ", meanage_rep, " Sampling Standard Deviation in Age Republicans: ",
```

```
## [1] "Mean of Age of Republicans:  57.9213483146067  Sampling Standard Deviation in Age Republicans:
```

The difference in means of ages of Democrats and Republicans is 3 with a small difference in sampling standard deviations too (0.3). However, from the graph below of the sampling distribution of the ages of democrats and republicans, we see a substancial area overlapping, so there is practical significance to reject the null hypothesis is not very high.

```r
library(ggpubr)

dems_reps <- usVoters %>%   filter(usVoters$pid1d==2 | usVoters$pid1d==3)
dems_reps$party <- ifelse(dems_reps$pid1d==2,'democrat','republican')
ggdensity(dems_reps, x = "age", fill = "party") +
  scale_fill_manual(values=c("#222a84","#cf1c10"))
```



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

Introducing the topic

Concept: To measure voters' belief about whether federal investigations of Russian election interference are baseless or not, we would use a nominal variable with values indicating true or

false. We will measure this for voters who identify their party inclination as independent.

Operationalization: We will use the 'pid1d' variable that captures the answer to "Generally speaking, do you usually think of yourself as a Democrat, a Republican?" to see if a voter identifies as independent. As mentioned in the earlier question's analysis, this variable takes one of six values. Here, we will be interested in only one of them: 3 which indicates independent.

For the belief about whether federal investigations of Russian election interference are baseless or not, there are two closely associated variables we can use: 1. 'russia16' which captures the answer to question "Do you think the Russian government probably interfered in the 2016 presidential?" 2. 'muellerinv' which captures the answer to question "Do you approve, disapprove, or neither approve nor disapprove of Robert Mueller' [ATTENTION HERE!]"

The 'russia16' variable takes one of three values: -7 No Answer 1 Russia probably interfered 2 This probably did not happen

While this variable does not capture the belief of the federal investigations being baseless or not, it is likely that if a voter things there was inteference (indicated by the variable 'russia16'), they will not find the investigations baseless and vice versa.

The 'muellerinv' variable takes one of the following eight values: -7 No Answer 1 Approve extremely strongly 2 Approve moderately strongly 3 Approve slightly 4 Neither approve nor disapprove 5 Disapprove slightly 6 Disapprove moderately strongly 7 Disapprove extremely strongly

Concerns or Gaps:

1. The data collected is for two questions that are not exactly what we are trying to analyze. They are the ones we think are closely related and informative about the variable we are trying to analyze.

2. Again, one person's interpretation of the Likert scale can be very different from another's. One person's 'Disapprove slightly' could be the same as another's 'Disapprove moderately strongly'.

Exploratory data analysis (EDA) of the relevant variables

Variables: russia16, pid1d

Types: pid1d is a nominal level variable, however, we are only using that to get our sample of independent voters. russia16 is also a nominal level variable. muellerinv is an ordinal variable since the values can be ordered from the level of approval to the level of disapproval.

Number of Entries

```
independents <- usVoters  %>% filter(usVoters$pid1d==3)

paste("Number of Entries for the variable russia16 of independents:", length(independents$russia16))
```

```
## [1] "Number of Entries for the variable russia16 of independents: 232"
```

```
paste("Number of Entries for the variable muellerinv of independents:", length(independents$muellerinv)
```

```
## [1] "Number of Entries for the variable muellerinv of independents: 232"
```

This makes sense because the possible values the value can take are only 1 or 2. From just looking at the mean, it seems that there are more 1s than 2s.

There are no Null Values for the variable russia16 or muellerinv for independents

```
paste("Number of Null Values for the variable russia16 of independents:", length(which(is.na(independen
```

```
## [1] "Number of Null Values for the variable russia16 of independents: 0"
```

```
paste("Number of Null Values for the variable muellerinv of independents:", length(which(is.na(independe
```

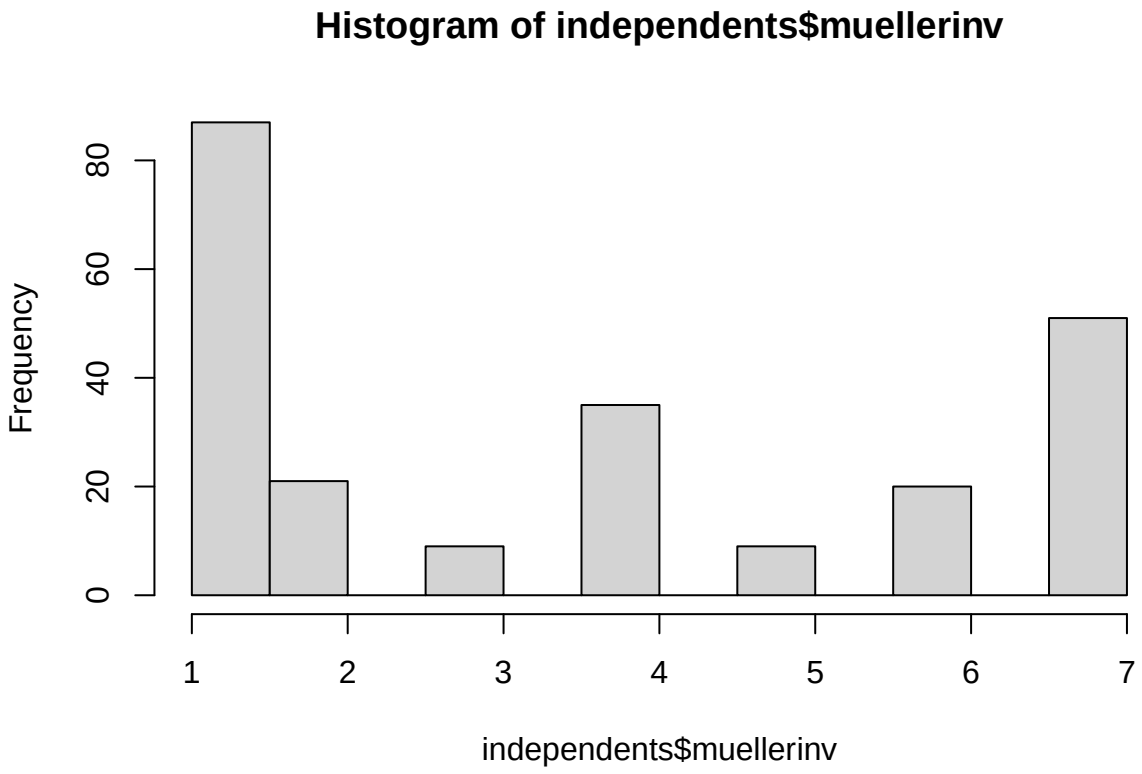
```
## [1] "Number of Null Values for the variable muellerinv of independents: 0"
```

There are no duplicate rows in the dataset of US Voters as we had already established while answering research question 1

We will not evaluate the histogram or skewness of the russia16 variable for independent voters because it is a nominal variable and the histogram or skewness will not inform us much.

We will plot the histogram of muellerinv for independent voters to see it's distribution

```
hist(independents$muellerinv)
```



We will observe the skewness of muellerinv of independent voters

```
skewness(independents$muellerinv)
```

```
## [1] 0.2966592
```

The distribution for muellerinv for independent voters seems symmetric about 4 and the skewness level is at a low level (skewness is between -0.5 and 0.5,).

As discussed earlier the sample is iid drawn.

Since mullerinv seems a closer informative of our question and it has an ordinal level, we will use that for our analysis.

Based on our EDA, selecting an appropriate hypothesis test

cw our choice of test down to the Wilcoxon rank-sum test.

Assumptions to be able to use the Wilcoxon rank-sum test are: 1. Ordinal variable 2. Data is paired and drawn from i.i.d samples

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

The Null Hypothesis we will be testing is: Independent voters are no more likely to believe that the federal investigations of Russian election interference are baselesss than they are to believe that the federal investigations of Russian election interference are not baseless

We will do a two-tailed tests even though we are interested in whether a majority of independent voters believe that the federal investigations of Russian election interference are baseless because: 1. We must choose what kind of test we are going to use before looking at the data. The use of one-tailed tests (which are lesser common) needs to be justified to large extents. We have to be cautious of the fact that it might not be widely accepted, for example, the audience might not believe that we started with a one-tail before looking at the data and think that we changed my test after running it once with two-tail and not getting a statistically significant result. Or the audience in general might not share the same opinion on my justifications for the one-tailed test. 2. With the one-tailed test, it is easier to reject the null hypothesis given all of the rejection region is on one side -> it is therefore associated with more skepticism than a one-tailed test.

We will consider the significance level of 0.05 since we are looking at a only a few hundred observations. Since the number of observations is not very large here, we will consider rejecting the null hypothesis if the p-value is <0.05.

Conducting the chosen test

```
wilcox.test(independents$muellerinv)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  independents$muellerinv
## V = 27028, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to 0
```

Since the p-value is less than 0.05, the test is statistically significant to reject the null hypothesis per the significance level we considered before doing the test.

Practical significance:

```
num_ind = length(independents$muellerinv)
paste("Number of independent voters: ",num_ind)
```

```
## [1] "Number of independent voters:  232"
```

```
approved <- independents %>% filter(independents$muellerinv<4 & independents$muellerinv!=-7)
neutral <- independents %>% filter(independents$muellerinv==4)
disapproved <- independents %>% filter(independents$muellerinv>4)

paste("Ratio of independent voters in the sample who approve of the Mueller Investigation: ",length(appr
```

```
## [1] "Ratio of independent voters in the sample who approve of the Mueller Investigation:  0.50431034
```

```
paste("Number of independent voters in the sample who were nuetral about of the Mueller Investigation:
```

```
## [1] "Number of independent voters in the sample who were nuetral about of the Mueller Investigation:
```

```
paste("Number of independent voters in the sample who disapprove of the Mueller Investigation: ",length
```

```
## [1] "Number of independent voters in the sample who disapprove of the Mueller Investigation:  0.3448
```

This shows there is a higher ratio of independent voters who approve of the Mueller Investigation. We can deduce that they must think the federal investigations of Russian election interference were not baseless. Therefore, there is practical significanct effect we see as well.

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

Introducing the topic

Concept: To measure the increase in turnout from 2016 to 2018, we the population of people who did not turnout in 2016 but did in 2018 since they will make the increase. We can also consider those who got eligible for voting in 2018 by their birth year. To gather which of the emotions, anger or fear, were driving factors for the increased turnout, we would have questions asking the respondents whether any or both of those feelings drove them to turnout to vote in 2018 when they did not in 2016.

Operationalization: Since we are concerned with the increase driven by anger or fear, we can safely discount the respondents who became eligible to vote for the 2018 election and were not eligble for the 2016 election. The variables 'turnout16' and 'turnout18' which capture answers to "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?" and "In the election held on November 6, did you definitely vote in person on election day, vote in person before Nov 6, vote by mail, did you definitely not vote, or are you not completely sure whether you voted in that election?" respectively are what we can use to build the segment for increase as discussed in the concept. We can also include those respondents who mentioned they probably did not vote in 2016 reflected the variable 'turnout16b' which captures the answer to "Do you think you probably voted or probably did not vote?" And for the 2018 election, we can include those you mentioned probably did vote in 2018 reflected in the variable 'turnout18ns' which captures the answer to "If you had to guess, would you say that you probably did vote in the election held on November 6, or probably did not vote in that election?"

There are two variables captured in the ANES 2018 Pilot Study that represent "Generally speaking, how do you feel about the way things are going in the country these days?" with respect to anger and fear. 'geangry' captures "How angry do you feel?" and 'geafraid' captures "How afraid do you feel?". The variables take one of the following six values depending on how intensely they feel the emotion being represented: -7 No Answer 1 Not at all 2 A little 3 Somewhat 4 Very 5 Extremely

Since we are interested in which of the two emotions were more effective in driving turnout, we will select the sample that provided an answer for both the variables.

Concerns or Gaps:

1. The data collected is for the feeling of anger and fear in general, not particularly whether it is effective in driving turnout. The question was not whether anger or fear drove the candidate to turnout in 2018 when they did not in 2016.
2. One person's interpretation of the Likert scale can be very different from another's, so if one person could represent their immense fear as 'very', represented by 4 on the scale, another person could represent the same as 'extremely', represented by 5 on the scale. Similarly for anger as well.

Exploratory data analysis (EDA) of the relevant variables

Variables: turnout16, turnout16b, turnout18, turnout18ns, geangry, geafraid

Types: turnout16, turnout16b, turnout18, turnout18ns are all nominal variables, but we are only using that to get our sample. geangry and geafraid are ordinal variables since can be ordered from the least strong feeling of the emotion to most strong.

Number of Entries

```
increased_voters <- A  %>% filter((A$turnout18<4 | (A$turnout18==4 & A$turnout18ns==1)) & (A$turnout16==

paste("Number of Entries for the variable geangry of increased_voters:", length(increased_voters$geangry
```

```
## [1] "Number of Entries for the variable geangry of increased_voters: 96"
```

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

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

```r
summary(increased_voters$geangry)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.896   4.000   5.000
```

```r
paste("Number of Entries for the variable geafraid of increased_voters:", length(increased_voters$geafra
```

```
## [1] "Number of Entries for the variable geafraid of increased_voters: 96"
```

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

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

```r
summary(increased_voters$geafraid)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   2.885   4.000   5.000
```

There are no Null Values for geangry and geafraid

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

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

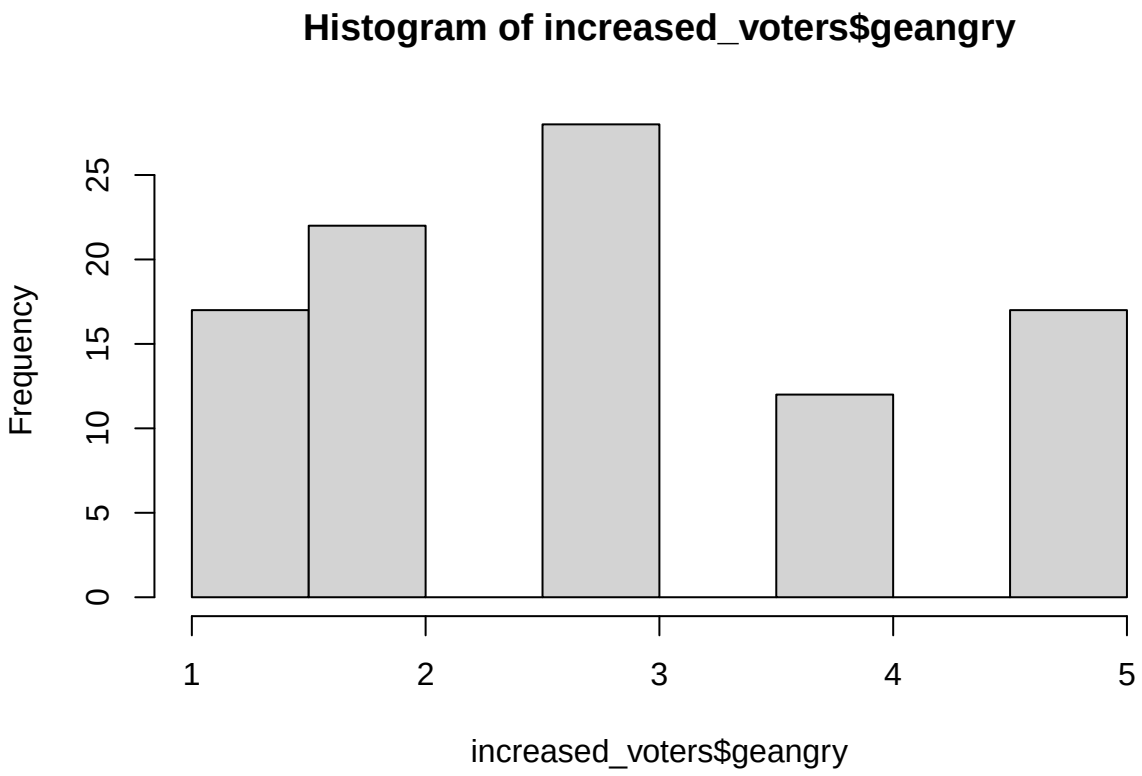
```r
paste("Number of Null Values for geafraid:", length(which(is.na(increased_voters$geafraid))))
```

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

There are no duplicate rows in the dataset of US Voters as we had already established while answering research question 1

Will draw a histogram of both geangry to observe the distribution and any outliers

```r
hist(increased_voters$geangry)
```



We will now observe the skewness of geangry variable

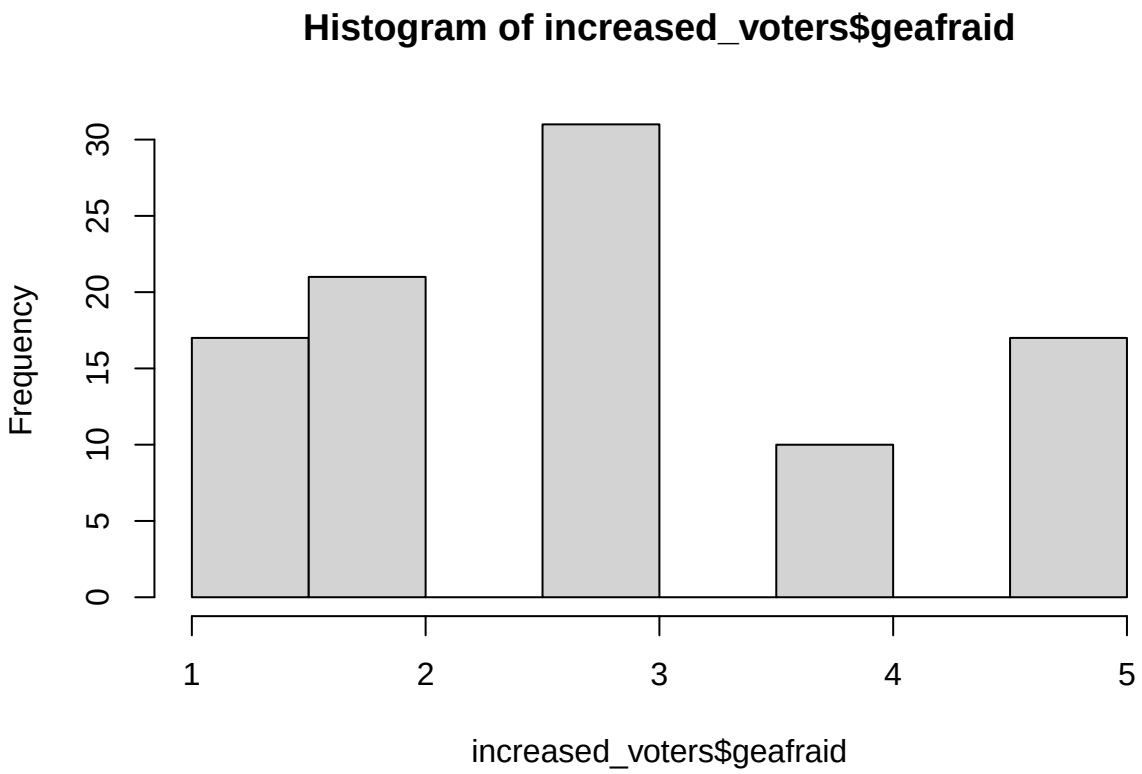
```
skewness(increased_voters$geangry)
```

```
## [1] 0.1884442
```

We see a symmetry about the median (3) in the distribution for geangry and the skewness level is also low (skewness is between -1 and -0.5 or between 0.5 and 1).

Will draw a histogram of both geafraid to observe the distribution and any outliers

```
hist(increased_voters$geafraid)
```



We will now observe the skewness of geafraid variable

```
skewness(increased_voters$geafraid)
```

```
## [1] 0.2084681
```

Again, we see a symmetry about the median (3) in the distribution for geafraid and the skewness level is also low (skewness is between -1 and -0.5 or between 0.5 and 1).

As discussed earlier the sample is iid drawn.

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. 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: Anger is no more or less effective than fear at driving increases in voter turnout from 2016 to 2018

We will do a two-tailed test since we are only interested in the alternative hypothesis: There is a difference in the effectiveness of anger and fear at driving increases in voter turnout from 2016 to 2018

We will consider the significance level of 0.05 since we are looking at a only a few hundred observations. Since the number of observations is not very large here, we will consider rejecting the null hypothesis if the p-value is <0.05.

Conducting the chosen test

```
more_anger = sum( increased_voters$geafraid < increased_voters$geangry)
trials = sum( increased_voters$geafraid < increased_voters$geangry | increased_voters$geafraid > increas
binom.test(more_anger , trials)
```

```
##
##  Exact binomial test
##
## data:  more_anger and trials
## number of successes = 24, number of trials = 49, p-value = 1
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.3442474 0.6366217
## sample estimates:
## probability of success
##              0.4897959
```

The p-value is greater than 0.05 indicating that we cannot reject the null hypothesis that Anger is no more or less effective than fear at driving increases in voter turnout from 2016 to 2018.

Practical significance: We see that the probability of increased voters having anger be more represented than fear comes up as 0.49 from the Sign Test. This also indicates that the practical significance of the effect is very low.

Question 5: Select a fifth question that you believe is important for understanding the behavior of voters

Clearly argue for the relevance of this question. (10 points)

In words, clearly state your research question and argue why it is important for understanding the recent voting behavior. Explain it as if you were presenting to an audience that includes technical and non technical members.

Explain how your variables are operationalized. Comment on any gaps that you can identify between your operational definitions and the concepts you are trying to study.

Perform EDA and select your hypothesis test (5 points)

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

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.

Based on your EDA, select an appropriate hypothesis test. Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Conduct your test. (2 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result.

Conclusion (3 points)

Clearly state the conclusion of your hypothesis test and how it relates to your research question.

Finally, briefly present your conclusion in words as if you were presenting to an audience that includes technical and non technical members.