W203 Lab 1: Comparing Means | Section 3

Srishti Mehta | Andi Morey Peterson | David Djambazov

10/15/2020

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. You are provided with data from the 2018 ANES Pilot Study.

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the ANES User's Guide and Codebook.

It is important to consider the way that the ANES sample was created. Survey participants are taken from the YouGov panel, which is an online system in which users earn rewards for completing questionnaires. This feature limits the extent to which results generalize to the U.S. population.

To partially account for differences between the YouGov panel and the U.S. Population, ANES assigns a survey weight to each observation. This weight estimates the degree to which a citizen with certain observed characteristics is over- or under-represented in the sample. For the purposes of this assignment, however, you are not asked to use the survey weights. (For groups with a strong interest in survey analysis, we recommend that you read about R's survey package. We will assign a very small number of bonus points (up to 3) to any group that correctly applies the survey weights and includes a clear explanation of how these work).

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

Following is an example of a question asked on the ANES survey:

> *How difficult was it for you to vote in this last election?*

The variable `votehard` records answers to this question, with the following encoding:

- -1 inapplicable, legitimate skip
- 1 Not difficult at all
- 2 A little difficult
- 3 Moderately difficult
- 4 Very difficult
- 5 Extremely difficult

To see the precise form of each question, take a look at the Questionnaire Specifications.

Assignment

You will use the ANES dataset to address five research questions. For each question, you will need to operationalize the concepts (selecting appropriate variables and possibly transforming them), conduct exploratory analysis, deal with non-response and other special codes, perform sanity checks, select an appropriate hypothesis test, conduct the test, and interpret your results. When selecting a hypothesis test, you may choose from the tests covered in the async videos and readings. These include both paired and unpaired t-tests, Wilcoxon rank-sum test, Wilcoxon signed-rank test, and sign test. You may select a one-tailed or two-tailed test.

Submission Guidelines

- Please organize your response according to the prompts in this notebook.
- Note that this is a group lab and your instructor will assign you to your team.
- Please limit your submission to 5000 words, not counting code or figures.
- Submit *one* report per group.
- Submit *both* your pdf report as well as your source (rmd) file.
- **Only analyses and comments included in your PDF report will be considered for grading.**
- Include names of group members on the front page of the submitted report.
- Naming structure of submitted files:
  - PDF report: [student_surname_1]_[student_surname_2][_*]_lab_1.pdf
  - R-markdown: [student_surname_1]_[student_surname_2][_*]_lab_1.rmd

David's survey generalization comments:

1. How well we can generalize the ANES sample to the entire population of the US, consequently our sub-sample to the population of US voters? There's an in-depth treatment of the topic in the dataset's introduction, including a discussion of the possible use of weights.

2. What is the effect of the opt-in nature of the survey? That is a complex topic that for now we will leave outside our analysis and make the assumption that opting-in is independent of the respondent variables we are interested in.

Research Questions

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

Introduce your topic briefly. (5 points)

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.

David's comments:

1. *Question:* Two relevant questions in the questionnaire are "How do you rate the police?" and "How do you rate journalists?". The answers are encoded as `ftpolice` and `ftjournal`. The method of collecting the answers is through a *thermometer widget* that allows the respondent to click on a "sentiment temperature" ranging in discrete integers from 0 to 100, 100 being "Very warm or favorable feeling", 0 being "Very cold or unfavorable feeling". -7 corresponds to No Answer. There are "-7" values in `ftjournal`, but not in `ftpolice`. Importantly, both variables are ordinal.

2. *Population:* The question is targeting the population of US voters. While there are a few different possible definitions (e.g. voter registration), since the survey was conducted in the immediate aftermath of the 2018 election, we can argue that the most solid definition of US voters for the purposes of this question are persons who have actually cast a vote in that election. The relevant field that indicates who among our sample are from the population of US voters, is then `turnout18`.

The field `turnout18` has 5 categorical responses numbered 1-5: 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

1-3 are clearly in the sample population, while 4 is definitely not. For responses 5, there is an additional field `turnout18ns`, matching the number of responses in `turnout18`, with categories as follows: -1 inapplicable, legitimate skip 1 Probably did vote 2 Probably did not vote

3. *Gaps:*

- To what extent an answer to "How do you rate x?" can be mapped to a measurement space of "respect for x"? Let's first try to answer the question "Do US voters rate higher the police or journalists?" and then discuss if that's sufficient to gauge respect.

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.

David's prelim EDA:

Extracting the right sample:

As we saw in the population discussion above, observations with values 1-3 from `turnout18` should be part of our sample. The question is what to do with the observations with value 5 (Not completely sure) in `turnout18` and 1 (Probably did vote) in `turnout18ns`.

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

Definitely Voted:

```
sum(A$turnout18 < 4)
```

```
## [1] 1842
```

Not completely sure & Probably did vote:

```
sum(A$turnout18 == 5 & A$turnout18ns == 1)
```

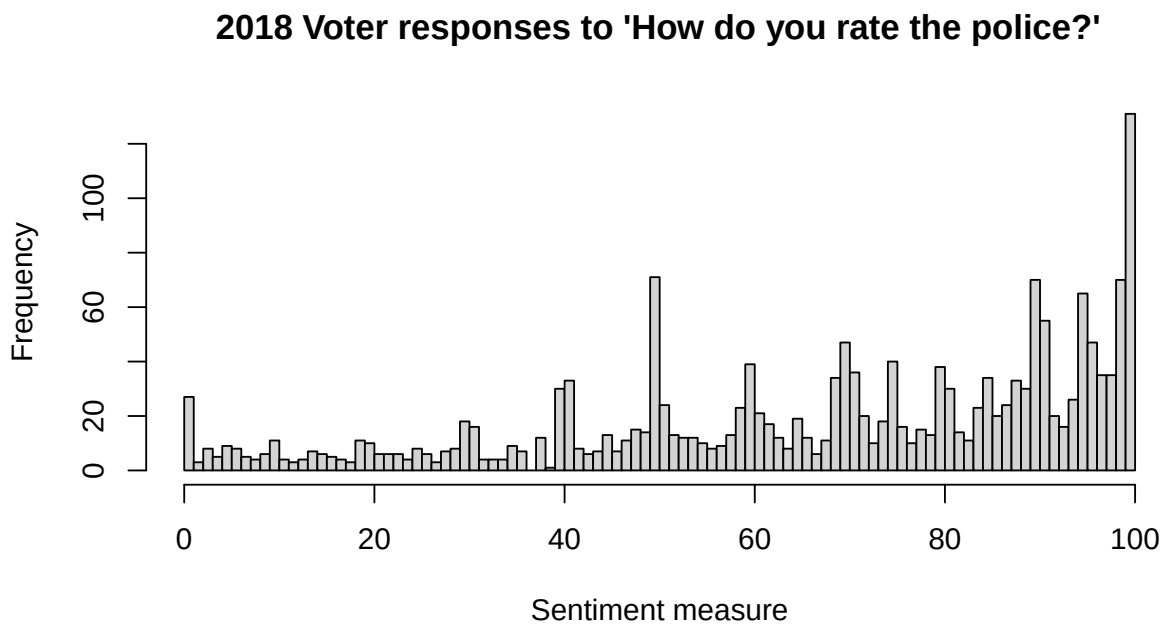
```
## [1] 18
```

About 1%. We can argue that a. that is a fairly small number, and b. 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 voters.

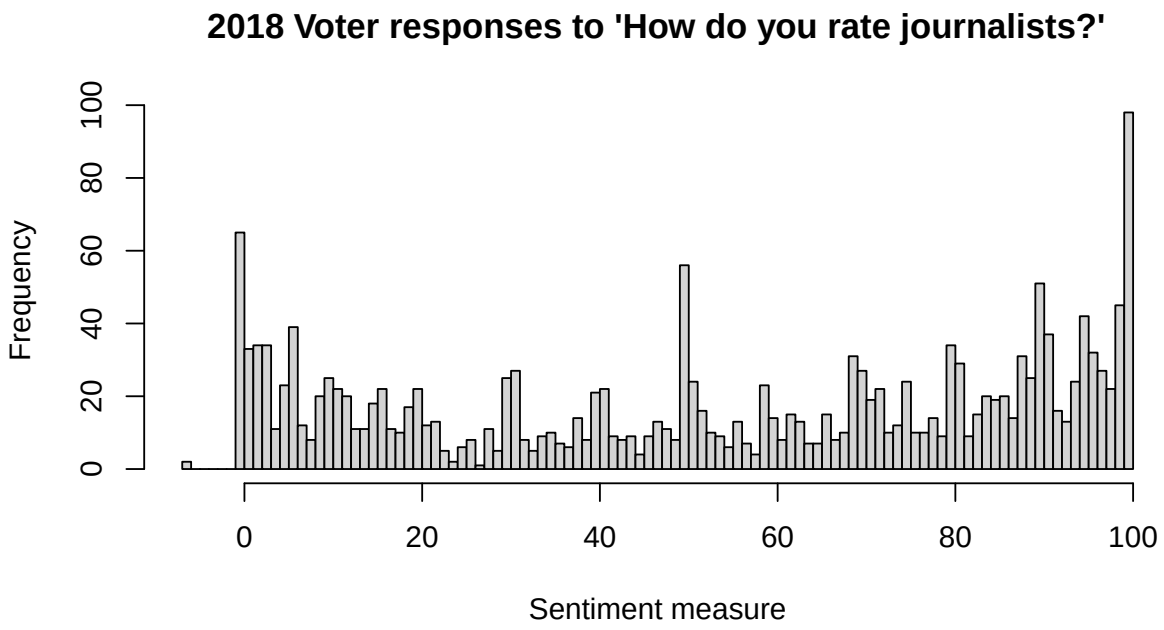
```
voter_sample = A[ which(A$turnout18 < 4),]
```

Histograms of the data

```
hist(voter_sample$ftpolice, breaks = 100,
     main = "2018 Voter responses to 'How do you rate the police?'",
     xlab = "Sentiment measure")
```

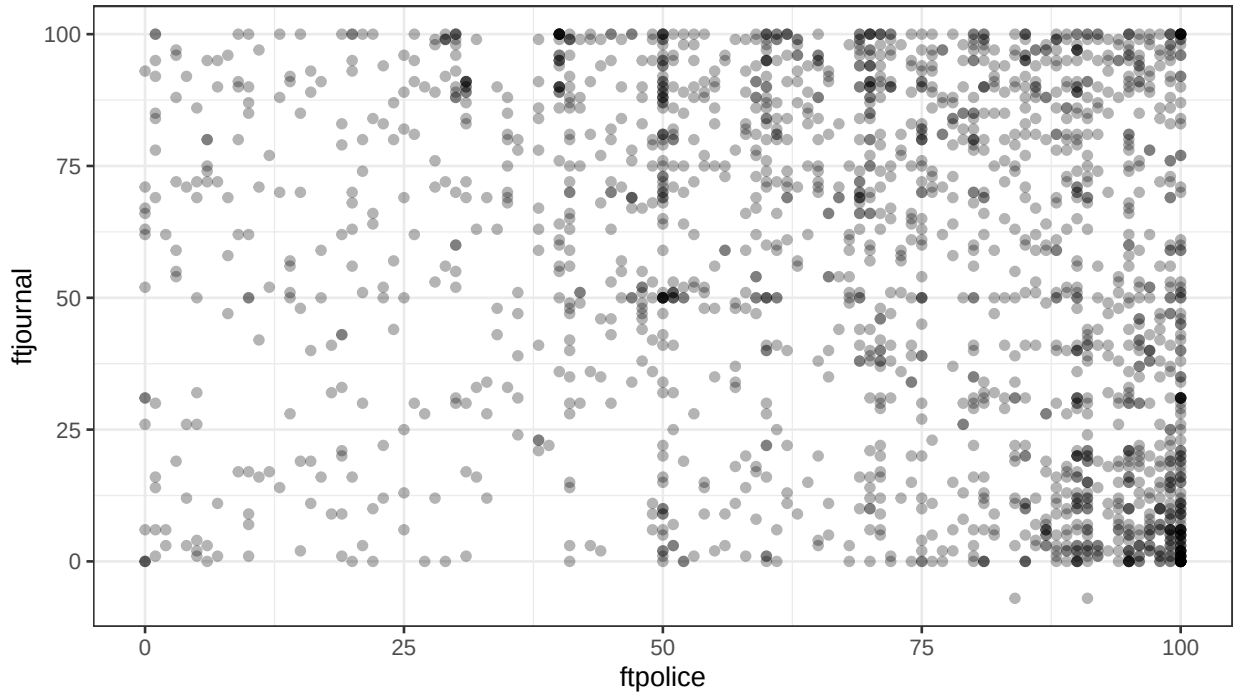



```
hist(voter_sample$ftjournal, breaks = 100,
     main = "2018 Voter responses to 'How do you rate journalists?'",
     xlab = "Sentiment measure")
```



Scatter Plot

```
ggplot(voter_sample, aes(x=ftpolice, y=ftjournal) ) +
  geom_point(alpha = 0.3) +
  theme_bw()
```



"No Answer" responses:

```r
sum(A$ftjournal == -7)
```

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

EDA Discussion:

The histograms of the two sample features reveal similar pictures. There are prominent peaks at the minimum value(0), the maximum value(100) and the middle value (50). The "-7" or No answer" responses in `ftjournal` are 2 out of 2500, so shouldn't have much effect on the analysis.

In terms of the relationship between the variables, from the scatter plot we see there isn't a clear cut relationship. If anything the variable look quite uncorrelated. However, reducing the opaqueness of the `geom_point` method reveals some interesting structure. Specifically, there are three general areas of concentration: 1. observations with high values for `ftpolice` and low values for `ftjournal`, 2. observations with high values for `ftjournal` and medium to high values for `ftpolice`, 3. observations with middle values for both `ftpolice` and `ftjournal`.

The spaces in between these three regions and in particular the combination of low values for `ftpolice` and high values for `ftjournal`, and low values for both seem to be less frequently observed. Since we're also getting a pair of values for each respondent (the unit of observation), we have paired data, which justifies using a paired test.

It is also reasonable to argue in favor of dropping the two observations that have "No Answer" (-7) in `ftjournal`, so that all pairs have both values.

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

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Data discussion: Since our sample's variables are ordinal and in addition exhibit some dependence, we would argue that using a sign test is the most appropriate way to test a hypothesis based on this data.

Null Hypothesis: The hypothesis we want to test is that Americans rate the police and journalists equally.

Test: Paired sign test.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

```r
q1_data <- voter_sample[which(voter_sample$ftjournal != -7),]
more_police <- sum( q1_data$ftjournal < q1_data$ftpolice)
trials <- sum( q1_data$ftjournal < q1_data$ftpolice | q1_data$ftjournal > q1_data$ftpolice)
binom.test(more_police, trials)
```

```
##
##  Exact binomial test
##
## data:  more_police and trials
## number of successes = 1006, number of trials = 1793, p-value =
## 2.544e-07
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
```

```
##  0.5377376 0.5842040
## sample estimates:
## probability of success
##              0.5610708
```

Conclusions: We have a statistically significant test with a p-value on the order of e-07. So we reject the null hypothesis that Americans rate the police and journalists equally. By extension, we reject the null hypothesis that Americans respect the police and journalists equally.

The practical significance can be gauged by considering that among respondents who express a difference, 56.1% rate the police higher than they do journalists. In terms of correlation we have:

```
(1006 - (1793 - 1006))/1793
```

```
## [1] 0.1221417
```

That's a pretty weak correlation (consistent with the intuition gained from the scatter plot), so we can say that while the data support the presence of a difference, it is not a very strong effect.

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

Introduce your topic briefly. (5 points)

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.

David's comments:

1. *Question:* Since the question asks for a straight comparison between the mean of a metric variable (age), here the only relevant field is `birthyr`, containing 74 unique values from 1927 to 2000.
2. *Population:* From question 1. we already have a definition of the sample of US voters. Here we need to further define of that sample, which observations can be attributed to "Republican" voters and which to "Democratic" voters.

One possibility would be to use the field about self-determination `pid1d`. Another would be to use `pid7x` (Party ID summary). However, it seems that the practical significance of these definitions would be rather limited as, on one hand, most voters do not self-define themselves as either and, on another, these identifications cannot guarantee actual voting behavior. Thus, it would be more interesting and relevant to compare the populations of those who have actually voted for the Republicans and the Democrats in the 2018 election. Let's see how we might parse that out.

In the 2018 election, there were three types of votes that have been captured by the survey: 1. For US House candidate: `house18p` -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3 something else

2. For US Senate candidate: `senate18p` -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3 another party 4 two different parties
3. For State governor: `gov18p` -7 No Answer -1 inapplicable, legitimate skip 1 Democrat 2 Republican 3 another party

Despite recent political polarization, a fair number of voters do "split the ticket" and vote for candidates from different parties for different offices, we have a few options how to define Democratic and Republican voters. The most restrictive definition would be observations that have cast one or more votes for Democrats, but none for Republicans and vice-versa.

A less restrictive approach would be to apply this criteria only to the US House and Senate elections and ignore the vote for governor. That might be wise considering that a number of heavily leaning Democratic states do have Republican governors (Massachussetts) and vice-versa (Kentucky). Under this definition splitting the ticket in the House and Senate races as well as voting for two different parties in the Senate race (in the case there were 2 Senate races in the same state) would indicate an independent voter.

3. *Gaps:*
4. Since exact age is not available and the variable that will be used for the test is the less granular birth year (`birthyr`), it could make resolving the two populations a bit harder, potentially leading to a failure to reject a small difference in age.
5. As discussed the result of the test is affected by our definition of Republican and Democratic voters.
6. Since we're looking at age, the opt-in and electronic nature of the survey could introduce a non-zero covariance between age and participation in the survey. If age is not independent of voting preference (which is exactly what we're trying to test for), that could directly affect the result of our test. Once again, that discussion is beyond the scope of the current analysis, so we'll operate under the assumption that any such effect can be neglected.

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.

David's prelim EDA:

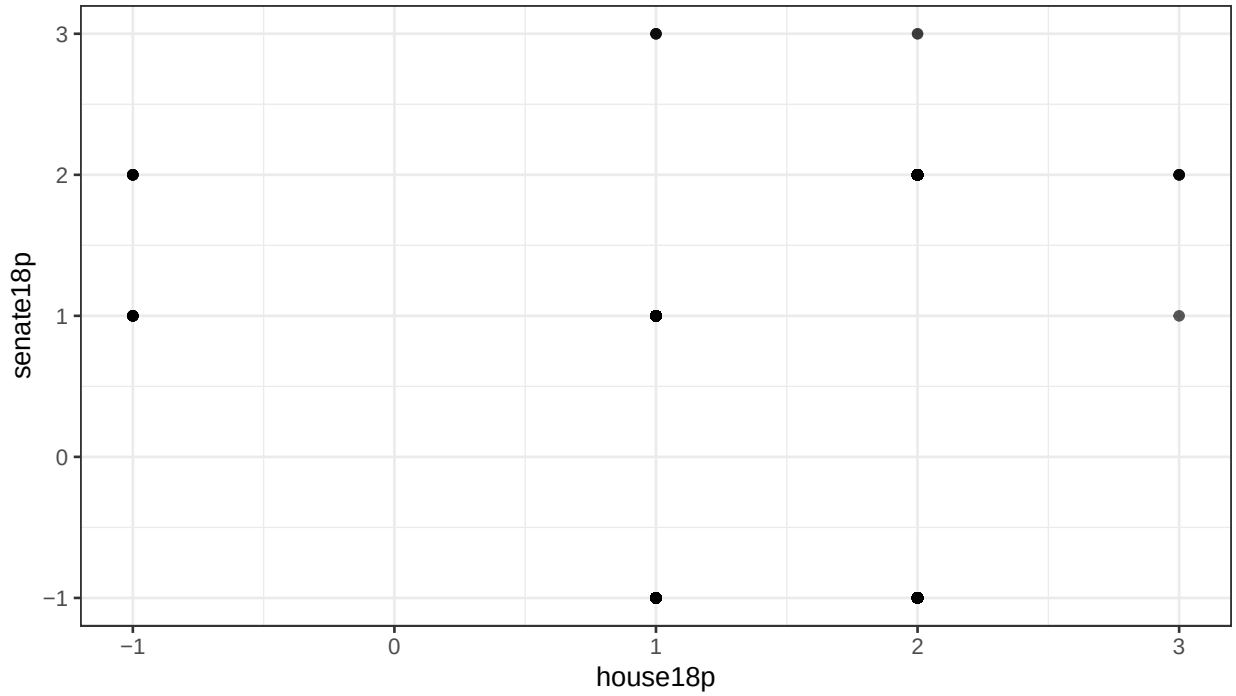
Winnowing the data: As discussed above, we will extract observations from the `voter_sample` from Question 1 for which one or more of the votes are for one of the parties and none are for the other.

```
partisan_voters <- voter_sample %>%
  filter(!((house18p != 1 & house18p != 2 &
              senate18p != 1 & senate18p != 2) |
              senate18p == 4)) %>%
  filter(!(house18p == 1 & senate18p == 2)) %>%
  filter(!(house18p == 2 & senate18p == 1))

partisan_voters$partisan <- factor(
  ifelse((partisan_voters$house18p == 1 | partisan_voters$senate18p == 1),
  "dem",
  "gop"))
```

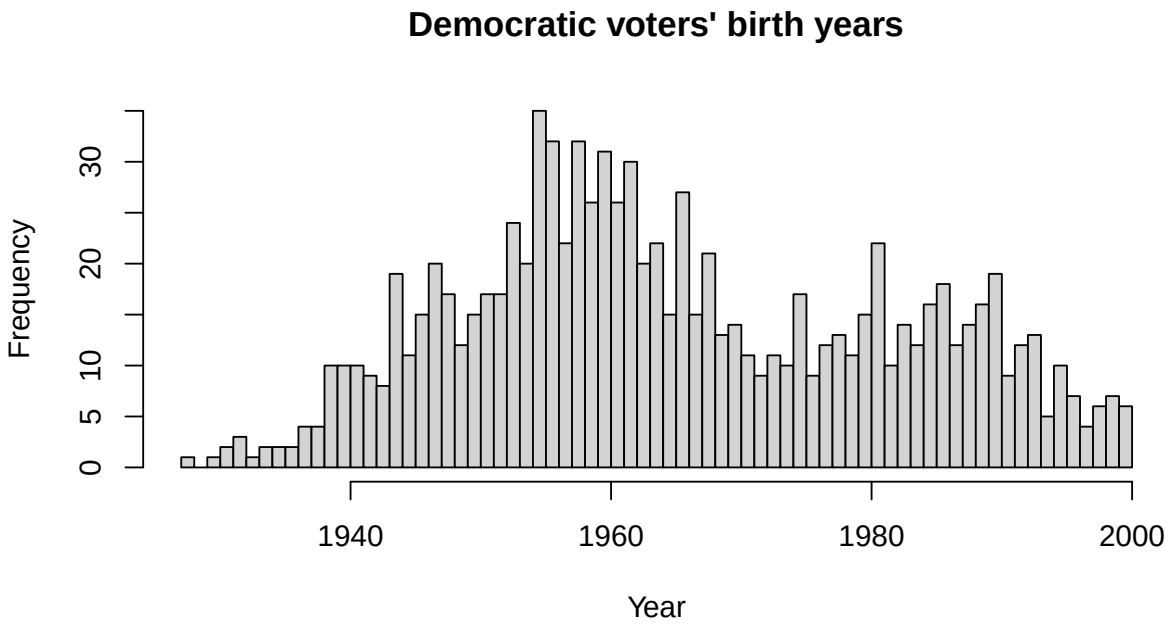
Let's do a quick visual check.

```
ggplot(partisan_voters, aes(x=house18p, y=senate18p) ) +
  geom_point(alpha = 0.3) +
  theme_bw()
```

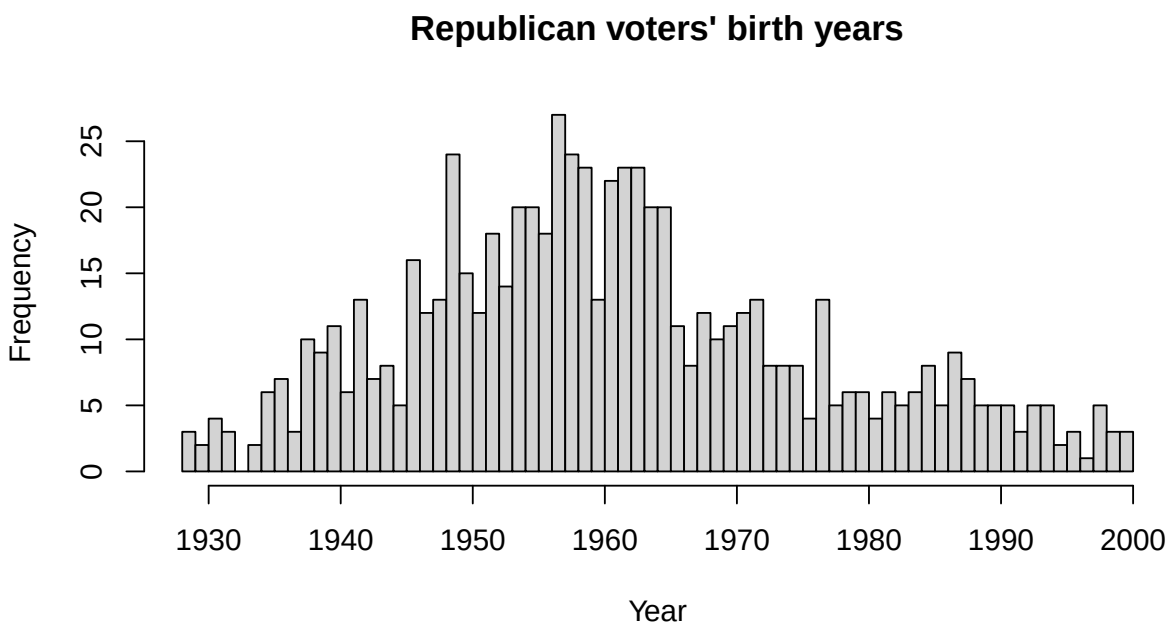


Histograms of the data

```
hist(partisan_voters[partisan_voters$partisan == "dem",]$birthyr, breaks = 100,
     main = "Democratic voters' birth years",
     xlab = "Year")
```



```r
hist(partisan_voters[partisan_voters$partisan == "gop",]$birthyr, breaks = 100,
     main = "Republican voters' birth years",
     xlab = "Year")
```



Both distributions seem very reasonable. The number of data points is also very robust for both populations.

```r
cat("Dem voters:", sum(partisan_voters$partisan == "dem"), ";",
      "GOP voters:", sum(partisan_voters$partisan == "gop"))
```

```
## Dem voters: 987 ; GOP voters: 706
```

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

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Data discussion: The measured variable `birthyr` is cardinal and is a single value for each unit of observation (respondent). In the discussion of the survey's design we have established validation for the i.i.d. variable assumption. We also see in the EDA that the sample distribution is well behaved and given the robust number of data points, we are justified in relying on the CLT to conduct a t-test to compare the mean birth years of the two samples (Democratic and Republican voters).

Null Hypothesis: The hypothesis we want to test is that Democratic voters and Republican voters were born at the same time as measured in years.

Test: Two sample t-test.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

```r
t.test(birthyr ~ partisan, data = partisan_voters)
```

```
##
##  Welch Two Sample t-test
##
## data:  birthyr by partisan
## t = 5.6633, df = 1562.7, p-value = 1.764e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  2.877915 5.927720
## sample estimates:
## mean in group dem mean in group gop
##          1965.675          1961.272
```

Conclusions: We have a statistically significant test with a t-statistic of 5.66, so we reject the null hypothesis. On average, Republican voters are older than Democratic voters.

In terms of practical significance the difference in mean birth year constitutes a good measure. Coming in at 4.4 years, it is perhaps not a generational difference, but meaningful never-the-less.

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)

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.

David's comments:

1. *Question:* The 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.
2. *Population:* We are looking to answer the question for the population of independent voters. At this point of the analysis, the most logical definition that can help us obtain a sample of that population would be observations that are in `voter sample` from Question 1, but not in `partisan_voter` from Question 2.
3. *Gaps:*

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.

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

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

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

Introduce your topic briefly. (5 points)

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.

1. *Question:* In terms of measuring feelings of fear and anger, the two relevant fields are 2 of the specific responses to the general question "Generally speaking, how do you feel about the way things are going in the country these days?" `geangry` ("How angry do you feel?") and `geafraid` ("How afraid do you feel?"). Both of those are ordinal variables ranging from 1 (Not at all) to 5 (Extremely).
2. *Population:* Since the question is about an increase of voter turnout, what we are interested in is the population of voters that could have voted in 2016 but did not, given that they did vote in 2018. So we can start with our `voter_sample` 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.
3. *Gaps:*

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.

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

Explain why your test is the most appropriate choice. List and evaluate all assumptions for your test.

Conduct your test. (5 points)

Explain (1) the statistical significance of your result, and (2) the practical significance of your result. Make sure you relate your findings to the original research question.

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.