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 responsants:

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)

General Study Comments (applicapble 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 min 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 not mentioned – the data did not account for people who are ineligable to vote to use a felony).

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

Conceptualize

The concept is in the question itself. Do US voters prefer police over journalists?

Operationalize

We could have directly asked the survey respondents to pick on or the other, but the survey didn't do this. Rather is other data we can use to operationalize this question. First there is the rating question in which respondents were asked to rate different groups such as police and journalists. "How do you rate the police?" and "How do you rate journalists?" These results reside in the variables $ftpolice$ and $ftjournal$. There is also the question "How concerned are you about violence against people who work in the news media?" However, since there is not a corresponding question about police, we cannot use that in any two-sample test.

In addition to this, the question asks about voters. The survey provides a few sample questions about the respondent to determine if that respondent was a voter. Variables $turnout18$ and $turnout18ns$ can be used to determine if they were a voter.

Exploratory Data Analysis

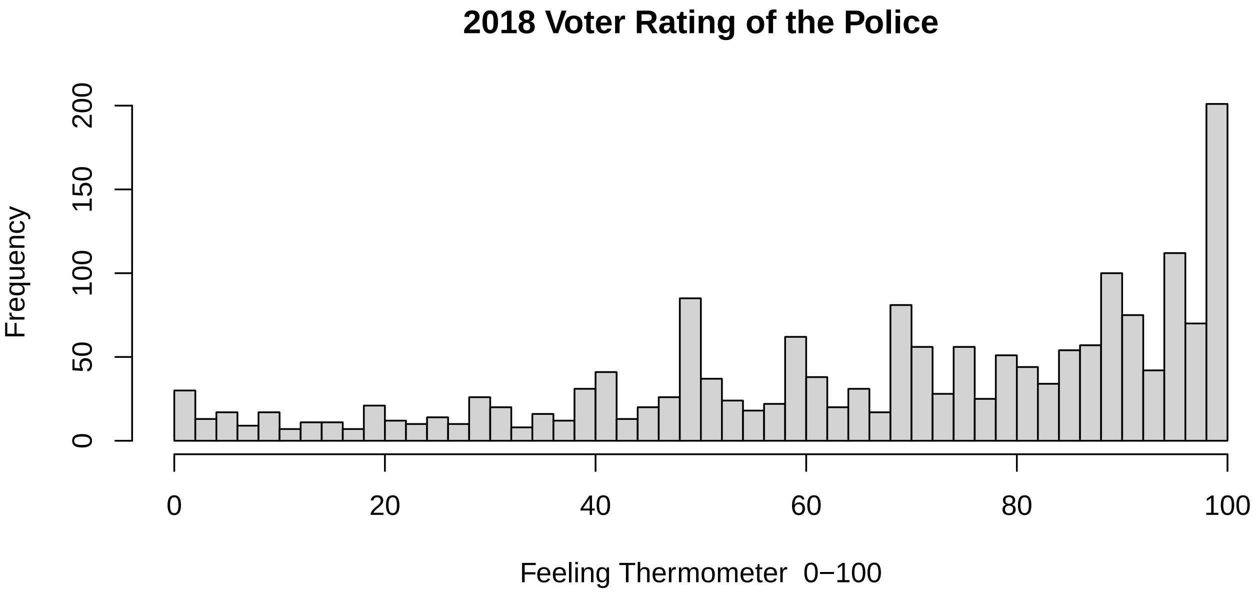
First, let's determine if we have enough "US Voters".

```
A$likelyvoter <-
  ifelse((A$turnout18<=3), 1,
         ifelse((A$turnout18==5)&&(A$turnout18ns=1),1,0))
paste("Number of Definite and Probable Voters:", sum(A$likelyvoter==1))
```

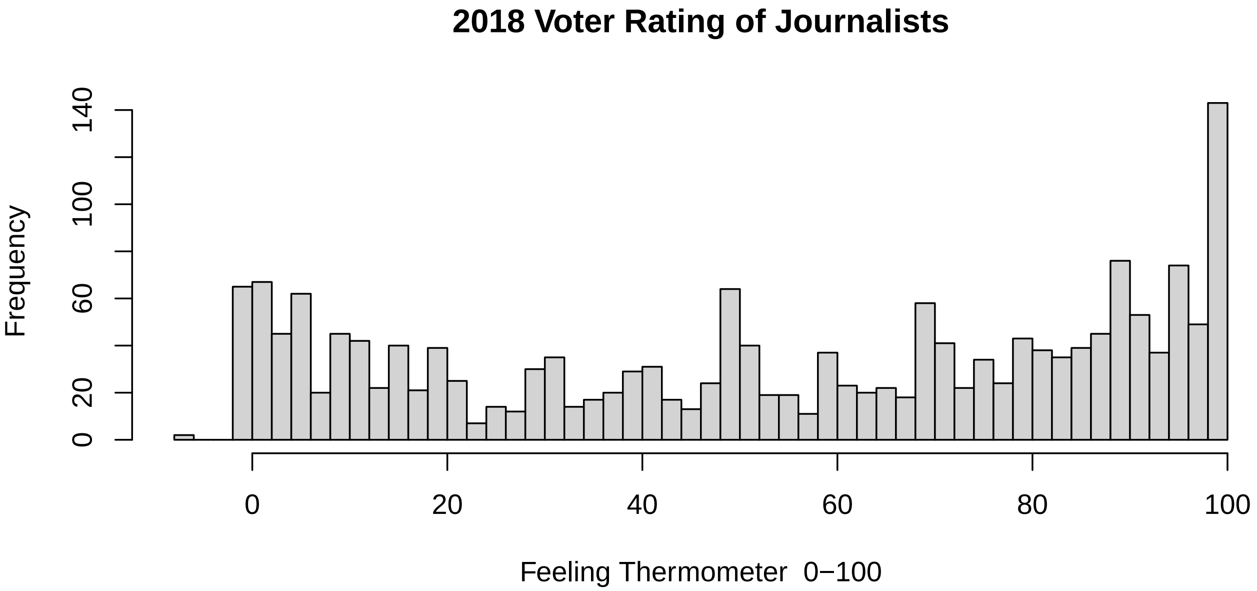
```
## [1] "Number of Definite and Probable Voters: 1842"
```

We have 1842 people who claimed they definitely or probably voted in which we will use to filter the other two variables $ftpolice$ and $ftjournal$. Now we need to see if the sample distribution of $ftpolice$ and $ftjournal$ are normal enough to use CLT.

```
filtered_voter18<-select(A, c("likelyvoter", "ftpolice", "ftjournal"))
filtered_voter18<-filter(filtered_voter18, likelyvoter==1)
hist(filtered_voter18$ftpolice, breaks = 50,
     main = "2018 Voter Rating of the Police",
     xlab = "Feeling Thermometer  0-100")
```



```
hist(filtered_voter18$ftjournal, breaks = 50,
     main = "2018 Voter Rating of Journalists",
     xlab = "Feeling Thermometer  0-100")
```



The histograms reveal that on both the police and journalists there are spikes at 100 and are not normal Because we will be comparing the means of both of these distributions, and because the number of observations are large enough at over 1800 it isn't a concern; we can rely on CLT. There are 2 appearances of -7 for the second variable, but since there's only 2 it shouldn't affect the data.

The Test

Since our sample's variables are ordinal, i.i.d (see general assumptions on page 1) and the data can be paired, we would argue that using a sign test is the most appropriate way test. Therefore the *Null Hypothesis* is that the difference in how 2018 US Voters rate police and how they rate journalists is equal to 0.

$$\mu_{ftpolice} - \mu_{ftjournal} = 0$$

```r
t.test(filtered_voter18$ftpolice, filtered_voter18$ftjournal)
```

```
##
##  Welch Two Sample t-test
##
## data:  filtered_voter18$ftpolice and filtered_voter18$ftjournal
## t = 13.857, df = 3507, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  12.08198 16.06460
## sample estimates:
## mean of x mean of y
##  68.49131  54.41802
```

Because the p-value is so low at $p < 2.2e^{-16}$, this test is highly significant and we can reject the null hypothesis that US Voters rated journalists and police the same. We can compare the means to understand the effect size. The mean rating for police is 68.491, while the mean rating for journalists is 54.418, so US voters, on average, rate the police 14 “temperature points” higher than journalists.

2. Are Republican voters older or younger than Democratic voters?

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

First, lets determine if we have enough Democrats and Replublicans. We will have to awknowlege that using only reponses 1 or 2 for *pid1d* and *pid2d* will remove non-responses, Indpendants, and “others” from the analysis.

```r
Age_and_Party <-select(A, c("pid1d", "pid2d", "birthyr"))
Age_and_Party$Party <-
  ifelse((Age_and_Party$pid1d==2 | Age_and_Party$pid2d == 2), 0,
  ifelse((Age_and_Party$pid1d==1 | Age_and_Party$pid1d == 1), 1, -1))
Age_and_Party<-filter(Age_and_Party, Party >= 0)
paste("Number of Republicans: ", sum(Age_and_Party$Party==0))
```

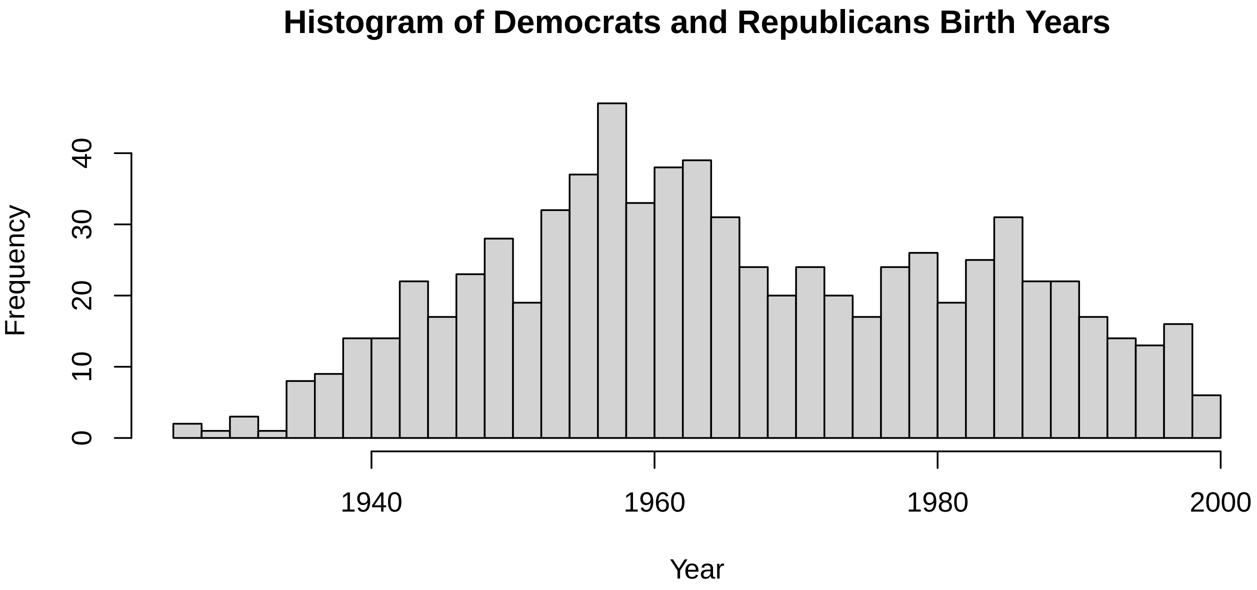
```
## [1] "Number of Republicans:  326"
```

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

```
## [1] "Number of Democrats:    432"
```

We have 758 people who claimed they are part of either party, Democrates and Republicans. Now we need to see if the sample distribution of their ages are normal enough to use CLT.

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



The historgram of ages has a somewhat bi-modal distribution however, with 758 reponses, we can rely on CLT. There are 2 appearances of -7 for the second variable, but since there's only 2 it shouldn't affect the data.

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 *Null Hypothsis* 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 = -1.2146, df = 691.37, p-value = 0.2249
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.8702977  0.9119644
## sample estimates:
## mean in group 0 mean in group 1
##        1965.500        1966.979
```

Here we do not have a statistically significant test is the p-value being near 0.3. We fail to reject the null hypthosis that there is an age difference between Democrats and Republicans.