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) 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 not mentioned – the data did not account for people who are ineligable to vote due to a felony).

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 conserative approach in considering US voters.

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

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

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

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

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

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

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

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

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

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

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

```
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 specfically 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.