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.

Voter sample calculation:

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

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.

Variables 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.

The key question is what does it mean "to believe that the federal investigations of Russian election interference are baseless"? My 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 `voter_sample` 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)

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 pull out our sample.

```
ind_voters2 <- voter_sample %>%
  filter(pid7x == 3 | pid7x == 4 | pid7x == 5)
```

And look at it.

```
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)

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

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 woule be interested in a statistically significant result in wither direction, we'll run a two-tailed 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.

```
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.

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 waranted and that motivates them to vote against the party that benefitted from the alledged 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?

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.

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 those.

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 that the turnout of the 2016 election. The reason is that 2016 was a presidential electon 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.

```
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 eleigible 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.

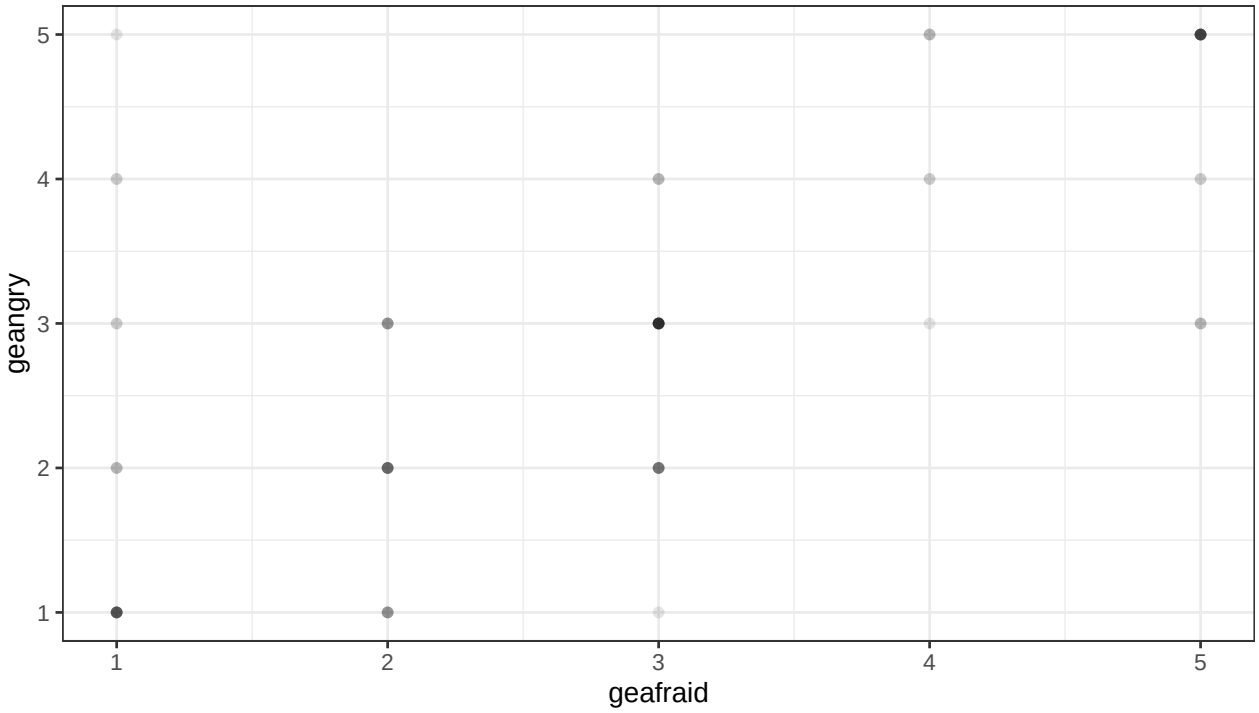
```
to_increase <- to_increase %>%
  filter(geafraid > 0) %>%
  filter(geangry > 0)

non_vote <- non_vote %>%
```

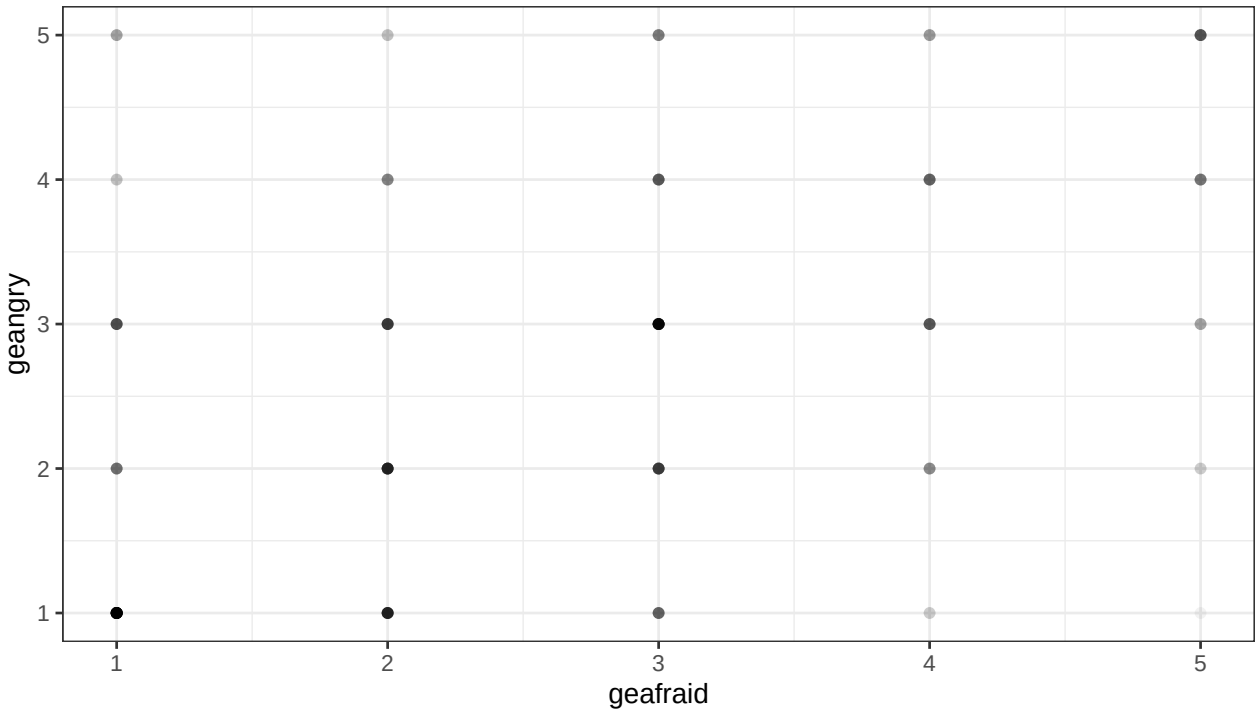
```
  filter(geafraid > 0) %>%
  filter(geangry > 0)
```

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)

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

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.