Lab 1: Hypothesis Testing

w203 Teaching Team

Overview

The American National Election Studies (ANES) conducts surveys of voters in the United States, with a flagship survey occurring immediately before and after each presidential election. While the post-election data for 2020 is not yet available, pre-election data is available as a preliminary release. In this lab, you will use the ANES data to answer questions about voters in the US.

This lab consists of three research questions. For each question, your team will conduct a statistical analysis and generate a written report in pdf format. This means that your team will create three separate reports, each one a complete analysis on its own.

This is an exercise in both statistics and professional communication. It is important that your techniques are properly executed; equally important is that your writing is clear and organized, and your argument well justified.

Your instructor will divide you into teams to work on this lab. Our goal is that each student learns as much as possible through conducting this lab. A divide-and-conquer approach might be the most expedient way to *finish* the work on this lab, but it might not be the way for each person to maximize their learning. Instead, we would like teams to collaborate and review each others' work, ask questions about approach, and work to improve writing, argument, and code.

This one week lab is due before your Unit 9 live session. You will find a separate place on Gradescope to submit each of your three responses, along with the source file used to create your pdf. Only one person from your team needs to upload a submission.

Data

Data for the lab should be drawn from the 2020 American National Election Studies (ANES). You can access this data at https://electionstudies.org. This is the official site of the ANES, a project that has been ongoing since 1948, and federally funded by the National Science Foundation since 1977.

To access the data, you will need to register for an account, confirm this account, and then login. The data that you need should come from the **2020 Time Series Study**.

You will note that there are two forms of data that are available, data that is stored in a `.dta` format, and data that is stored in a `.sav` format. Both of these are proprietary data formats (`.dta` for STATA, and `.sav` for SPSS). You will need to find an appropriate library to read this data into R; we recommend that you find a package that is within the "tidyverse".

While you're at the ANES website, you will also want to download the codebook, because all of the variables are marked as something like, `V200002` – which isn't very descriptive without the codebook.

For a glimpse into some of the intricacies that go into the design of this study, take a look at the introduction to the codebook.

Like many modern surveys, the ANES includes survey weights, which are used to correct for differences between how frequently demographic groups appear in the sample compared to the US population (often

ultimately relying on US census data). These weights are beyond the scope of our class and you are not expected to utilize them.

The Research Questions

The research question for each of the three parts of the lab are as follows:

1. Are Democratic voters older or younger than Republican voters in 2020?
2. Did Democratic voters or Republican voters report experiencing more difficulty voting in the 2020 election? (Exploratory, ungraded question: Were the types of reported difficulties the same, or different for respondents who identify with the different parties? We're not going to grade this, and it isn't a fully formed question here, but you [like us] might be interested in exploring this while you're writing up your report.)
3. Are people who believe that science is is important for making government decisions about COVID-19 more likely to disapprove of the way their governor is handling the pandemic?

Guidance From Political Scientists

This is beyond what we expect someone could know from the data or their background, so we'll share it here as guidance to students who are writing this report. Political identification in the US is a strange identity that loosely maps onto political ideology. While there is more written about this, see `./background_literature/petrocik_2009.pdf` as some guidance about how stated political identity might not match with revealed political identity at the ballot box.

As practical guidance, please treat individuals who "lean" in one direction or another as members of that party. This means that someone who "Leans Democratic" should be classified as as Democrat; and someone who "Leans Republican" should be classified as a Republican.

Report Guidelines

There is additional, specific guidance about testing in the Rubric.

General

For each of the three research questions, you will create a pdf created by a separate source .Rmd file.

- Each report should be a fully contained argument that does not rely on arguments made in other reports.
- Each report should be no more than 3 pages in standard latex formatting (i.e. `output: pdf_document`)
- Follow the .Rmd template that we have created for each question, using the prompts to guide you through the parts of an analysis. Make sure you fill in each prompt with all information requested.
- Each report should contain either a plot or a table that advances the argument.

Introduction

Begin each report with an introduction to motivate the analysis.

- Introduce the topic area and explain why the research question is interesting.
- The introduction must "do work," connecting the general topic to the specific techniques in the report.

Visual Design

Any plots or tables that you include must follow basic principles of visual design.

- A plot/figure must have a title that is informative.

- Variables must be labeled in plain language. As an example, `v20002` does not work for a label.
- A plot should have a good ratio of information to ink / space on the page. Do not select a large or complicated plot when a simple table conveys the same information directly.
- Do not include any plot (or R output in general), that you do not discuss in your narrative.
- The code that makes your plot/figure should be included in your report `.Rmd` file, but should not be shown in your final report. To accomplish this, you can use an `echo=FALSE` argument in the code chunk that produces the plot/figure.

Data Wrangling

To answer each research question, you will have to clean, tidy, and structure the data (A.K.A. wrangle).

- The code to wrangle data should be included with your deliverable somehow. If you choose to include it in your report `.Rmd` file, then it not be shown in the PDF of your final report. To accomplish this, you can use an `echo=FALSE` argument for the code chunk that does the wrangling.
  - A better practice – not strictly necessary for this lab – would be to write a function that loads and cleans *all* of the data that is being used by your team for its reports. This way, a single function can be run (and evaluated by your reader) for all the loading, cleaning, and manipulating.
- While we do not want to prohibit you from using additional tools for data manipulation, you should be able to complete this lab with no more than the base `stats` library, plus `dplyr` and `ggplot2` for data manipulation and plotting. Other tools within the tidyverse are available to use, but don't feel like you have to search them out.
- You will learn more by writing your own function than you would searching for a package that does one thing for your report.

Hypothesis Testing

To answer each research question, you will have to execute one of the statistical tests from the course.

- The code that executes your test *should* be shown in your report, because it makes very clear the specific test that you're conducting.
- You need to argue, from the statistical principles of the course, why the test you are conducting is the *most appropriate* way to answer the research question.
- Although you might not do this for a report at your organization, for this class please list every assumption from your test, and evaluate whether the data generating process actually meets this assumption.
- If you identify problems with some assumptions for your test, that does not mean that you should abandon the analysis or hide the problem. If these "limitations" exist, please describe them honestly, and provide your interpretation of the consequences for your test.
- While you can choose to display the results of your test in the report, you also *certainly* need to write about these results. This should be accomplished using inline code chunks, rather than by hard-coding / hard-writing output into your written report. An example of this is included in `lab_1_example_solution.Rmd`.

Test, results and interpretation

Please discuss whether any statistically significant results that you find are of *practical significance*. There are many ways to do this, but the best will provide your reader enough context to understand any measured differences in a scale appropriate to your variables. Explain the main takeaway of your analysis and how it relates to the broader context you identified in the introduction.

One way to self-assess whether you have succeeded in communicating the practical significance of any results that you find is to imagine that **all** results that you could possibly test would be statistically significant. If everything that you tested were significant, you would have to make an argument that *this particular* test either was (or was not) important within the broader context of the data.