Lab 1: Hypothesis Testing

w203 Teaching Team

1 Statistical Analysis

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. In this part, you will use the ANES data to address a question about voters in the US. Your team will conduct a statistical analysis and generate a written report in pdf format.

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.

2 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**.

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 situations in which members of one demographic group are more likely to respond to the survey than members of another demographic group. (The target proportions are ultimately based on US census data). The survey weights make it possible to generalize from the a population that represents people who take the survey to a population that represents the United States as a whole. These weights are beyond the scope of our class and you are not expected to utilize them. You will still be able to learn about a population model, even if applicability to the US population is limited.

3 The research question

Use the ANES data to address the following question:

> Did Democratic voters or Republican voters experience more difficulty voting in the 2020 election?

4 Guidance from political scientists

Political identification in the US is a complex phenomenon that is the topic of a large academic literature. See `./background_literature/petrocik_2009.pdf` for some guidance about how stated political identity might not match with revealed political identity at the ballot box.

As practical guidance:

1. Is it reasonable to use the vote that someone cast to identify their party preference in this case? What if someone had so difficult a time voting that they did not cast a ballot?
2. 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.

5 Report guidelines

This section provides some guidance for you as you write your report. In `rubric.md` we provide you with specific statements of how we will evaluate your report.

General guidance

- You should knit an .Rmd file to create your pdf report.
- Your report should be no more than 3 pages in standard latex formatting (i.e. `output: pdf_document`)
- You should assume your reader is familiar with statistics, but has no special knowledge of the ANES survey.
- Your report should contain either a plot or a table that advances the argument.

Introduction

- Begin your 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.

Conceptualization and Operationalization

How do you get from a research question to data? First, ensure that the concepts in your question are clear.

- Who or what is a voter?
- Who is a "Republican" and who is a "Democrat"?
- What is difficulty voting?

Only after you have informed your reader of what these concepts are can you then describe how you are going to *measure* these concepts.

- What would be the best **possible** method of measuring this concept? Is this method possible? Why or why not?
- What is the best **available** method of measuring this concept? Why have you opted to use this measurement instead of other possibilities? Map the concept definitions that you have written down onto the variables that you are going to use. Describe, precisely, how the variables were generated, if they come from survey data, provide the text of the question that the respondent is reacting to, not the variable name.
- What, if any, changes have you made to the dataset from how it was provided? Why did you make those changes, how much data was affected, and what are the consequences for any estimates that you produce?

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 your 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 should 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 your 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 it (assess whether the assumption is a reasonable reflection of the natural process that generated the data).
- 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.