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 you will create three separate reports, each one a complete analysis on its own. (This is especially important since different graders may grade each of your responses.)

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.

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.

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. Are Democratic voters more enthusiastic about Joe Biden or Kamala Harris?
3. Are survey respondents who have had someone in their home infected by COVID-19 more likely to disapprove of the way their governor is handling the pandemic?

Report Guidelines

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.

Visual Design

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

- 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 in your report `.Rmd` file, but should not be shown in your final report. To accomplish this, you can use an `echo=FALSE` argument for the code chunk that does the wrangling.
- 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 can also Let this soft constraint empower you to write code yourself, rather than searching for a package that does *one thing* only for your report.

Hypothesis Testing

To answer each research question, you will have to execute one of the statistical tests that you have learned in 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.
- In making this argument, you need to list the assumptions/requirements of the data, and either reason or demonstrate that the data does, in fact, meet these requirements.
- Even after you make the argument that you have conducted the *most* appropriate test, there may be some requirements of the data that are not met. If these exist, please name them, and provide your interpretation of what the consequences are for your test results.
- 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`.