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 3 parts. Each part is centered around a research question. For each question, you 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, but equally important that your writing is clear, well defended, and organized.

Your instructor will divide you into teams to work on this lab. To maximize your learning, we ask that you do not use a divide-and-conquer approach to finish the lab. Instead, all students should participate in all parts as much as possible.

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

- Each report must stand alone as a separate document. Each report should:
  - Be a separate pdf created by a separate source Rmd file.
  - Be a fully contained argument that does not rely on arguments made in other reports.
  - Be no more than 3 pages in standard latex formatting (i.e. `output: pdf_document`)
  - We have created a template Rmd file for each question, including prompts to guide you through the parts of an analysis. Make sure you fill in each prompt with all information requested.
  - Contain either a plot or a table that is informative about the distribution of the data. This plot or table should:
    * Have a caption or a title that is informative
    * Have variables listed in plain language. As an example, `v20002` does not work for a label, but `R Party Affiliation` *does*.
    * 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 in your `.Rmd` file.
- To answer your research question, you will have to clean, tidy, and structure the data generated by the the survey questions that you're using to answer your research question.
  - This code to clean, tidy, and structure 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 is cleaning, tidying and structuring in your `.Rmd` file.
  - While we do not want to prohibit you from using additional tools for data manipulation, *you should be able to complete this whole lab with no more than the base `stats` library, plus `dplyr` and `ggplot2` for data manipulation and plotting. Other tools within the tidyverse are free 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.
- To answer your research question, you will have to execute one of the statistical tests that you have learned in the course. for example, using a `t.test()`.
  - 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 name 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. Below is an example.

```r
test_result <- t.test(
  x = rnorm(n=10, mean=0, sd=1),
  y = rnorm(n=10, mean=0,sd=1)
)
```

The test for the difference between X and Y does no find a statistically significant difference between the two variables; the p-value is 0.61.