Lab 1

203 Teaching Team

June 16, 2025

Table of contents

Introduction

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 lab, 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 statistical techniques are properly executed; equally important is that your writing is clear and organized and your argument well justified.

The research question

You have enrolled in the MIDS program to continue your education. As instructors, we are proud of you for taking this step, and we believe that our program is good for the world! In the US context however, we have seen prominent debates about primary schools and colleges, what values the transmit to students, and their appropriate role in society.

As a starting point for understanding the current situation, we are tasking you with the following research question:

> Do Democratic and Republican voters have different views of education?

The ANES survey includes a number of variables that could be useful for answering this question. Your goal is to produce an argument that is both scientifically and statistically defensible, but also interesting within the larger national dialogue about truth, knowledge, and power.

Data

Data for the lab should be drawn from the 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 has been 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 **2024 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.[1] The survey weights make it possible to generalize from 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.[2] You will still be able to learn about a population model (representing people willing to take this survey), even if applicability to the US population is more limited.

Parameters

1. Your analysis must include one, and no more than two, hypothesis tests. For each comparison you make, you should use the most appropriate statistical test for the variables you have. In your write up, you should note why you believe this test you have chosen is the most appropriate, and write an evaluation for every test assumption.
2. **Before** you conduct any tests, or look at any plots, you should write down (i) all of the comparisons that you are going to make; and (ii) the order in which you will conduct them. The purpose is to prevent a fishing expedition, which is a form of cheating and leads to inflated error rates.
3. It is up to you to choose comparisons that would be informative of your research question, no just comparisons that are likely to be "statistically significant." Remember: negative test results can be interesting! If the test result is negative because the effect size is small, ask yourself who would be surprised by the small effect, or who might change the way they do their job because of it.

Guidance from political scientists

Political identification in the US is a complex phenomenon that is the topic of a large academic literature. Please read `./background_literature/petrocik_2009.pdf` for guidance about

[1] The target proportions are based on US census data, and the weights applied within the survey are derived from information collected as a part of the *Current Population Survey's* election supplement. There is some concern that these CPS estimates are overestimating the turnout rate among some racial/ethic groups (Ansolabehere, Fraga and Schaffner, 2022).

[2] An enthusiastic student team could choose to use these survey weights in their analysis, but in doing so, they will have to do additional reading so that they can defend the estimated statistics and the associated sampling based uncertainty. A difference in means might pull a function from the {estimatr} package, `estimatr::difference_in_means()` . If you need a weighted test for ranks, you're really getting deep into the world of statistics tests. You could look into the package {coin}, which uses a different inferential system than we have taught in class, namely permutation based inference.

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.

Deliverable

- You will turn in a pdf of your report.
  - Your report should be no more than **four** pages in standard R pdf output, and you should not change the document template to "make it fit" or "squeeze it in." If there is a page that contains only the name of the report, the authors and the date, that does not count as one of the four pages.
- You will provide access to your repository so that your instructor can read and execute your code. This repository should be sensibly structured and should allow your instructor to reproduce your analysis by simply knitting your report.

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 and the `tidyverse` for data manipulation and plotting. You *may* use other tools, but if you do, you should ensure that they are saved to your `renv::snapshot()` so that your instructor can fully replicate your work.

Evaluation Criteria

We present the following criteria to guide you to a professional-quality report. Moreover, these criteria are also the ones we will use to grade your report. The descriptions below are copied directly from our grading rubric.

(6 points) Introduction

An introduction that is scored in the top level has very successfully motivated the analysis for the reader. It will introduce the topic area and explain why the research question is interesting and to what audiences. It will identify the most salient motivations for the research question.

Finally, it will "do work" by connecting the general topic to the specific concepts in the research question and to the statistical techniques used in the report. After reading the introduction, the reader should be well primed to appreciate the statistical procedures that follow.

(6 points) Conceptualization and Operationalization

To receive full credit, your report must define the concepts in the research question, precisely explain how they are operationalized, and discuss any gaps between the conceptual and operational definitions. In particular, your report should clearly address the following questions: (1) Who is a voter? (2) Who is a "Republican" and who is a "Democrat"? (3) What is a view of education?

Only after you have informed your reader of what these concepts are can you then describe how you are going to *measure* these concepts. Be precise, for example, explaining how you treat different levels of a variable. You should also list how many observations you remove and for what reasons. If multiple variables could be used to represent a concept, you should point out the alternatives and give a reason for why you made your choice. Clearly explain any mismatch between the variables you use and the conceptual definition.

(4 points) Data Wrangling

A report that is scored in the top level on data wrangling will have succeeded – relative to expectations at this point in the course – to produce a modern, legible data pipeline from data to analysis.

The code you use to wrangle data should be placed in your Rmd file, but not shown in your pdf (use an `#| echo: false` argument in code chunks). Alternately, the wrangling can be placed into its own, more modular, `.R` file. The analysis should have a single source of truth for data, and avoid creating multiple data.frame objects as much as possible. Variable names should clearly communicate their meaning. Numbers in your narrative should be computed 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`.

A report that cannot be compiled by an instructor cannot score more than one point in this section.

(3 points) Data Understanding

To get full credit, your report should provide the reader with adequate background about the data source. You should assume that your reader may have never heard of ANES before. This part of your report may be brief, but it should give the reader enough information that they are prepared to place your results in context and engage with your discussion of assumptions.

If there are features of the data that require explanation, they should be covered either in writing or through the purposeful use of tables or figures.

(7 points) Plots and Tables

You are required to include at least one plot or table in your report (there is a preference for plots). A report that is scored in the top level will include plots that effectively transmit information, engage the reader's interest, maximize usability, and follow best practices of data visualization. Titles and labels will be informative and written in plain English, avoiding variable names or other artifacts of R code. Plots will have a good ratio of information to space or information to ink; a large or complicated plot will not be used when simple plot or table would show the same information more directly. Axis limits will be chosen to minimize visual distortion and avoid misleading the viewer. Plots will be free of visual artifacts created by binning. Colors and line types will be chosen to reinforce the meanings of variable levels and with thought given to accessibility for the visually-impaired.

(5 points) Stating a Hypothesis

A report that is scored in the top level will have stated the correct null hypothesis for each conducted test, using mathematically precise language.

(9 points) Test Selection and Evaluation of Assumptions

A report that is scored in the top level will select a fully appropriate test for the scenario, and evaluate **every** assumption for the test. Different teams may choose different tests and receive full credit; however, you will not receive full credit if your chosen test is clearly inferior to another test. You must include every assumption as listed in our async and cheat sheets. Each assumption must be discussed so that the reader can clearly judge how appropriate it is, though you should avoid giving a binary (valid/not valid) decision for each one. The reader should also understand the statistical consequences for any violated assumptions.

(9 points) Test Results and interpretation

A report that scores in the top level will have clearly interpreted the (1) statistical significance and the (2) practical significance of the results. The practical significance discussion should include a measure of effect size that is appropriate for the data. More generally, it must leave the reader with an understanding of the direction of the effect and whether the magnitude of the effect is meaningfully large or small in context. While many reports may not include any code, we ask that you include the code that executes your test in your report, because it makes very clear the specific test that you're conducting.

(5 points) Overall Effect

A report that scores in the top level will have met expectations for professionalism in data-based writing, reasoning and argument for this point in the course. It can be presented, as is, to another student in the course, and that student could read, interpret and take away the aims, intents, and conclusions of the report.