Lab 1: Hypothesis Testing

w203 Teaching Team

Overview

This is a team-based Lab. Your instructor will divide you into teams of around three students. To maximize learning, we would like all students to engage with every lab component, discussing strategy with teammates, reviewing solutions, and iterating on text and code.

The lab consists of two parts. Part 1 consists of foundational exercises similar to a homework. Part 2 is a written statistical analysis.

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

Part 1: Foundational Exercises

Proof and Short Answer

1. Your aunt (who is a professional magician), claims to have created a pair of magical coins that share a connection to each other that makes them land in the same way. The coins are always flipped at the same time. For a given flip $i \in \{1, 2, 3...\}$, let X_i be a Bernoulli random variable representing the outcome of the first coin, and let Y_i be a Bernoulli random variable representing the outcome of the second coin. You assume that each flip is independent of all other flips. You also assume that

$$P(X_i = 0) = P(X_i = 1) = P(Y_i = 0) = P(Y_i = 1) = 1/2$$

and write

$$P(X_i = Y_i) = p$$

Your aunt claims that $p > 1/2$.

You design a test to evaluate your aunt's claim. You flip the coins 3 times and your test statistic is the sum $X_1 + Y_1 + X_2 + Y_2 + X_3 + Y_3$

Your null hypothesis is that $p = 1/2$. You plan to reject the null if your test statistic is 0 or 6.

   a. What is the type 1 error rate of your test?

   b. What is the power of your test for the alternate hypothesis that $p = 3/4$?

2. Imagine that your organization surveys a set of customers to see how much they like your regular website, and how much they like your mobile website, both measured on 5-point Likert scales. If you were to run a paired t-test, what consequences would the violation of the metric scale assumption have for your interpretation of the test results?

Test Assumptions

For the following questions, your task is to evaluate the assumptions for the given test, using your background knowledge and examining the data. The test may or may not be the most appropriate test for the scenario. Do not perform the test.

1. The file `Happiness_WHR` is subsetted from the World Happiness Report, a yearly publication that uses data from the Gallup World Poll surveys. The variable `life ladder` is a measure of happiness, described in the FAQ as follows:

   > This is called the Cantril ladder: it asks respondents to think of a ladder, with the best possible life for them being a 10, and the worst possible life being a 0. They are then asked to rate their own current lives on that 0 to 10 scale. The rankings are from nationally representative samples, for the years 2018-2020.

You would like to know whether people in countries with high GDP per capita (higher than the mean) are more happy or less happy than people in countries with low GDP (lower than the mean).

List all assumptions for a two-sample t-test and evaluate them.

2. The file `legislators-current.csv` is taken from the congress-legislators project on Github (https://github.com/unitedstates/congress-legislators). You would like to test whether Democratic or Republican senators are older.

List all assumptions for a Wilcoxon rank-sum test and evaluate them.

3. The dataset `wine` can be accessed by installing the `wooldridge` package. It contains observations of variables related to wine consumption for 21 countries. You would like to use it to test whether countries have more deaths from heart disease or from liver disease.

List all assumptions for a signed-rank test and evaluate them.

4. The file `GSS_religion` is a subset of data from the 2004 General Social Survey (GSS). The variables `prottemp` and `cathtemp` are measurements of how a respondent feels towards protestants and towards catholics, respectively. The GSS questions are phrased as follows:

   > I'd like to get your feelings toward groups that are in the news these days. I will use something we call the feeling thermometer, and here is how it works: I'll read the names of a group and I'd like you to rate that group using the feeling thermometer. Ratings between 50 degrees and 100 degrees mean that you feel favorable and warm toward the group. Ratings between 0 degrees and 50 degrees mean that you don't feel favorable toward the group and that you don't care too much for that group. If we come to a group whose name you Don't recognize, you don't need to rate that group. Just tell me and we'll move on to the next one. IF you do recognize the name, but you don't feel particularly warm or cold toward the group, you would rate the group at the 50 degree mark. How would you rate this group using the thermometer?

You would like to test whether the US population feels more positive towards Protestants or towards Catholics.

List all assumptions for a paired t-test and evaluate them.

Part 2: 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.

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 Question

Use the ANES data to address the following question:

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

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, 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.

> I'm not sure why, given this guidance, students went and chose voting behavior to operationalize Dem and Rep. But perhaps we want this part to be more wide open so students have to spend time thinking of what operational definition makes the most sense. it seems like this guidance forecloses more possibilities.

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.

- Your report should be no more than 3 pages in standard latex formatting (i.e. `output: pdf_document`)
- You should assume your reader is familar with statitics, but has no special knowledge of the ANES survey.

- Follow the .Rmd template that we have created, using the prompts to guide you through the parts of an analysis. Make sure you fill in each prompt with all information requested.
- 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.

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 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 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.

> kind of want to delete the last paragraph. feels like double explaining what we already taught students.