BIS 315 Statistical Inference and Hypothesis Testing Lab

Make sure to begin by establishing a group agreement for the lab

1. Produce written agreements:
   • What is the distribution of work for the group members (in-class lab; any out-of-class lab work, if needed; writing of report)?
   • Do you want the same grade or individual grades? If individual, I will need to know the exact distribution of work and the separate writing segments.

2. How will you resolve a case in which someone is not fulfilling their agreement?

3. What are the expectations regarding communication (e.g., medium of communication, expected time to response)

Introduction

In Lab 1 you practiced the craft of generating, summarizing, and interpreting descriptive statistics by examining two distinct sets of variables. The first set related to the frequency with which U.S. adults socialize with friends, neighbors, and relatives, while the second set concerned the education level of U.S. adults in relation to their parents’ levels of education. In the Discussion section of the lab report, many of you raised an interesting set of hypotheses about how these results may be different for various contexts and subgroups in society. In Lab 2, we will test some of these hypotheses directly.

As with the first lab, I strongly encourage you to approach this assignment with intention and to set clear objectives for what you hope to get out of the experience. For example, if the concept of a p-value is unclear, this is a good opportunity to learn something about the concept. Or maybe you want to gain some clarity around the relationship between statistical significance and confidence intervals. My advice is to pick one or two things you hope to learn and focus on those concepts with intention. Whatever it might be, take 5 minutes or so to clarify with your partner the following questions:

- What do you hope to learn from this lab exercise focused on statistical inference and hypothesis testing? Are there certain concepts we have covered thus far that you hope to clarify through the assignment? Are there certain concepts that you understand but nevertheless hope to explore in more depth? Is there anything else you expect to learn?

Your work in this lab will be evaluated in a similar way to Lab 1. First, I will be looking to see that you completed all of the analyses that you were asked to run. The report does not ask you to summarize every bit of output, but I will be checking to see that you did, in fact, run every analysis. Second, I will be looking to see that you structured your report in the format outlined below. Third, and by far the most important, I will be looking to see that you can accurately interpret statistical output related to confidence intervals and t tests.

As always, please let me know if you have questions or if you are unsure about what is expected. I genuinely want this to be a productive learning experience (although that does not mean an experience free of frustration – at least some struggle should be part of the process).
Data
The work in this lab will center on the General Social Survey 2018 data set. The GSS is a cross-sectional survey conducted by the National Opinion Research Center (NORC, University of Chicago) that makes use of a full probability sampling of households in the United States. The data set contains a core set of demographic, attitudinal, and behavioral variables in each wave, and includes a variety of special topics (e.g., social network data) that appear in select years. I have done some initial recoding for the lab, so please use the version of the data set that is on the course website under November 21\textsuperscript{st} (https://www.josephferrare.com/s/GSS2018.sav). Make sure to save the data set to your own folder so that you can access it at a later time, if needed. In addition, download the syntax file from our course website under November 21\textsuperscript{st} (link is here: https://www.josephferrare.com/s/BIS315Lab2Syntax.sps). This file includes all the syntax you will need for the lab.

The lab exercise centers on inferences and hypothesis tests for two sets of variables:

1. Socializing: We will test specific hypotheses related to the frequency with which U.S. adults socialized with friends, relatives, and neighbors in 2018. As with Lab 1, the following will serve as the dependent variables (each GSS variable name is in bold type):
   - socrel\_r: This GSS variable categorizes the frequency with which U.S. adults spend a social evening with their relatives.
   - socfrend\_r: ...frequency with which U.S. adults spend a social evening with friends who live outside the neighborhood.
   - soccommun\_r: ...frequency with which U.S. adults spend a social evening with someone who lives in their neighborhood.

As we found in Lab 1, these variables were measured using ordinal scales, which means that measures such as the mean and standard deviation are less appropriate than median and IQR. However, for the purposes of this lab we are going to cautiously proceed to test differences in the mean categories for these variables. With that said, keep in mind what the mean actually means when interpreting the results.

NOTE: These questions were only asked among a subset of the sample, so you will observe many missing cases. You can assume that the subset was random.

We will test multiple hypotheses for each of the socializing variables. In particular, we will test for differences using the following variables:
   - **WestCoast**: a dummy variable that indicates whether or not the respondent lived on the West Coast at the time of the survey
   - **city**: a dummy variable that indicates whether or not the respondent lived in a city of more than 250,000 people when they were 16 years old\textsuperscript{1}
   - **fear**: indicates if the respondent is afraid to walk at night in their neighborhood
   - **age\_r2**: age groups of 18 – 45 and 45+

\textsuperscript{1} There was no data available for the current residence.
2. Education Attainment: Next, we will test hypotheses related to respondents’ level of education attainment.

The outcome (or dependent) variable of interest is:
- **educ**: measures the number of years of education completed by each respondent, ranging from 0 to 20

We will test multiple hypotheses to explore education attainment using the following variables:
- **mabachelor**: dichotomous variable indicating if the respondent’s mother has a bachelor’s degree or less than a bachelor’s degree
- **pabachelor**: dichotomous variable indicating if the respondent’s father has a bachelor’s degree or less than a bachelor’s degree
- **lowinc**: indicates if the respondent’s family had below average income when the respondent was age 16 (note: this relies on the respondent’s assessment, not actual tax records)
- **whiteblack**: indicates if the respondent identifies their race as White or Black
- **sex**: indicates if the respondent identifies as male or female

**Hypothesis Tests for Socializing Variables**

1. Before running any of the syntax, start by specifying and writing down the null and alternative hypothesis for each of the significance tests that will be run. In addition, write down what you expect the outcome of each significance test will be (i.e., do you expect to reject or fail to reject the null hypothesis?). The following significance tests will be run for each of the three socializing variables (relatives, friends, neighbors):
   a. West Coast v. non-West Coast
   b. City v. non-City
   c. Afraid to walk in neighborhood at night v. not afraid
   d. Age 18 – 45 v. 46+

   Keep in mind that you will specify three null and alternative hypotheses for each of the four tests (i.e., 12 null and alternative hypotheses total). Don’t forget to also write down you anticipated findings!

2. Run the syntax lines 3 – 34

3. The first few segments of output provide the descriptive statistics for the variables used in the analysis. Inspect these just to see what you are working with. Are there a lot of missing cases? For which variables? What are the overall means for the dependent variables?

---

2 The GSS measures race in a coarsely-grained way. There are variables in the data set that can be used to capture additional racial (and ethnic) groups, and I would be happy to work with anyone who is interested in exploring ways to incorporate additional racial or ethnic groups into their analysis for this lab.
4. Next, spend some time reviewing and interpreting the results of the independent samples t tests. Assuming that the significance level for each analysis is p<0.05, do you reject or fail to reject the null hypotheses? How do you interpret the p-values and confidence intervals for each test?
   a. Note: Enter your results into a table for ease of comparison and interpretation. I have created a template for you on the course website under Nov 21 where the rest of the files for the lab are located: https://www.josephferrare.com/s/Lab2Tables.xlsx
5. Now compare the results to your expected findings. Any surprises? Start to think about why you found what you did, considering the ways that these variables were measured, contexts that were not accounted for, and how your own individual (or even group-wide) biases may have shaped your initial intuition about these relationships.

Hypothesis Tests for Educational Attainment
1. Before running any of the syntax, start by specifying the null and alternative hypothesis for each of the significance tests that will be run. In addition, write down what you expect the outcome of each significance test will be (i.e., do you expect to reject or fail to reject the null hypothesis?). The following significance tests will be run:
   a. Bachelor’s-educated mother v. less than bachelor’s-educated mother
   b. Bachelor’s-educated father v. less than bachelor’s-educated father
   c. Low family income v. average or greater family income
   d. White v. Black
   e. Male v. Female

Keep in mind that you will specify a null and alternative hypothesis for each of the five tests (i.e., 5 null and alternative hypotheses total). Don’t forget to also write down you anticipated findings!

2. Run the syntax lines 37 – 73
3. The first few segments of output provide the descriptive statistics for the variables used in the analysis. Inspect these just to see what you are working with. Are there a lot of missing cases? For which variables? What is the overall mean and standard deviation for the dependent variable?
4. Next, spend some time reviewing and interpreting the results of the independent samples t tests. Assuming that the significance level for each analysis is p<0.05, do you reject or fail to reject the null hypotheses? How do you interpret the p-values and confidence intervals for each test?
   a. Note: As above, enter your results into the table provided for ease of comparison and interpretation.
5. Now compare the results to your expected findings. Any surprises? Start to think about why you found what you did, considering the ways that these variables were measured, contexts that were not accounted for, and how your own individual (or even group-wide) biases may have shaped your initial intuition about these relationships.
Reporting of Results
In this phase of the lab, you will engage in narrating the results and discussing possible explanations and limitations for what you observe. Your report should be single-spaced and in 12-pt font of your choosing. Please insert page numbers into the document. Please also include a title page that contains the names of all group members, the name of this assignment (“BIS 315 Hypothesis Testing Lab”), and the date of submission. I have included general guidelines for the length of each section. These are just guidelines, not laws. Structure the report in the following sections (the headings below are not guidelines, they are laws):

Introduction (2-3 paragraphs)
In this section, provide a general overview of the two sets of analyses that you completed. This can be done in 2 – 3 paragraphs. What were the general themes you were exploring through the lab? What did you expect to learn from the hypothesis tests about the ways U.S. adults spend time with friends, relatives, and neighbors? What did you expect to learn about education attainment in the United States? What were your assumptions about what you might find for both topics? Note that this should be different from the introduction in Lab 1 since your exploration in Lab 2 is based on hypothesis tests between groups rather than basic descriptive statistics (per Lab 1).

Data and Methods (about 1 page)
In about 1 page, describe the data and variables that you used in the analysis. The relevant information here should focus on an overview of the General Social Survey, the specific variables you used, and the sample size for the 2018 data set.

In addition, spend some time discussing the statistical tools you used to test your hypotheses. What assumptions are you making in running these tests? What is the general objective of these tools? What do they illuminate about data and what are the limitations?

Results
The results section should be divided into two sub-sections, one for each set of results.

Spending Time with Friends, Family, and Neighbors (about 1 page)
Address the following questions:

First, insert the table that was created during the lab. Title this table, “TABLE 1. T tests for socializing variables.” Then, proceed to narrate the key findings from these results in relation to your expectations discussed during the lab and summarized in the introduction. Which of the tests were statistically significant (i.e. rejection of the null hypothesis)? For the selected results that you highlight, interpret the p-values and situate the mean differences within the respective confidence intervals.

As with the report from Lab 1, in this results section you should be sticking closely to the what the results say, making sure to avoid any sweeping generalizations or inferences not supported by your data.
Discussion of Social Evening Analysis (about 2 paragraphs)
In this sub-section, discuss the general conclusions you can draw from your analysis. What is the overall story of the ways U.S. adults spend time with friends, family, and neighbors in relation to the comparisons tested in the hypothesis tests? What can you conclude? Were certain groups or geographic regions more or less likely to spend time with friends, relatives, and neighbors? If so, what might be some plausible explanations for these differences? What other variables would you need to test to deepen the story? To what extent might your own assumptions and individual biases shape the way you answer these questions? Is it possible that the individual biases of the designers of the GSS somehow crept into the ways the questions were worded and/or how the variables were measured? What are the potential implications for these forms of bias?

Education Attainment and Parental Education (2 – 3 paragraphs)
Address the following:

First, insert the table that was created during the lab for these variables. Title this table, “TABLE 2. T tests for educational attainment.” Then, proceed to narrate the key findings from these results in relation to your expectations discussed during the lab and summarized in the introduction. Which of the tests were statistically significant (i.e. rejection of the null hypothesis)? For the selected results that you highlight, interpret the p-values and situate the mean differences within the respective confidence intervals.

Discussion of Education Attainment Analysis (1 or 2 paragraphs)
In this sub-section, discuss your general conclusions about the findings. What can we say about educational attainment in relation to the comparisons tested? Overall, would you say that educational attainment is marked by equality or inequality? What is the basis of your claim? What other variables would you need to test to deepen the story? To what extent might your own assumptions and individual biases shape the way you answer these questions? Is it possible that the individual biases of the designers of the GSS somehow crept into the ways the questions were worded and/or how the variables were measured? What are the potential implications for these forms of bias?

Conclusion (about 2 paragraphs)
Aside from what you may have learned about socializing and education, what did you learn about hypothesis testing and data analysis more generally? Did this exercise clarify any of the concepts from the course? Did it make things more confusing? What questions did it raise? I’m looking for some reflection on what you learned about the course content.

Appendix A
Include all the output from both sets of analyses as an appendix within the report.

Appendix B
Include a scanned (or picture) copy of your group agreement.
BIS 315 Lab 2 Scoring Rubric: 75 Points
1. Evidence of completed analyses: 5 pts
2. Quality of introduction: 5 pts
3. Discussion of data & methods: 10 pts
4. Description & discussion of Results:
   Socializing Results:
   a. Accurately rejected or failed to reject null hypotheses: /5 pts …
      i. …and can accurately interpret p-values: /5 pts
      ii. …and can accurately interpret selected confidence intervals: /5 pts
   b. Can interpret these results in relation to your data – i.e., tell the story: /5 pts
   Education Results:
   a. Accurately rejected or failed to reject null hypotheses: /5 pts …
      i. …and can accurately interpret p-values: /5 pts
      ii. …and can accurately interpret selected confidence intervals: /5 pts
   b. Can interpret these results in relation to your data – i.e., tell the story: /5 pts
5. Depth of discussion of results for both sets of variables: /5 pts
6. Conclusion: /5 pts
7. Report followed assigned structure, including appendices: /5 pts