431 Lab 05

Deadline: See Course Calendar | Last Edited 2022-08-23 13:24:31

Table of contents

Deadline

Lab 05 has 9 questions, all of which you need to complete by the deadline specified on the Course Calendar.

- To receive full credit on a Lab, it must be received on Canvas no later than 59 minutes after the posted deadline. (This allows for small issues with uploading to Canvas to occur without penalty.)

Setting a Seed

You may or may not need to use a random seed in this assignment. If you need to use a random seed, please use `4312022`.

Learning Objectives

1. Be able to take information about a dataset, and an associated visualization, to identify the appropriate inferential test when comparing means.
2. Develop and appropriately interpret a confidence interval, in context, as derived from an analysis of categorical variables.
3. Use some of the concepts developed in Chapter 6 of Spiegelhalter's *The Art of Statistics* to evaluate a model's performance.

An Important Note

Your response to **every** question, whether we explicitly ask for it or not, should include a complete English sentence responding to the question. Code alone is not a sufficient response, even if the code is correct. Some responses might not need any code, but every response needs at least one complete sentence.

Part A. County Health Rankings (Questions 1-6)

We're going to revisit our County Health Rankings data, specifically using the same Midwest counties we worked with in Lab 02, but starting from an expanded `lab05_counties.csv` data set.

First, you'll want to read in the `lab05_counties` data and filter to the states of Ohio (OH), Indiana (IN), Illinois (IL), Michigan (MI), and Wisconsin (WI). The main variables we'll be looking at in this Lab are `metro` and `access_to_exercise_opportunities`, but you'll go ahead and still want to keep `state` and `county_name` as well. Determine whether there are any missing values in that set of variables. We'll call the resulting tibble `midwest05`.

You've been asked by the principal investigator of a study to examine if there are differences in the mean percentage of adults with access to exercise opportunities between metropolitan and non-metropolitan counties, but we'll assume (for now) that you only have data within the five states (OH, IN, IL, MI and WI) we've identified above.

Question 1 (10 points)

In Question 1, create a factor variable as part of the `midwest05` tibble which is better than the 0/1 that `metro` currently is, as well as adjust `access_to_exercise_opportunities` from a proportion to a percent and give it a shorter, more appealing name. Then please make an attractive and useful visualization which shows the distribution of the percent of adults with adequate access to exercise opportunities stratified by the county's metropolitan status. Finally, provide a couple of sentences describing your initial conclusions based on your visualization.

Question 2 (10 points)

In a few sentences, comment on the sampling approach used to create the data you used in Question 1. Are the data a random sample from the population(s) of interest? Is there at least a reasonable argument for generalizing from the sample to the population of all US counties? Or is there insufficient information provided on this point? How do you know?

Question 3 (5 points)

Are the data you developed in Question 1 matched / paired samples or independent samples? How do you know?

Question 4 (15 points)

Answer either part a or b of this question, based on your response to Question 3.

a. If we have paired samples, what does the distribution of sample paired differences tell us about which inferential procedure to use? Display an appropriate visualization that motivates your conclusions, and then describe those conclusions in complete English sentences.
b. If we have independent samples, what does the distribution of each individual sample tell us about which inferential procedure to use? Display an appropriate visualization that motivates your conclusions, and then describe those conclusions in complete English sentences.

Question 5 (10 points)

Produce an appropriate 95% confidence interval for a relevant population mean that addresses the key question from the study. Be sure to show and describe the R code that led to your selected confidence interval, and describe how your responses to prior Questions led you to select this approach. Save your interpretation of the results for Question 6.

Question 6 (10 points)

Interpret your confidence interval from Question 5 in the context of the request by the project's principal investigator using complete English sentences.

Part B. An Observational Study (Questions 7-9)

The `lab05_lind.Rds` dataset provided on our 431-data page comes from an observational study of 996 patients receiving an initial Percutaneous Coronary Intervention (PCI) at Ohio Heart Health, Christ Hospital, Cincinnati in 1997 and followed for at least 6 months by the staff of the Lindner Center.

The 698 patients thought to be more severely diseased were assigned to treatment with **abciximab** (an expensive, high-molecular-weight IIb/IIIa cascade blocker); while the remaining 298 patients received **usual care** with their initial PCI. Additional information on the lindner data set is available here .^1

> [1] Kereiakes DJ, Obenchain RL, Barber BL, et al. Abciximab provides cost effective survival advantage in high volume interventional practice. Am Heart J 2000; 140: 603-610.

Question 7 (10 points)

Ingest the `lab05_lind.Rds` data into R, and use them to develop an appropriate comparison of the relative risk of an `acutemi` for those receiving abciximab compared to those receiving usual care. Be sure to provide your code, and interpret your results in context in one or two English sentences. Use a 90% confidence level.

A couple of hints for Question 7:

1. You should be changing the variable type and labels to make the results more interpretable (perhaps with `fct_recode()`), as well as change the levels so we are obtaining the probability or odds of a myocardial infarction for those who received abciximab compared to those who received usual care in a contingency table with abciximab status in the rows and acute MI status in the columns.

2. An appropriate contingency table will have the value for subjects who have an acute MI and who are receiving abciximab in the top left, and that cell should contain between 100 and 150 subjects.

Question 8 (10 points)

Now develop an appropriate comparison of the difference in probability of an acute MI for those who have diabetes as compared to those who do not have diabetes. Again, use a 90% confidence level for this Question, provide all necessary code, and interpret your result in context using one or two English sentences.

Hint for Question 8: Make sure the top left cell of your contingency table includes subjects who have an acute MI and also have diabetes, with diabetes status in the rows of your table, thus following standard epidemiological format.

Question 9 (20 points)

Suppose that in a new test sample of 495 patients receiving an initial PCI (like those described in the Lindner Center data) that we obtain the following results for a model we have developed to predict six-month survival using information available at baseline.

- 405 were predicted to survive at least 6 months, and actually survived at least 6 months
- 74 were predicted not to survive at least 6 months, but did actually survive at least 6 months
- 9 were predicted not to survive at least 6 months and did not actually survive at least 6 months.

Specify the appropriate cross-tabulation for predicted and actual survival to 6 months, and then calculate and interpret the accuracy, sensitivity and specificity for the model described here.

Hint: I expect that a close reading of Chapter 6 in Spiegelhalter will be helpful here.

Include the session information

At the end of your R Markdown file, please include a new code chunk to provide the **session information**. You can use either the approach from either Lab 2 or Lab 3.

Submitting the Lab

As mentioned, you should build your entire response as an R Markdown file. Then use the Knit button in RStudio to create the resulting HTML document. Be sure to review the HTML result to ensure that it looks clean and clear, that the labels on your plots and other output are easy to read, and that it doesn't retain any unnecessary warning messages or other material that distracts from your work.

Submit **both** your revised R Markdown file **and** the HTML output file to the Lab 05 section in the Assignments folder in Canvas by the deadline specified in the Course Calendar. We will need both the R Markdown and HTML file submitted before we can grade your work.

Getting Help

You are encouraged to discuss Lab 05 with Professor Love, the teaching assistants or your colleagues, but your answer must be prepared by you alone. Don't be afraid to ask questions, using any of the methods described on our Contact Us page.

Grading

We will summarize some of the more interesting responses to Questions 2 and 5 after the Lab has been graded.

- This Lab will be graded on a scale from 0-100.
- Note that the teaching assistants will review your responses to all Questions carefully to assess clarity of writing, attention to detail, and adherence to grammatical and syntax requirements. Spelling, grammar, syntax and the rest all matter for grading purposes in this and all other assignments this term.

A detailed answer sketch for this Lab will be provided on the day after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.

Late Penalties for Lab Work

- Labs that are turned in 1-12 hours after the deadline will lose 10% of available points.
- Labs turned in more than 12 but less than 72 hours after the deadline will lose 25% of available points.
- No extensions to Lab deadlines will be permitted this semester. Labs turned in more than 72 hours after the deadline will receive no credit.

- Your lowest lab score (out of Labs 1-7) will be dropped before we calculate your lab grade.