431 Lab 05

Deadline: THURSDAY 2023-10-26 at Noon

431 Staff

2023-10-16

Table of contents

0.1 An Important Note: Answer 7 of the 10 Questions!

While there are ten questions listed in Lab 5, everyone will submit responses to **only seven** of the ten questions. In this Lab, you have two options:

- **Option 1**: Answer Questions 1, 2, 3, 4, 5, 9 and 10. (Skip Questions 6-8)
- **Option 2**: Answer Questions 1, 5, 6, 7, 8, 9 and 10. (Skip Questions 2-4)

Select **one** of these two options and submit the responses to your chosen seven (out of the ten available) questions. **There is no benefit to answering all 10 questions available - we will not grade any "extra" responses, so don't include them.**

0.2 Getting Help

You are welcome 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.

0.3 Learning Objectives

1. Demonstrate the use of the `haven` package to ingest a SAS transport file.
2. Be able to take information about a dataset, and an associated visualization, to identify the appropriate inferential test when comparing means using independent or paired samples.
3. Develop and appropriately interpret a confidence interval, in context, as derived from an analysis of categorical variables.
4. Use some concepts developed in Chapter 6 of Spiegelhalter's *The Art of Statistics* to evaluate a model's performance.

0.4 There is no template for Lab 05

Use a new section for each question on the Lab, and do not hide any of your code or results. Use a new section, as well, for the Session Information.

0.5 Our Labs Page

We encourage you in the strongest possible terms to **ask questions**, using any of the approaches described on our Contact Us page.

Visit our 431-Labs page for more information, including:

- how to get help with the Lab,
- our late policy,
- details on answer sketches and grading rubrics
- how your lab will be graded, and
- how we handle requests to regrade a Lab.

0.6 R Packages (I used these in my sketch)

```
library(broom)
library(Epi)
library(haven)
library(Hmisc)
library(janitor)
library(knitr)
library(mosaic)
library(patchwork)
library(tidyverse)

source("https://raw.githubusercontent.com/THOMASELOVE/431-data/main/data-and-code/Love-boo

theme_set(theme_bw())
opts_chunk$set(comment = NA)
```

0.7 Setting Seeds and the Bootstrap

If you need to set a seed, use `431` as your seed. If you need to fit a bootstrap, use 2000 replications.

Part A. NHANES 2001-2002 Data (Questions 1-8)

In Questions 1-8, we are going to work with data from the 2001-2002 administration of the National Health and Nutrition Examination Survey (NHANES). In particular, we will work with a data set we built that includes 1,256 respondents ages 60 and older to that survey, which comprises answers to one question from the Cognitive Functioning Questionnaire, and three questions from the Current Health Status Questionnaire.

First, we will use the `haven` package (part of the tidyverse, but not the core tidyverse, so it most be loaded with `library()` separately) to import the data in the `nh_lab5.xpt` file provided on our 431-data page.

- This is a SAS transport file (version 8) which is a common way SAS users can use to get data into R.
- Create a tibble using these data called `nh_lab5`, with the `read_xpt()` function in the `haven` package.

The variables available in that tibble (the blue links in the table below lead to descriptions of the data at NHANES) should be:

Variable	Description
SEQN	Respondent Code (should be treated as a character)
CFDRIGHT	Correct responses on Digit Symbol Substitution Test
HSQ500	Had a head or chest cold in the last 30 days? (1 = Yes, 2 = No)
HSQ470	# of days (in last 30) when physical health was not good
HSQ480	# of days (in last 30) when mental health was not good

Here is a summary of what the data should look like when initially imported…

```
> summary(nh_lab5)
```

```
    SEQN            CFDRIGHT        HSQ500        HSQ470         HSQ480
Length:1256      Min.   :  1.00  Min.   :1.000 Min.   : 0.000  Min.   : 0.000
Class :character 1st Qu.: 31.00  1st Qu.:2.000 1st Qu.: 0.000  1st Qu.: 0.000
Mode  :character Median : 42.00  Median :2.000 Median : 0.000  Median : 0.000
                 Mean   : 43.28  Mean   :1.826 Mean   : 2.299  Mean   : 1.304
                 3rd Qu.: 56.00  3rd Qu.:2.000 3rd Qu.: 2.000  3rd Qu.: 0.000
                 Max.   :100.00  Max.   :2.000 Max.   :28.000  Max.   :27.000
```

Tasks from the Principal Investigator

You've been asked by the principal investigator of a study to examine two issues and complete two estimation tasks, labeled a and b:

- **Task A**. How large are the differences in the mean number of correct responses on the Digit Symbol Substitution Test between respondents who have had a cold in the past 30 days and those who have not? Please provide a carefully labeled confidence interval, using 90% confidence, to address this issue.
- **Task B**. How large are the differences in the mean number of days a respondent has in the last 30 in which their physical health was not good vs. in which their mental health was not good? Please provide a carefully labeled confidence interval, again using 90% confidence, to address this issue.

1 Question 1 (6 points)

In Task A, are we dealing with independent or paired samples? Specify the reason for your choice in at least two complete, clear English sentences.

An Important Note: Answer FIVE of the first EIGHT Questions!

While there are ten questions listed in Lab 5, everyone will submit responses to **only seven** of the ten questions. In this Lab, you have two options:

- **Option 1**: Answer Questions 1, 2, 3, 4, 5, 9 and 10. (Skip Questions 6-8)
- **Option 2**: Answer Questions 1, 5, 6, 7, 8, 9 and 10. (Skip Questions 2-4)

Select **one** of these two options and submit the responses to your chosen seven (out of the ten available) questions. **There is no benefit to answering all 10 questions available - we will not grade any "extra" responses, so don't include them.**

2 Question 2 (6 points)

Convert the information from the relevant variable in Task A into a factor which has as its levels (Cold and Healthy), providing a sentence explaining your approach, along with your R code. Provide code which clearly specifies which group (Cold or Healthy) has more respondents.

Then provide an appropriate numerical summary of the Task A data that will allow you to calculate the point estimate of your eventual 90% confidence interval. Specify the value of that point estimate, including appropriate units, in a sentence.

3 Question 3 (6 points)

Next, build an appropriate visualization of the Task A data that lets you draw conclusions about whether a parametric confidence interval based on a t distribution, or a non-parametric confidence interval approach based on the bootstrap would be more appropriate. Make sure your visualization has an appropriate (non-default) title, and axis labels, and perhaps a subtitle or caption (if desired.)

Then, in at least two sentences, specify the choice of confidence interval estimate you plan to use, and motivate that choice through information from the visualization.

4 Question 4 (6 points)

Fit the 90% confidence interval you identified in Question 3 to the Task A data, and use it to provide a complete answer for the principal investigator to her question "How large are the differences in the mean number of correct responses on the Digit Symbol Substitution Test between respondents who have had a cold in the past 30 days and those who have not?" based on your confidence interval and your other findings in Questions 1-4. Your answer should include at least two complete sentences.

5 Question 5 (6 points)

In Task B, are we dealing with independent or paired samples? Specify the reason for your choice in at least two complete, clear English sentences.

An Important Note: Answer FIVE of the first EIGHT Questions!

While there are ten questions listed in Lab 5, everyone will submit responses to **only seven** of the ten questions. In this Lab, you have two options:

- **Option 1**: Answer Questions 1, 2, 3, 4, 5, 9 and 10. (Skip Questions 6-8)
- **Option 2**: Answer Questions 1, 5, 6, 7, 8, 9 and 10. (Skip Questions 2-4)

Select **one** of these two options and submit the responses to your chosen seven (out of the ten available) questions. **There is no benefit to answering all 10 questions available - we will not grade any "extra" responses, so don't include them.**

6 Question 6 (6 points)

Provide an appropriate numerical summary of the Task B data that will allow you to calculate the point estimate of your eventual 90% confidence interval. Specify the value of that point estimate, including appropriate units.

7 Question 7 (6 points)

Next, build an appropriate visualization of the Task B data that lets you draw conclusions about whether a parametric confidence interval based on a t distribution, or a non-parametric confidence interval approach based on the bootstrap would be more appropriate. Make sure your visualization has an appropriate (non-default) title, subtitle, and axis labels. Your visualization can show more than one plot, if you use patchwork to put them together. Our answer sketch shows three plots in our visualization for this question.

Then, in at least two sentences, specify the choice of confidence interval estimate you plan to use, and motivate that choice through information from the visualization.

8 Question 8 (6 points)

Fit the 90% confidence interval you identified in Question 7 and use it to provide a complete answer for the principal investigator to her question "How large are the differences in the mean number of days a respondent has in the last 30 in which their physical health was not good vs. in which their mental health was not good?" based on your confidence interval and your other findings in Questions 5-8. Your answer should include at least two complete sentences.

Part B. An Observational Study (Question 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.

The lindner data relate to 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.

9 Question 9 (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 at least two English sentences. Use a 90% confidence level.

A couple of hints for Question 9:

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.

10 Question 10 (10 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's *The Art of Statistics* will be necessary here.

11 Include the session information

At the end of your Quarto file, please include a new code chunk to provide the **session information**. You can use either the `xfun` or `sessioninfo` approach.

12 What Happens After Submission?

Submit **both** your Quarto 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 Quarto and HTML file submitted before we can grade your work. Remember to provide responses for exactly 7 of the 10 questions, as specified above.

There is no benefit to answering all 10 questions available - we will not grade any "extra" responses, so don't include them.

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

- This Lab will be graded on a scale from 0-50 points.
- You should answer seven of the ten questions: either questions 1, 2, 3, 4, 5, 9 and 10 or questions 1, 5, 6, 7, 8, 9 and 10. **Any additional responses submitted will be ignored**.
- 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.

A detailed answer sketch for all 10 questions in this Lab will be provided 48 hours after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.