431 Lab 5 Instructions

Fall 2025 - deadline in Course Calendar

Thomas E. Love

2025-08-08

Table of contents

> **!** Important
>
> - This Lab contains 10 questions, but you will complete 7.
> - The deadline for completing this Lab is posted in the Course Calendar.

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 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.

0.3 Getting Started

To start, create a directory on your computer for `lab5`. We suggest this be a directory you control, called `lab5`, and we recommend you create it as a subdirectory of a `2025-431` directory on your machine.

Now, open RStudio, and use the **File** ... **New Project** ... **Existing Directory** menu to create an R Project in your `lab5` directory in which you will do Lab 5.

0.4 There is no Quarto Template for Lab 5

In this Lab, you will prepare a report in the form of an HTML file, using Quarto. We have provided previous Lab 1 and Lab 2 Quarto document templates. Modify one of those to complete your work for Lab 5, or create something new that works similarly.

0.5 Getting Help

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

0.6 R Packages (I used these in my sketch)

```
library(haven)
library(easystats)
library(tidyverse)

source("data/Love-431.R")

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

0.7 Specifications for Responses

- If you need to set a seed, use `431` as your seed.
- If you need to fit a bootstrap, use 2000 replications.
- Use a **90%** confidence level throughout this Lab.

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 `lab5_nh.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 `lab5_nh`, 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(lab5_nh)
```

```
    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.

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 `lab5_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 `lab5_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 the section entitled "Assessing the Performance of an Algorithm" (where accuracy, sensitivity and specificity are defined) and in particular the material surrounding Table 6.1 from Chapter 6 of Spiegelhalter's *The Art of Statistics* will be necessary here.

11 Next-to-Last Section of your Lab Report: AI Usage

All students should include an AI Usage section in each assignment for this class. See the instructions from Lab 1 for more details.

12 Final Section of your Lab Report: Session Information

Include the session information as a final section in this Lab. I've done so at the bottom of this document.

13 Additional Notes and Instructions

13.1 Submitting this Lab

Submit this Lab via Canvas, using the Lab 5 assignment. Be sure to submit both files:

1. Your Quarto file (.qmd).
2. The HTML file you obtain by knitting the Quarto file (.html)

Be sure that your Quarto (and thus HTML) files include the AI information and session information as separate sections at the end of the document.

13.2 Grading this Lab

This Lab will be graded by the TAs and then reviewed by Dr. Love. Your grades will be available one week after the Lab deadline.

The maximum score on this Lab is 50 points.

As each Lab passes its deadline (as listed in the Course Calendar), we will:

- post the answer sketch (48 hours after the deadline) and draft grading rubric to our Shared Google Drive, and then
- post grades and any revisions to the grading rubric or answer sketch one week after the deadline to a location we will provide to you.

13.3 Emergencies and Late Policy

We do not grant extensions on Lab deadlines.

- 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.)
  - Labs that are turned in 1-48 hours after the deadline will lose 10 points for late work.
- No extensions to Lab deadlines will be made this semester. Labs turned in more than 48 hours after the deadline will receive no credit, since by then the Lab Sketch will be posted.
- Your lowest lab score (out of Labs 1-6) over the course of the semester will be dropped before we calculate your lab grade.

If you have an emergency that will keep you from submitting the Lab by even the late deadline of Friday at noon, please let Dr. Love know that (as soon as possible) via email and he will consider excusing you from the Lab.

13.4 Lab Regrade Requests

If, after your Lab is graded, you want Dr. Love to review the grading or correct a grading error, please follow the Lab Regrade Request policy posted on our Labs page.

14 Session Information

At the end of your Quarto file, you should run session information, like this.

```
xfun::session_info()
```

```
R version 4.5.1 (2025-06-13 ucrt)
Platform: x86_64-w64-mingw32/x64
Running under: Windows 11 x64 (build 26100)

Locale:
  LC_COLLATE=English_United States.utf8
  LC_CTYPE=English_United States.utf8
  LC_MONETARY=English_United States.utf8
  LC_NUMERIC=C
  LC_TIME=English_United States.utf8

Package version:
  askpass_1.2.1       backports_1.5.0     base64enc_0.1.3
  bayestestR_0.16.1   bit_4.6.0           bit64_4.6.0.1
  blob_1.2.4          broom_1.0.9         bslib_0.9.0
  cachem_1.1.0        callr_3.7.6         cellranger_1.1.0
  cli_3.6.5           clipr_0.8.0         coda_0.19-4.1
  codetools_0.2-20    compiler_4.5.1      conflicted_1.2.0
  correlation_0.8.8   cpp11_0.5.2         crayon_1.5.3
  curl_6.4.0          data.table_1.17.8   datasets_4.5.1
  datawizard_1.2.0    DBI_1.2.3           dbplyr_2.5.0
  digest_0.6.37       dplyr_1.1.4         dtplyr_1.3.1
  easystats_0.7.5     effectsize_1.0.1    emmeans_1.11.2
  estimability_1.5.1  evaluate_1.0.4      farver_2.1.2
  fastmap_1.2.0       fontawesome_0.5.3   forcats_1.0.0
```

fs_1.6.6	gargle_1.5.2	generics_0.1.4
ggplot2_3.5.2	glue_1.8.0	googledrive_2.1.1
googlesheets4_1.1.1	graphics_4.5.1	grDevices_4.5.1
grid_4.5.1	gtable_0.3.6	haven_2.5.5
highr_0.11	hms_1.1.3	htmltools_0.5.8.1
httr_1.4.7	ids_1.0.1	insight_1.3.1
isoband_0.2.7	jquerylib_0.1.4	jsonlite_2.0.0
knitr_1.50	labeling_0.4.3	lattice_0.22-7
lifecycle_1.0.4	lubridate_1.9.4	magrittr_2.0.3
MASS_7.3-65	Matrix_1.7-3	memoise_2.0.1
methods_4.5.1	mgcv_1.9.3	mime_0.13
modelbased_0.12.0	modelr_0.1.11	multcomp_1.4-28
mvtnorm_1.3-3	nlme_3.1.168	numDeriv_2016.8.1.1
openssl_2.3.3	parameters_0.27.0	patchwork_1.3.1
performance_0.15.0	pillar_1.11.0	pkgconfig_2.0.3
prettyunits_1.2.0	processx_3.8.6	progress_1.2.3
ps_1.9.1	purrr_1.1.0	R6_2.6.1
ragg_1.4.0	rappdirs_0.3.3	RColorBrewer_1.1-3
readr_2.1.5	readxl_1.4.5	rematch_2.0.0
rematch2_2.1.2	report_0.6.1	reprex_2.1.1
rlang_1.1.6	rmarkdown_2.29	rstudioapi_0.17.1
rvest_1.0.4	sandwich_3.1-1	sass_0.4.10
scales_1.4.0	see_0.11.0	selectr_0.4.2
splines_4.5.1	stats_4.5.1	stringi_1.8.7
stringr_1.5.1	survival_3.8-3	sys_3.4.3
systemfonts_1.2.3	textshaping_1.0.1	TH.data_1.1-3
tibble_3.3.0	tidyr_1.3.1	tidyselect_1.2.1
tidyverse_2.0.0	timechange_0.3.0	tinytex_0.57
tools_4.5.1	tzdb_0.5.0	utf8_1.2.6
utils_4.5.1	uuid_1.2.1	vctrs_0.6.5
viridisLite_0.4.2	vroom_1.6.5	withr_3.0.2
xfun_0.52	xml2_1.3.8	xtable_1.8-4
yaml_2.3.10	zoo_1.8-14	