431 Lab 2 Instructions

Fall 2025 - deadline in Course Calendar

Thomas E. Love

2025-08-07

Table of contents

❗ Important

- This Lab contains 3 tasks for you to complete.
- The deadline for completing this Lab is posted in the Course Calendar.

0.1 R Setup

Dr. Love used the follwing R setup to write his answer sketch for Lab 2.

```
library(infer)
library(knitr)
library(patchwork)
library(rstanarm)

library(easystats)
library(tidyverse)

theme_set(theme_bw())

source("Love-431.R")
```

0.2 Getting Started

To start, create a directory on your computer for `lab2`. We suggest this be a directory you control, called `lab2`, and we recommend you create it as a subdirectory of a `2025-431` directory on your machine.

- Into that `lab2` directory, you will download the Quarto Template for Lab 2, which is called `lab2-template.qmd`, and described below. I would then rename the file to include in your actual name in the file name, perhaps calling it `yourname-lab2.qmd`.

Now, open RStudio, and use the **File … New Project … Existing Directory** menu to create an R Project in your `lab2` directory in which you will do Lab 2.

0.3 The Quarto Template for Lab 2

In this Lab, you will prepare a report in the form of an HTML file, using Quarto. We have provided a Quarto document template called `lab2-template.qmd` that you should use to complete your work.

- The template is part of the Data and Code repository for the course. Follow the instructions posted there to download all of the files you'll need in a ZIP file, including the template, to an easy place to find them on your computer (we suggest a `431-data` subdirectory in your `2025-431` directory.) Then copy the template and the necessary data files (`lab2-task1.csv` and `lab2-task2.csv`) into the directory for Lab 2 that you created earlier.

Build your response to Tasks 1-3 using the Quarto template provided. Use the Render button in RStudio to compile your work and create the HTML output. You'll want to do this multiple times as you go, to identify potential problems quickly.

> ! Important
>
> Delete **all of the instructions** we provide to you in the template, in favor of your own words, before submitting your work. You are welcome to retain any or all of the R code we provide in the template as part of your response.

0.4 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.

0.5 Using AI / ChatGPT, etc.

If you decide to get help from a large language model (like ChatGPT) to help with your phrasing of ideas, or building code, OK, but you need to describe what you did carefully in the designated **AI Help** section (just before the Session Information) of your submission.

0.6 Learning Objectives for this Lab

1. Ingest data from a comma-separated version (.csv) file into R.
2. Obtain appropriate visual and numerical summaries of the distribution of a sample of paired differences.
3. Obtain appropriate visual and numerical summaries of the distribution of two independent samples.
4. Generate appropriate point and confidence interval estimates for means using the bootstrap as well as Bayes and least squares linear models.
5. Demonstrate your understanding of some of what you've been reading in David Spiegelhalter's *The Art of Statistics*.
6. Communicate effectively about your results in complete, clear English sentences.

1 Task 1 (19 points)

1.1 The Data

We will use the data in the `lab2-task1.csv` file, which contains the following five variables for each of 3,082 US counties. I developed these data through the County Health Rankings reports for two recent years.

Variable	Description
`FIPS`	Federal Information Processing Standard (FIPS) Code describing the state and county (default sorting order)
`state`	two-letter abbreviation for state
`county`	county name
`prem_2020`	Years of potential life lost before age 75 per 100,000 population (age-adjusted) in the county for 2018-2020
`prem_2016`	Years of potential life lost before age 75 per 100,000 population (age-adjusted) in the county for 2014-2016

The premature death rates are estimated by County Health Rankings using the National Center for Health Statistics' Mortality files. The first four observations in the data set are:

fips	state	county	prem_2020	prem_2016
FIPS_01001	AL	Autauga County	8027.4	9409.3
FIPS_01003	AL	Baldwin County	8118.4	7467.6
FIPS_01005	AL	Barbour County	12876.8	8929.5
FIPS_01007	AL	Bibb County	11191.5	11741.9

1.2 The Task

Complete these three subtasks:

a. Produce appropriate R code to add a column of paired (2018-2020) minus (2014-2016) differences to the data for the entire sample of 3082 counties. Then create a useful visualization of the distribution of the paired differences, as well as a useful numerical summary of those differences. Note: if you're building a histogram with a superimposed Normal curve of these differences, a binwidth of about 500 seems appropriate.

b. Use the complete sample to build a small table which specifies the point estimate and 95% confidence interval for the mean of the paired differences using three different estimation approaches, specifically:

- an ordinary least squares linear model (e.g. a paired t comparison),
- the bootstrap with seed set to `4311`, and using 1000 replications, and
- using a Bayesian linear model with seed set to `4312` using the default choice of weakly informative priors

c. In two to four sentences, compare the three results you obtained in task 1b in light of the visualizations you built in task 1a. What can we say about the center, spread and shape of the distribution that might be relevant to the estimation we did, and what can we conclude about the three different estimation methods as applied to this sample of 3082 counties?

2 Task 2 (19 points)

2.1 The Data

The `lab2-task2.csv` file contains these six variables for 1,328 US counties. I developed these data through the County Health Rankings report for 2023.

Variable	Description
`FIPS`	Code describing the state and county (default sorting order)
`state`	two-letter abbreviation for state
`county`	county name
`vote2020`	Voter Turnout (% of adult residents) for 2020 Presidential Election
`firearm_cat`	“Above 17” or “Below 10” from `firearm_rate`
`firearm_rate`	Annual Rate of Firearm Fatalities per 100,000 county residents, 2016-2020

Counties were excluded from the Task 2 sample if their `firearm_rate` was between 10 and 17 fatalities per 100,000 county residents. The firearm fatality rates are estimated by County Health Rankings using the National Center for Health Statistics' Mortality files. The first four observations in the data set are:

FIPS	state	county	vote2020	firearm_cat	firearm_rate
FIPS_01001	AL	Autauga County	66.2	Above 17	17.59
FIPS_01005	AL	Barbour County	54.0	Above 17	25.52
FIPS_01007	AL	Bibb County	54.6	Above 17	17.82
FIPS_01009	AL	Blount County	64.2	Above 17	20.40

2.2 The Task

Complete these three subtasks:

a. Produce appropriate R code to create a useful visualization of the distribution of (`vote2020`) the percentage of eligible adult voters who turned out for the 2020 presidential election, within each of the two `firearm_cat` subgroups, and also provide a useful numerical summary of `vote2020` within each firearm fatality subgroup.

b. Build a small table which specifies the point estimate and 95% confidence interval for the mean difference (Above 17 group - Below 10 group) in `vote2020` using three different estimation approaches, specifically:

   - an ordinary least squares linear model (e.g. a pooled t comparison),
   - a Welch t procedure,
   - the bootstrap with seed set to `4313`, and 1000 bootstrap replications, and
   - using a Bayesian linear model with seed set to `4314` using the default choice of weakly informative priors

c. In two to four sentences, compare the four results you obtained in task 2b in light of the visualizations and summaries you built in task 2a. What can we say about the center, spread and shape of the distribution and the number of counties in each subgroup that might be relevant to the estimation we did, and what can we conclude about the different estimation methods as applied to this sample of counties?

3 Task 3 (12 points)

By now, you should have read a substantial chunk of David Spiegelhalter's *The Art of Statistics*, including the introduction and Chapters 1-3 and 5. Reflecting on Chapter 3 of *The Art of Statistics*, please write a brief essay (100-200 words) that discusses the process of inductive inference and how that influences the conclusions we can draw from our work in Task 2. A major issue to consider is how the sample was taken.

Your response should be written using clear and complete English sentences and minimizing jargon.

4 Additional Notes and Instructions

4.1 Submitting this Lab

Submit this Lab via Canvas, using the Lab 2 assignment. Be sure to submit both files:

1. Your Quarto file (.qmd) built using our Lab 2 template.
2. The HTML file you obtain by knitting the Quarto file (.html)

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

4.2 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.

4.3 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.

4.4 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.

4.5 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:
  abind_1.4-8          askpass_1.2.1        backports_1.5.0
  base64enc_0.1-3      bayesplot_1.13.0     bayestestR_0.16.1
  BH_1.87.0.1          bit_4.6.0            bit64_4.6.0-1
  blob_1.2.4           boot_1.3-31          broom_1.0.9
  bslib_0.9.0          cachem_1.1.0         callr_3.7.6
  cellranger_1.1.0     checkmate_2.3.2      cli_3.6.5
  clipr_0.8.0          codetools_0.2-20     colourpicker_1.3.0
  commonmark_2.0.0     compiler_4.5.1       conflicted_1.2.0
  correlation_0.8.8    cpp11_0.5.2          crayon_1.5.3
  crosstalk_1.2.1      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         desc_1.4.3           digest_0.6.37
  distributional_0.5.0 dplyr_1.1.4          DT_0.33
  dtplyr_1.3.1         dygraphs_1.1.1.6     easystats_0.7.5
```

```
effectsize_1.0.1     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        ggridges_0.5.6       glue_1.8.0
googledrive_2.1.1    googlesheets4_1.1.1  graphics_4.5.1
grDevices_4.5.1      grid_4.5.1           gridExtra_2.3
gtable_0.3.6         gtools_3.9.5         haven_2.5.5
highr_0.11           hms_1.1.3            htmltools_0.5.8.1
htmlwidgets_1.6.4    httpuv_1.6.16        httr_1.4.7
ids_1.0.1            igraph_2.1.4         infer_1.0.9
inline_0.3.21        insight_1.3.1        isoband_0.2.7
jquerylib_0.1.4      jsonlite_2.0.0       knitr_1.50
labeling_0.4.3       later_1.4.2          lattice_0.22-7
lazyeval_0.2.2       lifecycle_1.0.4      litedown_0.7
lme4_1.1-37          loo_2.8.0            lubridate_1.9.4
magrittr_2.0.3       markdown_2.0         MASS_7.3-65
Matrix_1.7-3         matrixStats_1.5.0    memoise_2.0.1
methods_4.5.1        mgcv_1.9.3           mime_0.13
miniUI_0.1.2         minqa_1.2.8          modelbased_0.12.0
modelr_0.1.11        nlme_3.1-168         nloptr_2.2.1
numDeriv_2016.8.1.1  openssl_2.3.3        parallel_4.5.1
parameters_0.27.0    patchwork_1.3.1      performance_0.15.0
pillar_1.11.0        pkgbuild_1.4.8       pkgconfig_2.0.3
plyr_1.8.9           posterior_1.6.1      prettyunits_1.2.0
processx_3.8.6       progress_1.2.3       promises_1.3.3
ps_1.9.1             purrr_1.1.0          QuickJSR_1.8.0
R6_2.6.1             ragg_1.4.0           rappdirs_0.3.3
rbibutils_2.3        RColorBrewer_1.1-3   Rcpp_1.1.0
RcppEigen_0.3.4.0.2  RcppParallel_5.1.10  Rdpack_2.6.4
readr_2.1.5          readxl_1.4.5         reformulas_0.4.1
rematch_2.0.0        rematch2_2.1.2       report_0.6.1
reprex_2.1.1         reshape2_1.4.4       rlang_1.1.6
rmarkdown_2.29       rstan_2.32.7         rstanarm_2.32.1
rstantools_2.4.0     rstudioapi_0.17.1    rvest_1.0.4
sass_0.4.10          scales_1.4.0         see_0.11.0
selectr_0.4.2        shiny_1.11.1         shinyjs_2.1.0
shinystan_2.6.0      shinythemes_1.2.0    sourcetools_0.1.7.1
splines_4.5.1        StanHeaders_2.32.10  stats_4.5.1
stats4_4.5.1         stringi_1.8.7        stringr_1.5.1
survival_3.8-3       sys_3.4.3            systemfonts_1.2.3
tensorA_0.36.2.1     textshaping_1.0.1    threejs_0.3.4
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          V8_6.0.5
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        xts_0.14.1          yaml_2.3.10
zoo_1.8-14
```