431 Lab 1 Instructions

Fall 2025 - deadline in Course Calendar

Thomas E. Love

2025-08-08

Table of contents

> **!** Important
>
> - This Lab contains 5 tasks for you to complete.
> - The deadline for completing this Lab is posted in the Course Calendar.

0.1 R Setup

```
library(tidyverse)
```

You will use additional packages in developing your response.

0.2 Getting Started

To start, create a directory on your computer for `lab1`. We suggest this be a directory you control, called `lab1`, and we recommend you create it as a subdirectory of a `2025-431` directory on your machine.

- Into that `lab1` directory, you will download the Quarto Template for Lab 1, which is called `lab1-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-lab1.qmd`.

Now, open RStudio, and use the **File** … **New Project** … **Existing Directory** menu to create an R Project in your `lab1` directory in which you will do Lab 1.

0.3 The Quarto Template for Lab 1

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 `lab1-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 into the directory for Lab 1 that you created earlier.

Build your response to Tasks 1-4 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.

See the **AI Usage** section after the main questions in this Lab.

0.6 Learning Objectives for this Lab

1. Build a Quarto document using the template we have provided.
2. Manage a data set pre-loaded in R to obtain necessary results.
3. Visualize a batch of data, as well as multiple batches.
4. Obtain summary statistics describing an association, and the distribution of a sample of data, while excluding missing values.
5. Obtain a confidence interval for a population mean of a single sample, using a linear model.
6. Communicate effectively about your R results in complete, clear English sentences.

1 Task 1 (10 points)

1.1 The Data

> **Note**
>
> You will use the `cms_patient_experience` data for Tasks 1-4.

The first 6 rows of the `cms_patient_experience` data file (tibble) are printed below. These data are part of the `tidyr` package, which is part of the `tidyverse` meta-package, so the file should be available to you after you have loaded the `tidyverse`.

```
cms_patient_experience |> head()
```

```
# A tibble: 6 x 5
  org_pac_id org_nm                     measure_cd   measure_title      prf_rate
  <chr>      <chr>                      <chr>        <chr>                 <dbl>
1 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_1  CAHPS for MIPS SS~       63
2 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_2  CAHPS for MIPS SS~       87
3 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_3  CAHPS for MIPS SS~       86
4 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_5  CAHPS for MIPS SS~       57
5 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_8  CAHPS for MIPS SS~       85
6 0446157747 USC CARE MEDICAL GROUP INC CAHPS_GRP_12 CAHPS for MIPS SS~       24
```

The `cms_patient_experience` tibble contains some lightly cleaned data from the Centers for Medicare & Medicaid Services, specifically its "Hospice - Provider Data" as of a particular moment in time, which provides a list of hospice agencies along with some data on quality of patient care. More details are available at the tidyr reference page.

In particular, `cms_patient_experience` is a data frame (tibble) with 500 observations and five variables:

- `org_pac_id`: organization's ID number
- `org_nm`: organization's name
- `measure_cd`: measure code number
- `measure_title`: measure title
- `prf_rate`: measure performance rate

In all, 95 health care organizations are included in this sample. Some provide information on all five performance measures, but not all organizations have done so (in other words, there are missing data to deal with here.)

The performance rates (`prf_rate`) values are between 0 and 100 with higher values indicating stronger (better) performance according to the people who completed the surveys. There are no units to these measures other than "points" on a scale.

Here's a description of the measure code numbers and titles:

> **Note**
>
> - **CAHPS** is the Consumer Assessment of Healthcare Providers & Systems
> - **MIPS** is the Merit-based Incentive Payment System
> - **SSM** means Summary Survey Measure

measure_cd	measure_title
CAHPS_GRP_1	CAHPS for MIPS SSM: Getting Timely Care, Appointments, and Information
CAHPS_GRP_2	CAHPS for MIPS SSM: How Well Providers Communicate
CAHPS_GRP_3	CAHPS for MIPS SSM: Patient's Rating of Provider
CAHPS_GRP_5	CAHPS for MIPS SSM: Health Promotion and Education
CAHPS_GRP_8	CAHPS for MIPS SSM: Courteous and Helpful Office Staff
CAHPS_GRP_12	CAHPS for MIPS SSM: Stewardship of Patient Resources

1.2 The Task

Your Task 1 is to ingest the data into R, then obtain a numerical summary of the information provided on the SSM for "Patient's Rating of Provider" (`measure_cd`: CAHPS_GRP_3) across the available organizations.

> **What is needed in the Summary?**
>
> Your numerical summary should include, at least, the:
>
> - number of observations on this measure (including missing values)
> - number of missing values on this measure
> - minimum, maximum, first and third quartiles (25th and 75th percentiles) of the sample, excluding the missing values
> - mean and standard deviation of the sample, excluding the missing values
> - median and mad of the sample, excluding the missing values

Having generated these summaries (*rounding to one decimal place would be very appropriate here*), then write two or three complete English sentences comparing the sample mean and standard deviation to the sample median and mad, describing the implications of what you see in these numerical summaries. **See the hints on the next page**.

Hints for Task 1

1. You will need to use the `filter()` function as part of your response.
2. The score on the "Patient's Rating of Provider" for `USC CARE MEDICAL GROUP`, as we can see from the listing of the data above, is 86 points.
3. The sum of the number of missing values and the number of non-missing values for this measure should be 95.
4. You should find multiple missing values for this measure.
5. The sample median of the non-missing values for this measure should be 83.

2 Task 2 (8 points)

In Task 2, we again make use of the `cms_patient_experience` tibble from Task 1. Use a linear model (fit with `lm()`) to obtain a point estimate and 95% confidence interval for the mean `prf_rate` for the "Patient's Rating of Provider" (`measure_cd`: CAHPS_GRP_3) across the available organizations, with each estimate rounded to one decimal place. Be sure to specify in your response your point and interval estimate appropriately in a complete English sentence that also tells us how many (non-missing) observations were used to fit the model.

3 Task 3 (12 points)

In Task 3, we again use the `cms_patient_experience` tibble. Here, though, we aim to compare the scores on all of the six different measures using a comparison boxplot including violins.

Create an attractive, legible and easy to understand version of such a plot and display it, incorporating useful titles and axis labels.

Then write a few sentences which describe the conclusions you draw from the plot regarding the center and range of the sample values across the various measures.

4 Task 4 (15 points)

4.1 Building a Subset and Initial Plot

In the Lab 1 template, we provide the following code for you to use in building a scatterplot of the association between the "How Well Providers Communicate" measure and the "Courteous and Helpful Office Staff".

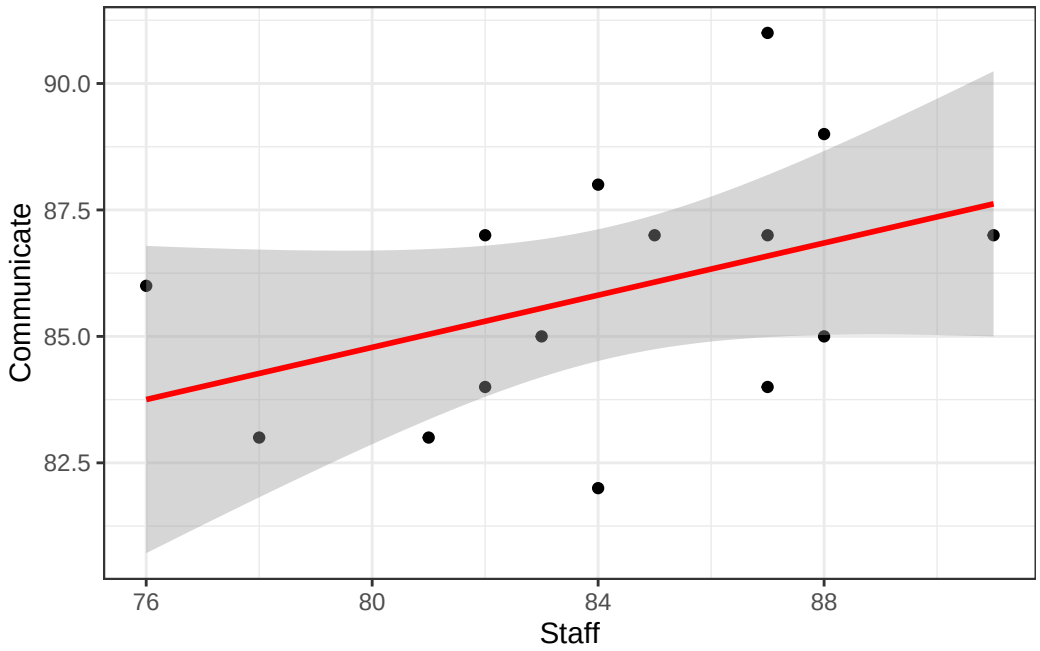
```
task4dat <- cms_patient_experience |>
  filter(measure_cd == "CAHPS_GRP_2" | measure_cd == "CAHPS_GRP_8") |>
  pivot_wider(names_from = c(measure_cd, measure_title),
              values_from = prf_rate) |>
  rename("Communicate" = "CAHPS_GRP_2_CAHPS for MIPS SSM: How Well Providers Communicate",
         "Staff" = "CAHPS_GRP_8_CAHPS for MIPS SSM: Courteous and Helpful Office Staff") |>
  drop_na()

head(task4dat)
```

```
# A tibble: 6 x 4
  org_pac_id org_nm                                Communicate Staff
  <chr>      <chr>                                       <dbl> <dbl>
1 0446157747 USC CARE MEDICAL GROUP INC                     87    85
2 0446162697 ASSOCIATION OF UNIVERSITY PHYSICIANS           85    88
3 0749333730 CAPE PHYSICIANS ASSOCIATES PA                  84    82
4 0840104360 ALLIANCE PHYSICIANS INC                        87    87
5 0840109864 REX HOSPITAL INC                               87    91
6 0840513552 SCL HEALTH MEDICAL GROUP DENVER LLC            83    78
```

Having rearranged the data into `task4dat`, the starting code for the scatterplot we want you to build is found on the next page.

```
ggplot(task4dat, aes(x = Staff, y = Communicate)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x, se = TRUE, col = "red") +
  theme_bw()
```



4.2 The Task

1. Specify and then interpret (in a couple of sentences) both the slope and intercept of the fitted line shown in the plot above. Round your estimates to two decimal places.
2. Replot the data, adding:
    a. a meaningful title which describes the key finding of the plot, and a subtitle which specifies the number of observations shown in the plot, and
    b. a loess smooth to the plot, in blue, without a standard error ribbon.

Display your resulting new plot, and then write a sentence or two about what you learn by comparing your blue and red smooths.

5 Task 5 (5 points)

Make a 30-second video to help us pronounce your name and learn something interesting about you. Recording yourself using Zoom is a solid approach. In this video, we should see your face clearly and hear you clearly, so make sure we will. Save and send the video as either an **.mp4** or **.mov** file.

You will be doing two things in the video.

1. First, say hello, and then state your name, loudly and clearly, so that the viewer can learn to pronounce it correctly. Please use both your given name and your family name. If you prefer to be called by a nickname, please specify that, too.
2. Then, tell us something about you that we don't know, and might be interested to learn. It can be anything at all. We're hoping to get to know you a little better, and this can be something about your professional or private life, or whatever you feel you would like to share with us. We're hoping to facilitate connections here.

Some tips for the video

- Your fellow students (as well as the Teaching Assistants and Dr. Love) will see your video later this semester, so keep that in mind as you decide what to tell us.
- Do not worry about adding elaborate theatricality, props or scenery. If you'd like to do so, that's fine, but just make sure that we can see you and hear what you are saying clearly. That's what matters.
- We're not kidding about the 30-second time limit. Do not exceed 30 seconds.
- In the Quarto file, all you need to do for Task 5 is the name of the video file you are submitting. The name `yourname-431lab1.mp4` is appropriate, after substituting in your name.

6 Section 6 of your Lab Report: AI Usage

All students should include an AI Usage section in each assignment for this class.

- 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 in some detail here. Use multiple clear and complete sentences.
- If you use nothing outside of the spell-checker and code completion tools within RStudio, please write that here, as follows:

In preparing this response, I made no use of AI outside of spell-check and code completion within RStudio.

7 Section 7 of your Lab Report: Session Information

At the end of your Quarto file, you should run session information. The code required is shown below.

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

The results of using this code are shown at the end of this document.

8 Additional Notes and Instructions

8.1 Submitting this Lab

Submit this Lab via Canvas, using the Lab 1 assignment. Be sure to submit all three files:

1. Your Quarto file (.qmd) built using our Lab 1 template.
2. The HTML file you obtain by knitting the Quarto file (.html)
3. Your video file (.mp4 or .mov preferred)

Be sure that your Quarto (and thus HTML) files include the **AI Help** and **Session information** sections at the end of the document, and that these section headings appear in the Table of Contents.

8.2 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.

8.3 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.

8.4 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.

9 Session Information

```
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
  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         compiler_4.5.1      conflicted_1.2.0
  cpp11_0.5.2         crayon_1.5.3        curl_6.4.0
  data.table_1.17.8   DBI_1.2.3           dbplyr_2.5.0
  digest_0.6.37       dplyr_1.1.4         dtplyr_1.3.1
  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           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           modelr_0.1.11       nlme_3.1-168
  openssl_2.3.3       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      reprex_2.1.1        rlang_1.1.6
  rmarkdown_2.29      rstudioapi_0.17.1   rvest_1.0.4
  sass_0.4.10         scales_1.4.0        selectr_0.4.2
  splines_4.5.1       stats_4.5.1         stringi_1.8.7
  stringr_1.5.1       sys_3.4.3           systemfonts_1.2.3
  textshaping_1.0.1   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
  yaml_2.3.10
```