431 Lab 4 Instructions

Fall 2025 - deadline in Course Calendar

Thomas E. Love

2025-10-09

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 `lab4`. We suggest this be a directory you control, called `lab4`, 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 `lab4` directory in which you will do Lab 4.

0.3 There is no Quarto Template for Lab 4

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 4, or create something new that works similarly.

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 Learning Objectives for this Lab

1. Ingest data from an Excel (.xlsx) file into R
2. Complete an appropriate analysis of a two-way contingency table.
3. Reflect on some key issues regarding study design discussed in Spiegelhalter's *The Art of Statistics*.
4. Compare distributions of three independent samples visually and numerically.
5. Build an appropriate analysis of variance model, and use it to motivate comparisons overall and across pairs of groups.
6. Communicate effectively about your results in complete, clear English sentences.

1 Task 1 (10 points)

1.1 The Data

For Tasks 1 and 2, we use the data in the `lab4-task1.xlsx` (Excel) file, which is derived from the LicoriceGargle example at the Cleveland Clinic Statistical Dataset Repository which provides two files to explain the data.

- a general description, and
- a data description.

The data were originally published in Ruetzler K et al. (2013) A randomized, double-blind comparison of licorice versus sugar-water gargle for prevention of postoperative sore throat and postextubation coughing *Anesth Analg* 117: 614-21. https://doi.org/10.1213/ane.0b013e318299a650.

From the Abstract to the Ruetzler paper, we have the following details:

> We enrolled 236 patients having elective thoracic surgery who required intubation with a double-lumen endotracheal tube. Patients were randomly assigned to gargle 5 minutes before induction of anesthesia for 1 minute with: (1) Extractum Liquiritiae Fluidum (licorice 0.5 g); or (2) Sirupus Simplex (sugar 5 g); each diluted in 30 mL water.

Our small sample of the original data includes only 235 rows and three variables:

Variable	Description
`subject`	Subject ID, added by Dr. Love
`intervention`	sugar or licorice (from `treat` in the CCF dataset)
`postop4`	Yes or No regarding sore throat pain at rest 4 hours after surgery (from `postOp4hour_throatpain`)

1.2 The Task

Compare the rates of post-operative sore throat pain among subjects who gargled with licorice as compared to those who gargled with sugar. In particular, create and share an appropriate statistical summary of the relationship, and use appropriate statistical methods to estimate the point estimate and 95% confidence intervals for an appropriate:

- relative risk,
- odds ratio, and
- risk difference.

In this case, do not use a Bayesian augmentation.

2 Task 2 (15 points)

Reflecting on what you've done in Task 1 and what you've read in Spiegelhalter's *The Art of Statistics*, please write an essay of 125-200 words which discusses the relationship between `postop4` and `intervention` that you observe in this sample.

In your essay, first, describe what you have learned about the association of licorice (vs. sugar) and sore throat pain 4 hours post-surgery from this work, avoiding jargon and also avoiding statements about statistical significance. Then, discuss how what we know about the study design tells us about what we might be able to conclude about whether gargling with licorice reduces pain at 4 hours post-surgery.

3 Task 3 (10 points)

3.1 The Data

Tasks 3 and 4 use data gathered in the `lab4-task3.xlsx` Excel file, which we have simulated to reflect a hypothetical clinical trial. In this trial, the investigators are testing out a new drug to see its effect on a subject's systolic blood pressure (SBP). For our purposes, if a subject's SBP is over 130 mm Hg they are considered to have hypertension.

The goal was to reduce the SBP from a baseline level by using a new drug (Treatment C), and comparing that to the current top-of-the-line drug (Treatment B), and the oldest drug (Treatment A). The trial focused specifically on non-Hispanic African-American women who were in long-term relationships and between the ages of 55 and 65 years old, with a minimal comorbidity profile.

The outcome of interest is the post-treatment systolic blood pressure (`sbp_follow`), and we are also given the subject's age, their pre-treatment systolic blood pressure (`sbp_baseline`) and whether or not the subject's partner has hypertension.

Variable	Description
`subjectid`	unique subject identifier
`group_rx`	treatment group where 1 = Group A, 2 = Group B, 3 = Group C
`partner`	whether or not the subject's partner also has hypertension
`age`	subject age in years at baseline
`sbp_baseline`	subject's baseline systolic blood pressure (mm Hg)
`sbp_follow`	subject's follow-up systolic blood pressure (mm Hg)

3.2 The Task

Ingest the data, and make `group_rx` into a factor variable called `group_f` which has levels "Group_A", "Group_B", and "Group_C" as appropriate.

Then produce one or two visualizations comparing the three treatment groups. Your goal should be to assess the assumptions of an analysis of variance which compares the SBP at follow-up (`sbp_follow`) means across the `group`s, ignoring all other information in the data.

Then, produce a numerical summary for `sbp_follow` within each group that includes, at minimum, the following summaries:

- sample size, mean, standard deviation, median, MAD, minimum and maximum

After you have completed this coding, write a paragraph describing your results. Specifically, your paragraph should tell us what conclusions can you draw about the center, spread and shape of the data in each group, and what you conclude about ANOVA model assumptions in light of your summaries.

3.2.1 Two Hints for Task 3

1. An appropriate way to describe center, spread and shape involves a graph (or set of them) that show us the distribution of the outcome within each group. There are lots of available options. You could compare (a) histograms, or compare (b) boxplots, or compare (c) Normal Q-Q plots, for example. You might want to look at more than one of these options, but I wouldn't think you would need all three, though. Don't include plots that you're not going to discuss in your paragraph describing your results.
2. The primary ANOVA assumptions that we're looking at in Task 3 are linearity, homogeneity of variance, and Normality. You could do this with `check_model()` on the linear model supporting your ANOVA, or you could do this without `check_model()` through the combination of (a) a discussion at the plot of the `sbp_follow` data by `group` you developed earlier in Task 3, and (b) a Normal Q-Q plot of the residuals from the linear model supporting your ANOVA. Again, though, you shouldn't include plots that you're not going to discuss in your description of the results.

4 Task 4 (10 points)

Now complete the comparison of the SBP at follow-up means of the three treatment groups (A, B and C) described in Task 3 using an analysis of variance, backed by an OLS linear model.

What conclusions do you draw about the groups (a) overall, and (b) in appropriate pairwise pre-planned comparisons[1], using a **90%** confidence level? Write a paragraph to describe your results, using clear and complete English sentences and minimizing jargon.

5 Task 5 (5 points)

Here, you will share something useful that you learned from reading our recommended sections of R for Data Science, including a specific quote from the book.

> **Note**
>
> By now, we hope you have read the introduction and sections 1-11, 16 and 28 of R for Data Science.

Your response should include a useful quote or other reference from the book (specify the section number for the quote), plus two or three clear and complete English sentences that tell us why you found this to be particularly helpful.

6 Section 6 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.

7 Section 7 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.

8 Additional Notes and Instructions

8.1 Submitting this Lab

Submit this Lab via Canvas, using the Lab 4 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)

[1] A Tukey HSD approach is appropriate here.

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.

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.

8.5 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:
  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_1.0.10        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_7.0.0
  data.table_1.17.8   DBI_1.2.3           dbplyr_2.5.1
  digest_0.6.37       dplyr_1.1.4         dtplyr_1.3.2
  evaluate_1.0.5      farver_2.1.2        fastmap_1.2.0
  fontawesome_0.5.3   forcats_1.0.1       fs_1.6.6
  gargle_1.6.0        generics_0.1.4      ggplot2_4.0.0
  glue_1.8.0          googledrive_2.1.2   googlesheets4_1.1.2
  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
  lifecycle_1.0.4     lubridate_1.9.4     magrittr_2.0.4
  memoise_2.0.1       methods_4.5.1       mime_0.13
  modelr_0.1.11       openssl_2.3.4       pillar_1.11.1
  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_1.5.0          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_2.30	rstudioapi_0.17.1
rvest_1.0.5	S7_0.2.0	sass_0.4.10
scales_1.4.0	selectr_0.4.2	stats_4.5.1
stringi_1.8.7	stringr_1.5.2	sys_3.4.3
systemfonts_1.3.1	textshaping_1.0.3	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_1.6.6	withr_3.0.2	xfun_0.53
xml2_1.4.0	yaml_2.3.10	