431 Lab 03

Deadline: Tuesday 2023-09-26 at Noon

431 Staff

2023-08-10

Table of contents

0.1 Getting Help

Lab 03 has 6 questions, all of which you need to complete by the deadline specified on the Course Calendar.

You are welcome to discuss Lab 03 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.

0.2 Learning Objectives

1. Continue to practice and refine visualizing data in an informative way, with attention to the center, shape, and spread.
2. Use various approaches to obtain numeric summaries to assess distributions
3. Use visualization to describe the relationship between two variables.
4. Use visualizations to assess the adequacy of a Normal distribution model.

0.3 Use the Lab 3 Template

To start, create a directory on your computer for `lab03` following the suggestions we provided in the Lab 02 instructions.

- Into that `lab03` directory, download the Quarto Template for Lab 03 from our 431-data site. The template is called `431-lab03-template-2023.qmd`. This should be useful to you in the same way as the Template for Lab 02 was. You should then rename the file to `yourname-lab03.qmd`.

As in Lab 02, after you've downloaded the relevant data file and template file to your `lab03` directory, open RStudio, and use the **File … New Project … Existing Directory** menu to create an R Project in that directory in which you will do Lab 03.

Build your response in the Quarto file, and 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.

Be sure to remove all of the instructions included in the original template before submitting your work, and also be sure to review the HTML result to ensure that it looks clean and clear, that the labels on your plots and other output are easy to read, and that it doesn't retain any unnecessary warning messages or other material that distracts from your work.

0.4 Submitting the Lab

Submit **both** your revised Quarto file **and** the HTML output file to the Lab 03 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.

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

0.5 Our Labs 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 The Data for Lab 03

There are two sets of data used in this lab.

0.6.1 The `penguins` data

In Questions 1-3, we'll be using the `penguins` data (note: use the `penguins` tibble, and not the `penguins_raw` tibble for this Lab) contained in the `palmerpenguins` package in R. The complete citation is …

Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. R package version 0.1.0. https://allisonhorst.github.io/palmerpenguins/. doi: 10.5281/zenodo.3960218.

There, you'll find a nice cartoon of the three species of penguin contained in the data and a detailed description of the bill measurements that are worth your time.

0.6.2 County Health Rankings 2022 information in `lab03-data.csv`

In Questions 4-6, we'll be using additional data from the 2022 County Health Rankings Data that we saw back in Lab 02. We have compiled a new data set called `lab03-data.csv` which you will find in our 431-data repository. This data file contains seven variables of interest for *all* 3,082 U.S. counties ranked by County Health Rankings in their 2022 report.

1 Question 1 (8 points)

We are interested in comparing the bill depth (in mm), `bill_depth_mm`, of the penguins included in these data, on the basis of the `island` where they were measured. Create a visualization of these variables that will help us evaluate the center, shape, and spread of the bill depths within each of the islands, using appropriate labels for all axes, and a useful title for the visualization. Be sure to account for any missing data in an appropriate and clear way.

2 Question 2 (8 points)

Use R code to identify the mean, standard deviation, median and interquartile range of the penguin's bill depth (in grams) associated with each of the three islands. Again, be sure to account for any missing data in an appropriate and clear way. Then, in 2-3 sentences, briefly compare the center and spread of the bill depths across the islands.

3 Question 3 (8 points)

Now build a visualization of the bill depths for all of the penguins with available measurements, regardless of island. In studying this variable, does it seem that bill depth follows a Normal distribution? Would it be more appropriate to examine the mean or median in this setting, and why?

4 Question 4 (8 points)

Take a random sample of 750 counties without replacement from the County Health Rankings Data provided in `lab03-data.csv`. As part of this work, use the command:

```
set.seed(1)
```

so that each of us selects the same sample of data. Name this random sample `lab3sample` in R, and be sure that your sample includes: `state`, `county`, `adult_obesity`, and `food_insecurity`.

Once this is done, demonstrate that Cuyahoga County in Ohio is in your sample and that the median of `adult_obesity` across all 750 sampled counties is 0.361.

4.1 Notes

- The `adult_obesity` variable comes from the CDC Diabetes Interactive Atlas for 2019, and estimates the proportion of adult county residents (ages 18 and over) who have a body mass index (BMI) greater than or equal to 30 kg/m^2, and this is age-adjusted.
- The `food_insecurity` variable comes from the Map the Meal Gap project for 2019, and estimates the proportion of the county's residents who lack adequate access to food.
- The value for Cuyahoga County's `adult_obesity` should be 0.367, and the value for `food_insecurity` should be 0.139, if you're interested.
- The `adult_obesity` and `food_insecurity` variables are reported in the data as proportions. Please leave them as proportions for this lab, rather than converting to percentages.

5 Question 5 (10 points)

Now we'd like to examine the relationship between `food_insecurity` (our predictor) and `adult_obesity` (our outcome) across our 750 sampled counties. Build a useful, well-labeled visualization which, in addition to showing the association of these variables, **also** identifies where Cuyahoga County's results fall on the plot.

Does a linear model seem appropriate to describe the association of these variables? Why or why not? Hint: you will likely want to include a smooth of some sort in your visualization.

6 Question 6 (8 points)

Reflecting on Chapter 4 of Spiegelhalter's *The Art of Statistics*, please write an essay of no more than 150 words which discusses the relationship between `adult_obesity` and `food_insecurity` we observe in our sample of counties. Specifically, discuss whether or not we can conclude that lack of access to adequate food causes obesity.

In your response, be sure to also discuss whether the method that we used to identify these data (our 750 county random sample) strengthens or weakens your conclusion(s).

7 Session Information

At the end of your Quarto file, please provide the **session information**.

```
sessioninfo::session_info()
```

As in previous labs, I've hidden the results in these instructions, but we'll want to see your full results in evaluating your work. Again, providing the session information helps with reproducibility. It lets us see what packages you have loaded on your machine, and some other information about your R session that can be helpful in understanding problems as they occur.

8 Hint: Some Places to Look for Help

In addition to the slides discussed in class, we recommend the following resources as being helpful for completing this Lab's coding tasks:

- The early chapters in the Course Notes
- The chapter on Exploratory Data Analysis within R for Data Science, and
- The R Graphics Cookbook.

9 What Happens After Submission?

We will summarize some of the more interesting responses to Question 7 after the Lab has been graded.

- This Lab will be graded on a scale from 0-50.
- 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. Spelling, grammar, syntax and the rest all matter for grading purposes in this and all other assignments this term.

A detailed answer sketch for 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.