431 Lab 03

Deadline: See Course Calendar | Last Edited 2022-09-28 11:44:16

Table of contents

Deadline

Lab 03 has 10 questions, all of which you need to complete by the deadline specified on the Course Calendar.

- 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.)

Getting Help

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.

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.

An Important Note

Your response to **every** question, whether we explicitly ask for it or not, should include a complete English sentence responding to the question. Code alone is not a sufficient response, even if the code is correct. Some responses might not need any code, but every response needs at least one complete sentence.

Getting Started with Lab 03

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

- Into that `lab03` directory, you will download the R Markdown Template for Lab 03, which is called `YOURNAME-Lab03.Rmd`. This should be useful to you in the same way as the Template for Lab 02 was. After you download the template file to your directory, you will want to rename it to substitute in your actual name in the file name, rather than `YOURNAME`.

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

The R Markdown Template for Lab 03

We have provided an R Markdown document template for this assignment called `YOURNAME-Lab03.Rmd` 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 `2021-431` directory.) Then copy the template into your directory for Lab 03, specifically, that you created earlier. That's probably the easiest approach.
- The template provides some coding hints, which we hope you'll make use of.

You should build your response to all ten questions as an R Markdown file using the `YOURNAME-lab03.Rmd` template provided. Use the Knit 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.

The Data for Lab 03

There are two sets of data used in this lab.

The `penguins` data

In Questions 1-5, 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.

Additional information on the data are provided by Allison Horst at the github site linked above. In particular, 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.

County Health Rankings data

In Questions 6-9, we'll be using additional data from the 2021 County Health Rankings Data that we saw back in Lab 02. We have compiled a larger data set called `lab03_counties.csv` which you will find in our Data and Code repository. This data file contains all of the counties and a dozen of the available variables in the full County Health Rankings data.

Part A: Palmer Penguins (Questions 1-5)

Question 1 (8 points)

We are interested in the body mass (in grams), `body_mass_g`, of the penguins included in these data. Create a visualization of this variable that will help us evaluate its center, shape, and spread, using appropriate labels for all axes, and a useful title for the visualization.

Question 2 (8 points)

Identify the mean, standard deviation, median and interquartile range of the penguin's body mass (in grams). Do not simply report R output but (in addition to presenting your code and the resulting output) state, in complete English sentences, your results.

Question 3 (8 points)

Given your visualization in Question 1 and the numeric summary in Question 2, please discuss the center, shape, and spread of the body mass of the penguins. Does it seem that body mass follows a Normal distribution? Would it be more appropriate to examine the mean or median in this setting, and why?

Question 4 (8 points)

The dataset also contains information on the penguin's species (`species`). Now, build a visualization and a numeric summary to examine body mass across the three species types.

Question 5 (8 points)

Given your findings in Question 4, what can we conclude about the body mass across species?

Part B: County Health Rankings (Questions 6-9)

Question 6 (10 points)

Take a random sample of 750 counties from the County Health Rankings Data provided in `lab03_counties.csv`. As part of this work, use the command:

```
set.seed(20212022)
```

so that each of us selects the same sample of data. Name this random sample `chr_sample` in R. Select only the following variables: `state`, `county_name`, `adult_obesity`, and `food_insecurity`.

Once this is done, demonstrate that Cuyahoga County in Ohio is in your sample and that the mean of `adult_obesity` across all 750 sampled counties is 0.3345.

Note 1

The value for Cuyahoga County's `adult_obesity` should in fact be 0.318, and the value for `food_insecurity` should be 0.159, if you're interested.

Note 2

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.

Question 7 (10 points)

We are interested in looking at `adult_obesity` as an outcome. Build a visualization and describe the center, spread, and shape of the distribution of this variable. Does a Normal model for this distribution seem appropriate? Why or why not?

Question 8 (10 points)

Now we'd like to examine the relationship between `food_insecurity` (our predictor) and `adult_obesity` (our outcome). Build a useful, well-labeled visualization and then use it to motivate your description of the relationship between these two variables in complete sentences. 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.

Question 9 (10 points)

Create a new figure (perhaps building on the one you built in Question 8) so that this new Figure identifies where Cuyahoga County falls when examining the relationship between `adult_obesity` and `food_insecurity`. Briefly interpret Cuyahoga County's position on each variable relative to the others in the random sample.

Part C: Spiegelhalter Reaction

Question 10 (20 points)

Reflecting on Chapter 4 of *The Art of Statistics*, please write an essay of no more than 100 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 observation random sample) strengthens or weakens your conclusion(s).

Include the session information

At the end of your R Markdown file, please include a new code chunk to provide the **session information**. This time, we'll use an alternative approach to get this (as compared to what we did in Lab 02):

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

```
- Session info -------------------------------------------------------------
 setting  value
 version  R version 4.2.1 (2022-06-23 ucrt)
 os       Windows 10 x64 (build 22000)
 system   x86_64, mingw32
 ui       RTerm
 language (EN)
 collate  English_United States.utf8
 ctype    English_United States.utf8
 tz       America/New_York
 date     2022-09-28
 pandoc   2.18 @ C:/Program Files/RStudio/bin/quarto/bin/tools/ (via rmarkdown)

- Packages -----------------------------------------------------------------
 package     * version date (UTC) lib source
 cli           3.3.0   2022-04-25 [1] CRAN (R 4.2.1)
 digest        0.6.29  2021-12-01 [1] CRAN (R 4.2.0)
 evaluate      0.16    2022-08-09 [1] CRAN (R 4.2.1)
 fastmap       1.1.0   2021-01-25 [1] CRAN (R 4.2.0)
 htmltools     0.5.3   2022-07-18 [1] CRAN (R 4.2.1)
 jsonlite      1.8.0   2022-02-22 [1] CRAN (R 4.2.0)
 knitr         1.40    2022-08-24 [1] CRAN (R 4.2.1)
 magrittr      2.0.3   2022-03-30 [1] CRAN (R 4.2.0)
 rlang         1.0.5   2022-08-31 [1] CRAN (R 4.2.1)
 rmarkdown     2.16    2022-08-24 [1] CRAN (R 4.2.1)
 rstudioapi    0.14    2022-08-22 [1] CRAN (R 4.2.1)
 sessioninfo   1.2.2   2021-12-06 [1] CRAN (R 4.2.0)
 stringi       1.7.8   2022-07-11 [1] CRAN (R 4.2.1)
 stringr       1.4.1   2022-08-20 [1] CRAN (R 4.2.1)
 xfun          0.33    2022-09-12 [1] CRAN (R 4.2.1)
 yaml          2.3.5   2022-02-21 [1] CRAN (R 4.2.0)

 [1] C:/Users/thoma/AppData/Local/R/win-library/4.2
 [2] C:/Program Files/R/R-4.2.1/library
```

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 any problems you run into. The `sessioninfo::session_info()` command shown above is part of the template for this Lab, and is a good alternative in many cases to the approach we took in Lab 02. One or the other of these session information runs should appear at the end of all of your lab and project work for this course.

Submitting the Lab

As mentioned, you should build your entire response as an R Markdown file using the `YOURNAME-lab03.Rmd` template provided. Then use the Knit button in RStudio to create the resulting HTML document. 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.

Submit **both** your revised R Markdown 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 R Markdown 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.

Grading

We will summarize some of the more interesting responses to Question 10 after the Lab has been graded.

- This Lab will be graded on a scale from 0-100.
- 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 on the day after the submission deadline, and a grading rubric will be provided when the grades are made available, approximately one week after the submission deadline.

Late Penalties for Lab Work

- Labs that are turned in 1-12 hours after the deadline will lose 10% of available points.
- Labs turned in more than 12 but less than 72 hours after the deadline will lose 25% of available points.
- No extensions to Lab deadlines will be permitted this semester. Labs turned in more than 72 hours after the deadline will receive no credit.
- Your lowest lab score (out of Labs 1-7) will be dropped before we calculate your lab grade.