

BIOL/STATS 2244

Lab 2: *Data*

Objectives

The Lab 2 Assignment provides an opportunity to experience aspects of the *Data* stage of the PPDAC scientific inquiry framework. Specifically, this assignment focuses on aspects of

- i. collecting data according to a previously described *Plan*,
- ii. exploring and cleaning data in advance of further analysis,
- iii. applying your understanding of vocabulary and concepts associated with summarizing and exploring data, and,
- iv. summarizing data in appropriate scientific formats.

To achieve these objectives, students will need to draw on content from virtual lab meeting, and from course material covered in the first three lab topics (Intro to Lab, Lab 1 and Lab 2) and lecture topics up to and including topic 5: Summarizing and Exploring Data.

Background for Assignment

In the first Lab Assignment, you were introduced to a little background information that led/explained the logic behind the following **Research Objective** for 2244 labs:

Characterize the effect of practicing yoga on the balance/stability of visually impaired people.

In Lab Assignment 1, you were tasked with characterizing the *Problem* of interest, and creating a *Plan* for a sampling and study design to collect data to address a specific research question related to this Objective. Rather than have each of you implement your described *Plan*, we will use published data from a research study with a similar objective:

Jeter PE, Haaz Moonaz S, Bittner AK, Dagnelie G (2015) Ashtanga-Based Yoga Therapy Increases the Sensory Contribution to Postural Stability in Visually-Impaired Persons at Risk for Falls as Measured by the Wii Balance Board: A Pilot Randomized Controlled Trial. PLoS ONE 10(6):e0129646.
doi:10.1371/journal.pone.0129646.

An annotated pdf of the paper is provided with the lab material for this assignment. You should read up to page 10 (the Abstract, Introduction and Methods (except sections on Secondary Outcomes, Instrumentation and Statistical Analysis)). It is not necessary for you to understand every technical detail of the paper, but you should understand

- the sampling design,
- study design,
- methods of randomization, replication, and control
- how the explanatory and response variables were measured (it is okay to be somewhat confused about the calculations of the Stability Indices, but you should understand what they are comparing.)

Data for this lab are presented in two files, *balance_data.csv* contains the raw primary outcomes data and *demographic.csv* contains some demographic data about the participants. The data description file

includes details about symbols and units used in the two data files. **You should familiarize yourself with the information in that file and/or keep it handy as you work with the data.** It will likely be useful as you answer the questions on this Assignment, as well as some questions on future Assignments in the course.

Instructions for Assignment

To help you approach this Assignment in a logical/organized fashion, you are encouraged to follow these steps (in the order presented). Ask questions about any step if you don't understand the information!

1. Read the Abstract, Introduction and the sections of the Methods so you understand the sampling and study design, and the dataset with which we are working.
2. Read through the remainder of this Assignment file, including the “Reminders/Tips for Success”, the Assignment Questions, and the comments related to Marking Rubrics so that you know what you are being asked to do. Be sure to ask questions (e.g. during your section's Zoom meeting, by OWL message to your lab TA, or on the Forum) if you need clarification!
3. Open the “Answer template” file. Use this file to type/enter your answers to the Assignment Questions; it is set up with the proper headings for this Assignment; you just need to input your answers (use whatever space you need to do so). Refer back to any feedback you received from the “Practice Lab Assignment” on the formatting of your assignments.
4. **Import** the data file provided, “balance_data.csv” and “demographic_data.csv” into R or R Studio as a dataframes called *balance_data* and *demographic_data*, respectively. If you refer to these names in your R code at any point in your assignment, we will assume that they are the files that you were provided. This way, if you choose to save the data file under a different name that we don't recognize (i.e. other than *balance_data* or *demographic_data*), you won't be penalized if you always refer to it in code as *balance_data* or *demographic_data*. It might be useful to check the structure of the data using a function like *str()* to ensure it imported properly and/or is consistent with the quality check information provided in the Data Description file.
5. Explore the data a little bit to get familiar with it. To do this, consider the data types of the variables, and use appropriate numerical and/or graphical summaries to explore the distribution of individual variables, and/or the relationship between variables. You may also find it helpful to simply *View()* the datafile (e.g. scroll through the rows to get a visual sense of the data).
6. Answer the Assignment Questions (below).

Reminders/Tips for Success

1. **Make interpreting your R code easy and ensure that it is functional.** The best approach is to generate an R script file for each numbered Question that requires use of R in any way. Put everything in that R script file that takes you from attaching the dataset to completing the question. Annotate your more complicated lines of code with #comments. Then, simply copy/paste the contents of that script file into your answer (and include the Output if applicable) when asked for R code. We should then be able to copy/paste what you've included, and the code should run without problems!
2. **We only know names/variables that are part of the original data files.** So, when we are looking through/grading your R code, if we see a name for a variable or datafile that is NOT part of the original data files (e.g. you've renamed a variable or created a new one), you must provide information on what that R object represents. As discussed during the second in-lab session, this is best achieved by simply providing the line(s) of R code (plus #comments!) used to create that new object. Failing to provide such new R object definitions is one way to not be awarded full marks for your R code.

3. Your answers should be written *specifically* for the research study/context (in terms of variables, sample, units, measurements, etc.) with which we are dealing. For example, it's insufficient to talk about the "response variable"; we should be talking about the actual name of that variable. Being specific means using the *context* of the research; a sentence like, "the distribution of *student heights* in my sample of 2244 students needs to be symmetric" is explicit about *what* needs to be symmetric AND uses the *vocabulary* of the study.
4. **Demonstrate your understanding of course content through application, not definition.** A question like Question 3b, which asks you to compare and discuss something with reference to particular course concepts (e.g. sampling variability) requires an *application* of those course concepts to the current scenario/situation. That means that simply providing the definition of those concepts is not going to result in points awarded for the Question. Your answers should demonstrate you understand that concept and why it applies (or doesn't!) or is useful in the particular context.

Assignment Questions

Question 1.

The first step in working with a dataset is "cleaning" the data. This involves inspecting the dataset for potential errors and ensuring that our software has imported the data completely, and in a manner consistent with the nature of the data.

- a) Use the `str()` function in R (as introduced in Online R Module 1) to quickly inspect the dataset. Using your knowledge of types of variables (e.g. quantitative, ordinal, ratio, etc.) and the labelling of data types in R (e.g. character, factor, numeric [num], integer [int]) to check that all variables have been imported and identified correctly by R (hint: go review the variable descriptions from the Data Description file so you know what kind of variable each should be, then compare this to what R has identified). Write R code that will re-class any variables that are mis-classed by R.

What do we mean about "incorrect data types" in R?

When we import a .csv file into R, R automatically looks at the data and labels the type of data for each variable as either character, factor, numeric, or integer. However, R sometimes gets the labelling wrong.

e.g. Suppose you have a variable called "bmi" (measured in kg/m²) which you know is a quantitative, ratio, and continuous variable. As a consequence, it should be identified as a "numeric" data type by R. However, R may have incorrectly identified this variable during import as "factor". Consequently, 'bmi' would be a variable that was incorrectly identified.

- b) Use R to create graphs to help you inspect the dataset for outliers and missing values. You will discover that at least one variable has an outlier (or more than one). Identify ONE (1) **variable** in which there is an outlier in the dataset.
- c) Briefly identify the most plausible **reason** for *why the outlier exists* in the variable you identified, and explain why you suspect that is the reason.
- d) Briefly describe the most appropriate **method** of dealing with the outlier you identified, and justify your decision.
- e) Report the **R code(s) plus output** that you used that explicitly helped you find/identify the outlier for the variable that you have discussed. If you graphed more than one variable separately, you

only need to include the code that produced the graph which showed the outlier(s) you identified in 1a.

Question 2.

- a) In Online Lab Module on data manipulation, you were introduced to the distinction between datafiles organized in the 'wide' vs. 'long' formats. In what **format** is our *balance_data.csv* file? Briefly describe/explain the specific **elements** of the datafile's structure that led you to your answer.
- b) Convert the data to the other format. Include the R code you used to convert and show the `str()` of the resulting data frame.

Question 3.

In most research involving humans, researchers describe the 'demographics' of their sample (i.e. age, gender, and other characteristics, etc.). This description often occurs because the researchers want to highlight the degree to which their sample represents (or possibly deviates from) the population of interest, or, to identify the similarity/differences between their sample and samples used in related research. According to the US Census Bureau, the population of the United States is 50.7% female and has a median age of 38.2 years. The National Research Council (US) Working Group on Mobility Aids for the Visually Impaired and Blind reports that the median age of legally blind Americans is greater than 64 years. A systematic review study by Abou_Gareeb and colleagues (Abou-Gareeb I, Lewallen S, Bassett K, Courtright P. Gender and blindness: A meta-analysis of population-based prevalence surveys. *Ophthalmic Epidemiol.* 2001;8:39–56.) found that after controlling for age, women are approximately 40% more likely to be legally blind than men.

- a) Use R to calculate the median age and proportion female of all study participants. *Report and describe* the values you calculated, and the R code(s) you used to compute the summaries.
- b) Briefly **compare and discuss** whether you think our sample has characteristics consistent with the population of interest based on the information as described above. Be sure to apply your understanding of sampling strategies, sampling variability, and sampling error in your discussion.

Question 4.

When conducting studies that involve comparing two or more groups, researchers commonly compare the groups at a 'baseline' stage of the study. In this question, you will undertake this type of comparison *visually* with a graph.

- a. Create a SINGLE (i.e. one pane) **graph** in R, that can be used to compare groups in our data to address the following question: *Directly before beginning the Yoga treatment, do the individuals in the randomly created experimental groups differ in with respect to their MTV of COP in the Firm_EO condition?* Make sure that all axes on your graph are correctly labelled including units.

- b. Provide all the **R code** you used to generate your graph (with any # comments necessary to help us interpret the code; be sure to include any preliminary data manipulation, variable creation, etc. that you did!).
- c. Graphs that appear in scientific papers are accompanied by legends to fully describe all variables, axes and symbols included in the graph. For example:

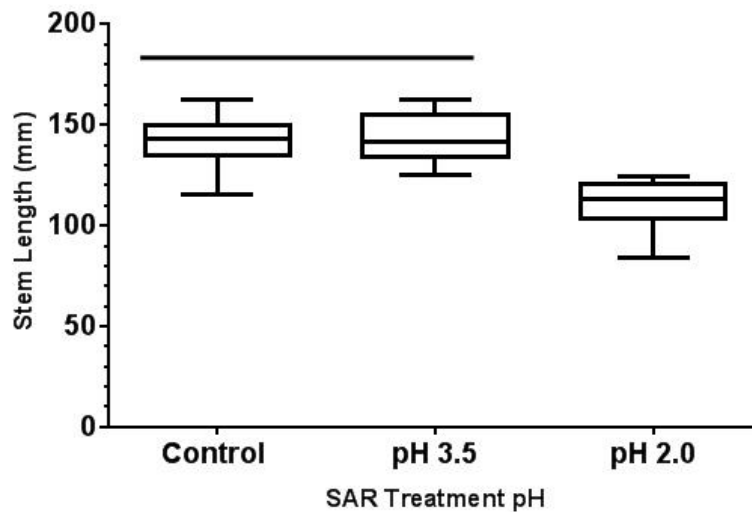


Figure 1. Distributions of stem length of 30-day old sunflower seedlings (*Helianthus annuus*) following 20 daily treatments with water (control, $n=20$), or simulated acid rain (SAR) solutions ($n=15$ for pH 3.5, $n=22$ for pH 2.0) in a greenhouse setting. SAR solutions prepared from tapwater with 2 M sulfuric acid: 1 M nitric acid. Line over bars indicates no significant differences. Kruskal-Wallis Test ($p<0.0001$) and Dunn's post hoc test.

Write a suitable figure legend for your figure.

Marking Rubric

Unlike Assignment 1 (where many of the questions were open-ended), many of the questions in Assignment 2 have correct vs. incorrect answers and/or approaches. Consequently, the marking scheme for evaluating your answers to certain questions may often have a single ‘right’ answer/approach for which we are looking. However, how we use R to explore, summarize, and analyse our data can, to some degree, vary in technique. That is, there may be more than one way to ask R to complete a particular ‘task’.

So, what does this mean for a student trying to understand expectations when completing this Assignment? In addition to the “Reminders/Tips for Success” provided on page 2 of this file, consider the following general criteria for different types of questions/markings; these criteria will likely play a heavy role in evaluating the answers submitted for the Assignment:

Criteria for R code and output

- ✓ Selection of data, variables, and subsets is relevant to the question or task;
- ✓ Choice of R functions is relevant and appropriate (demonstrating an understanding of the analysis being conducted) for the question, task, and/or type of data being summarized/analyzed;
- ✓ Reported R code for any numbered question is complete and would function (i.e. reproduce the output included/described in the answer) if it were copied/pasted into R, and run on the *labdata.csv* data (assuming we had first successfully imported that data and saved it as a dataframe called *labdata*).
- ✓ Reported R code uses brief *#comments* to help interpret the purpose of more complex commands/functions

Criteria for ‘other’ questions (i.e. identifying, describing, explaining, discussing, etc.)

- ✓ *Knowledge*: use and application of relevant statistical vocabulary/concepts demonstrates an accurate understanding of those concepts; that is, the vocabulary/concepts are *used/applied* in a manner that is consistent with the definition/understanding. The use moves beyond simply defining the concepts, but actually applies them to the situation. This criterion also connects to whether an answer is consistent with any expectations/guidelines communicated in course content (e.g. lectures).
- ✓ *Connections/Justification*: Answer demonstrates (through explanation and/or description, where appropriate) the relevance or relationship of choices made/vocabulary used to the question(s) or situation. That is, it’s clear WHY you have made the choices you did and these choices make logical sense.
- ✓ *Completeness*: Answer provides sufficient detail (whether in written answers or visual content) that another, knowledgeable individual can understand and/or recognize the application of the concepts, without ambiguity or doubt. This also refers to whether the answer has addressed all aspects of the question.
- ✓ *Communication*: Answer uses clear and concise language, and thoughtful organization of ideas to facilitate readability and understanding. That is, we do not have to re-read your answer multiple times to try to understand what you are saying.

Other comments

Remember that these Assignments are meant to assess your understanding of course content. So, many of the questions relate directly to content covered/discussed in lecture and lab. Reviewing these ‘resources’ while working on the Assignment may be quite beneficial/informative!