Biostat 561: Homework 2

Instructor: Amy Willis, Biostatistics, UW

10 April, 2019

Instructions and preliminaries

See the specification for Homework 1 regarding academic integrity, expectations about responses (i.e., RMarkdown), and where to find answers to questions on this homework that were not discussed in class.

Please complete Question 0 before Monday 15 April at 2:30pm, and attend office hours if you have trouble. The remaining questions are due Wednesday 17 April at 2:30pm sharp.

You will need to install and load the `tidyverse` for this homework.

Question 0: Easy version control with git

1. Go to your materials folder that you created last week and type `git pull`. This should download Lecture 2 and Homework 2 to your local copy. You should repeat this command every week to give you the latest course information and resources.
2. Complete the following questions in your responses folder, then use the workflow from last week to upload Homework 2 to github classroom as a submission to `hw2`. Solutions must be completed using RMarkdown. Submit both your `Rmd` and your `pdf` files. There is no need to show *all* code and *all* output – use your judgement and be succinct.

This is the last homework where these instructions will be formally repeated, but all future homeworks should be submitted in this way.

Link to Homework 2 submissiohn: https://classroom.github.com/a/fbGsK0bd

Question 1: Getting started with the tidyverse

The dataset `faithful` gives waiting time between eruptions and the duration of the eruption for the Old Faithful geiser in Yellowstone.

a) Use `apply` (without any piping) to calculate the mean waiting time and duration of the eruptions.

b) Use the pipe operator in conjunction with `apply` to calculate the mean waiting time and duration.

c) Using the pipe operator and `summarise` (or `summarize`, whatever...), calculate the mean eruption duration.

d) Calculate the mean waiting time and duration. You need only the pipe operator and one other function from `dplyr`. Find this function at http://dplyr.tidyverse.org/reference/index.html.

e) `faithful` is a data frame. How many lines are output by default when you show a data frame in console? Convert `faithful` to a tibble. How many lines are output by default when you show a tibble? What other information is given?

f) `filter()` and `subset()` perform similar functions. List 3 differences between them. Which will you intend to use?

g) Using the dataset `airquality`, calculate the mean and standard deviation of Ozone, stratifying observations with temperatures of 85 degrees or more (versus less than 85 degrees).

h) What's the difference between `mutate` and `transmute`? Given example of when you would use one versus the other.

i) The temperature variable in `airquality` is currently in degrees Fahrenheit. Modify it in place to the same measurement in Celsius, rounded to the nearest degree.

Question 2: Function composition

Install and load the package `magrittr`. Make extensive use of the documentation `http://magrittr.tidyverse.org/` in answering these questions.

a) What does this do: `f <- . %>% cos %>% sin`?

b) What is the compound assignment pipe operator, `%<>%`?

Question 3: Tidy data: a reflective exercise

Read:

- Karl W. Broman & Kara H. Woo (2018) Data Organization in Spreadsheets, The American Statistician, 72:1, 2-10, DOI: 10.1080/00031305.2017.1375989
  - https://www.tandfonline.com/doi/pdf/10.1080/00031305.2017.1375989
- Wickham, Hadley. "Tidy data." Journal of Statistical Software 59.10 (2014): 1-23.
  - https://www.jstatsoft.org/article/view/v059i10

Describe a situation where you have either created or had to analyse messy data. What was the format of the data? Did you make it tidy before analysing it? Or did you just plough along and work with the messy form?

If you've never created or had to analyse messy data, either (a) Think a little harder about data that you have analysed, because most data is not tidy, or (b) Find a nontrivial example of messy data in a subject area that interests you, and process it into a tidy form.

Question 4: Joining multiple datasets

The authors of Brooks et al (2015) (DOI: 10.1186/s12866-015-0351-6) did a fantastic job of making their data available (you don't need to read this paper). Here are 3 spreadsheets containing relevant information from their experiments:

- `brooks_sample_data.csv`: Every row is an experiment, with columns indicating an identifier for the sample, the `Type` of experiment, and other information associated with sequencing.
- `brooks_true.csv`: Every row is an experiment, and the columns `Lcrispatus`, `Liners`, `Gvaginalis`, `Avaginae`, `Pbivia`, `Samnii`, and `GroupBStrep` (these are species) indicate the actual relative abundance of the species in that experiment.
- `brooks_observed.csv`: Every row is an experiment, and all columns except the first denote species. The entries indicate the number of times the species was observed in the experiment.

This data is included in the lecture2 directory and should have been downloaded with your homework.

(a) Create a single tibble called `brooks` containing all the information from `brooks_sample_data.csv`, `brooks_true.csv` and `brooks_observed.csv`. You will need to look at the data to decide how to best

join them. There should be no redundancy (i.e. information duplicated in the columns), and there should be exactly one experiment per row.

- *See the lecture slides for the reference page for joining tibbles*

(b) How many plates were there in the experiment? How many barcodes were included on each plate?

(c) What is the mean and standard deviation of the total number of species observed in each experiment? What is the mean and standard deviation of the total number of species observed in each experiment of each `Type`?

- *There are multiple ways to do this, and you may choose any tidy way to do it. Personally, I would do this using* `gather`