MAR536: Biological Statistics II

R Laboratory Exercise 6

February 22, 2023

Open a new R markdown file. Save it. (name it lastname_lab6.Rmd or something similar)
At the top of the script, add comments with your name and lab 6. Work in pairs or individually. Submit your Rmd and any other files via myCourses before lab next week.

Exercise 1

Write code that uses one of the map functions to:

a. Compute the mean of every numeric column in `palmerpenguins::penguins`.

b. Determine the type of each column in `nycflights13::flights`.

c. Compute the number of unique values in each column of `palmerpenguins::penguins`.

Exercise 2

1. Create a data frame of samples from the `palmerpenguins::penguins` dataset, that contains 3 Adelies, 6 Gentoos, and 4 Chinstraps.
   (your new data frame will have 13 rows, with 3, 6, and 4 of the three species) (hint: use a nested or split dataframe, `map2()`, and `slice_sample()`)
2. We have data from several years of crab surveys. The data for each year is contained in separate ".csv" files.

We would like to read these data into R, and combine them into a single data frame so we can inspect and plot them.

a. Write code to read these data into R, and combine them into a single dataframe.

b-d. Then produce 3 plots (of your choice) summarizing the full dataset. Include "b", "c", and "d" in the title of your plots.

- you can use the following to get an object containing a list of files in a folder

```
data_path = "../data/crabs"  # directory where the files are located
files <- dir(path = data_path, pattern = "*.csv",
             full.names = TRUE) # names of files ending in ".csv"
files
##  [1] "../data/crabs/CRABS_2001.csv" "../data/crabs/CRABS_2002.csv"
##  [3] "../data/crabs/CRABS_2003.csv" "../data/crabs/CRABS_2004.csv"
##  [5] "../data/crabs/CRABS_2005.csv" "../data/crabs/CRABS_2006.csv"
##  [7] "../data/crabs/CRABS_2007.csv" "../data/crabs/CRABS_2008.csv"
##  [9] "../data/crabs/CRABS_2009.csv" "../data/crabs/CRABS_2010.csv"
## [11] "../data/crabs/CRABS_2011.csv" "../data/crabs/CRABS_2012.csv"
## [13] "../data/crabs/CRABS_2013.csv" "../data/crabs/CRABS_2014.csv"
## [15] "../data/crabs/CRABS_2015.csv" "../data/crabs/CRABS_2016.csv"
## [17] "../data/crabs/CRABS_2017.csv" "../data/crabs/CRABS_2018.csv"
## [19] "../data/crabs/CRABS_2019.csv"
```

- look at the help for 'dir' for additional functionality

Lab Exercise 3/3 - Eukaryote genes

`data/eukaryotes.tsv` contains a NCBI Eukaryotic genome dataset, with basic information about the genomic content of all eukaryotes that were uploaded to the NCBI Genome database.
It contains accession numbers, information about the quality of the genome and stats such a average genome size and GC-content.
Use `glimpse()` and other data exploration to get familiar with the data. Then use `map_*` functions to answer the following:

1. How many different organisms are there in the dataset?
2. How many different institutes (centers) submitted a genome?
3. The data seem to be grouped in groups. How many groups are there?
4. How many sub groups are there?
5. How many different organisms are there per group?
6. How many different institutes (centers) submitted a genome per group?
7. How many sub groups are there per group?

We might hypothesisze that "The bigger the size of a genome, the higher the number of proteins".

8. Fit a linear model of `log10_proteins ~ log10_size_mb` for each group.
9. Extract the `R^2` for each model and print these for each group.
10. Assess the validity of your modeling approach.
11. Obtain and plot predictions for each group for genome sizes 0.5, 123, and 500 MB.
12. How do you interpret the results in terms of the original hypothesis?

BONUS use residual bootstrapping to obtain distributions for the predictions made in part `11`.