BMR 617

Exploring Quantitative Data

Exploring quantitative data

- Remember quantitative data are *numerical* data
  - Typically represent a measure of some quantity
- Given a set of quantitative data, questions we might ask are:
  - What is a *typical value* of the data?
  - How spread out are the data around the typical value?
- We can do this both numerically and visually
  - Both can give us an initial “feel” for the data

Aside: Data wrangling

- It's common to find that a large portion of the time you spend in data analysis is in getting the data into a form where the software can perform the analysis
- Informally known as *"data wrangling"*
- Can be a frustrating, time-consuming, boring part of data analysis…
- Recently, Hadley Wickham, a statistician from New Zealand, developed a collection of R *packages* for data wrangling, collectively known as the *tidyverse*
  - Make your data *tidy*

Introducing the tidyverse

- Open R studio
- Under the “Tools” menu, choose “Install Packages”
- In the “Packages” field, type “Tidyverse”
- Leave the other fields as the default values and press “Install”
- This will install the *tidyverse* package
- You only need to do this step once; if you exit R and restart, the package will still be installed

Loading the library

- We'll put the code for today's work in a script
- This means we can easily save our work, and redo the same analysis later
- In RStudio, select "File", "New File", and "R Script"
- This will create a new tab in the top left, labelled "Untitled"
  - You can save it with a meaningful name in a folder of your choosing at any time
- In the new script, enter the code `library(tidyverse)`
- This will load the library into the current session, so its functionality will be available
  - We need to do this once per session in which we need the library
- Click next to this command in the script, and press the "Run" button

Loading some sample data

- We'll load a data set that we'll use frequently throughout this course
- This data set comprises metabolic data from a mouse experiment in Dr. Kim's lab
- Two strains of mouse, C57BL/6 ("B6") and Tallyho ("TH") were fed three different diets (standard Chow, the control; a low-fat, high-calorie diet "LF", and a high-fat diet "HF")
- Various metabolic measurements were taken from each mouse after 16 weeks on the diet
- Load the data with the tidyverse `read_csv` function

```
met <- read_csv("https://denvirlab.marshall.edu/BMR617-2021/data/TH-B6-metabolic.csv")
```

Tidyverse tables

- Run the command from the previous slide
- This will create a new object in your environment called “met”
- In RStudio, this appears in the “Environment” tab under “Data”
- Click on “met” in the environment tab to view the data
- This appears as a table in a tab next to your script
  - In tidyverse, this data type is called a “tibble”
- In “Environment”, expand met by clicking on the small arrow next to the name
  - Note how it describes the types of each column

Some simple data wrangling

- Look at the “MouseID” column in the table
  - If you like, you can access this directly in your console by typing
    ```
    met[ , “MouseID”]
    ```
- Each ID contains three pieces of information
  - The strain, diet, and a numeric ID
- To make the data “tidy”, we should separate these into three columns
- Type the following into the console:
  ```
  separate(met, "MouseID", sep="-", into=c("Strain", "Diet", "ID"))
  ```
- When you’re convinced it gives you what you want, save it to the script
  - You can do this by selecting it in the “history” and pressing the “To Source” button
- Edit the command in the source so that it stores the new version back in the met object:
  ```
  met <- separate(met, "MouseID", sep="-", into=c("Strain", "Diet", "ID"))
  ```
- Run that command, and check met in the data viewer

Summary statistics

- For quantitative data, we commonly consider two forms of *average*:
- The *mean* is the sum of all values, divided by the number of values in the data set
  - It gives a sense of the “typical” value
  - If you replaced all values with the mean, the total would remain the same
- The *median* is the middle value
  - Put all the values in order, and choose the one in the middle

Mean and median in R

- Let's focus on just one group of mice. We can do this by filtering the data: (another tidyverse function):
  `th_chow <- filter(met, Strain==“TH” & Diet==“Chow”)`
- We can “pull” the cholesterol values from this filtered table:
  `th_chow_chol <- pull(th_chow, Cholesterol)`
- To find the mean cholesterol for this group, use
  `mean(th_chow_chol)`
- And to find the median
  `median(th_chow_chol)`
- Repeat to find the mean and median Cholesterol for the TH HF group
- Which group appears to have the higher Cholesterol?

Mean versus median

- Which *measure of central tendency* should we use? Mean or median?
- The mean has more useful mathematical properties
  - We will discuss these briefly later in the course
  - Allows us to do more powerful statistics, such as hypothesis testing
- However, the mean is not ”robust to outliers”
  - If one of our values was accidentally recorded with a large error (say by a factor of 10), this would greatly impact the mean
- The median, by contrast, is very robust to outliers
  - Even multiplying a value by a factor of 10 in error might not change the median at all

Measures of spread

- As well as knowing what a “typical” value in our data set looks like, we should also ask how representative this typical value is of the data set
- To do this, we can measure the “spread” of the data: how far is a value in the data set from our average
- There are two ways to do this:
  - Measure the *standard deviation*. Roughly speaking, this is the average distance of a data point from the mean of the data. This makes most sense when we are working with the mean.
  - Measure the *interquartile range.* This is the spread of the “middle half” of the data. This makes most sense when we are working with the median.

The standard deviation

- The standard deviation is computed by taking the sum of the squares of the difference between each data point and the mean, dividing by the number of data points, and then taking the square root
- In R, we can do
  `sd(th_chow_chol)`
- What is the standard deviation of the cholesterol in the TH HF group?

The interquartile range

- The interquartile range is the difference between the value that is the 25th percentile and the 75th percentile in the data
- Experiment with the following in R:

```
quantile(th_chow_chol, 0.25)
quantile(th_chow_chol, 0.75)
IQR(th_chow_chol)
```

- Note that the IQR is the difference between the first two values.
- What is the IQR for cholesterol in the TH HF group?