DAY 2: DATA COLLECTION & NUMERICAL SUMMARIES

BSTA 511/611 Fall 2023, OHSU

Meike Niederhausen, PhD

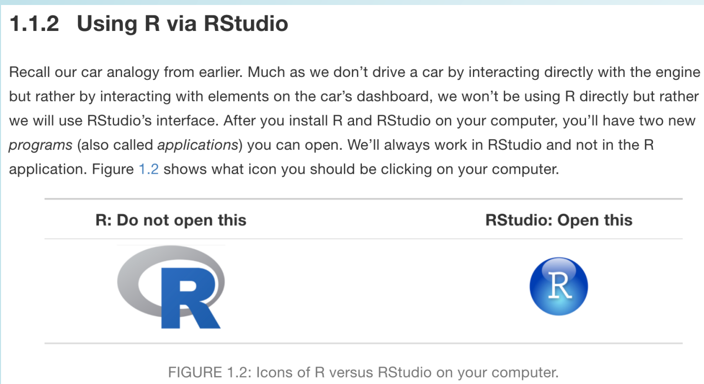
2023-10-02

GOALS FOR TODAY

- (1.3) **Data collection principles**
  - Population vs. sample
  - Sampling methods
  - Experiments vs. Observational studies
- (1.2) **Intro to Data**
  - Data types
  - How are data stored in R?
  - Working with data in R
- (1.4) **Summarizing numerical data**
  - Mean, median, mode, SD, IQR, range, 5 number summary
  - Empirical Rule
  - robust statistics
- **R packages** -> **install for next class!!!**

RECAP OF LAST TIME

- Open RStudio on your computer (not R!)



1.1.2 Using R via RStudio

Recall our car analogy from earlier. Much as we don't drive a car by interacting directly with the engine but rather by interacting with elements on the car's dashboard, we won't be using R directly but rather we will use RStudio's interface. After you install R and RStudio on your computer, you'll have two new *programs* (also called *applications*) you can open. We'll always work in RStudio and not in the R application. Figure 1.2 shows what icon you should be clicking on your computer.

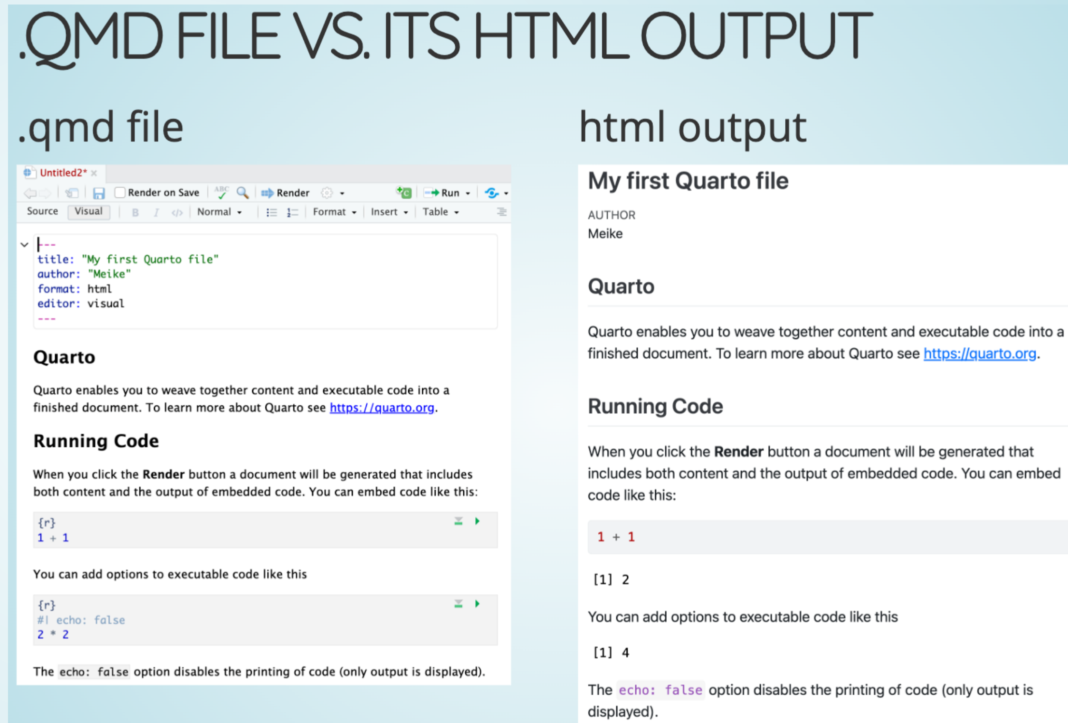
R: Do not open this	RStudio: Open this

FIGURE 1.2: Icons of R versus RStudio on your computer.

Modern Dive

- Basic math using R

- Creating and rendering Quarto files



- Formatting text & headers
- Code chunks

USEFUL KEYBOARD SHORTCUTS

Full list of keyboard shortcuts

action	mac	windows/linux
Run code in qmd (or script)	cmd + enter	ctrl + enter
`<-`	option + -	alt + -
interrupt currently running command	esc	esc
in console, retrieve previously run code	up/down	up/down
keyboard shortcut help	option + shift + k	alt + shift + k

PRACTICE

Try typing code below in your qmd (with shortcut) and evaluating it:

```
y <- 5
y
```

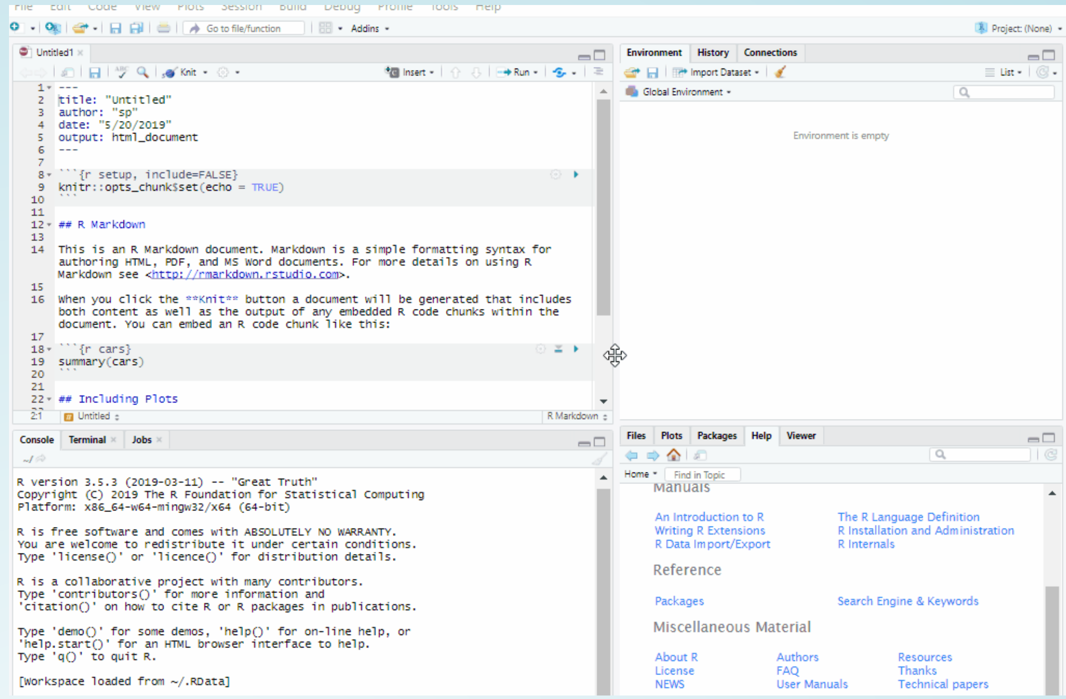
ANOTHER RESOURCE FOR AN INTRODUCTION TO R

- If you would like another perspective on what we covered the first week, you might find **Danielle Navarro's** online book ***Learning Statistics with R*** to be helpful.
- Download free pdf: https://learningstatisticswithr.com/
- See Sections 3.1-3.7.1 for some of the topics we covered on first day

MORITZ'S TIP OF THE DAY

Customize your RStudio interface!

https://www.pipinghotdata.com/posts/2020-09-07-introducing-the-rstudio-ide-and-r-markdown/#background



(1.3) DATA COLLECTION PRINCIPLES

- Population vs. sample
- Sampling methods
- Experiments vs. Observational studies

POPULATION VS. SAMPLE

(TARGET) POPULATION

- group of interest being studied
- group from which the sample is selected
  - studies often have *inclusion* and/or *exclusion* criteria

SAMPLE

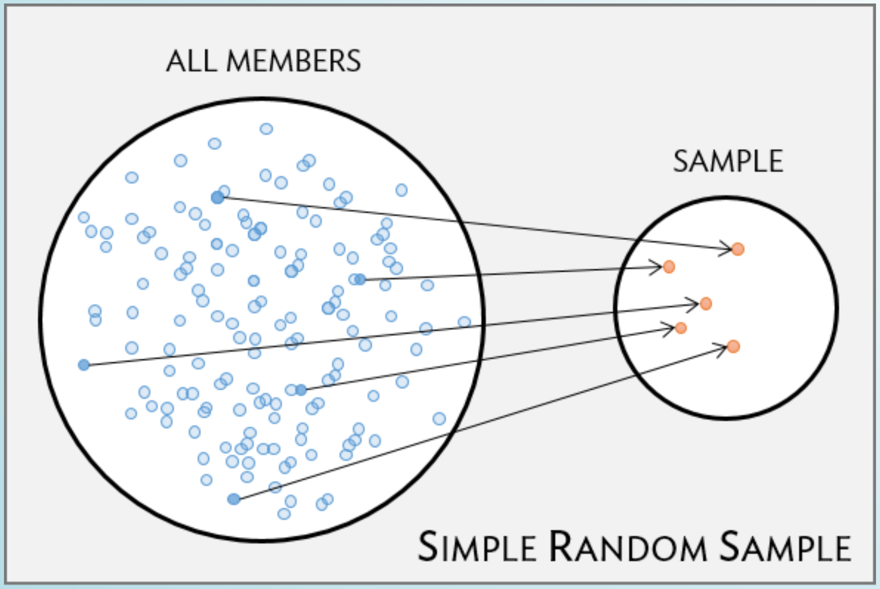
- group on which data are collected
- often a small subset of the population

SAMPLING METHODS (1/4)

Goal is to get a **representative** sample of the population:
the characteristics of the sample are similar to the characteristics of the population

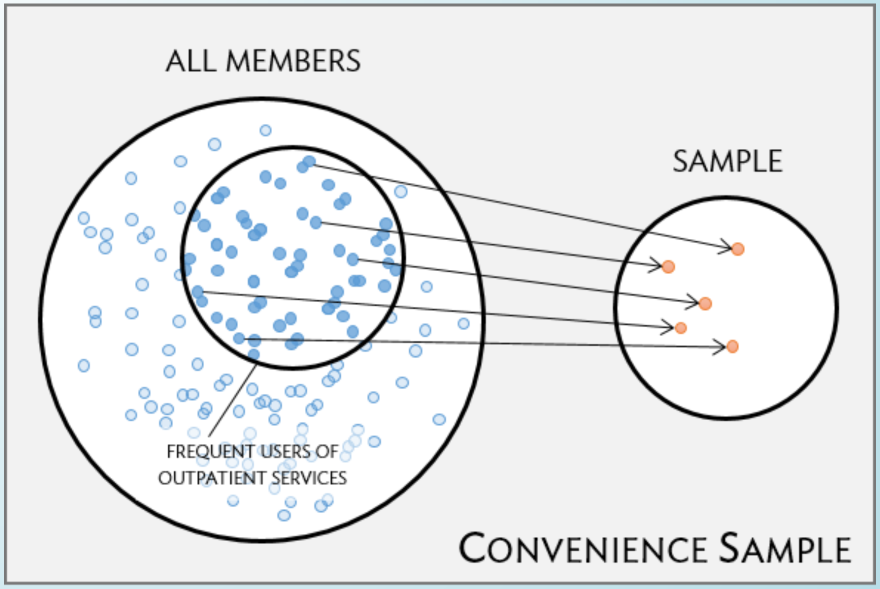
Simple random sample (SRS)

- each individual of a population has the *same chance* of being sampled
- randomly sampled
- considered best way to sample



Convenience sample

- easily accessible individuals are *more likely* to be included in the sample than other individuals
- a common "pitfall"

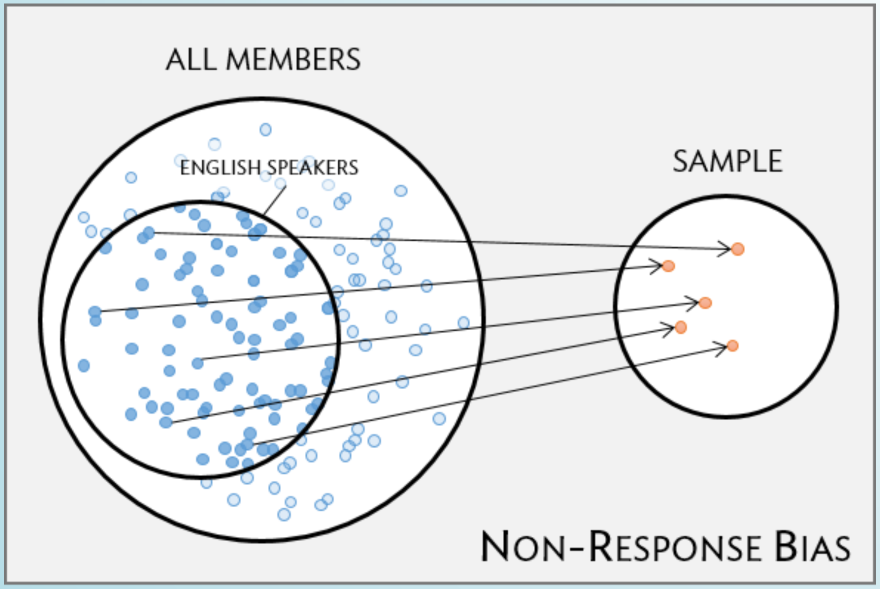


SAMPLING METHODS (2/4)

Good sampling plans don't guarantee samples representative of the population

Non-response bias

- non-response rates can be high
- are all groups within a population being reached?
- unrepresentative sample => skewed results

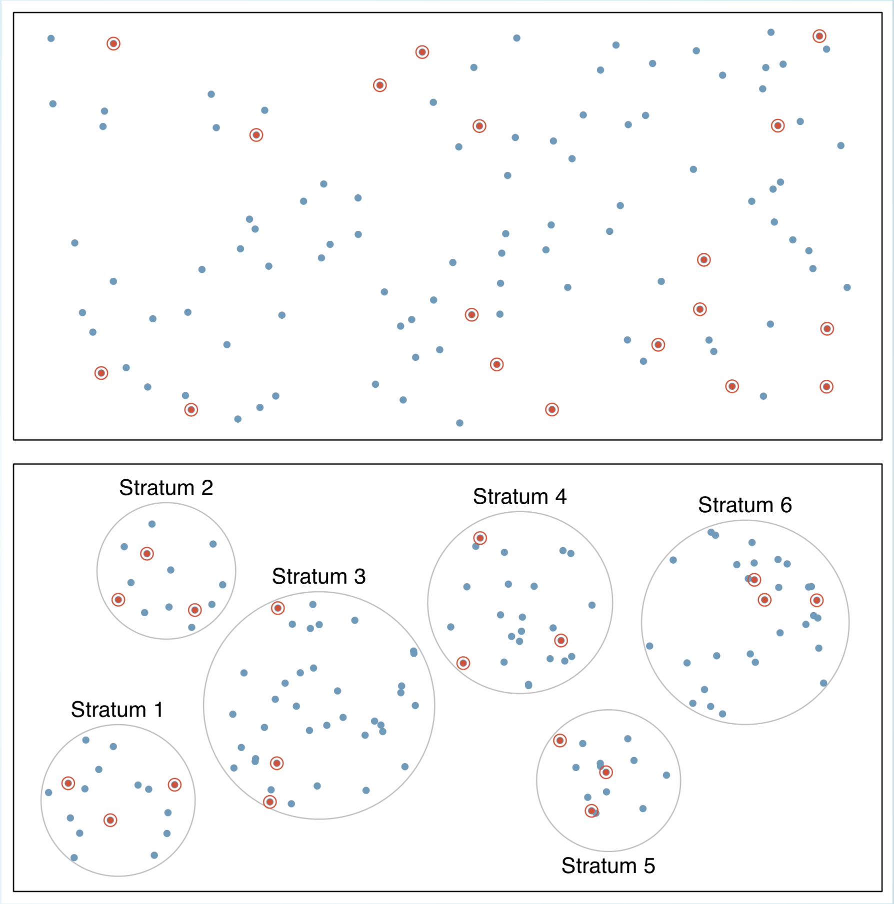


"Random" samples can be unrepresentative by random chance

- In a SRS each case in the population has an equal chance of being included in the sample
- But by random chance alone a random sample might contain a higher proportion of one group over another
- Ex: a SRS might by chance include 70% men (unlikely, but theoretically possible)

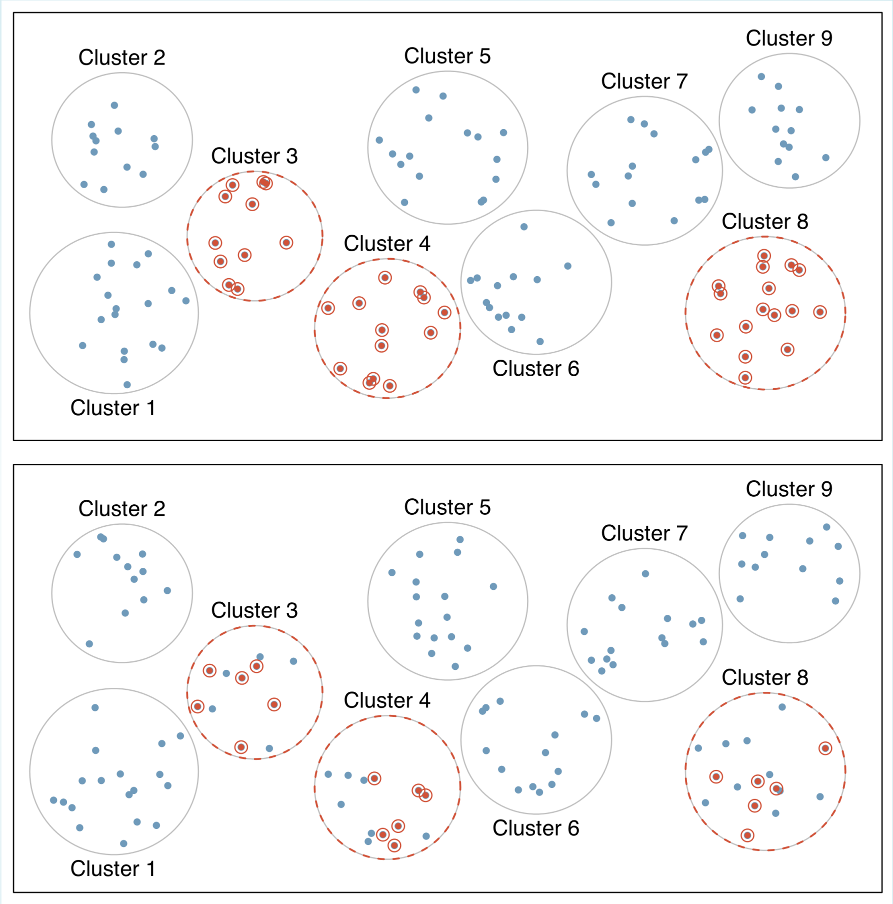
SAMPLING METHODS (3/4)

- **Simple random sample (SRS)**
  - each individual of a population has the *same chance* of being sampled
  - *statistical methods taught in this class assume a SRS!*
- **Stratified sampling**
  - divide population into groups (strata) before selecting cases within each stratum (often via SRS)
  - usually cases within a strata are similar, but are different from other strata with respect to the outcome of interest, such as gender or age groups



SAMPLING METHODS (4/4)

- **Cluster sample**
  - first divide population into groups (clusters)
  - then sample a fixed number of clusters, and include *all* observations from chosen clusters
  - clusters are often hospitals, clinicians, schools, etc., where each cluster will have similar services/ policies/ etc.
  - cases within clusters usually very diverse
- **Multistage sample**
  - similar to a cluster sample, but select a random sample within each selected cluster instead of all individuals



EXPERIMENTS (1/2)

- Researchers assign individuals to different **treatment** or **intervention groups**
  - **control group**: often receive a **placebo** or usual care
  - different treatment groups are often called **study arms**
- **Randomization**
  - group assignment is usually random to ensure similar (balanced) study arms for all variables (observed and unobserved)
  - randomization allows study arm differences in outcomes to be attributed to treatment rather than variability in patient characteristics
    - treatment is the only systematic difference between groups
    - establish causality
  - **blocking (stratification)**: group individuals into blocks (strata) before randomizing if there are certain characteristics that may influence the outcome other than treatment (i.e. gender, age group)

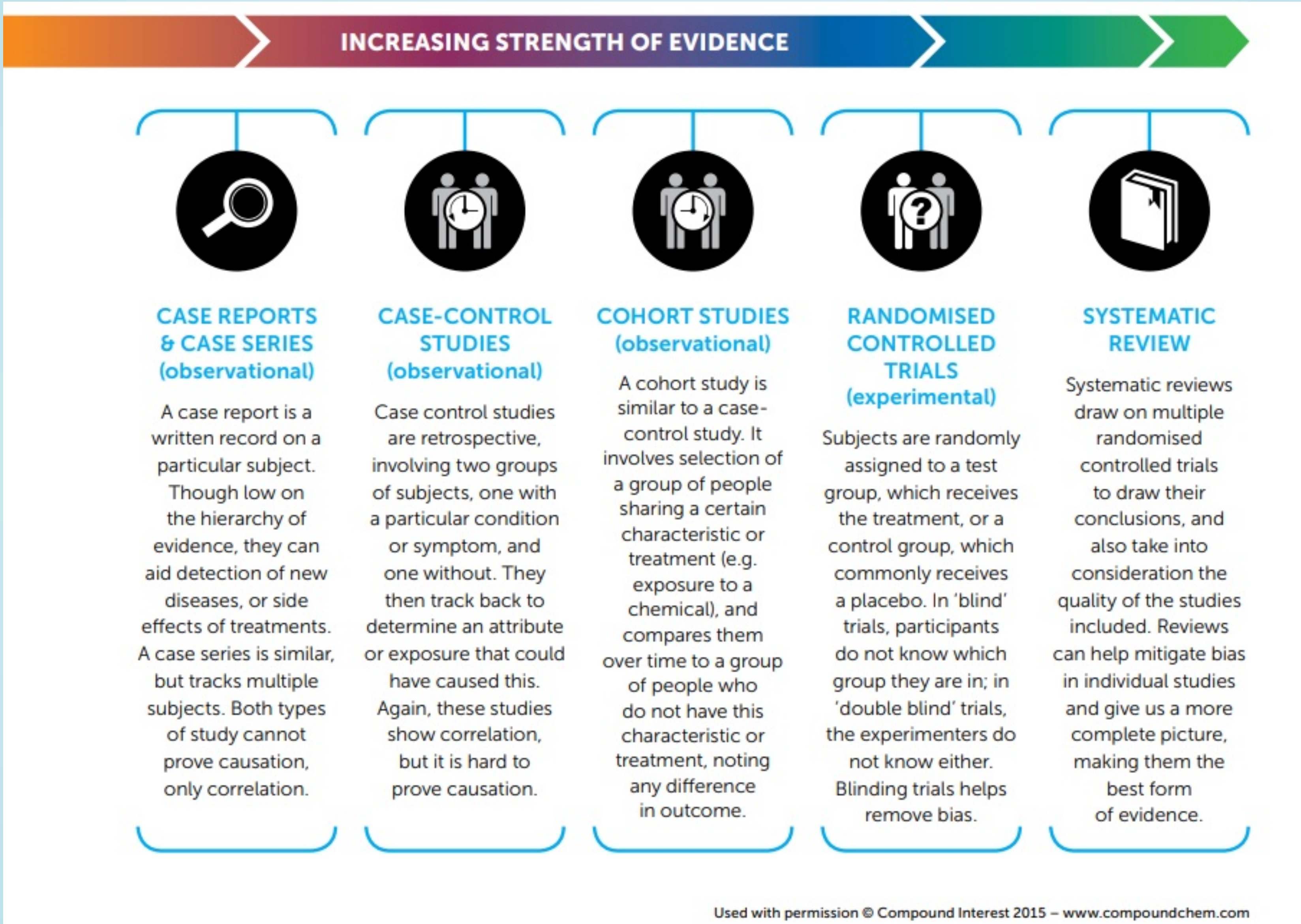
EXPERIMENTS (2/2)

- **Replication**
  - accomplished by collecting a sufficiently large sample
  - results usually more reliable with a large sample size
    - often less variability
    - more likely to be representative of population
- Some studies are not ethical to carry out as experiments

OBSERVATIONAL STUDIES

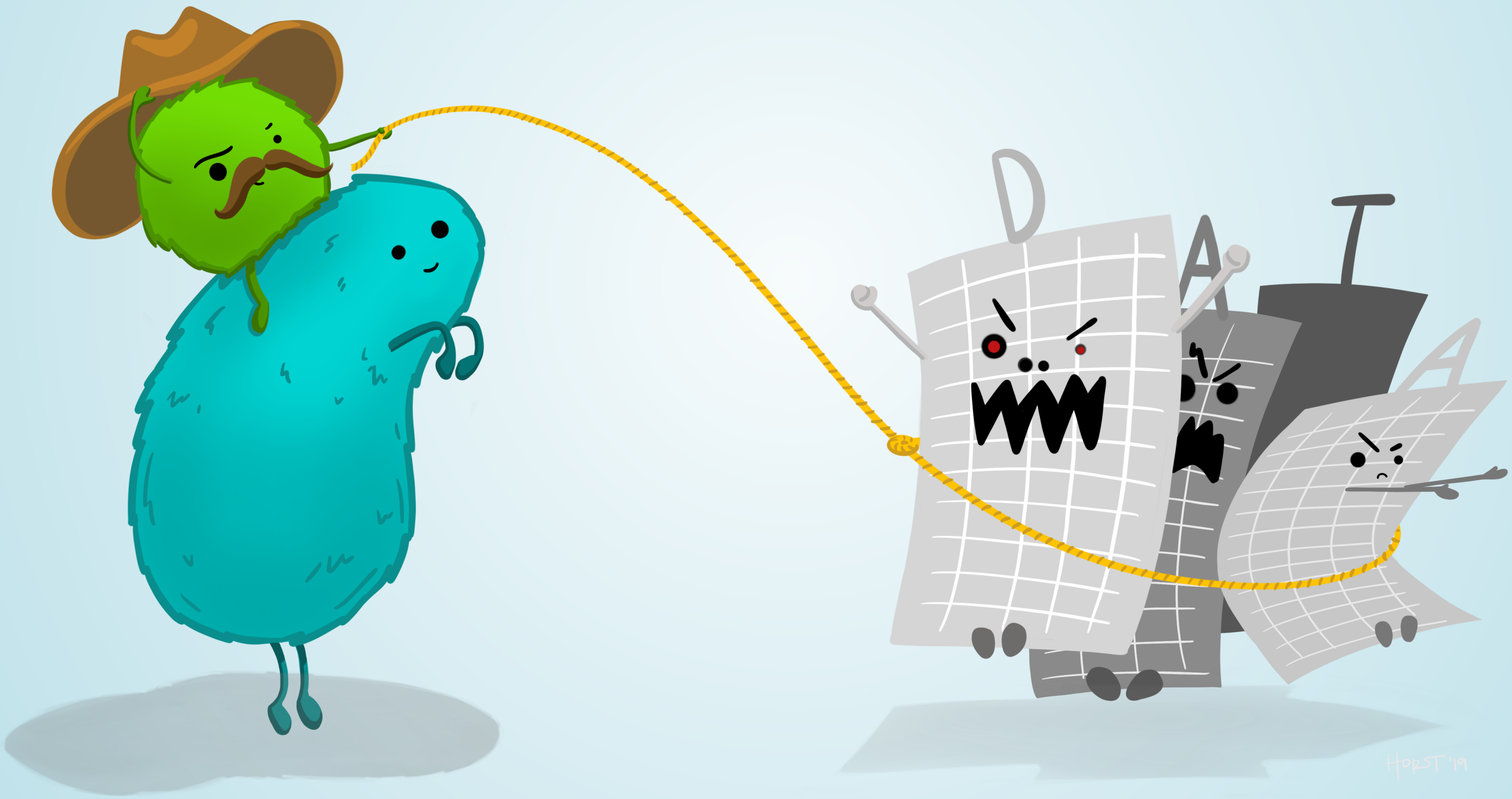
- data are observed and recorded without interference
- often done via surveys, electronic health records, or medical chart reviews
- cohorts
- associations between variables can be established, but not causality
  - Individuals with different characteristics may also differ in other ways that influence response
- confounding variables (lurking variable)
  - variables associated with both the explanatory and response variables
- prospective vs. retrospective studies

COMPARING STUDY DESIGNS



Science Media Centre

(1.2) INTRO TO DATA



Artwork by @allison_horst

HOW ARE DATA STORED, HOW DO WE USE THEM?

- Often, data are in an Excel sheet, or a plain text file (.csv, .txt)
- .csv files open in Excel automatically, but actually are plain text
- Usually, columns are variables/measures and rows are observations (i.e. a person's measurements)

DATA IN R

- We can import data from many file types, including .csv, .txt., and .xlsx
  - We will cover this on a later date
- Once imported, R typically stores data as **data frames**, or **tibbles** if using the `tidyverse` package (more on this later).
  - For our purposes, these are essentially the same, and I will tend to use the terms interchangeably.
  - These are examples of what we call **object types** in R.

DATA FRAME EXAMPLE

```r
df <- data.frame(
  IDs=1:3,
  gender=c("male", "female", "Male"),
  age=c(28, 35.5, 31),
  trt = c("control", "1", "1"),
  Veteran = c(FALSE, TRUE, TRUE)
  )
df
```

```
  IDs gender  age     trt Veteran
1   1   male 28.0 control   FALSE
2   2 female 35.5       1    TRUE
3   3   Male 31.0       1    TRUE
```

- **Vectors** vs. **data frames**
  - a data frame is a collection (or array or table) of vectors

- Different columns can be of different data types (i.e. numeric vs. text)
- Both numeric and text can be stored within a column (stored together as *text*).
- Vectors and data frames are examples of ***objects*** in R.
  - There are other types of R objects to store data, such as matrices, lists.

OBSERVATIONS & VARIABLES

```
df
```

```
  IDs gender  age     trt Veteran
1   1   male 28.0 control   FALSE
2   2 female 35.5       1    TRUE
3   3   Male 31.0       1    TRUE
```

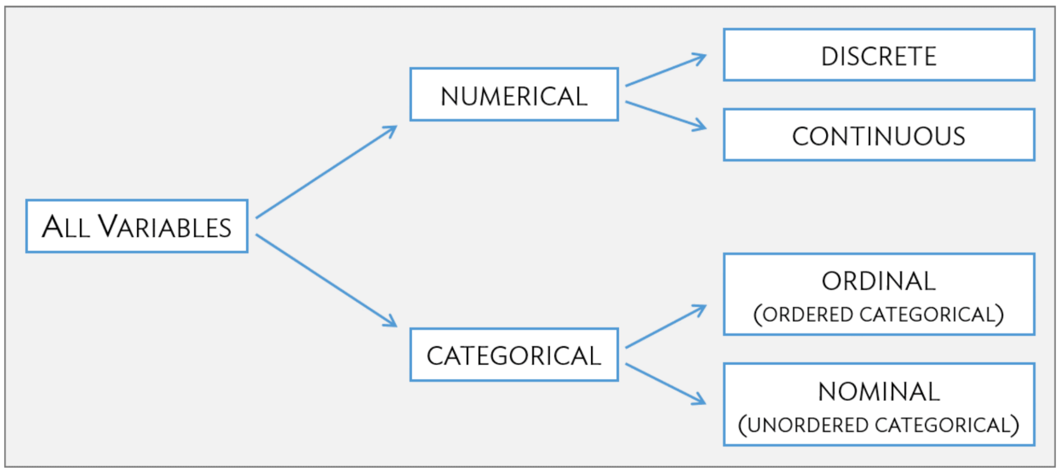


Figure 1.8: Breakdown of variables into their respective types.

ISLBS

- Book refers to a dataset as a *data matrix*
- Rows are usually **observations**
- Columns are usually **variables**
- **How many observations are in this dataset?**
- **What are the variable types in this dataset?**

VARIABLE (COLUMN) TYPES

R type	variable type	description
integer	discrete	integer-valued numbers
double or numeric	continuous	numbers that are decimals
factor	categorical	categorical variables stored with levels (groups)
character	categorical	text, “strings”
logical	categorical	boolean (TRUE, FALSE)

- View the **structure** of our data frame to see what the variable types are:

```
str(df)
```

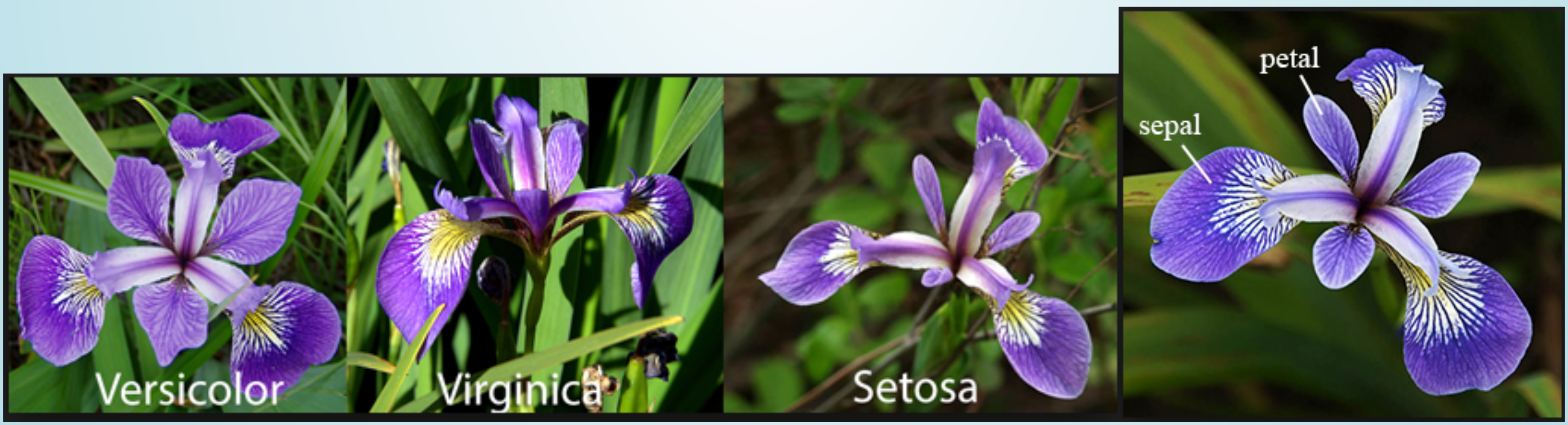
```
'data.frame':	3 obs. of  5 variables:
 $ IDs     : int  1 2 3
 $ gender  : chr  "male" "female" "Male"
 $ age     : num  28 35.5 31
 $ trt     : chr  "control" "1" "1"
 $ Veteran: logi  FALSE TRUE TRUE
```

FISHER'S (OR ANDERSON'S) IRIS DATA SET

Data description:

- n = 150
- 3 species of Iris flowers (Setosa, Virginica, and Versicolour)
  - 50 measurements of each type of Iris
- **variables**:
  - sepal length, sepal width, petal length, petal width, and species

Can the iris species be determined by these variables?



VIEW THE `iris` DATASET

- The `iris` dataset is already pre-loaded in *base* R and ready to use.
- Type the following command in the console window
  - *Warning: this command cannot be rendered. It will give an error.*

```
View(iris)
```

A new tab in the scripting window should appear with the `iris` dataset.

iris

Filter

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

DATA STRUCTURE

- What are the different **variable types** in this data set?

```
str(iris)   # structure of data
```

```
'data.frame':	150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1
...
```

DATA SET SUMMARY

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
       Species  
 setosa    :50  
 versicolor:50  
 virginica :50  
```

DATA SET INFO

```
dim(iris)
```

```
[1] 150   5
```

```
nrow(iris)
```

```
[1] 150
```

```
ncol(iris)
```

```
[1] 5
```

```
names(iris)
```

```
[1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

VIEW THE BEGINNING OR END OF A DATASET

```r
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```r
tail(iris)
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
145          6.7         3.3          5.7         2.5 virginica
146          6.7         3.0          5.2         2.3 virginica
147          6.3         2.5          5.0         1.9 virginica
148          6.5         3.0          5.2         2.0 virginica
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

SPECIFY HOW MANY ROWS TO VIEW AT BEGINNING OR END OF A DATASET

```r
head(iris, 3)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
```

```r
tail(iris, 2)
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width   Species
149          6.2         3.4          5.4         2.3 virginica
150          5.9         3.0          5.1         1.8 virginica
```

THE $

- Suppose we want to single out the column of petal width values.
- One way to do this is to use the `$`
  - `DatSetName$VariableName`

```
iris$Petal.Width
```

```
  [1] 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 0.2 0.2 0.1 0.1 0.2 0.4 0.4 0.3
 [19] 0.3 0.3 0.2 0.4 0.2 0.5 0.2 0.2 0.4 0.2 0.2 0.2 0.2 0.4 0.1 0.2 0.2 0.2
 [37] 0.2 0.1 0.2 0.2 0.3 0.3 0.2 0.6 0.4 0.3 0.2 0.2 0.2 0.2 1.4 1.5 1.5 1.3
 [55] 1.5 1.3 1.6 1.0 1.3 1.4 1.0 1.5 1.0 1.4 1.3 1.4 1.5 1.0 1.5 1.1 1.8 1.3
 [73] 1.5 1.2 1.3 1.4 1.4 1.7 1.5 1.0 1.1 1.0 1.2 1.6 1.5 1.6 1.5 1.3 1.3 1.3
 [91] 1.2 1.4 1.2 1.0 1.3 1.2 1.3 1.3 1.1 1.3 2.5 1.9 2.1 1.8 2.2 2.1 1.7 1.8
[109] 1.8 2.5 2.0 1.9 2.1 2.0 2.4 2.3 1.8 2.2 2.3 1.5 2.3 2.0 2.0 1.8 2.1 1.8
[127] 1.8 1.8 2.1 1.6 1.9 2.0 2.2 1.5 1.4 2.3 2.4 1.8 1.8 2.1 2.4 2.3 1.9 2.3
[145] 2.5 2.3 1.9 2.0 2.3 1.8
```

EXAMPLE USING THE $

The $ is helpful if you want to create a new dataset for just that one variable, or, more commonly, if you want to calculate summary statistics for that one variable.

```
mean(iris$Petal.Width)
```

```
[1] 1.199333
```

```
sd(iris$Petal.Width)
```

```
[1] 0.7622377
```

```
median(iris$Petal.Width)
```

```
[1] 1.3
```

INLINE CODE

- With markdown you can also report **R code output inline** with the text instead of using a chunk.

Text in editor:

```
The mean petal width for all 3 species combined
is `r round(mean(iris$Petal.Width),1)`
(SD = `r round(sd(iris$Petal.Width),1)`) cm.
```

Output:

The mean petal width for all 3 species combined is 1.2 (SD = 0.8) cm.

- Reporting summary statistics this way in a report, makes the numbers computationally reproducible.
- For example, if this were for an abstract and a year later you are wondering where the numbers came from, your R code will tell you exactly which dataset was used to calculate the values.

(1.4) SUMMARIZING NUMERICAL DATA

Measures of center & spread

https://xkcd.com/937/

TABLE 1 EXAMPLE

Table 1. Patient characteristics, overall and by concordance

		Total N=204	Discordant N=40	Concordant N=164	p-value
Site, n (%)	OHSU	122 (62.7%)	26 (65.0%)	96 (62.2%)	0.86
	VA	76 (37.3%)	14 (35.0%)	62 (37.8%)	
Gender, n (%)	Male	85 (41.7%)	18 (45.0%)	67 (40.9%)	0.72
	Female	119 (58.3%)	22 (55.0%)	97 (59.1%)	
Age (years), mean (SD)		57.2 (14.2)	58.2 (15.1)	56.9 (14.0)	0.62
Language, n (%)	English	168 (84.4%)	35 (92.1%)	133 (82.6%)	0.21
	Spanish	31 (15.6%)	3 (7.9%)	28 (17.4%)	
Limited English language proficiency, n (%)		30 (15.1%)	3 (7.9%)	27 (16.8%)	0.17
Coupled, n (%)		110 (57.9%)	22 (61.1%)	88 (57.1%)	0.71
Education, n (%)	High school or less	60 (31.6%)	15 (40.5%)	45 (29.4%)	0.24
	Some college or more	130 (68.4%)	22 (59.5%)	108 (70.6%)	
Income, >$40,000, n (%)	Less than $40,000	85 (45.5%)	12 (33.3%)	73 (48.3%)	0.14
	Greater than $40,000	102 (54.5%)	24 (66.7%)	78 (51.7%)	
People in household, median (IQR)		2 (2-4)	2 (2-3)	2 (2-4)	0.92
Race/Ethnicity, n (%)	White	123 (68.3%)	25 (78.1%)	98 (66.2%)	0.62
	Black	6 (3.3%)	0 (0.0%)	6 (4.1%)	
	Latinx/Hispanic	39 (21.7%)	6 (18.8%)	33 (22.3%)	
	Other	12 (6.7%)	1 (3.1%)	11 (7.4%)	
Limited health literacy, n (%)		55 (28.6%)	13 (35.1%)	42 (27.1%)	0.42
Disease duration (years), median (IQR)		8 (4-16)	13 (5-21)	7 (4-15)	0.039
Number of medications, median (IQR)		1 (1-2)	1 (0-2)	1 (1-2)	0.10
Depressive symptoms, n (%)		38 (20.8%)	3 (8.1%)	35 (24.0%)	0.040
PTSD, n (%)		13 (7.1%)	2 (5.6%)	11 (7.5%)	1.00
Self-efficacy score, mean (SD)		6.3 (2.1)	6.3 (2.1)	6.3 (2.1)	0.96
Trust in Physician, n (%)		106 (53.8%)	19 (51.4%)	87 (%)	0.74
Disease activity score (CDAI), mean (SD)		12.8 (10.5)	10.5 (9.7)	13.2 (10.8)	0.21
Medication Adherence, n (%)	High	63 (33.5%)	7 (20.6%)	56 (36.4%)	0.11
	Low/Medium	125 (66.5%)	27 (79.4%)	98 (63.6%)	

Abbreviations: IQR, interquartile range; PTSD, post-traumatic stress disorder; SD, standard deviation; OHSU, Oregon Health & Science University; VA, Veterans Affairs; CDAI, Clinical Disease Activity Index

Are We on the Same Page?: A Cross-Sectional Study of Patient-Clinician Goal Concordance in Rheumatoid Arthritis
J Barton et al.
Arthritis Care & Research.
2021 Sep 27
https://pubmed.ncbi.n

MEASURES OF CENTER: MEAN

Sample mean: the average value of observations

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{i=1}^{n} \frac{x_i}{n}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values in a sample

Example: What is the mean age in the toy dataset `df` defined earlier?

```
df
```

```
  IDs gender  age     trt Veteran
1   1   male 28.0 control   FALSE
2   2 female 35.5       1    TRUE
3   3   Male 31.0       1    TRUE
```

```
mean(df$age)
```

```
[1] 31.5
```

MEASURES OF CENTER: MEDIAN

- The **median** is the middle value of the observations in a sample.
- The median is the 50th percentile, meaning
  - 50% of observations lie below and
  - 50% of observations lie above the median.

```r
df$age
```

```
[1] 28.0 35.5 31.0
```

```r
median(df$age)
```

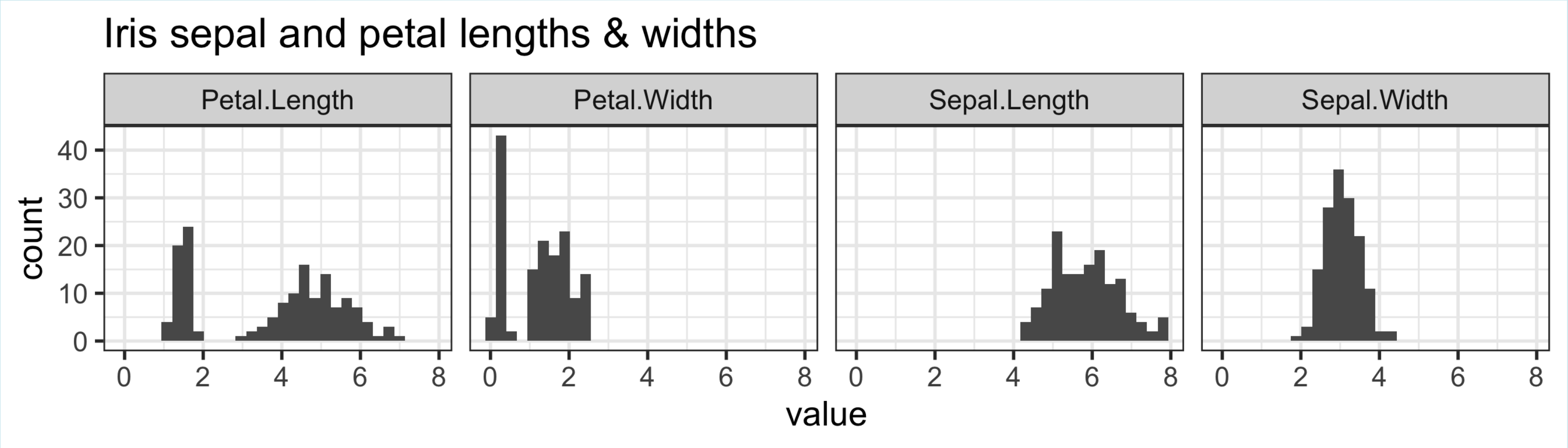
```
[1] 31
```

```r
median(c(df$age, 67))
```

```
[1] 33.25
```

- If the number of observations is
  - odd: the median is the middle observed value
  - even: the median is the average of the two middle observed values

MEASURES OF CENTER: MEAN VS. MEDIAN




```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width   
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100  
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300  
 Median :5.800   Median :3.000   Median :4.350   Median :1.300  
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199  
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800  
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500  
       Species  
 setosa    :50  
 versicolor:50  
 virginica :50  
```

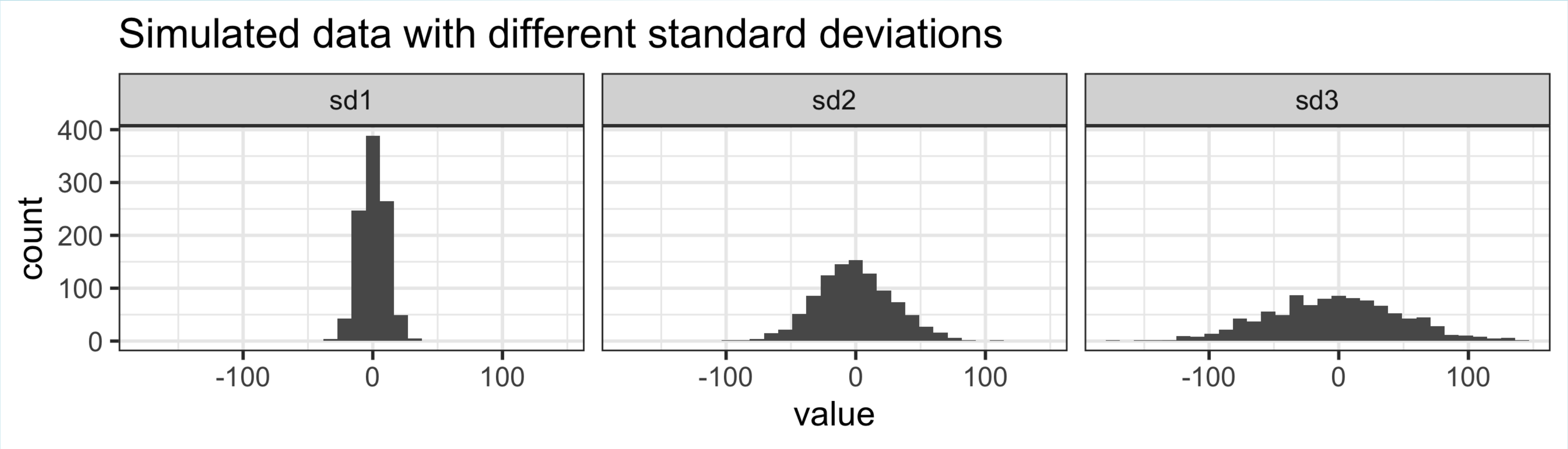
MEASURES OF CENTER: MODE

mode: the most frequent value in a dataset

MEASURES OF SPREAD: STANDARD DEVIATION (SD) (1/3)

standard deviation is (approximately) the average distance between a typical observation and the mean

- An observation's **deviation** is the distance between its value x and the sample mean $\overline{x}$: deviation = $x - \overline{x}$.



MEASURES OF SPREAD: SD (2/3)

- The **sample variance** s^2 is the sum of squared deviations divided by the number of observations minus 1.

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$$

where $x_1, x_2, \ldots, x_n$ represent the n observed values.

- The **standard deviation** s is the square root of the variance.

$$s = \sqrt{\frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1}} = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

MEASURES OF SPREAD: SD (3/3)

Let's calculate the sample standard deviation for our toy example

```r
df$age
```

```
[1] 28.0 35.5 31.0
```

$$s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}} =$$

```r
mean(df$age)
```

```
[1] 31.5
```

```r
sd(df$age)
```

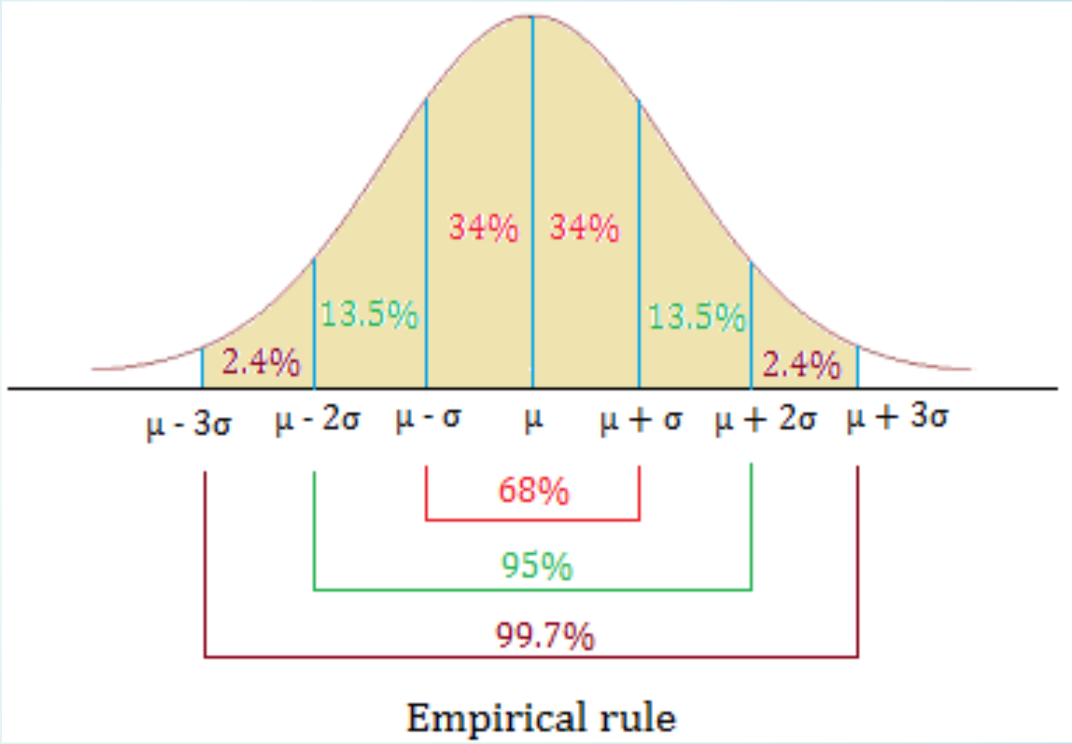
```
[1] 3.774917
```

EMPIRICAL RULE: ONE WAY TO THINK ABOUT THE SD (1/2)

For symmetric bell-shaped data, about

- 68% of the data are within 1 SD of the mean
- 95% of the data are within 2 SD's of the mean
- 99.7% of the data are within 3 SD's of the mean

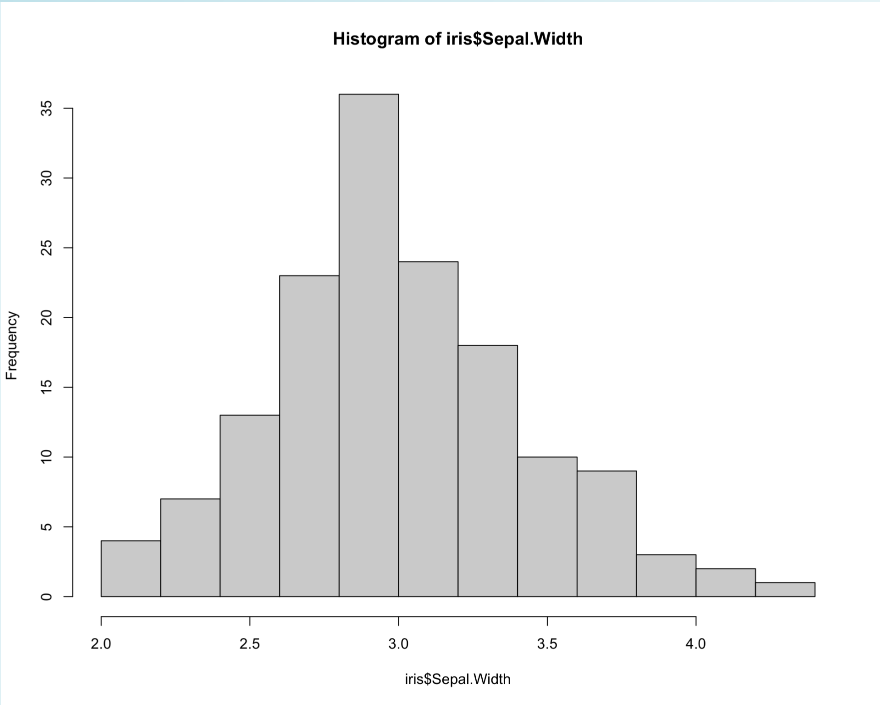
These percentages are based off of percentages of a true normal distribution.



https://statistics-made-easy.com/empirical-rule/

EMPIRICAL RULE: ONE WAY TO THINK ABOUT THE SD (2/2)

```r
hist(iris$Sepal.Width)
```




```r
mean(iris$Sepal.Width)
```

```
[1] 3.057333
```

```r
sd(iris$Sepal.Width)
```

```
[1] 0.4358663
```

MEASURES OF SPREAD: INTERQUARTILE RANGE (IQR) (1/2)

The p^{th} percentile is the observation such that $p\%$ of the remaining observations fall below this observation.

- The *first quartile* Q_1 is the 25^{th} percentile.
- The *second quartile* Q_2, i.e., the median, is the 50^{th} percentile.
- The *third quartile* Q_3 is the 75^{th} percentile.

The **interquartile range (IQR)** is the distance between the third and first quartiles.

$$IQR = Q_3 - Q_1$$

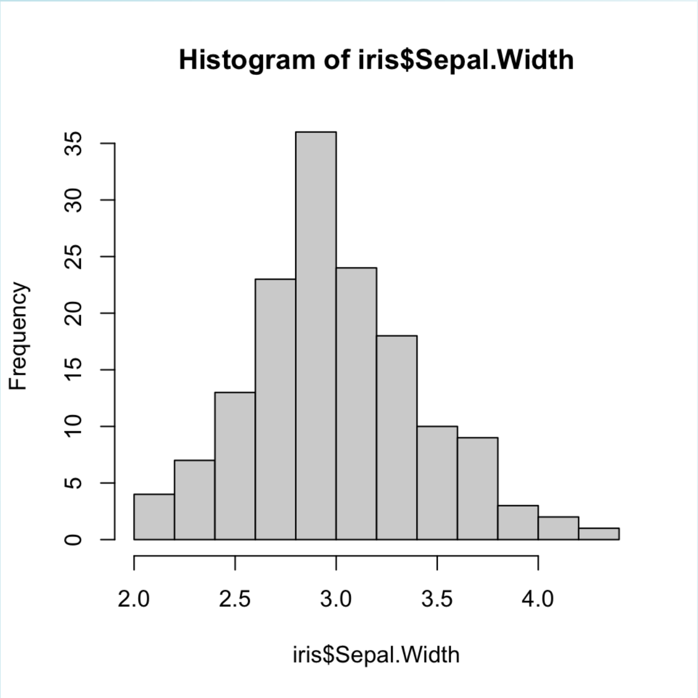
- IQR is the width of the *middle half* of the data

MEASURES OF SPREAD: IQR (2/2)

5 number summary

```
1  summary(iris$Sepal.Width)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   2.800   3.000   3.057   3.300   4.400 
```



What is the IQR of the sepal widths?

```
1  quantile(iris$Sepal.Width, c(.25, .75))
```

```
25% 75% 
2.8 3.3 
```

```
1  diff(quantile(iris$Sepal.Width, c(.25, .75)))
```

```
75% 
0.5 
```

```
1  IQR(iris$Sepal.Width)
```

```
[1] 0.5
```

ROBUST ESTIMATES

Summary statistics are called **robust estimates** if extreme observations have little effect on their values

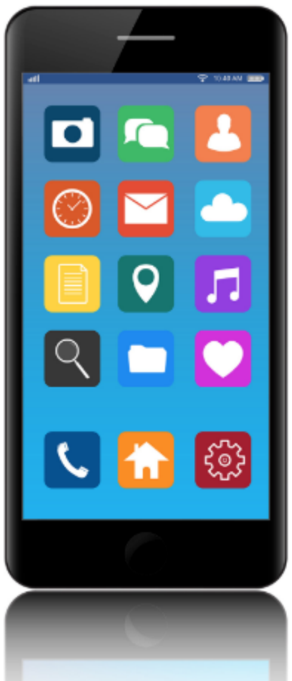
estimate	**robust?**
mean	
median	
mode	
standard deviaiton	
IQR	
range	

R PACKAGES

stringr
forcats
tidyverse
dplyr
tidyr
rmarkdown
lubridate
TIBBLE
purrr
readr
Horst

R PACKAGES

A good analogy for R packages is that they are like apps you can download onto a mobile phone:

R: A new phone	R Packages: Apps you can download
	

ModernDive Figure 1.4

INSTALLING PACKAGES

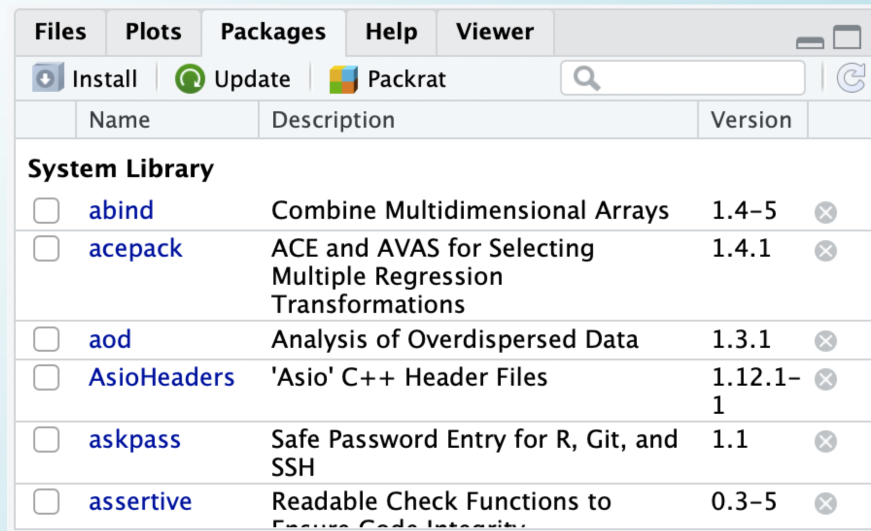
- Packages contain additional functions and data

Two options to install packages:

1. `install.packages()` *or*
2. The "Packages" tab in Files/Plots/Packages/Help/Viewer window

```
install.packages("dplyr")    # only do this ONCE, use quotes
```

- **Only install packages once** *(unless you want to update them)*
- Installed from Comprehensive R Archive Network (CRAN) = package mothership



VIDEO ON INSTALLING PACKAGES

- Danielle Navarro's YouTube video on ***Installing and loading R packages***: https://www.youtube.com/watch?v=kpHZVyDvEhQ

LOAD PACKAGES WITH `library()` COMMAND

- Tip: at the top of your Rmd file, create a chunk that loads all of the R packages you want to use in that file.
- Use the `library()` command to load each required package.
- Packages need to be reloaded *every* time you open Rstudio.

```
library(dplyr)     # run this every time you open Rstudio
```

- You can use a function without loading the package with `PackageName::CommandName`

```
dplyr::arrange(iris, Petal.Width)    # what does arrange do?
```

```
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            4.9         3.1          1.5         0.1     setosa
2            4.8         3.0          1.4         0.1     setosa
3            4.3         3.0          1.1         0.1     setosa
4            5.2         4.1          1.5         0.1     setosa
5            4.9         3.6          1.4         0.1     setosa
6            5.1         3.5          1.4         0.2     setosa
7            4.9         3.0          1.4         0.2     setosa
8            4.7         3.2          1.3         0.2     setosa
```

INSTALL THE PACAKGES LISTED BELOW BEFORE DAY 3

- `knitr`
  - this might actually already be installed
  - check your packages list
- `tidyverse`
  - this is actually a bundle of packages
  - *Warning: it will take a while to install!!!*
  - see more info at https://tidyverse.tidyverse.org/
- `rstatix`
  - for summary statistics of a dataset
- `janitor`
  - for cleaning and exploring data
- `ggridges`
  - for creating ridgeline plots
- `devtools`
  - used to create R packages
  - for our purposes, needed to install some packages
- `oi_biostat_data`
  - this package is on github
  - **see the next slide for directions on how to install `oi_biostat_data`**

DIRECTIONS FOR INSTALLING PACKAGE `oibiostat`

- The textbook's datasets are in the R package `oibiostat`
- Explanation of code below
  - Installation of `oibiostat` package requires first installing `devtools` package
  - The code `devtools::install_github()` tells R to use the command `install_github()` from the `devtools` package without loading the entire package and all of its commands (which `library(devtools)` would do).

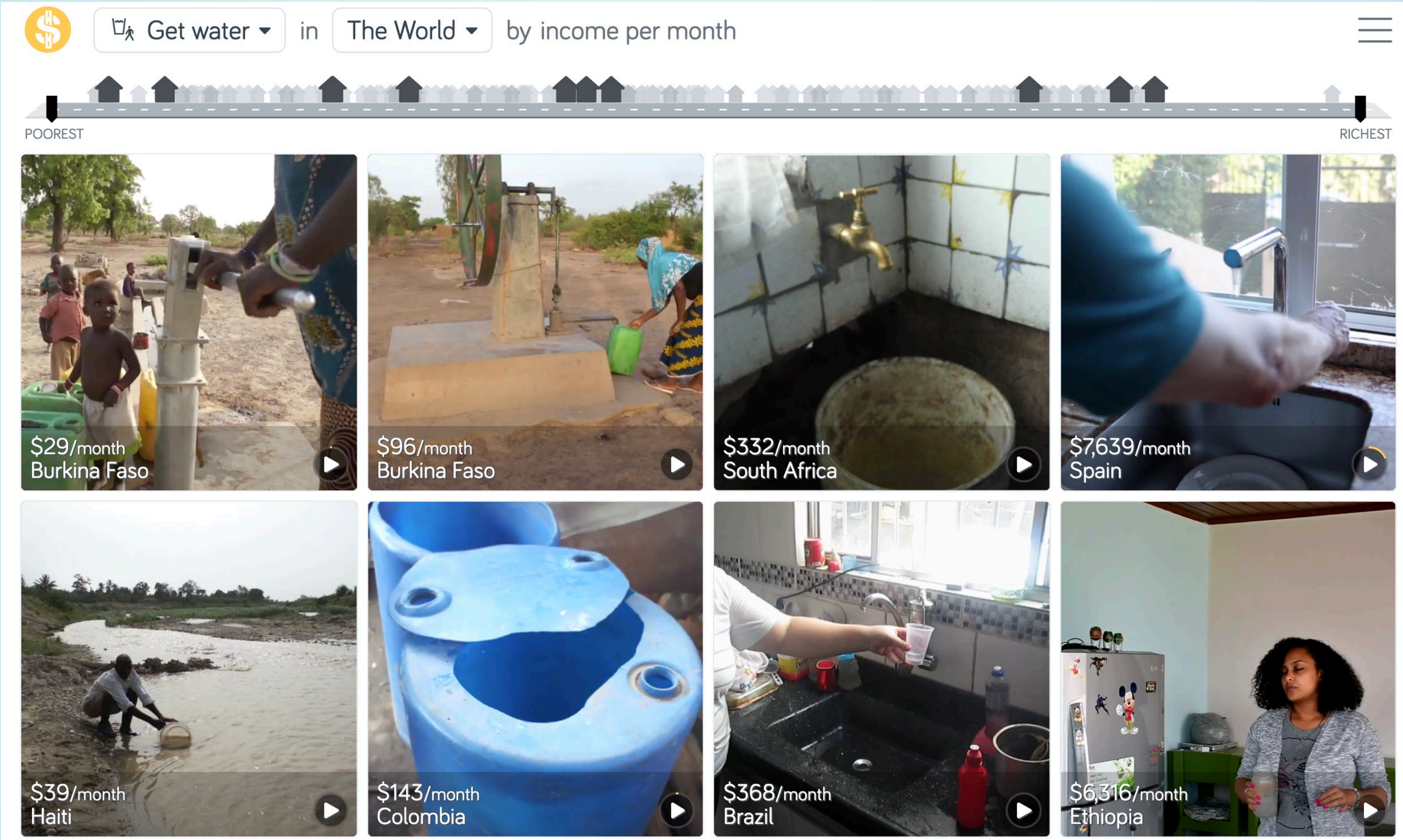
```
install.packages("devtools")
devtools::install_github("OI-Biostat/oi_biostat_data", force = TRUE)
```

- After running the code above, put `#` in front of the commands so that RStudio doesn't evaluate them when rendering.
- Now load the `oibiostat` package
  - the code below needs to be run *every time* you restart R or knit an Rmd file

```
library(oibiostat)
```

A VISUAL DATASET

Compare water sources across the world by country and family income



Gapminder Dollarstreet

Check out Gapminder's Dollar Street for many more examples:
https://www.gapminder.org/dollar-street