What is Data?

Author: Nicholas G Reich

This material is part of the ***statsTeachR*** *project*

Newton showed that the book of nature is written in the language of mathematics. Some chapters ... boil down to a clear-cut equation; but scholars who attempted to reduce biology, economics, and psychology to neat Newtonian equations have discovered that these fields have a level of complexity that makes such an aspiration futile.

This did not mean, however, that they gave up on mathematics.

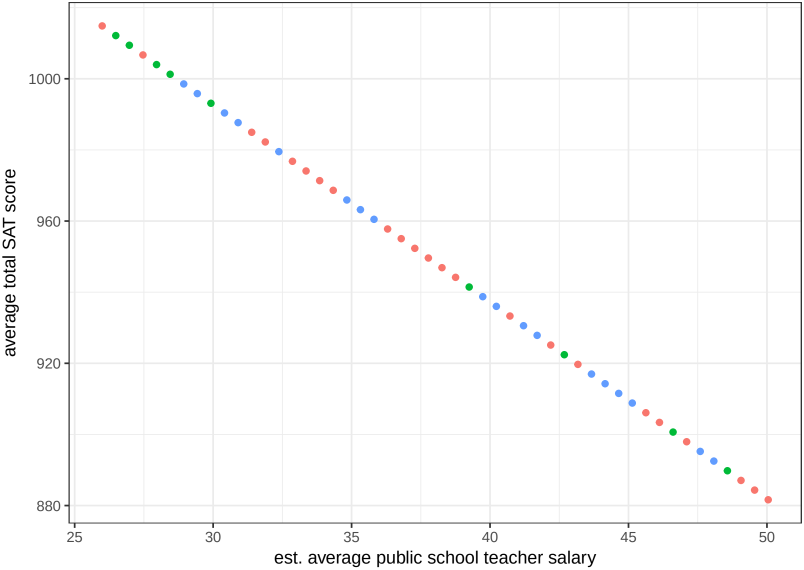
A new branch of mathematics was developed over the last 200 years to deal with the more complex aspects of reality: statistics.

- Yuval Noah Harari

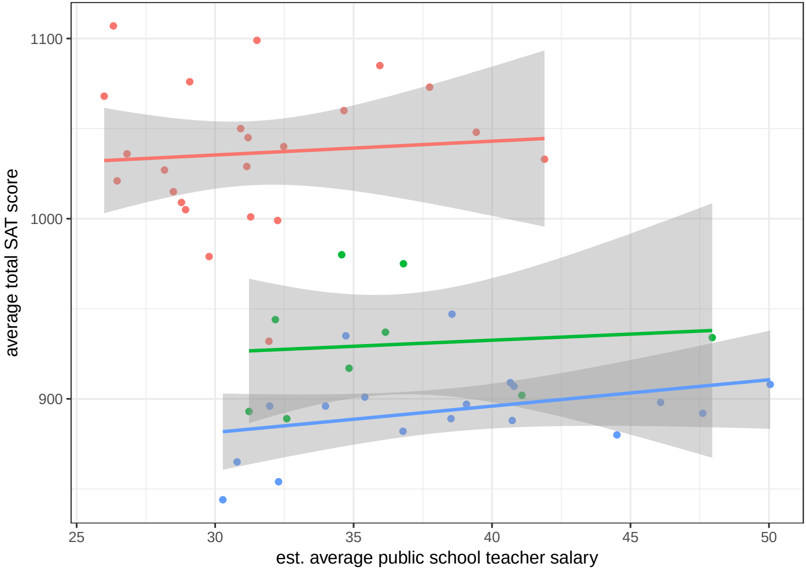
Sapiens: A Brief History of Humankind

Statistics brings data into focus

Statistics does not eliminate noise



Statistics speaks a language of uncertainty



Data are measurements from our imperfect, noisy world.

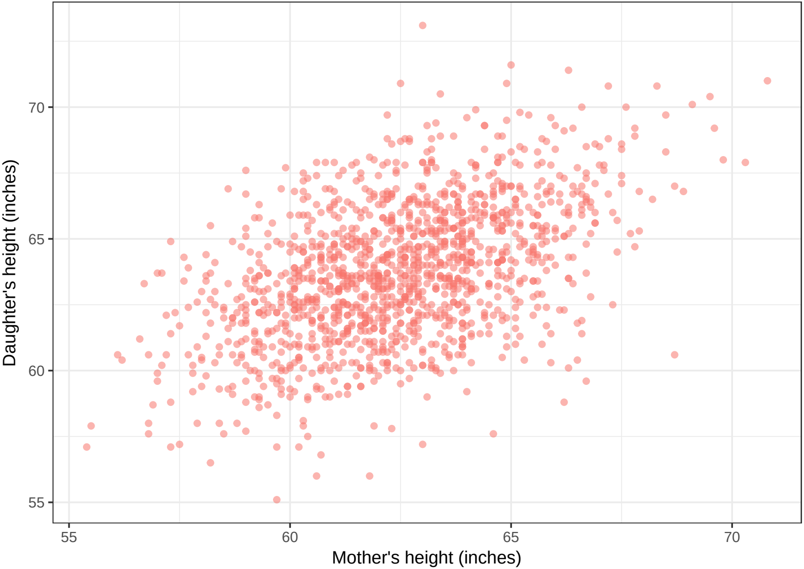
Key questions for any data analysis

What population do your cases represent?

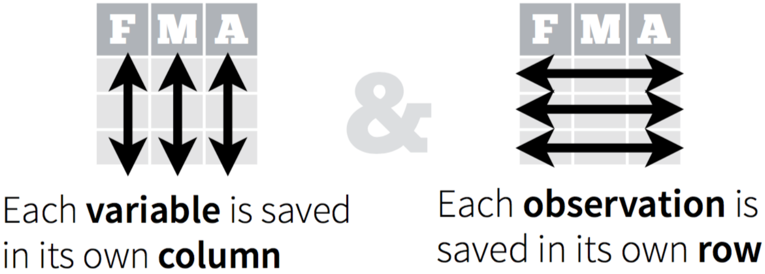
What variables do you have measurements on?

What are some sources of noise/variability?

Where does the noise come from?



Tidy Data

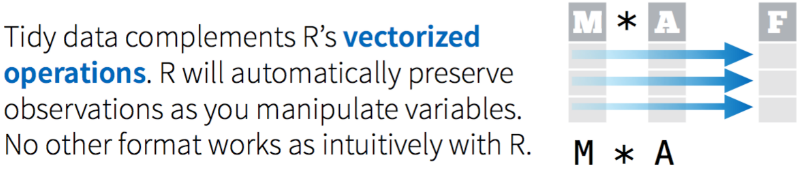


Each **variable** is saved in its own **column**

Each **observation** is saved in its own **row**

dplyr and tidyr cheatsheet

Tidy Data



Tidy data complements R’s **vectorized operations**. R will automatically preserve observations as you manipulate variables. No other format works as intuitively with R.

dplyr and tidyr cheatsheet

Sampling in a small population

```
groupA <- c("A", "A", "A", "A", "A", "A", "A")
groupB <- c("B", "B", "B")
population <- c(groupA, groupB)
sample(population, size = 5, replace=FALSE)

## [1] "A" "B" "A" "B" "B"

sample(population, size = 5, replace=FALSE)

## [1] "A" "B" "A" "A" "A"

sample(population, size = 5, replace=FALSE)

## [1] "B" "A" "A" "A" "A"
```

Sampling in a large population

```
groupA <- rep("A", 1000)
groupB <- rep("B", 500)
population <- c(groupA, groupB)
sample1 <- sample(population, size = 100,
                  replace=FALSE)
table(sample1)

## sample1
##  A  B
## 73 27
```

Sampling in a large population (with bias)

```
## with a biased sample
weights <- c(rep(1,1000), rep(3, 500))
sample2 <- sample(population, size = 100,
                  replace=FALSE, prob=weights)
table(sample2)

## sample2
##  A  B
## 42 58
```

Your project

- What types of variables are you collecting?
- Who are you collecting data on?
- What population are you trying to draw conclusions about?
- Do you expect your sample to be representative of the population? Why or why not?