On statistics, sampling, and data structures

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

The three challenges of statistical inference are:

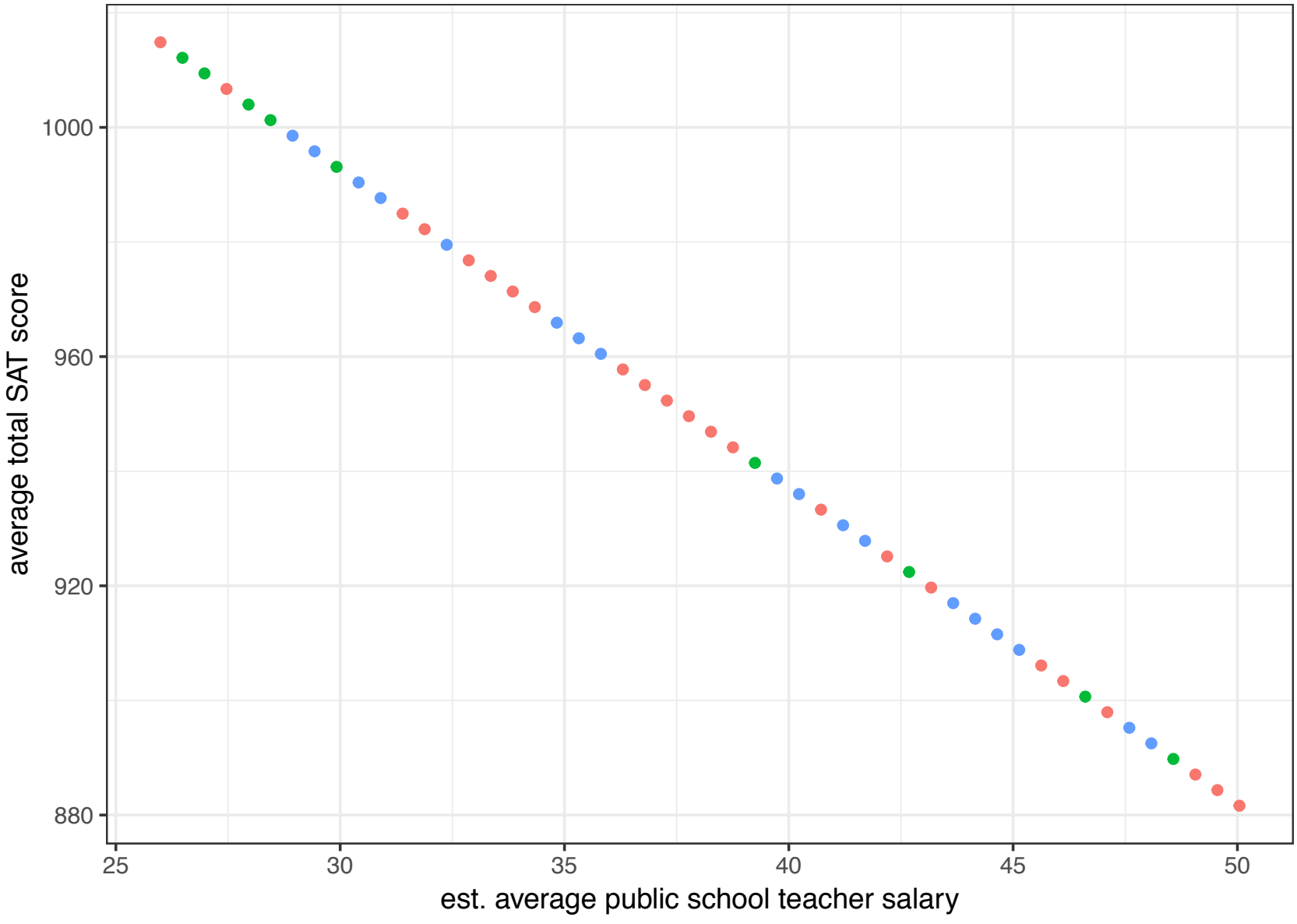
1. *Generalizing from sample to population* and from past to future, problems which are associated with survey sampling and forecasting but actually arise in nearly every application of statistical inference;
2. *Generalizing from control to treatment group*, a problem which is associated with causal inference, which is implicitly or explicitly part of the interpretation of most regressions we have seen; and
3. *Generalizing from observed measurements to the underlying constructs of interest*, as most of the time our data do not quite record exactly what we would ideally like to study.

All three of these challenges can be framed as problems of prediction (for new people or new items that are not in the sample, for future outcomes under different potentially assigned treatments, and for underlying constructs of interest, if they could be measured exactly).

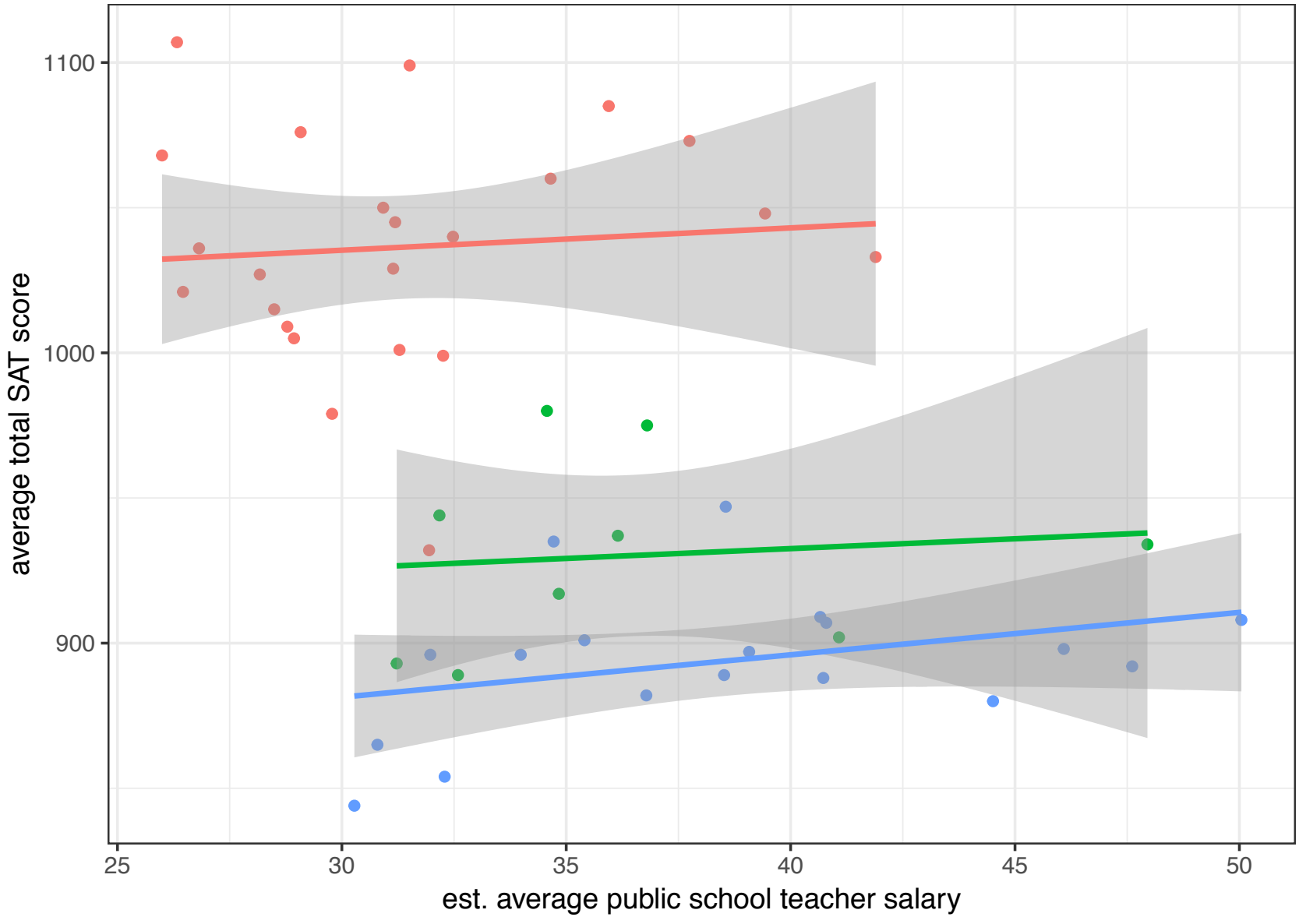
Andrew Gelman's blog

Statistics brings data into focus

Statistics does not eliminate noise



Statistics speaks a language of uncertainty



Language of uncertainty: real-world politics edition

Quotes from Nate Silver, Sunday Nov 6, 2016

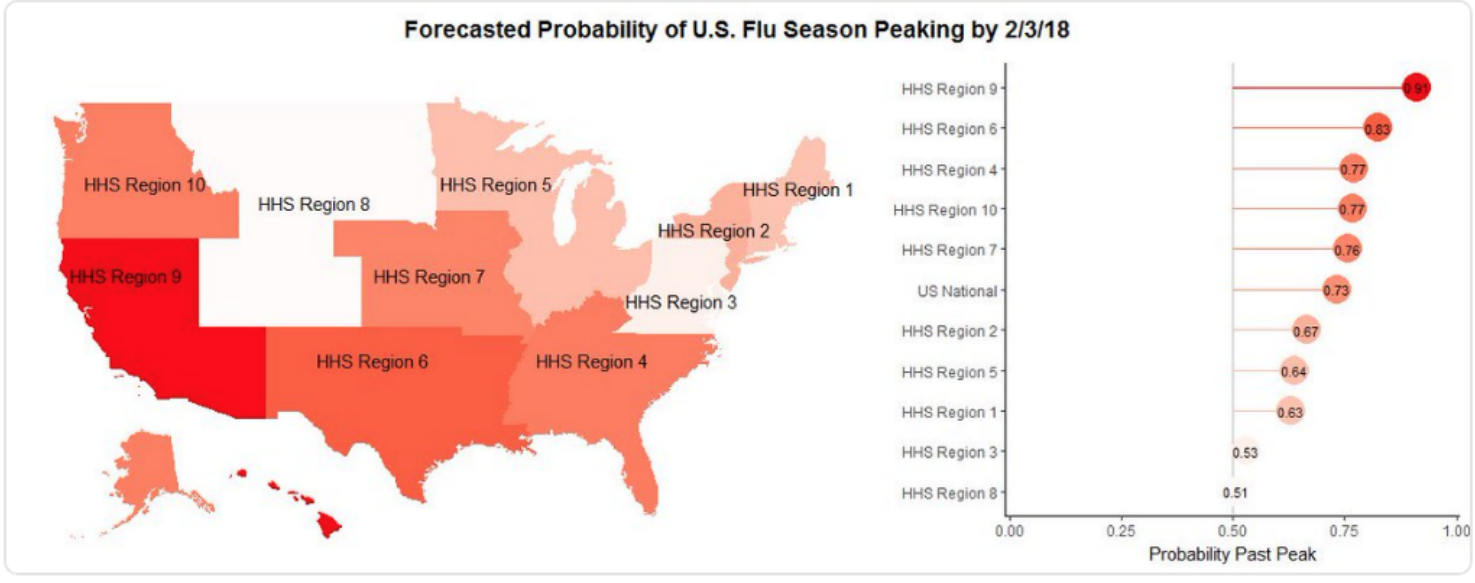
- "... it shouldnt be hard to see how Clinton could lose. She's up by about 3 percentage points nationally, and 3-point polling errors happen fairly often, including in the last two federal elections. Obama beat his polls by about 3 points in 2012, whereas Republicans beat their polls by 3 to 4 points in the 2014 midterms."
- "Right now, the tipping-point state in our forecast the state that would provide the decisive 270th electoral vote if the polls got things exactly right is New Hampshire. ... Clintons doing a little bit worse in the tipping-point state than she is overall a sign that she might win the popular vote but lose the Electoral College."
- "To be honest, Im kind of confused as to why people think its heretical for our model to give Trump a 1-in-3 chance which does make him a fairly significant underdog, after all. ... the public polls specifically including the highest-quality public polls show a tight race in which turnout and late-deciding voters will determine the difference between a clear Clinton win, a narrow Clinton win and Trump finding his way to 270 electoral votes."

FiveThirtyEight.com

Language of uncertainty: flu forecasting edition

Nicholas G. Reich @reichlab · Jan 30

Latest #FluSight forecasts still showing a fair amount of uncertainty over whether we have seen the #flu season peak in all regions of the US, especially mountain west and mid-Atlantic. Still quite a bit of flu ahead, even if the worst is past. flusightnetwork.io



2 3

Data are measurements from our imperfect, noisy world.

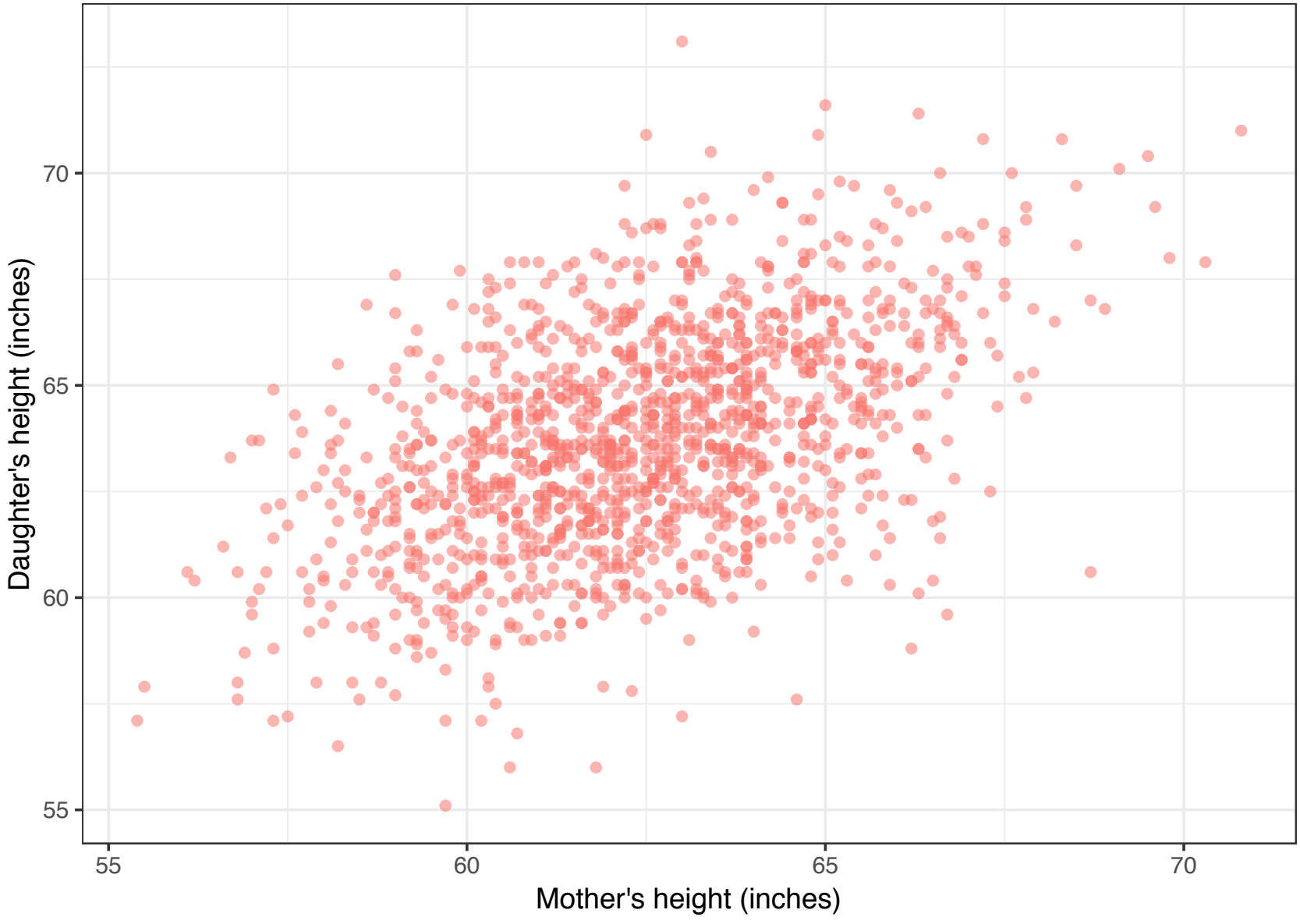
Key questions for any data analysis

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?

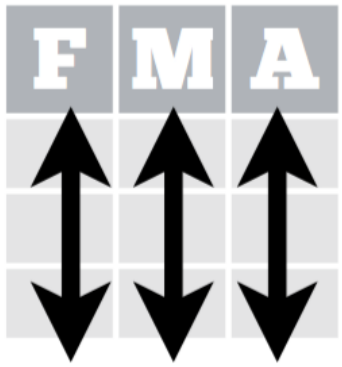


Tidy data

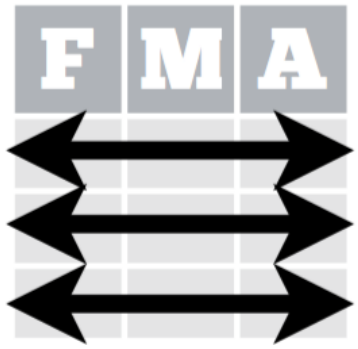
Tidy Data

- Tidy data is a very useful term. It was defined (although not conceived) by Hadley Wickham in this paper.
- It is not the only acceptable format for data (there are times when other formats, such as a 'wide-format' dataset may be needed), but it is a very common and widely used data structure.
- And, it plays very nicely with R.

Tidy Data



&



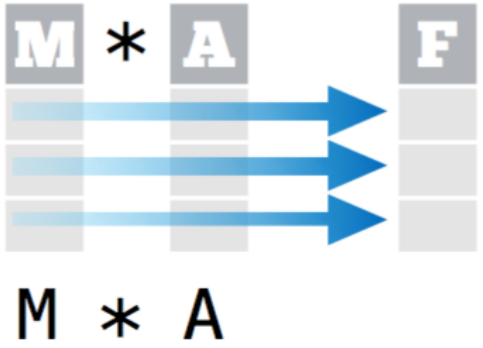
Each **variable** is saved in its own **column**

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

Understanding Sampling

Sample size is tied to how much info is in your data

Say we have a population of 1000 people where 600 have characteristic X and 400 do not. We want to estimate how many people have characteristic X.

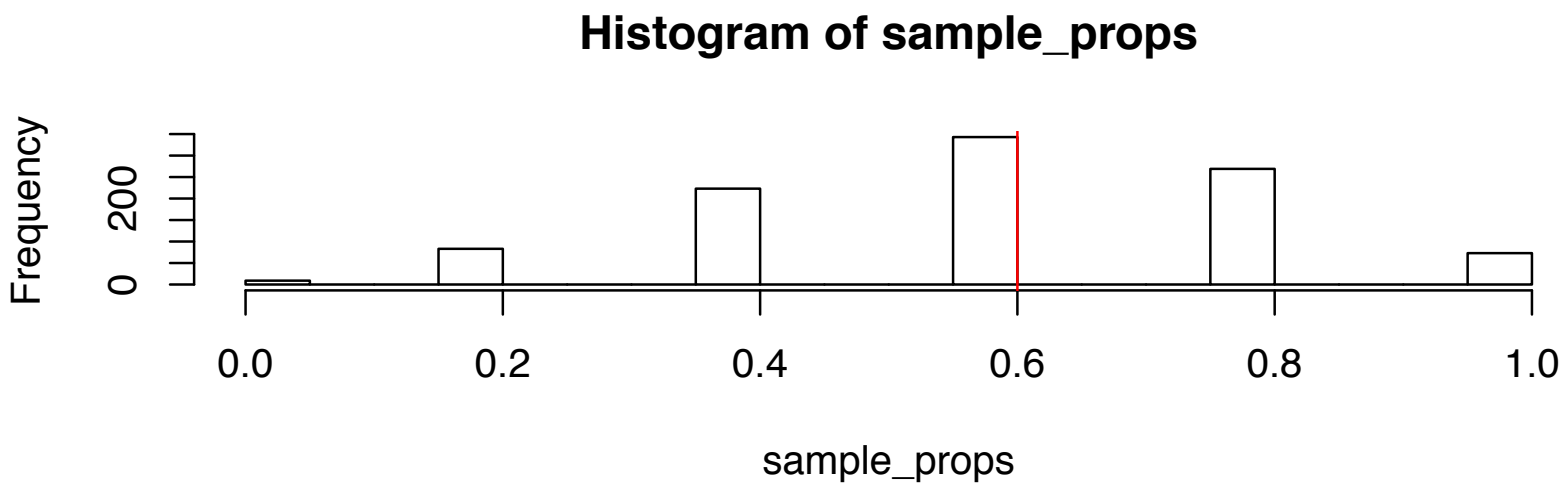
```
groupA <- rep("A", 600)
groupB <- rep("B", 400)
population <- c(groupA, groupB)
sample(population, size = 5, replace=FALSE)

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

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

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

How many people do you think we need to sample randomly to attain "adequate" precision on our estimate?

The size of your sample matters!

Results from hypothetically sampling 5 people from the population 1000 times...

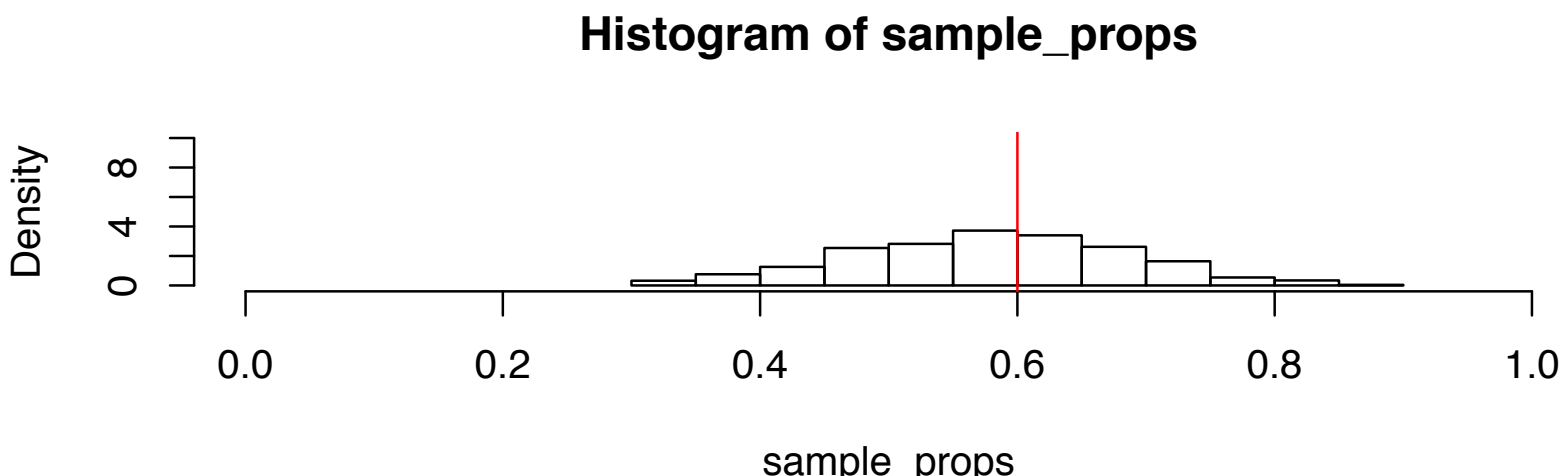
```
nsim <- 1000
sample_props <- rep(NA, nsim)
for(i in 1:nsim) {
  tmp <- sample(population, size = 5, replace=FALSE)
  sample_props[i] <- sum(tmp=="A")/5
}
hist(sample_props, breaks=20, xlim=c(0,1))
abline(v=3/5, col="red")
```



The size of your sample matters!

Results from hypothetically sampling 20 people from the population 1000 times...

```
nsim <- 1000
nsamp <- 20
sample_props <- rep(NA, nsim)
for(i in 1:nsim) {
  tmp <- sample(population, size = nsamp, replace=FALSE)
  sample_props[i] <- sum(tmp=="A")/nsamp
}
hist(sample_props, breaks=20, xlim=c(0,1), ylim=c(0,10), freq =
abline(v=3/5, col="red")
```



Beware of bias in your sampling!

What if people without characteristic X were 2 times as likely to be sampled than those with it?

```
weights <- c(rep(1,600), rep(2, 400))
sample2 <- sample(population, size = 100,
                  replace=FALSE, prob=weights)
table(sample2)

## sample2
##  A  B
## 43 57
```

In your groups: What fraction of your sample do you expect to have characteristic X?

Your project

- What types of variables are you collecting?
- What is the appropriate "tidy data" format for your data?
- 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?