# moderndive: statistical inference via the tidyverse

Albert Y. Kim

Jordan Moody

Ziwei “Crystal” Zang

Starry Zhou

ModernDive

USCOTS 2019
State College PA
May 15-16

Smith College
SDS
Statistical &
Data Sciences

# About us!

also my collaborators…

**Jenny Smetzer**

**Chester Ismay**

# About you!

Say hi to your nearest neighbors!
You’ll be learning together!

# My Context for moderndive

**My students:**

- Undergraduate-only liberal arts college
- Service intro stats course for all majors, all years
- Calculus is a pre-req only in name
- 13 weeks x (3 x 70min lectures + 75min lab)
- 29/40 had never coded in R prior

**My goals:**

- Goal 1: Sampling for inference
- Goal 2: Modeling with regression

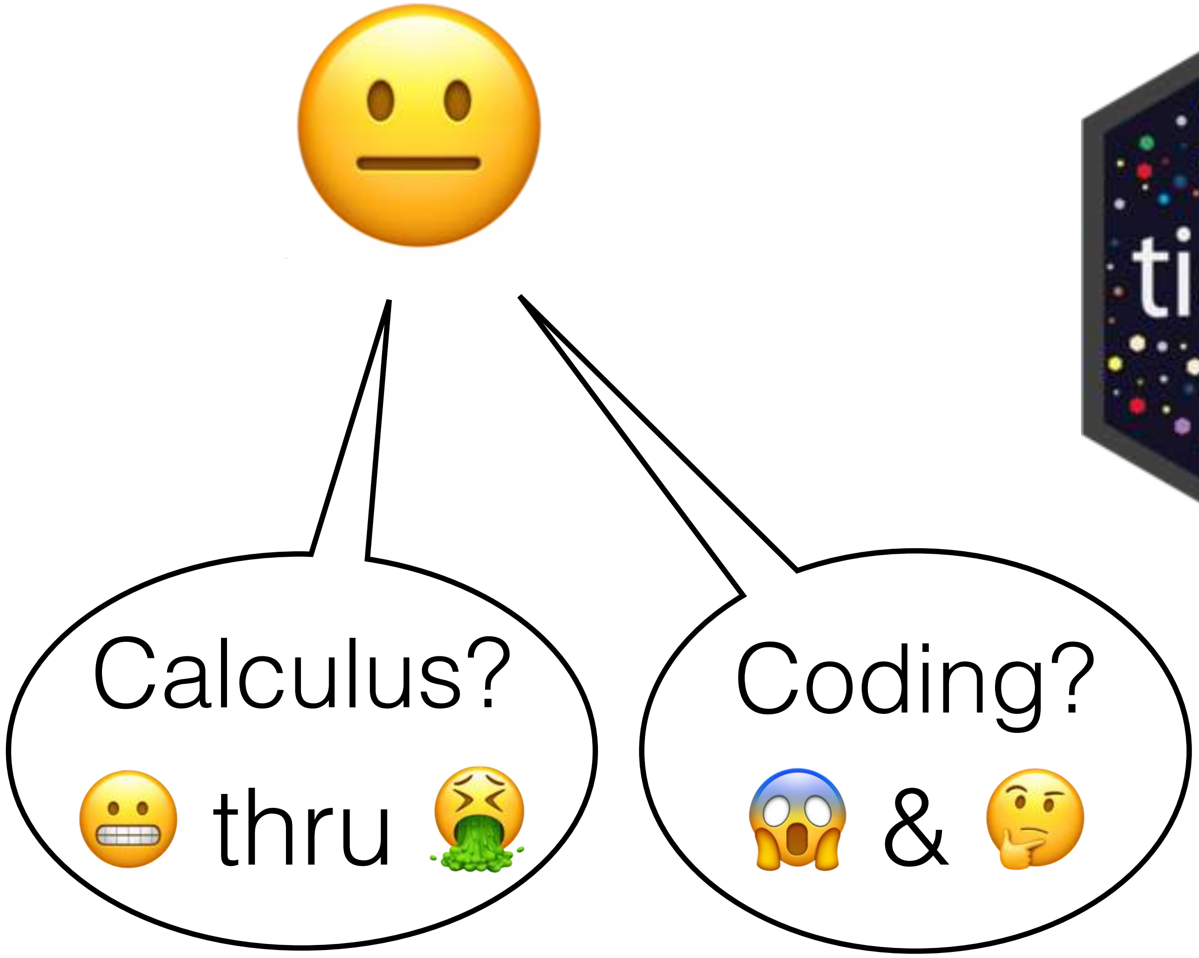
# Getting from Point A to Point B

Point A:
Modal 1st time
stats student

via the
`tidyverse`

→

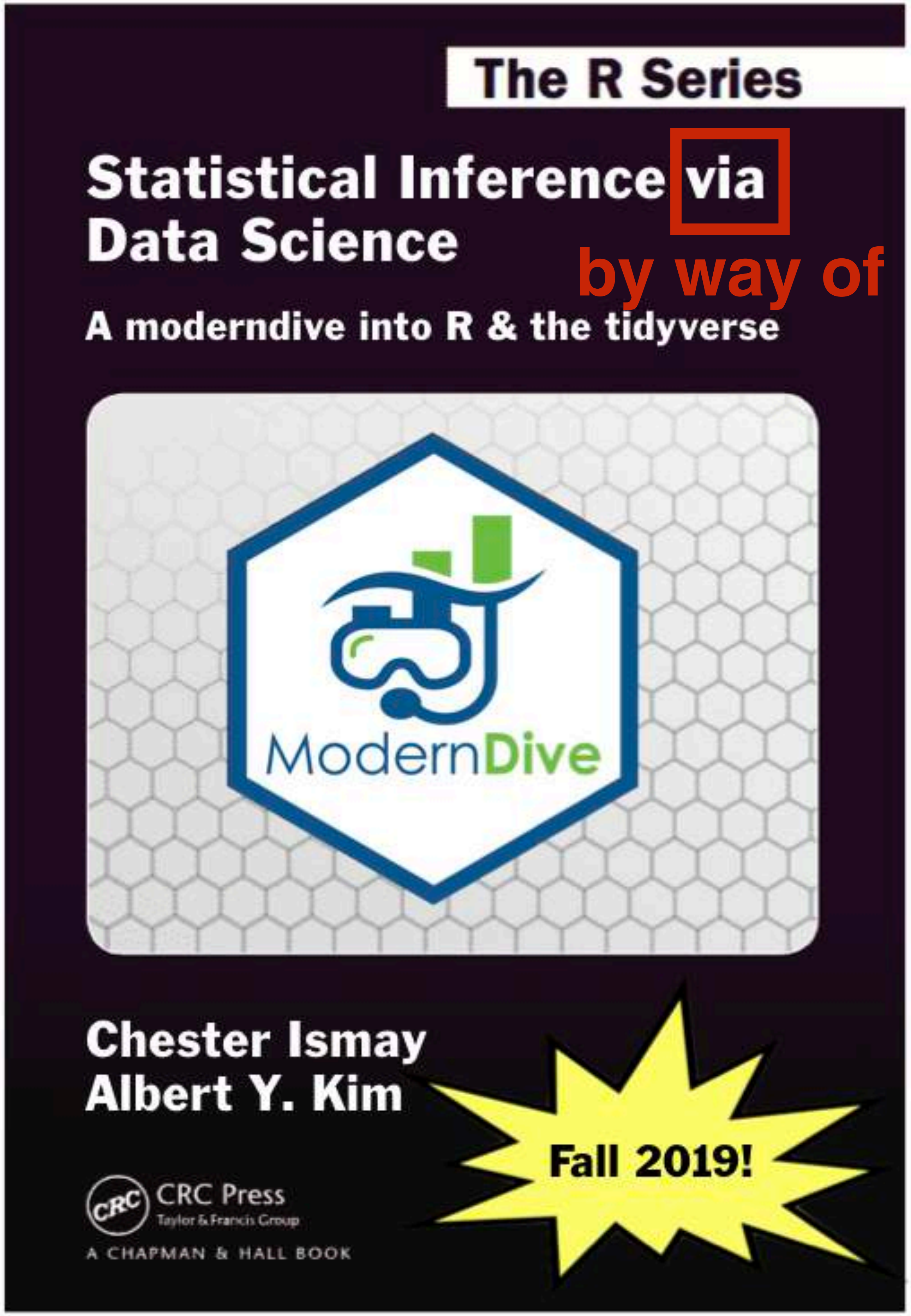
Point B:
Two goals



Calculus?
😬 thru 🤮

Coding?
😱 & 🤔

1. Sampling for inference
2. Modeling with regression



The R Series

## Statistical Inference via Data Science

by way of

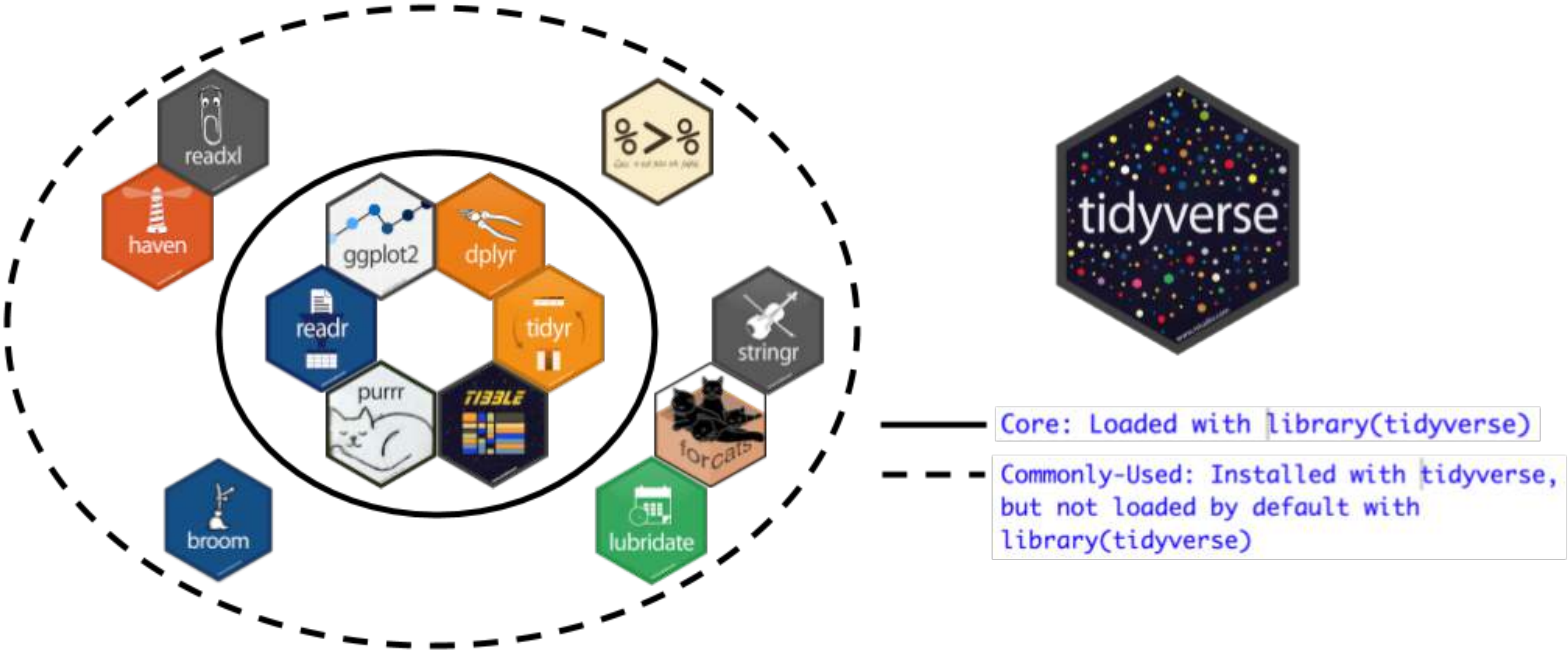
A moderndive into R & the tidyverse

Chester Ismay
Albert Y. Kim

Fall 2019!

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

# What is the tidyverse?



Core: Loaded with library(tidyverse)

Commonly-Used: Installed with tidyverse, but not loaded by default with library(tidyverse)

- ggplot2 for data visualization
- dplyr for data wrangling
- readr for data importing

# Why tidyverse? Some principles

From tidy tools manifesto:

1. Reuse existing data structures
2. Compose simple functions with the pipe
3. Embrace functional programming
4. Design for humans

Say what?

1. Don’t reinvent the wheel!
2. Break down tasks step-by-step!
3. What is the goal of your code?
4. Make code understandable

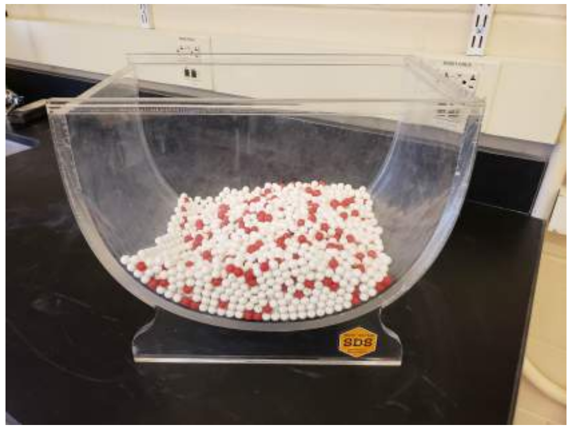
# Why tidyverse for stats newbies?

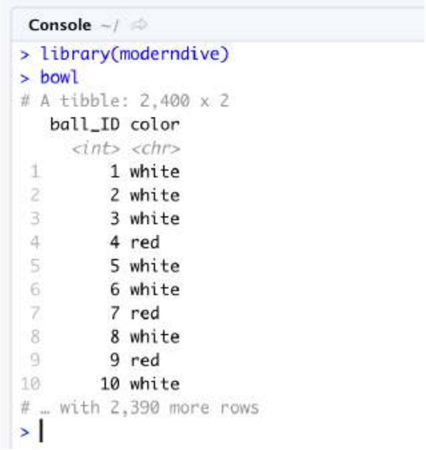
- *In my opinion* it’s easier to learn than base R. Others too.
- It scales. You leverage an entire ecosystem of online developers and support: Google & StackOverflow
- Satisfy learning goals *while learning tools they can use beyond the classroom*

# Goal 1: Sampling for Inference

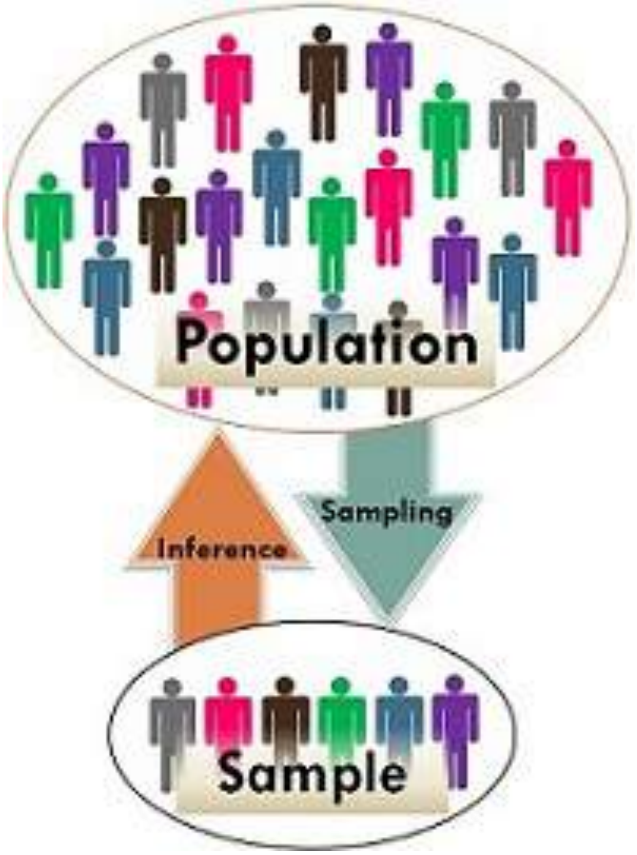
1. Tactile Sampling → 2. Virtual Sampling → 3. Theoretical

Population

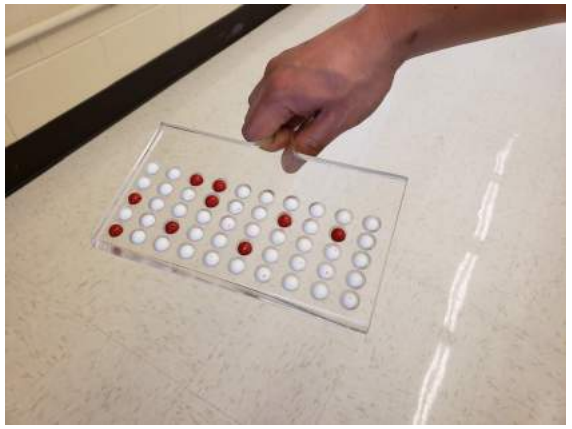


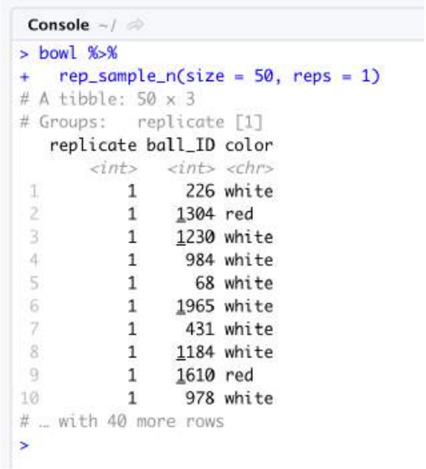


```
Console ~/
> library(moderndive)
> bowl
# A tibble: 2,400 x 2
   ball_ID color
     <int> <chr>
 1       1 white
 2       2 white
 3       3 white
 4       4 red
 5       5 white
 6       6 white
 7       7 red
 8       8 white
 9       9 red
10      10 white
# … with 2,390 more rows
> |
```



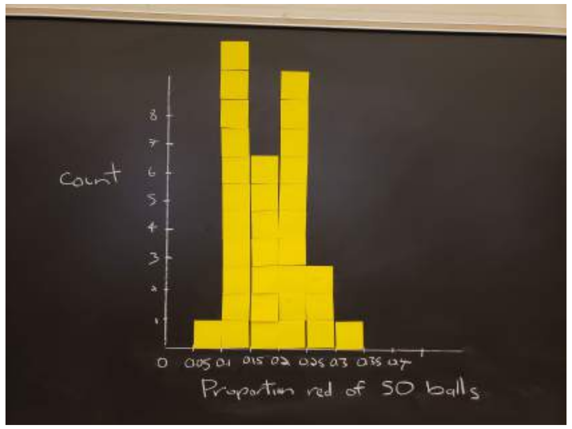
Sample

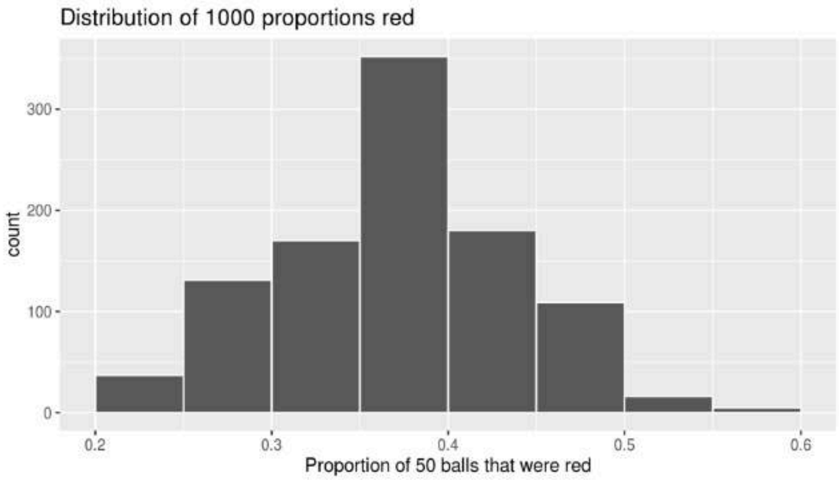


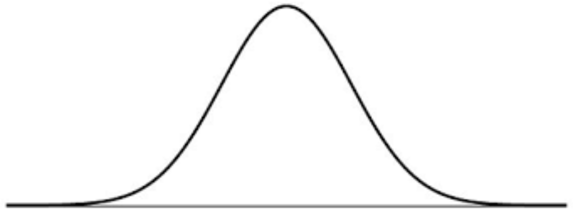


```
Console ~/
> bowl %>%
+   rep_sample_n(size = 50, reps = 1)
# A tibble: 50 x 3
# Groups:   replicate [1]
   replicate ball_ID color
       <int>   <int> <chr>
 1         1     226 white
 2         1    1304 red
 3         1    1230 white
 4         1     984 white
 5         1      68 white
 6         1    1965 white
 7         1     431 white
 8         1    1184 white
 9         1    1610 red
10         1     978 white
# … with 40 more rows
>
```

Sampling Distributions & Standard Errors



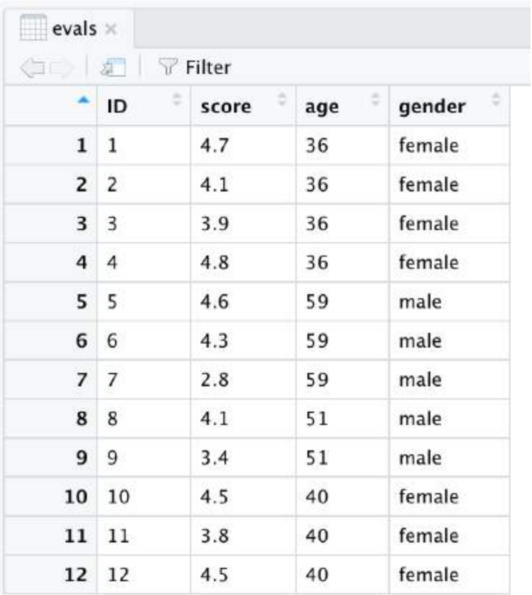




$$SE = \sqrt{\frac{p(1-p)}{n}}$$

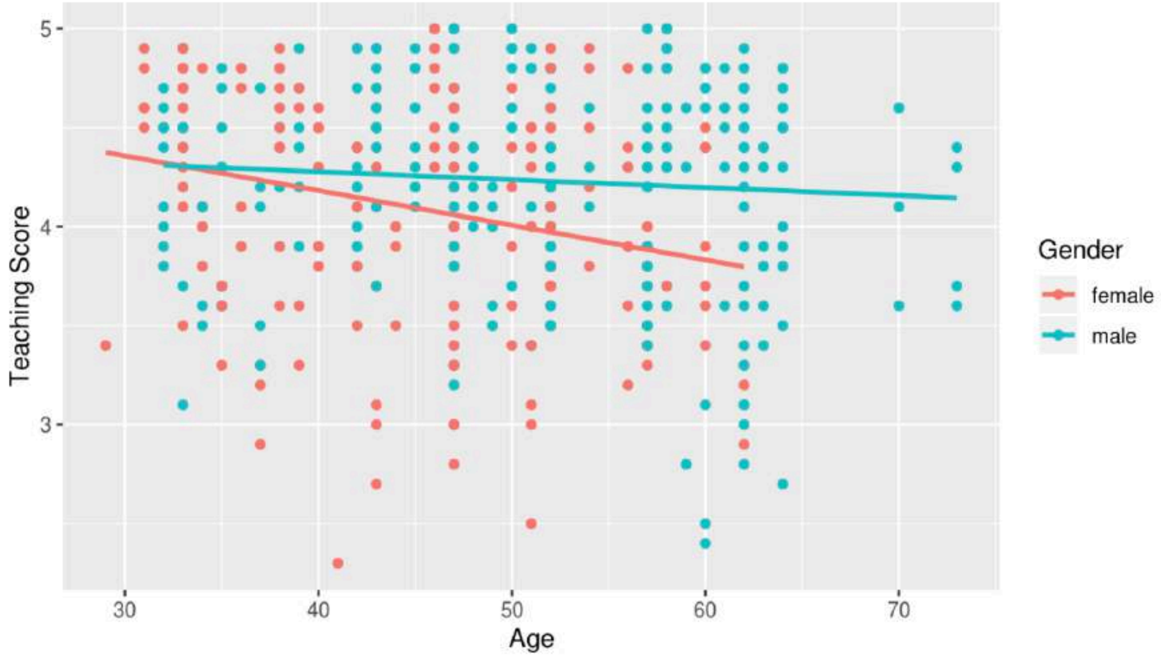
# Goal 2: Modeling with Regression

## 1. Data

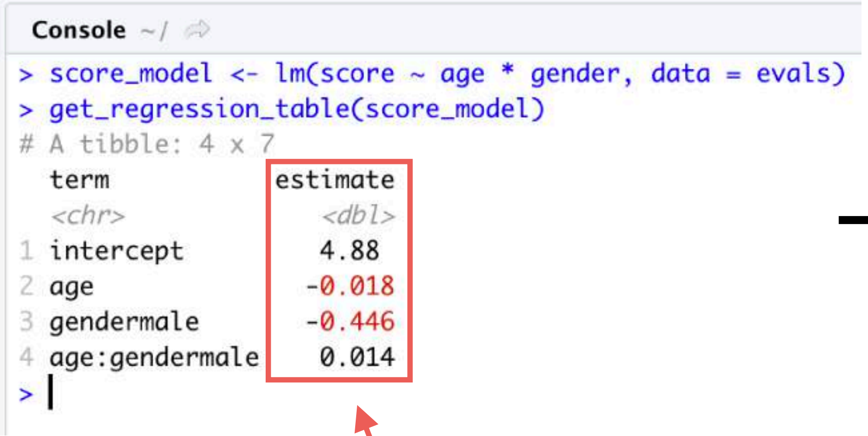


| | ID | score | age | gender |
|---|---|---|---|---|
| 1 | 1 | 4.7 | 36 | female |
| 2 | 2 | 4.1 | 36 | female |
| 3 | 3 | 3.9 | 36 | female |
| 4 | 4 | 4.8 | 36 | female |
| 5 | 5 | 4.6 | 59 | male |
| 6 | 6 | 4.3 | 59 | male |
| 7 | 7 | 2.8 | 59 | male |
| 8 | 8 | 4.1 | 51 | male |
| 9 | 9 | 3.4 | 51 | male |
| 10 | 10 | 4.5 | 40 | female |
| 11 | 11 | 3.8 | 40 | female |
| 12 | 12 | 4.5 | 40 | female |

## 2. Exploratory Data Analysis

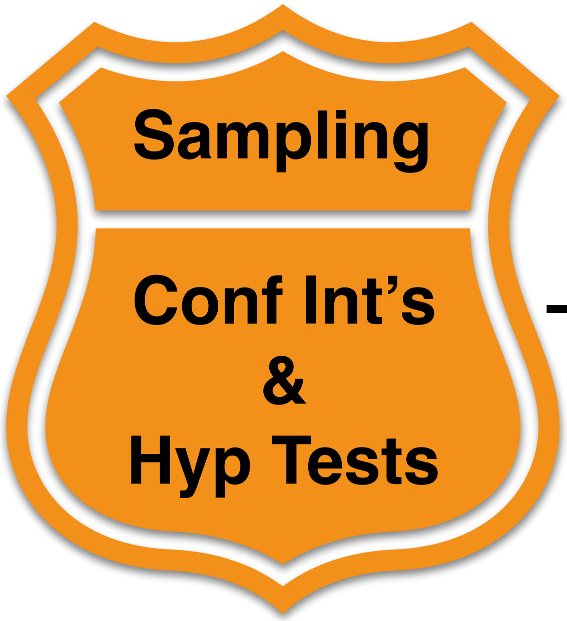


## 3. Regression Coeff



```
> score_model <- lm(score ~ age * gender, data = evals)
> get_regression_table(score_model)
# A tibble: 4 x 7
  term           estimate
  <chr>             <dbl>
1 intercept          4.88
2 age              -0.018
3 gendermale       -0.446
4 age:gendermale    0.014
> |
```

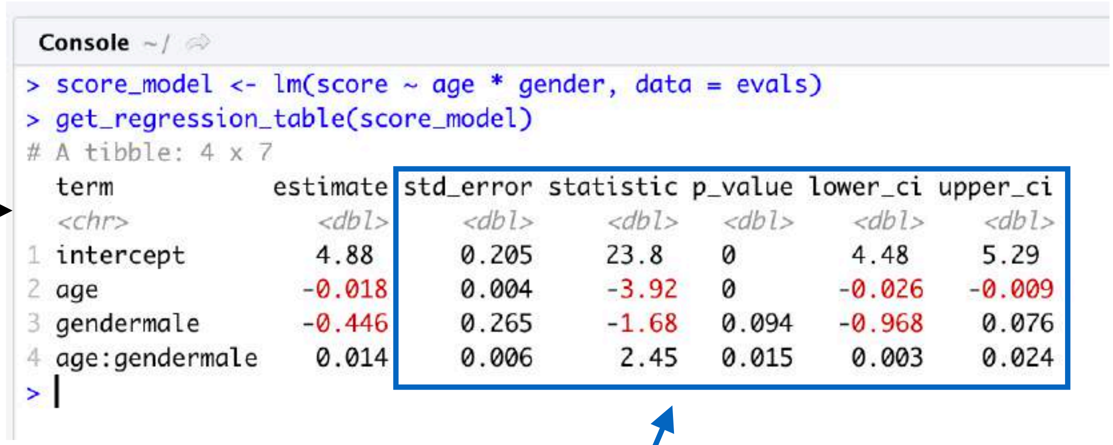
Early: Descriptive regression



Sampling

Conf Int's
&
Hyp Tests

## 4. Regression Table

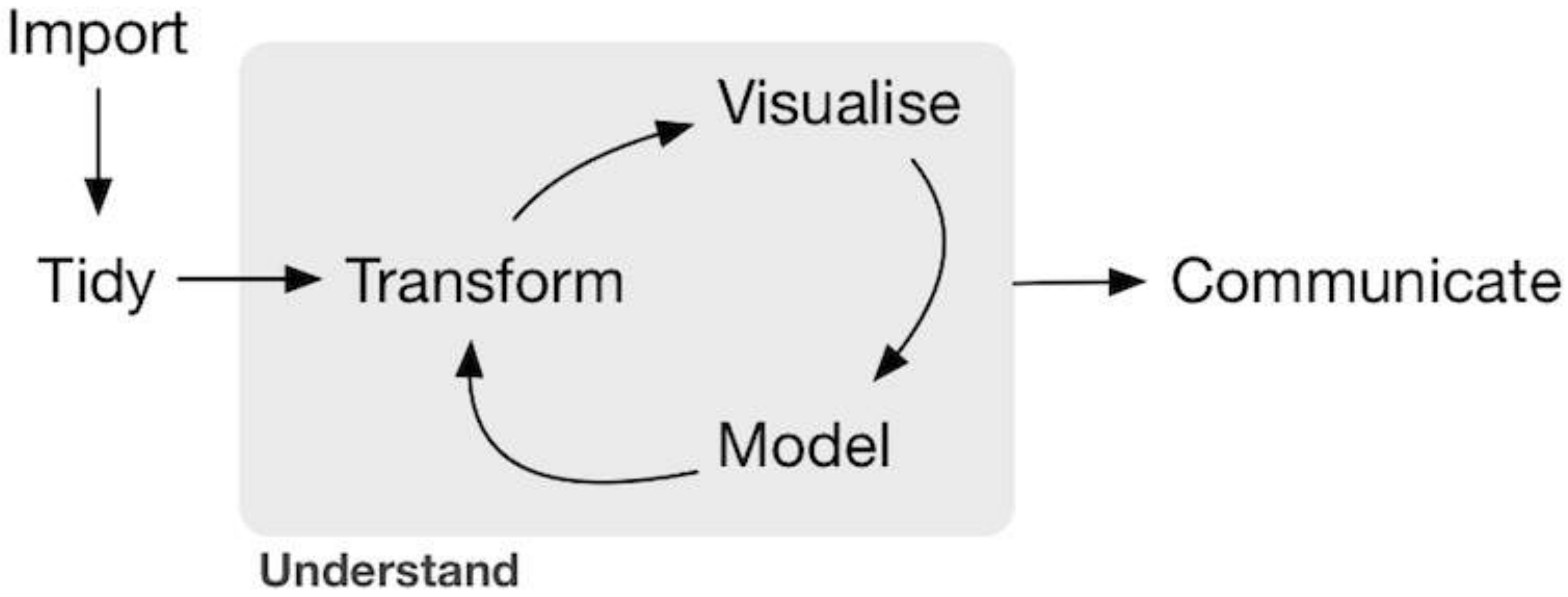


```
> score_model <- lm(score ~ age * gender, data = evals)
> get_regression_table(score_model)
# A tibble: 4 x 7
  term           estimate std_error statistic p_value lower_ci upper_ci
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept          4.88     0.205     23.8    0         4.48     5.29
2 age              -0.018     0.004     -3.92   0        -0.026   -0.009
3 gendermale       -0.446     0.265     -1.68   0.094    -0.968    0.076
4 age:gendermale    0.014     0.006      2.45   0.015     0.003    0.024
> |
```

Later: Inference for Regression

# End Deliverable

Final project that “plays the whole game”
of data/science pipeline:



Example template given to students this semester, based on work by Alexis Cohen, Andrianne Dao, & Isabel Gomez last semester.

# Schedule

See Google Doc available at bit.ly/USCOTS2019

# Keep in mind throughout…

- You are not currently you, but you are currently your students *as best as you can imagine*.
- In other words, these exercises are meant for your students!
- Ultimately where do I start? *Start small!*

Questions?

Let's Go!

# Activity: Your Birthday

# For Every Chapter…

Slides on:

1. What are we doing ❓
2. Why are we doing this 🤔
3. Opinions
4. Potential pitfalls ⚠️

Followed by activities:

1. Chalk talk, pen/paper, or tactile exercise
2. Replicating exercise on computer
3. Exercise
4. Discussion

# Chapter 2: Exploring data

1. What are we doing ❓
    - Getting used to workspace via data exploration
2. Why are we doing this 🤔
    - Getting them over initial 😱 of coding
3. Our opinions
    - Stress importance of looking at RAW data values. Removing these layers of abstraction.
4. Potential pitfalls ⚠️
    - Difference of R vs RStudio
    - Installing/loading packages
    - Error messages, warning messages, regular messages
    - Coding: both student self doubt & lowered instructor expectations

# Chapter 3: Visualizing Data

1. What are we doing ❓
    - Creating (colored) scatterplots, histograms, boxplots
2. Why are we doing this 🤔
    - Equip students with tools for both our goals
    - Exploratory data analysis!!!
3. Our opinions
    - Viewing all graphics through lens of the Grammar of Graphics (via `ggplot2`)
4. Potential pitfalls ⚠️
    - Histograms & boxplots involve transformations of raw values
    - Coding ramps up: Reassure students! Encourage them to not code from scratch, rather copy/paste/tweak

# Chapter 4: Data Wrangling

1. What are we doing ❓
   - Learning the pipe operator **%>%**
   - Wrangling/transforming data
2. Why are we doing this 🤔
   - Equip students with tools for both our goals
3. Our opinions
   - To *completely* shield students from *any* data wrangling is to betray true nature of work in our fields
4. Potential pitfalls ⚠️
   - How much wrangling should you require vs you curate yourself?

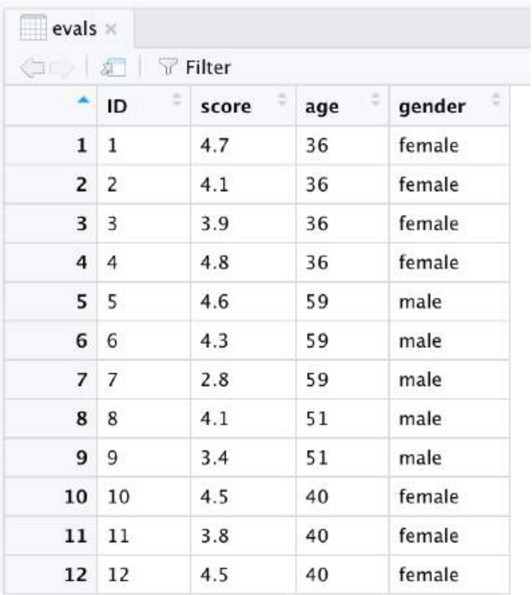
# Chapter 5: “Tidy” data

1. What are we doing ❓
   - tidyverse gets its name from fact that all inputs/outputs are assumed to be *tidy data frames*
   - Importing data via `readr::read_csv()`
2. Why are we doing this 🤔
   - Students have their own Excel/Google Sheets data
   - Will have to convert from wide to tidy/long format
3. Our opinions
   - This chapter can be skipped if
     A. You only provide tidy/long data
     B. You have your students publish .csv files to Google Sheets
4. Potential pitfalls ⚠️
   - Working directories!

# Chapter 6: Simple regression

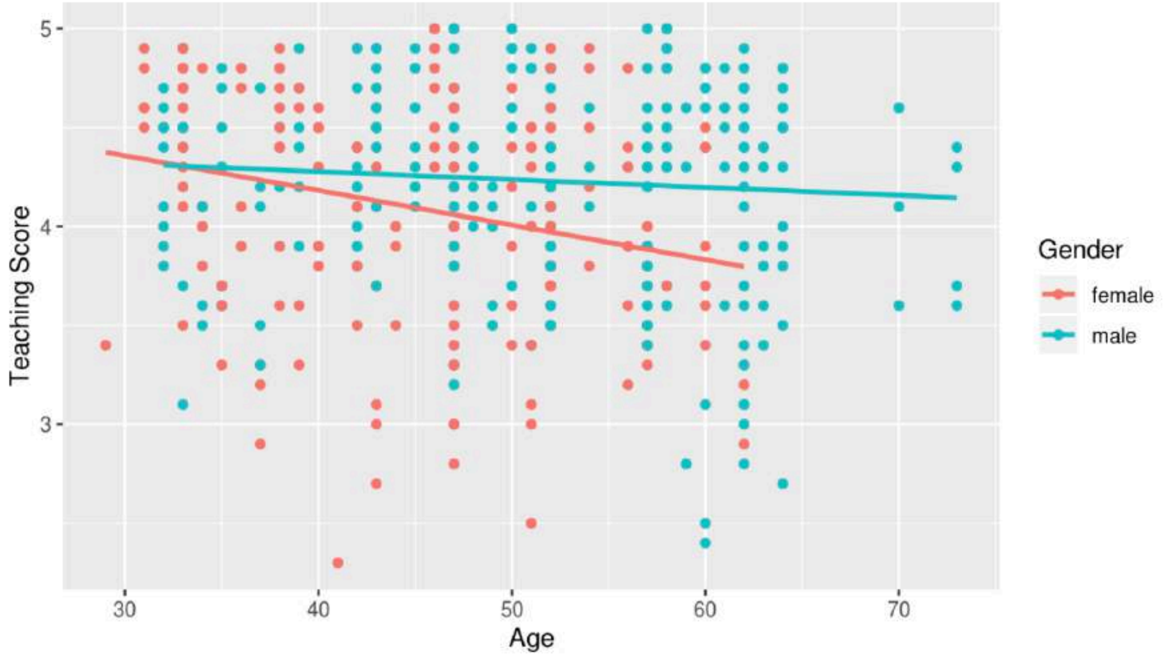
# Goal 2: Modeling with Regression

## 1. Data

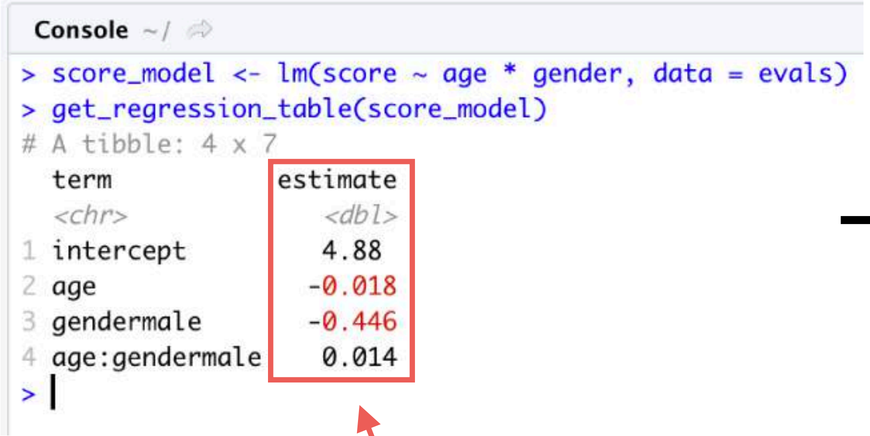


|  | ID | score | age | gender |
|---|---|---|---|---|
| 1 | 1 | 4.7 | 36 | female |
| 2 | 2 | 4.1 | 36 | female |
| 3 | 3 | 3.9 | 36 | female |
| 4 | 4 | 4.8 | 36 | female |
| 5 | 5 | 4.6 | 59 | male |
| 6 | 6 | 4.3 | 59 | male |
| 7 | 7 | 2.8 | 59 | male |
| 8 | 8 | 4.1 | 51 | male |
| 9 | 9 | 3.4 | 51 | male |
| 10 | 10 | 4.5 | 40 | female |
| 11 | 11 | 3.8 | 40 | female |
| 12 | 12 | 4.5 | 40 | female |

## 2. Exploratory Data Analysis



## 3. Regression Coeff



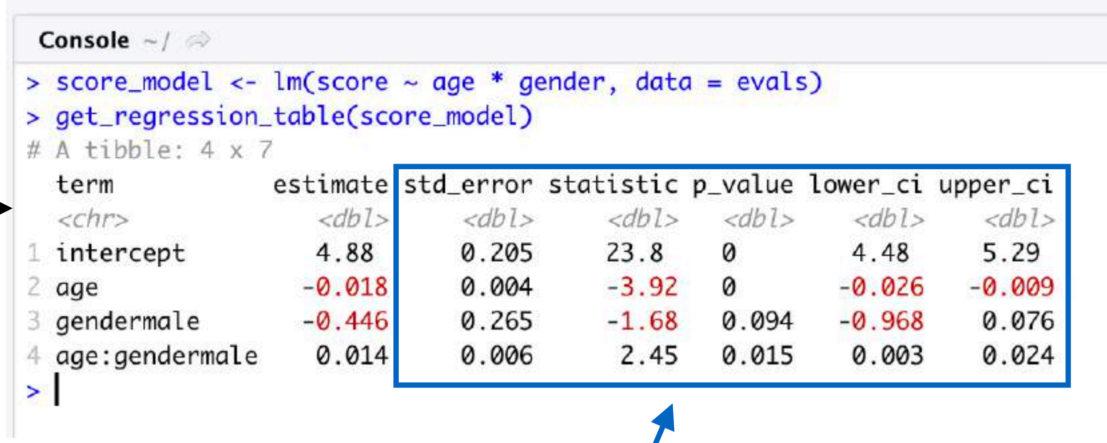
```
> score_model <- lm(score ~ age * gender, data = evals)
> get_regression_table(score_model)
# A tibble: 4 x 7
  term           estimate
  <chr>             <dbl>
1 intercept          4.88
2 age              -0.018
3 gendermale       -0.446
4 age:gendermale    0.014
> |
```

Early: Descriptive regression

Sampling

Conf Int's
&
Hyp Tests

## 4. Regression Table



```
> score_model <- lm(score ~ age * gender, data = evals)
> get_regression_table(score_model)
# A tibble: 4 x 7
  term           estimate std_error statistic p_value lower_ci upper_ci
  <chr>             <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept          4.88     0.205      23.8   0         4.48     5.29
2 age              -0.018     0.004     -3.92   0       -0.026   -0.009
3 gendermale       -0.446     0.265     -1.68   0.094   -0.968    0.076
4 age:gendermale    0.014     0.006      2.45   0.015    0.003    0.024
> |
```

Later: Inference for Regression

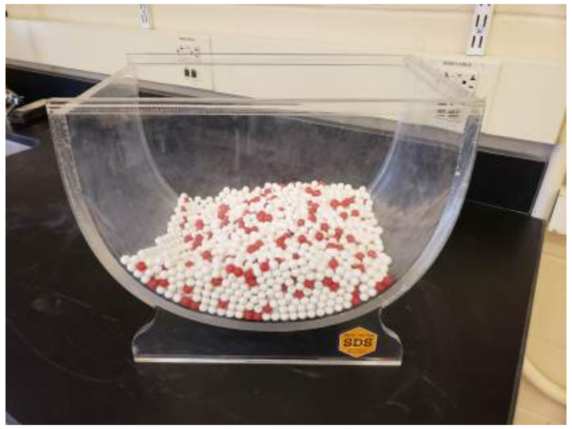
1. What are we doing ❓
   - *Descriptive* simple linear regression & regression with single categorical x only.
2. Why are we doing this 🤔
   - Multivariate thinking per GAISE guidelines & modeling
3. Our opinions
   - Separate descriptive vs inference so we can introduce it early, not at end of term 🥵
   - `moderndive::get_regression_table()` function has CI's, no p-value 🌟s
   - Much of world's data is categorical, to skip is to do students a disservice
   - Introduce causal inference
4. Potential pitfalls ⚠️
   - Understanding regression with categorical x

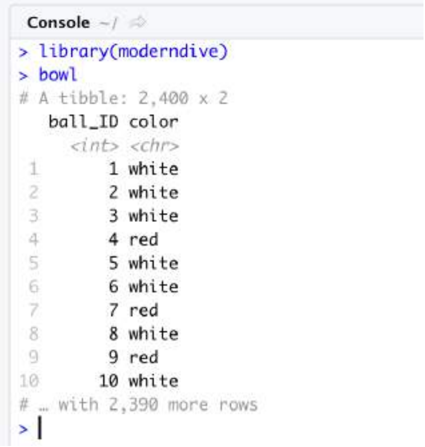
# Chapter 7: Multiple regression

# Goal 1: Sampling for Inference

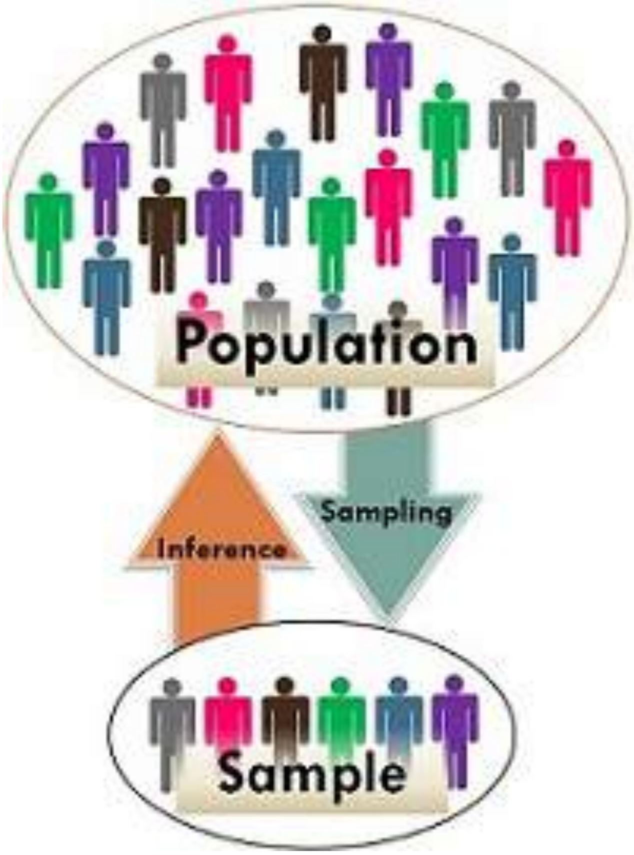
1. Tactile Sampling → 2. Virtual Sampling → 3. Theoretical

Population

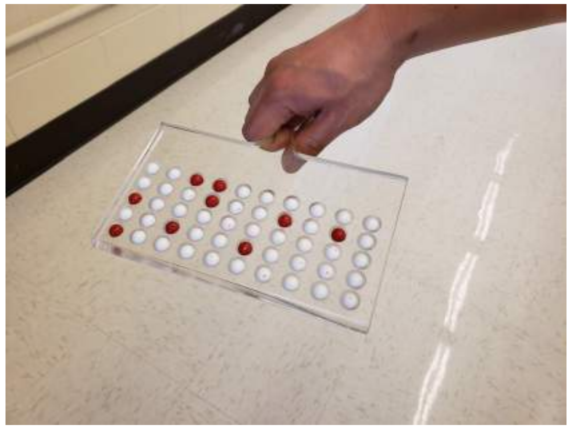


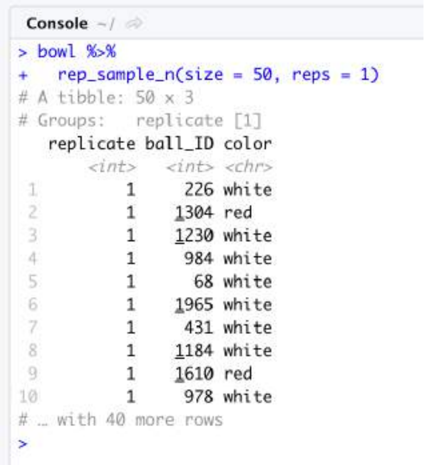


```
Console ~/
> library(moderndive)
> bowl
# A tibble: 2,400 x 2
   ball_ID color
     <int> <chr>
 1       1 white
 2       2 white
 3       3 white
 4       4 red
 5       5 white
 6       6 white
 7       7 red
 8       8 white
 9       9 red
10      10 white
# … with 2,390 more rows
> |
```



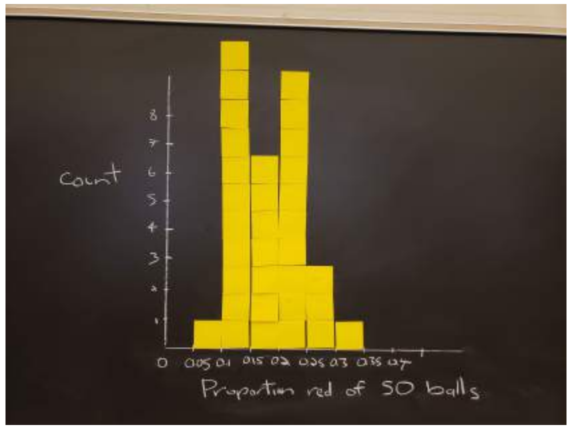
Sample

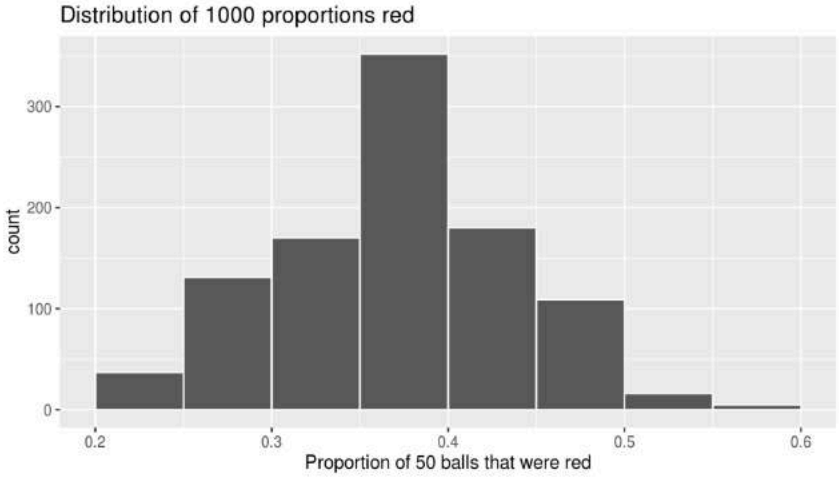


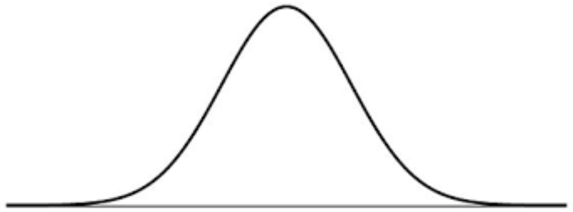


```
Console ~/
> bowl %>%
+   rep_sample_n(size = 50, reps = 1)
# A tibble: 50 x 3
# Groups:   replicate [1]
   replicate ball_ID color
       <int>   <int> <chr>
 1         1     226 white
 2         1    1304 red
 3         1    1230 white
 4         1     984 white
 5         1      68 white
 6         1    1965 white
 7         1     431 white
 8         1    1184 white
 9         1    1610 red
10         1     978 white
# … with 40 more rows
>
```

Sampling Distributions & Standard Errors

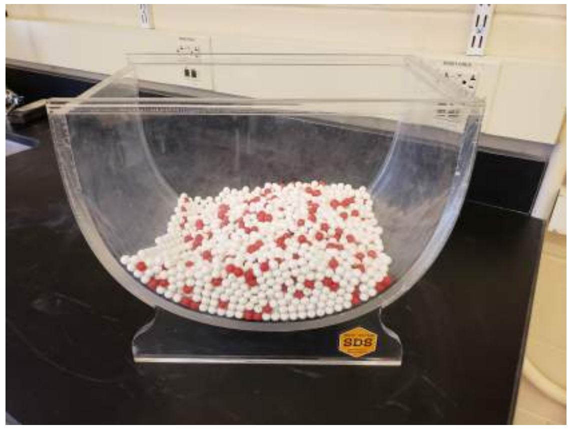


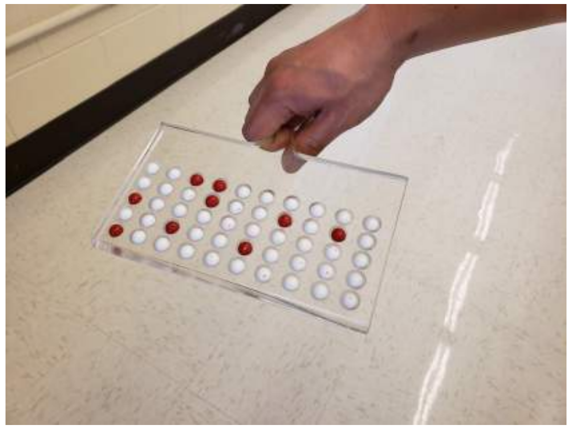


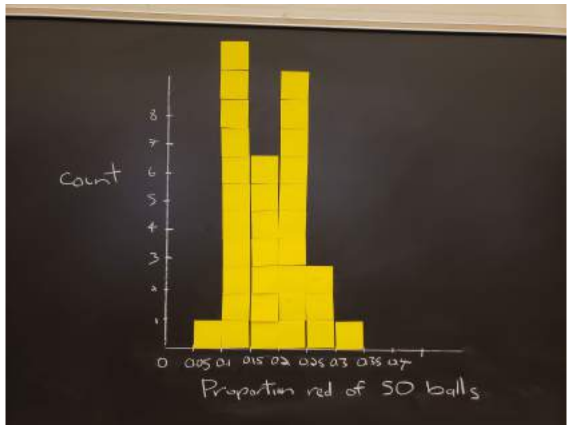


$$SE = \sqrt{\frac{p(1-p)}{n}}$$

# Statistical Inference

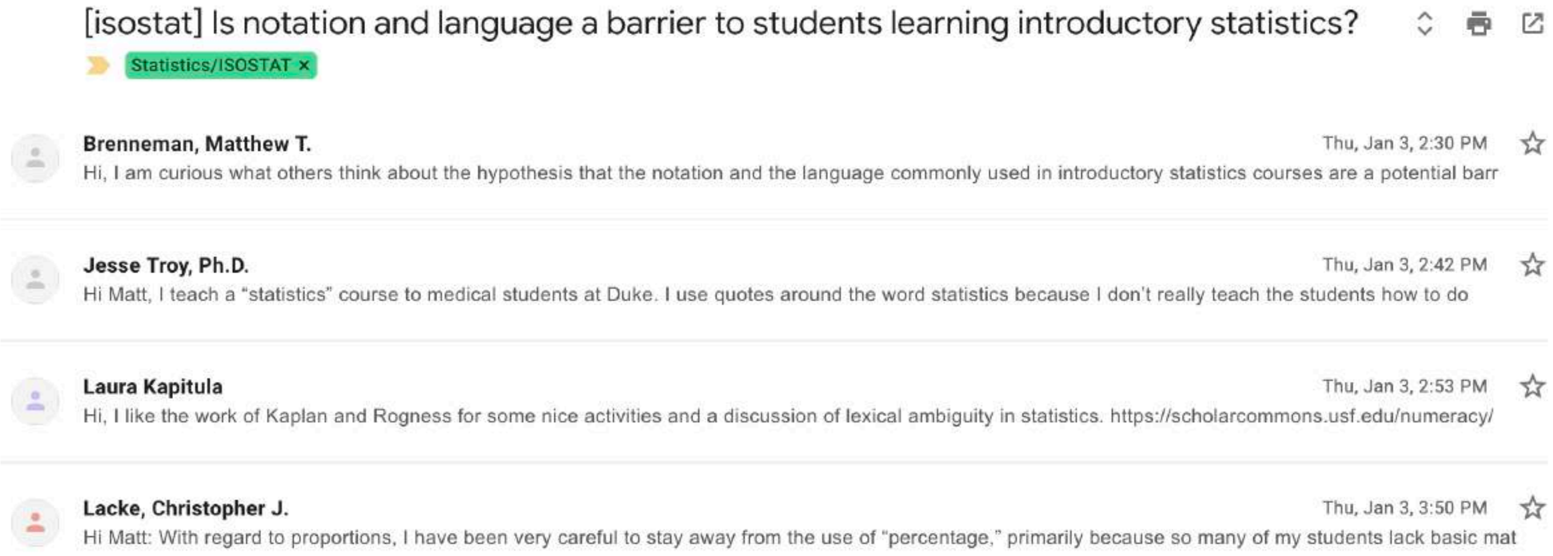






Issues:

1. Easily lost in many layers of abstraction
2. 😱 Notation, terminology, & definitions 😱



[isostat] Is notation and language a barrier to students learning introductory statistics?

Statistics/ISOSTAT ×

**Brenneman, Matthew T.** Thu, Jan 3, 2:30 PM
Hi, I am curious what others think about the hypothesis that the notation and the language commonly used in introductory statistics courses are a potential barr

**Jesse Troy, Ph.D.** Thu, Jan 3, 2:42 PM
Hi Matt, I teach a "statistics" course to medical students at Duke. I use quotes around the word statistics because I don't really teach the students how to do

**Laura Kapitula** Thu, Jan 3, 2:53 PM
Hi, I like the work of Kaplan and Rogness for some nice activities and a discussion of lexical ambiguity in statistics. https://scholarcommons.usf.edu/numeracy/

**Lacke, Christopher J.** Thu, Jan 3, 3:50 PM
Hi Matt: With regard to proportions, I have been very careful to stay away from the use of "percentage," primarily because so many of my students lack basic mat

# Hypothesis Testing

# Transference & Generalizability

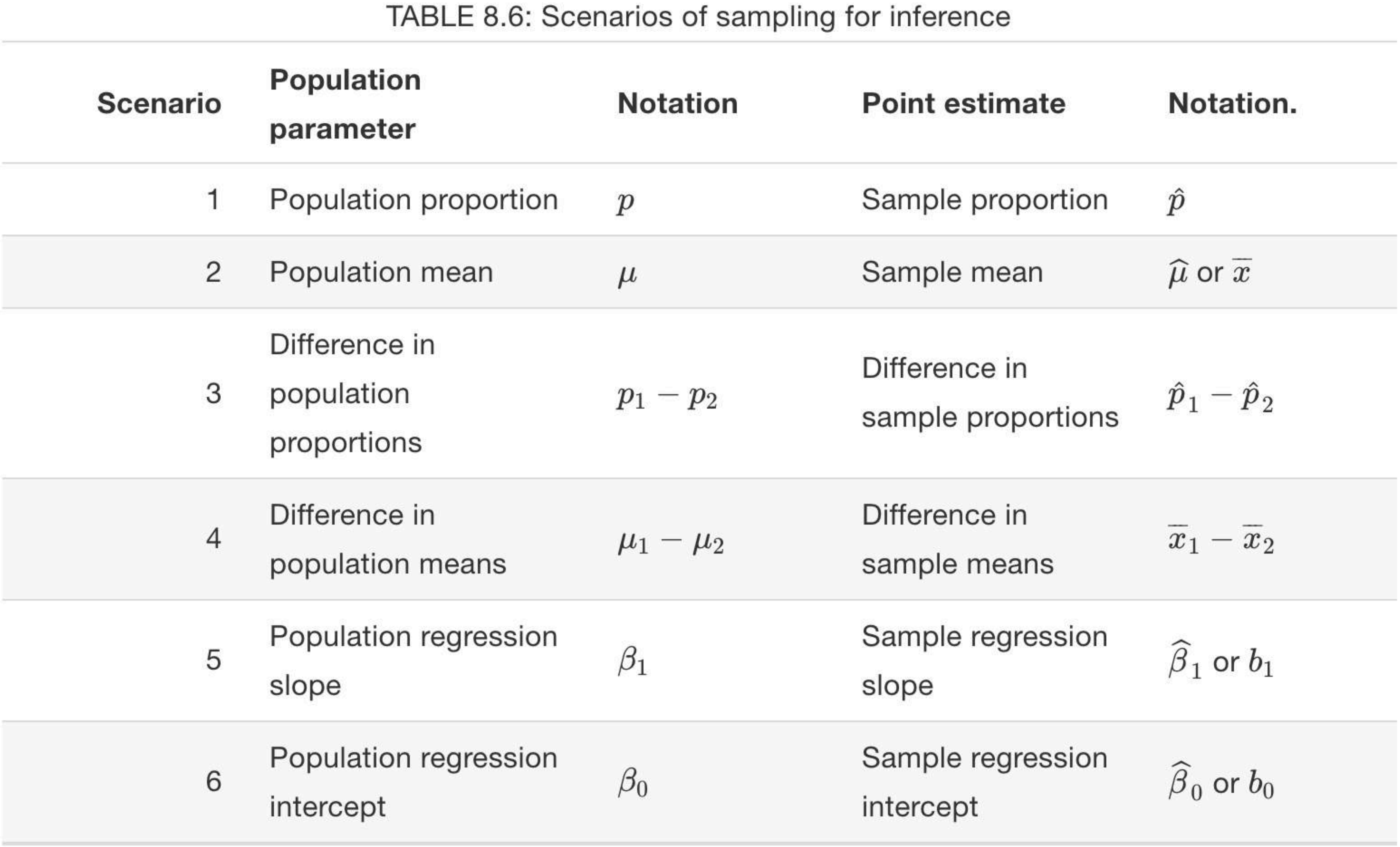
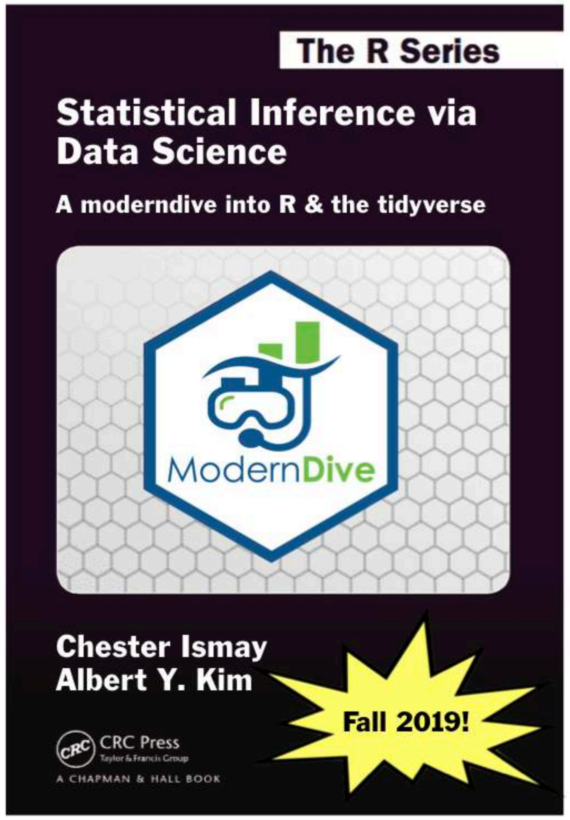


TABLE 8.6: Scenarios of sampling for inference

| Scenario | Population parameter | Notation | Point estimate | Notation. |
|---|---|---|---|---|
| 1 | Population proportion | $p$ | Sample proportion | $\hat{p}$ |
| 2 | Population mean | $\mu$ | Sample mean | $\widehat{\mu}$ or $\overline{x}$ |
| 3 | Difference in population proportions | $p_1 - p_2$ | Difference in sample proportions | $\hat{p}_1 - \hat{p}_2$ |
| 4 | Difference in population means | $\mu_1 - \mu_2$ | Difference in sample means | $\overline{x}_1 - \overline{x}_2$ |
| 5 | Population regression slope | $\beta_1$ | Sample regression slope | $\widehat{\beta}_1$ or $b_1$ |
| 6 | Population regression intercept | $\beta_0$ | Sample regression intercept | $\widehat{\beta}_0$ or $b_0$ |

# Timeline

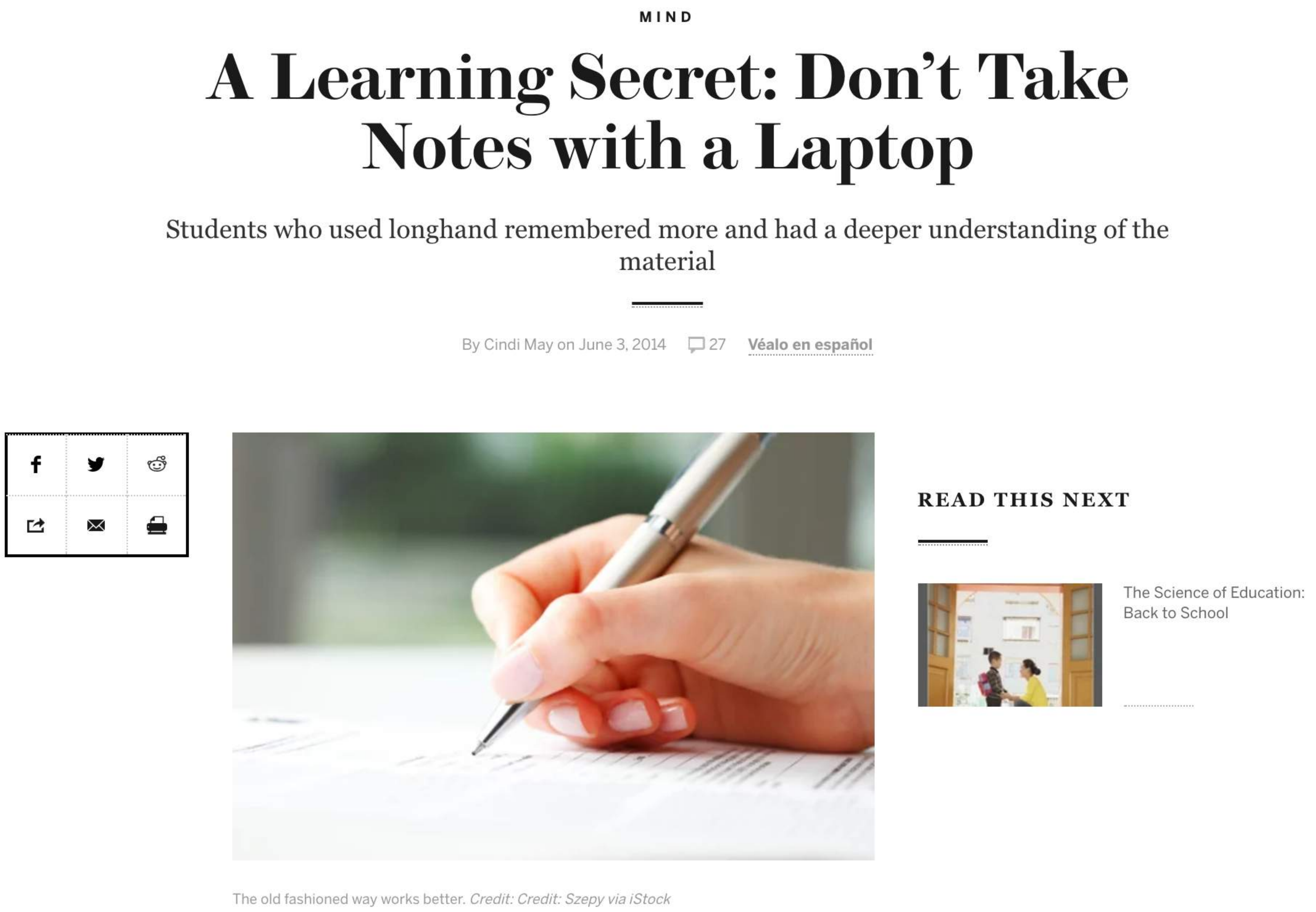
- **Now:** Follow development version "under construction" at moderndive.netlify.com
- **Mid-June**: Preview of print edition available on moderndive.com
- **Late-July**: Posting labs and example final project samples
- **Fall 2019**: Print edition available!



# Tips!

- Start small
- Teaching/learning code is a psychological battle
- You will run into problems

# The debate continues…



MIND

## A Learning Secret: Don’t Take Notes with a Laptop

Students who used longhand remembered more and had a deeper understanding of the material

By Cindi May on June 3, 2014 27 Véalo en español

The old fashioned way works better. *Credit: Credit: Szepy via iStock*

# Coding

Cobb (2015) argued there are two possible computational engines for statistics:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_{\overline{X}_1 - \overline{X}_2}} = \frac{\overline{X}_1 - \overline{X}_2}{S_{\overline{X}_1 - \overline{X}_2}}$$

$$S_{\overline{X}_1 - \overline{X}_2} = \sqrt{\frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}\left[\frac{1}{N_1} + \frac{1}{N_2}\right]}$$

```
a = replaceAll(",", " ", a); a = a.
return a.split(" "); } $("#unique").
var a = array_from_string($("#fin").
).val(), c = use_unique(array_
).val())); if (c < 2 * b - 1) {
c), this.trigger("click"); }
!= a[b] && " " != a[b] ||
"User_logged").val(); c = array_
c.length;b++) { -1 != a.indexOf(
; for (b = 0;b < c.length;
$("#User_logged").val(
.click(function(
```

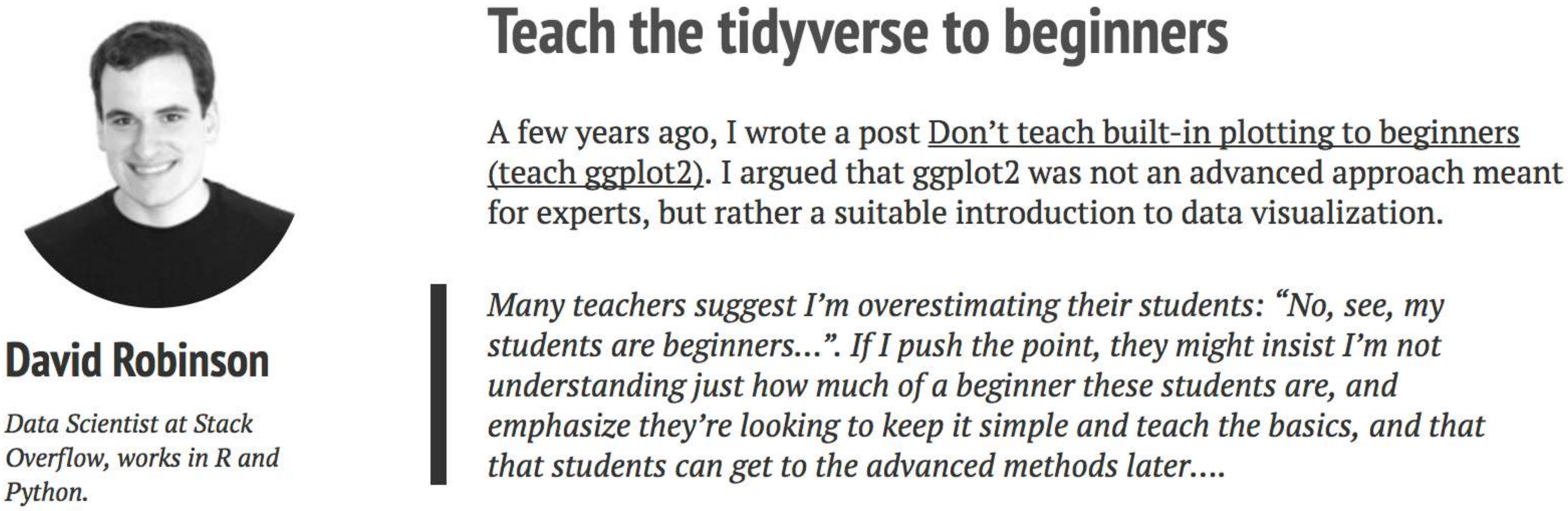
# Teaching/Learning Code

- Learn how a practitioner would learn:
  the “Copy/paste/tweak approach”
- Borrow elements of “flipped classroom”: how to use time we’re all in the same room together?

# Teaching Coding: The Battle is Psychological

- “Don’t code from scratch, take the copy/paste/tweak approach!”
- “Computers are stupid!”
- “Learning to code is similar to learning a language”

# New Tools Specific for Data Science



## Teach the tidyverse to beginners

**David Robinson**

*Data Scientist at Stack Overflow, works in R and Python.*

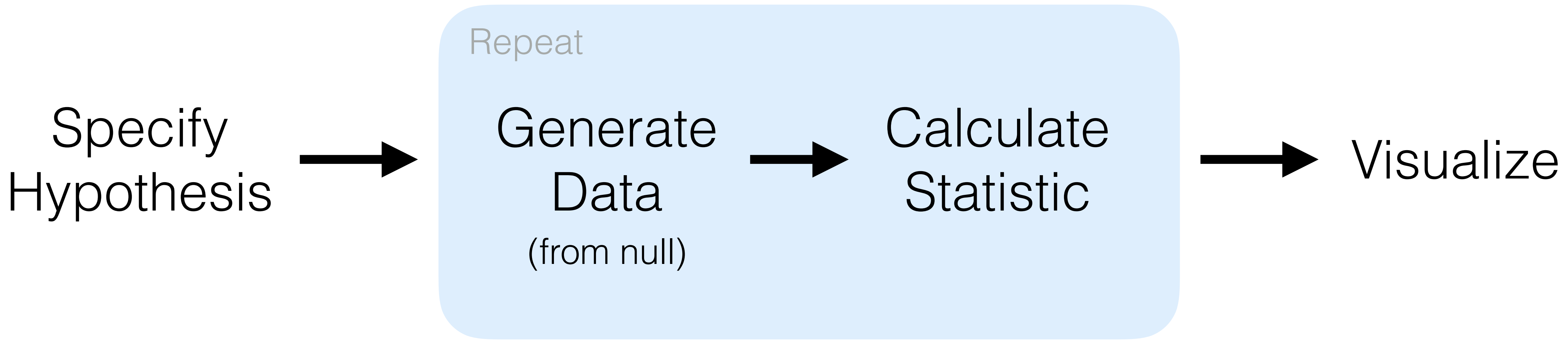
A few years ago, I wrote a post Don’t teach built-in plotting to beginners (teach ggplot2). I argued that ggplot2 was not an advanced approach meant for experts, but rather a suitable introduction to data visualization.

> *Many teachers suggest I’m overestimating their students: “No, see, my students are beginners…”. If I push the point, they might insist I’m not understanding just how much of a beginner these students are, and emphasize they’re looking to keep it simple and teach the basics, and that that students can get to the advanced methods later….*

DataCamp

# `infer` package for tidy statistical inference

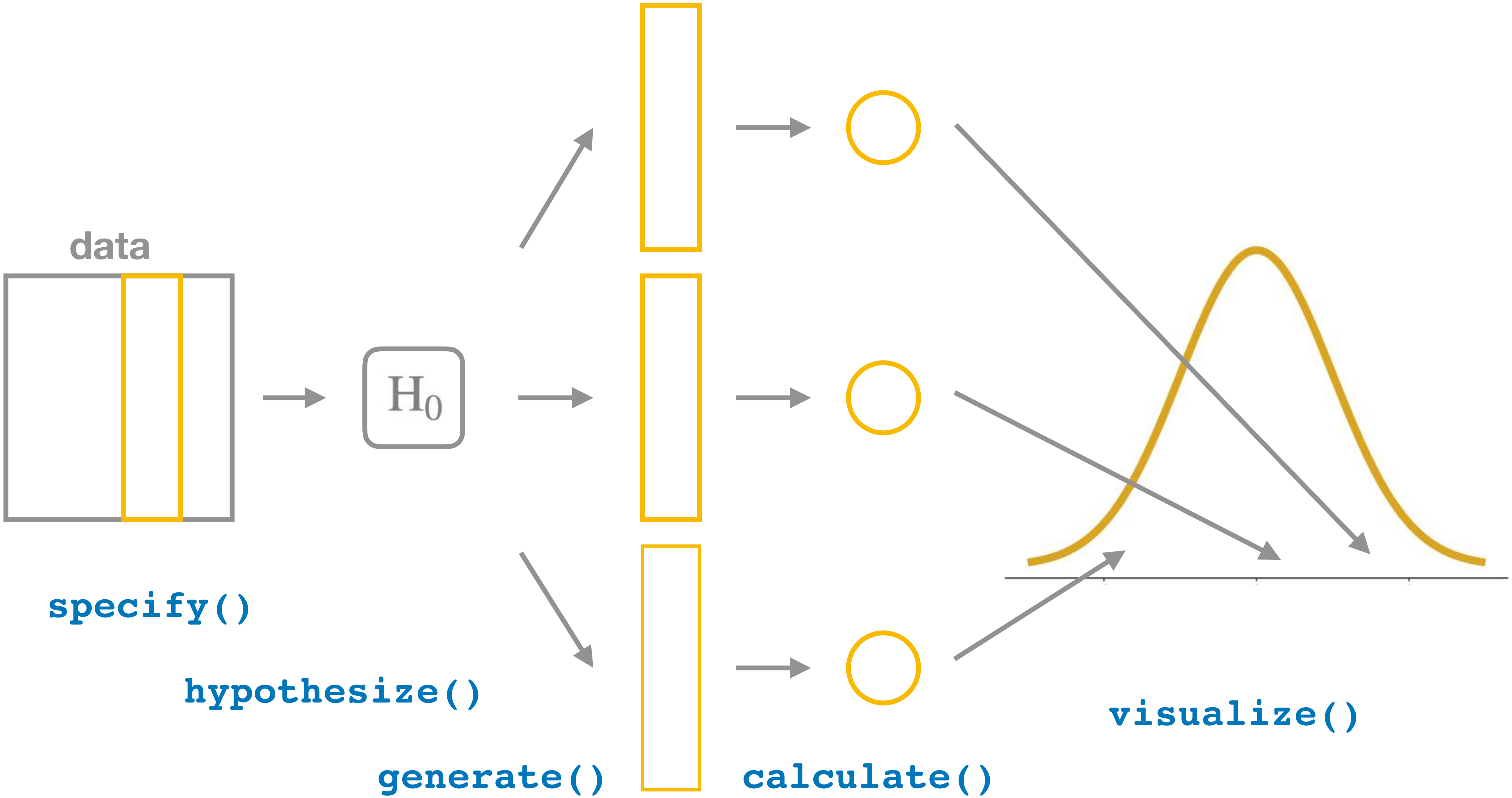
[http://infer.netlify.com/](http://infer.netlify.com/)



Specify Hypothesis → Generate Data (from null) → Calculate Statistic → Visualize

```
hypothesize(null)  %>%  generate(reps)  %>%  calculate(stat)  %>%  visualize()
```

# Hypothesis Testing



`specify()`

`hypothesize()`

`generate()`

`calculate()`

`visualize()`