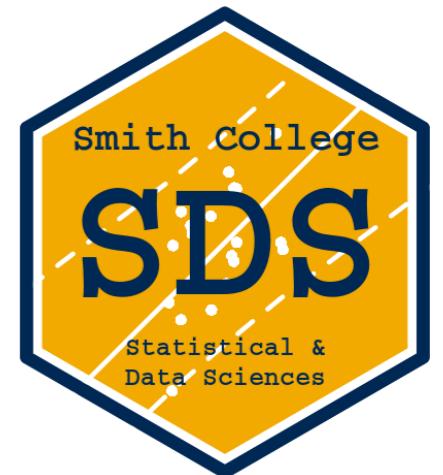
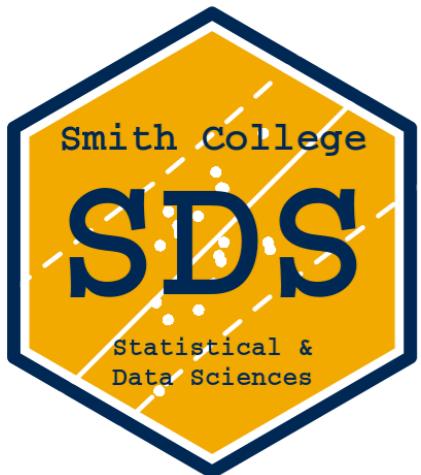


# Statistical Inference via Data Science: A ModernDive into R & the Tidyverse



**Albert Y. Kim**  
**UBC Statistics**  
**Vancouver BC Canada**  
**Tuesday May 19, 2020**



Slides available at [twitter.com/rudeboybert](https://twitter.com/rudeboybert)



Statistical inference **via**  
data science...

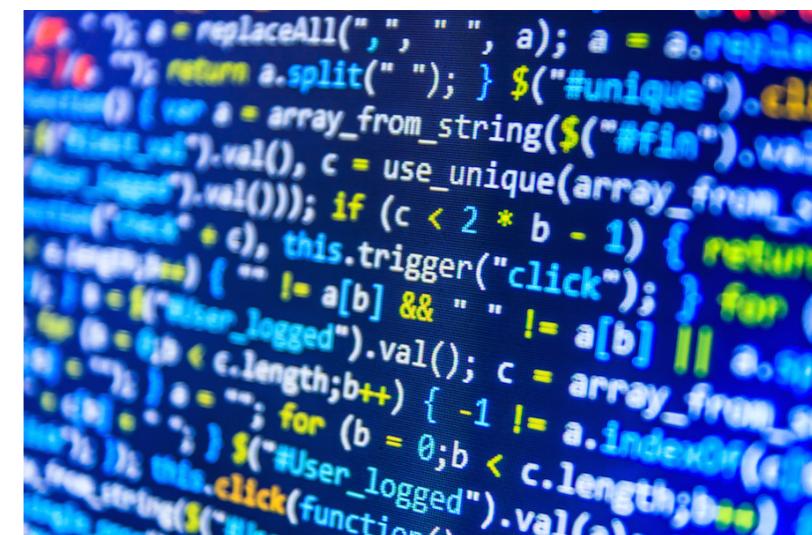
# Guiding Paper

“Mere Renovation is Too Little Too Late: We Need to Rethink Our Undergraduate Curriculum from the Ground Up” by Cobb RIP  (TAS 2015)

- Minimize prerequisites to research
- Substitute “mathematics” with “computation” as the *engine of statistics*

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

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



# My ideal pathway for Intro Stats Students

Have them:

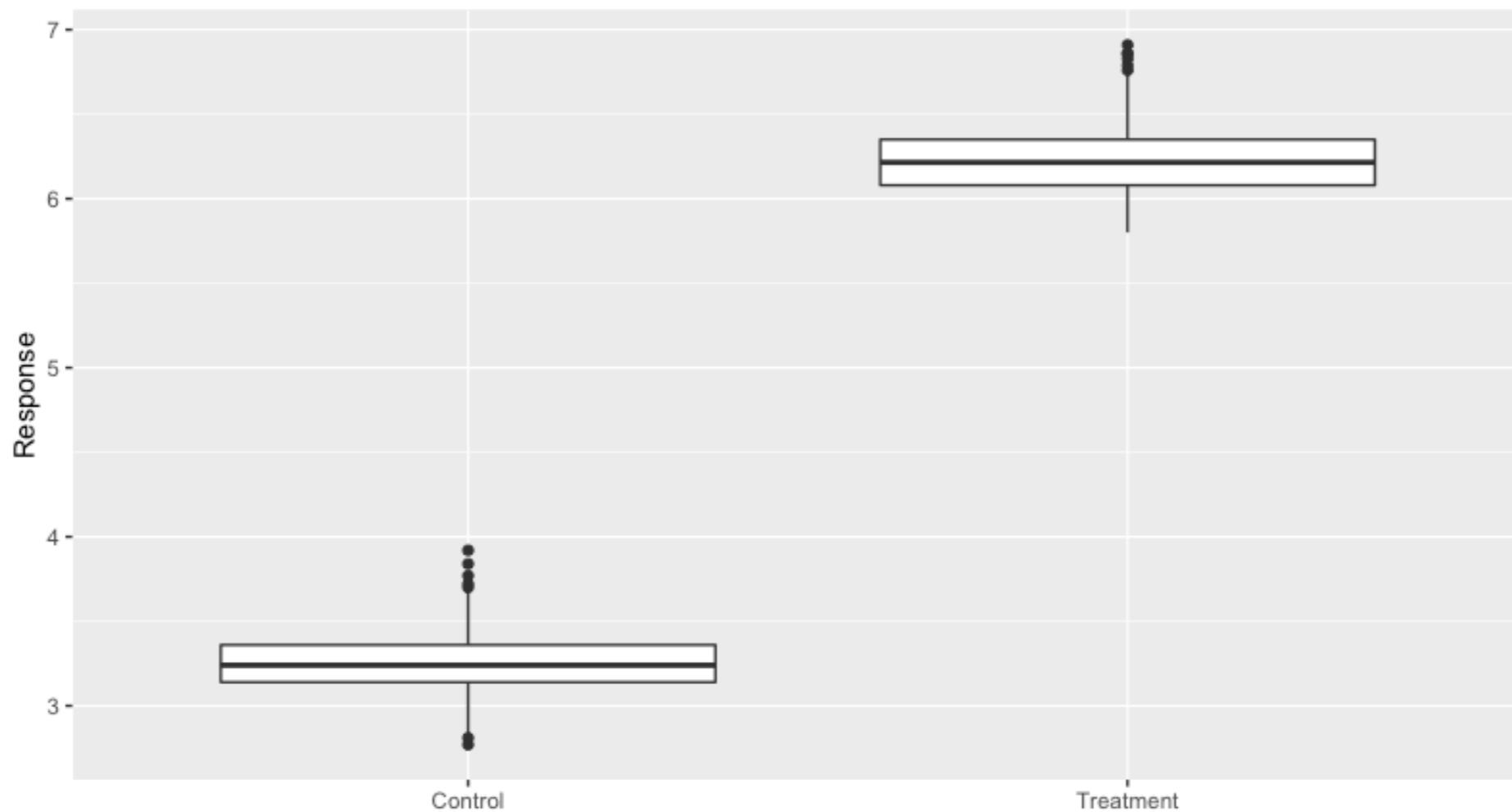
1. Develop a minimally viable “data science” toolbox
2. Build intuition for statistical inference by learning simulation-based methods that they implement themselves using these tools
3. Bridge the gap between simulation-based & traditional asymptotic-based inference

**i.e. Teach Data Science first, Statistics second**

# Ch3 Data Viz with *ggplot2*

“You don’t need no PhD in Stats, just EDA”

Question: Is there a difference in response?



Versus just saying: “The p-value is 0!”

# Ch2,4,5 Data Wrangling, Importing, “Tidy”

Have students practice:

- Looking at raw data values using `View()`
- Functional programming with the pipe `%>%`  
i.e. develop *algorithmic thinking*
- Thinking of data in terms of “tidy” *data frames* that can be transformed with `filter()`, `mutate()`,  
`group_by()` `%>%` `summarize()`

state	year	voted
AK	2016	TRUE
AL	2016	TRUE
AR	2016	TRUE
AZ	2016	TRUE
CA	2016	TRUE
CO	2016	TRUE
CT	2016	TRUE
DC	2016	TRUE
DE	2016	TRUE
FL	2016	TRUE
GA	2016	TRUE
HI	2016	TRUE
IA	2016	TRUE

VS

all the same type (numeric)

year
2015
2013
2011
2016
2018

vector

list

state_data
“Arizona”
2015
733375
TRUE
<tbl_df [12,4]>

character

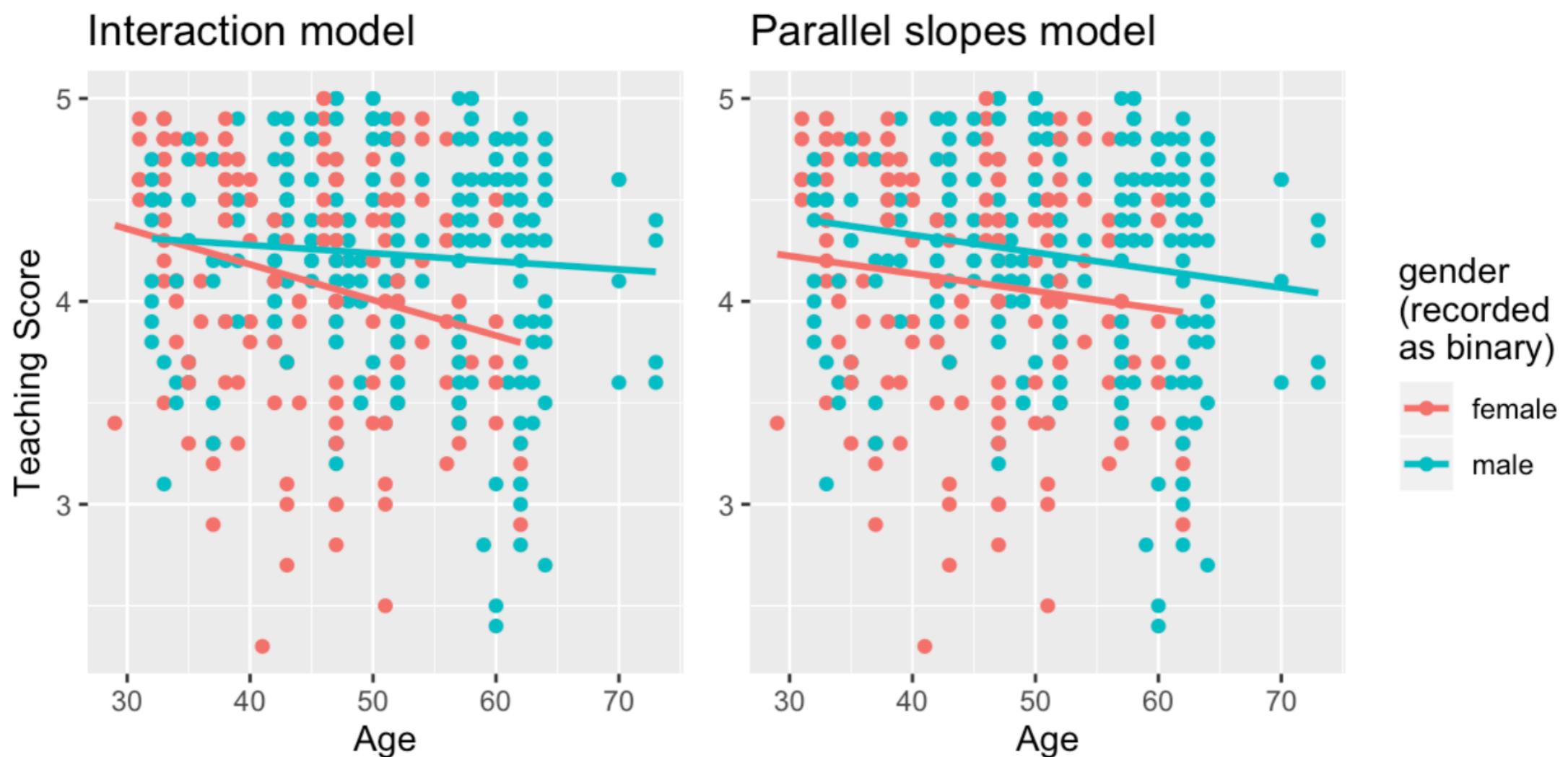
numeric

logical

list

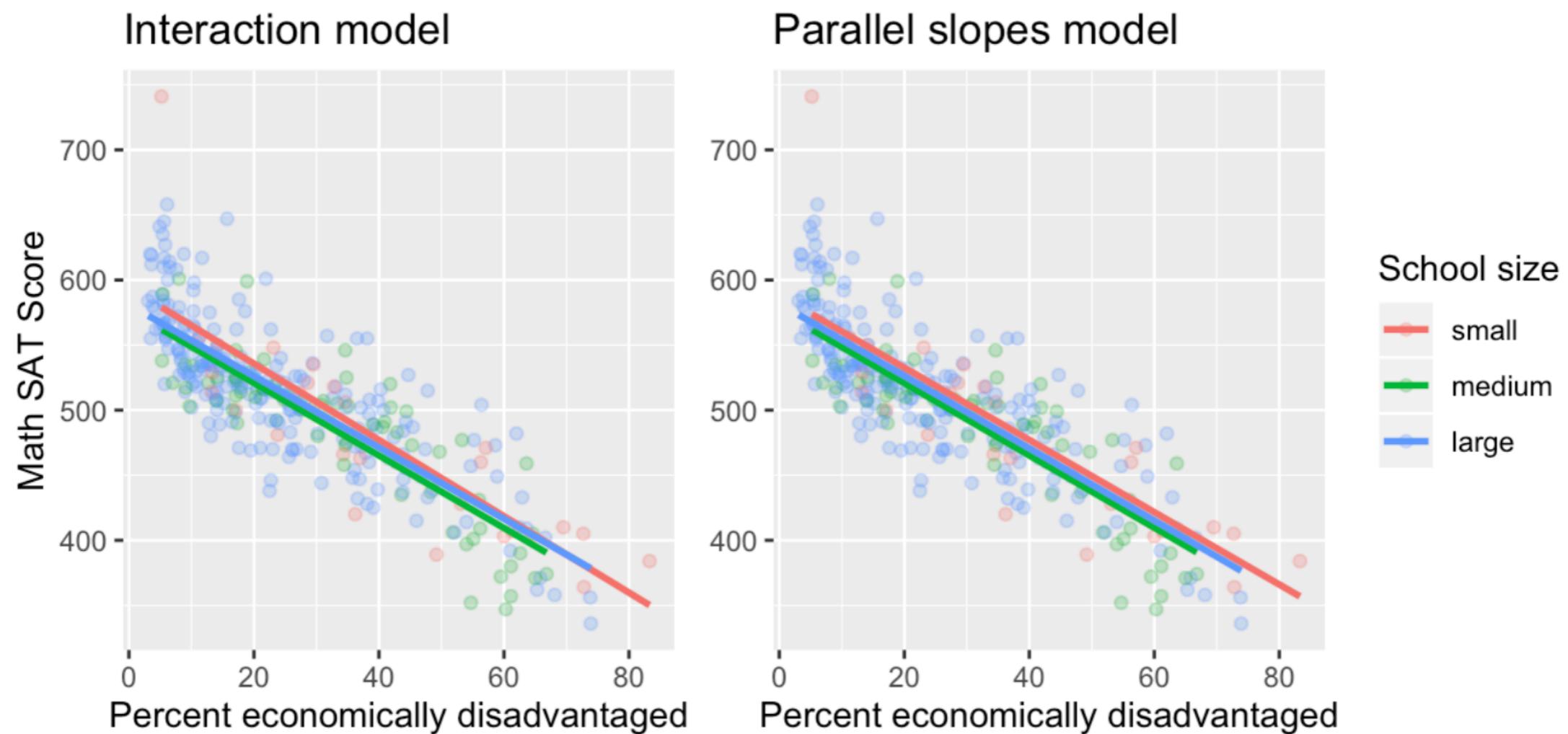
# An aside: Ch6+7 Linear Regression

“Visual” model selection using teaching evals data



# An aside: Ch6+7 Linear Regression

“Visual” model selection using 2017 MA Public HS Data



# My ideal pathway for Intro Stats

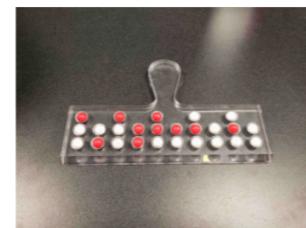
Have them:

1. Develop a minimally viable “data science” toolbox
2. **Build intuition for statistical inference by learning simulation-based methods that they implement themselves using these tools**
3. Bridge the gap between simulation-based & traditional asymptotic-based inference

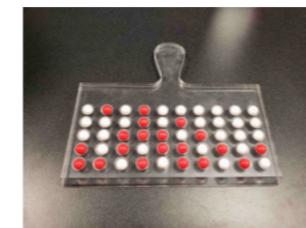
# Ch7 Sampling



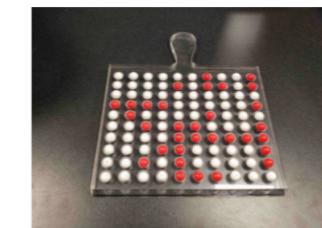
A shovel with 25 slots



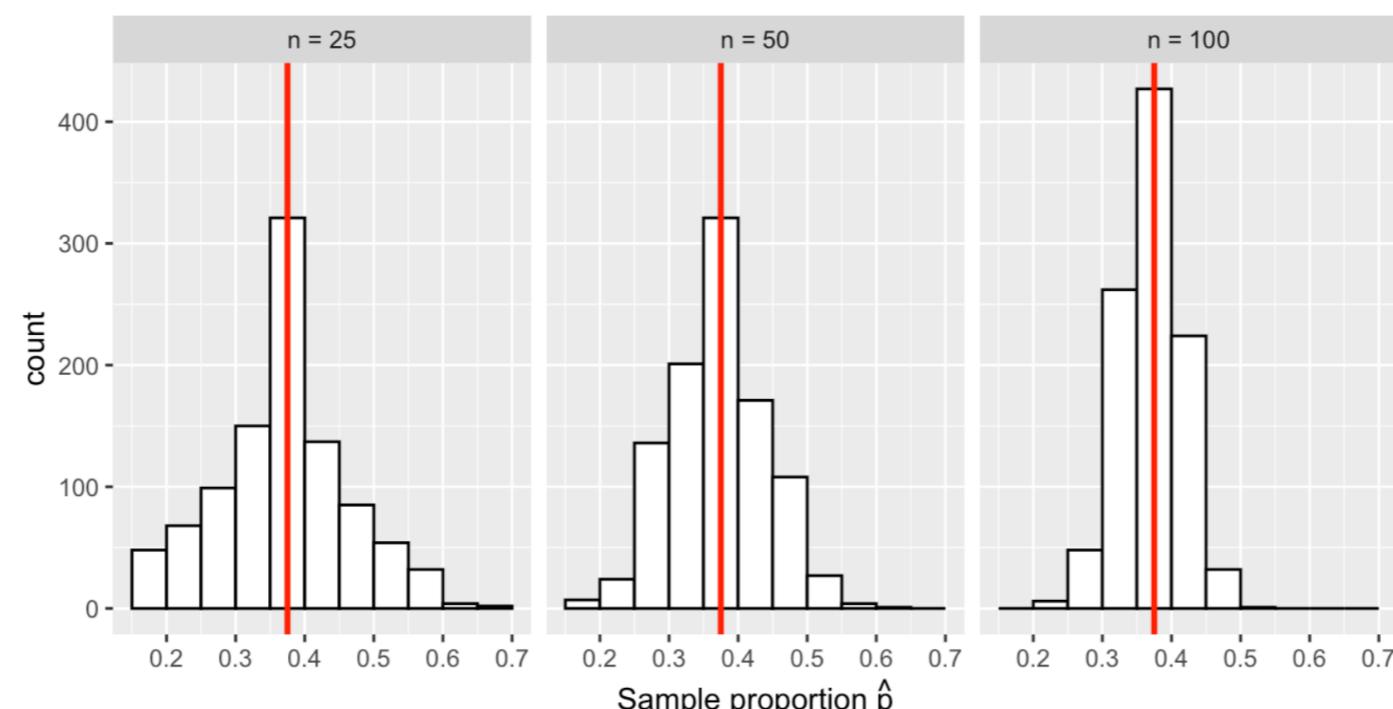
A shovel with 50 slots



A shovel with 100 slots



Sampling distributions of  $\hat{p}$  based on  $n = 25, 50, 100$ .

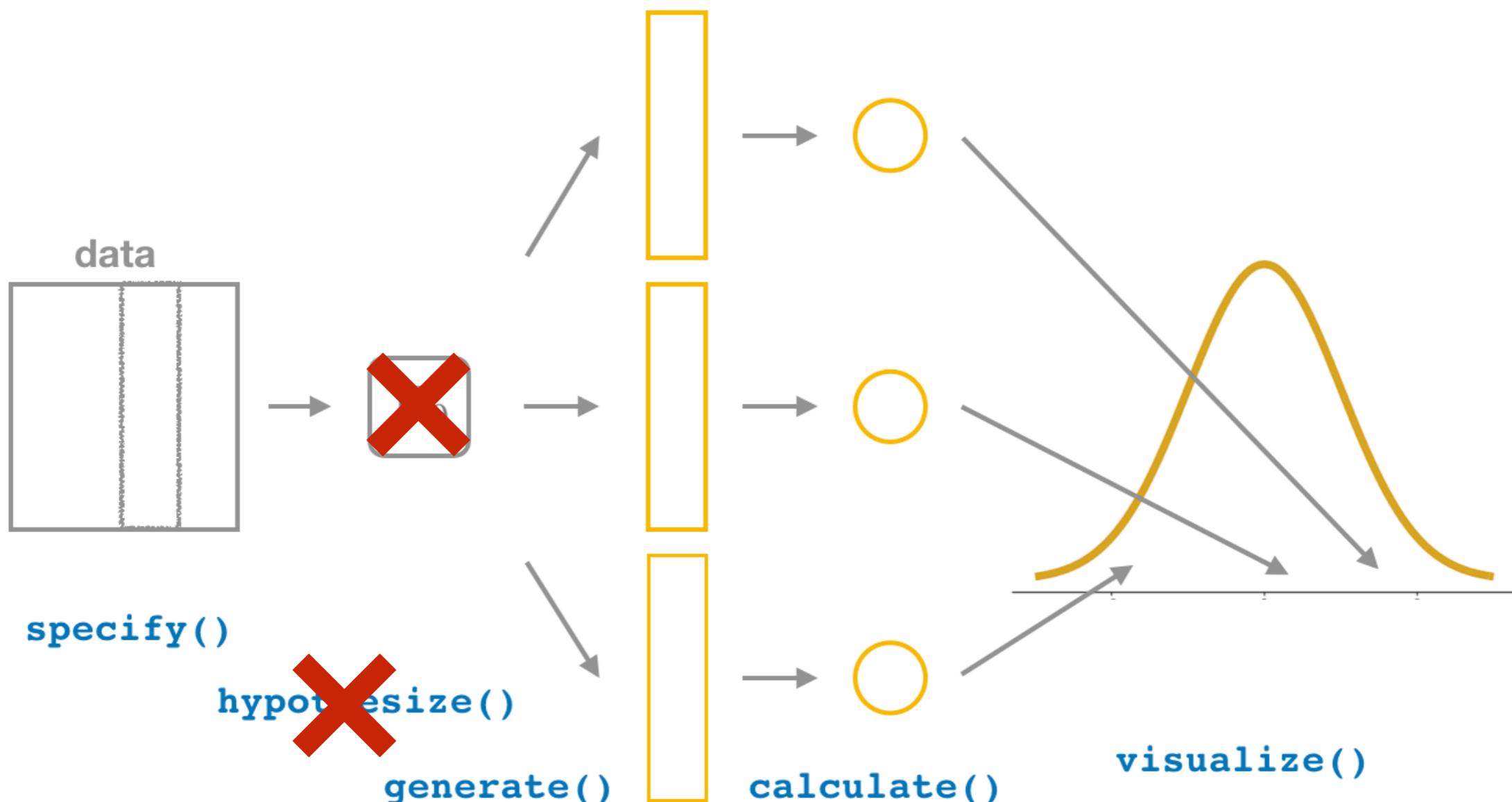


# Sampling Scenarios Covered

TABLE 7.5: Scenarios of sampling for inference

Scenario	Population parameter	Notation	Point estimate	Symbol(s)
1	Population proportion	$p$	Sample proportion	$\hat{p}$
2	Population mean	$\mu$	Sample mean	$\bar{x}$ or $\hat{\mu}$
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	$\bar{x}_1 - \bar{x}_2$
5	Population regression slope	$\beta_1$	Fitted regression slope	$b_1$ or $\hat{\beta}_1$

# infer package for “tidy” statistical inference



**Skip this step for  
confidence intervals**

# Ch8 What is mean year of all 🇺🇸 pennies?

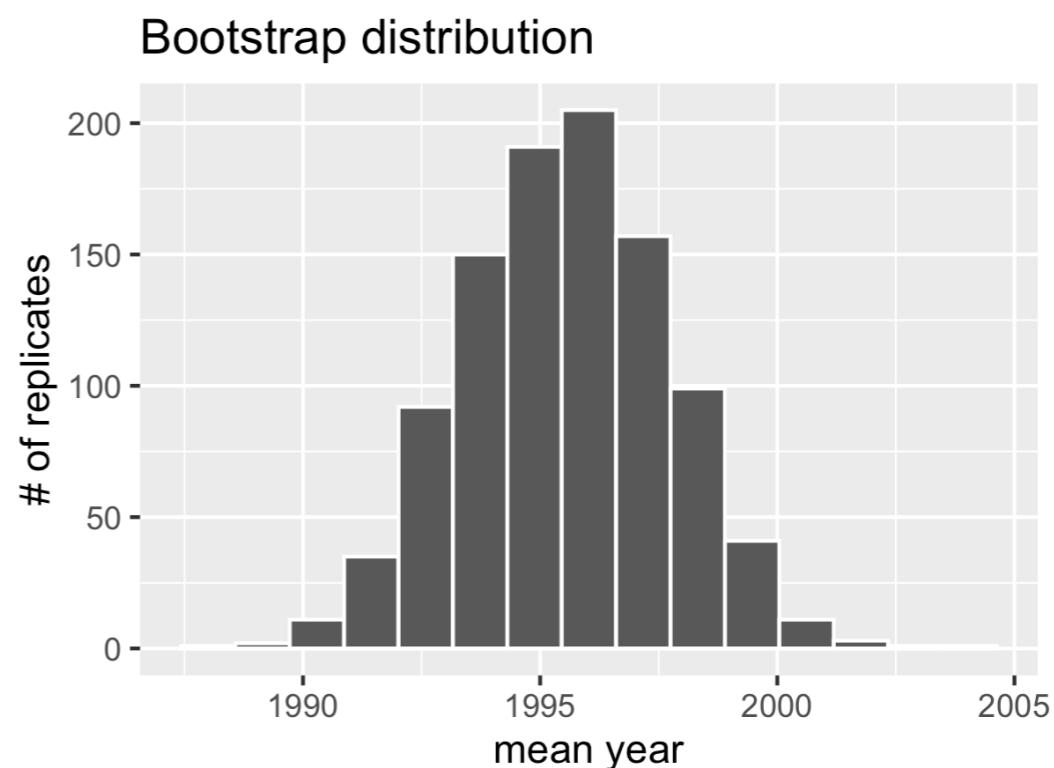


```
> library(moderndive)
> pennies_sample
# A tibble: 50 x 2
  ID    year
  <int> <dbl>
1 1     2002
2 2     1986
3 3     2017
4 4     1988
5 5     2008
6 6     1983
7 7     2008
8 8     1996
9 9     2004
10 10    2000
# ... with 40 more rows
```

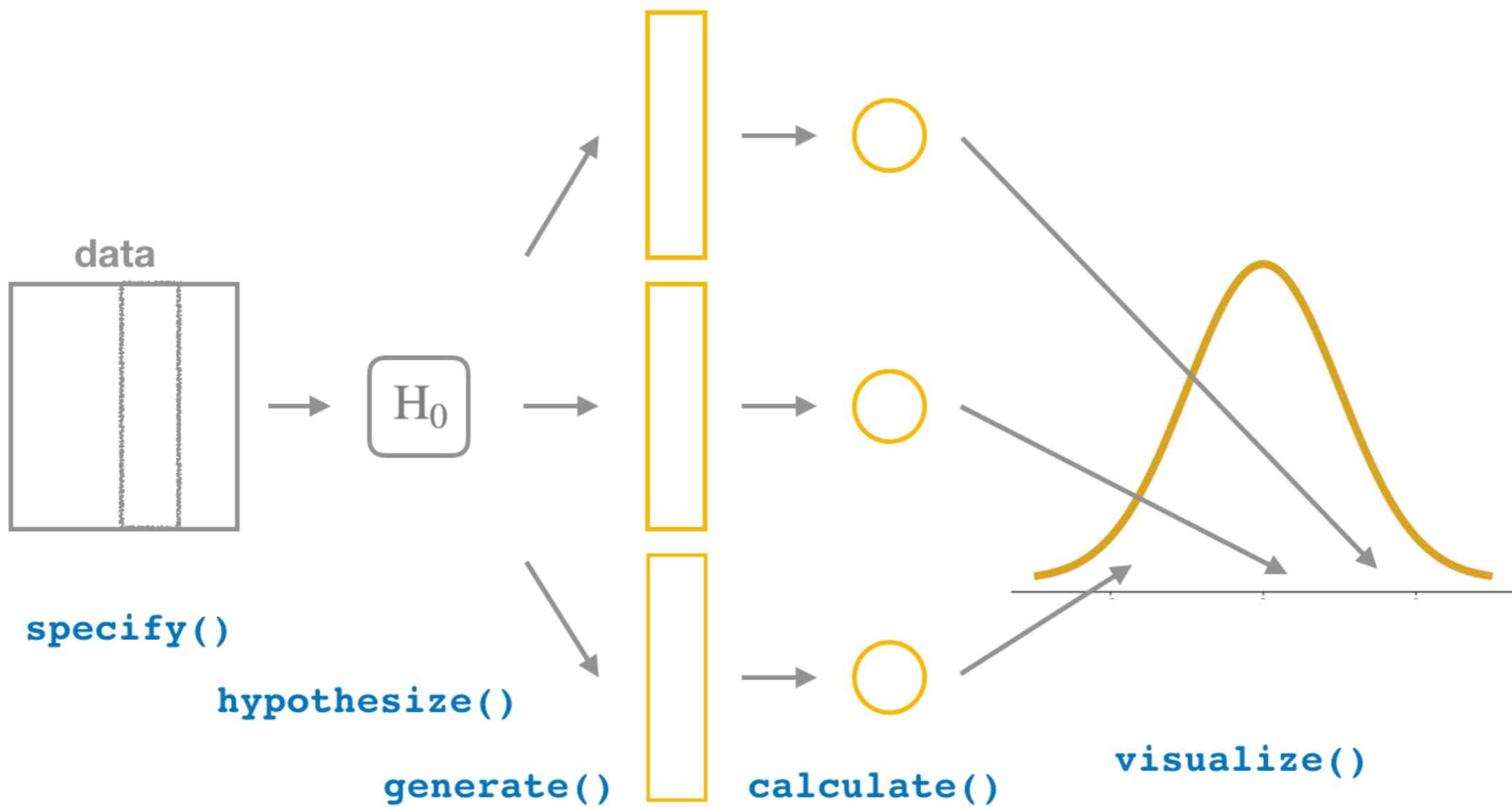
Using bootstrap resampling with replacement:

```
library(tidyverse)
library(infer)

pennies_sample %>%
  specify(response = year) %>%
  generate(reps = 1000) %>%
  calculate(stat = "mean")
```



# infer package for “tidy” statistical inference



# Ch9 Does gender affect promotions at banks?

Using 48 identical résumés: Diff of 29.2%

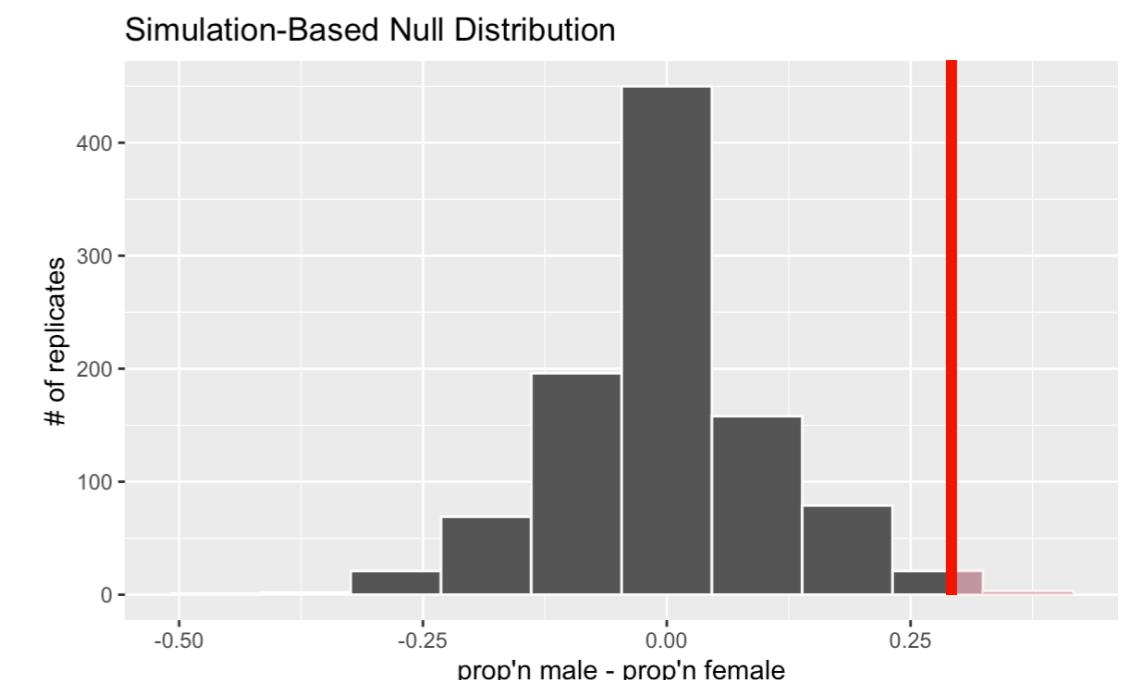


```
> library(moderndive)
> promotions
# A tibble: 48 x 3
  id decision gender
  <int> <fct>   <fct>
1 11 promoted male
2 3 promoted male
3 7 promoted male
4 44 not     female
5 30 promoted female
6 38 not     male
7 2 promoted male
8 24 promoted female
9 1 promoted male
10 20 promoted male
# ... with 38 more rows
```

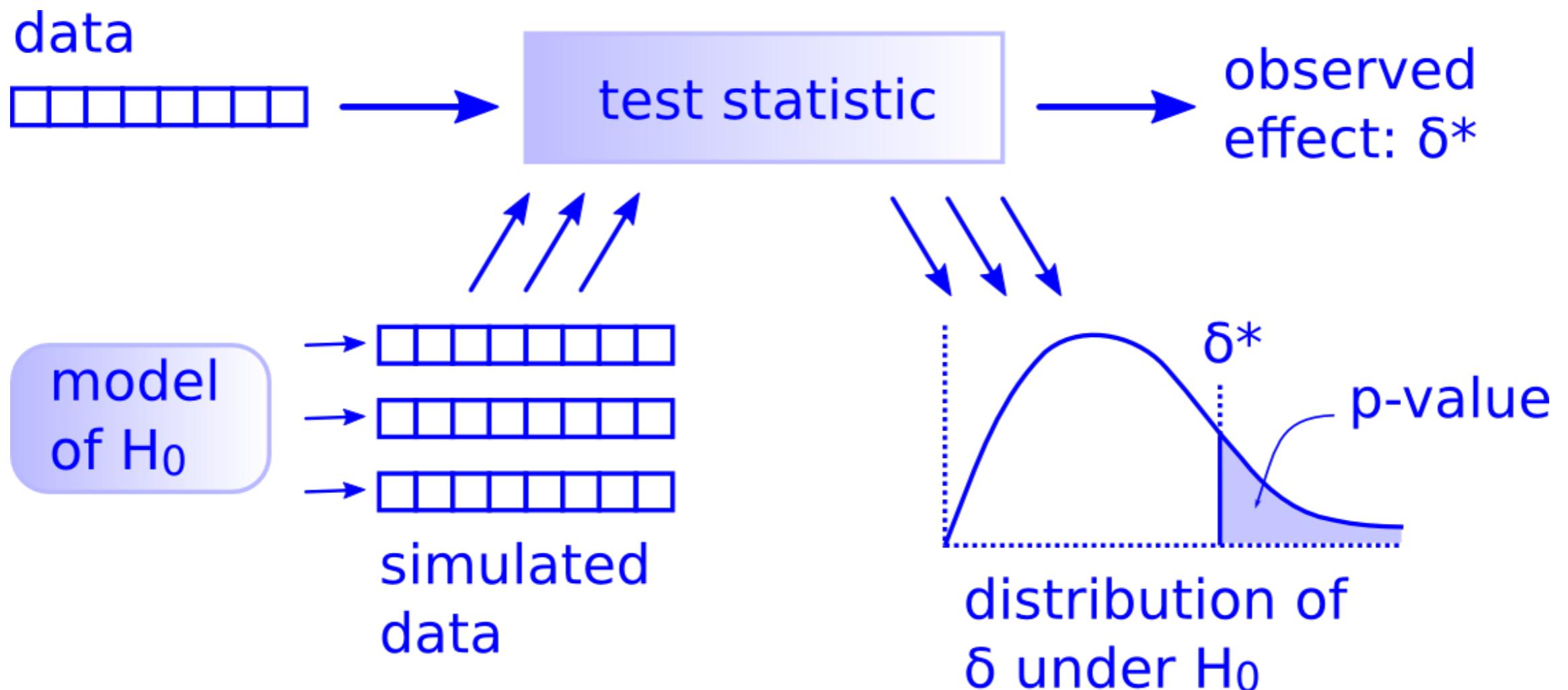
```
> # Under H0: p_m - p_f = 0
> promotions
# A tibble: 48 x 4
  id decision gender gender_shuffle
  <int> <fct>   <fct>   <fct>
1 11 promoted male male
2 3 promoted male female
3 7 promoted male male
4 44 not     female female
5 30 promoted female male
6 38 not     male female
7 2 promoted male male
8 24 promoted female male
9 1 promoted male male
10 20 promoted male male
# ... with 38 more rows
```

Using permutation “shuffling” assuming  $H_0$ :

```
# Diff args given that variables are binary:
promotions %>%
  specify(formula = decision ~ gender,
          success = "promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props",
            order = c("male", "female"))
```



# “There is only one test”



# Ch10 Revisit Regression

# My ideal pathway for Intro Stats

Have them:

1. Develop a minimally viable “data science” toolbox
2. Build intuition for statistical inference by learning simulation-based methods that *they implement themselves* using these tools
3. **Bridge the gap between simulation-based & traditional asymptotic-based inference**

# How to make room for data science

In a single term course:

1. Drop all probability theory
2. Drop asymptotic theory in favor of simulation based inference
3. Introduce the “There is only one test” framework and tell students it’s true
4. De-emphasize  $\chi^2$  tests & ANOVA

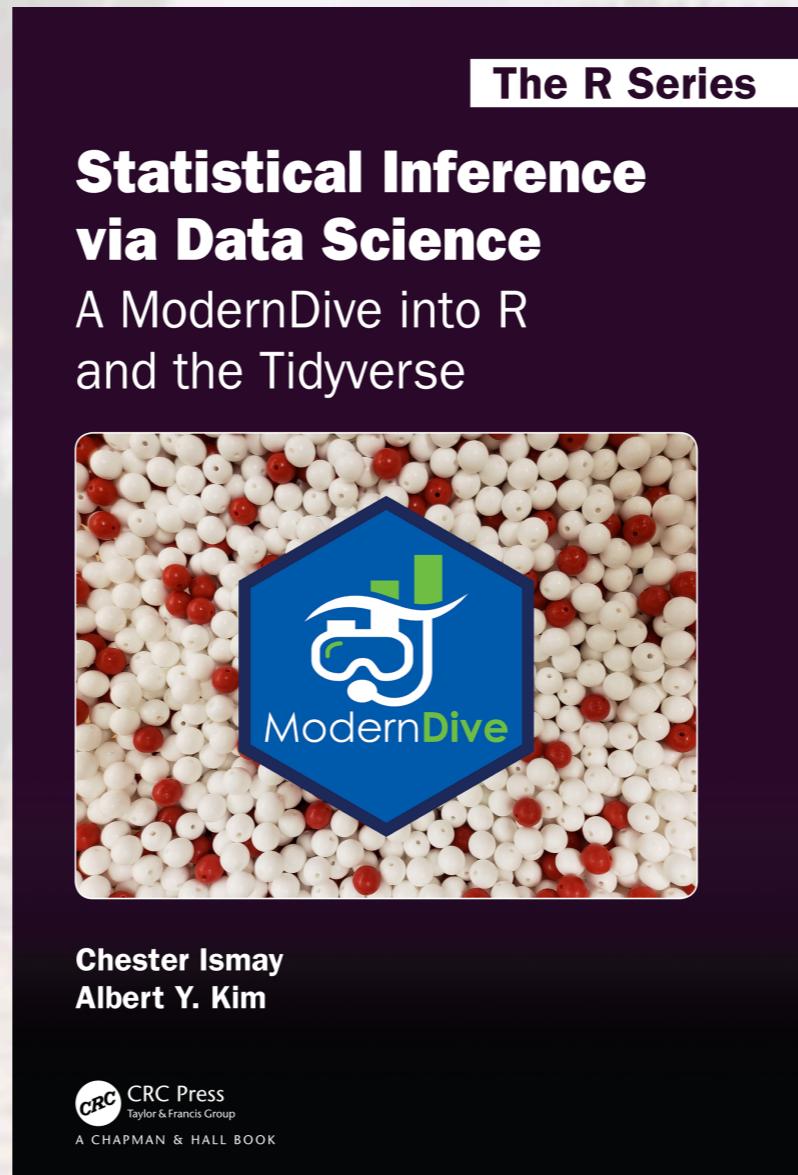
When you have room for data science

In a multi-term sequence of courses:

1. *Cover probability theory: distributions, z-scores*
2. *Make explicit connections between asymptotic theory & simulation based inference*
3. *Repeatedly go thru “There is only one test” framework & convince students it’s true*
4. *Cover  $\chi^2$  tests & ANOVA as case studies*

# Currently in ModernDive

# For more info check out:



- Available free online at [moderndive.com](http://moderndive.com)
- Print copies on sale at CRC Press website:  
Use discount code ASA18
- Slides available at [twitter.com/rudeboybert](https://twitter.com/rudeboybert)

# My ideal pathway for Intro Stats

1. Develop a minimally viable “data science” toolbox *then*
2. Build intuition for statistical inference by teaching simulation-based methods using these tools *then*
3. Bridge the gap between simulation-based inference & traditional asymptotic-based inference

# Why tidyverse?