moderndive: statistical inference via the tidyverse

Albert Y. Kim

@rudeboybert

Cal Poly Statistics Department
San Luis Obispo
May 23

My Co-Authors

Chester Ismay:
Textbook co-author

Jenny Smetzer:
Labs author

Background

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:

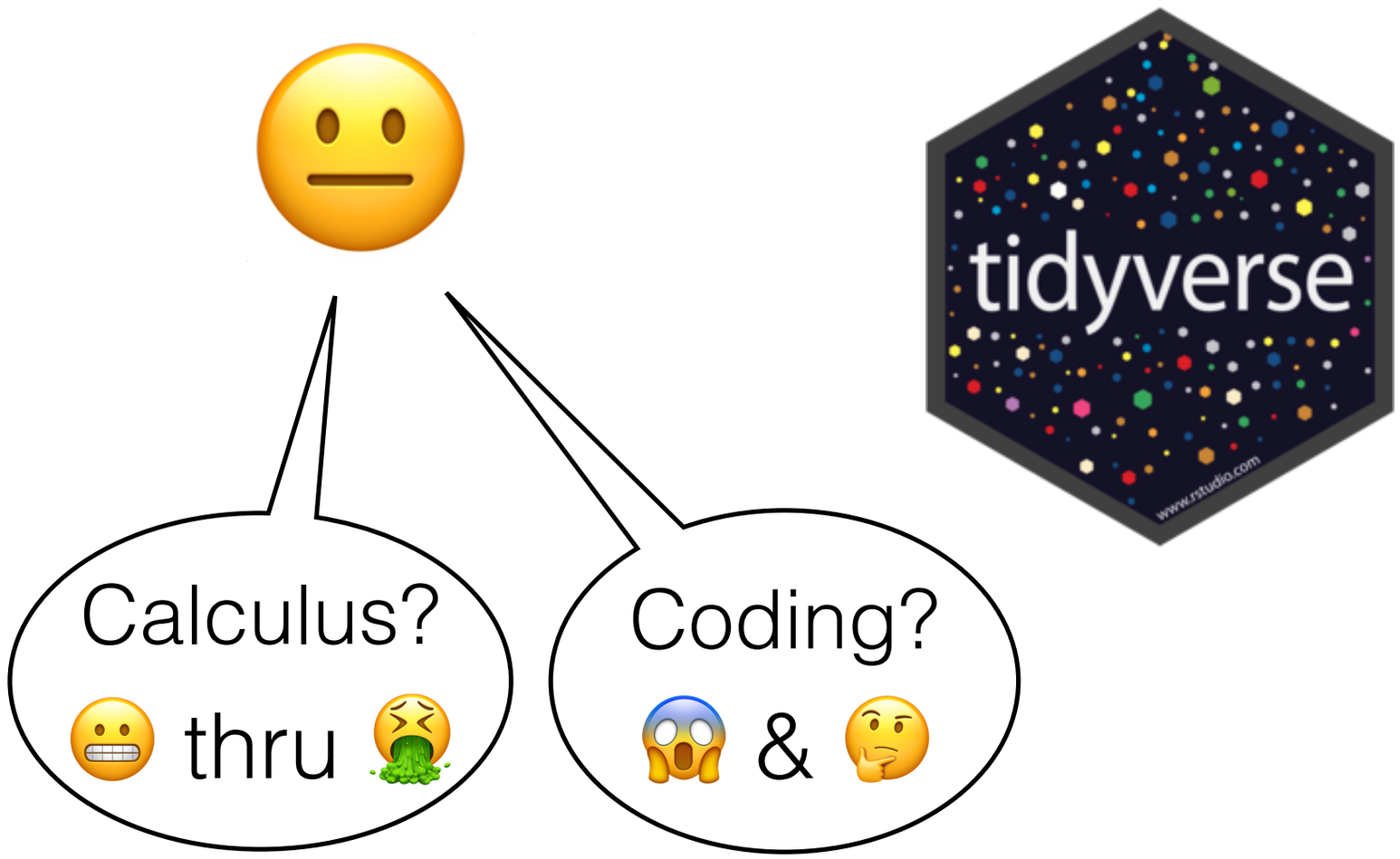
- Goal 1: Modeling with regression
- Goal 2: Sampling for inference

Getting from Point A to Point B

Point A:
Modal 1st time stats student

via the
`tidyverse`

Point B:
Two goals



1. Modeling with regression
2. Sampling for inference

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 in general?

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. Breakdown large tasks into steps using **%>%** "then"
3. What is the goal of your code?
4. Make code understandable to humans

Why tidyverse for stats newbies?

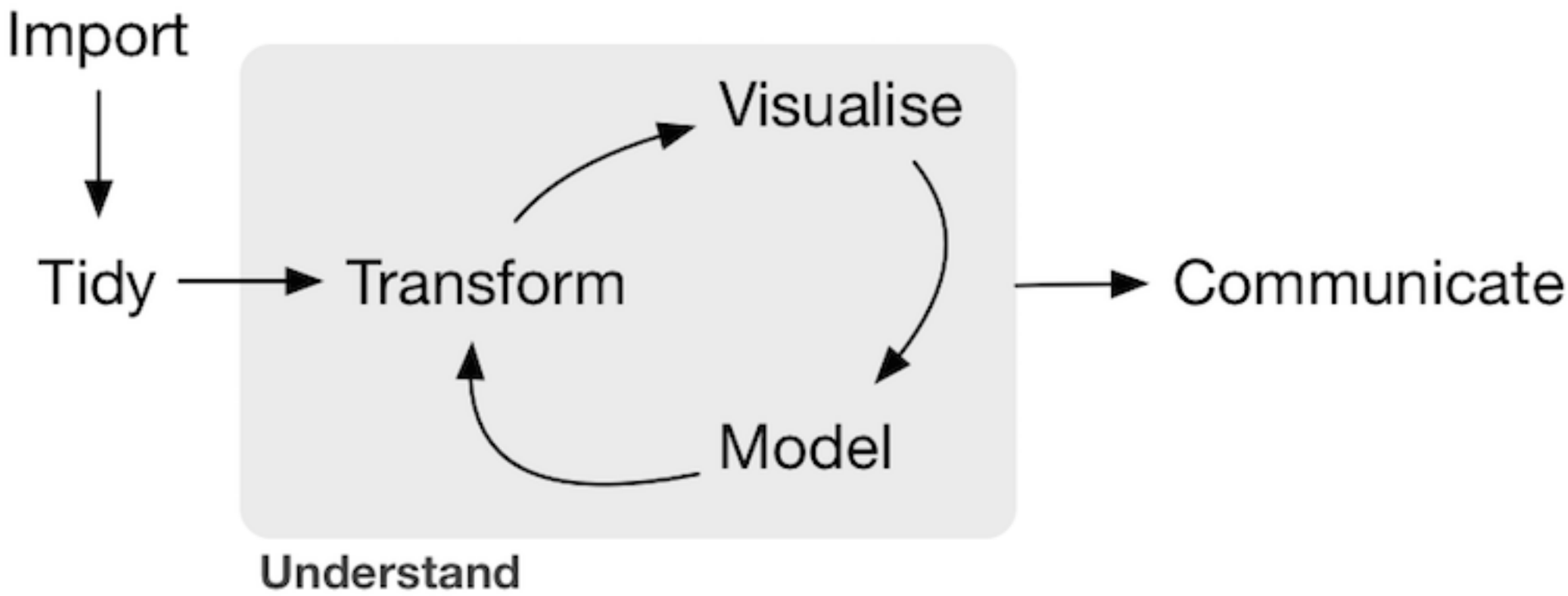
- IMO it's easier to learn than base R. Others too.
- It scales. You leverage an entire ecosystem of online developers and support: Google & StackOverflow
- Satisfy learning goals *while learning tools they can use beyond the classroom.*

End Deliverable of Course

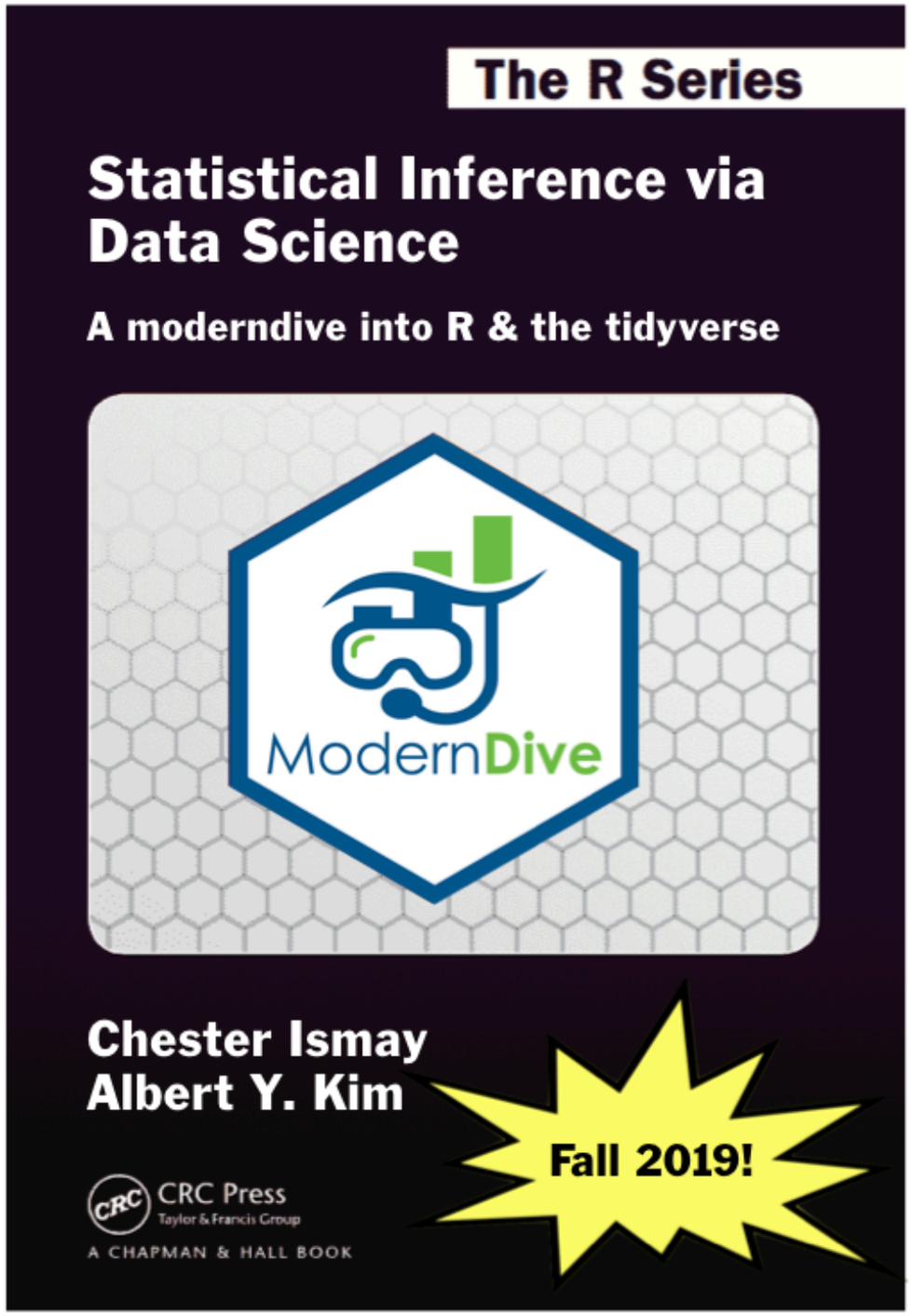
- Think of how youths learn to play sports…
- IMO stats newbies should learn to *"play the whole game"* in simplified form first
  - **%>%** add layers of complexity…
  - **%>%** add more layers of complexity…
  - **%>%** add more layers of complexity…
- Do this instead of learning individual components in isolation

End Deliverable of Course

Final project that "plays the whole game"
of *all components* of data/science pipeline:



Example template given to students this semester,
based on work by students
Alexis C., Andrianne D., & Isabel G.



Development version at moderndive.netlify.com

Part I: Data Science via the `tidyverse`

Chapters 2 - 5

Chapter 2: Getting Started

R: Engine	RStudio: Dashboard
	

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

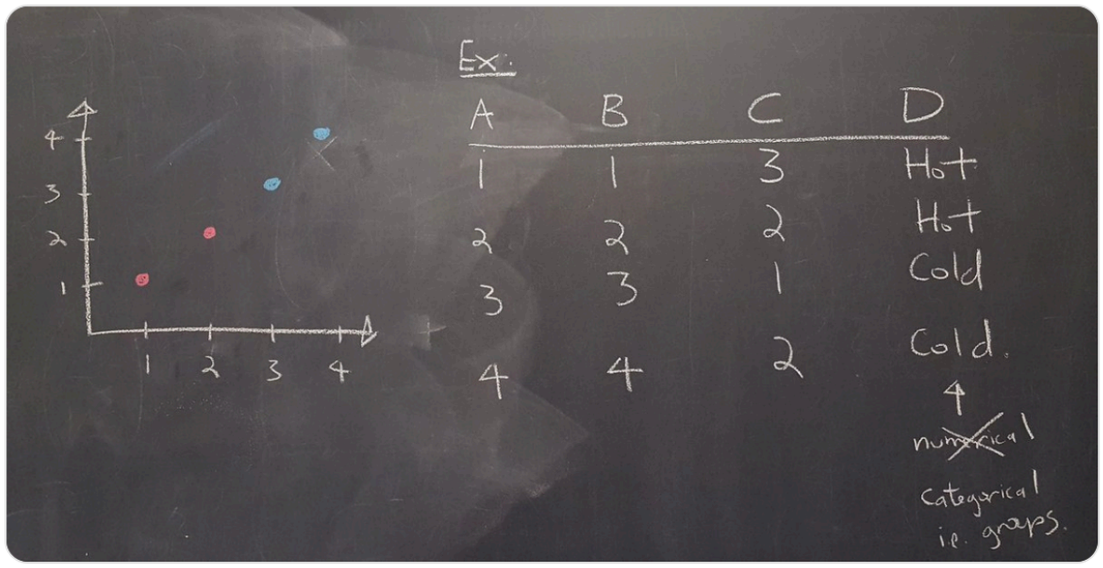
- IMO RStudio's best function: `View()`
- Getting students over initial 😱 of coding
- Think piece:
  "Why women in psychology can't program"

Chapter 3: Data Viz via `ggplot2`

Often said “Intro students can’t learn ggplot”

Albert Y. Kim
@rudeboybert

Intro stats & data science #chalktalk of grammar of graphics + homage to @katyperry today, #ggplot2 tomorrow #rstats



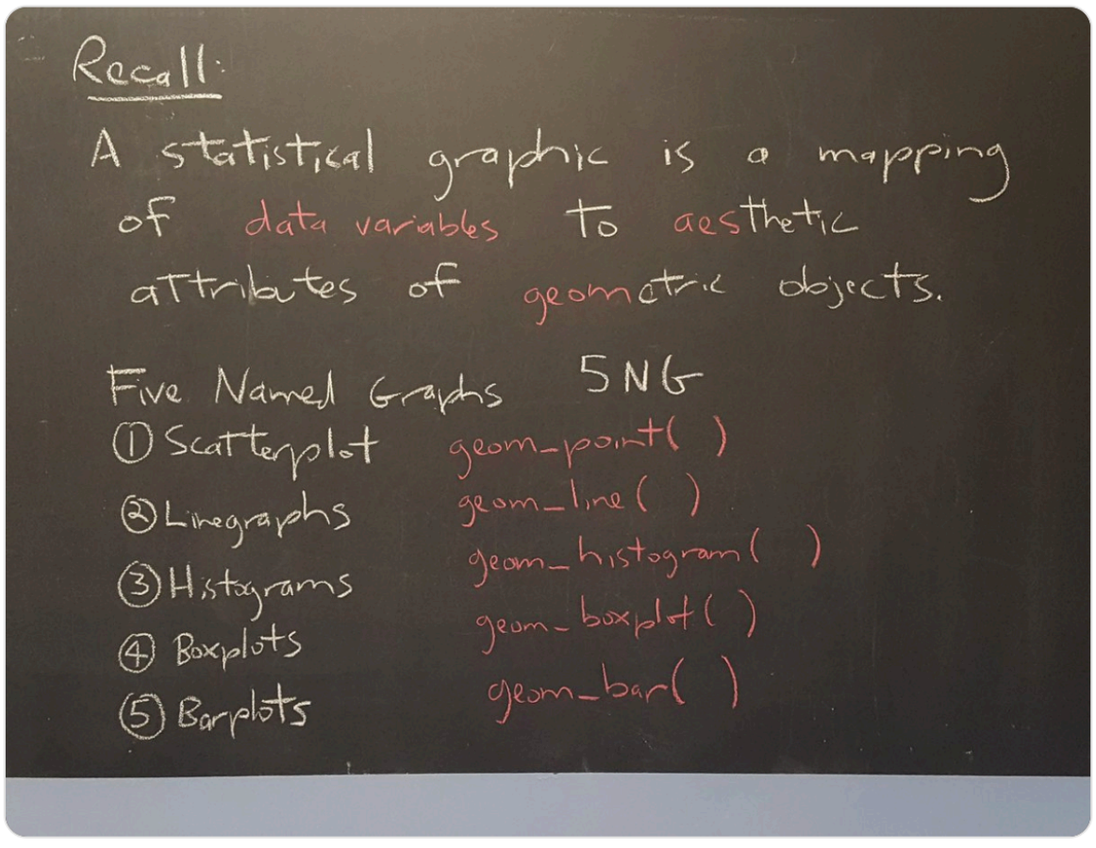
11:58 AM - 11 Sep 2017 from Amherst College

5 Retweets **29** Likes

3 5 29

Albert Y. Kim
@rudeboybert

#chalktalk of #GrammarOfGraphics definition of "statistical graphic" + @ModernDive's "Five Named Graphs" #5NG #ggplot2



12:50 PM - 12 Sep 2017 from Amherst College

15 Retweets **61** Likes

15 61

Chapter 4: Data Wrangling via `dplyr`
Chapter 5: "Tidy" Data via `tidyr`

- Essential: **%>%** operator as it's needed later.
- Balance of how much students wrangling do vs how much you do for them?
- To *completely* shield students from *any* data wrangling is to betray true nature of work in our fields
- One proposed balance is in "tame" data & fivethirtyeight package paper (Kim, Ismay, Chunn)

Part II: Data Modeling via moderndive

Chapters 6, 7, & 11

Goal 1: Modeling with Regression

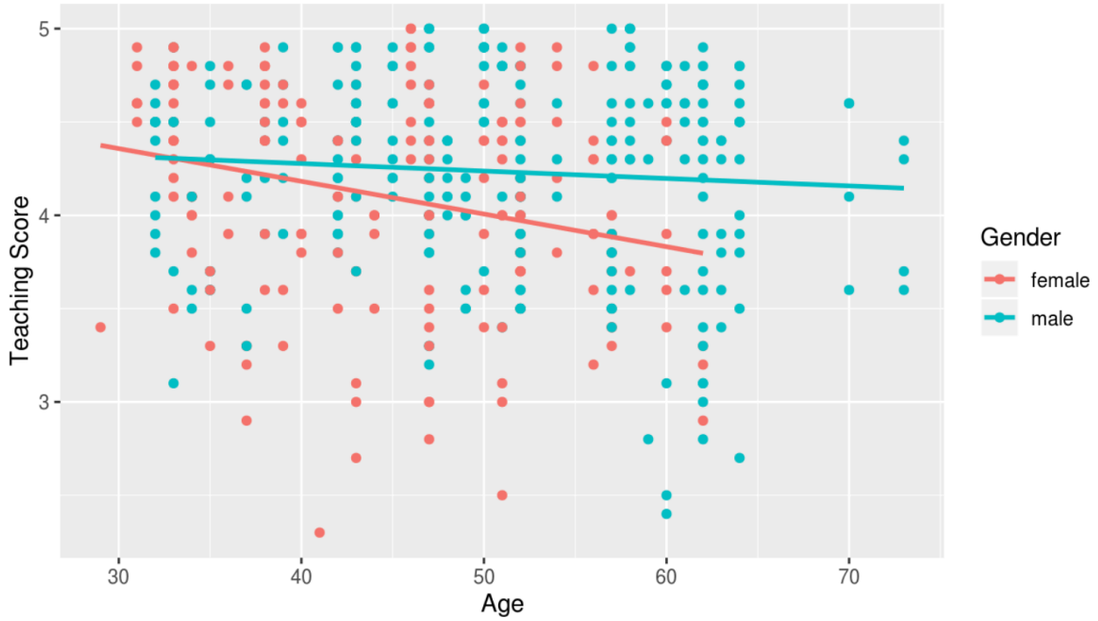
1. Data: `evals`

evals

Filter

	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

Console ~/

```
> 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

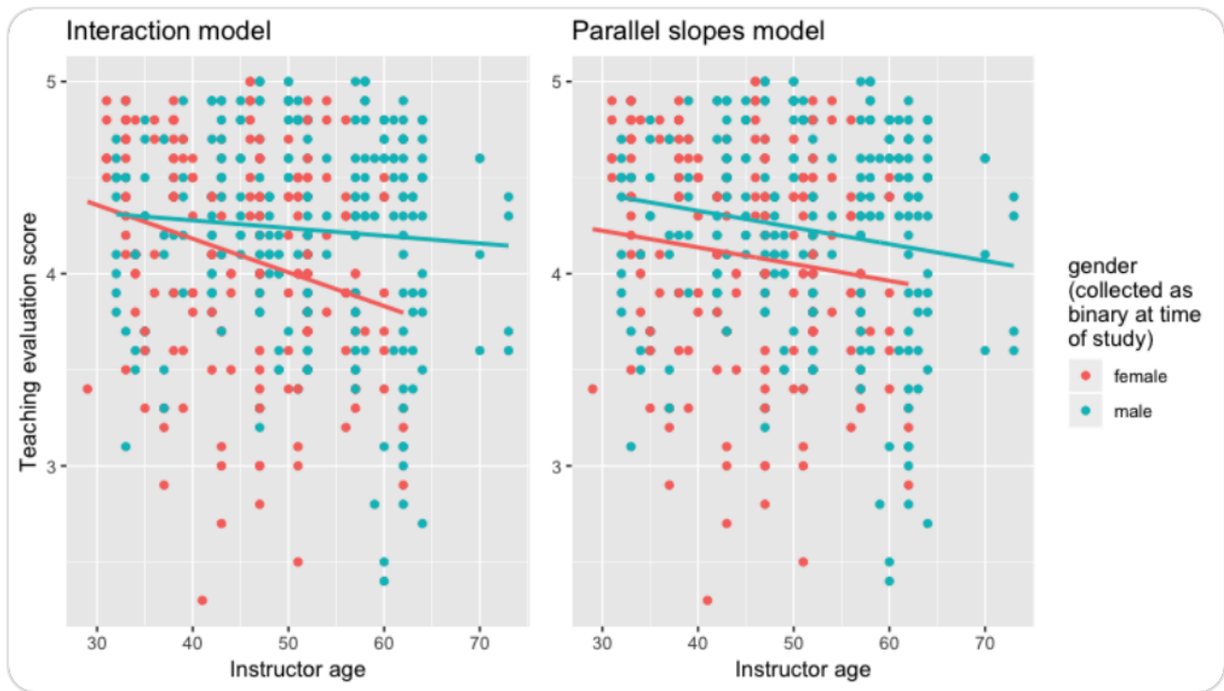
More later!

Also model selection!

Is there a way to teach
✅ model selection
✅ model complexity vs parsimony
✅ occam's razor
To intro stats students? 👶🏿👶🏽👶

YES! Via data viz 📊📈📉 & EDA 🔍!

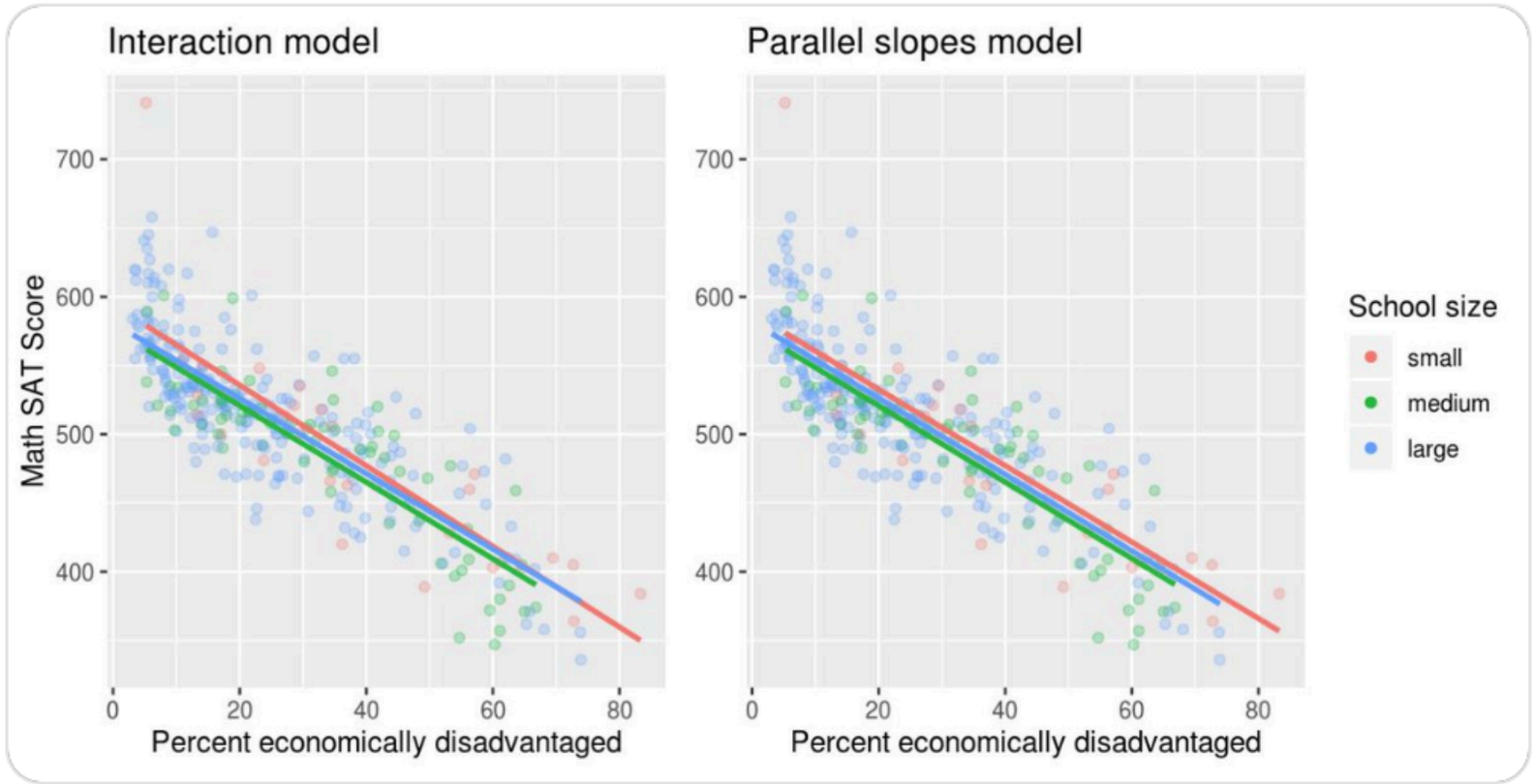
First show a case study where "interaction model" >>> "parallel slopes model"! 1/4



More model selection

ModernDive @ModernDive · Apr 19

Next show a case study where "interaction model" vs "parallel slopes model" is "I dunno?!? They look kinda the same to me?!? 🤷🏻‍♀️🤷🏻‍♂️🤷🏾‍♂️🤷‍♀️"? 2/4

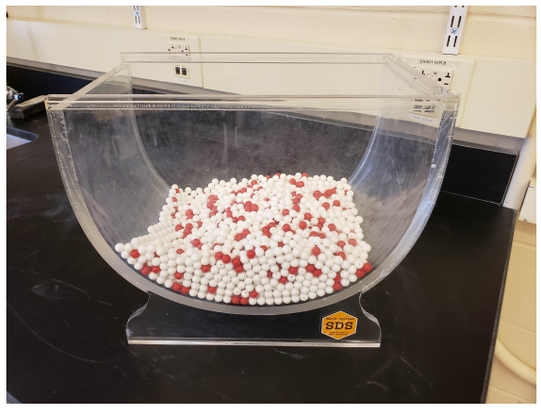
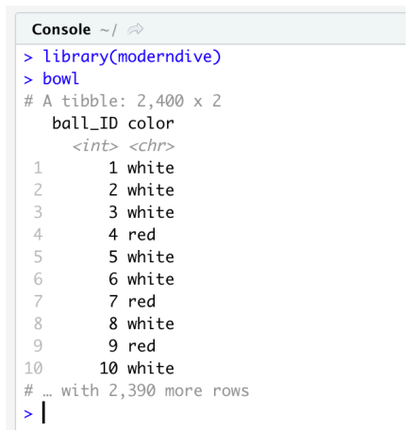
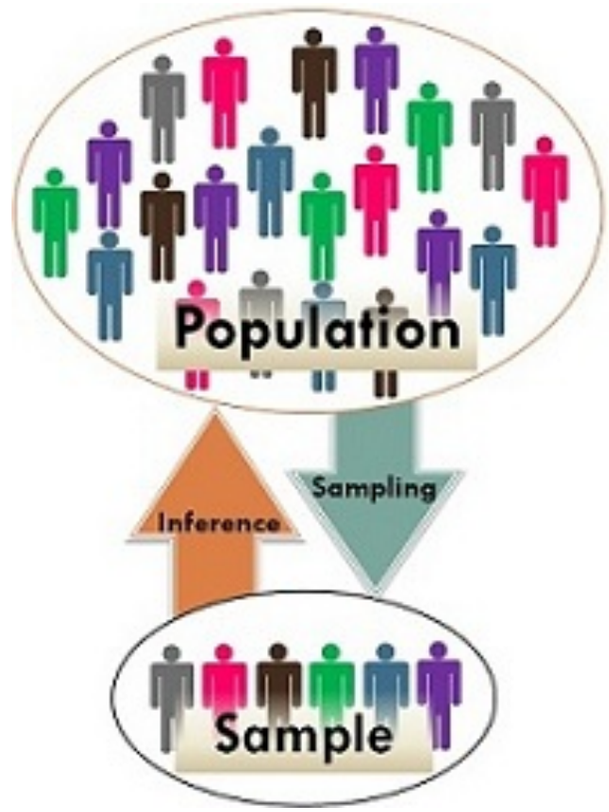
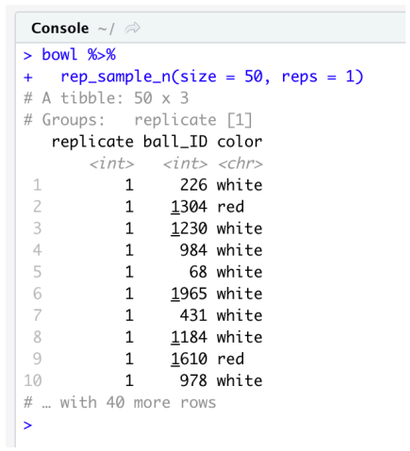
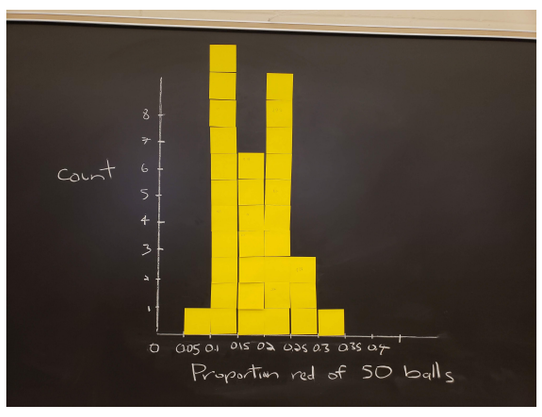
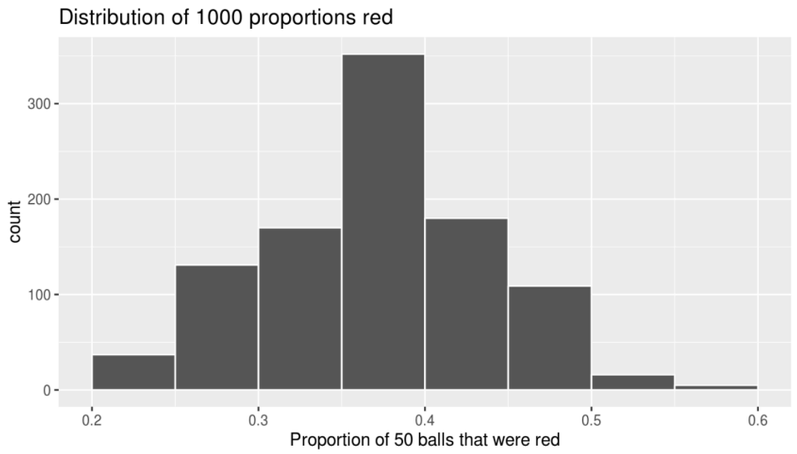
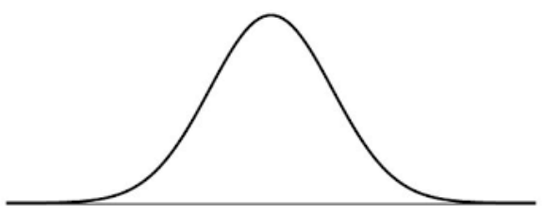


1 1 1

Part III: Statistical Inference via `infer`

Chapters 8 - 11

Goal 2: Sampling for Inference

	1. Tactile Sampling →	2. Virtual Sampling →	3. Theoretical
Population			
Sample			
Sampling Distributions & Standard Errors			 $SE = \sqrt{\frac{p(1-p)}{n}}$

Chapter 8: Sampling

Terminology, definitions, & notation 😱

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

Statistics/ISOSTAT ×

[redacted] 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

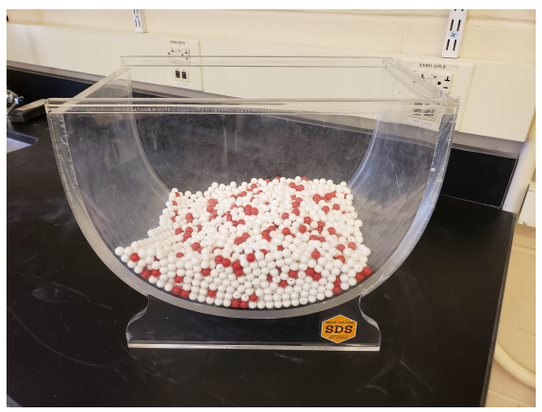
[redacted] Thu, Jan 3, 2:42 PM
Hi Matt, I teach a “statistics” course to medical students at Duke. I use quotes around the word statistics because I don’t really teach the students how to do

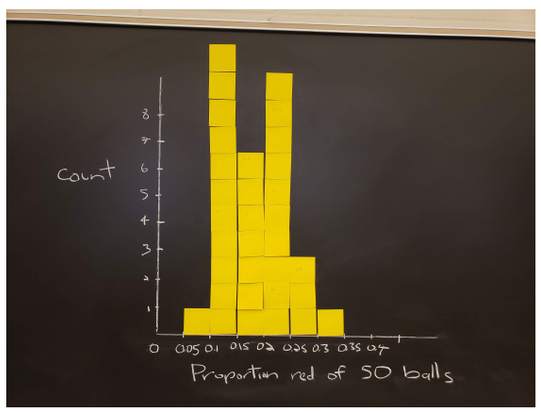
[redacted] Thu, Jan 3, 2:53 PM
Hi, I like the work of Kaplan and Rogness for some nice activities and a discussion of lexical ambiguity in statistics. https://scholarcommons.usf.edu/numeracy/

[redacted] 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

[redacted] Thu, Jan 3, 4:10 PM
I don't think the issue is using percentages but rather using percentages while giving students a formula for proportions;-)

Our approach: Do this first…





Terminology, definitions, & notation

Then this…

TABLE 8.6: Scenarios of sampling for inference

Scenario	Population parameter	Notation	Point estimate	Notation.
1	Population proportion	p	Sample proportion	$\hat{p}$

Terminology, definitions, & notation

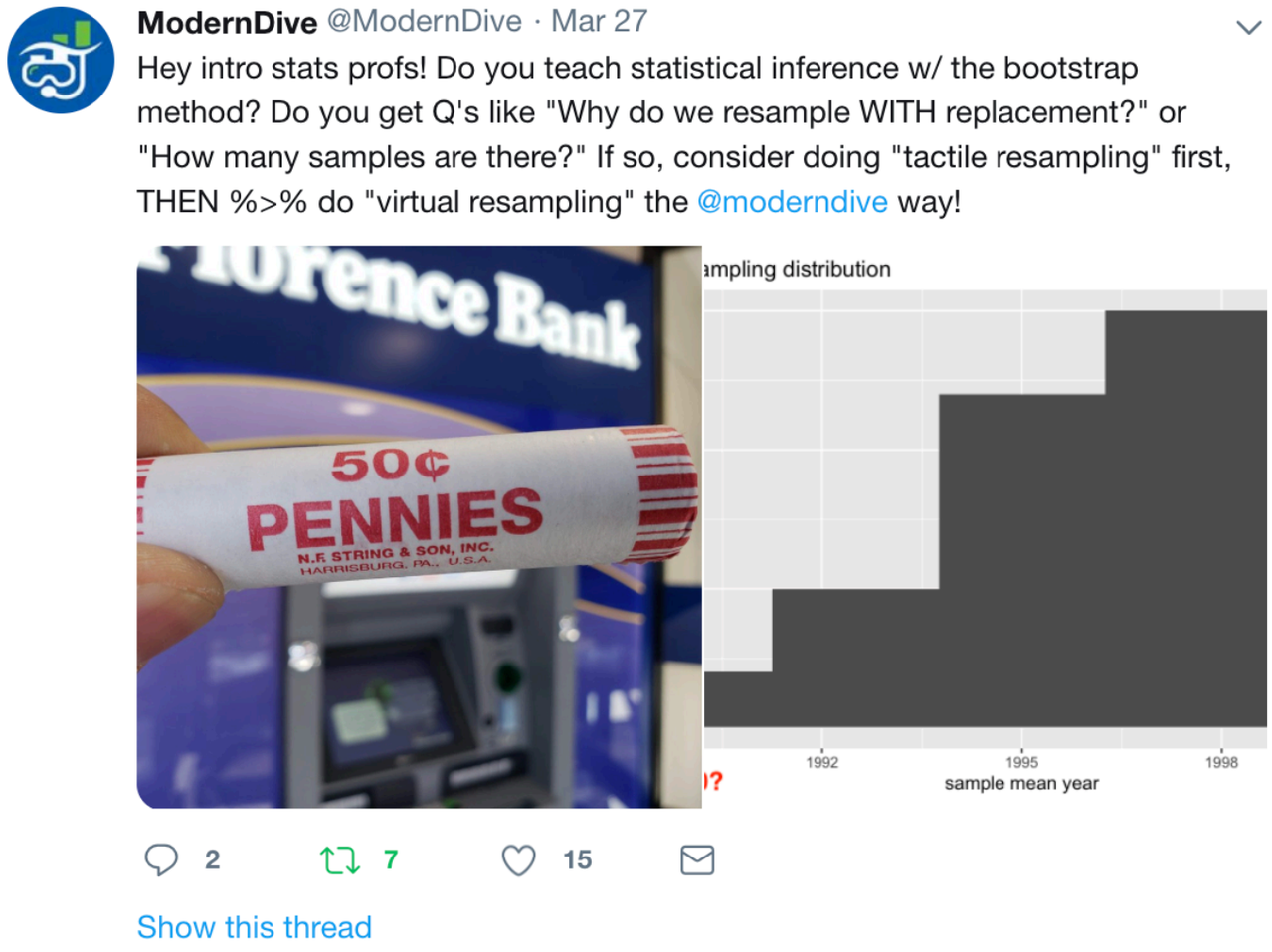
Then this… Then generalize & transfer…

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	μ	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	β_1	Sample regression slope	$\widehat{\beta}_1$ or b_1
6	Population regression intercept	β_0	Sample regression intercept	$\widehat{\beta}_0$ or b_0

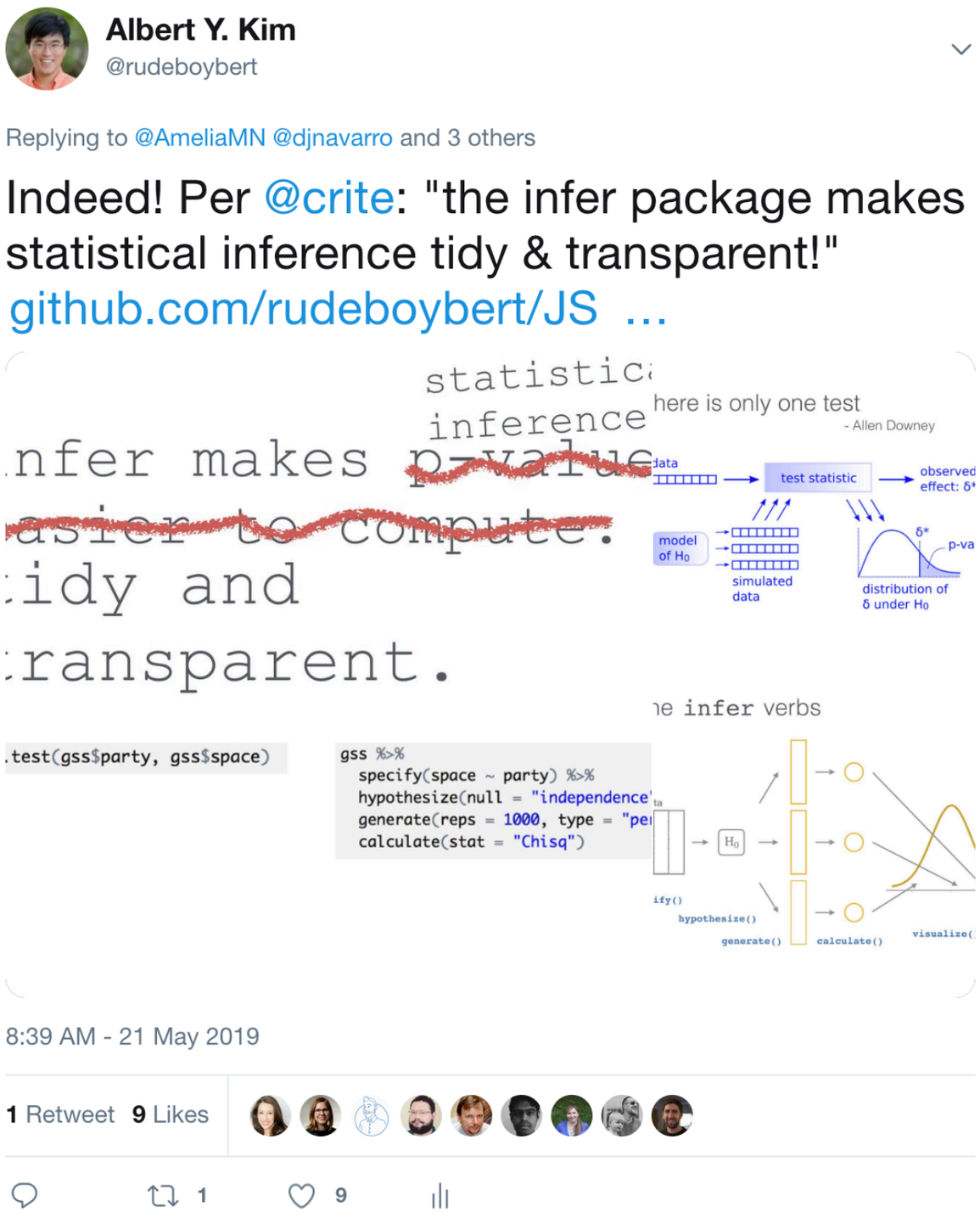
From moderndive Ch 8.5.2

Chap 9: Confidence Intervals

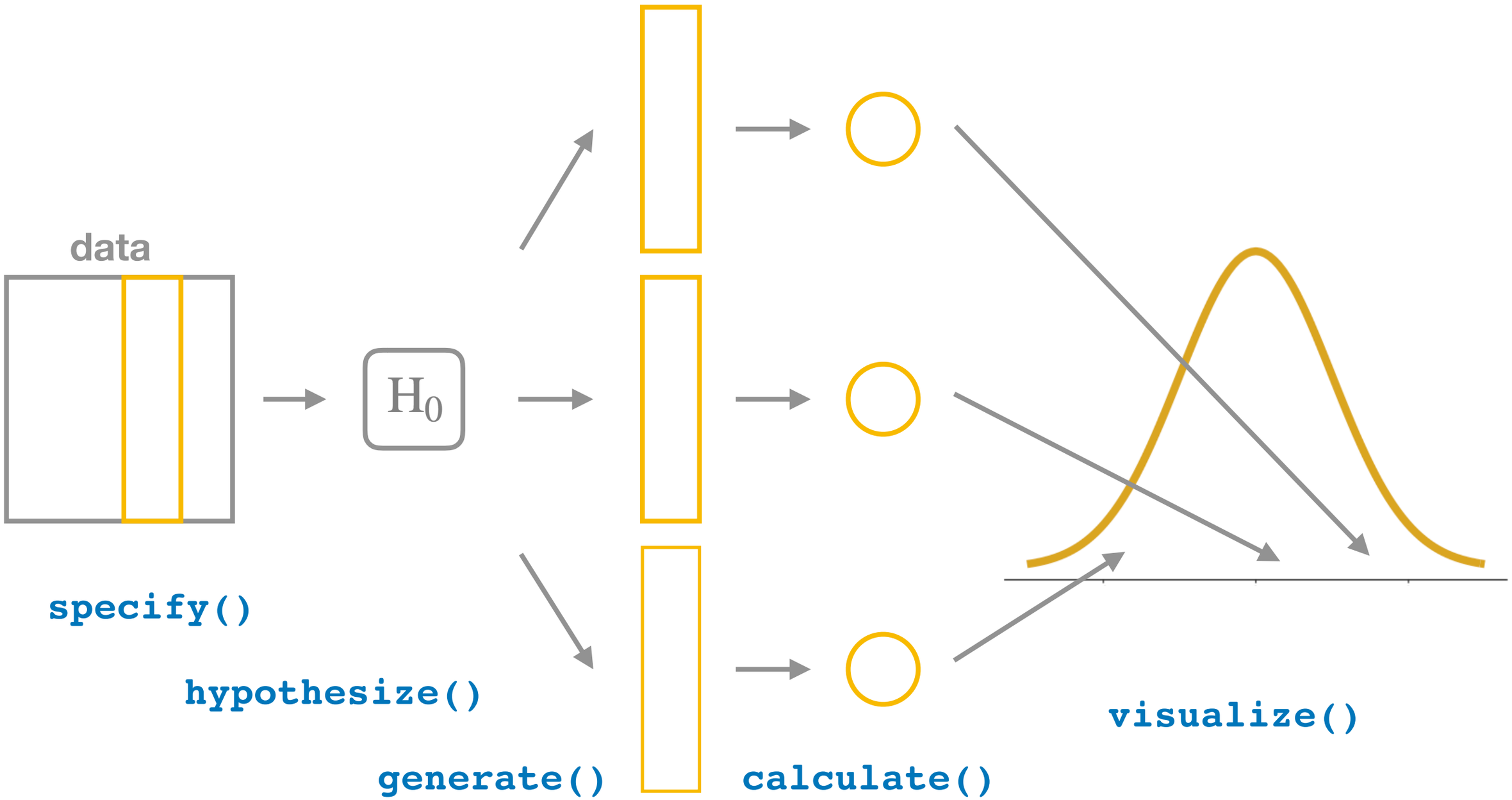


1. What are we doing ❓
   - Studying effect of sampling variation on estimates
   - Studying effect of sample size on sampling variation
2. Why are we doing this 🤔
   - So students don't get lost in abstraction & never lose 👀 on what statistical inference is about.

Chap 10: Hypothesis Testing via `infer`



Hypothesis Testing



`infer` package

- Live code demo of constructing null distribution
- Comparing *the what* vs *the how*
  - The what is the same as Rossman/Chance applets & StatKey by Lock5
  - *The how* is different: "Getting under the hood" via `tidyverse`
- More on *the what*
  - Convincing students there is only one test
  - Bridging gap with traditional formula-based methods/approximations. Ex: Central Limit Theorem

Goal 1: Modeling with Regression

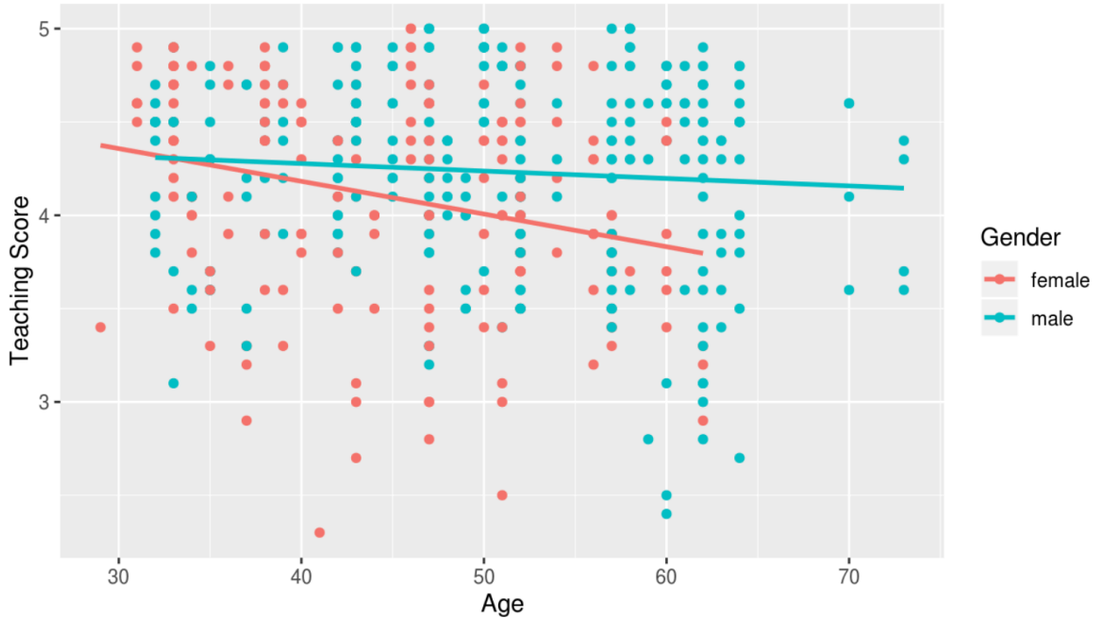
1. Data: `evals`

evals

Filter

	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

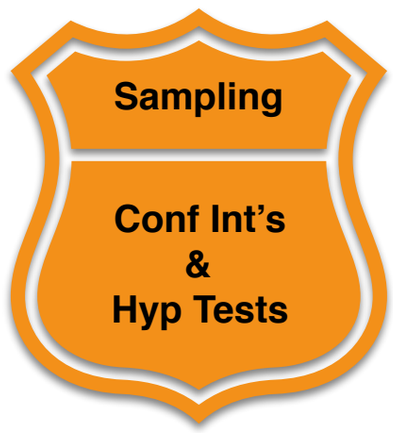
2. Exploratory Data Analysis



3. Regression Coeff

```
Console ~/
> 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



4. Regression Table

```
Console ~/
> 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

Regression `wrapper()` functions

Console ~/

```
> library(tidyverse)
> library(moderndive)
> # Convert to tibble
> mtcars <- mtcars %>%
+   as_tibble(rownames_to_column(mtcars))
> # Fit lm
> mpg_model <- lm(mpg ~ hp, data = mtcars)
> # Two options
> summary(mpg_model)

Call:
lm(formula = mpg ~ hp, data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7121 -2.1122 -0.8854  1.5819  8.2360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.09886    1.63392  18.421  < 2e-16 ***
hp          -0.06823    0.01012  -6.742 1.79e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.863 on 30 degrees of freedom
Multiple R-squared:  0.6024,    Adjusted R-squared:  0.5892
F-statistic: 45.46 on 1 and 30 DF,  p-value: 1.788e-07

> get_regression_table(mpg_model)
# A tibble: 2 x 7
  term      estimate std_error statistic p_value lower_ci upper_ci
  <chr>        <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept   30.1       1.63      18.4        0   26.8      33.4
2 hp          -0.068     0.01      -6.74       0   -0.089    -0.048
>
```

summary() encourages p-value stargazing!

Why not a tibble w/ CI's?

Regression `wrapper()` functions

ModernDive @ModernDive · Mar 13

"Hold up, isn't that just broom::tidy()?" You betcha! But we made things novice friendly by renaming everything, even the function names! Lay 👀 on the get_regression_points() wrapper to broom::augment()!

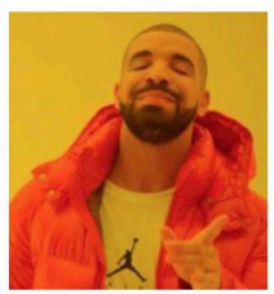
Make partial residual plots from scratch instead of w/ plot.lm()!

Console ~/

```
> library(broom)
> augment(mpg_model)
# A tibble: 32 x 9
     mpg    hp .fitted .se.fit .resid    .hat .sigma   .cooksd .std.resid
   <dbl> <dbl>   <dbl>   <dbl>  <dbl>   <dbl>  <dbl>     <dbl>      <dbl>
 1  21     110    22.6   0.777 -1.59  0.0405    3.92 0.00374       -0.421
 2  21     110    22.6   0.777 -1.59  0.0405    3.92 0.00374       -0.421
 3  22.8    93    23.8   0.873 -0.954 0.0510    3.92 0.00173       -0.253
 4  21.4   110    22.6   0.777 -1.19  0.0405    3.92 0.00210       -0.315
 5  18.7   175    18.2   0.741  0.541 0.0368    3.93 0.000389       0.143
 6  18.1   105    22.9   0.803 -4.83  0.0432    3.82 0.0369        -1.28
 7  14.3   245    13.4   1.21   0.917 0.0976    3.92 0.00338        0.250
 8  24.4    62    25.9   1.10  -1.47  0.0805    3.92 0.00688       -0.396
 9  22.8    95    23.6   0.860 -0.817 0.0496    3.93 0.00123       -0.217
10  19.2   123    21.7   0.724 -2.51  0.0351    3.90 0.00794       -0.661
# … with 22 more rows
> get_regression_points(mpg_model, ID = "rowname")
# A tibble: 32 x 5
   rowname             mpg    hp mpg_hat residual
   <chr>             <dbl> <dbl>   <dbl>    <dbl>
 1 Mazda RX4          21     110    22.6   -1.59
 2 Mazda RX4 Wag      21     110    22.6   -1.59
 3 Datsun 710         22.8    93    23.8   -0.954
 4 Hornet 4 Drive     21.4   110    22.6   -1.19
 5 Hornet Sportabout  18.7   175    18.2    0.541
 6 Valiant            18.1   105    22.9   -4.84
 7 Duster 360         14.3   245    13.4    0.917
 8 Merc 240D          24.4    62    25.9   -1.47
 9 Merc 230           22.8    95    23.6   -0.817
10 Merc 280           19.2   123    21.7   -2.51
# … with 22 more rows
>
```

Gah! What are these names? 😱 😱 😱

"Ahh! Much better!"



1 9

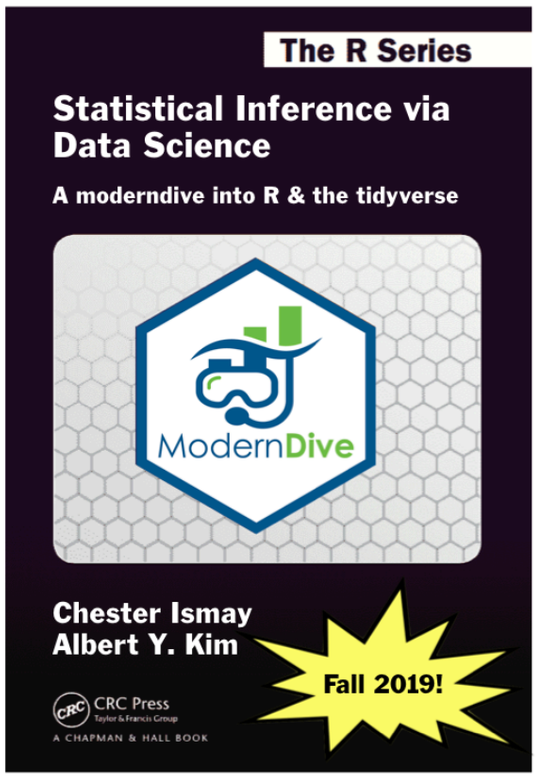
Conclusion

Resources

- Two versions of moderndive
  1. Development (being edited): moderndive.netlify.com
  2. Latest release (updated x2 yearly): moderndive.com
- On GitHub at github.com/moderndive/
  1. **bookdown** source code for book
  2. **moderndive** package source code
- Course webpage from Spring 2019
- moderndive mailing list: eepurl.com/cBkItf

Timeline

- **Now:** Development version on moderndive.netlify.com being edited:
  - Ch9 on CI, Ch10 on HT need cleaning
  - 🚧Ch11 on inference for regression 🚧
- **Late-June**: Preview of print edition available on moderndive.com
- **Late-July**: Posting labs/problems sets & example final project samples
- **Fall 2019**: Print edition available!



Thank you!