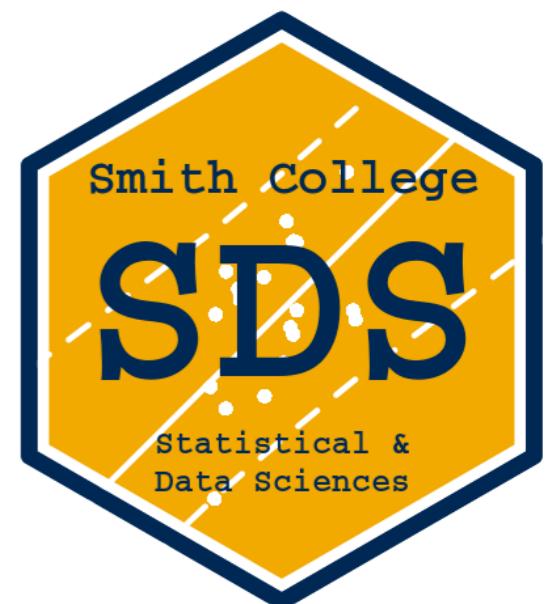


moderndive: statistical inference via the tidyverse



Albert Y. Kim
[@rudeboybert](https://twitter.com/rudeboybert)

Cal Poly Statistics Department
San Luis Obispo
May 23



My Co-Authors



Chester Ismay:
Textbook co-author

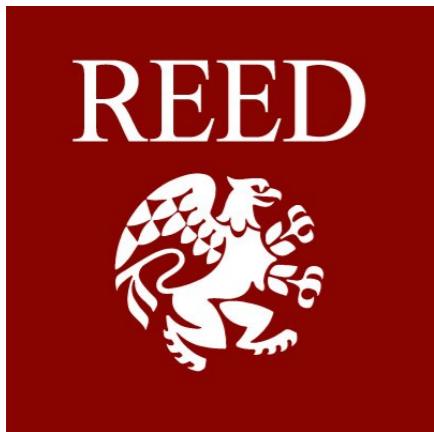


Jenny Smetzer:
Labs author

Background



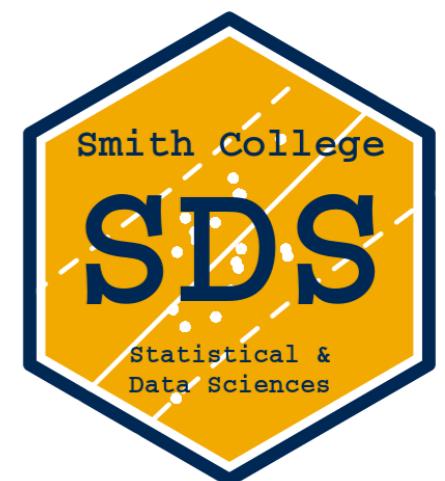
Google



Middlebury



AMHERST
COLLEGE



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

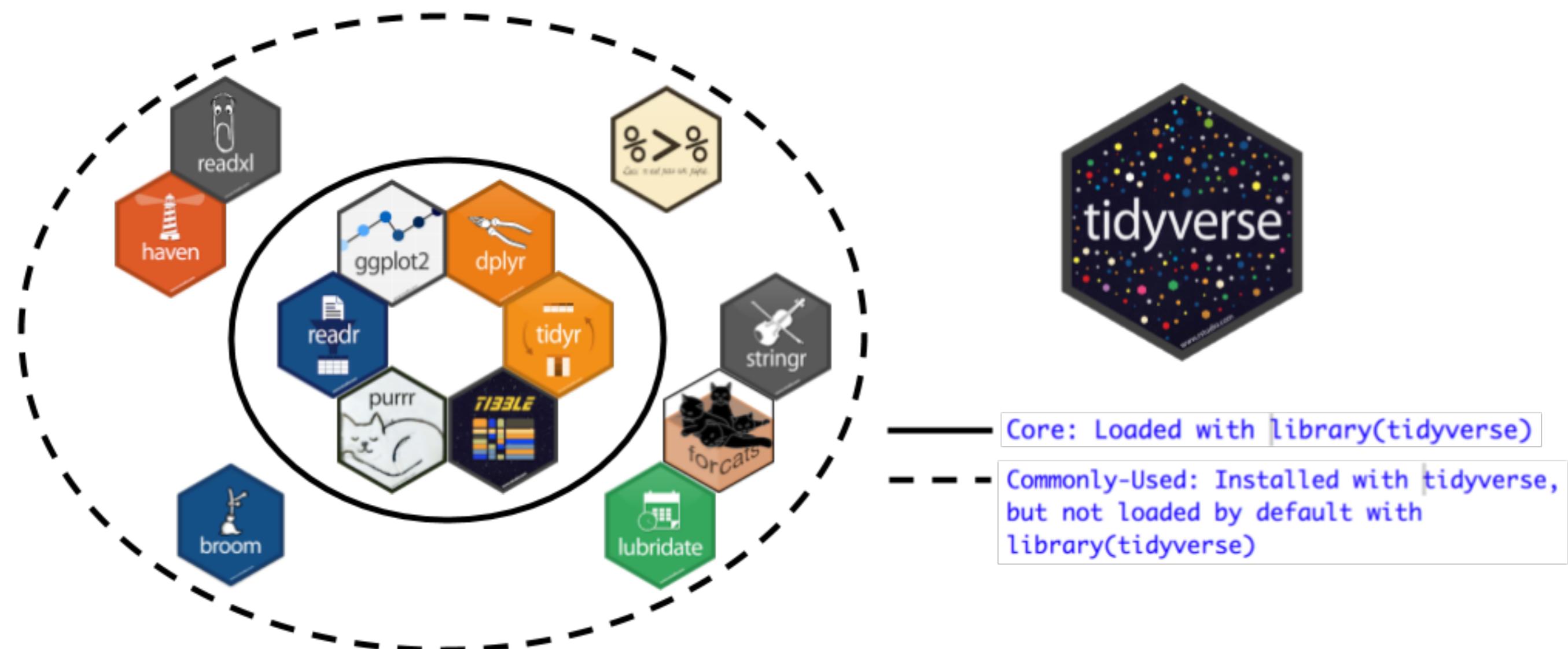
1. Modeling with regression
2. Sampling for inference



Calculus?
😁 thru 🤢

Coding?
😱 & 🤔

What is the tidyverse?



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

Why tidyverse in general?

From [tidy tools manifesto](#): Say what?

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

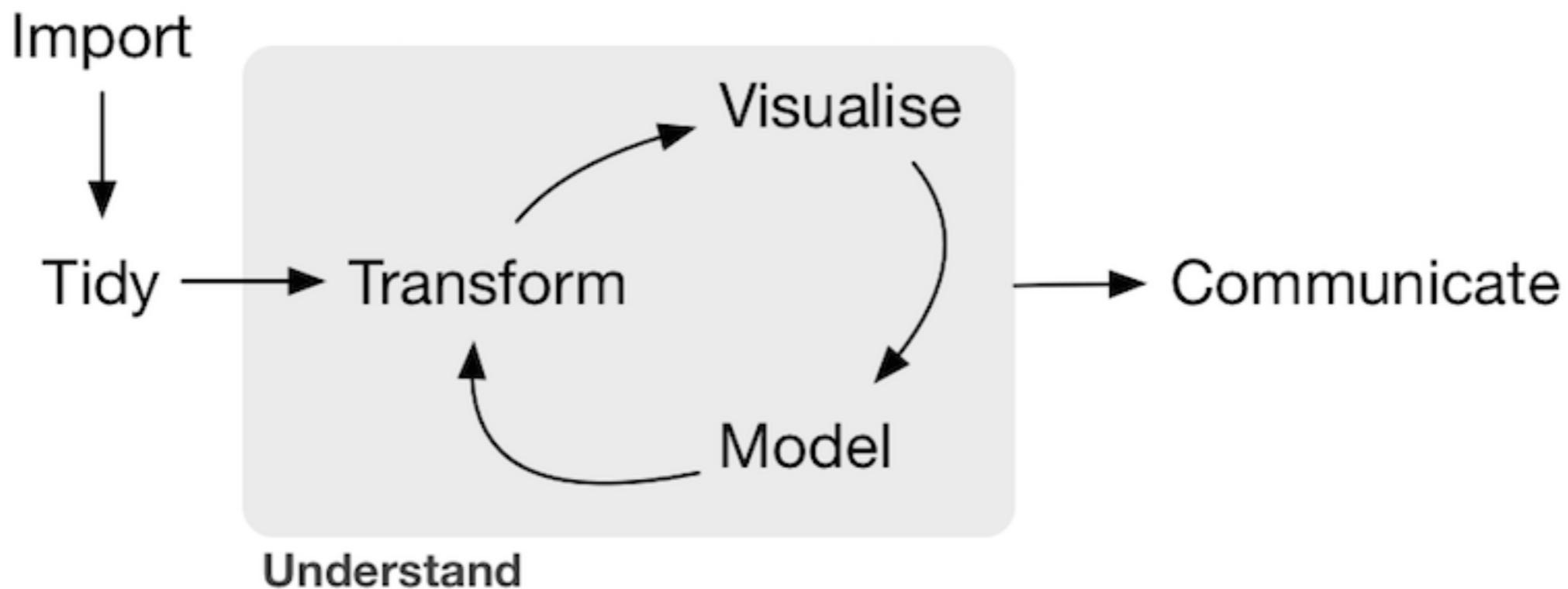
- IMO it's easier to learn than base R. [Others too.](#)
- It scales. You leverage an entire ecosystem of online developers and support: Google & StackOverflow
- Satisfy learning goals *while learning tools they can use beyond the classroom.*

Why tidyverse for stats newbies?

- 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 layers of complexity...
 - %>% add layers of complexity...
- Do this instead of learning individual components in isolation

End Deliverable

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.

The R Series

Statistical Inference via Data Science

A moderndive into R & the tidyverse



**Chester Ismay
Albert Y. Kim**

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Fall 2019!

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



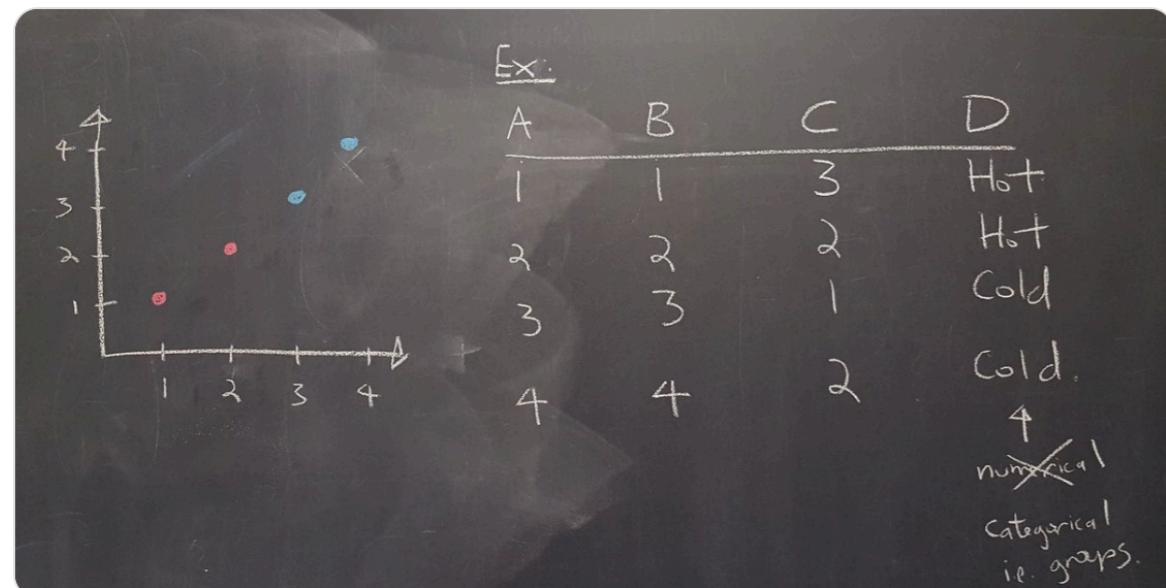
- IMO RStudio's best function: `View()`
- Getting students over initial 😱 of coding

Chapter 3: Data Viz via ggplot2



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

Recall:

A statistical graphic is a mapping of data variables to aesthetic attributes of geometric objects.

Five Named Graphs 5NG

- (1) Scatterplots geom_point()
- (2) Linegraphs geom_line()
- (3) Histograms geom_histogram()
- (4) Boxplots geom_boxplot()
- (5) Barplots geom_bar()

12:50 PM - 12 Sep 2017 from Amherst College

15 Retweets 61 Likes



3 15 61

Chapter 4: Data Wrangling via dplyr

Chapter 5: “Tidy” Data via tidyverse

- Essential: `%>%` operator. Needed later.
- Balance of how much students do vs how much you curate yourself?
- To *completely* shield students from *any* data wrangling is to betray [true nature of work in our fields](#)
- One proposed balance point in [“tame” data & fivethirtyeight package](#) paper

Part II: Data Modeling via moderndive

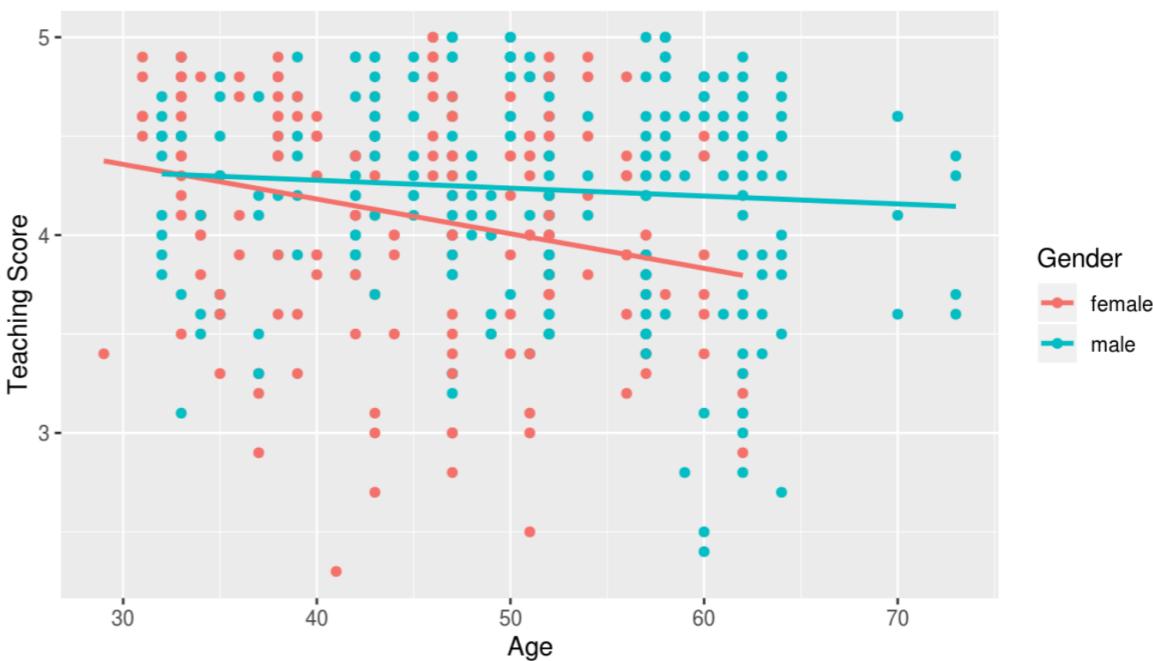
Chapters 6, 7, & 11

Goal 1: Modeling with Regression

1. Data: evals

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

More later!

Early: Descriptive regression

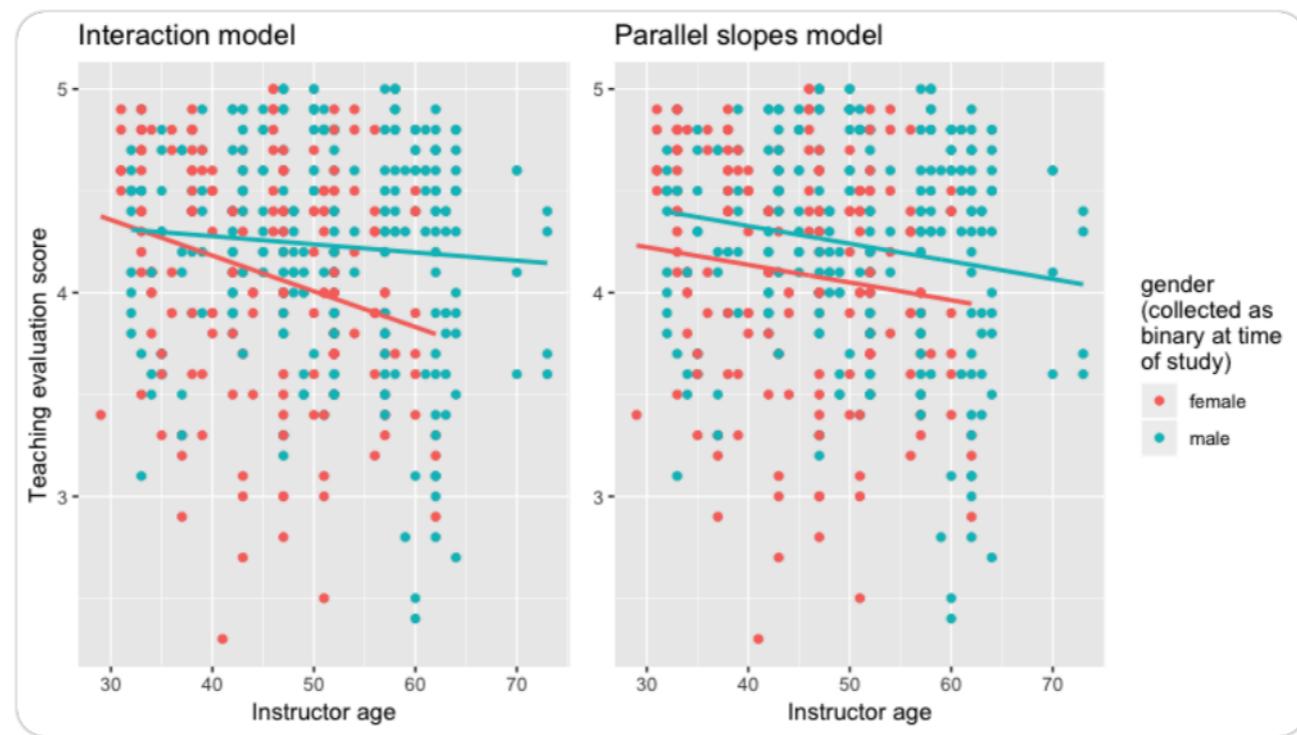
Including model selection...

- 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

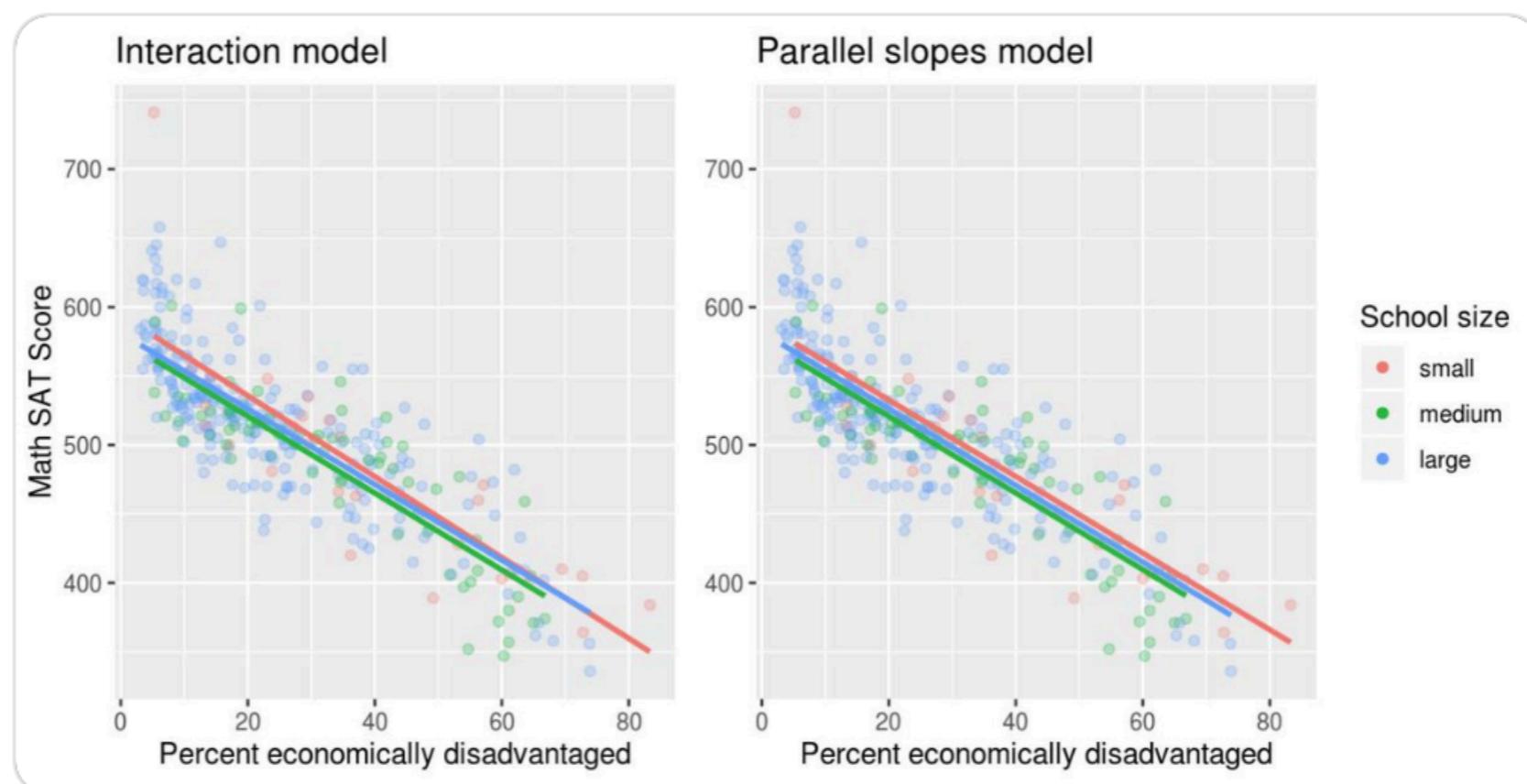


Including model selection...



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

1



ModernDive @ModernDive · Apr 19

Then ask them to consider:

"Is the extra 😬 COMPLEXITY 😬 of the interaction model warranted?"

or

"Should I favor the 😊 SIMPLER 😊 parallel slopes model?" 3/4

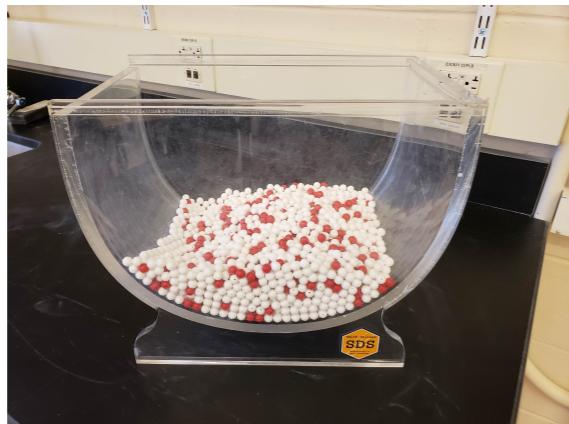
Part III: Statistical Inference via infer

Chapters 8 - 11

Goal 2: Sampling for Inference

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

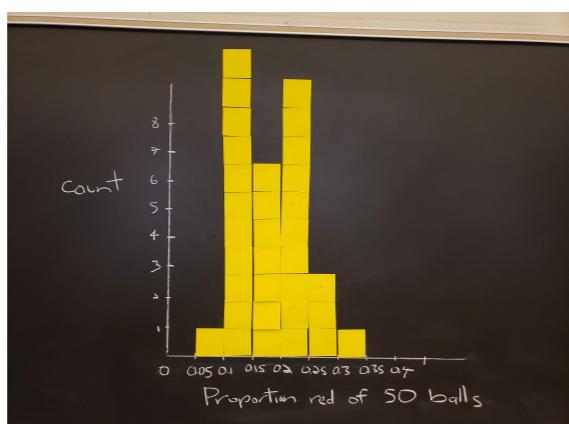
Population



Sample

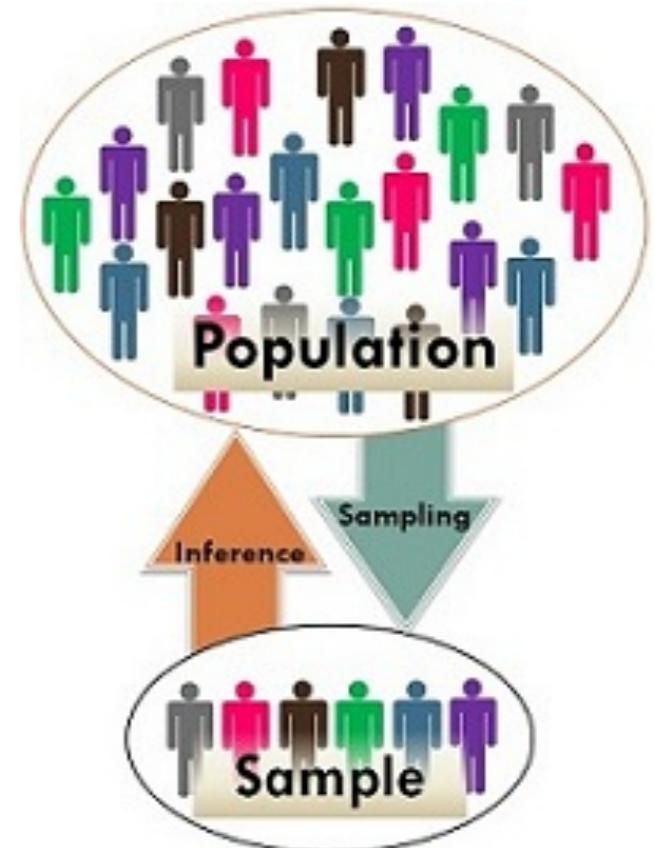
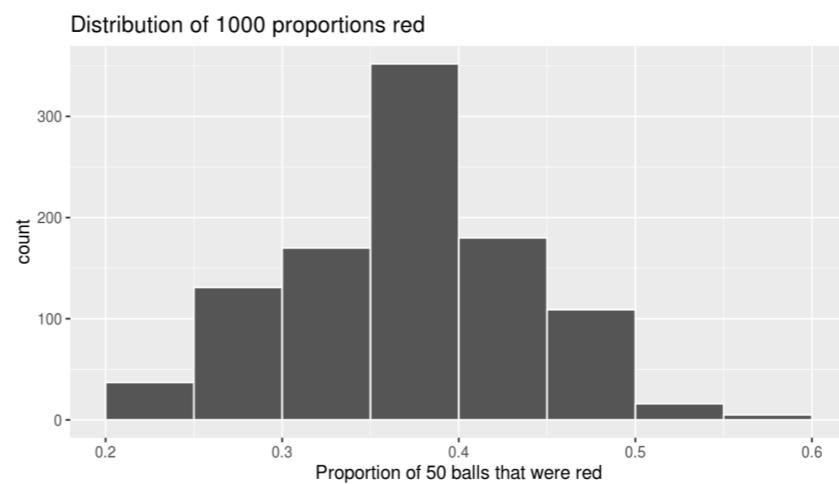


Sampling
Distributions &
Standard Errors



```
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
> |
```

```
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     1  white
2       1     1  red
3       1     1  white
4       1     1  white
5       1     1  white
6       1     1  white
7       1     1  white
8       1     1  white
9       1     1  red
10      1     1  white
# ... with 40 more rows
> |
```



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

Chapter 8: Sampling

Terminology, definitions, & notation 😱

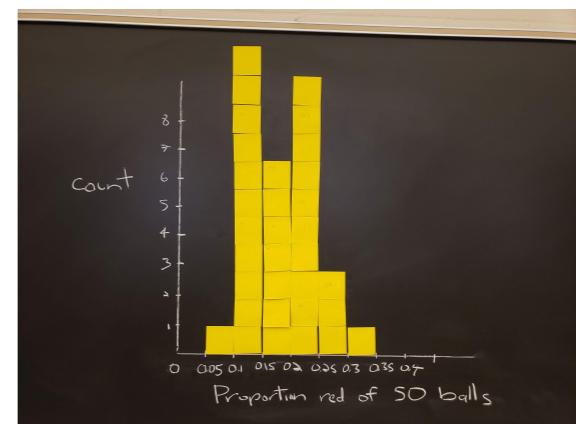
[isostat] Is notation and language a barrier to students learning introductory statistics?



▶ Statistics/ISOSTAT ×

- 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
- 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
- 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/>
- 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
- 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

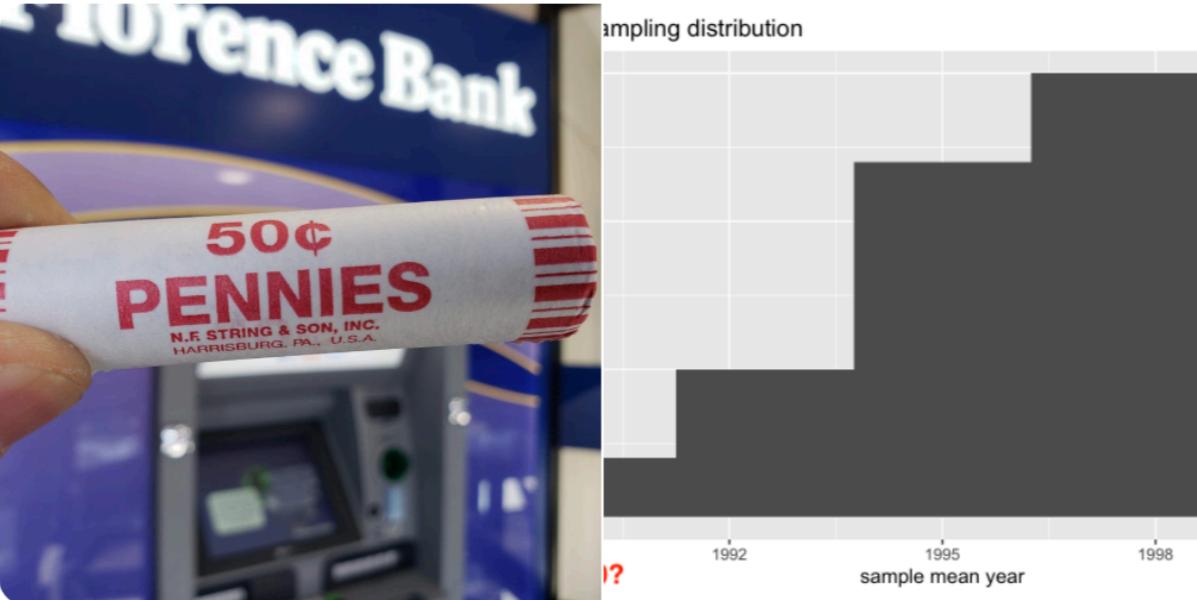
Scenario	Population parameter	Notation	Point estimate	Notation.
1	Population proportion	p	Sample proportion	\hat{p}
2	Population mean	μ	Sample mean	$\hat{\mu}$ or \bar{x}
3	Difference in population proportions	$p_1 - p_2$	Difference in sample proportions	$\hat{p}_1 - \hat{p}_2$
4	Difference in population means	$\mu_1 - \mu_2$	Difference in sample means	$\bar{x}_1 - \bar{x}_2$
5	Population regression slope	β_1	Sample regression slope	$\hat{\beta}_1$ or b_1
6	Population regression intercept	β_0	Sample regression intercept	$\hat{\beta}_0$ or b_0

From moderndive Ch 8.5.2

Chap 9: Confidence Intervals

ModernDive @ModernDive · Mar 27

Hey intro stats profs! Do you teach statistical inference w/ the bootstrap method? Do you get Q's like "Why do we resample WITH replacement?" or "How many samples are there?" If so, consider doing "tactile resampling" first, THEN %>% do "virtual resampling" the [@moderndive](#) way!



The image contains two parts: a photograph of a hand holding a roll of pennies labeled '50¢ PENNIES N.F. STRING & SON, INC. HARRISBURG, PA., U.S.A.' in front of a bank ATM, and a histogram titled 'Sampling distribution' showing the distribution of sample mean years. The x-axis is labeled 'sample mean year' with ticks at 1992, 1995, and 1998. The distribution is skewed right, with the highest frequency in the bin around 1996.

2 7 15

Show this thread

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`



Albert Y. Kim

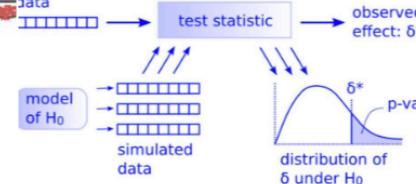
@rudeboybert



Replying to @AmeliaMN @djnavarro and 3 others

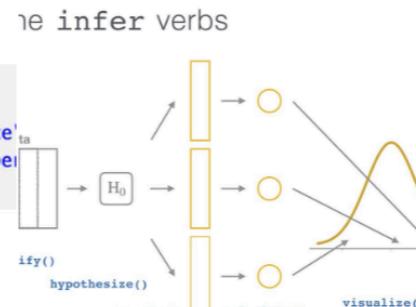
Indeed! Per [@crite](#): "the `infer` package makes statistical inference tidy & transparent!"
[github.com/rudeboybert/JS ...](https://github.com/rudeboybert/JS...)

statistic
inference
here is only one test
- Allen Downey
`infer` makes p-value
easier to compute.
tidy and
transparent.



.test(gss\$party, gss\$space)

```
gss %>%  
  specify(space ~ party) %>%  
  hypothesize(null = "independence")  
  generate(reps = 1000, type = "perm")  
  calculate(stat = "Chisq")
```

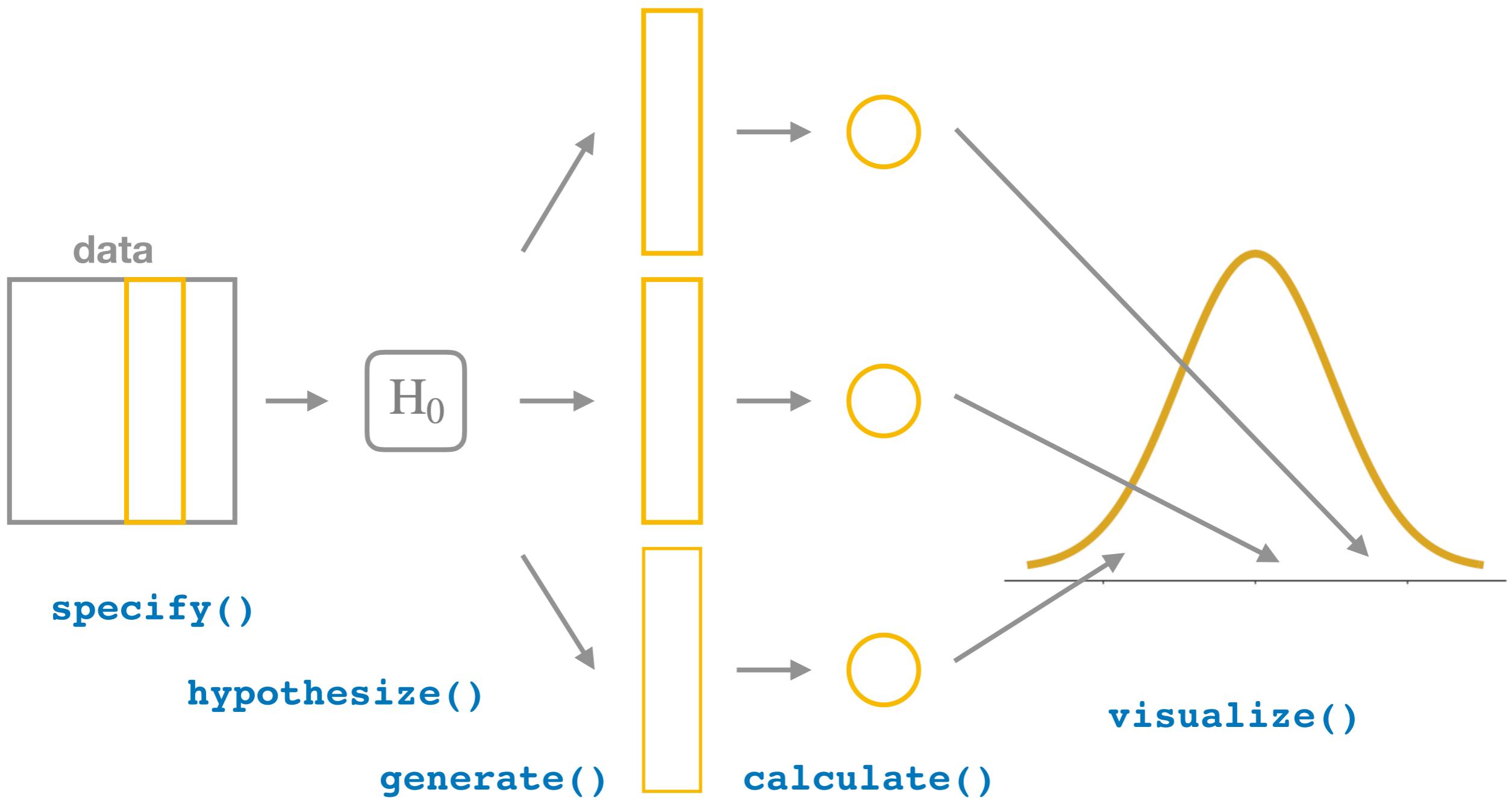


8:39 AM - 21 May 2019

1 Retweet 9 Likes



Hypothesis Testing



infer package

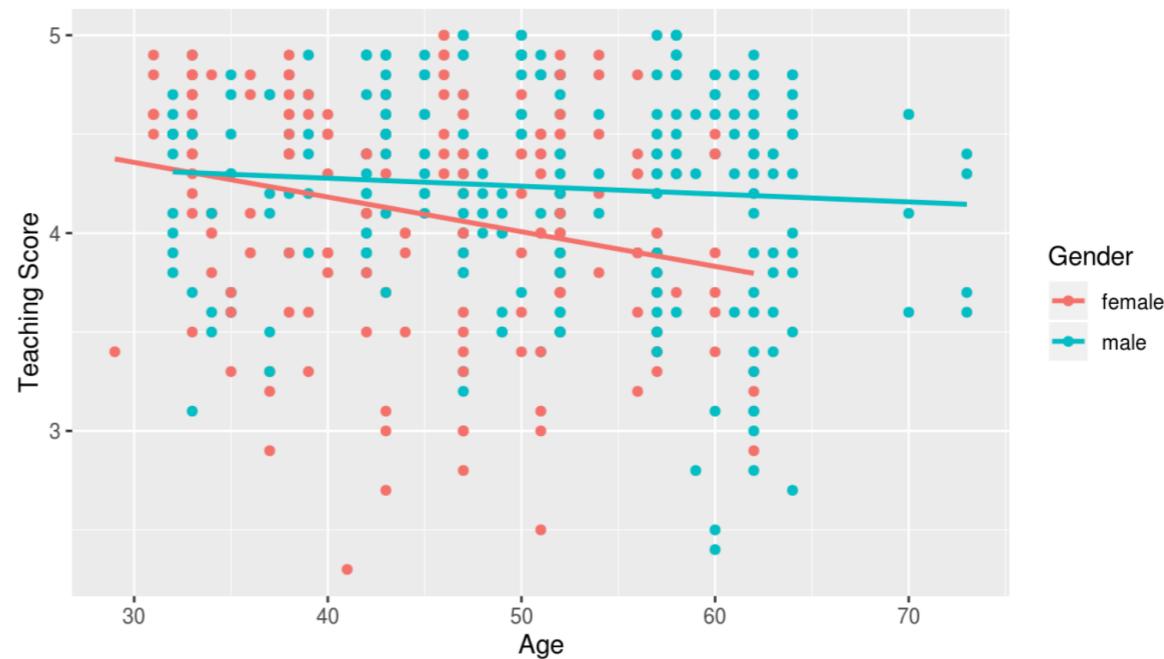
- Live [code demo](#)
- Comparing *the what* vs *the how*
 - The *what* is the same as Rossman/Chance [applets](#) & [StatKey](#) by Lock5
 - *The how* is different: “Getting under the hood” via **tidyverse**
- More on *the what*
 - Convincing students [there is only one test](#)
 - [Bridging gap](#) with traditional formula-based methods/approximations. Ex: Central Limit Theorem

Goal 1: Modeling with Regression

1. Data: evals

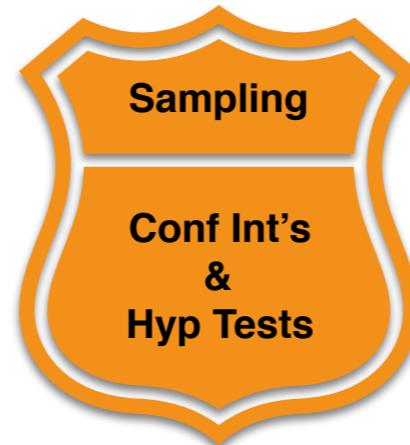
ID	score	age	gender
1	4.7	36	female
2	4.1	36	female
3	3.9	36	female
4	4.8	36	female
5	4.6	59	male
6	4.3	59	male
7	2.8	59	male
8	4.1	51	male
9	3.4	51	male
10	4.5	40	female
11	3.8	40	female
12	4.5	40	female

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
> |
```



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
> |
```

Early: Descriptive regression

Later: Inference for Regression

A few 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)
```

A few 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()!

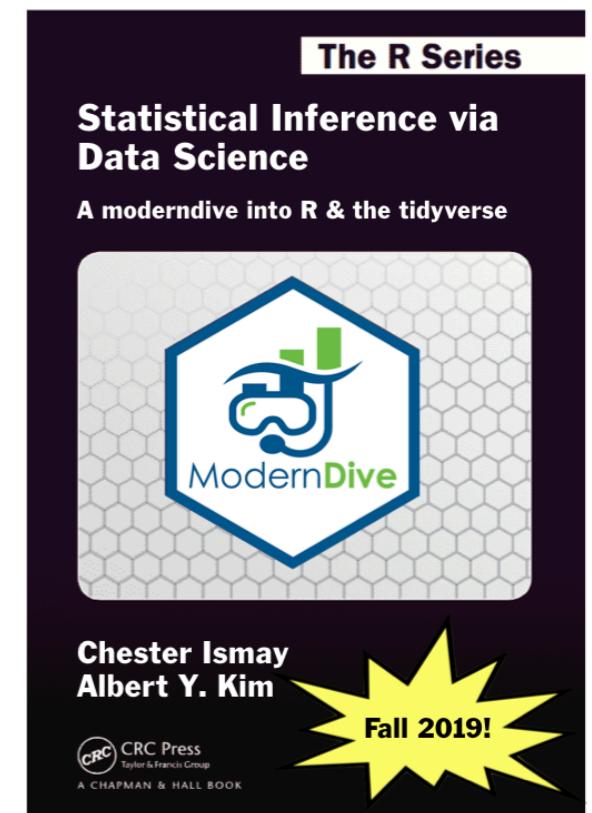
Conclusion

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/cBkltf

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!

Starting Small: Some Suggestions

- Ch6: Use `get_regression_table()` instead of `summary()`
- Ch5 + Ch2: Publish (non-sensitive) data to .csv via Google Sheets and import with `read_csv()`.
- Ch3: Spend time covering [Grammar of Graphics](#) & do all plots in course via `ggplot2`
- Ch8 + Ch5 + Ch3: Use data frame + `%>%` + `rep_sample_n()` to make a visualization of a sampling distribution from scratch!
- Ch5: Have them do an EDA via `group_by()` `%>%` `summarize()` to get two means + two-sample t-test
- Ch3 + Ch5 + Ch10: Jump straight into `infer` package