moderndive: statistical inference via the tidyverse

Albert Y. Kim
[@rudeboybert]

Statistical Society of Canada
Calgary, Alberta
May 29, 2019

My Context for moderndive

My students:

- Undergraduate-only women's 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

😐

Calculus?
😬 thru 🤮

Coding?
😱 & 🤔

via the
`tidyverse`

→

tidyverse

Point B:
Two goals

1. Modeling with regression
2. Sampling for inference

What is the tidyverse?

From [tidyverse.org](https://www.tidyverse.org):

R packages for data science

The tidyverse is an opinionated **collection of R packages** designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

Install the complete tidyverse with:

```
install.packages("tidyverse")
```

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.*

Typical Lecture

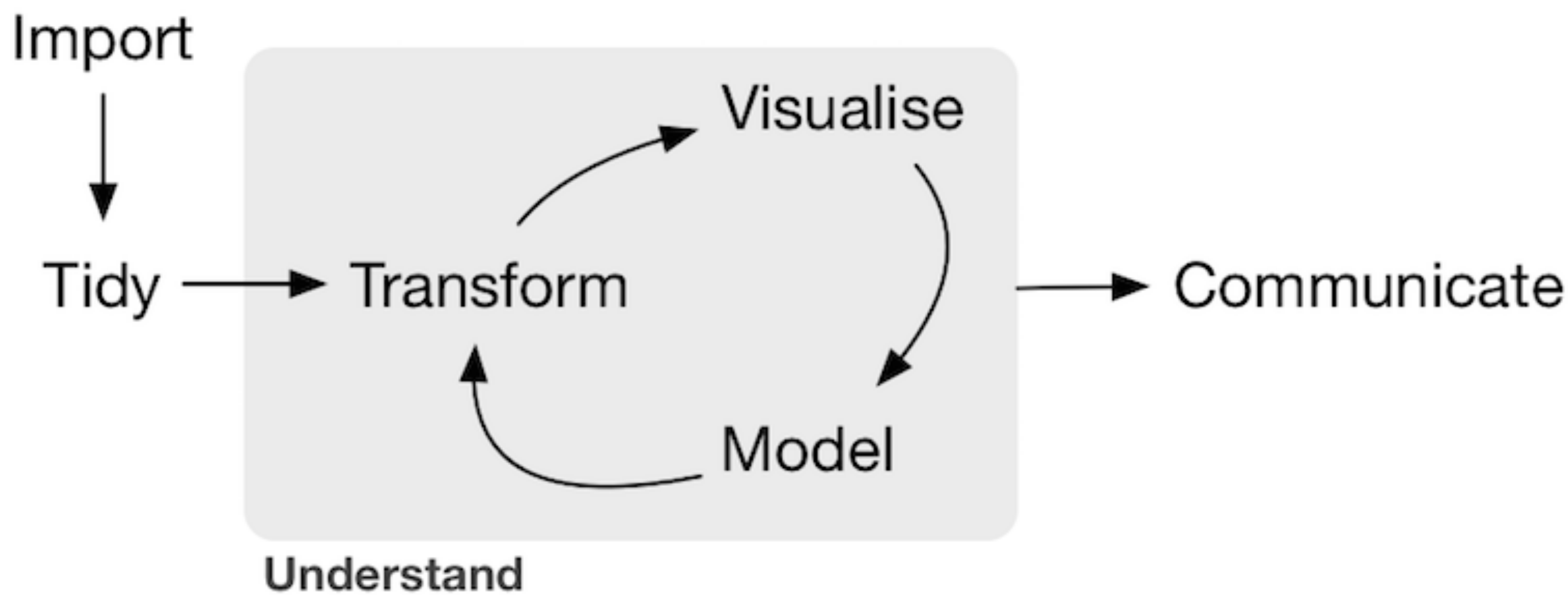
- 15-20min of lecturing + 50-55min of open exercises
- Borrow elements of “flipped classroom”: how to use time we’re all in the same room together?

End Deliverable of Course

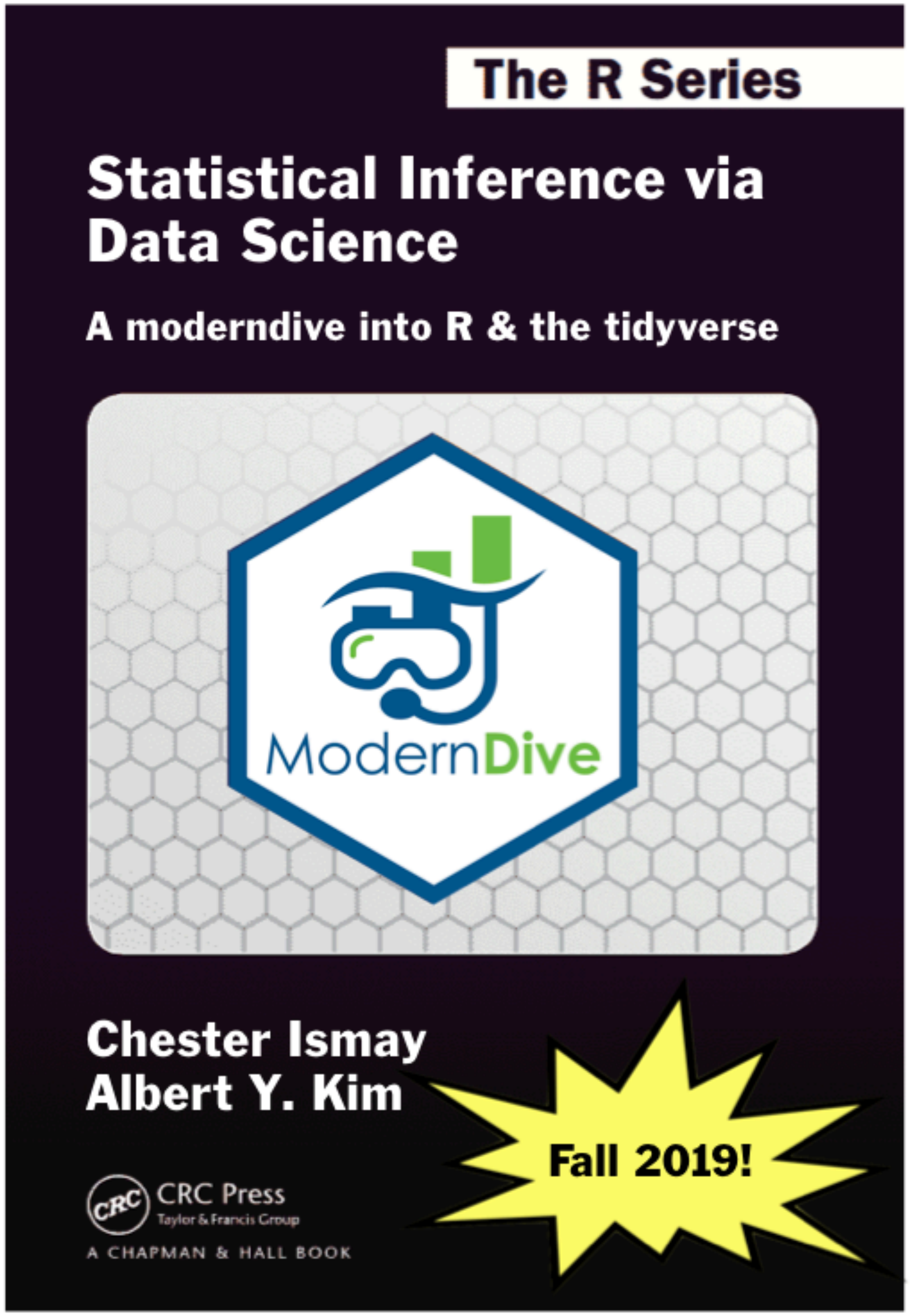
- Think of how youths learn to play sports…
- IMO stats newbies should learn to *"play the whole game"* in simplified form first
    - %>% then add layers of complexity…
    - %>% then add more layers of complexity…
    - %>% then add more layers of complexity…
- Do this instead of learning individual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



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
	GET IT ON Google Play Download on the App Store

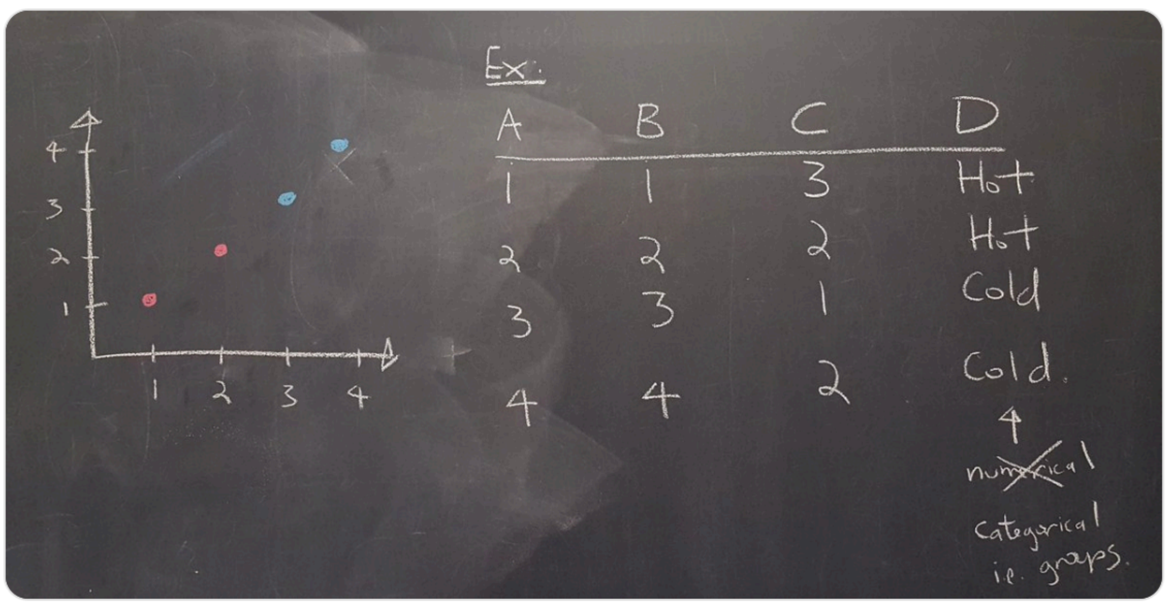
Getting students over initial 😱 of coding

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



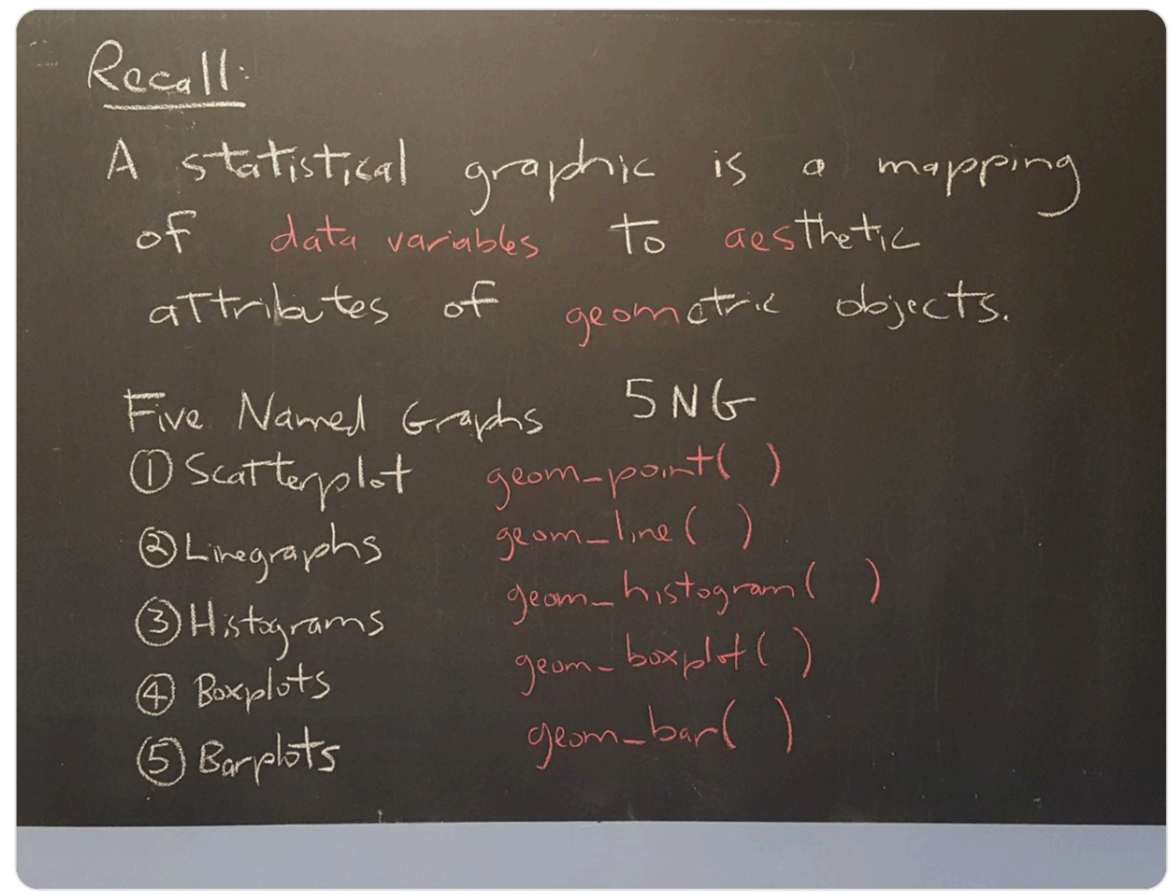
11:58 AM - 11 Sep 2017 from Amherst College

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.
- How much data "taming" is necessary?

Part II: Data Modeling via `moderndive`

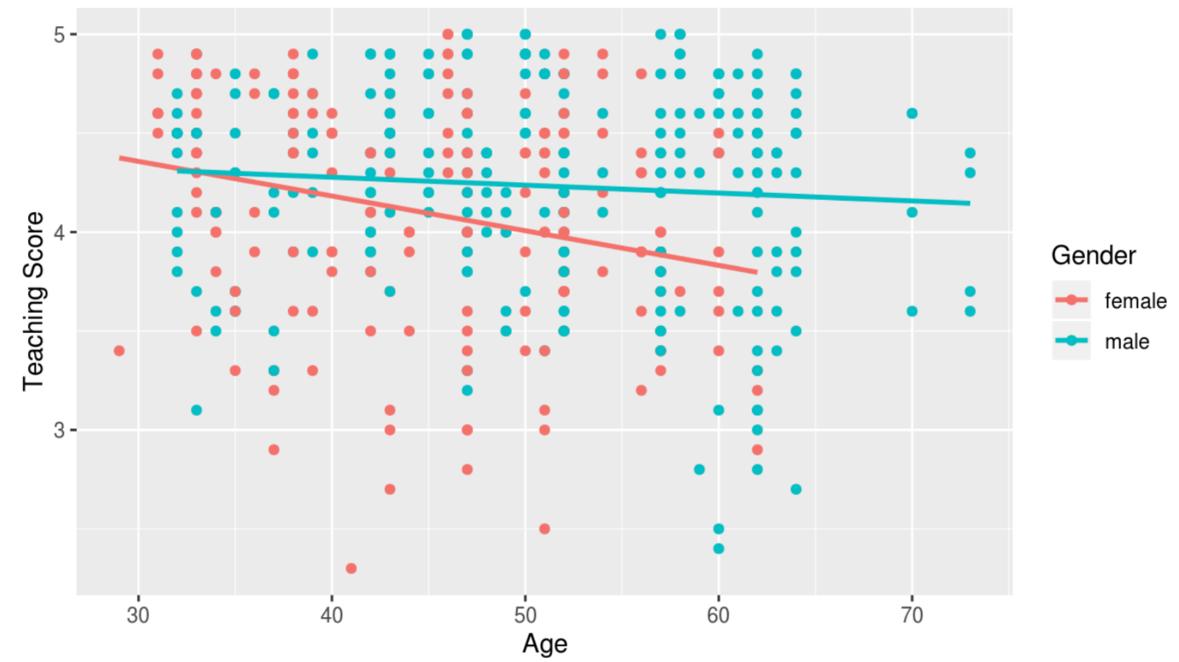
Chapters 6, 7, & 11

Goal 1: Modeling with Regression

1. Data: `evals`

	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

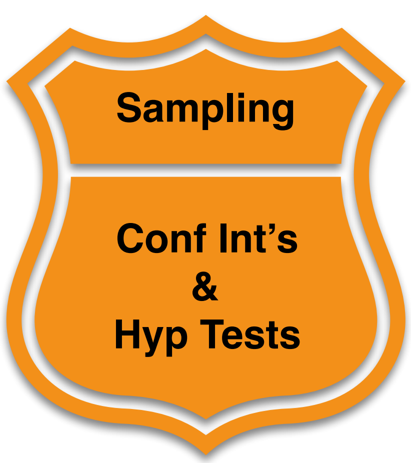
2. Exploratory Data Analysis



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



4. Regression Table

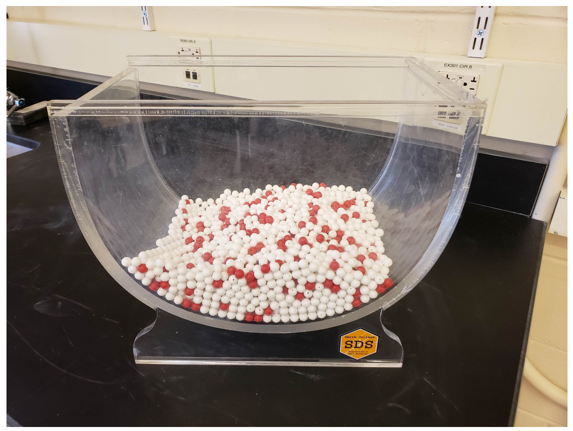
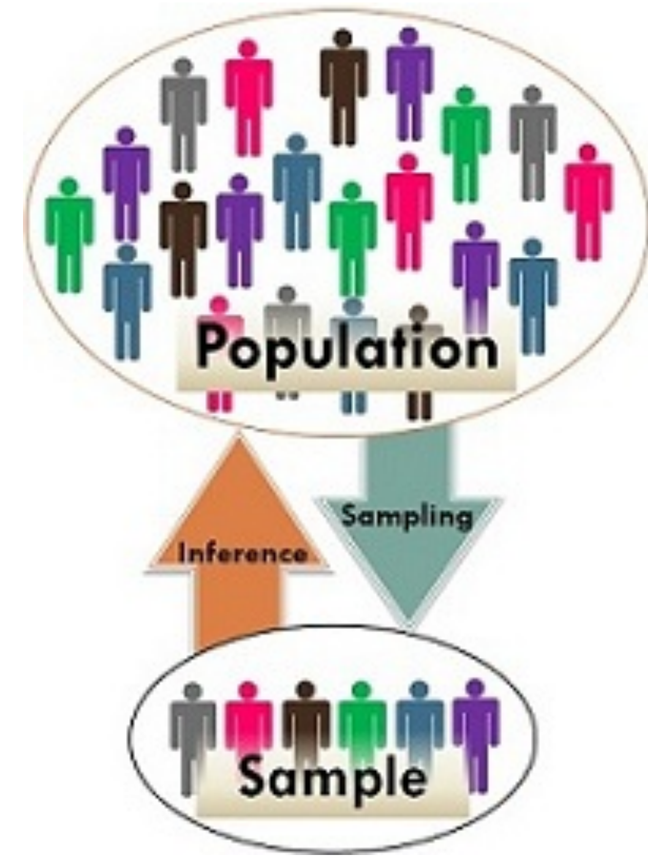
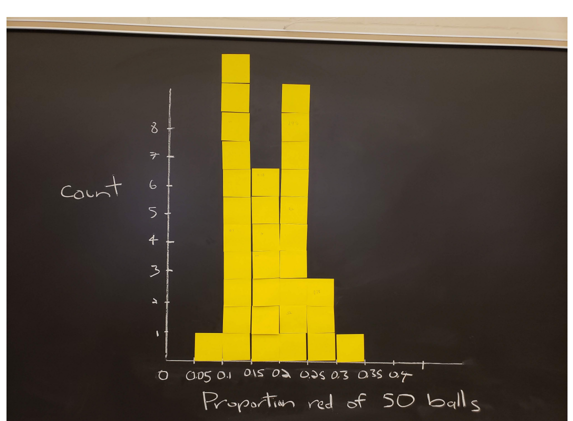
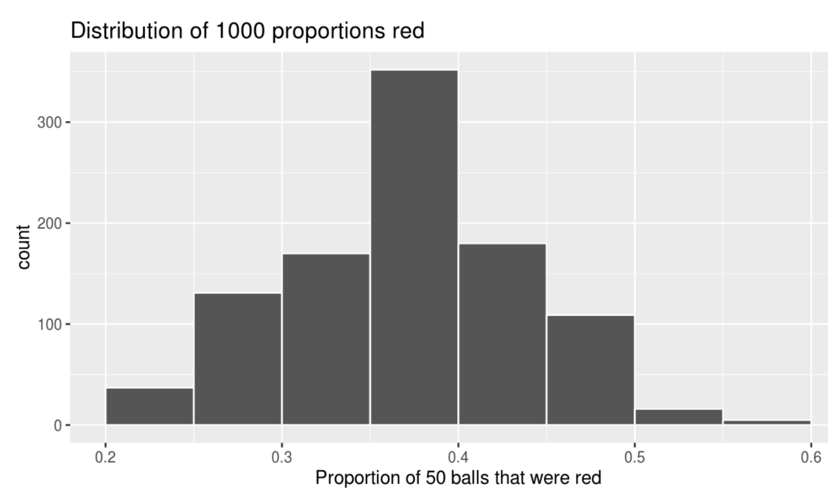
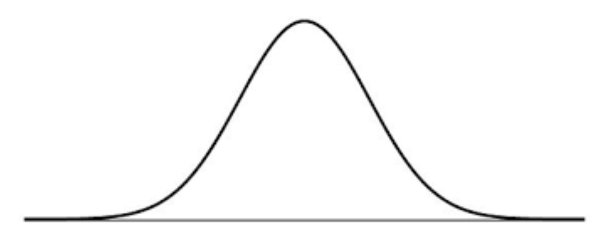
```
> score_model <- lm(score ~ age * gender, data = evals)
> get_regression_table(score_model)
# A tibble: 4 x 7
  term            estimate std_error statistic p_value lower_ci upper_ci
  <chr>              <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept           4.88     0.205     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

Part III: Statistical Inference via `infer`

Chapters 8 - 11

Goal 2: Sampling for Inference

	1. Tactile Sampling →	2. Virtual Sampling →	3. Theoretical
Population		see console 1 below	
Sample		see console 2 below	(see above)
Sampling Distributions & Standard Errors			

Console 1:

```
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 2:

```
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
>
```

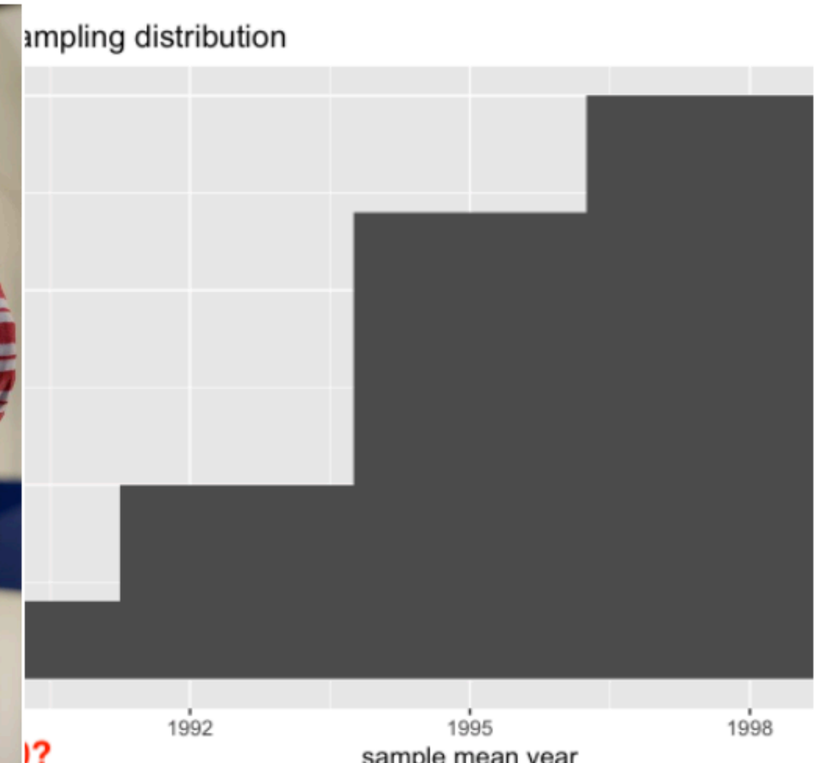
$$SE = \sqrt{\frac{p(1-p)}{n}}$$

Chap 9: Confidence Intervals

Click me!!!

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!

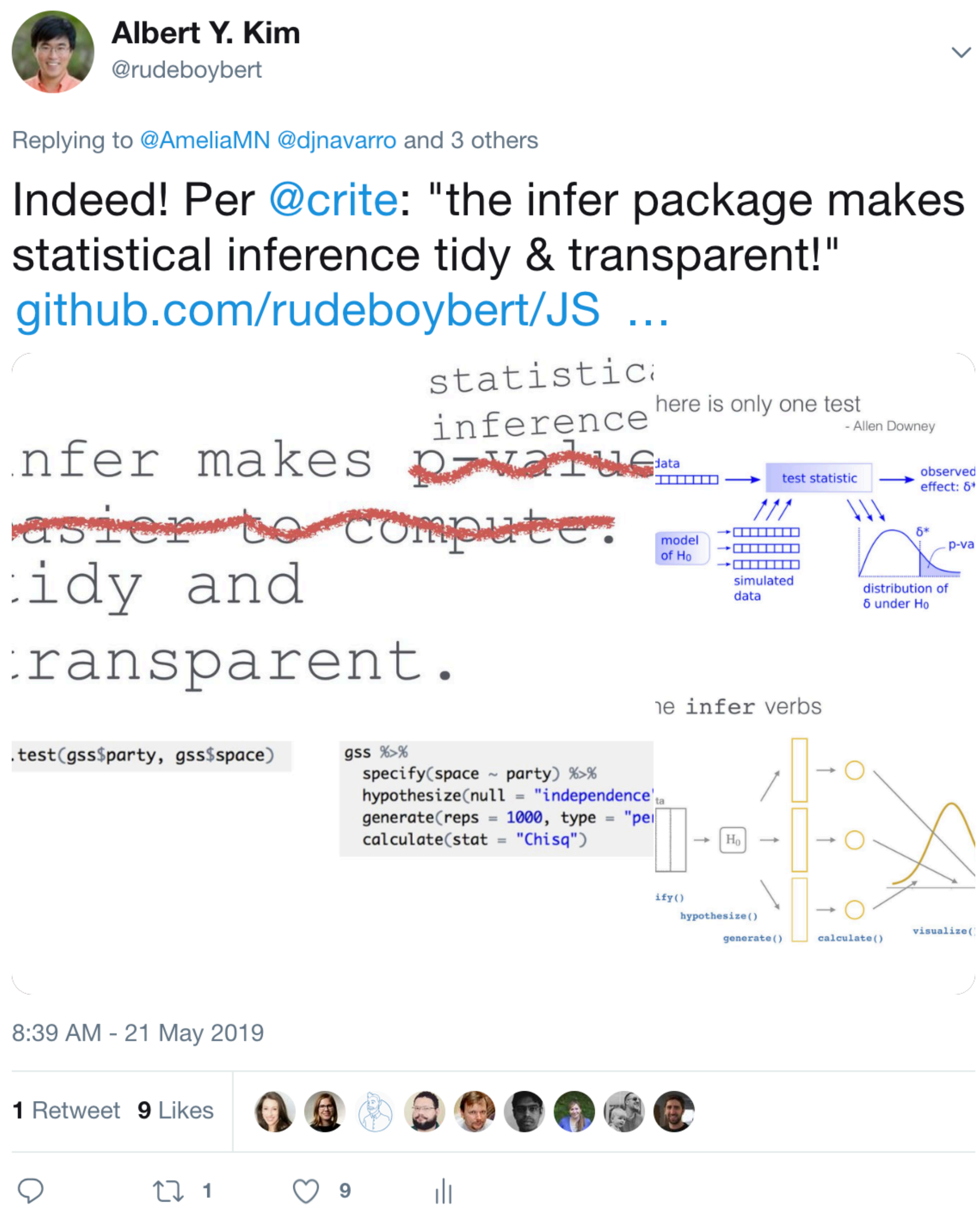


2 7 15

Show this thread

Chap 10: Hypothesis Testing via `infer`

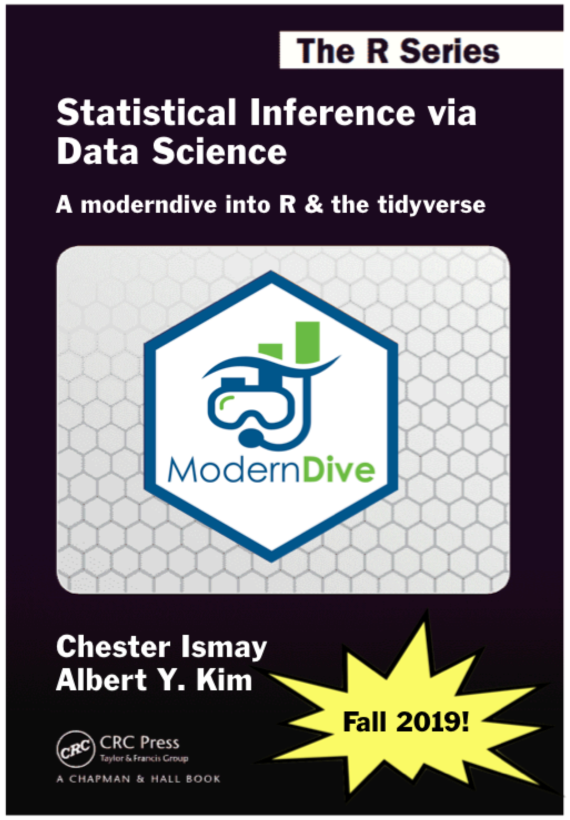
Click me!!!



Conclusion

Timeline

- **Now:** Development version on moderndive.netlify.com being edited.
- 🚧 Ch9-11 on CI, HT, & 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!



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

Thank you!

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