moderndive: statistical inference via the tidyverse

Albert Y. Kim | **Jordan Moody** | **Ziwei “Crystal” Zang** | **Starry Zhou**

USCOTS 2019
State College PA
May 15-16

About us!

also my collaborators…

Jenny Smetzer

Chester Ismay

About you!

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

Workshop Materials

- Schedule can be found at bit.ly/USCOTS2019
- All files can be found on Google Drive

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:

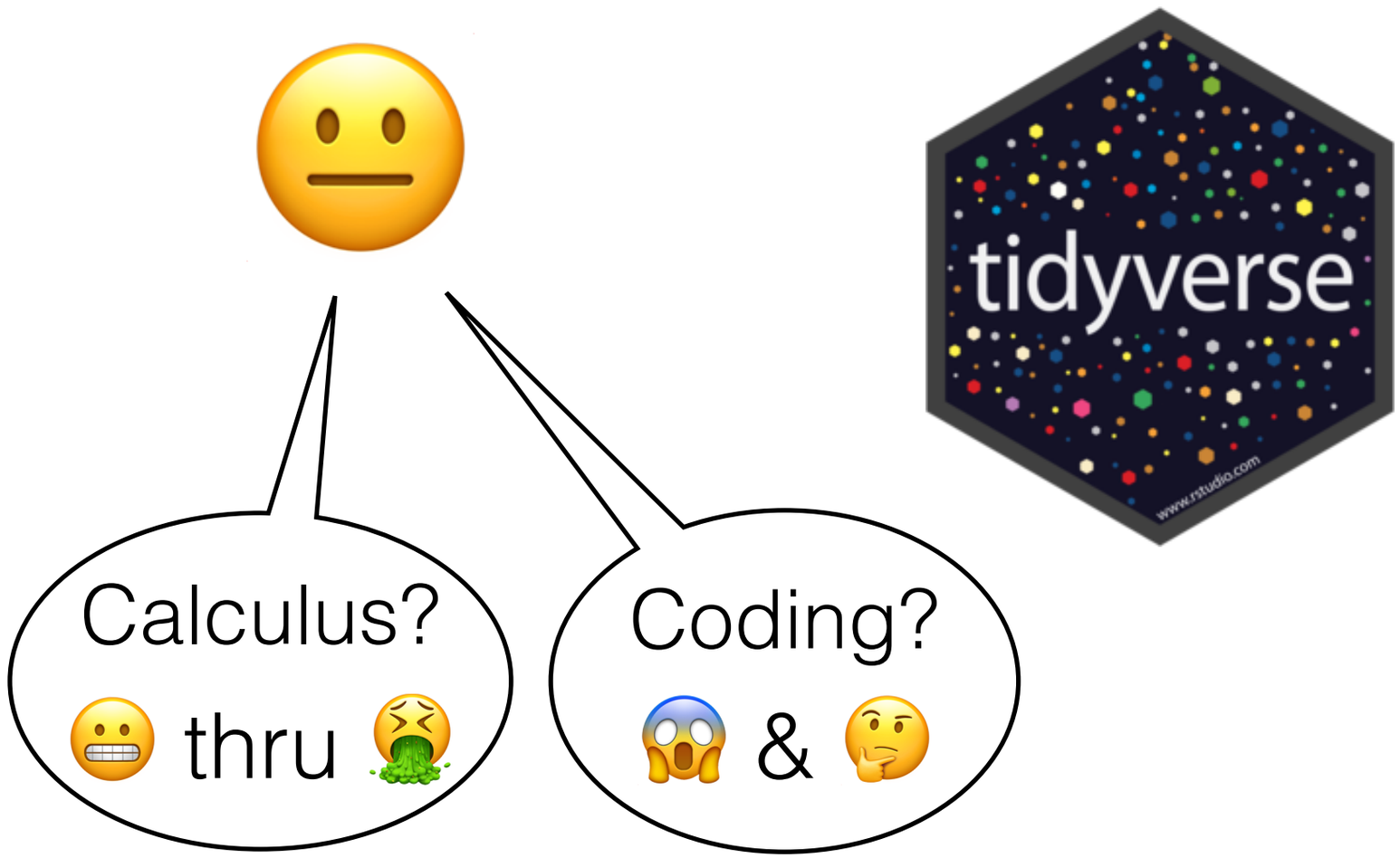
- Goal 1: Sampling for inference
- Goal 2: Modeling with regression

Getting from Point A to Point B

Point A:
Modal 1st time
stats student

via the
`tidyverse`

Point B:
Two goals



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
ModernDive
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?

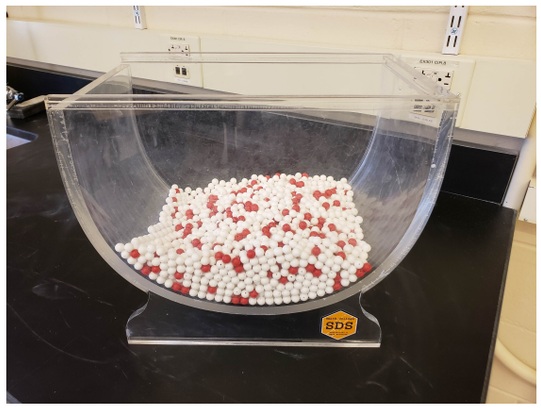
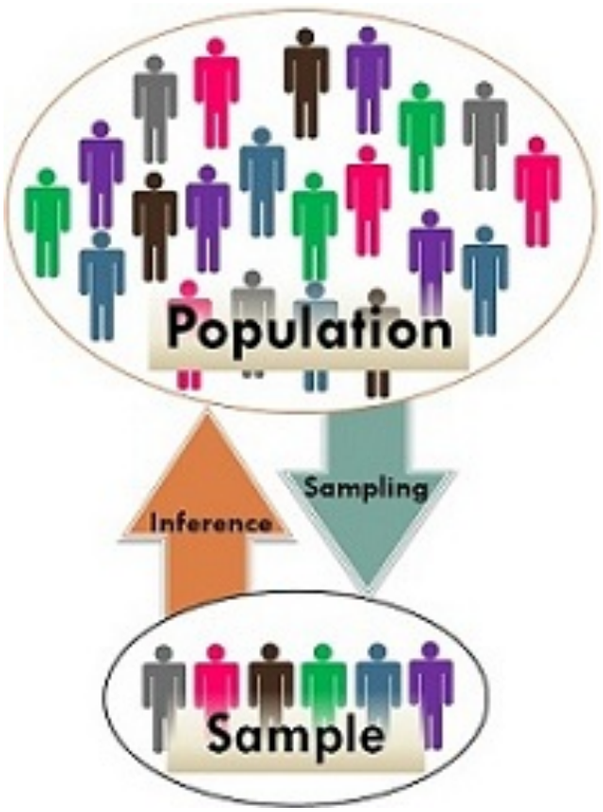
1. Don't reinvent the wheel!
2. Break down tasks step-by-step!
3. What is the goal of your code?
4. Make code understandable

Why tidyverse for stats newbies?

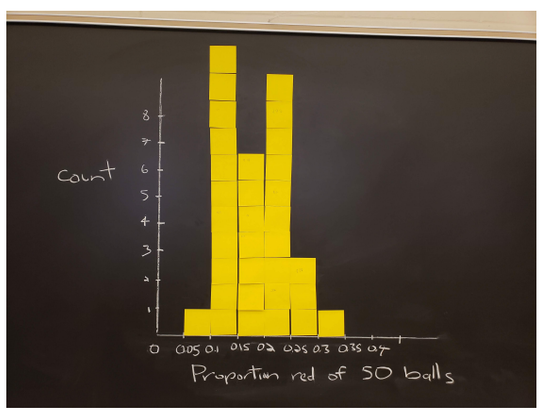
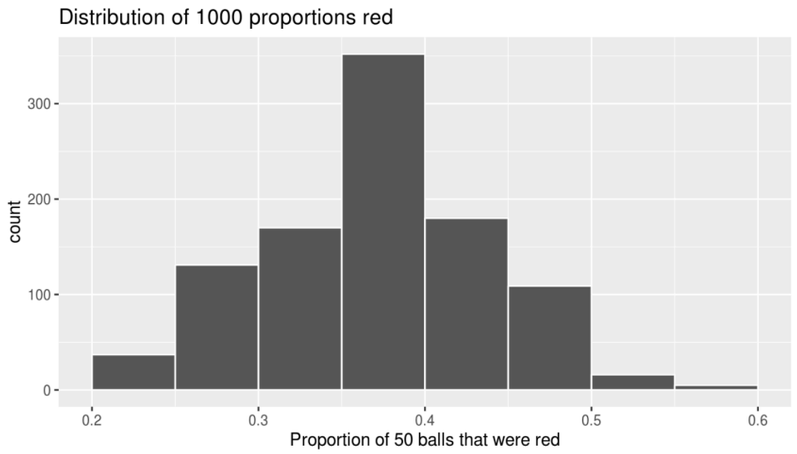
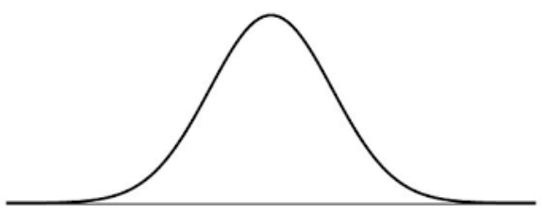
- *In my opinion* it's easier to learn than base R. Others too.
- It scales. You leverage an entire ecosystem of online developers and support: Google & StackOverflow
- Satisfy learning goals *while learning tools they can use beyond the classroom*

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

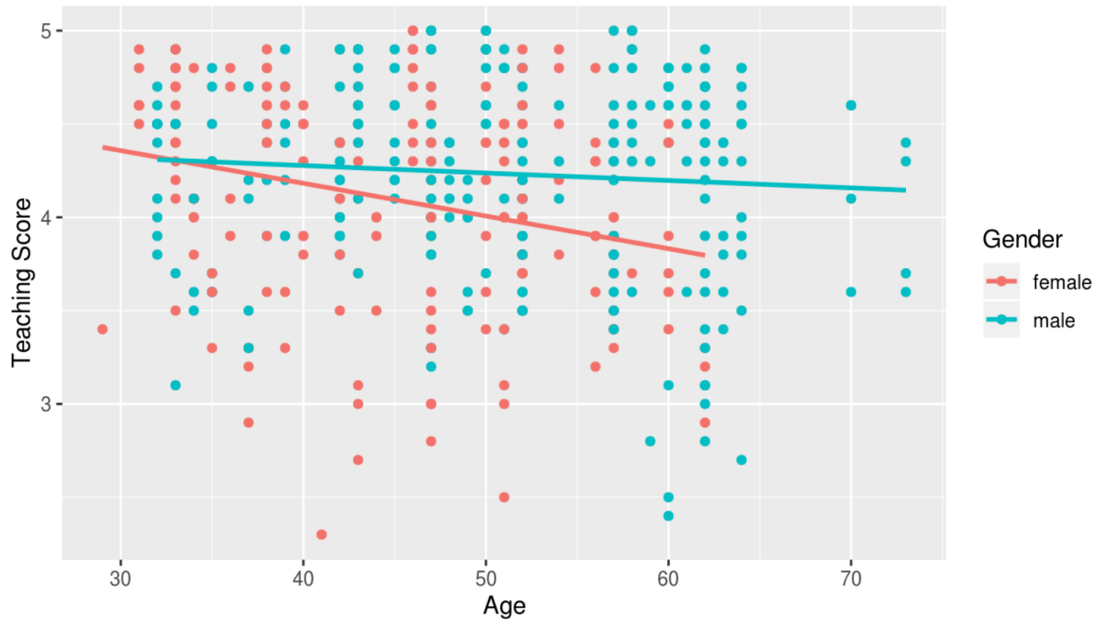
1. Data

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

Sampling

Conf Int's & Hyp Tests

4. Regression Table

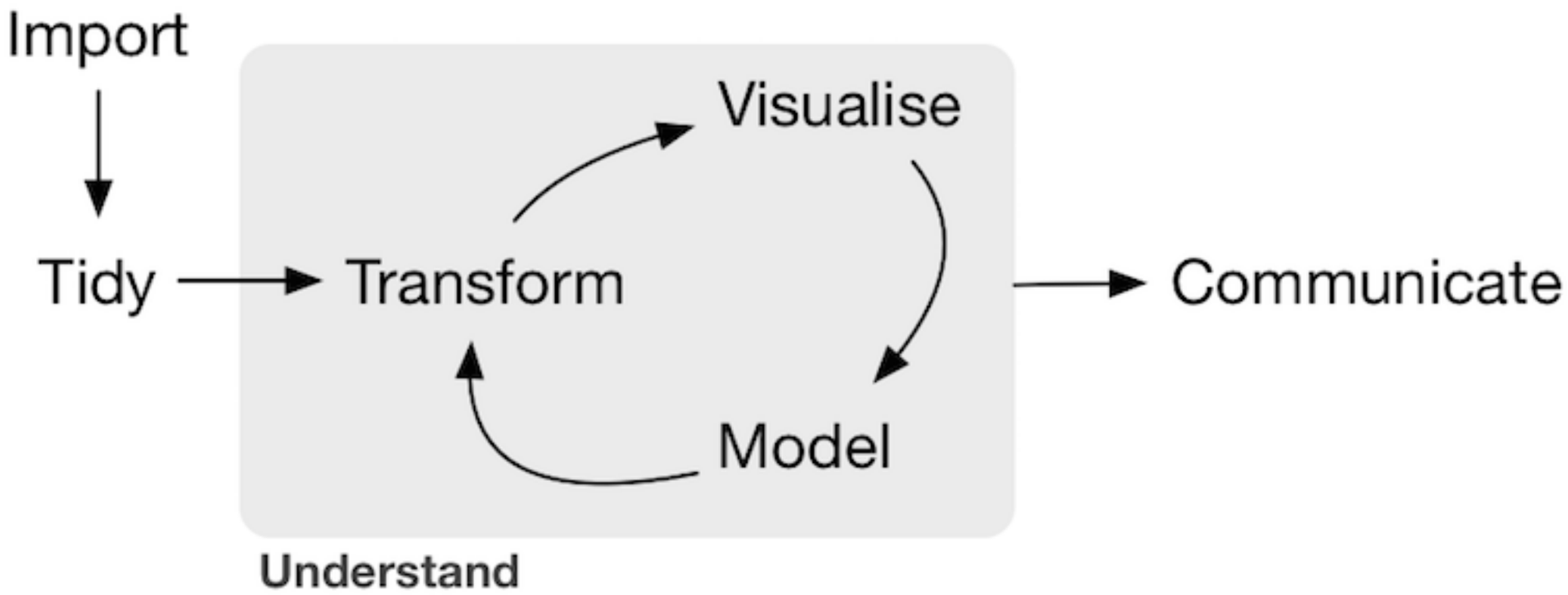
Console ~/

```
> 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

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

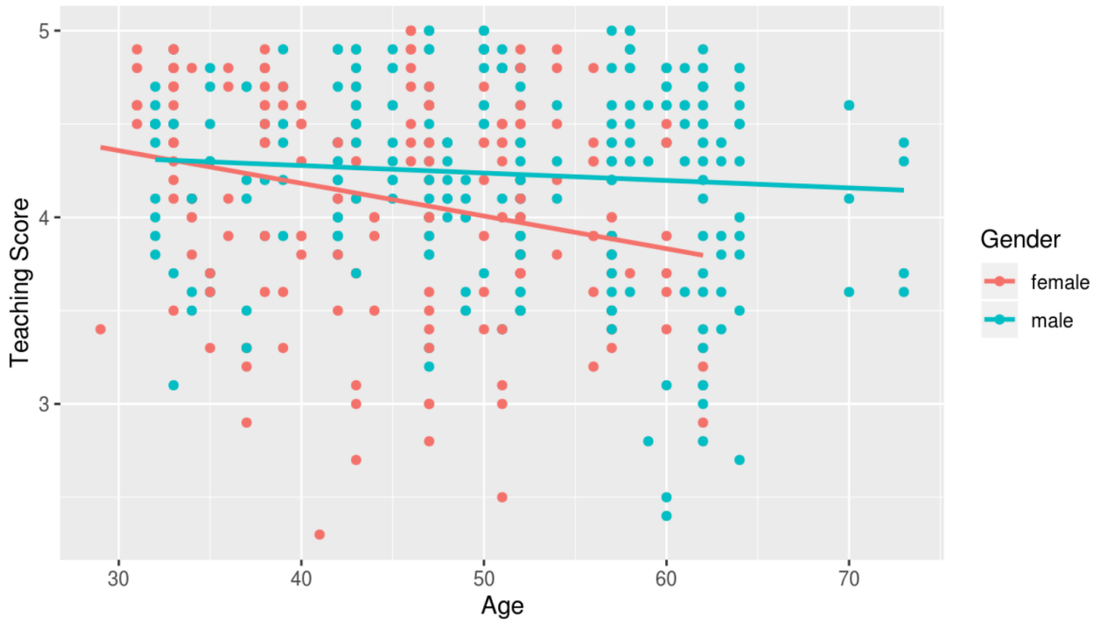
1. Data

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

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

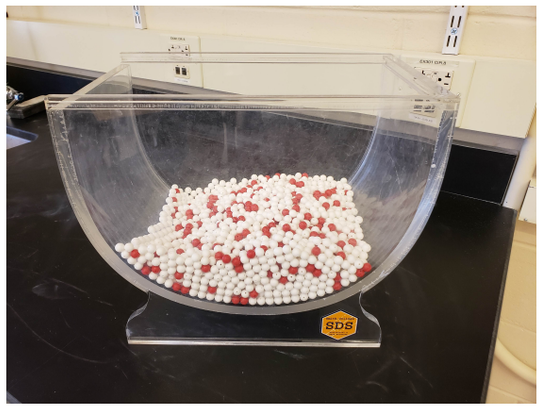
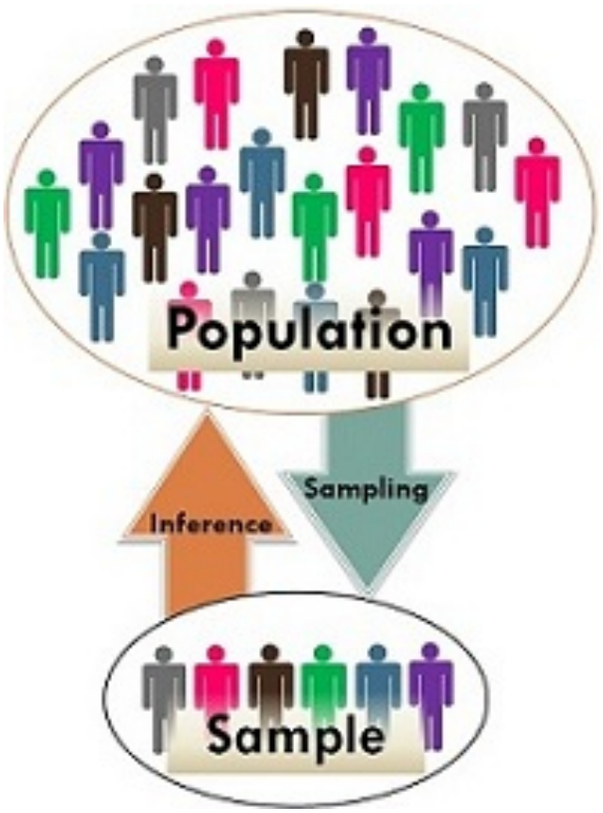
Chapter 7: Multiple regression

1. What are we doing ❓
   - *Descriptive* multiple regression & regression with 1 num & 1 categ x.
2. Why are we doing this 🤔
   - 👶 Baby's first model selection! 👶
   - Occam's Razor between interaction and parallel slopes model
3. Our opinions
   - Equation for fitted values w/ indicator functions is 😱
   - [1 num & 1 categ x] is more important than [2 num x]
4. Potential pitfalls ⚠️
   - Interaction model: interpreting offsets in intercept + differences in slope
   - How to plot parallel slopes model

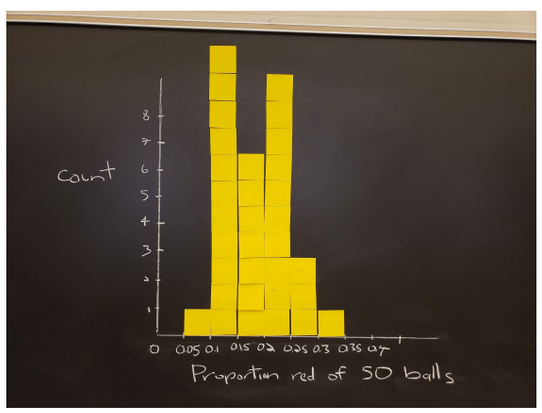
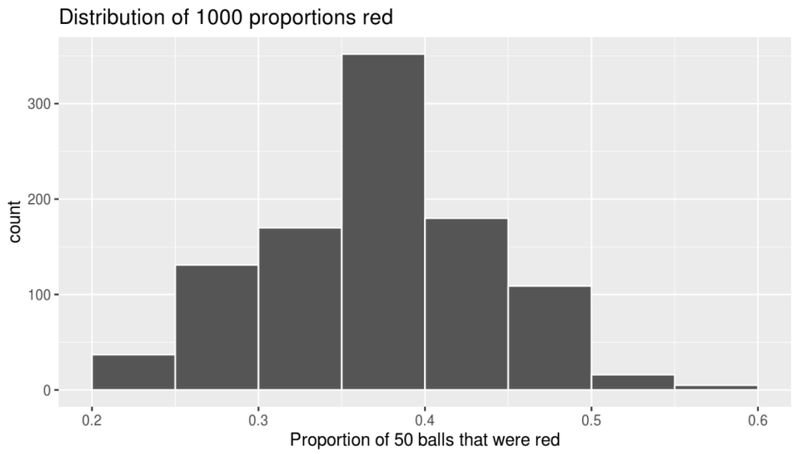
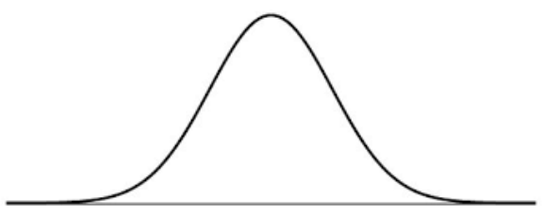
Chapter 8: Sampling

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

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.
3. Our opinions
   - Have some mental anchor for all statistical inference: tactile sampling exercise
4. Potential pitfalls ⚠️
   - Terminology, notation, & definitions related to 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

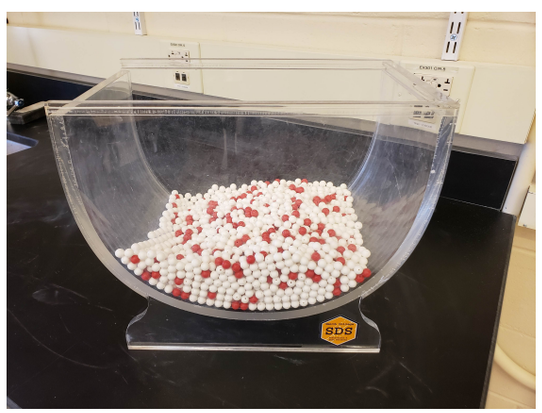
[redacted] Thu, Jan 3, 2:42 PM
Hi Matt, I teach a “statistics” course to medical students at Duke. I use quotes around the word statistics because I don’t really teach the students how to do

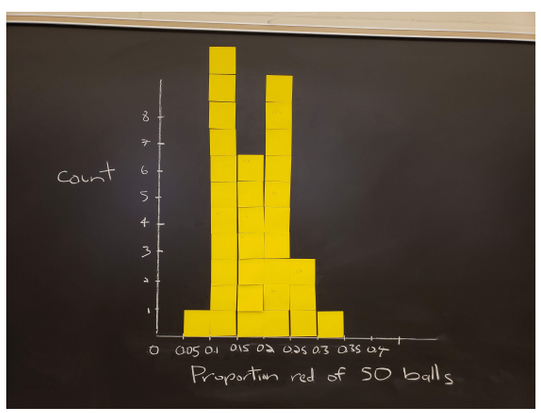
[redacted] Thu, Jan 3, 2:53 PM
Hi, I like the work of Kaplan and Rogness for some nice activities and a discussion of lexical ambiguity in statistics. https://scholarcommons.usf.edu/numeracy/

[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;-)

My 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 generalization & transference…

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

Chapter 9: Confidence Intervals

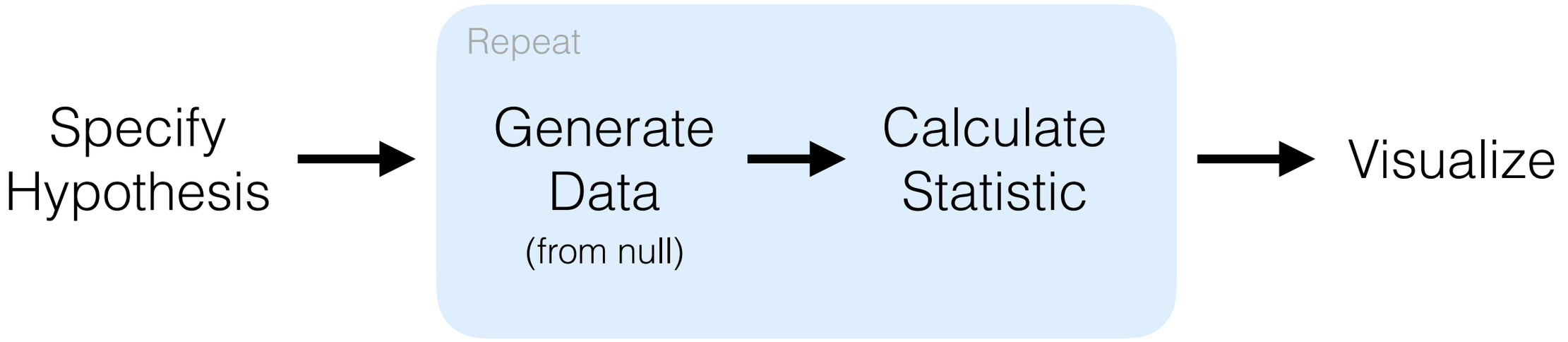
1. What are we doing ❓
   - Introducing bootstrap REsampling
   - Constructing confidence intervals
2. Why are we doing this 🤔
   - Convince students what needs to happen in real life (IRL) when you have *only one sample*
   - Where is sampling variation in CI's?
3. Our opinions
   - Have some mental anchor for all statistical inference: tactile REsampling exercise
4. Potential pitfalls ⚠️
   - "Bootstrap resampling distribution is an approximation to sampling distribution"
   - Population from a *superpopulation?*
   - Bridging gap with traditional formula-based methods

Chapter 10: Hypothesis Testing

1. What are we doing ❓
   - Introducing permutation REsampling
   - Conducting hypothesis tests
2. Why are we doing this 🤔
   - Convince students what needs to happen in real life (IRL) when you have *only one sample*
   - Where is sampling variation in HT's?
   - Convincing students there is only one test
3. Our opinions
   - I hate hypothesis testing, but they are still widely used
4. Potential pitfalls ⚠️
   - Terminology, notation, & definitions related to HT
   - Bridging gap with traditional formula-based methods

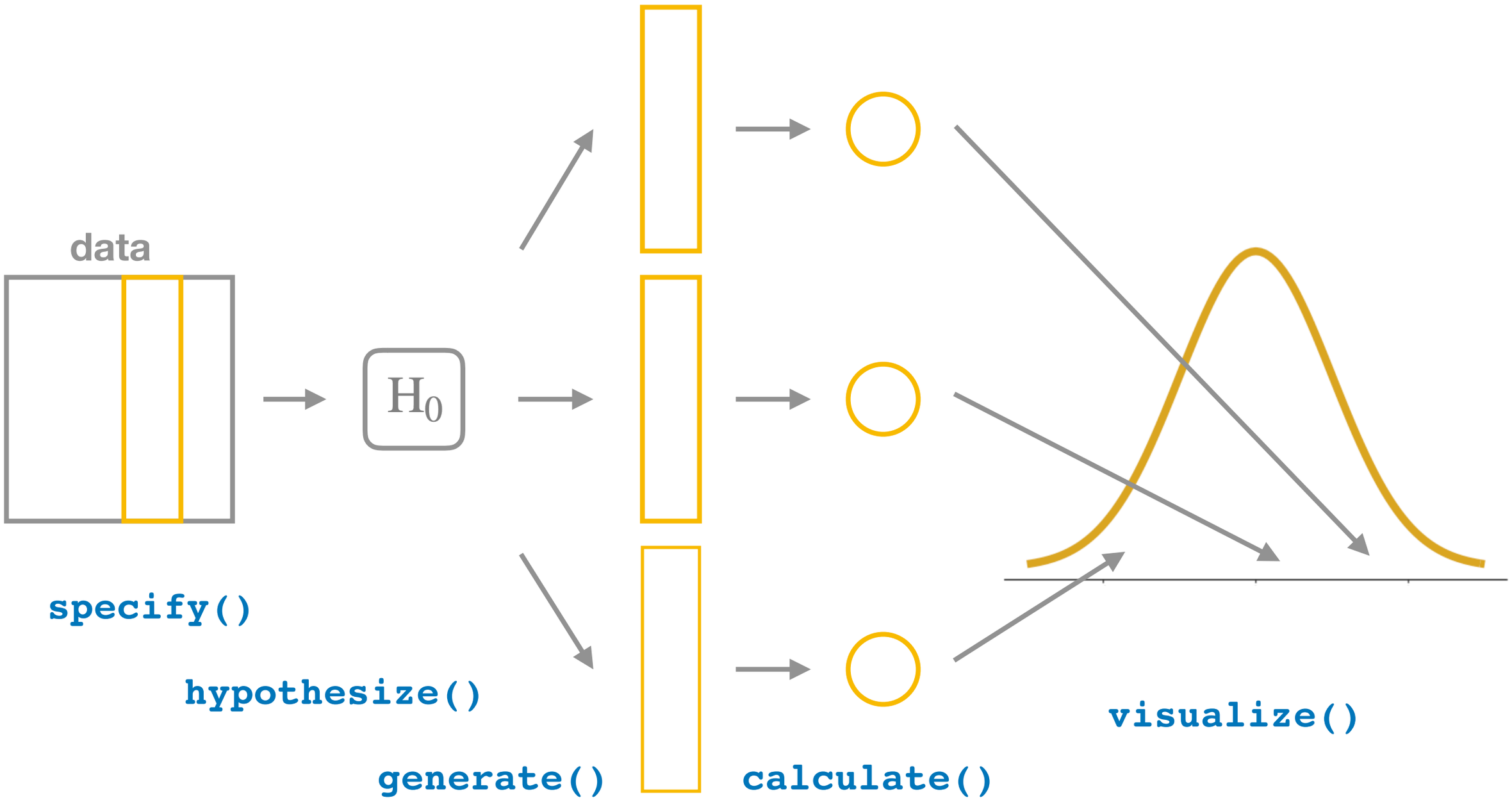
`infer` package for tidy statistical inference

http://infer.netlify.com/




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

Hypothesis Testing



Chapter 11: Inference for Regression

Goal 2: Modeling with Regression

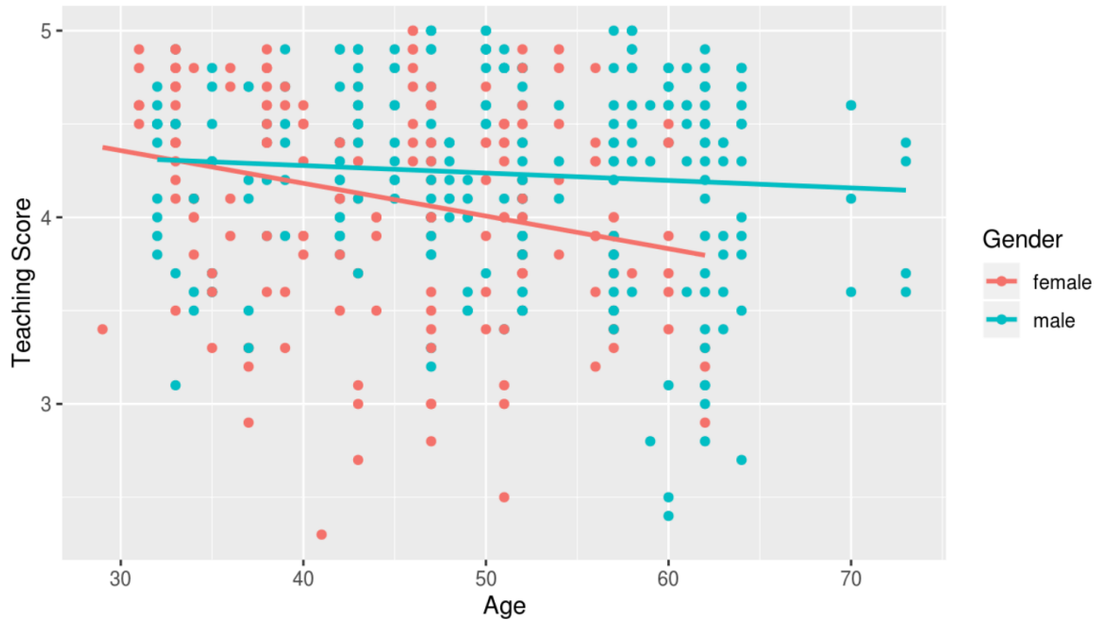
1. Data

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
>
```

Sampling

Conf Int's & Hyp Tests

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
>
```

1. What are we doing ❓
   - Getting students to interpret regression thru an inferential lens
   - Worth doing resampling for regression? I'm not sure.
2. Why are we doing this 🤔
   - Convince students what needs to happen in real life (IRL) when you have *only one sample*
   - Where is sampling variation in regression?
3. Our opinions
   - Use EDA + `get_regression_points()` to do your own residual analysis, not `base::plot(model)`
4. Potential pitfalls ⚠️
   - "Does R use simulation or a formula for p-values/CI's in a regression table?"

Conclusion

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

Resources

- Always two versions of moderndive available
  1. Development version (being edited): moderndive.netlify.com
  2. Latest release (updated x2 per year): moderndive.com
- On GitHub at github.com/moderndive/
  1. bookdown source code for book
  2. `moderndive` package source code
- Join our mailing list at 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 🚧
- **Mid-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!