moderndive: statistical inference via the tidyverse




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

USCOTS 2019
State College PA
May 15-16

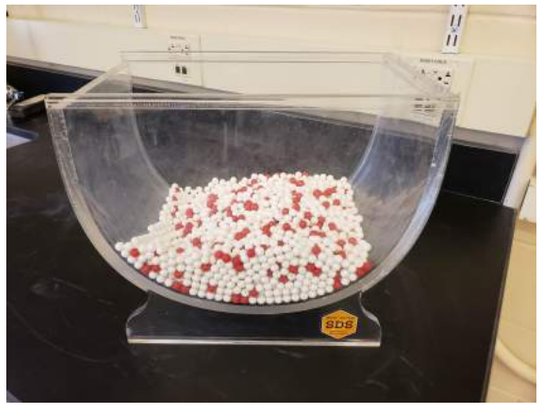
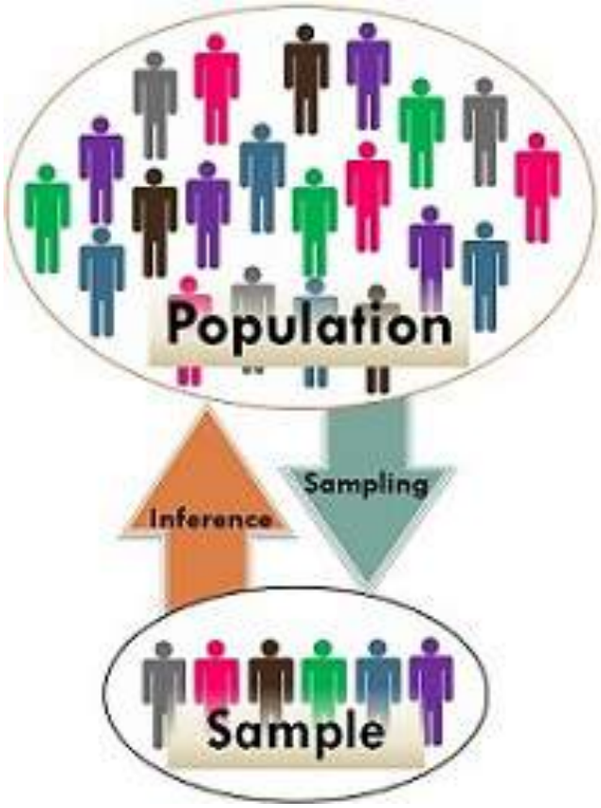
Chapter 7: Multiple regression

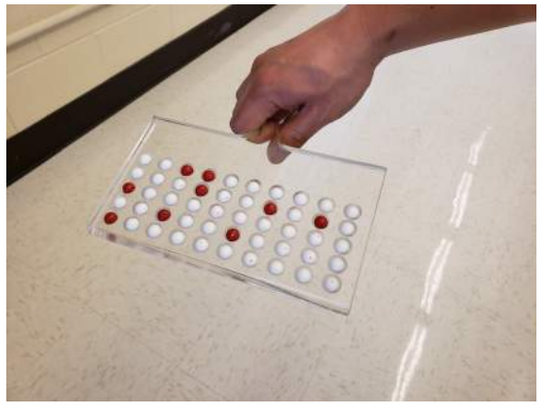
1. What are we doing ❓
   - *Descriptive* multiple regression & regression with 1 num & 1 categ x.
2. Why are we doing this 🤔
   - 👶 Baby's first model selection! 👶
   - Occam's Razor between interaction and parallel slopes model
3. Our opinions
   - Equation for fitted values w/ indicator functions 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

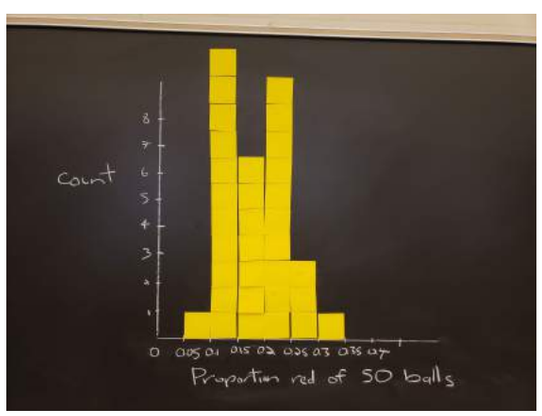
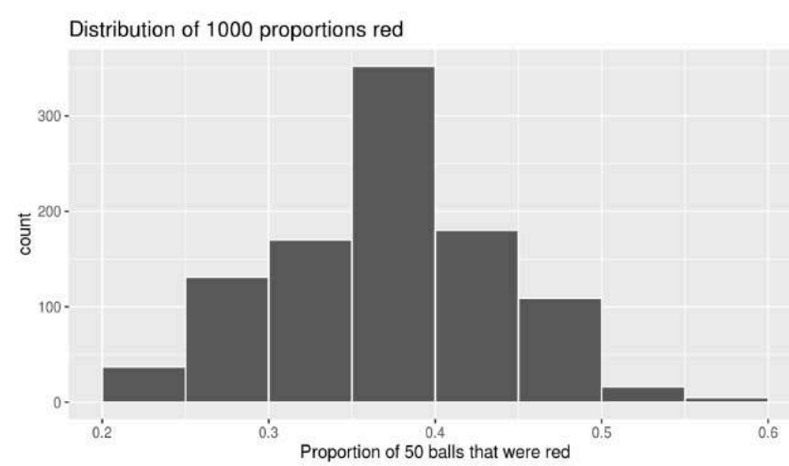
Chapter 8: Sampling

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/

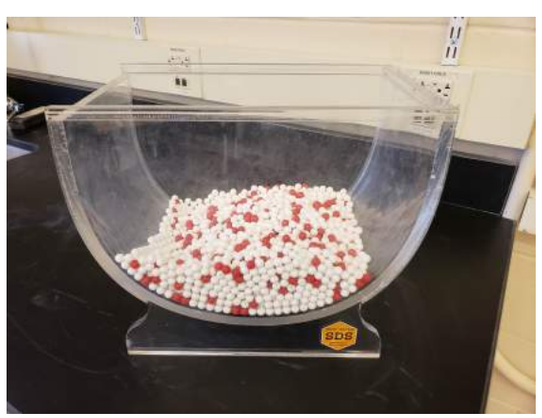
[redacted] Thu, Jan 3, 3:50 PM

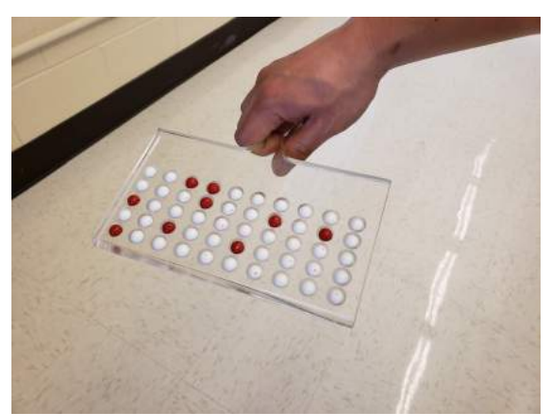
Hi Matt: With regard to proportions, I have been very careful to stay away from the use of "percentage," primarily because so many of my students lack basic mat

[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

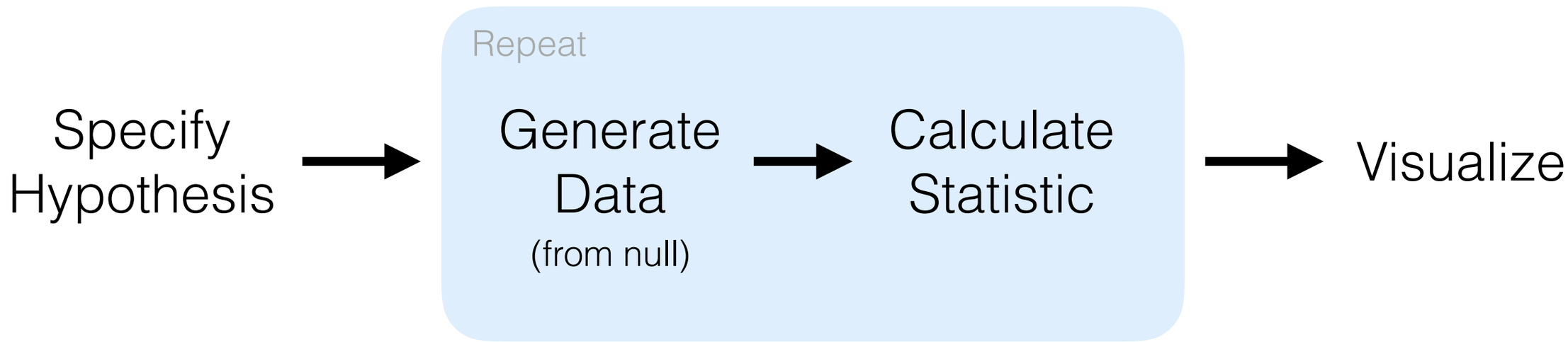
1. What are we doing ❓
   - Introducing bootstrap REsampling
   - Constructing confidence intervals
2. Why are we doing this 🤔
   - Convince students what needs to happen in real life (IRL) when you have <u>*only one sample*</u>
   - Where is sampling variation in CI's?
3. Our opinions
   - Have some 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

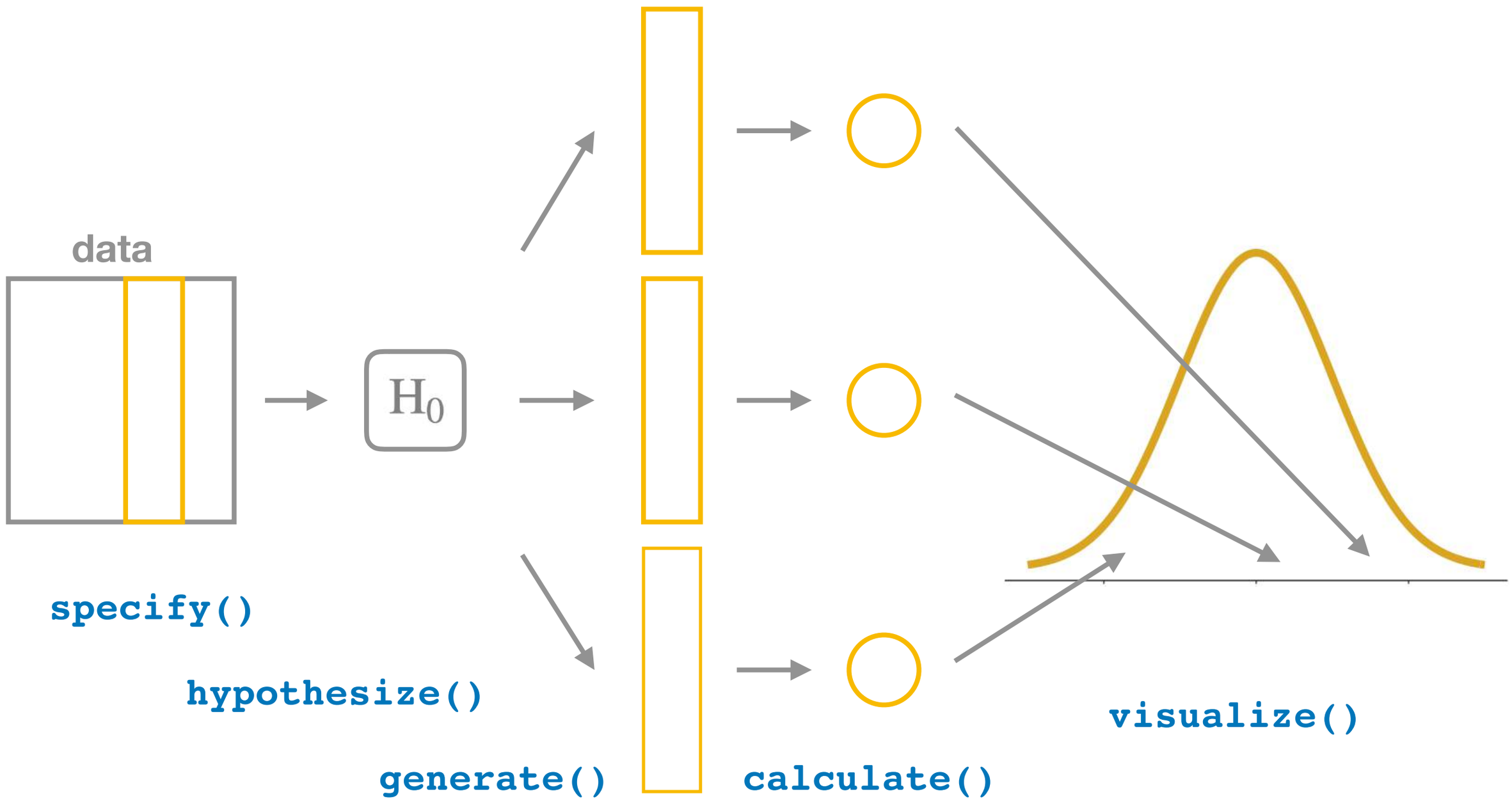
`infer` package for tidy statistical inference

http://infer.netlify.com/




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

Hypothesis Testing



Chapter 11: Inference for Regression

Goal 2: Modeling with Regression

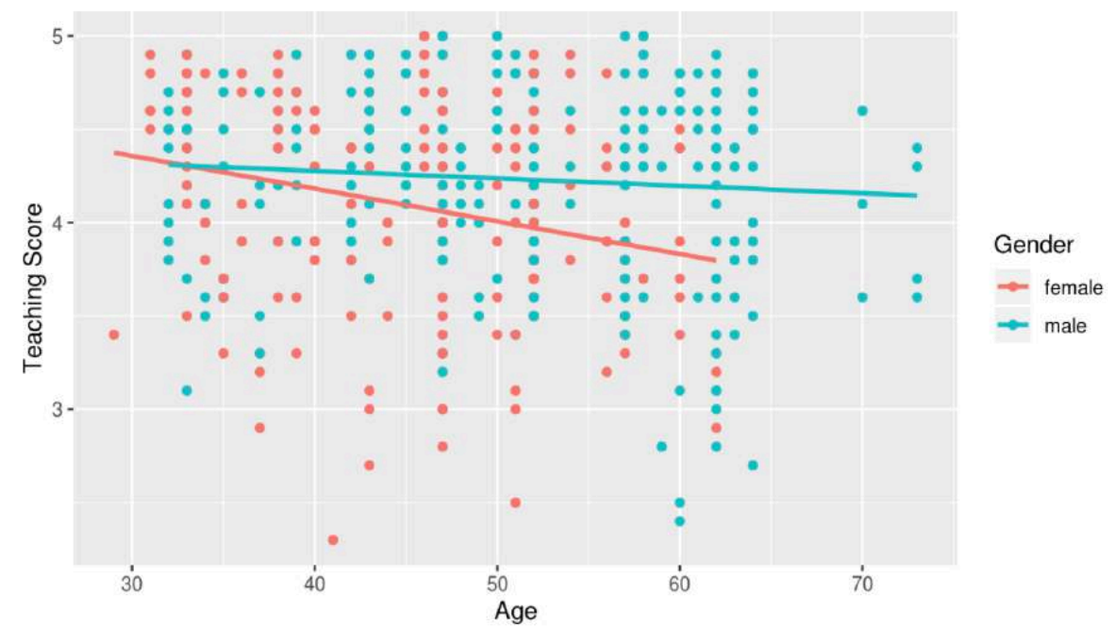
1. Data

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 <u>*only one sample*</u>
   - 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

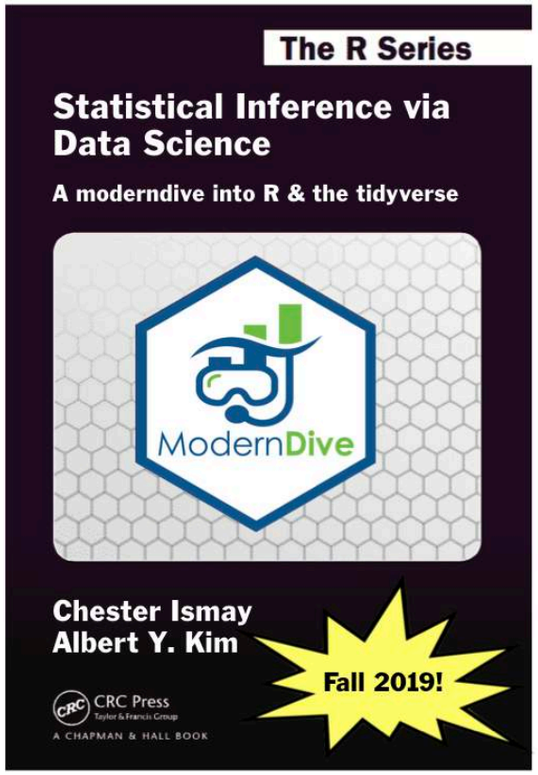
- Ch6: Use `get_regression_table()` instead of `summary()`
- Ch5 + Ch2: Publish (non-sensitive) data to .csv via Google Sheets and import with `read_csv()`.
- Ch3: Spend time covering Grammar of Graphics & do all plots in course via `ggplot2`
- Ch8 + Ch5 + Ch3: Use data frame + `%>%` + `rep_sample_n()` 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!