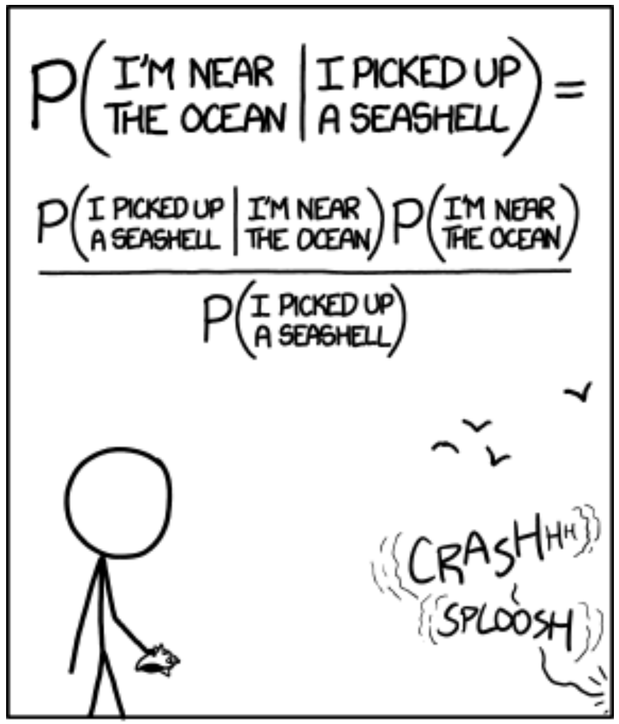
Summary and course outlook



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND *DON'T* HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.

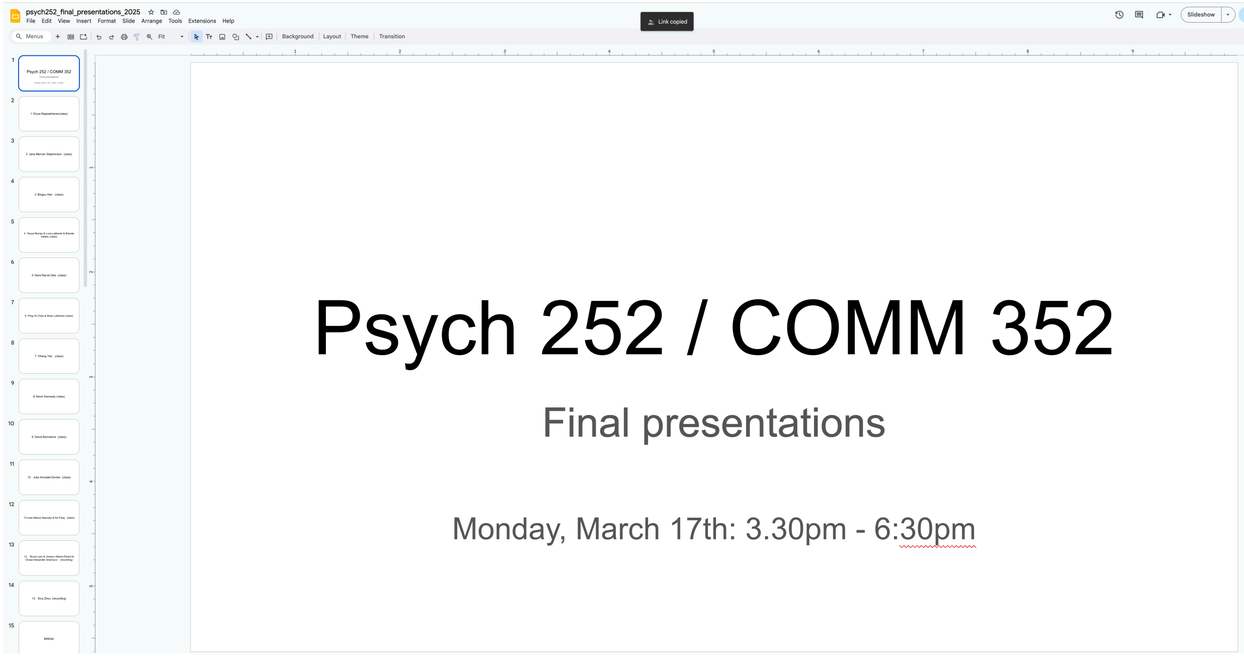
03/12/2025

Logistics

Final presentations

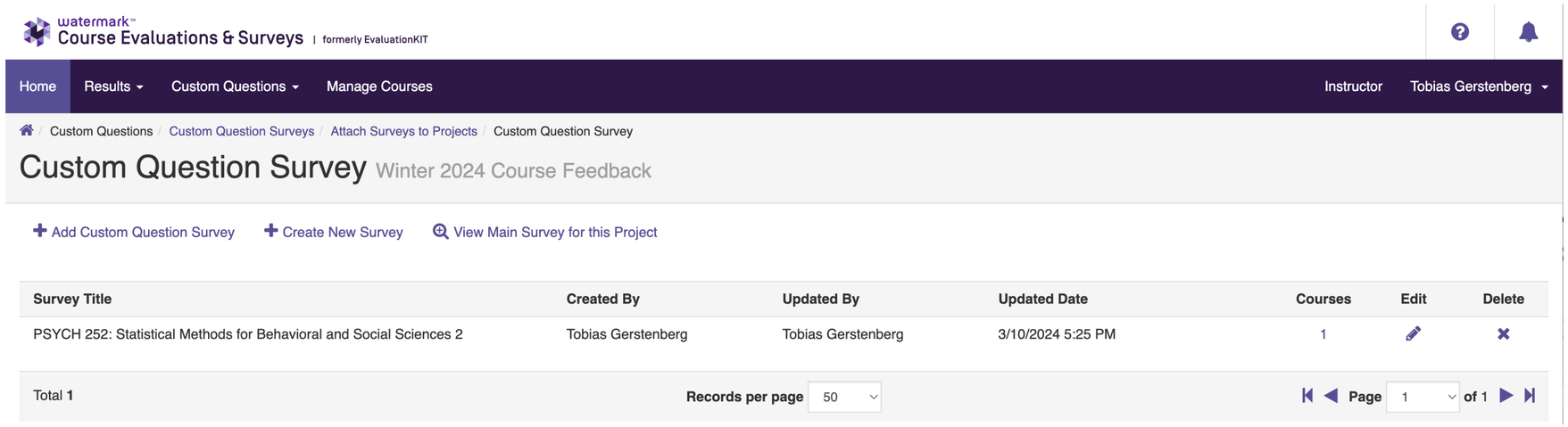
index	name	choice	duration	start	end
1	Divya Rajasekharan	class	4	15:36	15:40
2	Jane Mercier Stephenson	class	4	15:40	15:44
3	Bingxu Han	class	4	15:44	15:48
4	Tonya Murray & Luna Laliberte & Brenda Valdes	class	8	15:48	15:56
5	Nora Ranck Dee	class	4	15:56	16:00
6	Ping-Ya Chao & Brian Lattimore	class	6	16:00	16:06
7	Yiheng Yao	class	4	16:06	16:10
8	Kevin Kennedy	class	4	16:10	16:14
9	David Barnstone	class	4	16:14	16:18
10	Julia Annadel Donlon	class	4	16:18	16:22
11	Linas Marius Nasvytis & Ke Fang	class	6	16:22	16:28
12	Bruno Lam & Jocelyn Allaina Ricard & Chase Alexander Antonacci	recording	8	16:28	16:36
13	Siva Zhou	recording	4	16:36	16:40
	BREAK		6	16:40	16:46
14	Danilo Symonette	remote	4	16:46	16:50
15	Beleicia Benita Bullock	class	4	16:50	16:54
16	Shane Muldowney	class	4	16:54	16:58
17	Julia Proshan	class	4	16:58	17:02
18	Jonah Rosemeier	class	4	17:02	17:06
19	Isabella Caterina Aslarus	class	4	17:06	17:10
20	Sarah Sampaio Izabel	class	4	17:10	17:14
21	Grace Brown	class	4	17:14	17:18
22	Yulia Venichenko	class	4	17:18	17:22
23	Noah Vinoya	class	4	17:22	17:26
24	Micaela Maria Bonilla	class	4	17:26	17:30
25	Ziyu Ren & Hogkai Mao & Simon Huang	recording	8	17:30	17:38
26	Kavindya Thennakoon	recording	4	17:38	17:42
	BREAK		6	17:42	17:48
27	Clara Maria Bacmeister	remote	4	17:48	17:52
28	Marcos Santiago Rojas Pino	remote	4	17:52	17:56
29	Caroline Kaicher	class	4	17:56	18:00
30	Ramya Kumar	class	4	18:00	18:04
31	Devin Chuyi Moua	class	4	18:04	18:08
32	Kathrine Mia Whitman & Shashanka Subrahmanya	class	6	18:08	18:14
33	Carla Colina	class	4	18:14	18:18
34	Gabrianna Barcelo	class	4	18:18	18:22
35	Jeongyeon Kim	recording	4	18:22	18:26
36	Shuman Wang	recording	4	18:26	18:30

schedule



slide deck

Course evaluations



Thank you for evaluating the course!!

http://course-evaluations.stanford.edu

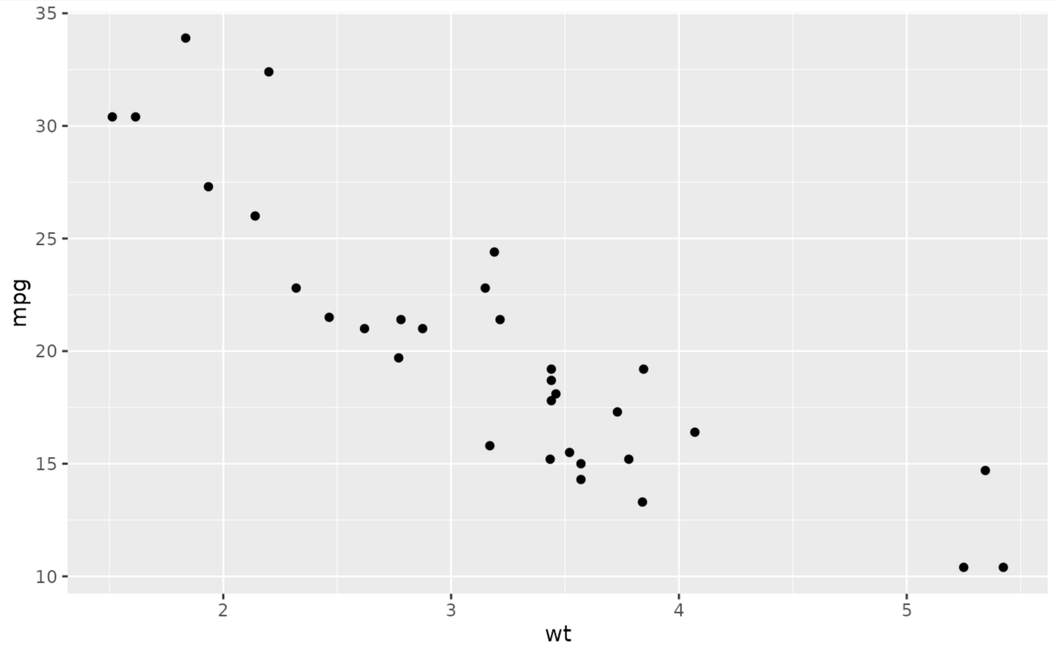
Guest lecture: Friday

Alice Xue	Catherine Garton	Justin Yang	Satchel Grant	Verity Lua
				

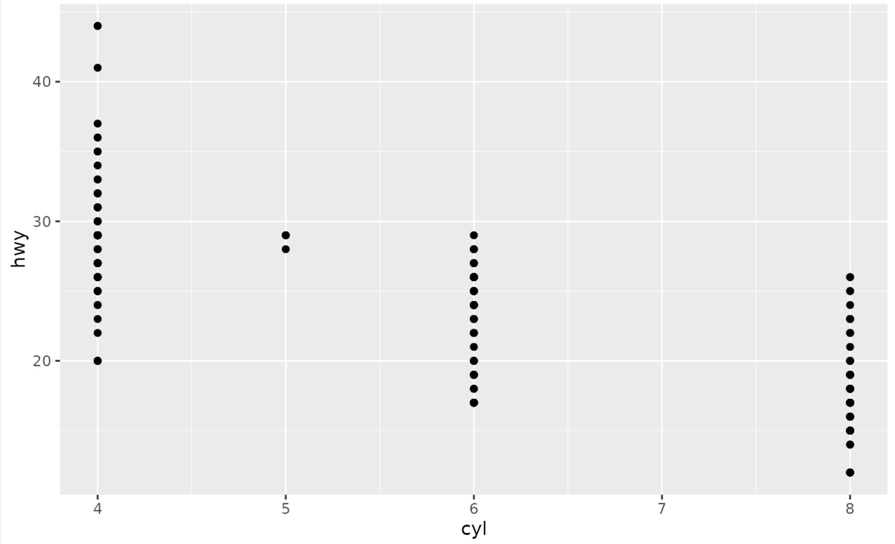
Things that came up

geom_point() vs. geom_jitter()

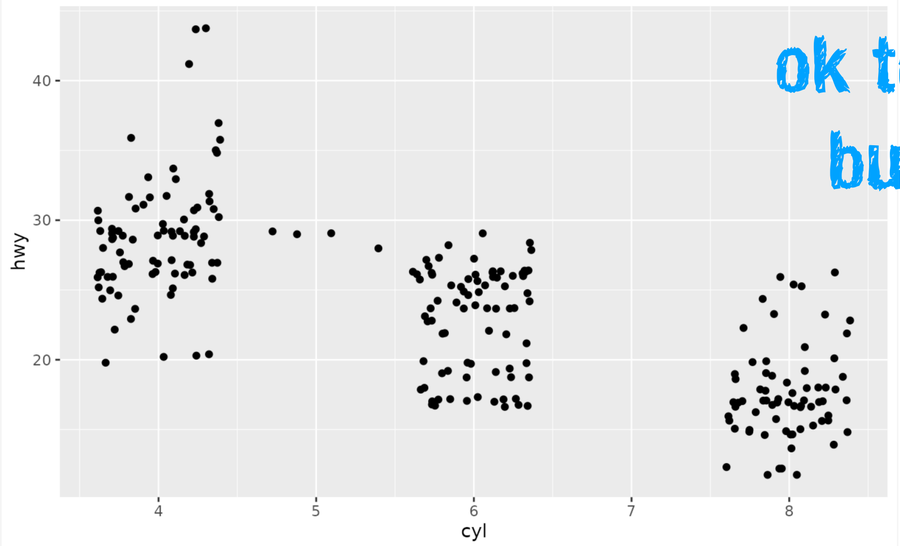
```
p <- ggplot(mtcars, aes(wt, mpg))
p + geom_point()
```




```
p <- ggplot(mpg, aes(cyl, hwy))
p + geom_point()
```




```
p + geom_jitter()
```



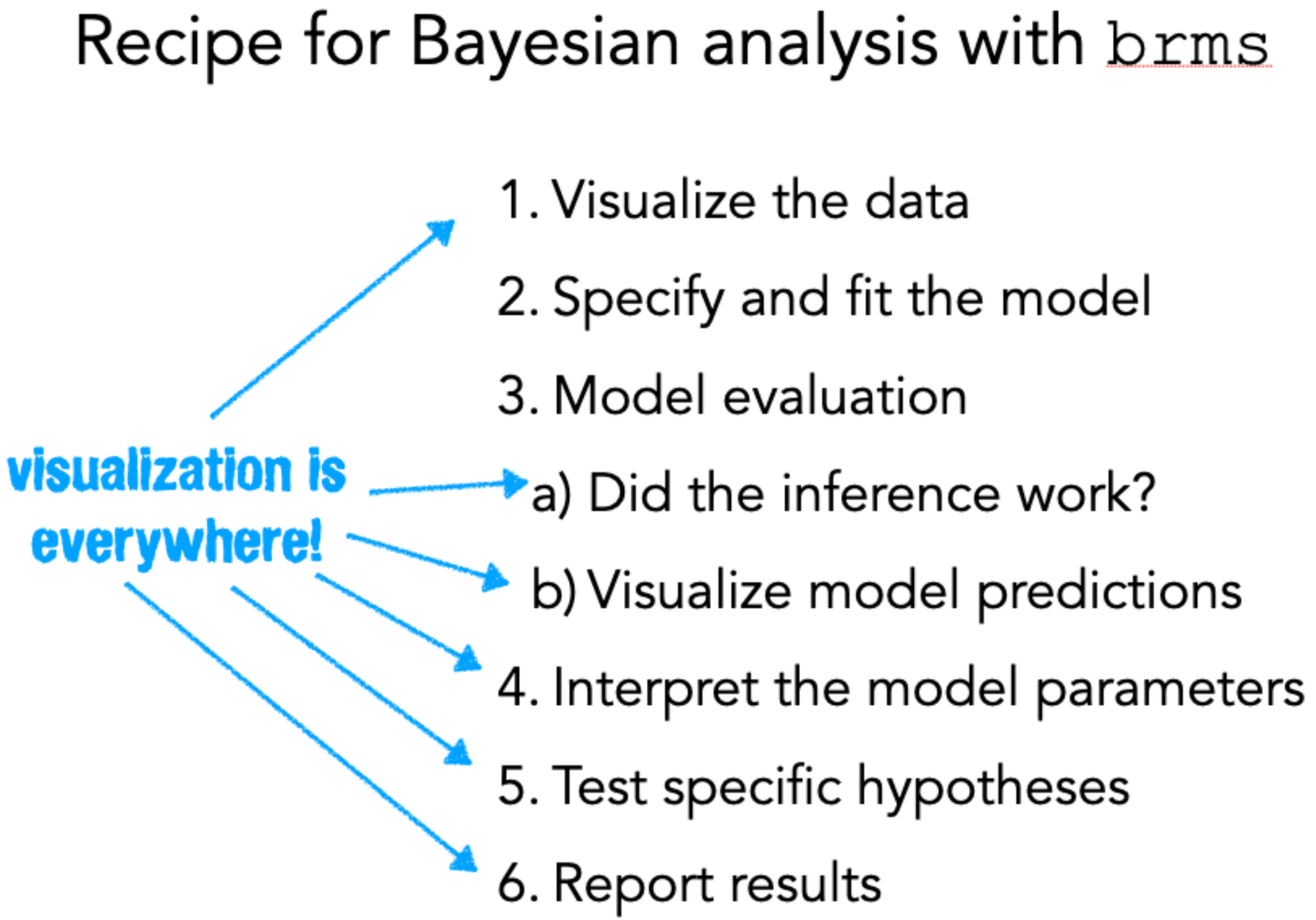
ok to jitter on x, but not on y!

Plan for today

- Quick recap
- What we've learned
- Nilam's 5 highlights
- Tobi's 5 highlights
- What shall I do now?
- Thank you!

Quick recap

Quick recap: Bayesian data analysis



Quick recap: Sleep data

1. Specify and fit the model

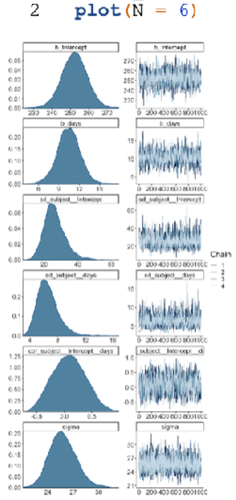
```
fit.brm_sleep = brm(formula = reaction ~ 1 + days + (1 + days | subject),
                    data = df.sleep,
                    seed = 1,
                    file = "cache/brm_sleep")
```

43

a) Did the inference work?

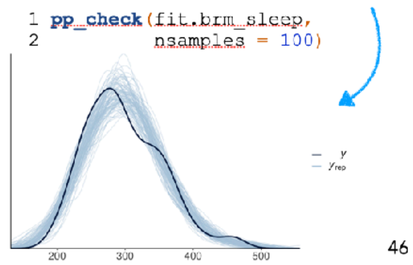
```
fit.brm_sleep %>%
  plot(N = 6)
```

these look good!



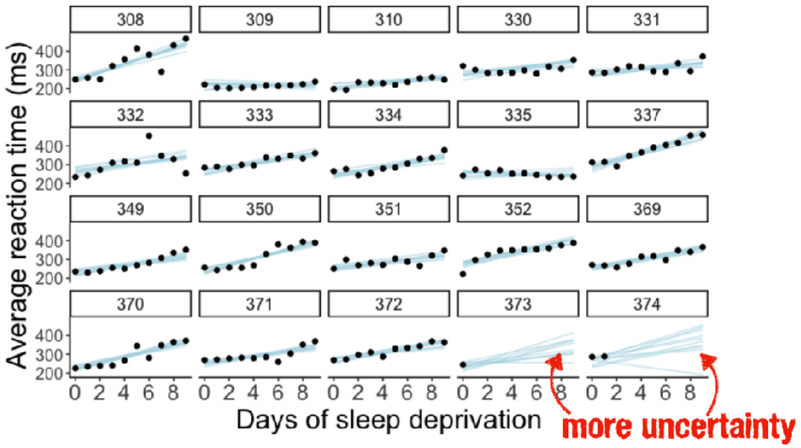
also looks good!

```
pp_check(fit.brm_sleep,
         nsamples = 100)
```



46

b) Visualize the model predictions



10 random samples from the posterior distribution

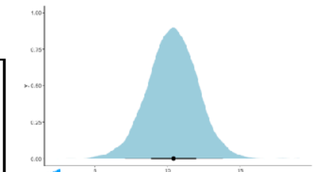
48

5. Test specific hypotheses

Did reaction times increase with the number of days of sleep deprivation?

```
fit.brm_sleep %>%
  summary()
```

```
Family: gaussian
  Links: mu = identity; sigma = identity
Formula: reaction ~ 1 + days + (1 + days | subject)
   Data: df.sleep (Number of observations: 180)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Group-Level Effects:
~subject (Number of levels: 20)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)          25.16      6.25    15.05    40.54 1.00     1879     2463
sd(days)                6.26      1.52     4.14    10.13 1.00     1145     1925
cor(Intercept,days)     0.06      0.25    -0.46     0.57 1.00      993     1526

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept   252.18      6.86   238.47   265.42 1.00     1826     2766
days         10.46      1.69     7.13    13.78 1.00     1903     1782

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma    25.77      1.57    22.83    29.14 1.00     3864     2773

Samples were drawn using sampling(NUTS). For each parameter, Bulk_Sample
is a crude measure of effective sample size, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```



54

Quick recap: Titanic data

1. Specify and fit the model

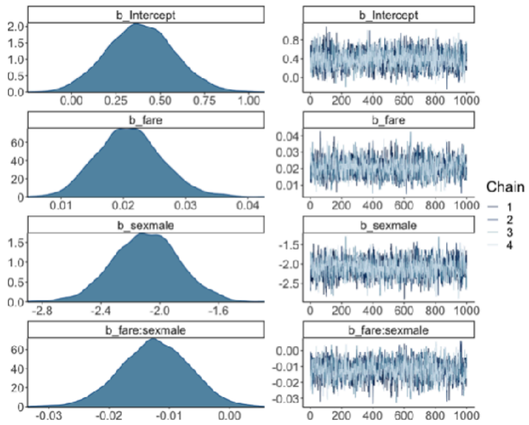
```
fit.brm_titanic = brm(formula = survived ~ 1 + fare * sex,
                      family = "bernoulli",
                      data = df.titanic,
                      file = "cache/brm_titanic",
                      seed = 1)
```

just need to change the family

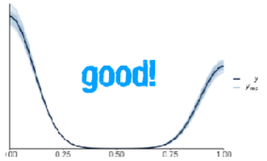
63

a) Did the inference work?

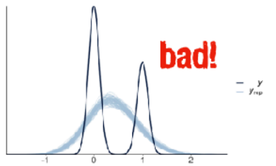
```
fit.brm_titanic %>%
  plot()
```




```
pp_check(fit.brm_titanic,
         nsamples = 100)
```



model with Gaussian family

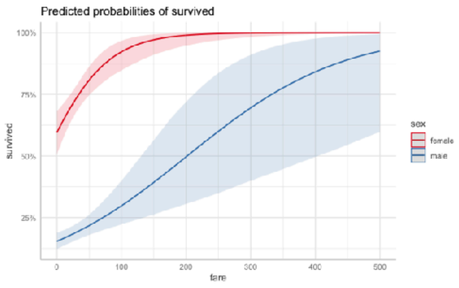


66

4. Interpret the model parameters

```
fit.brm_titanic %>%
  ggpredict(terms = c("fare [0:500]", "sex"))
```

```
# Predicted probabilities of survived
# x = fare

# sex = female

  x | Predicted |       95% CI
-------------------------------
  0 |      0.60 | [0.52, 0.68]
 83 |      0.89 | [0.82, 0.95]
167 |      0.98 | [0.93, 1.00]
250 |      1.00 | [0.97, 1.00]
333 |      1.00 | [0.99, 1.00]
500 |      1.00 | [1.00, 1.00]

# sex = male

  x | Predicted |       95% CI
-------------------------------
  0 |      0.15 | [0.12, 0.19]
 83 |      0.27 | [0.21, 0.35]
167 |      0.43 | [0.28, 0.62]
250 |      0.60 | [0.35, 0.84]
333 |      0.75 | [0.43, 0.94]
500 |      0.93 | [0.60, 0.99]
```



70

5. Test specific hypotheses

Was the effect of fare on survival different for men vs women?

```
fit.brm_titanic %>%
  emtrends(specs = pairwise ~ sex,
           var = "fare")
```

```
$emtrends
 sex    fare.trend lower.HPD upper.HPD
 female    0.02083   0.01129    0.0316
 male      0.00845   0.00385    0.0135

Point estimate displayed: median
HPD interval probability: 0.95

$contrasts
 contrast      estimate lower.HPD upper.HPD
 female - male   0.0124  0.000884    0.0232

Point estimate displayed: median
HPD interval probability: 0.95
```

the chance of survival increased more with fare for female than male passengers

73

Going beyond: Evidence for the Null

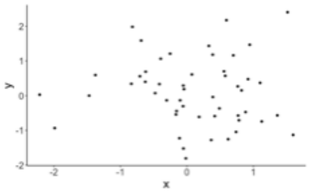
Bayes factor

$$BF_{01} = \frac{p(D \mid H_0)}{p(D \mid H_1)}$$

probability of the data given H_0

probability of the data given H_1

Evidence for the null hypothesis



```
1 fit.brm_loo1 = brm(formula = y ~ 1, data = df.loo)
2
3 fit.brm_loo2 = brm(formula = y ~ 1 + x, data = df.loo)
4
5 fit.brm_loo1 = add_criterion(fit.brm_loo1, criterion = "loo")
6
7 fit.brm_loo2 = add_criterion(fit.brm_loo2, criterion = "loo")
```

approximate leave-one-out cross-validation

```
loo_compare(fit.brm_loo1, fit.brm_loo2)
```

	elpd_diff	se_diff
fit.brm_loo1	0.0	0.0
fit.brm_loo2	-1.1	0.5

```
model_weights(fit.brm_loo1, fit.brm_loo2)
```

fit.brm_loo1	fit.brm_loo2
99.99999	0.00001

62

Going beyond

I only want positive coefficients!

```
brm(formula = how_much_i_love_stats ~ 1 + tobi + ari + beth + satchel + shawn,
    data = df.stats_love)
```

coefficients in the model

```
# priors
priors = c(set_prior("normal(0,10)", class = "b", lb = 0))
```

lower bound = 0

```
brm(formula = how_much_i_love_stats ~ 1 + tobi + ari + beth + satchel + shawn,
    prior = priors,
    data = df.stats_love)
```

Unequal variance aka heteroscedasticity

```
fit.brm1 = brm(formula = bf(response ~ group,
                            sigma ~ group),
               data = df.variance,
               file = "cache/brm1",
               seed = 1)
```

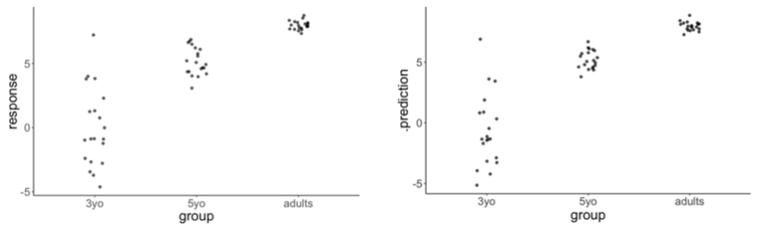
modeling both the means and variances

```
 Family: gaussian
  Links: mu = identity; sigma = log
Formula: response ~ group
         sigma ~ group
   Data: df.variance (Number of observations: 60)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup samples = 4000

Population-Level Effects:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept          -0.01      0.73    -1.41     1.51 1.01     1107     1072
sigma_Intercept     1.15      0.17     0.85     1.51 1.00     1991     1922
group5yo            5.18      0.77     3.60     6.65 1.00     1252     1327
groupadults         7.98      0.74     6.47     9.37 1.01     1110     1079
sigma_group5yo     -1.05      0.24    -1.51    -0.57 1.00     2249     2420
sigma_groupadults  -2.19      0.24    -2.66    -1.74 1.00     2171     2427

Samples were drawn using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Unequal variance aka heteroscedasticity

```
df.variance %>%
  add_predicted_draws(model = fit.brm1,
                      n = 1) %>%
  ggplot(aes(x = group, y = .prediction)) +
  geom_jitter(height = 0,
              width = 0.1,
              alpha = 0.7)
```

original data → predicted data



these predictions look good!

What we've learned

Learning goals

What you will learn

You will learn how to **use R** to …

- read, wrangle, and analyze data
- make publication-ready plots

Understand the philosophy behind null **hypothesis significance testing (NHST)** and **Bayesian statistics** through …

- running computer simulations and visualizing the results

Formulate **research questions as statistical models** and …

- determine which models work for different situations

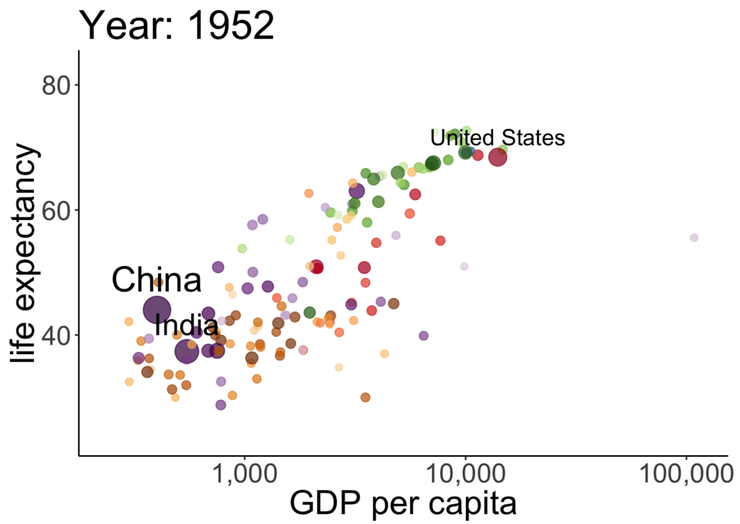
Communicate what you have learned about your data …

- in short presentations in class, showcasing your visualization and analysis
- in written reports

Contribute to open and **reproducible science** through …

- adopting good coding practices
- sharing your data and research reports online

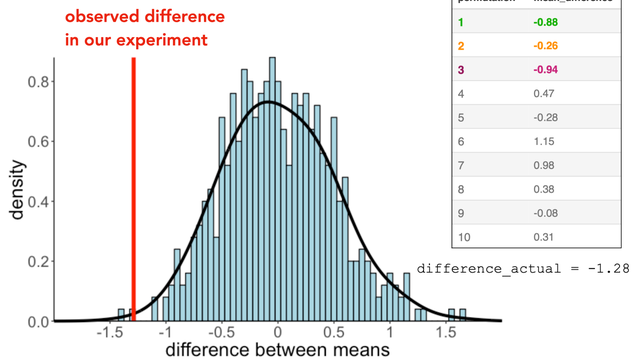
You will learn to use **R**



visualization

data wrangling

Permutation test



df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

difference_actual = -1.28

```
1 #calculate p-value of our observed result
2 df.permutations %>%
3   summarize(p_value = sum(mean_difference <= difference_actual)/n())
```

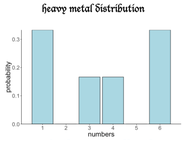
p-value = .002

64

simulation

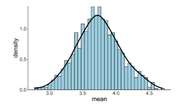
Philosophy behind **frequentist** and Bayesian stats

3 distributions in statistical inference



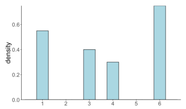
population distribution

- unknown
- our target for inference
- e.g. we might be interested in the mean of the population distribution



sampling distribution

- bridge between sample and population
- derived mathematically / computationally
- asymptotic distribution theory or resampling approaches
- shows how test statistic varies between samples



sample distribution

- our observed sample
- we compute statistics of interest (mean, variance, correlation, ...)
- make an inference about the population via the sampling distribution

45

sampling distribution

Permutation test

observed data

participant	condition	performance
1	control	4.25
2	control	5.87
3	control	3.83
4	control	8.69
5	control	6.16
26	experimental	4.42
27	experimental	4.27
28	experimental	2.29
29	experimental	3.78
30	experimental	5.13

1

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	control	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

2

participant	condition	performance
1	experimental	4.25
2	control	5.87
3	experimental	3.83
4	experimental	8.69
5	experimental	6.16
26	control	4.42
27	control	4.27
28	control	2.29
29	control	3.78
30	experimental	5.13

3

participant	condition	performance
1	control	4.25
2	experimental	5.87
3	control	3.83
4	experimental	8.69
5	control	6.16
26	control	4.42
27	experimental	4.27
28	control	2.29
29	experimental	3.78
30	experimental	5.13

⋮

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

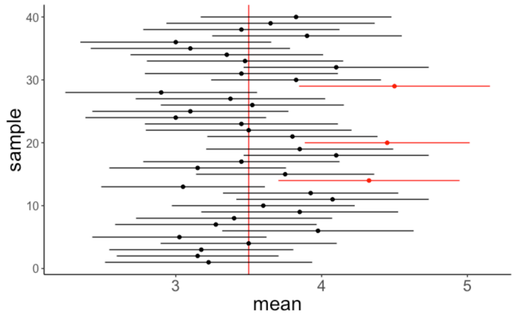
p-value

61

95% confidence interval

Definition

"If we were to repeat the experiment over and over, then 95% of the time the confidence interval contains the estimate of interest."



Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust Misinterpretation of Confidence Intervals. *Psychonomic Bulletin & Review*, 21(5), 1157-1164.

24

confidence interval

The general procedure

1. Define H_0 as Model C (compact) and H_1 as Model A (augmented)
2. Fit model parameters to the data
3. **Calculate the proportional reduction of error (PRE) in our sample**
4. Decide whether the augmented model is **worth it** by comparing the observed PRE in our sample to the sampling distribution of PRE (assuming that H_0 is true)

model comparison

Philosophy behind frequentist and **Bayesian** stats

Clue guide to probability

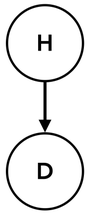
$$p(B \mid A) = \frac{p(A \mid B) \cdot p(B)}{p(A)}$$

$$\text{posterior } p(H \mid D) = \frac{\overbrace{p(D \mid H)}^{\text{likelihood}} \cdot \overbrace{p(H)}^{\text{prior}}}{p(D)}$$

subjective probability interpretation

H = Hypothesis

D = Data



formal framework for learning from data

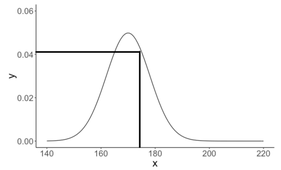
updating our prior belief $p(H)$, to a posterior belief $p(H \mid D)$ given some data

Bayes' theorem

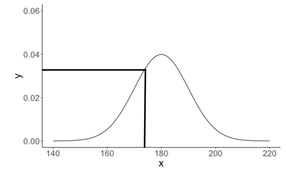
Summer camp

prior $p(\text{chess}) = \frac{1}{3}$ $p(\text{basketball}) = \frac{2}{3}$

likelihood



`dnorm(175, mean = 170, sd = 8)` = 0.041



`dnorm(175, mean = 180, sd = 10)` = 0.035

posterior

$$p(\text{basketball} \mid 175) = \frac{p(175 \mid \text{basketball}) \cdot p(\text{basketball})}{p(175 \mid \text{basketball}) \cdot p(\text{basketball}) + p(175 \mid \text{chess}) \cdot p(\text{chess})}$$

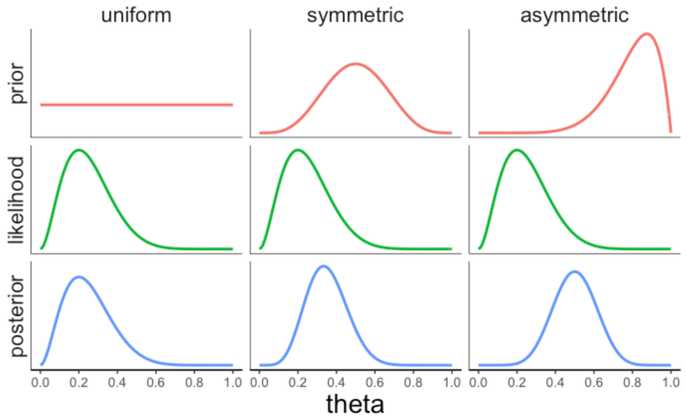
$$p(\text{basketball} \mid 175) = \frac{0.035 \cdot 2/3}{0.035 \cdot 2/3 + 0.041 \cdot 1/3} \approx 0.63$$

send the kid to the basketball gym!

Bayesian inference

The effect of the **prior**

same data, different priors



prior, likelihood, posterior

Recipe for Bayesian analysis with `brms`

visualization is everywhere!

1. Visualize the data
2. Specify and fit the model
3. Model evaluation
   a) Did the inference work?
   b) Visualize model predictions
4. Interpret the model parameters
5. Test specific hypotheses
6. Report results

Bayesian data analysis

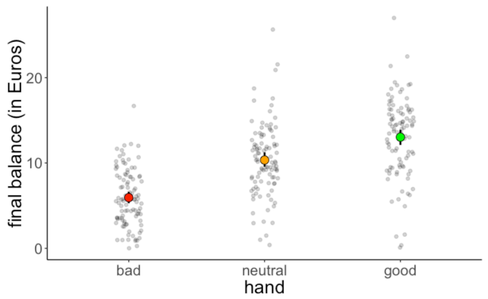
Formulate research questions as statistical models

Do better hands win more money?

participant	skill	hand	limit	balance
1	expert	bad	fixed	4.00
2	expert	bad	fixed	5.55
26	expert	bad	none	5.52
27	expert	bad	none	8.28
51	expert	neutral	fixed	11.74
52	expert	neutral	fixed	10.04
76	expert	neutral	none	21.55
77	expert	neutral	none	3.12
101	expert	good	fixed	10.86
102	expert	good	fixed	8.68

hand = {bad, neutral, good}

Visualize the data first

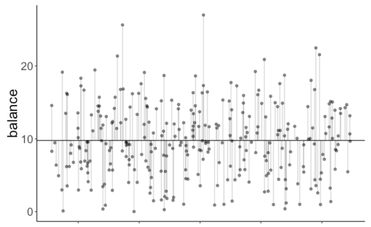


H_0: Card quality does not affect the final balance.

Model C

$$\text{balance}_i = \beta_0 + \epsilon_i$$

Model prediction



SSE(C) = 7580

Fitted model

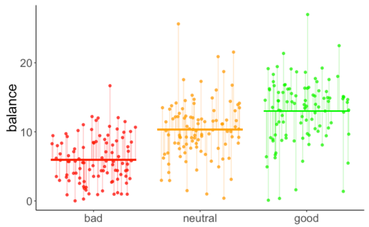
$$\widehat{\text{balance}}_i = 9.77$$

H_1: Card quality affects the final balance.

Model A

$$\text{balance}_i = \beta_0 + \beta_1 \text{hand\_neutral}_i + \beta_2 \text{hand\_good}_i + \epsilon$$

Model prediction



SSE(A) = 5021

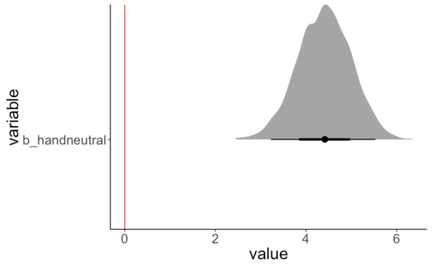
Fitted model

$$\widehat{\text{balance}}_i = 5.94 + 4.41 \cdot \text{hand\_neutral}_i + 7.08 \cdot \text{hand\_good}_i$$

31

Asking questions based on the posterior

Do neutral hands earn more money than bad hands?



What's the probability that handneutral is less than 0?

```
1 hypothesis(fit.brm,
2            hypothesis = "handneutral < 0")
```

p = 0

Communicate what you've learned about your data

2.1: Regressions (1.5 point)

Run three linear regression models that estimate the relationships between `quarter_of_birth` and `log_weekly_wage`, `education` and `log_weekly_wage`, and `quarter_of_birth` and `education`. Then run a multiple regression model that predicts `log_weekly_wage` based on both `quarter_of_birth` and `education`. Print the summaries of all four regressions and comment on their significances.

```
### YOUR CODE HERE ###
lm1 = lm(formula = log_weekly_wage ~ quarter_of_birth,
         data = df.qob)
lm2 = lm(formula = log_weekly_wage ~ education,
         data = df.qob)
lm3 = lm(formula = education ~ quarter_of_birth,
         data = df.qob)

lm_multi = lm(formula = log_weekly_wage ~ quarter_of_birth + education,
              data = df.qob)

summary(lm1)
```

```
Call:
lm(formula = log_weekly_wage ~ quarter_of_birth, data = df.qob)

Residuals:
    Min      1Q  Median      3Q     Max
-8.2511 -0.2629  0.0605  0.3720  4.2099

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.887183   0.007114 827.584   <2e-16 ***
quarter_of_birth 0.007366   0.002913   2.529   0.0115 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6908 on 49998 degrees of freedom
Multiple R-squared:  0.0001279, Adjusted R-squared:  0.0001079
F-statistic: 6.394 on 1 and 49998 DF,  p-value: 0.01145
```

homework/midterm

A catchy project title goes here

My team's name goes here

The team members' names go here

2021-03-19 12:44:19

The final project is due on **Friday, March 20th at 8pm**.

Here are some guidelines:

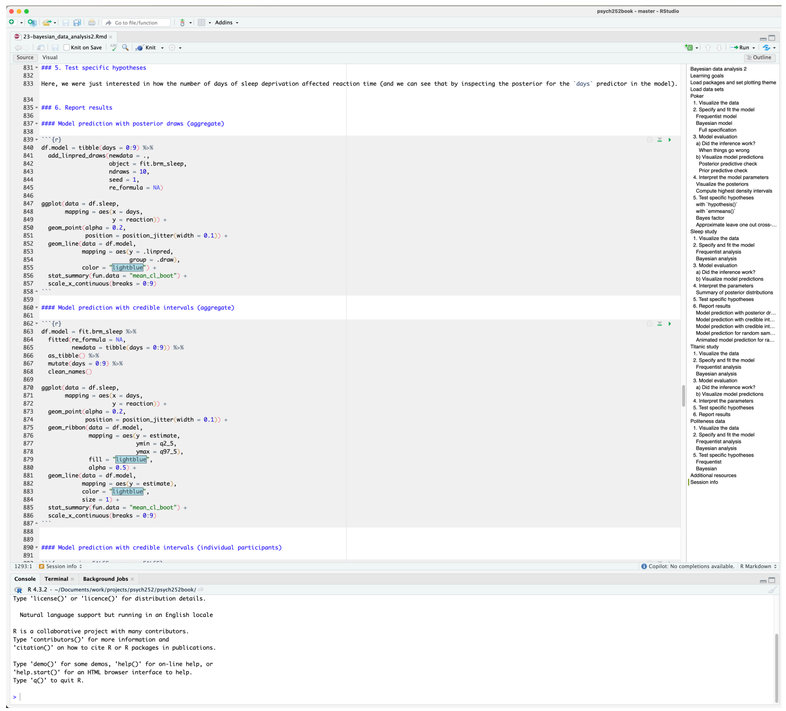
- The length of the final report should be around 2000 words per person in the group.
- All the code should be contained in this RMarkdown file (from reading in the messy data file, to making beautiful plots).
- Feel free to make the final report look like an actual paper. So you can hide all the code chunks that do data wrangling etc. from the output (by setting the code chunk option to `echo=F`), and only show the figures and tables that you need to explain your work.
- Show us what you've learned :) We're excited to read it!

For more information on how to do stuff in RMarkdown, check out:

- bookdown documentation
- RMarkdown code chunks
- RMarkdown cheat sheet
- Citations

final project
(proposal, presentation, report)

Contribute to open and reproducible science



RStudio

final-projects Private template

Edit Pins · Unwatch 4 · Fork 0 · Star 0 · Use this template

master · 1 Branch · 0 Tags · Go to file t · Add file · Code

tobiasgerstenberg updated due dates — 8a7191e · 3 weeks ago · 16 Commits

code/R	updated due dates	3 weeks ago
data	initial commit	5 years ago
figures	some updates to final project starter code	4 years ago
papers	initial commit	5 years ago
presentation	initial commit	5 years ago
writeup	updated dates and minor wording changes	2 years ago
.gitignore	initial commit	5 years ago
README.md	some updates to final project starter code	4 years ago

README

Final project

Starter code for your final project.

General points

- for folder and file names:
  - don't use white space in either folder or filenames, use an underscore "_" instead
  - (almost always) use lower case only
- always use relative paths in your code
  - for example, to save a figure from an R script inside the `code/R/` folder the path should be "../../figures/figure_name.pdf"
- keep your folder structure organized
  - we recommend adhering to the folder structure in this repository
  - more complex projects may have additional folders such as `videos/`, `papers/`, ...
- note: some of the folders are empty except for a `.keep` file
  - the `.keep` file is just there to make sure that github includes the otherwise empty folder
  - feel free to delete the `.keep` file once you've added another file to that folder

Repository structure

```
├── code
│   └── R
├── data
├── figures
├── papers
├── presentation
└── writeup
    ├── final_report
    └── proposal
```

About

Starter code for the final project

Readme
Activity
Custom properties
0 stars
4 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Contributors 2

tobiasgerstenberg Tobias Gerstenberg
sarahawu Sarah A. Wu

Languages

HTML 99.4% · TeX 0.6%

github

What we've covered

visualization and data wrangling

probability, simulation, causality

linear model

logistic regression

power analysis

model comparison

linear mixed effects models

Bayesian data analysis

Day	Date	Topic
Monday	January 6th	Introduction
Wednesday	January 8th	Visualization 1
Friday	January 10th	Visualization 2
Monday	January 13th	Data wrangling 1
Wednesday	January 15th	Data wrangling 2
Friday	January 17th	Probability
Monday	January 20th	Martin Luther King Jr. Day
Wednesday	January 22nd	Simulation 1
Friday	January 24th	Simulation 2
Monday	January 27th	Modeling data
Wednesday	January 29th	Linear model 1
Friday	January 31st	Linear model 2
Monday	February 3rd	Linear model 3
Wednesday	February 5th	Linear model 4
Friday	February 7th	Generalized linear model
Monday	February 10th	Power analysis
Wednesday	February 12th	No class (due to Midterm)
Friday	February 14th	Model comparison
Monday	February 17th	President's Day
Wednesday	February 19th	Linear mixed effects models 1
Friday	February 21st	Linear mixed effects models 2
Monday	February 24th	Linear mixed effects models 3
Wednesday	February 26th	Linear mixed effects models 4
Friday	February 28th	Causation
Monday	March 3rd	Bayesian data analysis 1
Wednesday	March 5th	Bayesian data analysis 2
Friday	March 7th	Bayesian data analysis 3
Monday	March 10th	Bayesian data analysis 4
Wednesday	March 12th	Summary and course outlook
Friday	March 14th	TA presentations

Nilam's 5 highlights

Collaborative Learning/Teaching

5.

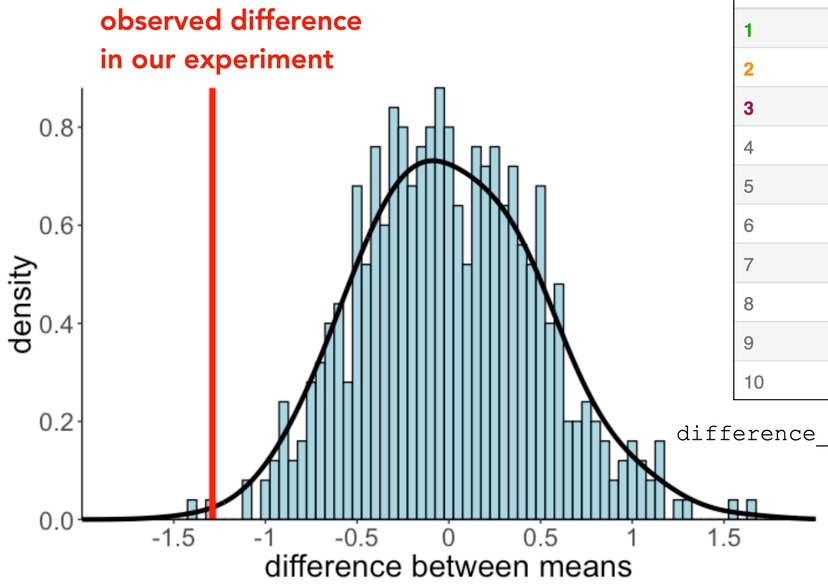
Sooo many gains!

- New ways of coding / thinking
- Up-weighting other parts of process
- Openness to try
- Patience
- Deep kindness

Thank you!

Simulation: Develop and Test ideas quickly

Permutation test



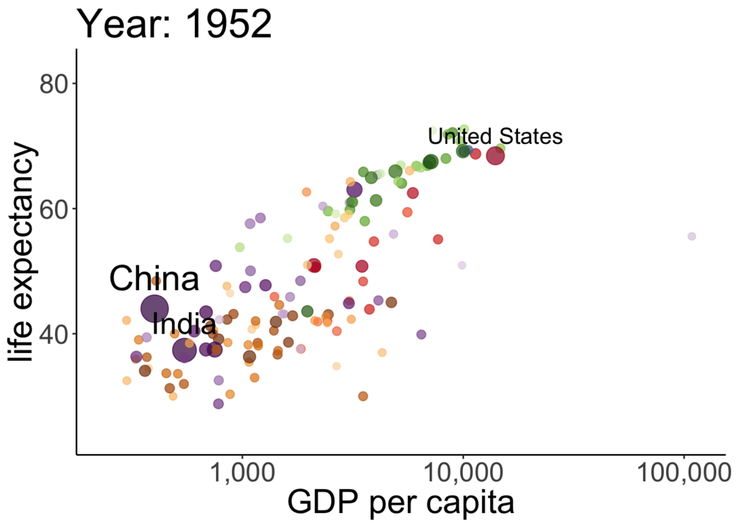
df.permutations

permutation	mean_difference
1	-0.88
2	-0.26
3	-0.94
4	0.47
5	-0.28
6	1.15
7	0.98
8	0.38
9	-0.08
10	0.31

difference_actual = -1.28

```
1 #calculate p-value of our observed result
2 df.permutations %>%
3   summarize(p_value = sum(mean_difference <= difference_actual)/n())
```

p-value = .002



visualization

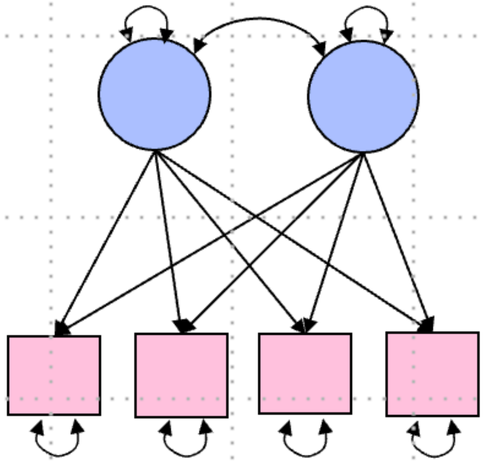
simulation

... a way of understanding

data wrangling

Fluidity across modalities: Code, Math, Reporting

Graphical



Mathematical – *lingua franca of science*

$$Y_{ti} = \beta_{0i} + \beta_{1i} time_{ti} + e_{ti}$$

$$\beta_{0i} = \gamma_{00} + u_{0i}$$

$$\beta_{1i} = \gamma_{10} + u_{1i}$$

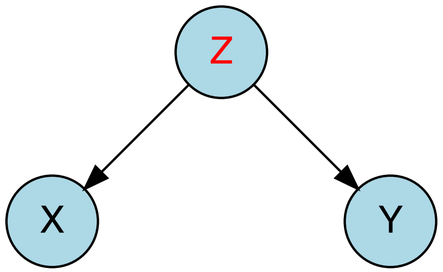
Programming Scripts

```
fit1 <- lme(fixed= y ~ 1 + time,
            random= ~ 1 + time|id,
            data= mydata)
summary(fit1)

fit2 <- lme(fixed= y ~ 1 + time + x,
            random= ~ 1 + time + x|id,
            data= mydata)
summary(fit2)
```

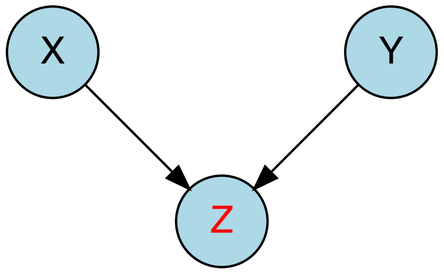
Narrative Prose – Text

We employed a multilevel model of change with time since WBRT as a predictor to determine the average trajectory of WMC accumulation. This model was then expanded to identify factors (i.e., age, WBRT dose, hypertension, hyperglycemia, smoking, and diabetes) influencing the rate of WMC accumulation following WBRT. Non-significant factors were trimmed. Results from the final model are given in Table 2.

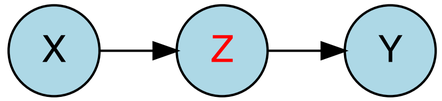
Causal Theorizing for Design & Analysis



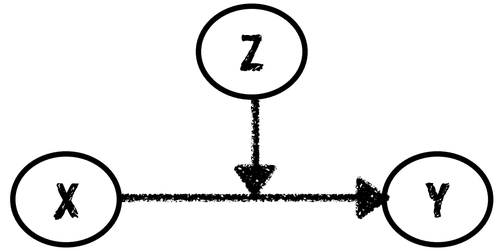
common cause



common effect

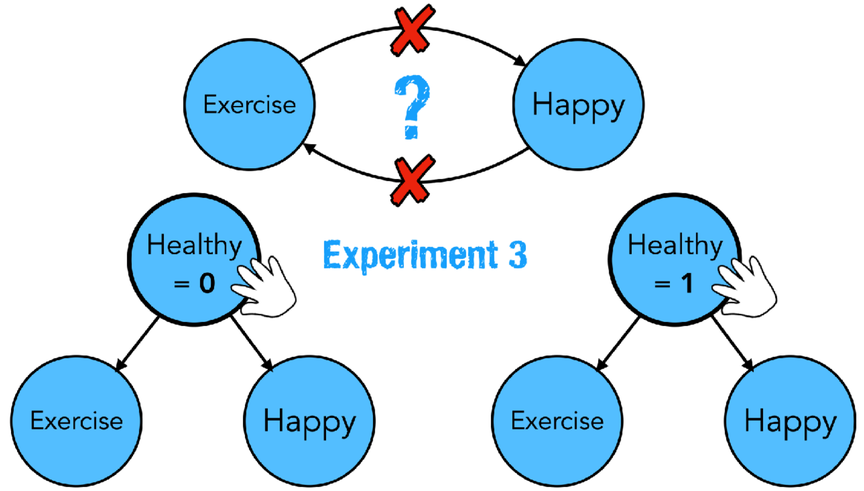


causal chain



moderation

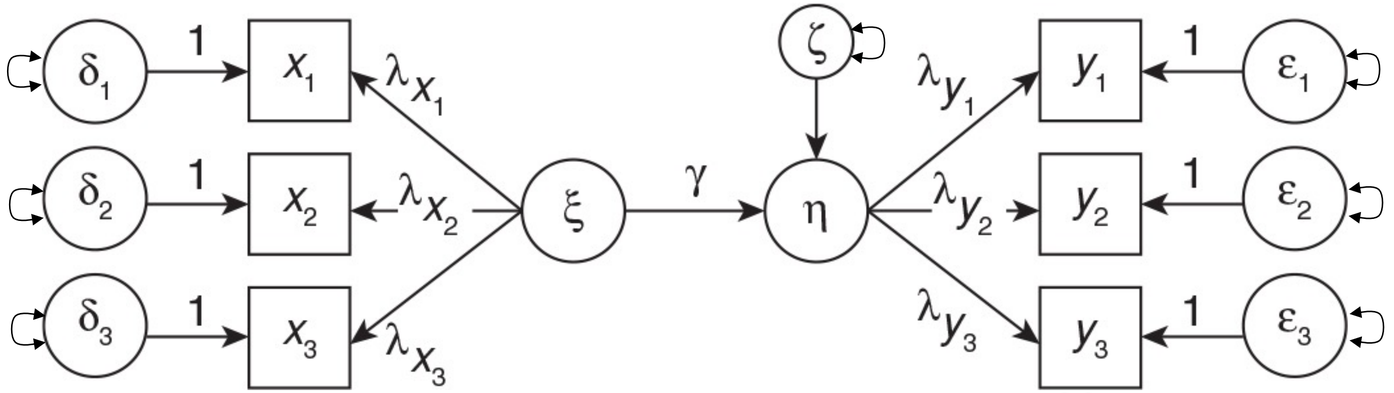
Inferring causal structure through intervention



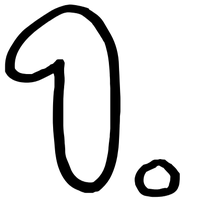
$p(\text{Happy} \mid \text{do}(\text{Healthy} = 0)) = 0.05$ $p(\text{Happy} \mid \text{do}(\text{Healthy} = 1)) = 0.5$

$p(\text{Exercise} \mid \text{do}(\text{Healthy} = 0)) = 0.1$ $p(\text{Exercise} \mid \text{do}(\text{Healthy} = 1)) = 0.75$

SEM & Latent Variable Modeling



Bayesian / Statistical Thinking & Doing

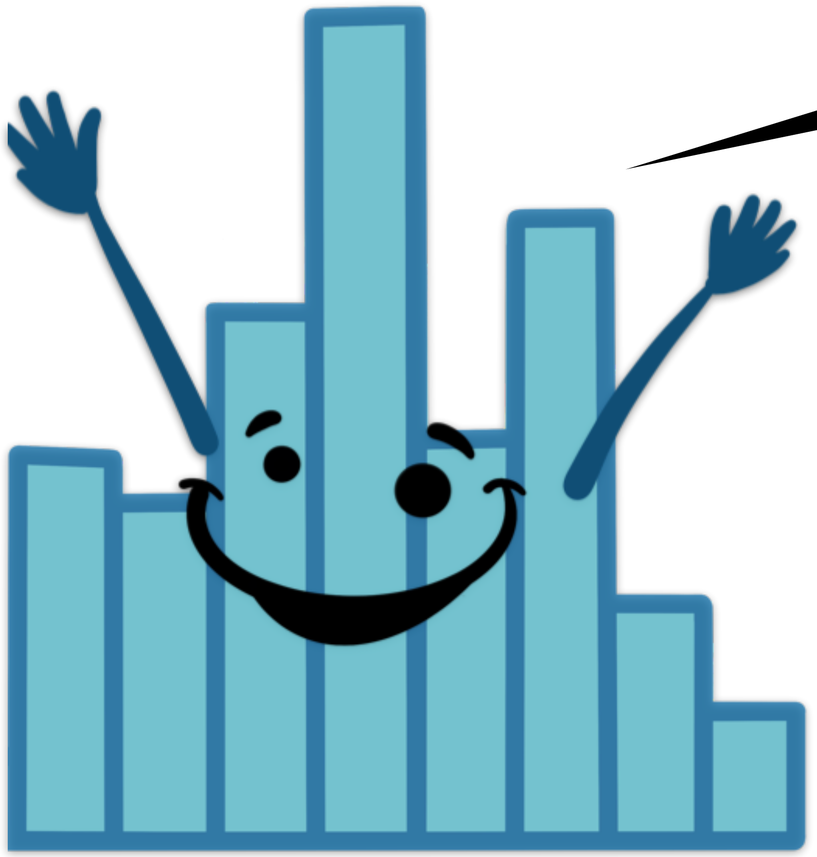


1. **Visualize the data**
2. Specify and fit the model
   a) **Define priors**
   b) Define formulas
   c) Choose estimation method
3. Model evaluation
   a) Did the inference work?
   b) Visualize model predictions
4. **Interpret the model parameters**
5. Test specific hypotheses
6. Report results
7. **Update Priors ~ Update formula ~ Update Data**

Statistical Modeling ...
... a way of developing Theories
... a way of designing Studies
... a way of interpreting Evidence

02:00

stretch break!



Tobi's 5 highlights

Consistent coding style

5.

```
# ggplot call with global aesthetics
ggplot(data = data,
       mapping = aes(x = cause,
                     y = effect)) +
  # add geometric objects (geoms)
  geom_point() +
  stat_summary(fun = "mean", geom = "point") +
  ... +
  # add text objects
  geom_text() +
  annotate() +
  # adjust axes and coordinates
  scale_x_continuous() +
  scale_y_continuous() +
  coord_cartesian() +
  # define plot title, and axis titles
  labs(title = "Title",
       x = "Cause",
       y = "Effect") +
  # change global aspects of the plot
  theme(text = element_text(size = 20),
        plot.margin = margin(t = 1, b = 1, l = 0.5, r = 0.5, unit = "cm")) +
  # save the plot
  ggsave(filename = "super_nice_plot.pdf",
         width = 8,
         height = 6)
```

what?

how?

add some text?

"local" adjustments

"global" adjustments

save the beauty!

Interpreting parameters

H_0: The relationship between income and balance is the same for students and non-students.

H_1: The relationship between income and balance differs between students and non-students.

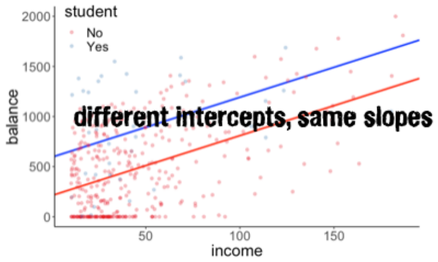
Model C

$$\text{balance}_i = \beta_0 + \beta_1 \text{income}_i + \beta_2 \text{student}_i + \epsilon_i$$

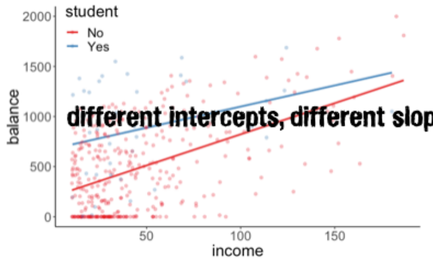
Model A

$$\widehat{\text{balance}}_i = b_0 + b_1 \text{income}_i + b_2 \text{student}_i + b_3(\text{income}_i \times \text{student}_i) + \epsilon_i$$

Model prediction



Model prediction



Fitted model

$$\widehat{\text{balance}}_i = 211.14 + 5.98 \cdot \text{income}_i + 382.67 \cdot \text{student}_i$$

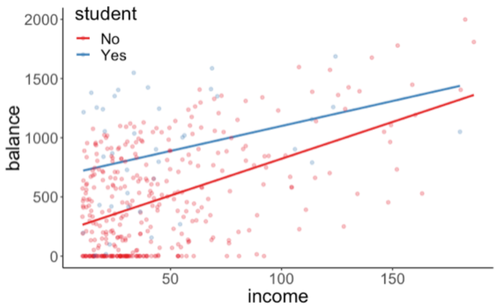
Fitted model

$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i)$$

```
Call:
lm(formula = balance ~ 1 + income + student, data = df.credit)

Residuals:
    Min      1Q  Median      3Q     Max
-762.37 -331.38  -45.04  323.60  818.28

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 211.1430    32.4572   6.505 2.34e-10 ***
income        5.9843     0.5566  10.751  < 2e-16 ***
studentYes  382.6705    65.3108   5.859 9.78e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.8 on 397 degrees of freedom
Multiple R-squared:  0.2775, Adjusted R-squared:  0.2738
F-statistic: 76.22 on 2 and 397 DF,  p-value: < 2.2e-16
```

effect of income for non-students

Interpretation



$$\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot \text{student}_i - 2.00 \cdot (\text{income}_i \times \text{student}_i) + 0$$

if student = "No" $\widehat{\text{balance}}_i = 200.62 + 6.22 \cdot \text{income}_i + (476.68 \cdot 0) - 2.00 \cdot (\text{income}_i \times 0)$

if student = "Yes"

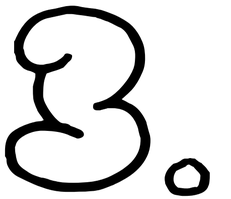
$$\begin{aligned}\widehat{\text{balance}}_i &= 200.62 + 6.22 \cdot \text{income}_i + 476.68 \cdot 1 - 2.00 \cdot (\text{income}_i \times 1) \\ &= 677.3 + 6.22 \cdot \text{income}_i - 2.00 \cdot \text{income}_i \\ &= 677.3 + 4.22 \cdot \text{income}_i\end{aligned}$$

22

```
Call:
lm(formula = balance ~ income + student + income:student, data = df.credit)

Residuals:
    Min      1Q  Median      3Q     Max
-773.39 -325.70  -41.13  321.65  814.04

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       200.6232    33.6984   5.953 5.79e-09 ***
income              6.2182     0.5921  10.502  < 2e-16 ***
studentYes        476.6758   104.3512   4.568 6.59e-06 ***
income:studentYes  -1.9992     1.7313  -1.155    0.249
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 391.6 on 396 degrees of freedom
Multiple R-squared:  0.2799, Adjusted R-squared:  0.2744
F-statistic:  51.3 on 3 and 396 DF,  p-value: < 2.2e-16
```

effect of income for non-students

`joint_tests()`

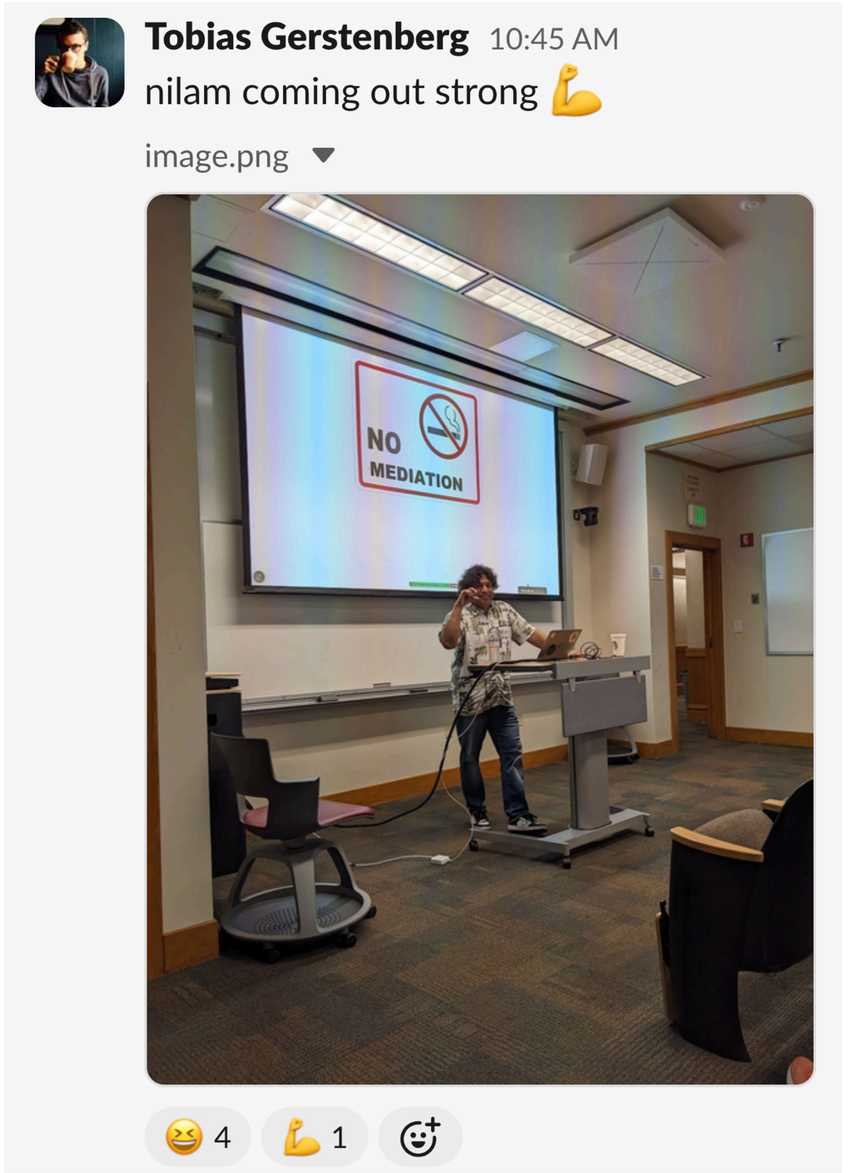


Unbalanced design

- There are different kinds of ANOVAs, for which the sums of squares are calculated differently.
- This makes a difference when we have an unbalanced design (i.e. the number of participants is not the same for each cell in our design).

`joint_tests()` is your friend!

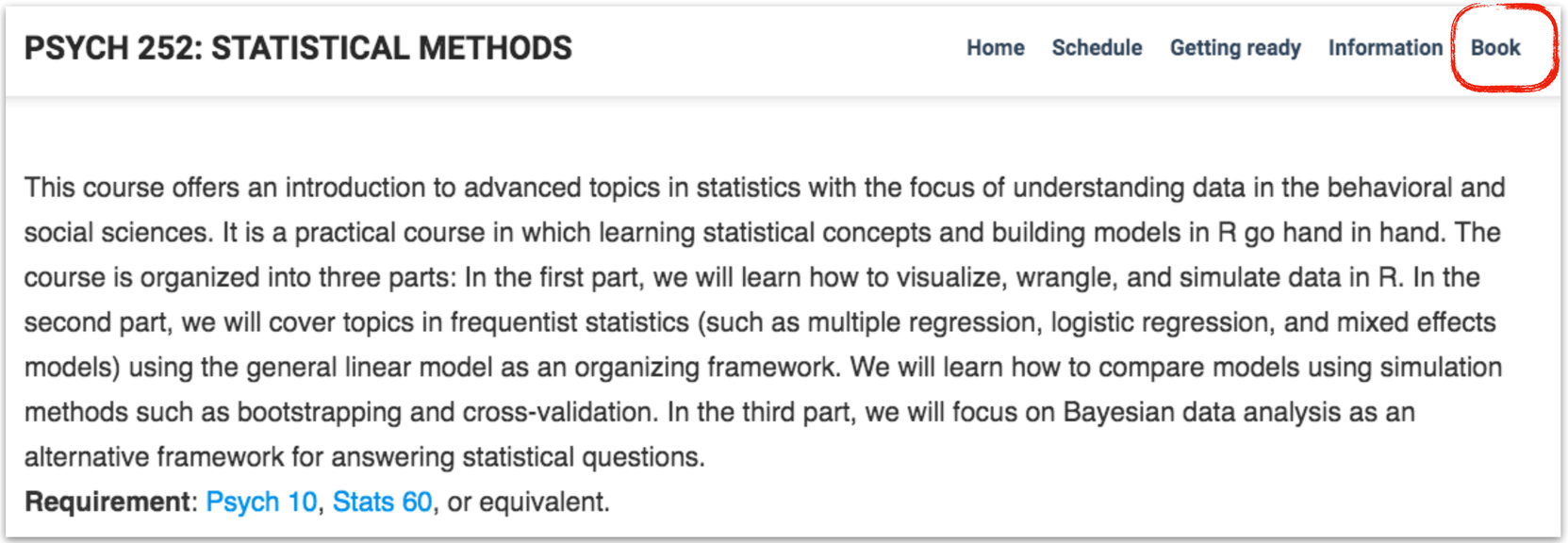
Causation & mediation

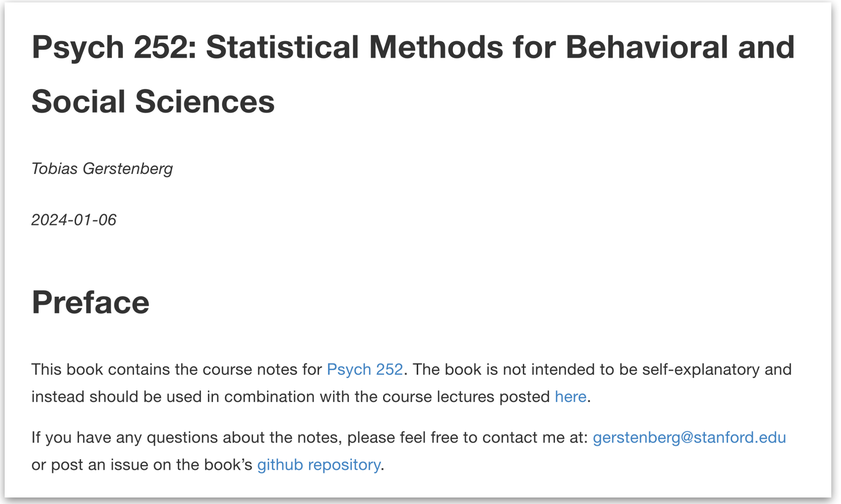


Bayesian data analysis

What shall I do now?

We'll keep updating the course notes!

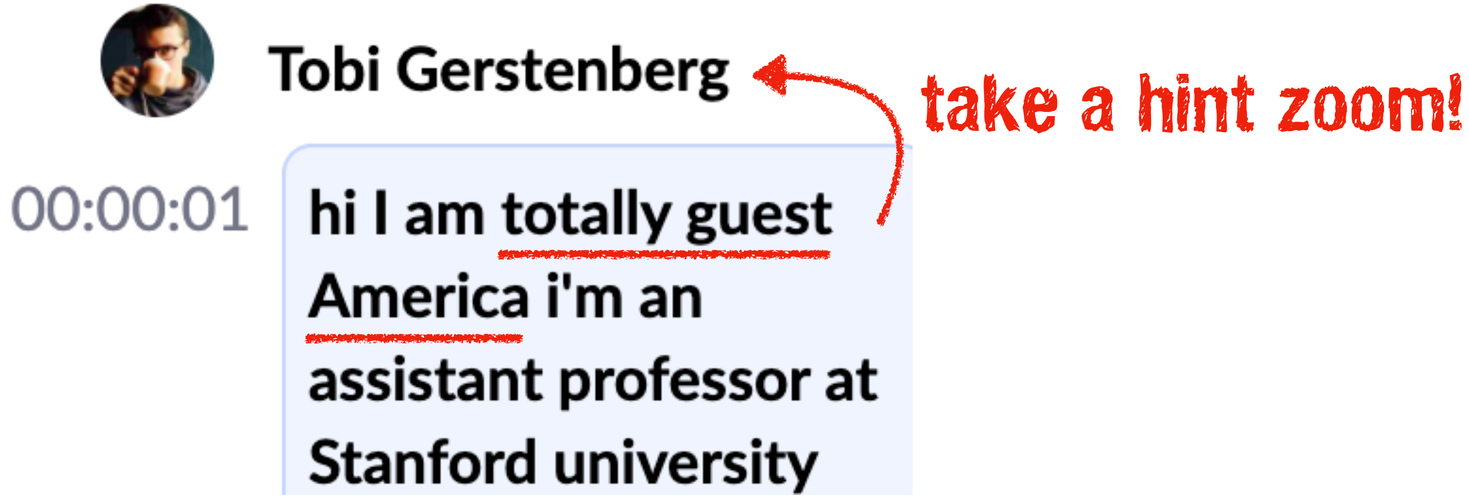




https://psych252.github.io/

you'll still have access to the lecture recordings

including the super accurate captions!



Datacamp (available until ~ mid june)

Take some more methods classes

PSYCH 289: Longitudinal Data Analysis in Social Science Research (COMM 365)

This course offers a project-based orientation to methodological issues associated with the analysis of multivariate and/or longitudinal data in the social sciences. General areas to be covered include the manipulation/organization/description of the types of empirical data obtained in social science research, and the application/implementation of multivariate analysis techniques to those data. Students will, through hands-on analysis of their data, acquire experiences in the formulation of research questions and study designs that are appropriately tethered to a variety of advanced analytical methods.

Terms: Spr | Units: 3 | Repeatable for credit

Instructors: Ram, N. (PI) ; Abdelrahim, S. (TA) ; Tan, A. (TA)

Schedule for PSYCH 289

PSYCH 254A: Advanced Statistical Modeling for Behavioral and Neural Sciences

This class will teach you how to formulate, train, test, and compare your own custom statistical and mechanistic models for behavioral and neural data. The core of the class is the "universal procedure" of modern modeling that has emerged over the past 10 years, involving: (1) formulating your hypothesis space as a parameterized model, (2) optimizing model parameters to fit data with gradient methods, and (3) fairly evaluating the fitted model using cross-validation. The first part of the class will build understanding by recreating within this framework standard models you may already have encountered, such as regularized linear regression, GLMs, SVMs and logistic regression, linear mixed models, PCA and factor analysis, structural equation modeling, and simple neural networks. The second part of the class will focus on helping you workshop custom models for your own research problems. Prereqs: a working knowledge of Python programming, and Psych 251/253 (or similar courses). A few math tools will be used (derivatives and gradients, and some linear algebra), but we will help you get up to speed on these as part of the class.

Terms: Spr | Units: 3

Instructors: Yamins, D. (PI)

Schedule for PSYCH 254A

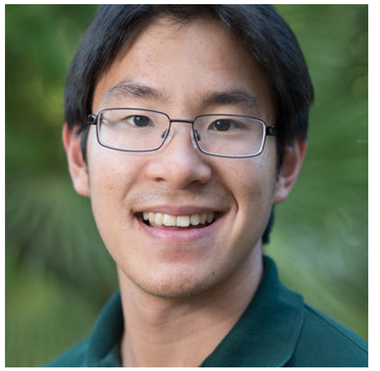
PSYCH 290: Natural Language Processing in the Social Sciences (SOC 281, SYMSYS 195T)

Digital communications (including social media) are the largest data sets of our time, and most of them are text. Social scientists need to be able to digest small and big data sets alike, process them and extract psychological insight. This applied and project-focused course introduces students to a Python codebase developed to facilitate text analysis in the social sciences (see dlatk.wwbp.org -- knowledge of Python is helpful but not required). The goal is to practice these methods in guided tutorials and project-based work so that the students can apply them to their own research contexts and be prepared to write up the results for publication. The course will provide best practices, as well as access to and familiarity with a Linux-based server environment to process text, including the extraction of words and phrases, topics, and psychological dictionaries. We will also practice the use of machine learning based on text data for psychological assessment, and the further statistic more »

Terms: Spr | Units: 3

Instructors: Eichstaedt, J. (PI) ; Lim (Chun Hui), C. (TA)

Schedule for PSYCH 290



PSYCH 139: Data Science and the Science of Learning (DATASCI 194L, DATASCI 294L, EDUC 139)

This advanced seminar will explore the application of analytic techniques from modern data science to advance the science of human learning. Students will have opportunities to work with real datasets from educational contexts and engage with contemporary research in the learning sciences, culminating in a final project. Enrollment in this course is by instructor permission only. All students interested in enrolling in this course must complete this application to be eligible to enroll: https://forms.gle/dfcm5LJReBgJ6yDf6. It is recommended that interested students submit this application by Monday, March 24th 11:59PM PT for full consideration, but applications will be reviewed on an ongoing basis through Friday, April 4th 11:59PM PT

Terms: Spr | Units: 3

Instructors: Fan, J. (PI) ; Sun, D. (PI)

Schedule for PSYCH 139

and many more ...

- **EDUC 326**: Advanced regression analyses
- **EDUC 423B**: Introduction to Data Science II: Machine learning (SOC 302B) (overview of machine learning techniques)
- **EDUC 430A**: Experimental Research Design and Analysis (learn how to do field experiments and causal inference)
- **EDUC 430B**: Quasi-Experimental Research Design & Analysis (SOC 258B) ((seeking to) get causal inference without doing experiments)
- **MS&E 226**: Fundamentals of Data Science: Prediction, Inference, Causality (a bit redundant with this class but great if you want to reinforce this knowledge and get an intro to ML)
- **MS&E 231**: Introduction to Computational Social Science (SOC 278) (I heard this was very good. it hasn't been offered for a couple years though)
- **STATS 209A**: Topics in Causal Inference (MS&E 327) (haven’t taken but seems like a good intro to causal inference)
- **STATS 216**: Introduction to Statistical Learning
- **CS 109**: Probability for computer scientists
- **CS 228**: Probabilistic Graphical Models: Principles and Techniques

What shall I not do?

Email the psych252 teaching team for stats questions I have in the future

We'd love to hear from you, but we can't help with stats questions.

Getting help with stats

for anyone

Consulting Services

The Department of Statistics offers a free online consulting service to members of the broader research community during each Stanford academic quarter.

Under the supervision of a senior faculty member, Statistics graduate students arrange Zoom meetings with clients to help with statistical research questions in areas such as:

- Experimental design and data acquisition
- Data exploration, analysis, and interpretation
- Modeling data and model fitting
- Statistical inference for estimation, testing, and prediction

Students taking statistics courses should understand that **this is not a tutoring service**.

https://statistics.stanford.edu/resources/consulting

for psych grads

Shawn Schwartz
stschwartz@stanford.edu

Ari Beller
abeller@stanford.edu

Veronica Boyce
vboyce@stanford.edu

Thanks

Teaching Team

Alice Xue	Catherine Garton	Justin Yang	Satchel Grant	Verity Lua
				

All of you!

Beleicia Benita Bullock
Bingxu Han
Bruno Lam
Jocelyn Allaina Ricard
Chase Alexander Antonacci
Carla Colina
Caroline Kaicher
Clara Maria Bacmeister
Danilo Symonette
David Barnstone
Devin Chuyi Moua
Divya Rajasekharan
Gabrianna Barcelo
Grace Brown
Isabella Caterina Aslarus
Jane Mercier Stephenson
Jeongyeon Kim
Jonah Rosemeier
Julia Annadel Donlon
Julia Proshan
Kathrine Mia Whitman
Shashanka Subrahmanya
Kavindya Thennakoon
Kevin Kennedy
Linas Marius Nasvytis
Ke Fang
Marcos Santiago Rojas Pino
Micaela Maria Bonilla
Noah Vinoya
Nora Ranck Dee
Ping-Ya Chao
Brian Lattimore
Ramya Kumar
Sarah Sampaio Izabel
Shane Muldowney
Shuman Wang
Siva Zhou
Tonya Murray
Luna Laliberte
Brenda Valdes
Yiheng Yao
Yulia Venichenko
Ziyu Ren
Hogkai Mao
Simon Huang