Scream, Shout, and yelp® for Causal Inference

PH252: Causal Inference

Asem Berkalieva
Phil Boileau
Edie Espejo
Naomi Wilcox

Presentation Overview

1. Scientific Question
2. Causal model
3. Causal question
4. Our data
5. Observed data & link to causal model
6. Identifying our parameter
7. Statistical Model and Estimand
8. Estimation
9. Interpretation
10. Limitations and future work

Scientific Question

Do Yelp reviews influence restaurant closure?

Causal Model

$$W_{age} = f_{age}(W_{type}, W_{chain}, U_{W_{age}})$$

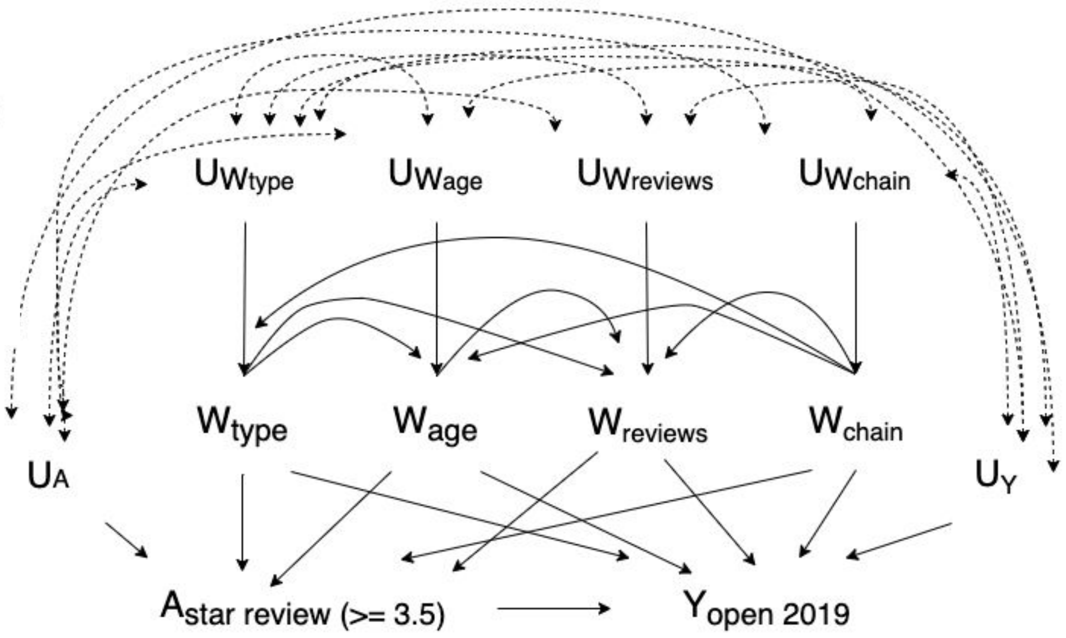
$$W_{type} = f_{type}(W_{chain}, U_{W_{type}})$$

$$W_{reviews} = f_{reviews}(W_{age}, W_{type}, W_{chain}, U_{W_{reviews}})$$

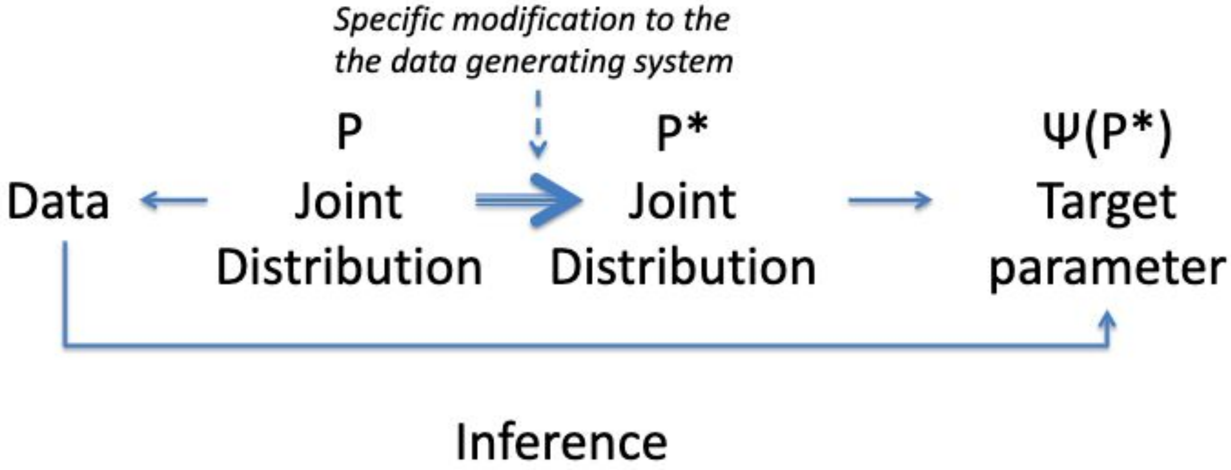
$$W_{chain} = f_{chain}(U_{W_{chain}})$$

$$A = f_A(W, U_A)$$

$$Y = f_Y(W, A, U_Y)$$



From Statistical to Causal Analysis



Causal Question

Question: What is the effect of average Yelp review on two-year survival in Las Vegas restaurants?

Intervention variable: threshold of 3.5 stars

Counterfactuals:

- Y_1: Restaurant survival at year 2 having received an average Yelp rating above or equal to 3.5 stars
- Y_0: Restaurant survival at year 2 having received an average Yelp rating below 3.5 stars

Our Parameter

Target Causal Parameter: Average treatment effect

The difference in counterfactual probability of 2 year survival had all restaurants received an average Yelp rating above or equal to 3.5 stars and the counterfactual probability of 2 year survival had all restaurants received an average Yelp rating below 3.5 stars:

$$\Psi^F = E_{U,X}[Y_1] - E_{U,X}[Y_0] = P_{U,X}(Y_1 = 1) - P_{U,X}(Y_0 = 1)$$

Observed Data and Link to Causal Model

- We assume our observed data were generated by sampling 3,644 i.i.d. times from a data generating process compatible with our causal model

$$O = (W_{age}, W_{type}, W_{reviews}, W_{chain}, A, Y) \sim \mathbb{P}_0$$

- The distribution of U and the structural equations F identify the distribution of X, and thus, the observed data
- This is the link between the causal model and the statistical model
- Statistical model is non-parametric

Data

- Our datasets
  - 2017 Yelp challenge dataset
  - 2019 Yelp challenge dataset

- Our columns
  - Stars above 3.5 on Yelp
  - Open > 2,606 days
  - Review count > 65 reviews
  - American restaurant (Yes/no)
  - Chain restaurant (Yes/no)

Table 1. Characteristics of 3,644 Las Vegas restaurants reviewed on Yelp by survival status in 2019.

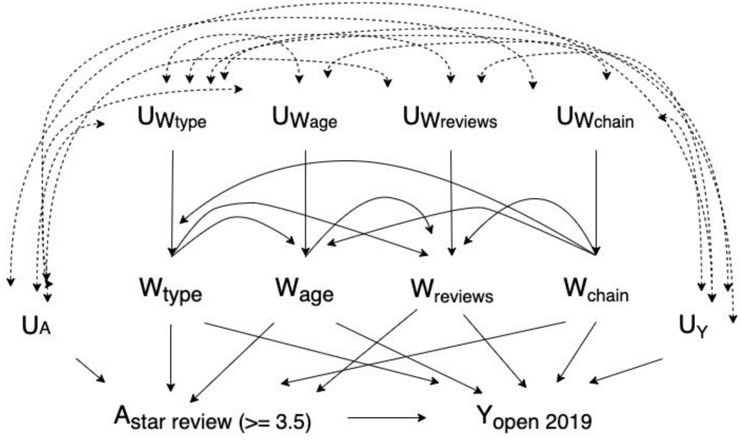
Variable n (%)	Closed in 2019 248 (7)	Open in 2019 3396 (93)	p-value
Number of stars			0.00
< 3.5	58 (4.3)	1298 (95.7)	
≥ 3.5	190 (8.3)	2098 (91.7)	
Days open			0.00
≤ 2606	158 (8.7)	1664 (91.3)	
> 2606	90 (4.9)	1732 (95.1)	
Number of reviews			0.55
≤ 65	129 (7.1)	1695 (92.9)	
> 65	119 (6.5)	1701 (93.5)	
American restaurant			0.19
No	187 (7.2)	2423 (92.8)	
Yes	61 (5.9)	973 (94.1)	
Chain restaurant			0.00
No	209 (8.9)	2151 (91.1)	
Yes	39 (3.0)	1245 (97.0)	

Values are N (%).
Fisher's exact test was used for categorical variables.
Median values were selected as cut-off points.

Identifiability

Positivity assumption: Met in theory and practice (more on this later)

Backdoor criterion: Not satisfied due to lack of independence assumptions.



Positivity assumption

Positivity assumption: Met in theory and (informally) in practice

- $P(A=a|W=w)$ is defined for all possible values (a,w) -- no zero cells
- Each treatment of interest occurs with some positive probability for each possible covariate history (though some have less variation than others)

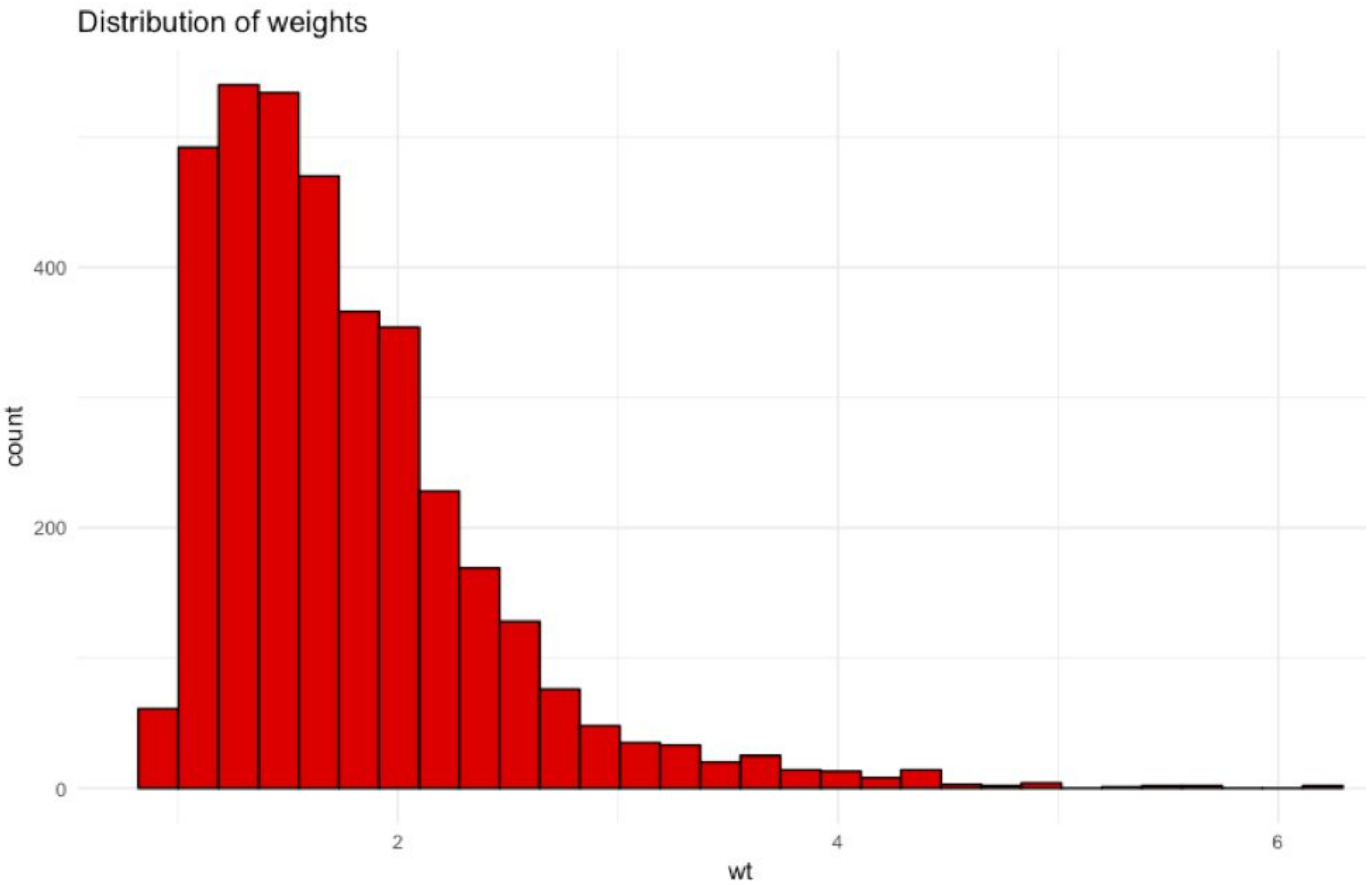


Figure 1. Distribution of weights used in IPTW estimation

Modeling

Working SCM:

- Augment SCM with additional assumptions to continue analysis
- Working SCM assumes independence of exogenous variables

Estimand:

$$\Psi(P_0) = E_W[E_0[Y|A=1,W] - E_0[Y|A=0,W]]$$

$$= E_W[Pr_0[Y=1|A=1,W] - Pr_0[Y=1|A=0,W]]$$

Estimation

Table 2. Results obtained for each estimation method

Method	*Resulting value $\Psi(P_0)$
Simple Substitution	-0.024
IPTW	-0.011
Stabilized IPTW	-0.028
TMLE	-0.027
TMLE: Asymptotic Variance	0.316
TMLE: 95% CI / p-value	(-0.045, -0.008) / p-value = 0.004

*The average effect of having a Yelp review score above or equal to 3.5 stars on the probability of two-year restaurant survival in Las Vegas.

Non-parametric Bootstrap

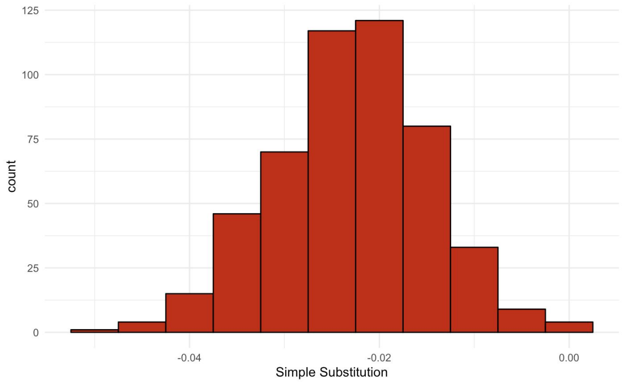


Figure 2. Bootstrapped distribution of simple substitution estimator

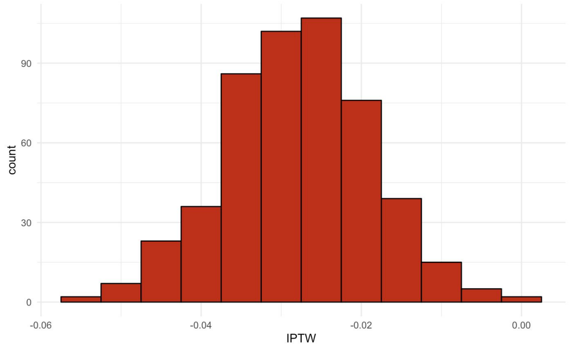


Figure 3. Bootstrapped distribution of s-IPTW estimator

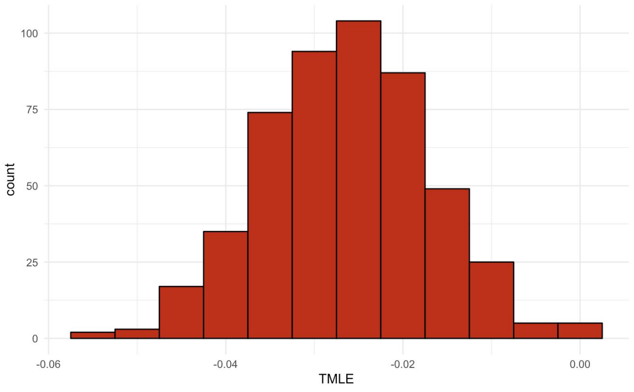
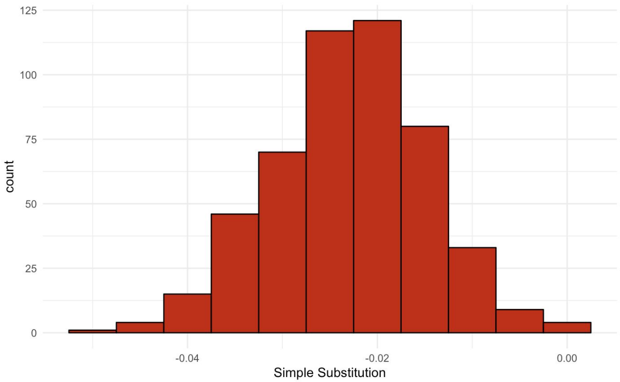
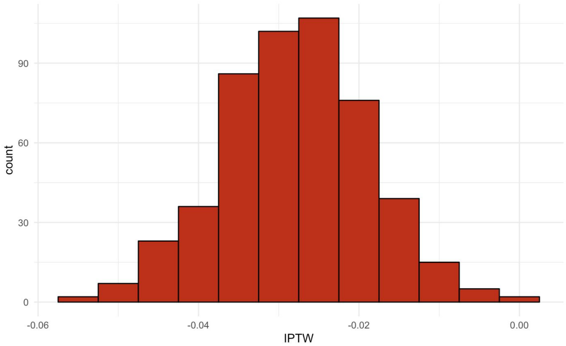
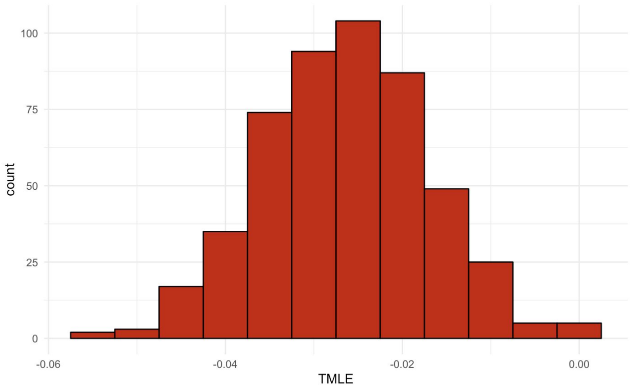


Figure 4. Bootstrapped distribution of TMLE estimator

B = 500 bootstraps


125
100
75
50
25
0
count
-0.04
-0.02
0.00
Simple Substitution

90
60
30
0
count
-0.06
-0.04
-0.02
0.00
IPTW

100
75
50
25
0
count
-0.06
-0.04
-0.02
0.00
TMLE

Non-parametric Bootstrap

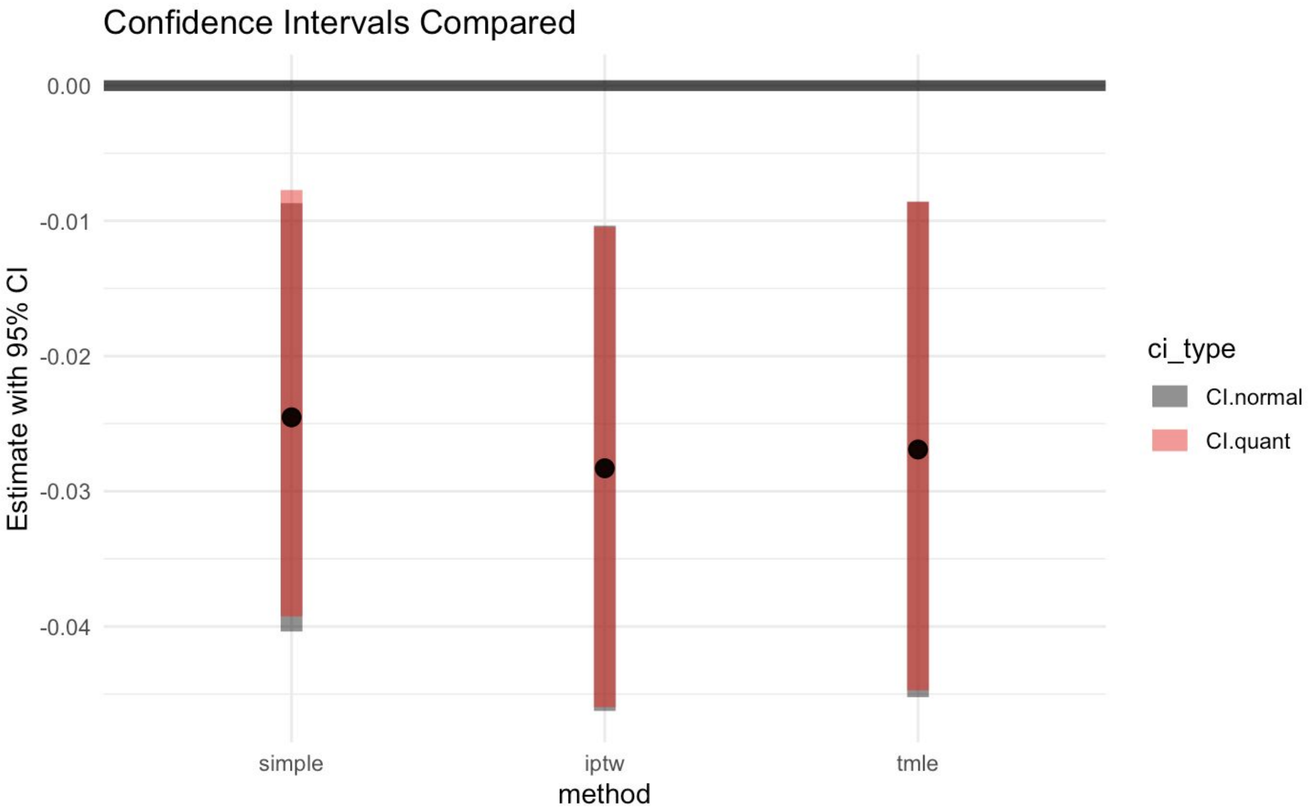


Figure 5. Bootstrap CI coverage for the three methods

Non-parametric Bootstrap

Table 3. Confidence intervals obtained for each bootstrap

Method	Normal-based confidence interval	Quantile confidence interval
Simple Substitution	(-0.0404, -0.0087)	(-0.0393, -0.0077)
Stabilized IPTW	(-0.0462, -0.0104)	(-0.0460, -0.0105)
TMLE	(-0.0452, -0.0086)	(-0.0447, -0.0086)

Statistical Interpretation

- The average effect of having a Yelp review score above or equal to 3.5 stars on the probability of two-year restaurant survival in Las Vegas is about -0.028 according to TMLE methods.

- After controlling for baseline covariates, the marginal difference in the probability of survival among restaurants with a Yelp review score above or equal to 3.5 stars and Yelp review score less than 3.5 stars was -0.028.

- Bootstrap CIs (testing the hypothesis that the effect is 0) do not contain 0.

Causal Interpretation

- If causal model + convenience assumptions are true, then:
    - Under the causal assumptions, the probability of survival is 2.8% lower if the restaurant had a Yelp review score above or equal to 3.5.

- Convenience assumptions were made

Limitations

- Treated all variables as binary
  - Above/below median is not informative for covariates with wide ranges
- Removed certain covariates (loss of information)
- Using a working model
  - Exogeneous variables have some sort of dependence structure
  - We ignored this
  - We made some convenience assumptions (no unmeasured confounding)
- Quality of data
  - Not collecting all the data we can (unmeasured covariates)
- Data is spatial
- Reviews may not be representative of restaurant quality

Conclusion

- We expected that having 3.5 stars or more on Yelp would help 2-year-survival.

- We don't think that our results are representative of the truth because:
  - We did not mine the data to incorporate spatial aspects
  - We removed a lot of information by using binary variables only

- The assumptions we had to make were too extreme for the problem at hand

Future Work

- Extend to continuous and spatial covariates
- Better understand the system that governs restaurant closure in order to make less assumptions and work with more variables
- How would our estimators vary if we chose another city?
- What would it mean to intervene on Yelp reviews in the real world?
  - Ex: Incentivize 5-star Yelp reviews with discounts

Team contributions

- Asem Berkalieva:
  - Sections 3, 5; Data preparation; Bootstrap; Interpretation
- Philippe Boileau:
  - Sections 1, 2, 6, 7; SuperLearner estimation (G-comp, IPTW, stabilized IPTW, TMLE); Interpretation
- Edie Espejo:
  - Sections 4; Data preparation; G-comp formula and TMLE estimation; Bootstrap; Interpretation
- Naomi Wilcox:
  - Sections 5; Practical positivity assumption analysis; Interpretation; DAG; Table 1

Limitations

Good Practice

–Get more data

–Do the best job you can with data you have, and understand limitations

•Formal Causal Framework:

–Which data and/or how to change design

–What additional assumptions are needed: **“convenience assumptions”?**

–Estimand that comes closest to answering our question with the data we have

•Use lack of identifiability to inform interpretation and future studies