Data-driven methods
Introduction

INFO/STSCI/ILRST 3900: Causal Inference

14 Nov 2023

Learning goals for today

At the end of class, you will be able to:

1. discuss how targeted treatments could aid policy
2. recognize the dangers of selecting on effect size
3. implement a general-purpose solution
4. evaluate an estimator by simulation

Targeting interventions

Use data to **discover subgroups** most affected by treatment

Targeting interventions

Use data to **discover subgroups** most affected by treatment

- to whom should a company show an ad?

Targeting interventions

Use data to **discover subgroups** most affected by treatment

- to whom should a company show an ad?
- on which doors should a campaign volunteer knock?

Targeting interventions

Use data to **discover subgroups** most affected by treatment

- to whom should a company show an ad?
- on which doors should a campaign volunteer knock?
- which students should be offered a scholarship?

Targeting interventions

WATCH fitbit GARMIN

A simulated setting

Randomized treatment in simulated population

A simulated setting

Randomized treatment in simulated population

X	Pre-treatment covariate	How much do you love fall? 1 = least, 10 = most

A simulated setting

Randomized treatment in simulated population

X	Pre-treatment covariate	How much do you love fall? 1 = least, 10 = most
A	Treatment (randomized)	Encouragement to take a walk

A simulated setting

Randomized treatment in simulated population

X	Pre-treatment covariate	How much do you love fall? 1 = least, 10 = most
A	Treatment (randomized)	Encouragement to take a walk
Y	Outcome	Minutes active in the day

A simulated setting

Randomized treatment in simulated population

X	Pre-treatment covariate	How much do you love fall? 1 = least, 10 = most
A	Treatment (randomized)	Encouragement to take a walk
Y	Outcome	Minutes active in the day

We want to **discover the group** most affected

A simulated setting

Randomized treatment in simulated population

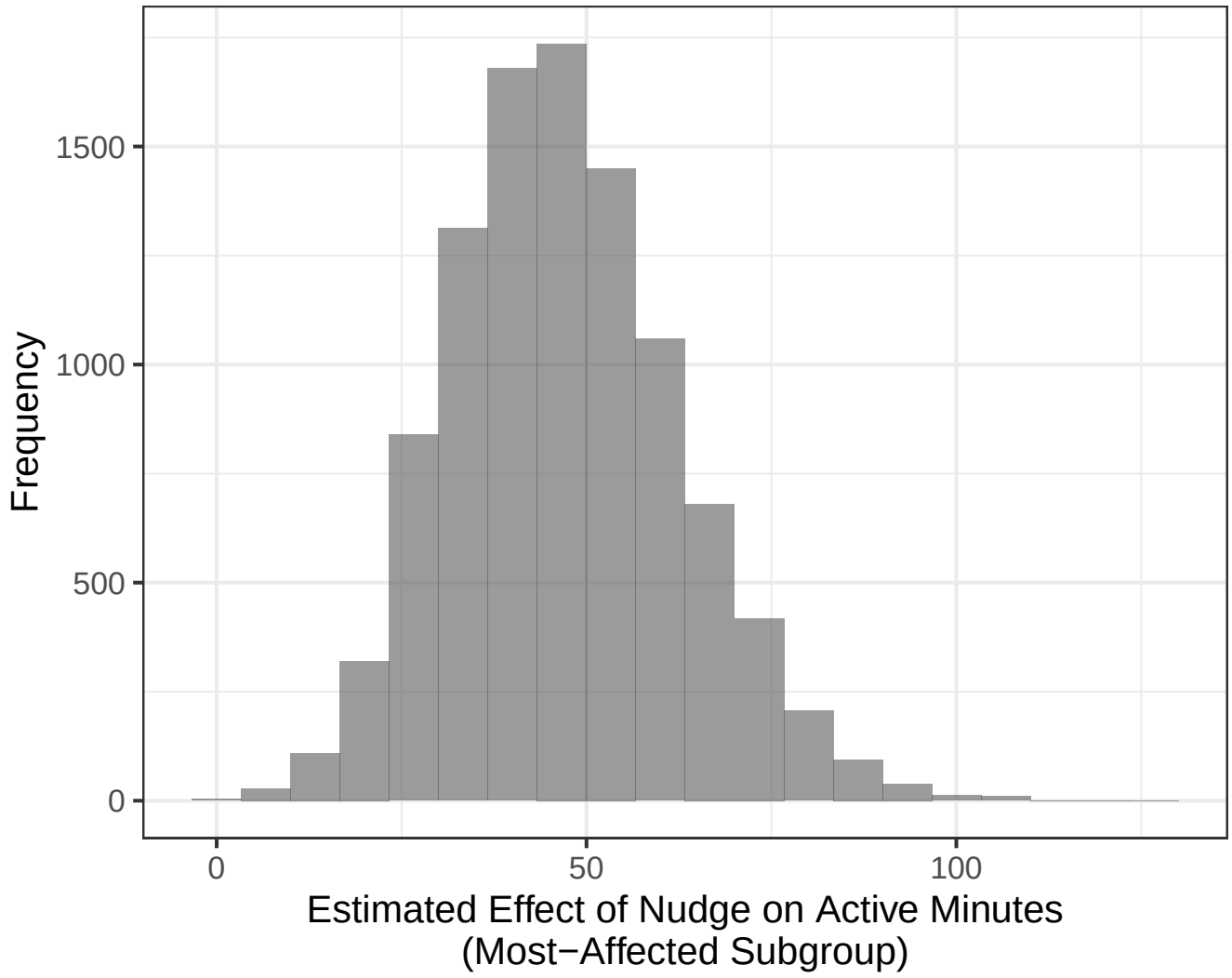
X	Pre-treatment covariate	How much do you love fall? 1 = least, 10 = most
A	Treatment (randomized)	Encouragement to take a walk
Y	Outcome	Minutes active in the day

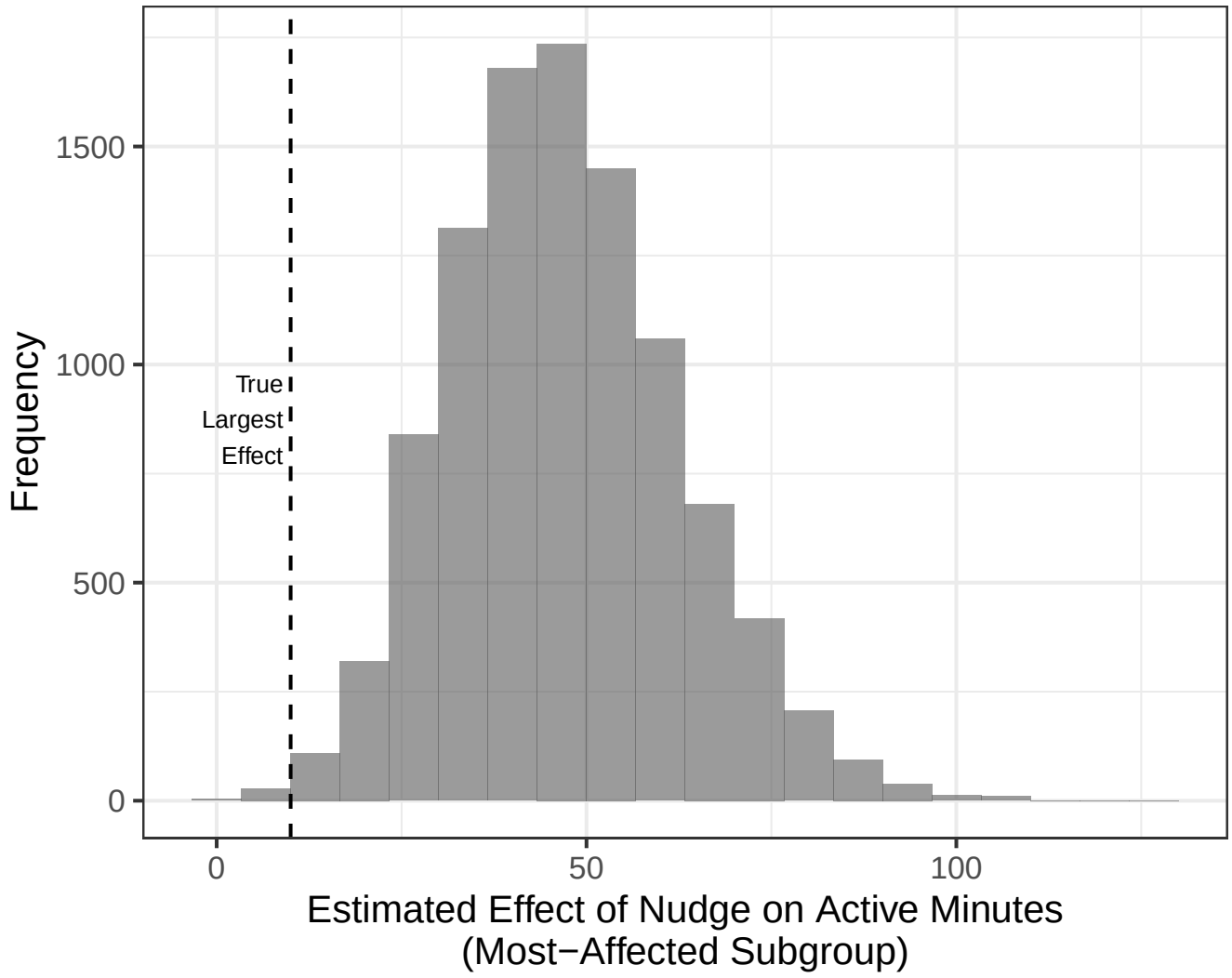
We want to **discover the group** most affected

1. Visit course website to download data
2. Draw a sample of size 50
3. Nonparametrically estimate

$$\tau_x = \mathsf{E}(Y^1 - Y^0 \mid X = x)$$

   for $x = 1, \ldots, 10$
4. Tell me your **highest** estimate


Frequency
1500
1000
500
0
0
50
100
Estimated Effect of Nudge on Active Minutes
(Most−Affected Subgroup)


Frequency
1500
1000
500
0
True
Largest
Effect
0
50
100
Estimated Effect of Nudge on Active Minutes
(Most−Affected Subgroup)

What went wrong?

- estimated 10 effects, $\hat{\tau}_x = \hat{\mathrm{E}}(Y^1 - Y^0 \mid X = x)$

What went wrong?

- estimated 10 effects, $\hat{\tau}_x = \hat{\mathrm{E}}(Y^1 - Y^0 \mid X = x)$
- each was a mixture of signal and noise

$$\hat{\tau}_x = \underbrace{\tau_x}_{\substack{\textbf{signal} \\ \text{true} \\ \text{effect}}} + \underbrace{\epsilon_x}_{\substack{\textbf{noise} \\ \text{sampling} \\ \text{variability}}}$$

What went wrong?

- estimated 10 effects, $\hat{\tau}_x = \hat{\mathrm{E}}(Y^1 - Y^0 \mid X = x)$
- each was a mixture of signal and noise

$$\hat{\tau}_x = \underbrace{\tau_x}_{\substack{\textbf{signal} \\ \text{true} \\ \text{effect}}} + \underbrace{\epsilon_x}_{\substack{\textbf{noise} \\ \text{sampling} \\ \text{variability}}}$$

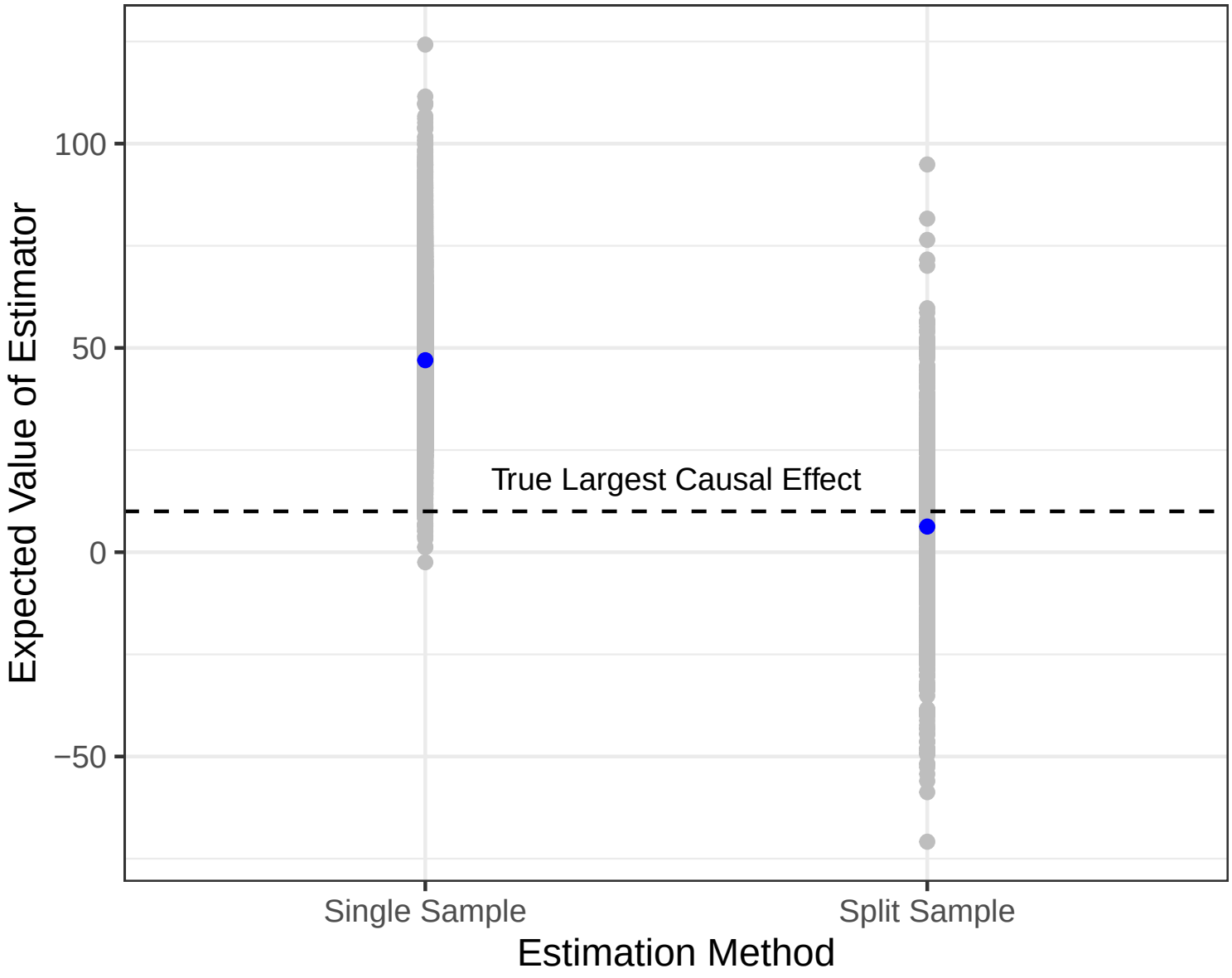
- we inadvertently picked a subgroup with high positive noise

What could we do?

1. first draw a **selection sample**
   - select the most responsive subgroup
2. then draw an **estimation sample**
   - estimate the effect for that subgroup

Why this works

The subgroup that we choose is no longer a function of randomness in the Y values we use to estimate the reported effect


Expected Value of Estimator
100
50
0
−50
True Largest Causal Effect
Single Sample
Split Sample
Estimation Method

What we just did:

Evaluate an estimator by simulation

What we just did:
Evaluate an estimator by simulation

In the real world

- we often observe only one sample
- we never observe counterfactuals

What we just did:
Evaluate an estimator by simulation

In the real world

- we often observe only one sample
- we never observe counterfactuals

In simulated data

- we can observe many samples
- we get to know the truth

What we just did:
Evaluate an estimator by simulation

Strategy when faced with a statistical problem:

What we just did:
Evaluate an estimator by simulation

Strategy when faced with a statistical problem:

1. simulate data where you know the truth

What we just did:
Evaluate an estimator by simulation

Strategy when faced with a statistical problem:

1. simulate data where you know the truth
2. apply your estimator to many simulations

What we just did:
Evaluate an estimator by simulation

Strategy when faced with a statistical problem:

1. simulate data where you know the truth
2. apply your estimator to many simulations
3. see if you recover the truth

Targeting interventions

WATCH fitbit GARMIN

Can you think of an example?

- in what setting might we want to target a treatment to responsive groups?
- how would today's discussion apply to that setting?

Learning goals for today

At the end of class, you will be able to:

1. discuss how targeted treatments could aid policy
2. recognize the dangers of selecting on effect size
3. implement a general-purpose solution
4. evaluate an estimator by simulation