In-person session 6

February 19, 2026

PMAP 8521: Program evaluation
Andrew Young School of Policy Studies

Plan for today

Validity and *p*-hacking

DAGs

p-values and confidence intervals

Validity and *p*-hacking

What's going on with the assignments?

andhs.co/nullworlds

What is *p*-hacking?

andhs.co/hack

Are the results from p-hacking actually a threat to validity?

Is a little exploratory p-hacking okay?

Do people actually post their preregistrations?

Yes!

OSF

See this and this for examples

As Predicted

See this

DAGs

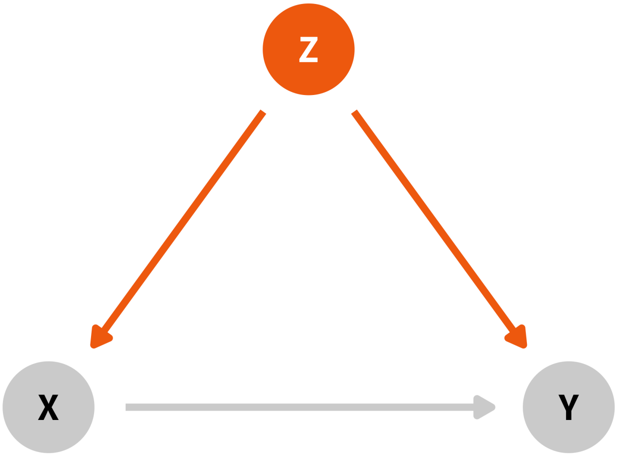
What about cycles?

Example time-based DAG

How do I know which of these is which?

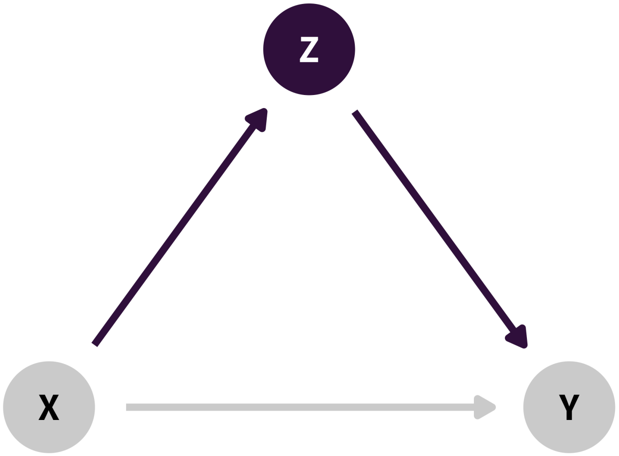
Confounder

(Fork)



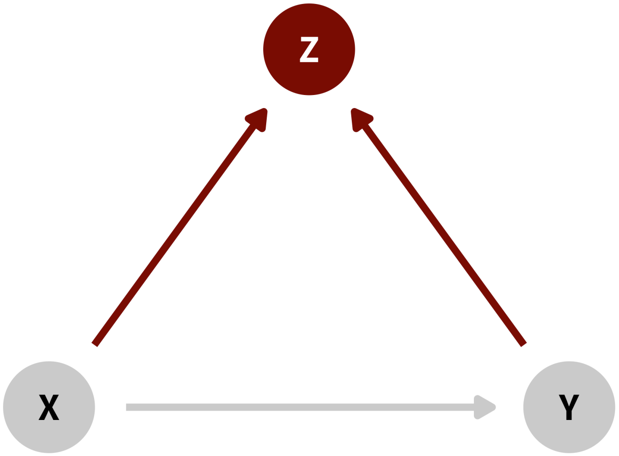
Mediator

(Chain)

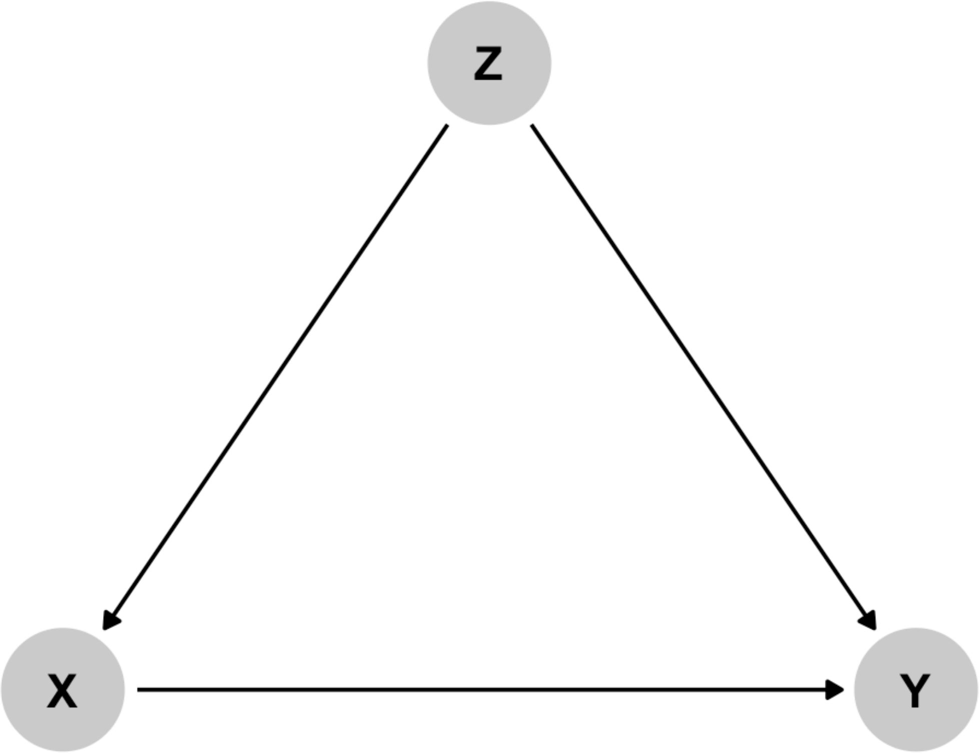


Collider

(Inverted fork)



LUTRON


Z
X
Y

andhs.co/dags

d-separation

Except for the one arrow between X and Y,
no statistical association can flow between X and Y

This is **identification**—
all alternative stories are ruled out
and the relationship is isolated

What's the difference between logic models and DAGs?

Can't I just remake my logic model in Dagitty and be done?

DAGs vs. logic models

DAGs are a *statistical* tool

Describe a data-generating process and isolate/identify relationships

Logic models are a *managerial* tool

Oversee the inner workings of a program and its theory

DAGs and identification

DAGs are a statistical tool, but they don't tell you what statistical method to use

DAGs help you with the **identification strategy**

Easiest identification

Identification through research design

RCTs

When treatment is randomized, delete all arrows going into it

Most other identification

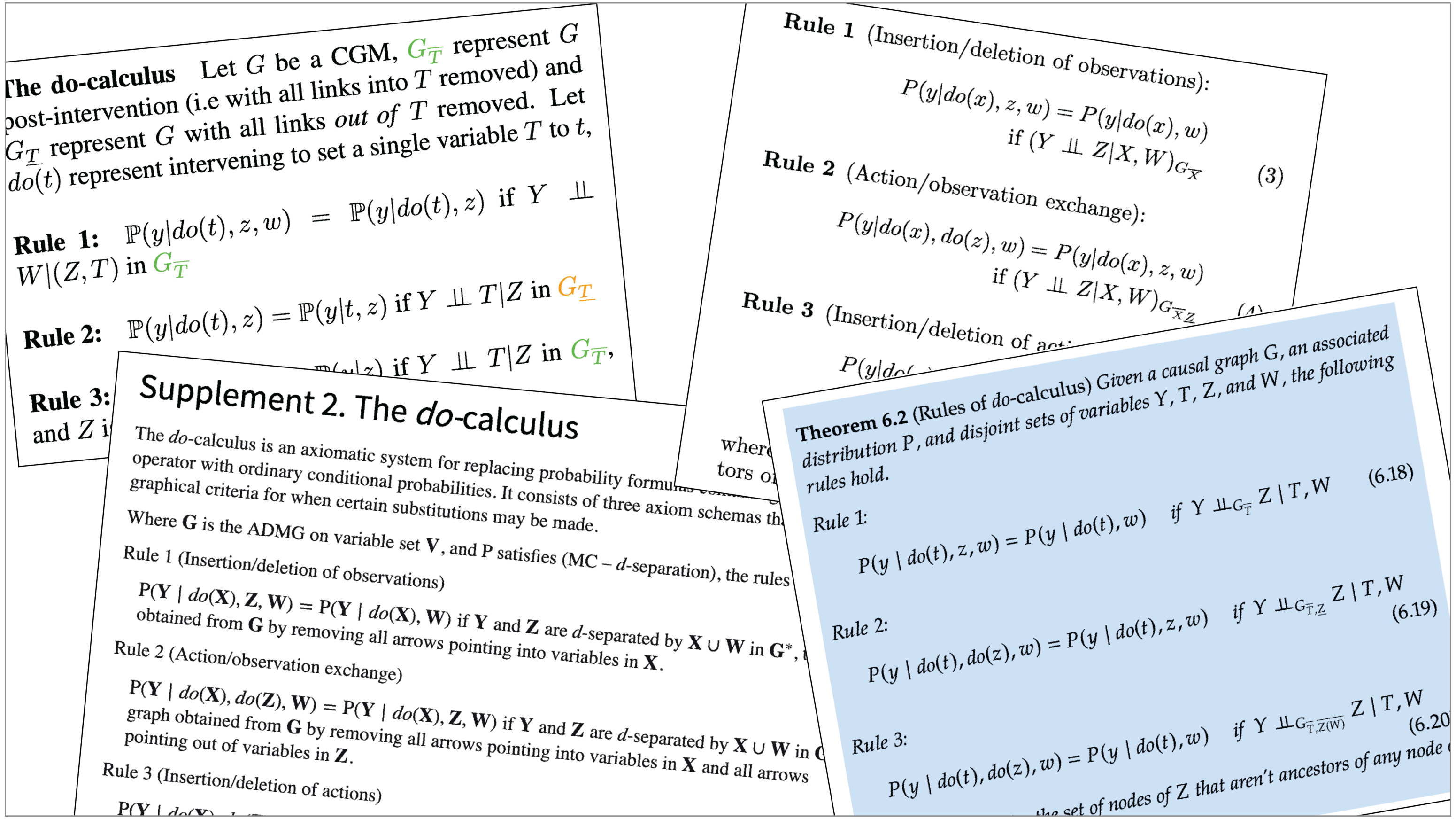
Identification through do-calculus

Rules for treating causal interventions as observations

Backdoor adjustment and frontdoor adjustment are special common patterns of do-calculus

Where can we learn more about *do*-calculus?

Here!



The do-calculus Let G be a CGM, $G_{\overline{T}}$ represent G post-intervention (i.e with all links into T removed) and $G_{\underline{T}}$ represent G with all links *out of* T removed. Let $do(t)$ represent intervening to set a single variable T to t,

Rule 1: $\mathbb{P}(y|do(t),z,w) = \mathbb{P}(y|do(t),z)$ if $Y \perp\!\!\!\perp W|(Z,T)$ in $G_{\overline{T}}$

Rule 2: $\mathbb{P}(y|do(t),z) = \mathbb{P}(y|t,z)$ if $Y \perp\!\!\!\perp T|Z$ in $G_{\underline{T}}$

Rule 3: $\mathbb{P}(y|z)$ if $Y \perp\!\!\!\perp T|Z$ in $G_{\overline{T}}$,

and Z i

Rule 1 (Insertion/deletion of observations):

$$P(y|do(x),z,w) = P(y|do(x),w) \quad \text{if } (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}}} \qquad (3)$$

Rule 2 (Action/observation exchange):

$$P(y|do(x),do(z),w) = P(y|do(x),z,w) \quad \text{if } (Y \perp\!\!\!\perp Z|X,W)_{G_{\overline{X}\underline{Z}}} \qquad (4)$$

Rule 3 (Insertion/deletion of act

$P(y|do($

where

tors o

Supplement 2. The *do*-calculus

The *do*-calculus is an axiomatic system for replacing probability formulas operator with ordinary conditional probabilities. It consists of three axiom schemas th graphical criteria for when certain substitutions may be made.

Where **G** is the ADMG on variable set **V**, and P satisfies (MC – *d*-separation), the rules

Rule 1 (Insertion/deletion of observations)

P(**Y** | *do*(**X**), **Z**, **W**) = P(**Y** | *do*(**X**), **W**) if **Y** and **Z** are *d*-separated by **X** ∪ **W** in **G***, obtained from **G** by removing all arrows pointing into variables in **X**.

Rule 2 (Action/observation exchange)

P(**Y** | *do*(**X**), *do*(**Z**), **W**) = P(**Y** | *do*(**X**), **Z**, **W**) if **Y** and **Z** are *d*-separated by **X** ∪ **W** in graph obtained from **G** by removing all arrows pointing into variables in **X** and all arrows pointing out of variables in **Z**.

Rule 3 (Insertion/deletion of actions)

P(**Y** | *do*(**X**)

Theorem 6.2 (Rules of do-calculus) *Given a causal graph G, an associated distribution P, and disjoint sets of variables Y, T, Z, and W, the following rules hold.*

Rule 1:

$$P(y \mid do(t), z, w) = P(y \mid do(t), w) \quad \textit{if } Y \perp\!\!\!\perp_{G_{\overline{T}}} Z \mid T, W \qquad (6.18)$$

Rule 2:

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), z, w) \quad \textit{if } Y \perp\!\!\!\perp_{G_{\overline{T},\underline{Z}}} Z \mid T, W \qquad (6.19)$$

Rule 3:

$$P(y \mid do(t), do(z), w) = P(y \mid do(t), w) \quad \textit{if } Y \perp\!\!\!\perp_{G_{\overline{T},\overline{Z(W)}}} Z \mid T, W \qquad (6.20)$$

the set of nodes of Z that aren't ancestors of any node

Adjusting for backdoor confounding

$$\overbrace{\mathbf{E}(y \mid \mathrm{do}(x))}^{\text{Causal effect of } x \text{ on } y} = \underbrace{\sum_{z}}_{\text{Sum across all values of } z} \overbrace{\mathbf{E}(y \mid x, z)}^{\text{Conditional mean of } y\text{, given } x \text{ and } z\ldots} \times \overbrace{\mathbf{P}(z)}^{\ldots \text{ weighted by } z}$$

When things are identified, there are still arrows leading into Y. What do we do with those? How do you explain those relationships?

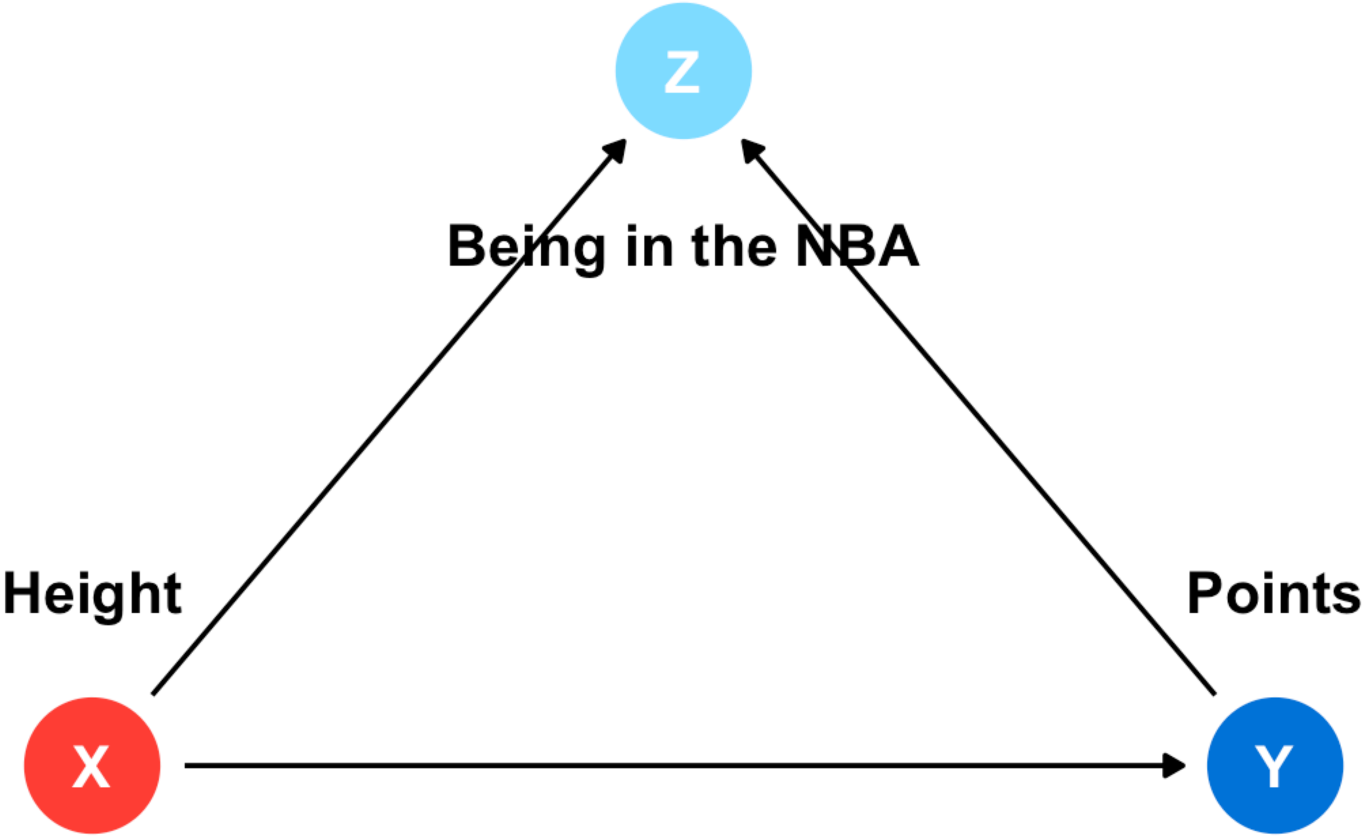
Outcomes have multiple causes. How do you justify that your proposed cause is the most causal factor?

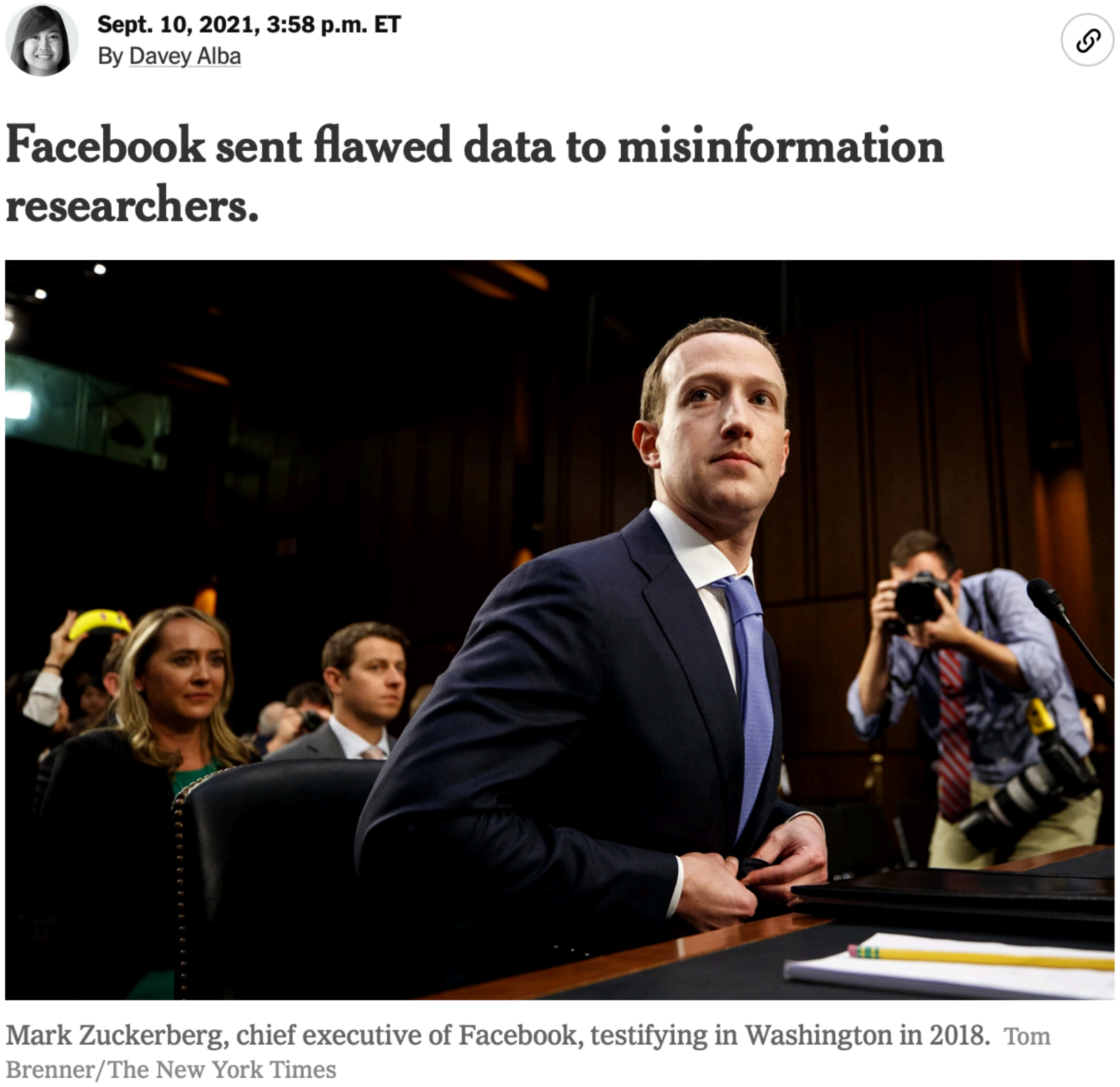
How exactly do we close backdoors?

What is front door adjustment??

How exactly do colliders mess up your results?

It looks like you can still get the effect of X on Y


Z
Being in the NBA
Height
Points
X
Y

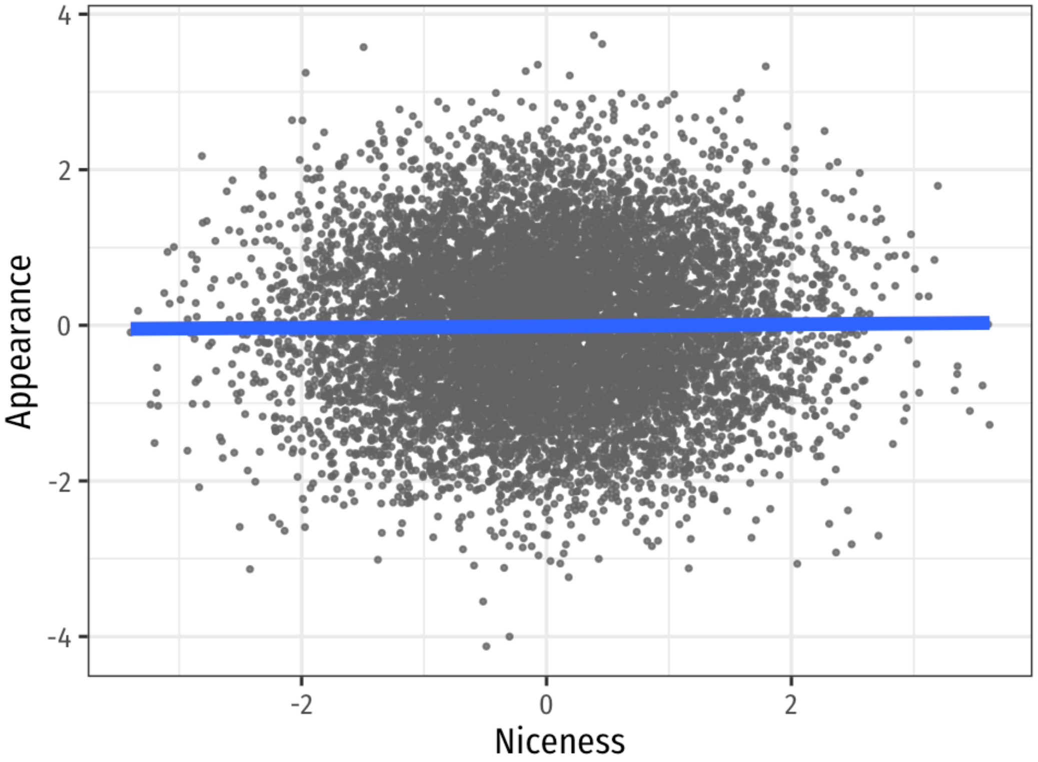


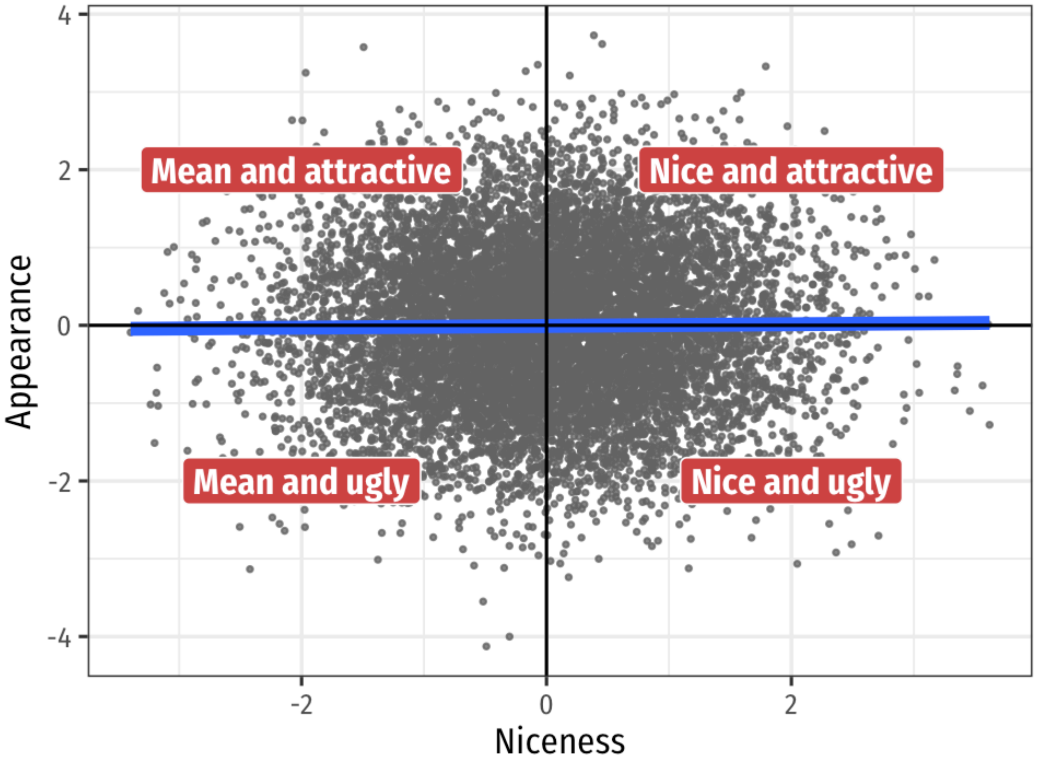
Sept. 10, 2021, 3:58 p.m. ET
By Davey Alba

Facebook sent flawed data to misinformation researchers.

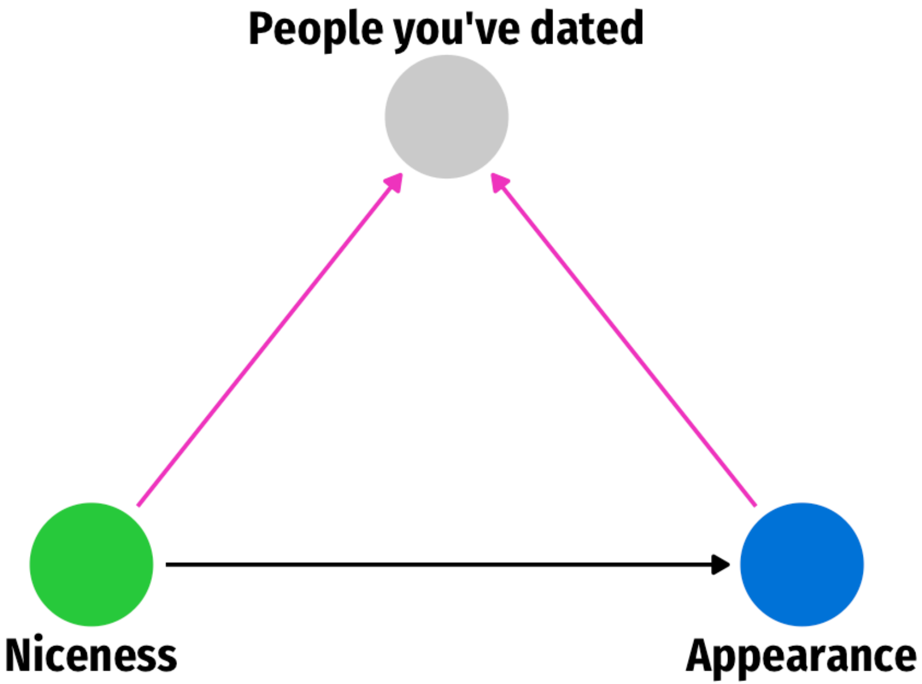
Mark Zuckerberg, chief executive of Facebook, testifying in Washington in 2018. Tom Brenner/The New York Times

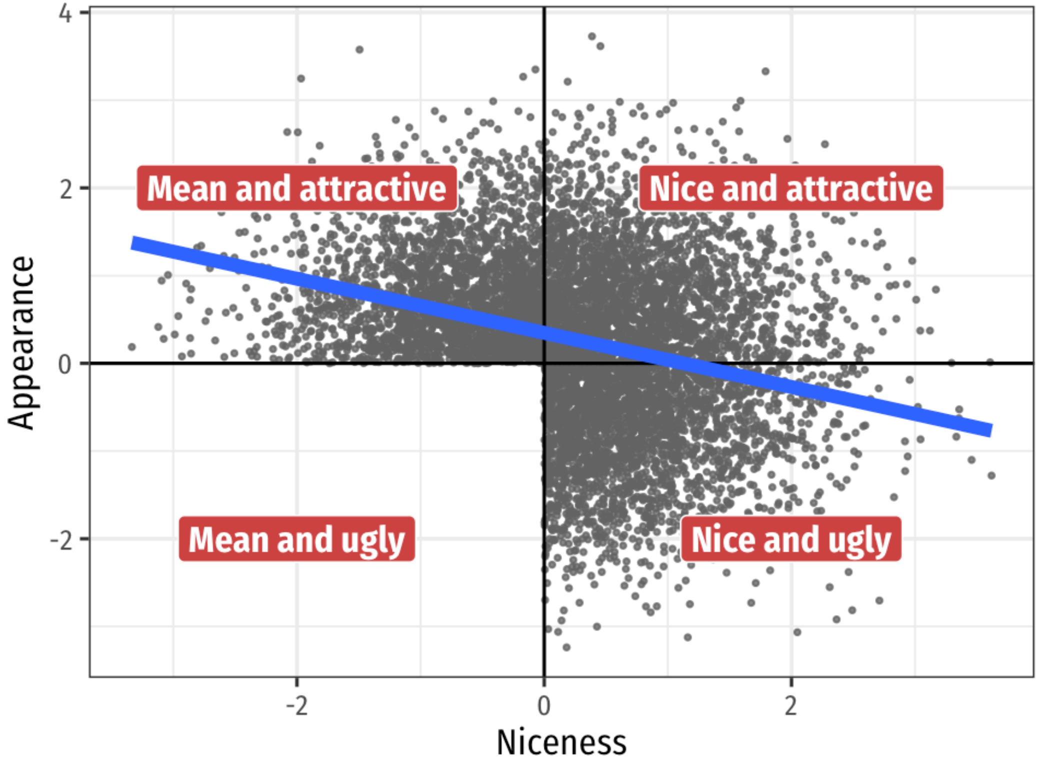
Does niceness improve appearance?



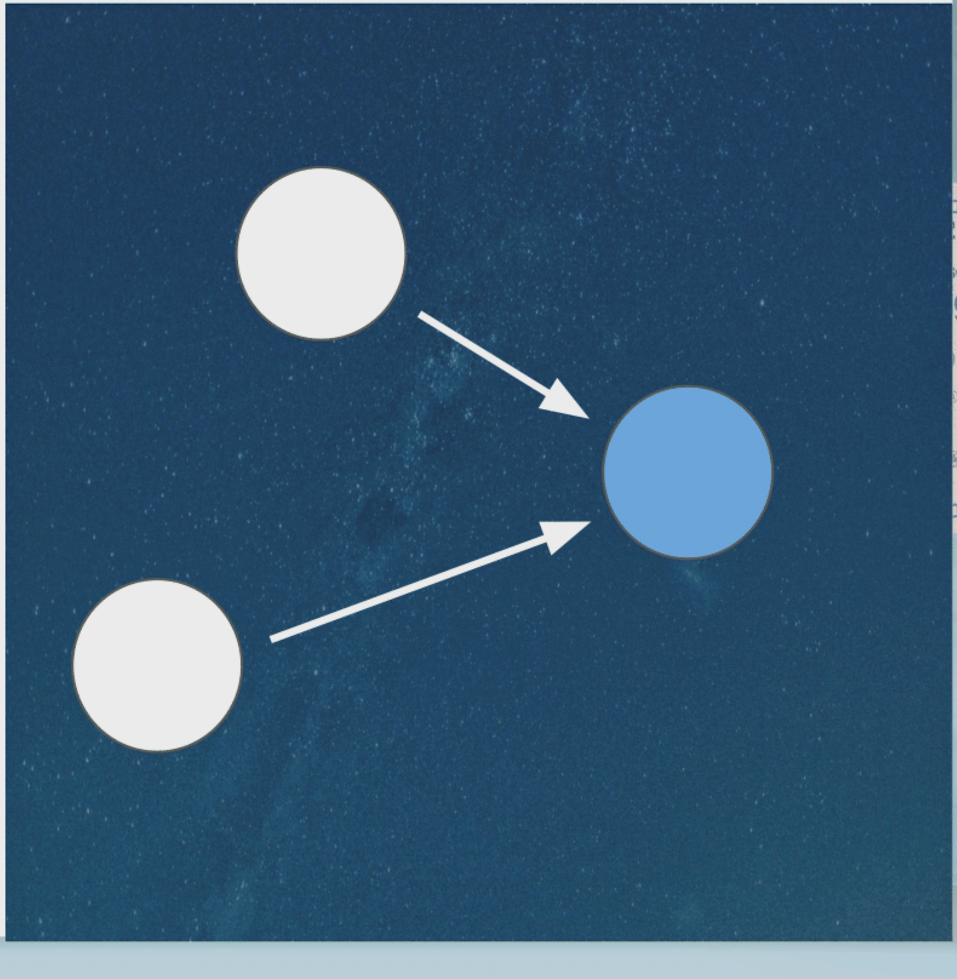


Collider distorts the true effect!



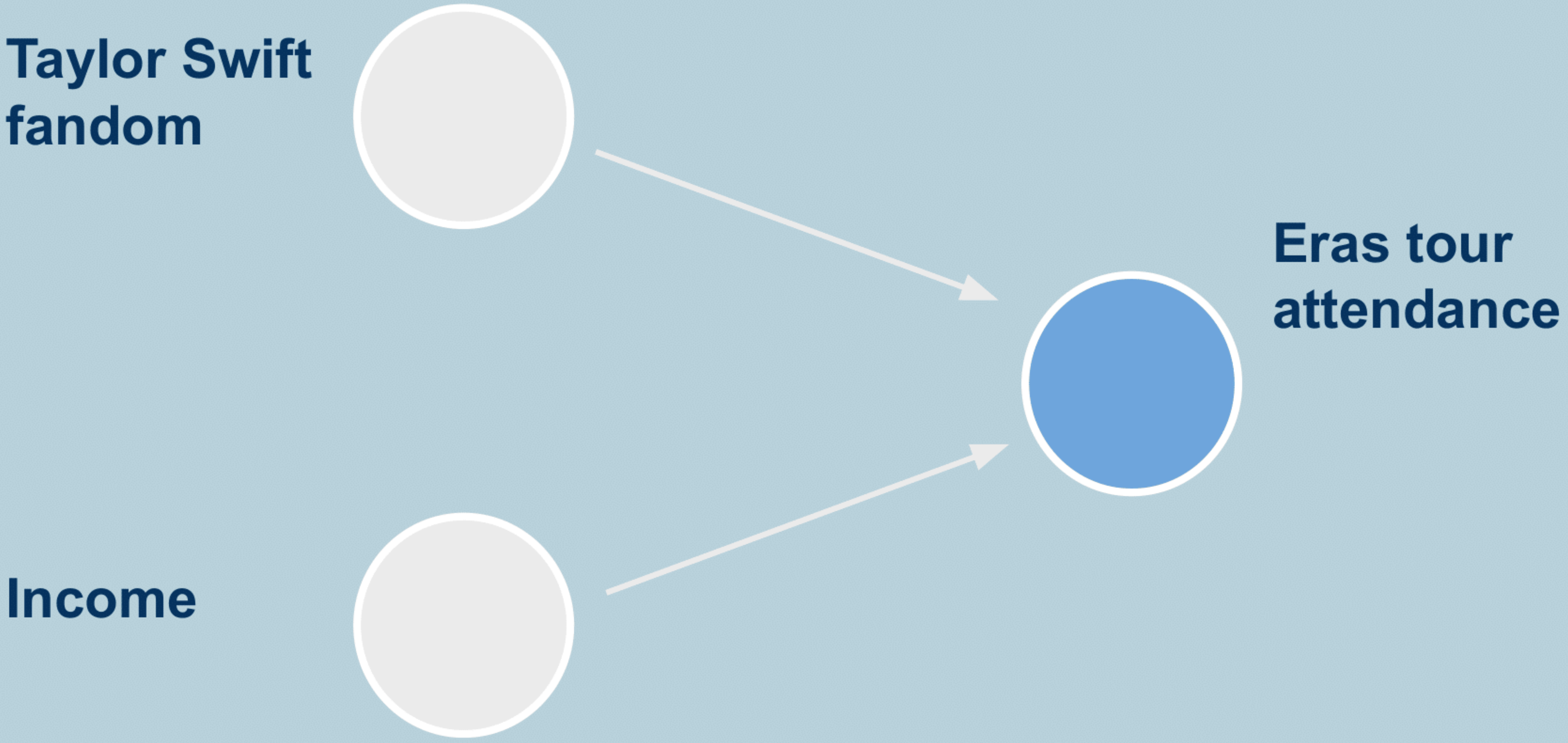


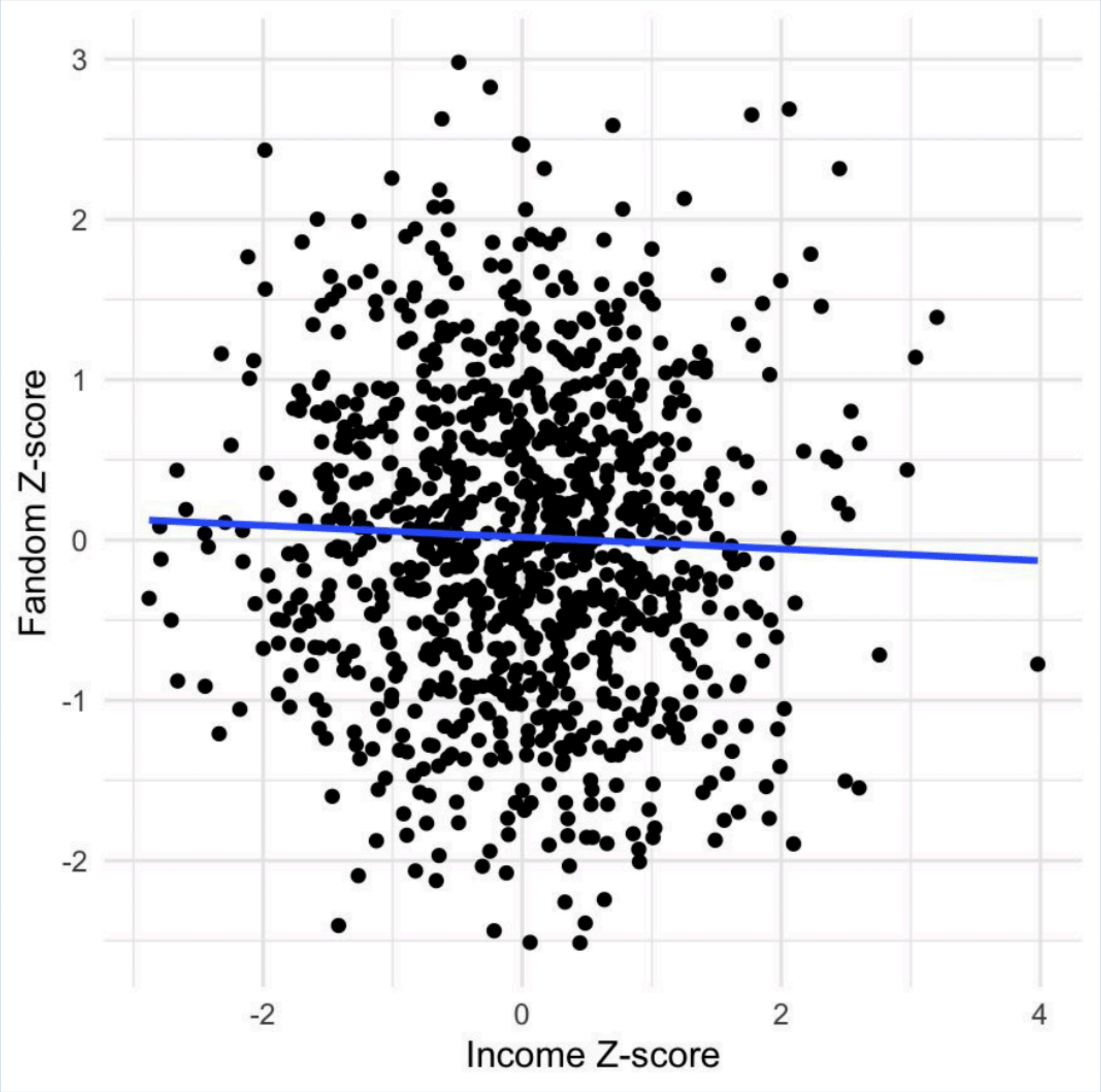
Colliders



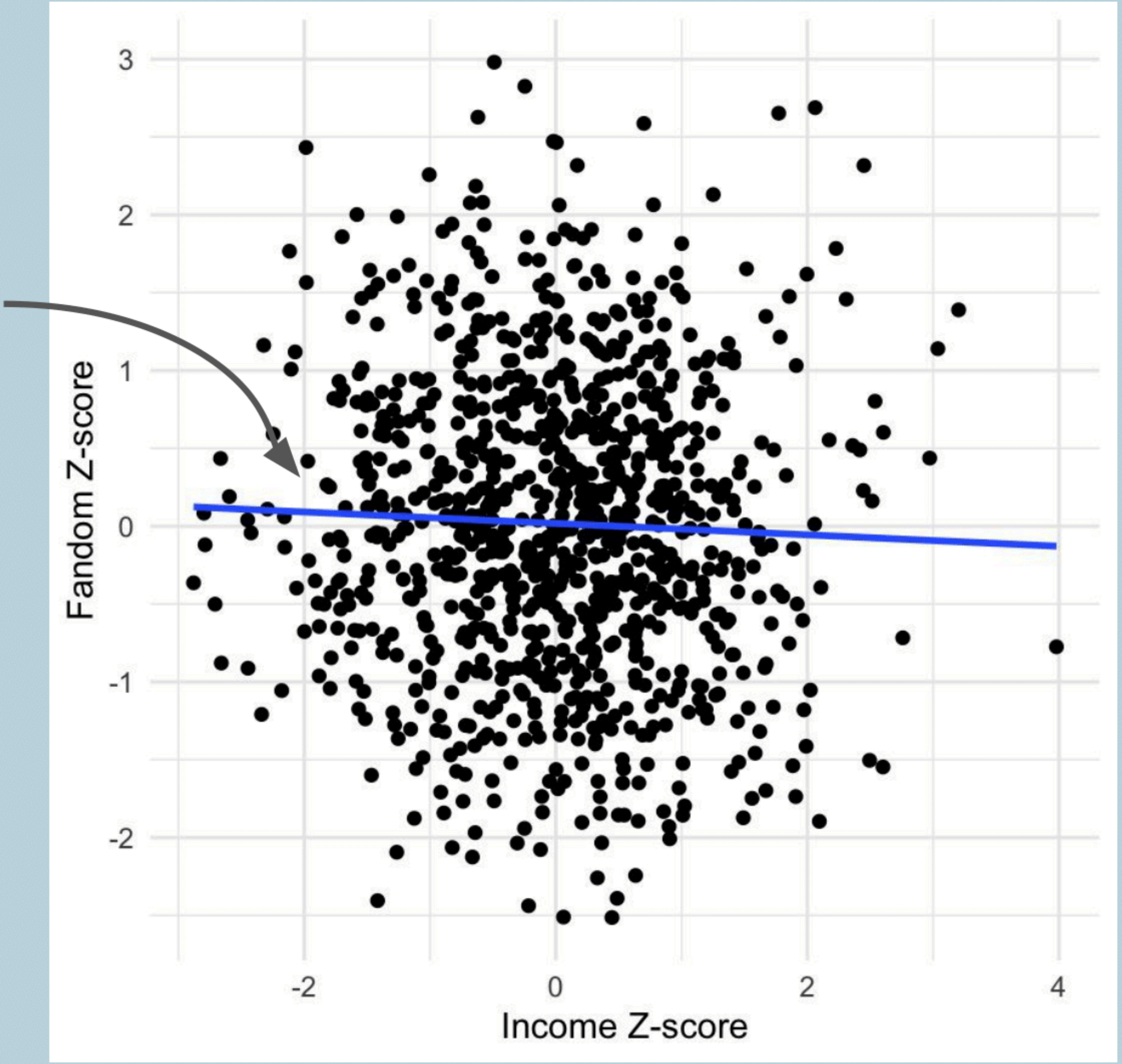
It’s ME hi I’m the collider it’s ME

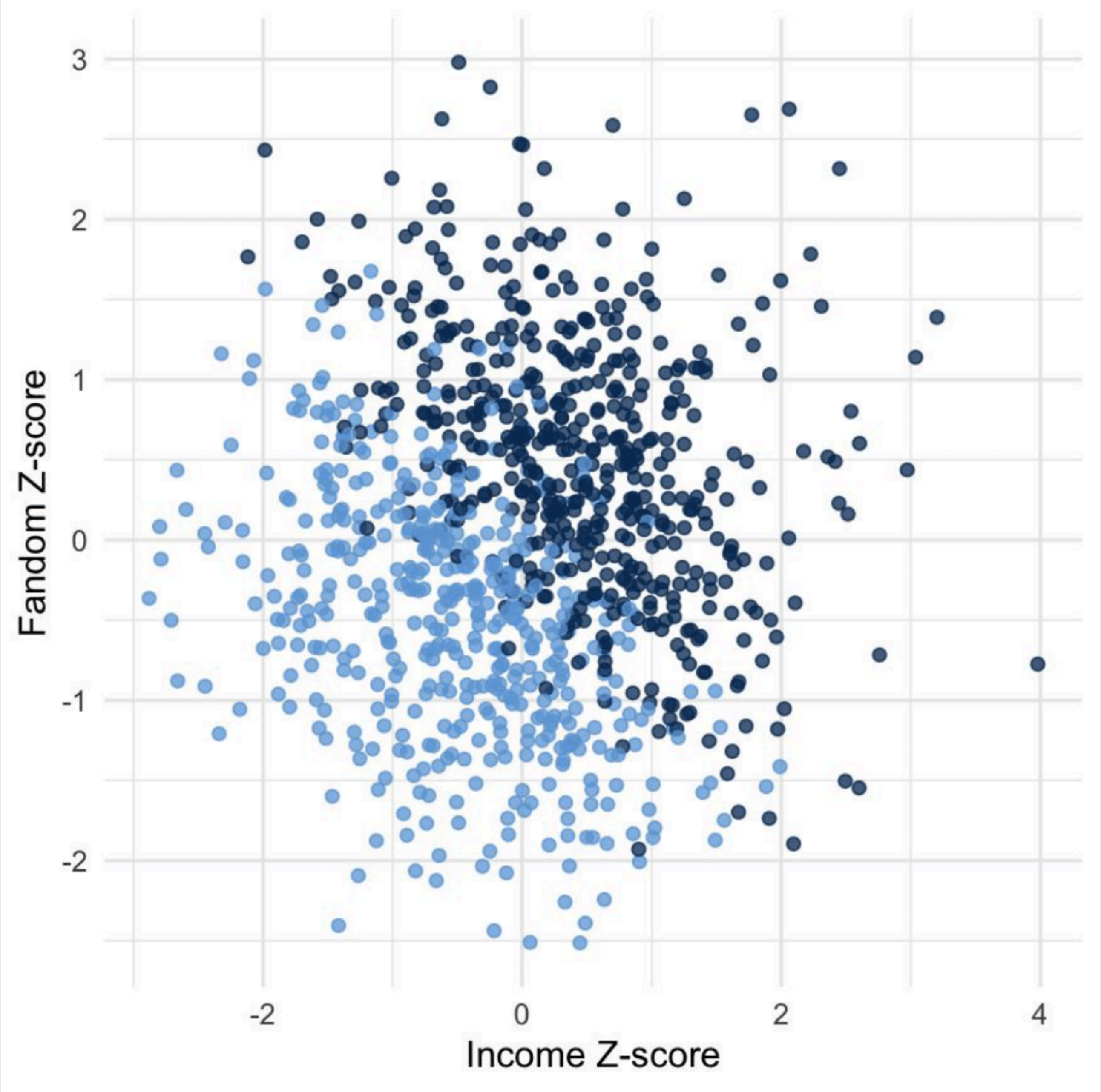
Lucy D’Agostino McGowan
Wake Forest University

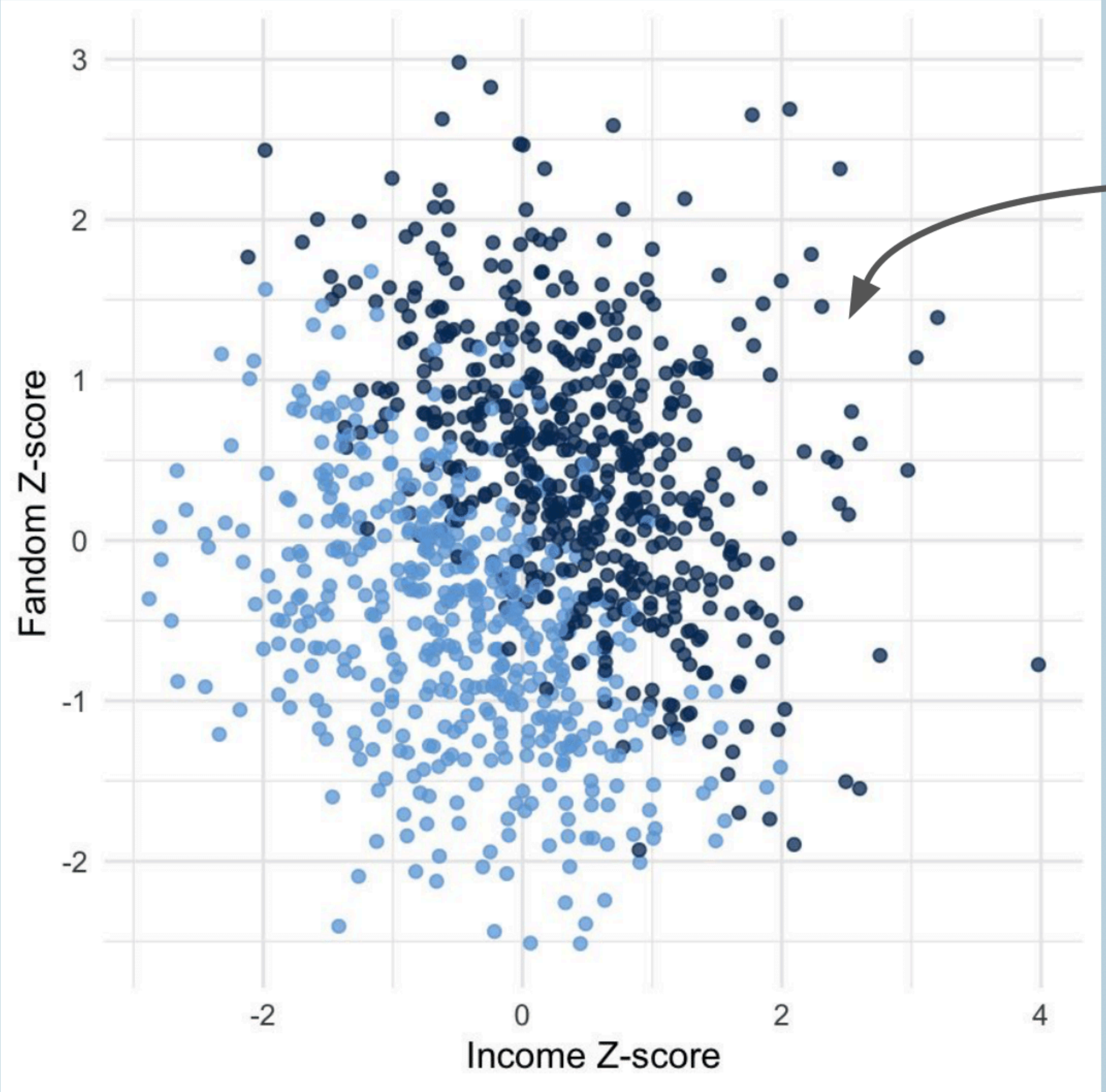

Taylor Swift fandom
Income
Eras tour attendance


3
2
1
0
-1
-2
Fandom Z-score
-2
0
2
4
Income Z-score

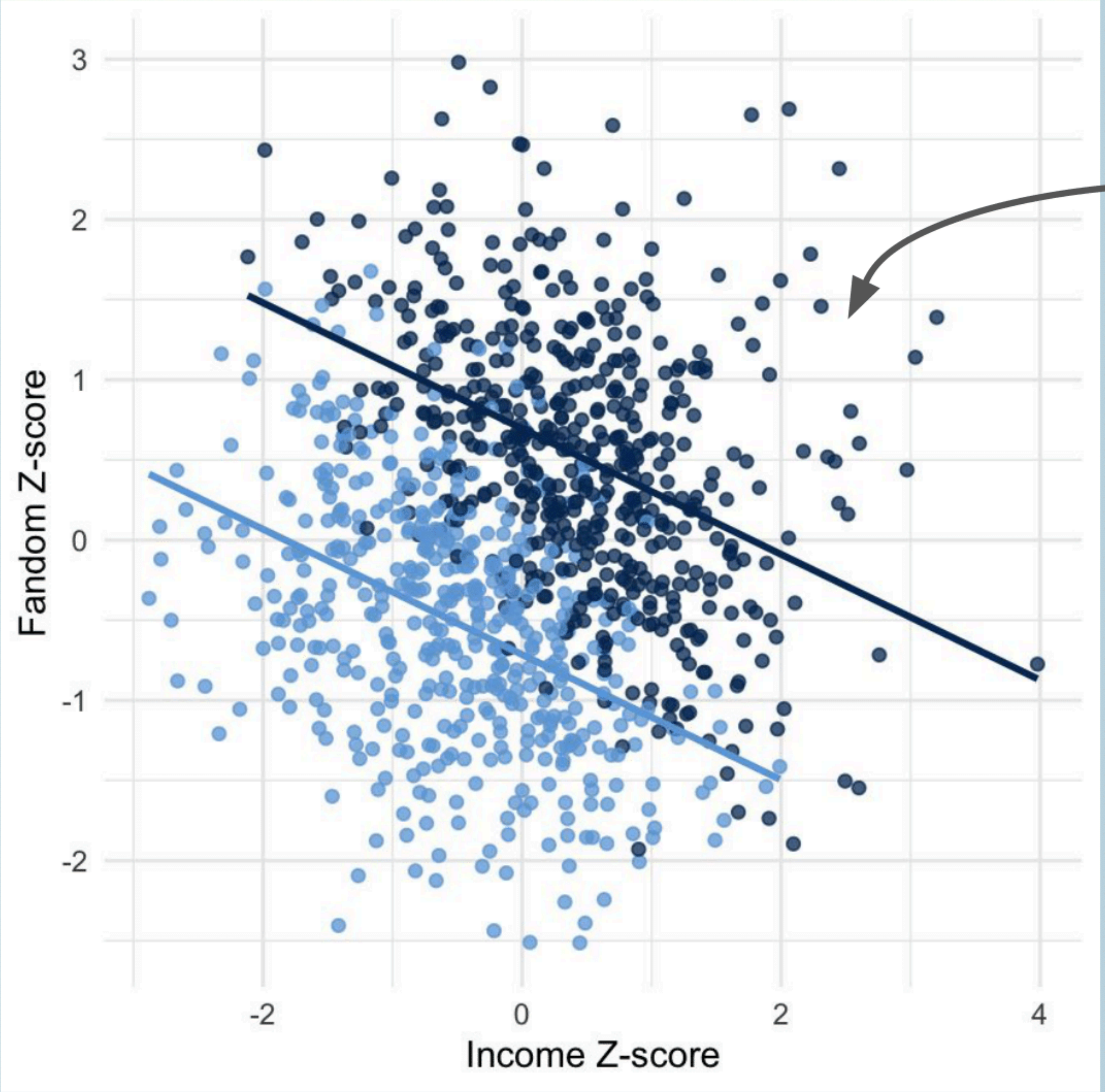
No relationship between income and Taylor Swift fandom



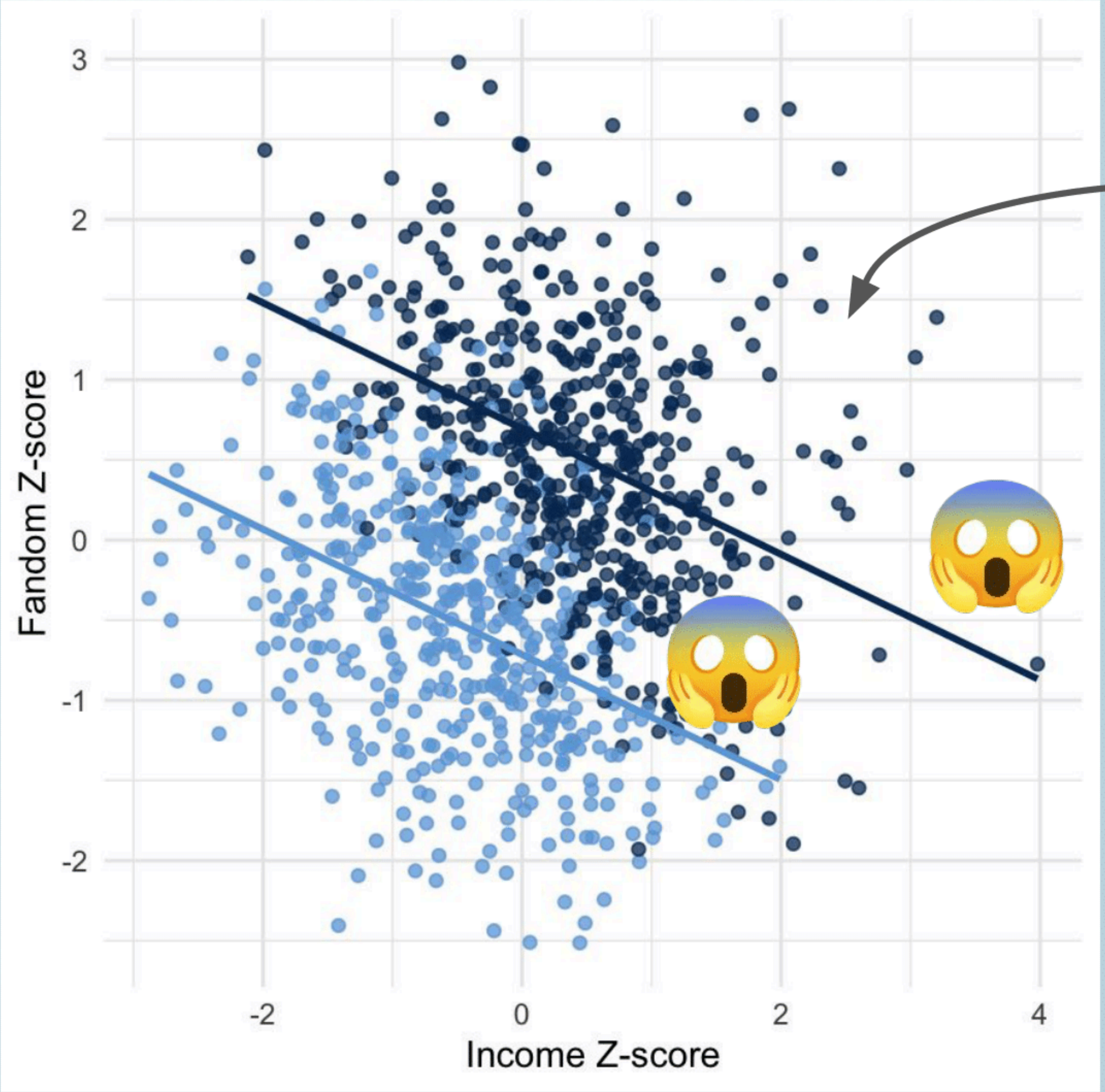

3
2
1
0
-1
-2
Fandom Z-score
-2
0
2
4
Income Z-score



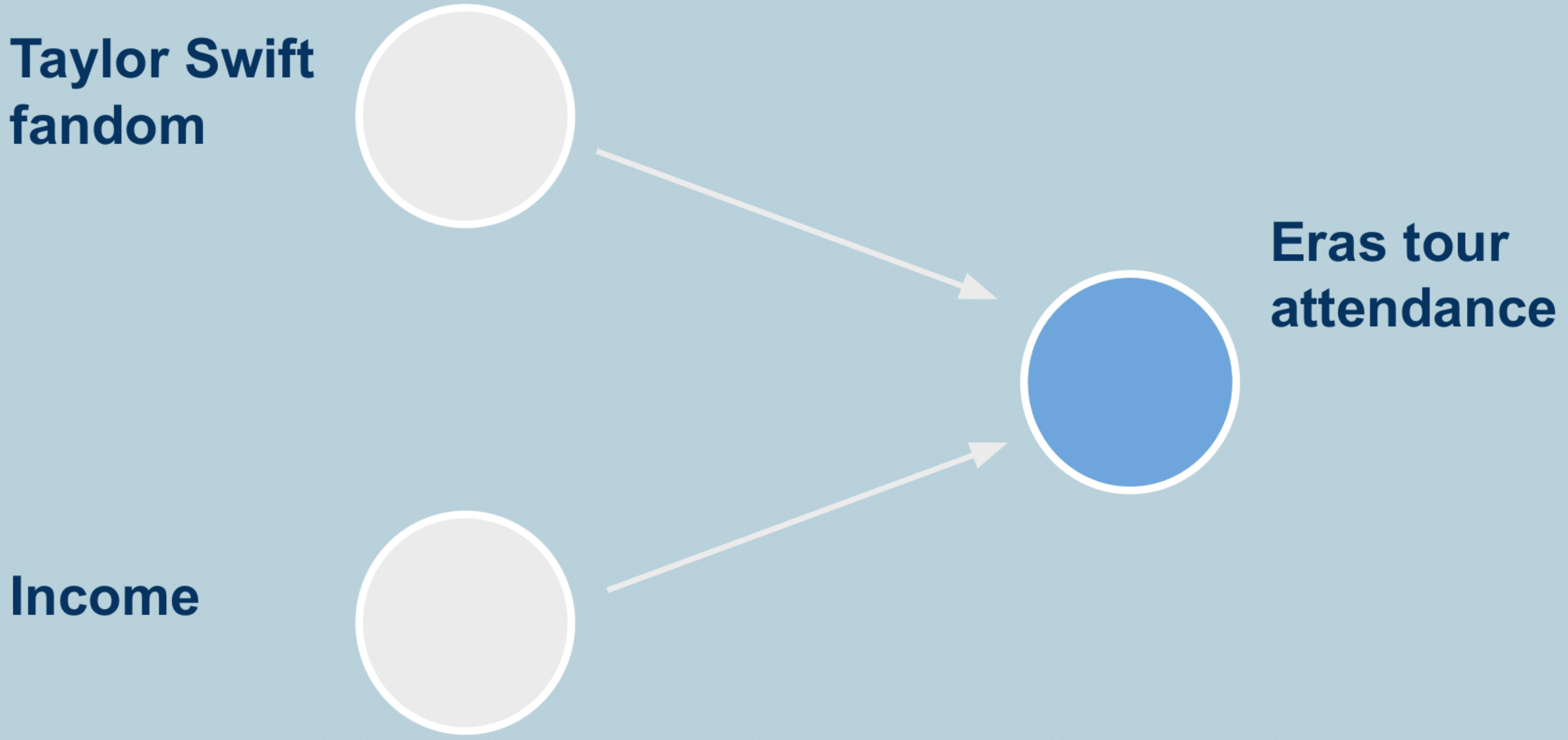
Attended Era's tour



Attended Era’s tour



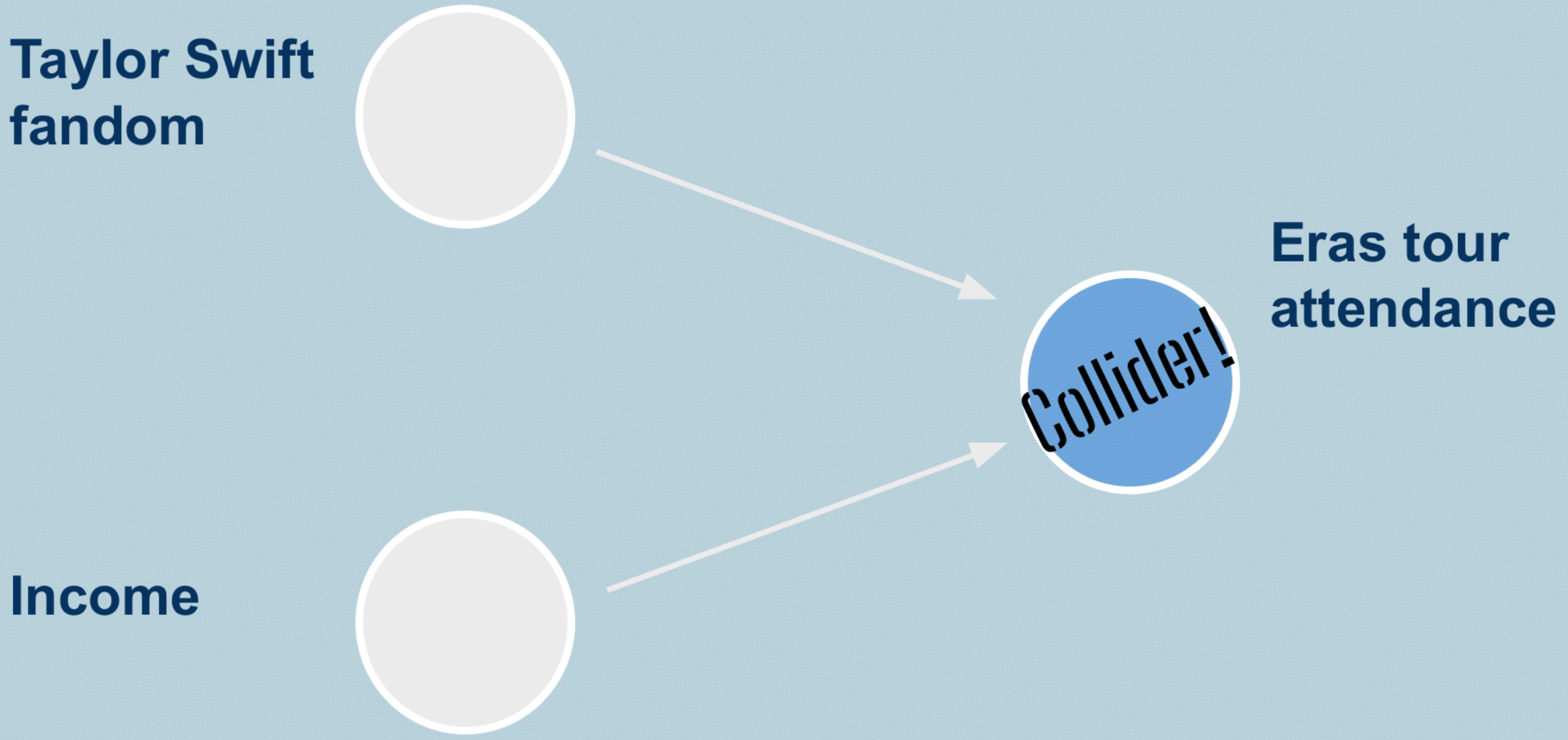
Attended Era's tour



Taylor Swift fandom

Income

Eras tour attendance


Taylor Swift fandom
Income
Eras tour attendance
Collider!

What we want:

$$\widehat{\text{fandom}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{income}$$

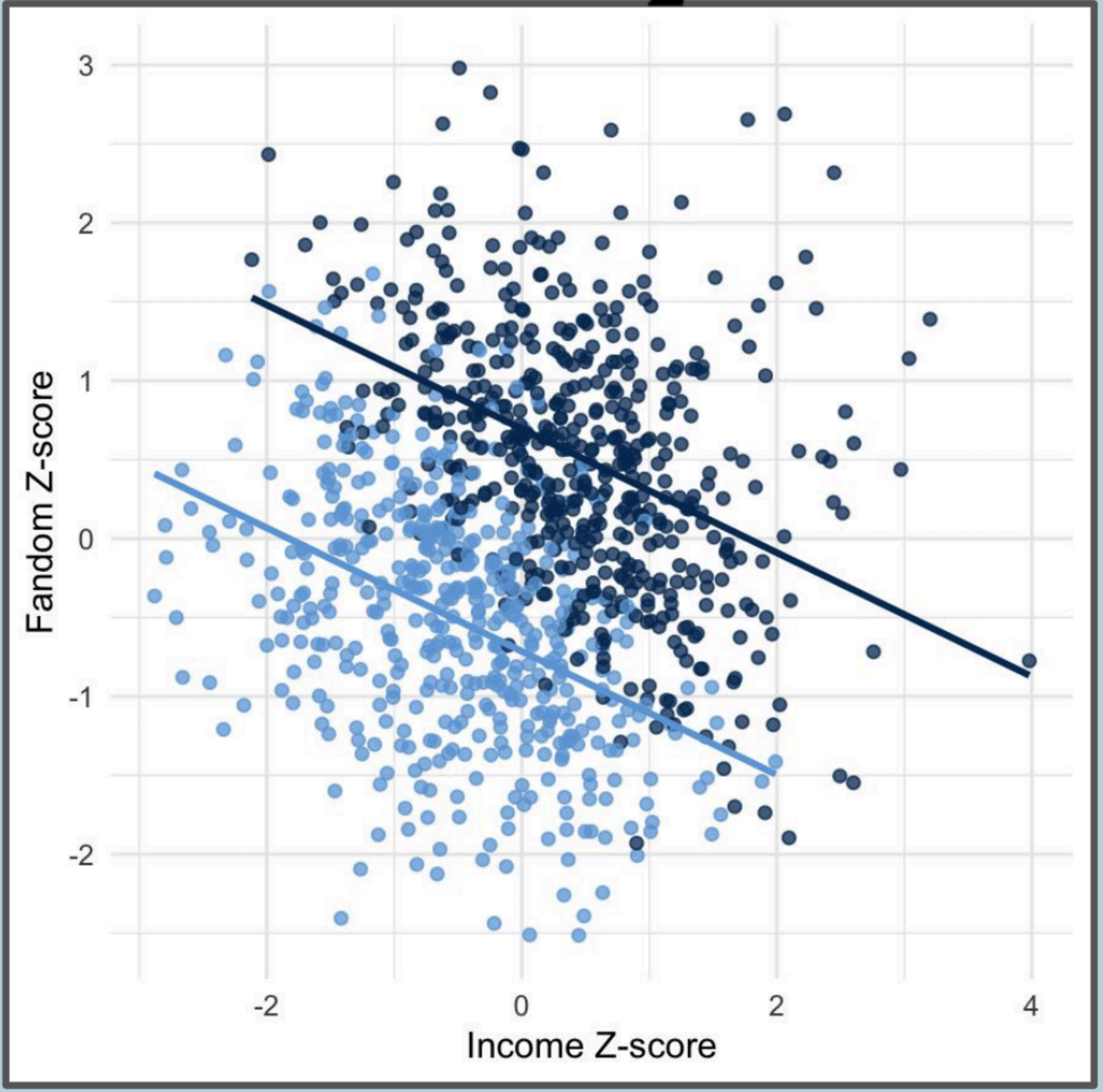
What we want:

$$\widehat{\text{fandom}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{income}$$

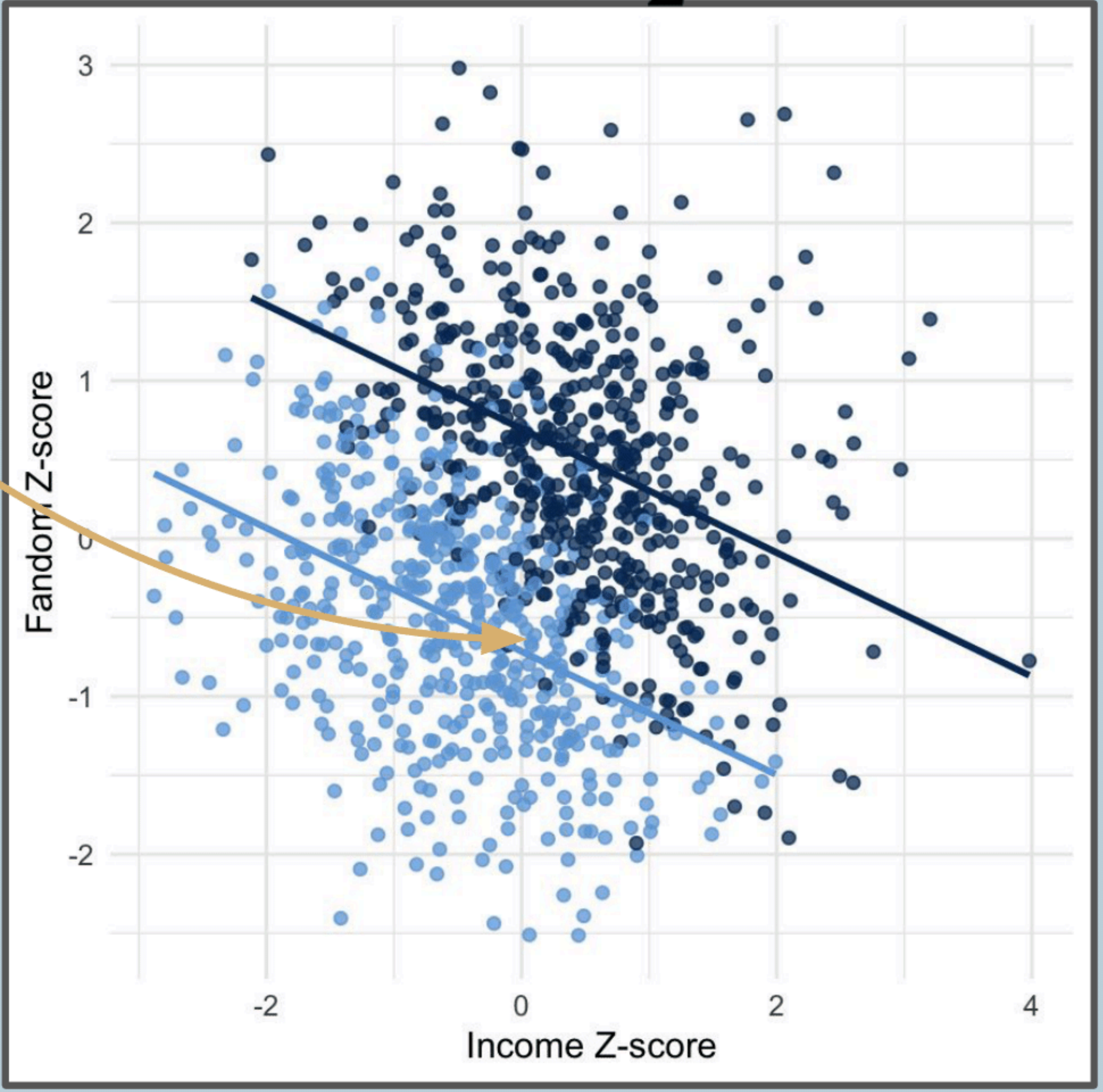
What we have:

$$\widehat{\text{fandom}} = \widehat{\beta}_0^* + \widehat{\beta}_1^* \text{income} + \widehat{\beta}_2^* \text{eras tour}$$

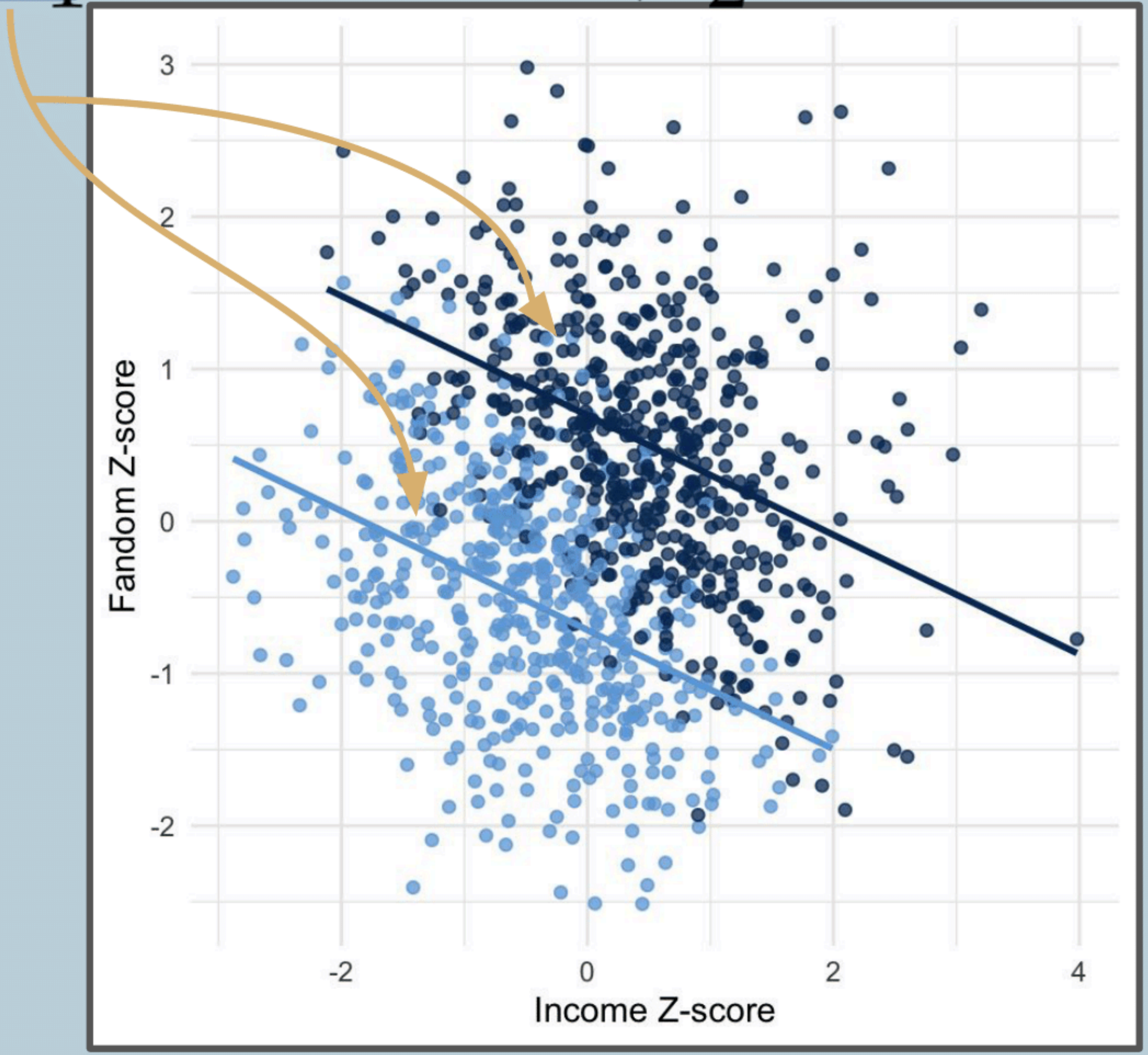
$$\widehat{\text{fandom}} = \widehat{\beta}_0^* + \widehat{\beta}_1^* \text{income} + \widehat{\beta}_2^* \text{eras tour}$$



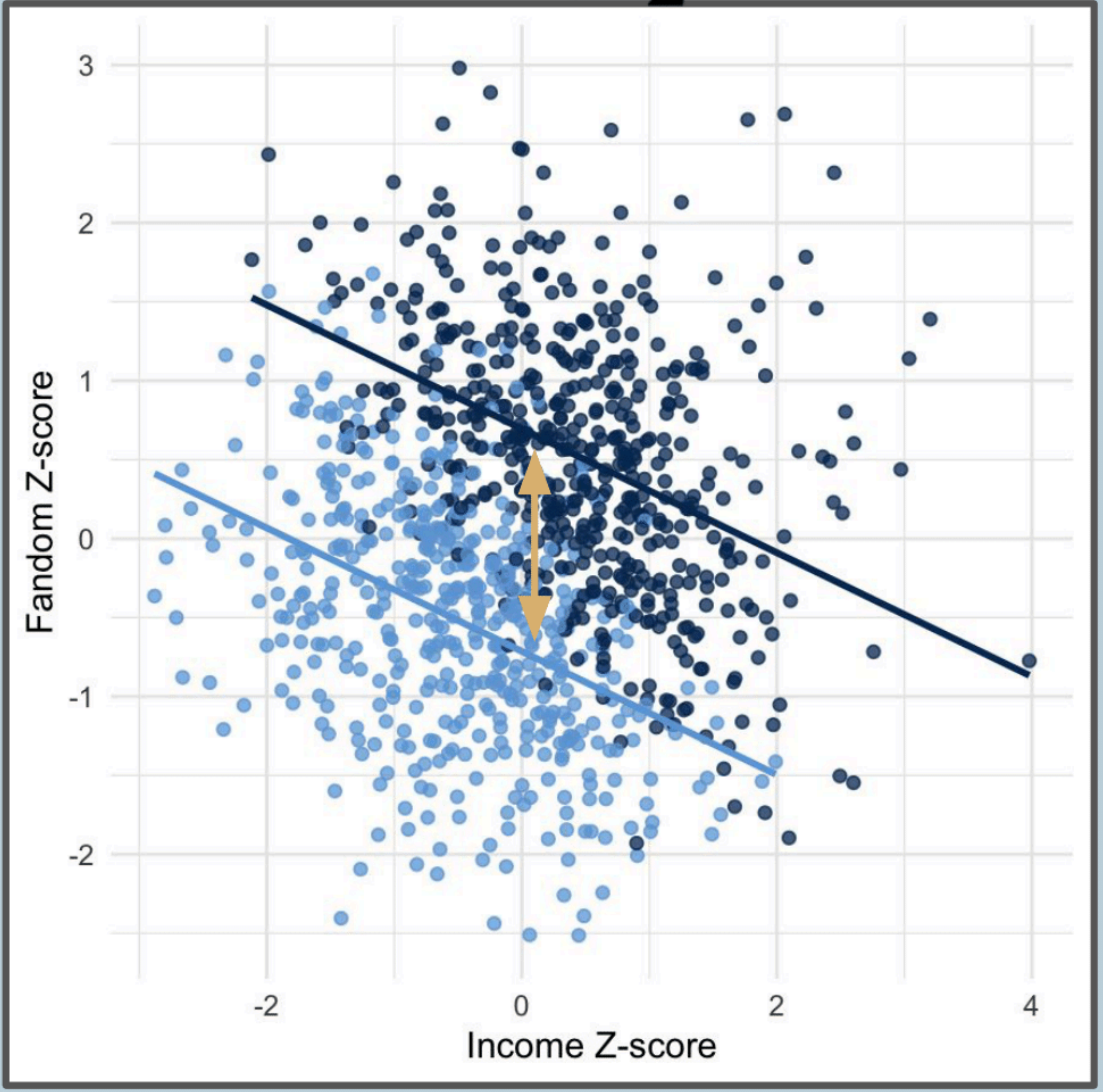
$$\widehat{\text{fandom}} = \boxed{\widehat{\beta}_0^*} + \widehat{\beta}_1^* \text{income} + \widehat{\beta}_2^* \text{eras tour}$$



$$\widehat{\text{fandom}} = \widehat{\beta}_0^* + \boxed{\widehat{\beta}_1^*}\text{income} + \widehat{\beta}_2^*\text{eras tour}$$

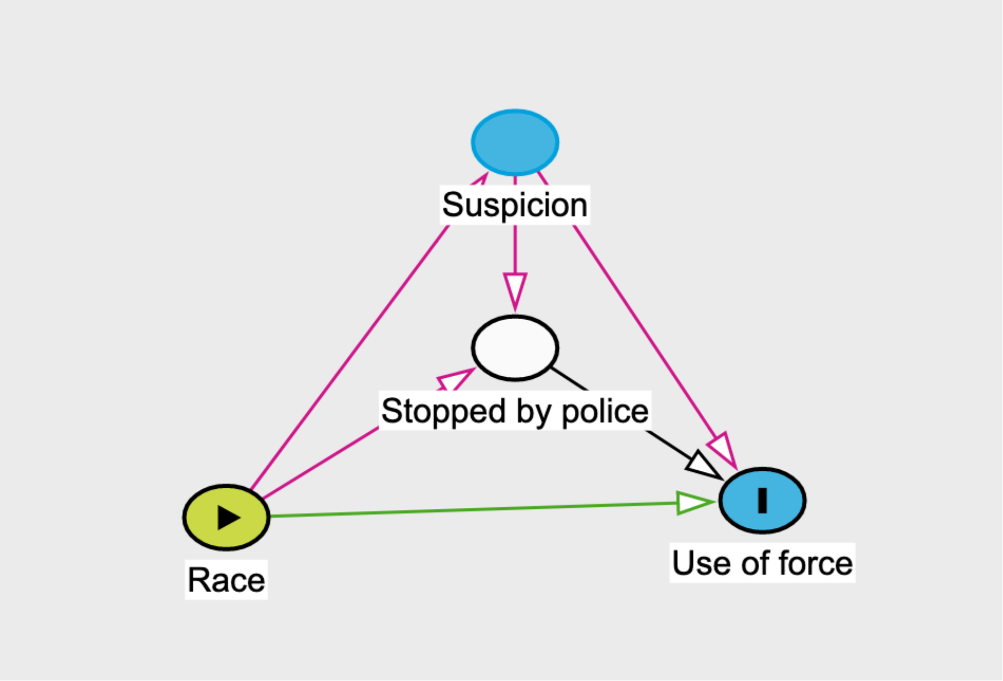


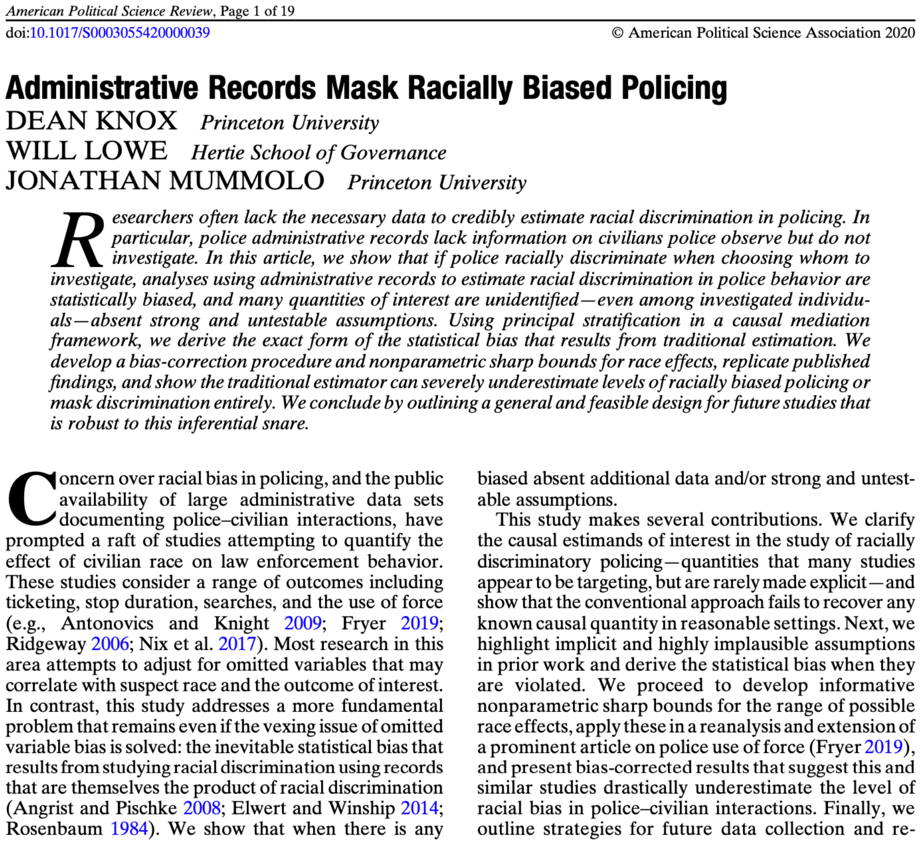
$$\widehat{\text{fandom}} = \widehat{\beta}_0^* + \widehat{\beta}_1^*\text{income} + \widehat{\beta}_2^*\text{eras tour}$$



Effect of race on police use of force using administrative data

Effect of race on police use of force using administrative data





American Political Science Review, Page 1 of 19

doi:10.1017/S0003055420000039 © American Political Science Association 2020

Administrative Records Mask Racially Biased Policing

DEAN KNOX *Princeton University*

WILL LOWE *Hertie School of Governance*

JONATHAN MUMMOLO *Princeton University*

Researchers often lack the necessary data to credibly estimate racial discrimination in policing. In particular, police administrative records lack information on civilians police observe but do not investigate. In this article, we show that if police racially discriminate when choosing whom to investigate, analyses using administrative records to estimate racial discrimination in police behavior are statistically biased, and many quantities of interest are unidentified—even among investigated individuals—absent strong and untestable assumptions. Using principal stratification in a causal mediation framework, we derive the exact form of the statistical bias that results from traditional estimation. We develop a bias-correction procedure and nonparametric sharp bounds for race effects, replicate published findings, and show the traditional estimator can severely underestimate levels of racially biased policing or mask discrimination entirely. We conclude by outlining a general and feasible design for future studies that is robust to this inferential snare.

Concern over racial bias in policing, and the public availability of large administrative data sets documenting police–civilian interactions, have prompted a raft of studies attempting to quantify the effect of civilian race on law enforcement behavior. These studies consider a range of outcomes including ticketing, stop duration, searches, and the use of force (e.g., Antonovics and Knight 2009; Fryer 2019; Ridgeway 2006; Nix et al. 2017). Most research in this area attempts to adjust for omitted variables that may correlate with suspect race and the outcome of interest. In contrast, this study addresses a more fundamental problem that remains even if the vexing issue of omitted variable bias is solved: the inevitable statistical bias that results from studying racial discrimination using records that are themselves the product of racial discrimination (Angrist and Pischke 2008; Elwert and Winship 2014; Rosenbaum 1984). We show that when there is any biased absent additional data and/or strong and untestable assumptions.

This study makes several contributions. We clarify the causal estimands of interest in the study of racially discriminatory policing—quantities that many studies appear to be targeting, but are rarely made explicit—and show that the conventional approach fails to recover any known causal quantity in reasonable settings. Next, we highlight implicit and highly implausible assumptions in prior work and derive the statistical bias when they are violated. We proceed to develop informative nonparametric sharp bounds for the range of possible race effects, apply these in a reanalysis and extension of a prominent article on police use of force (Fryer 2019), and present bias-corrected results that suggest this and similar studies drastically underestimate the level of racial bias in police–civilian interactions. Finally, we outline strategies for future data collection and re-

Can we make another DAG together?

Potential outcomes vs. do() notation

Expectations

$$\mathrm{E}(\cdot), \mathbf{E}(\cdot), \mathbb{E}(\cdot) \quad \text{vs.} \quad \mathrm{P}(\cdot)$$

Basically a fancy way of saying "average"

andhs.co/potential-outcomes

Causal effects with potential outcomes

Potential outcomes notation:

$$\delta \;= \tfrac{1}{n} \textstyle\sum_{i=1}^{n} Y_i(1) - Y_i(0)$$

or alternatively with $\mathbf{E}$

$$\delta \;= \mathbf{E}[Y_i(1) - Y_i(0)]$$

Causal effects with do()

Pearl notation:

$$\delta \; = \mathbf{E}[Y_i \mid \text{do}(X = 1) - Y_i \mid \text{do}(X = 0)]$$

or more simply

$$\delta \; = \mathbf{E}[Y_i \mid \text{do}(X)]$$

$$\mathbf{E}[Y_i \mid \mathrm{do}(X)]$$

$$=$$

$$\mathbf{E}[\boldsymbol{Y}_i(1) - \boldsymbol{Y}_i(0)]$$

We can't see this

$$\mathbf{E}[Y_i \mid \text{do}(X)] \quad \text{or} \quad \mathbf{E}[Y_i(1) - Y_i(0)]$$

So we find the average causal effect (ACE)

$$\hat{\delta} = \mathbf{E}[Y_i \mid X = 1] - \mathbf{E}[Y_i \mid X = 0]$$

$$\underbrace{\overbrace{\mathbf{E}(y \mid \mathrm{do}(x))}^{\text{The average population-level change in } y \text{ when } \textit{directly intervening} \text{ (or doing) } x}}_{\text{Causation}} \neq \underbrace{\overbrace{\mathbf{E}(y \mid x)}^{\text{The average population-level change in } y \text{ when accounting for } \textit{observed } x}}_{\text{Correlation}}$$

p-values and confidence intervals

In the absence of *p*-values, I'm confused about how we report… significance?

Imbens and *p*-values

Nobody really cares about *p*-values

Decision makers want to know
a number or a range of numbers—
some sort of effect and uncertainty

Nobody cares how likely a number would be
in an imaginary null world!

Imbens's solution

Report point estimates and some sort of range

> "It would be preferable if reporting standards emphasized confidence intervals or standard errors, and, even better, Bayesian posterior intervals."

Point estimate

The single number you calculate (mean, coefficient, etc.)

Uncertainty

A range of possible values

Population parameter

Truth = Greek letter

An single unknown number that is true for the entire population

Proportion of left-handed students at GSU

Median rent of apartments in Atlanta

Proportion of red M&Ms produced in a factory

Treatment effect of your program

Samples and estimates

We take a sample and make a guess

This single value is a *point estimate*

(This is the Greek letter with a hat)

Variability

**You have an estimate,
but how different might that
estimate be if you take another sample?**

Nets and confidence intervals

How confident are we that the sample picked up the population parameter?

Confidence interval is a net

We can be X% confident that our net is picking up that population parameter

If we took 100 samples, at least 95 of them would have the true population parameter in their 95% confidence intervals

> A city manager wants to know the true average property value of single-owner homes in her city. She takes a random sample of 200 houses and builds a 95% confidence interval. The interval is ($180,000, $300,000).

We're 95% confident that the interval ($180,000, $300,000) captured the true mean value

WARNING

It is way too tempting to say "We're 95% sure that the population parameter is X"

People do this all the time! People with PhDs!

YOU will do this too

Nets

**If you took lots of samples,
95% of their confidence intervals
would have the single true value in them**

Frequentism

This kind of statistics is called "frequentism"

The population parameter θ is fixed and singular while the data can vary

$$P(\text{Data} \mid \theta)$$

You can do an experiment over and over again; take more and more samples and polls

Frequentist confidence intervals

"We are 95% confident that this net captures the true population parameter"

~~"There's a 95% chance that the true value falls in this range"~~

Bayesian statistics

Rev. Thomas Bayes

$$P(\theta \mid \text{Data})$$

$$P(\text{H} \mid \text{E}) = \frac{P(\text{H}) \times P(\text{E} \mid \text{H})}{P(\text{E})}$$

$$P(\mathrm{H} \mid \mathrm{E}) = \frac{P(\mathrm{H}) \times P(\mathrm{E} \mid \mathrm{H})}{P(\mathrm{E})}$$

$$P(\text{Hypothesis} \mid \text{Evidence}) = \frac{P(\text{Hypothesis}) \times P(\text{Evidence} \mid \text{Hypothesis})}{P(\text{Evidence})}$$

But the math is too hard!

So we simulate!

(Monte Carlo Markov Chains, or MCMC)

Bayesianism and parameters

**In the world of frequentism,
there's a fixed population parameter
and the data can hypothetically vary**

$$P(\text{Data} \mid \theta)$$

**In the world of Bayesianism,
the data is fixed (you collected it just once!)
and the population parameter can vary**

$$P(\theta \mid \text{Data})$$

Bayesian credible intervals

(AKA posterior intervals)

"Given the data, there is a 95% probability that the true population parameter falls in the credible interval"

Intervals

Frequentism

There's a 95% probability that the range contains the true value

Probability of the range

Few people naturally think like this

Bayesianism

There's a 95% probability that the true value falls in this range

Probability of the actual value

People *do* naturally think like this!

Thinking Bayesianly

We all think Bayesianly,
even if you've never heard of Bayesian stats

Every time you look at a confidence interval, you inherently think that the parameter is around that value, but that's wrong!

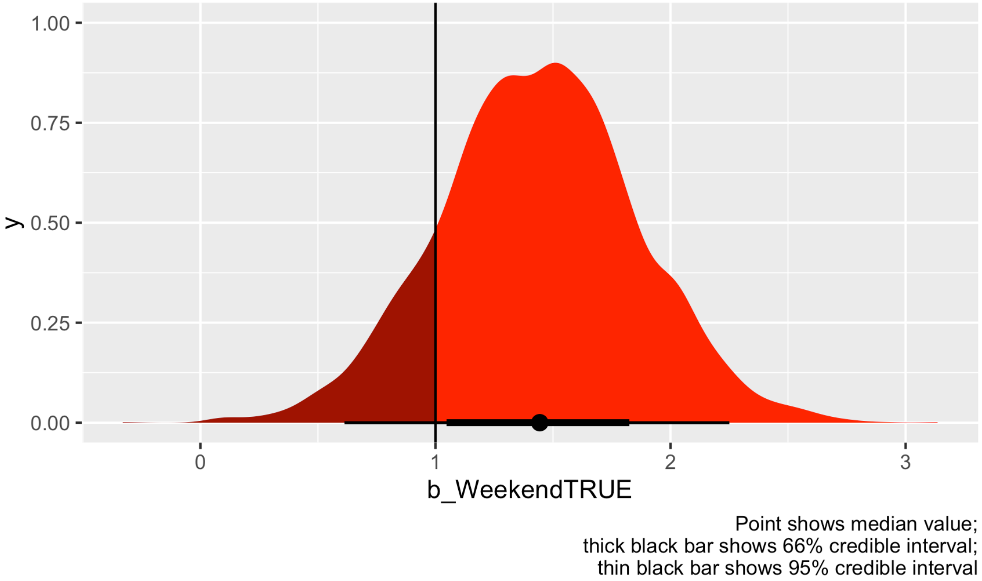
BUT Imbens cites research that that's actually generally okay

Often credible intervals are super similar to confidence intervals

Bayesian inference

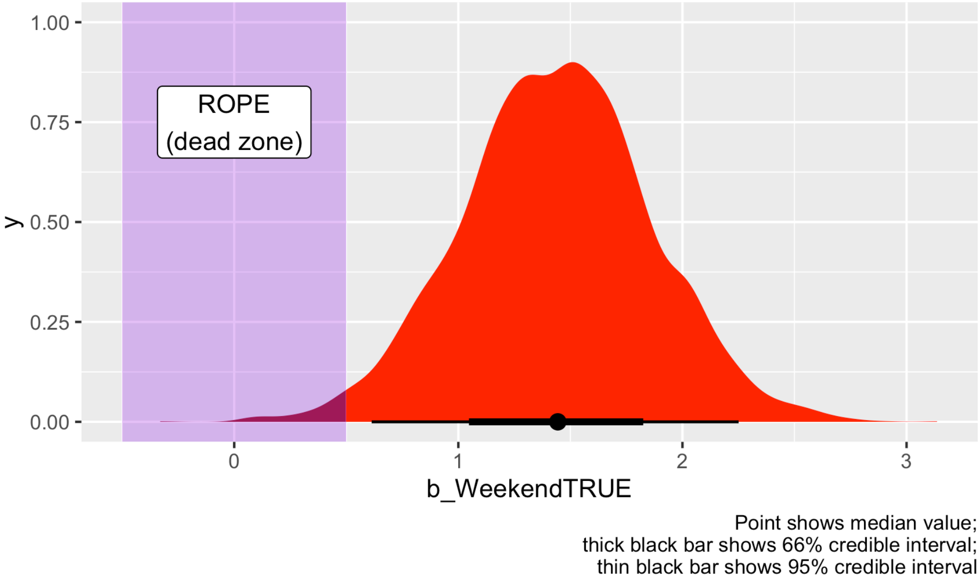
Inference without *p*-values!

Probability of direction



Point shows median value;
thick black bar shows 66% credible interval;
thin black bar shows 95% credible interval

Region of practical equivalence (ROPE)



Point shows median value;
thick black bar shows 66% credible interval;
thin black bar shows 95% credible interval