Inference & Causality

Week 5 Session 10 • November 18, 2025

Lecturer: Narges Chinichian

IU University of Applied Sciences, Berlin

Unit 5

Fallacies in Causal Reasoning

Special Cases: Birth weight paradox, M-bias, Berkson's paradox, Simpson's Paradox

Recap from Last session

1. Mediator Fallacy
2. Collider Bias

The Mediator Fallacy

The mediator fallacy occurs when we condition on a mediator instead of holding it constant using an intervention.

A mediator sits in the middle of a causal chain:

X → M → Y

If we condition on M in data (e.g., only look at cases with M = some value), we inadvertently block the real causal path from X to Y.

Furthermore, conditioning on M can also open spurious paths through other variables, creating misleading or fake associations that do not reflect the true causal structure.

Key Message 1

Conditioning on a mediator is not the same as **holding it constant through an intervention**.

Key Message 2

This type of conditioning creates bias, leading to what is known as the mediator fallacy.

Controlled Direct Effect (CDE)

The correct way to "hold M constant" is to intervene on it.

Defining CDE

Instead of conditioning, we ask: "What happens to Y if we force M to a fixed value (like M = m) and change X?" This is called the Controlled Direct Effect (CDE).

The Intervention

We literally fix M using a hypothetical intervention (do(M = m)) and compare outcomes for different values of X.

Isolating the Path

This isolates the direct path from X → Y while the mediator is frozen at a chosen value.

Why it Matters

This is what people think they are doing when they condition on M — but conditioning is not intervention.

Natural Direct Effect (NDE)

What if fixing the mediator to an artificial value makes no sense? (example: forcing all students to apply to the Physics department)

Natural Value

The Natural Direct Effect fixes the mediator at the value it would have taken naturally, under a baseline version of X.

Intervene on X

Then, we intervene on X (e.g., pretend they had a different gender) while keeping the mediator's natural choice.

NDE Definition

NDE = direct effect of X on Y when the mediator stays at its natural value.

Mediator's Choice

We let people choose the mediator value they naturally would (e.g., the department they would apply to).

Compare Outcomes

We compare the outcomes under these conditions.

Real-world Use

Used in cases like the Berkeley admissions paradox, where forcing the mediator to an artificial value would be unrealistic.

Mediation fallacy vocab

Mediator Fallacy	Incorrectly tries to isolate the direct effect	**Conditions** on M (statistical filtering)	*Never — this creates bias*	"Conditioning on a mediator instead of holding it constant" (mediator fallacy)
Controlled Direct Effect (CDE)	Measures the direct effect of X on Y when we *force* M to a specific value	**Intervene:** do(M = m)	When fixing the mediator to a chosen value makes sense	Defined explicitly as an intervention on both X and M (CDE)
Natural Direct Effect (NDE)	Measures the direct effect of X on Y while keeping M at the value it would naturally take	**Freeze M at its natural value** under baseline X	When forcing M is unnatural or unrealistic (e.g., Berkeley departments)	NDE introduced with counterfactuals using students choosing their natural department

Concept	Meaning	Formula
Mediator Fallacy	Mistakenly conditioning on the mediator	-
CDE	Direct effect when mediator is *forced* to m	$P(Y \mid do(X = x), do(M = m)) - P(Y \mid do(X = x'), do(M = m))$
NDE	Direct effect when mediator is held at natural baseline value	$P(Y_{M=M_{x'}} \vert do(X = x)) - P(Y_{M=M_{x'}} \vert do(X = x'))$

Overview of Today

1. Special Cases of Collider Bias: Birth weight paradox, M-bias and Berkson's paradox
2. Simpson's paradox (confounder case)
3. Summary and Quiz

Special Cases of Collider Bias

While we've explored the general concept of collider bias, certain manifestations of this phenomenon have been historically important and developed their own specific names due to their distinct contexts and implications.

Birth-weight paradox

M-Bias

Berkson's Paradox

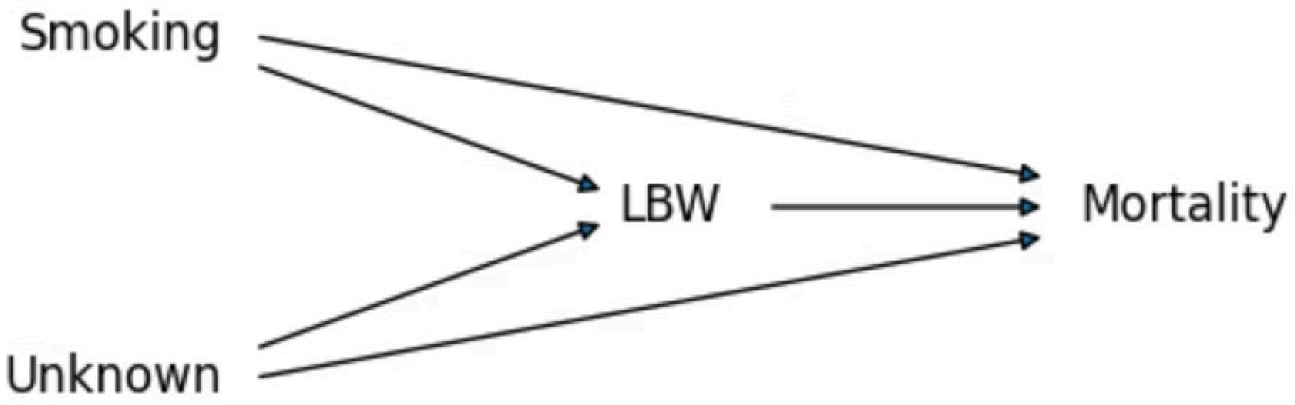
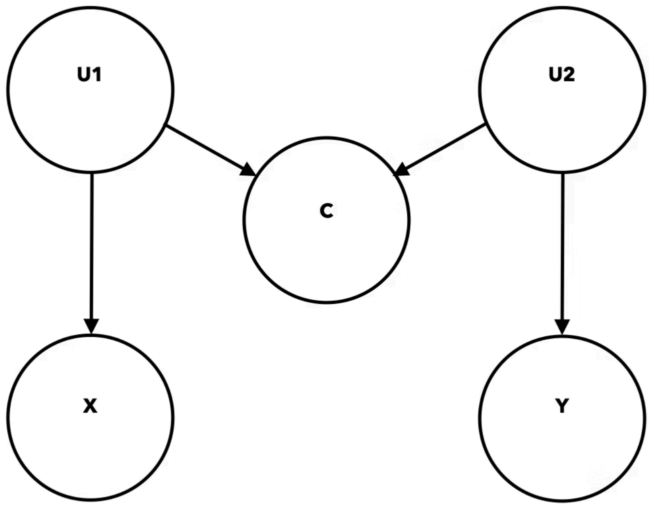


image credit: Allen Downey: https://allendowney.github.io/ProbablyOverthinkingIt/birthweight.html

Birth-weight Paradox

A counterintuitive finding where low birth weight seems to reduce infant mortality risk within specific subgroups (such as babies of smoking mothers), despite low birth weight typically being associated with increased mortality in the general population.

Key insight: This paradox occurs because birth weight acts as a collider; both maternal factors (like smoking) and genetic factors influence birth weight, and conditioning on birth weight creates spurious associations.



M-Bias

M-bias is a form of **collider bias**, not confounding.

It arises when:

- Two unmeasured variables (U1 and U2) influence X and Y **separately,** and both also influence a common collider (C).
- The path between X and Y is **naturally blocked** at the collider C.
- **Conditioning on C** (or anything influenced by C) *opens* this blocked path and creates a **spurious association** between X and Y.
- This produces bias even though C is **not** on the causal path from X to Y.

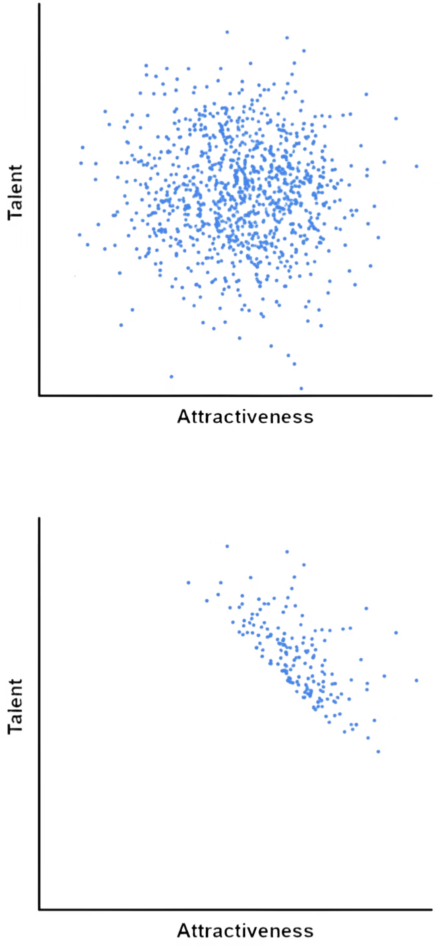
Key insight: Adjusting for variables that seem relevant can sometimes introduce bias rather than reduce it, particularly when dealing with colliders and unmeasured common causes.

Berkson's Paradox

Berkson's paradox is a **special case** that shows up in real-world selection processes.

Definition:

Apparent **negative correlation** between two positive traits, caused by conditioning on a selection variable.



"Collider Bias Diagram" by Belbury, CC BY-SA 4.0, via Wikimedia Commons. Source: https://commons.wikimedia.org/wiki/File:Collider_bias.png

Berkson's Example (Original)

A hospital study:

Disease A increases hospitalization

Disease B increases hospitalization

Inside hospital patients: → A and B appear negatively correlated

But in the population, they are independent.

The ugly five-star restaurant

There are two highly rated restaurants side by side; you can't try the food, but you see the restaurants' atmosphere from outside. Which of the restaurants should you choose for the best food experience?

Source: modified adoption of http://brilliant.org/wiki/berksons-paradox/ example.

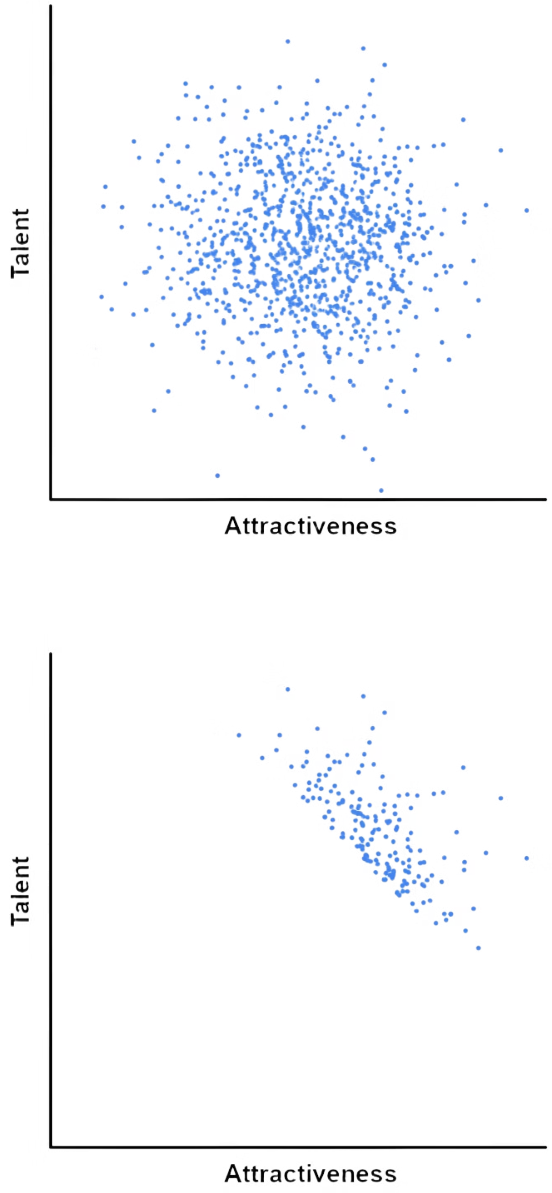
Why Are Handsome Men Such Jerks?

(Article by Jordan Ellenberg on Slate)

Visualizing Berkson's Paradox

Trait1 → Selection ← Trait2

Conditioning on selection creates negative association.



[1] "Collider Bias Diagram" by Belbury, CC BY-SA 4.0, via Wikimedia Commons. Source: https://commons.wikimedia.org/wiki/File:Collider_bias.png

Simpson's Paradox

Simpson's paradox is a statistical paradox where:

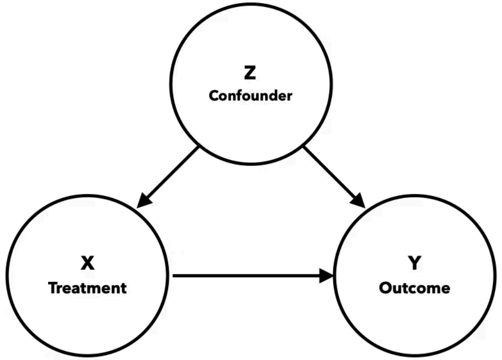
- The **direction of an association reverses** when data are **aggregated vs stratified**, due to the presence of a **confounder** that affects both the predictor and the outcome.

In other words:

Sometimes a pattern looks one way **when you look at everyone together**, but it looks **the opposite** when you look at the **subgroups separately**.

This is called **Simpson's Paradox**.

The DAG Behind Simpson's Paradox



Understanding the causal structure using a Directed Acyclic Graph (DAG) helps illuminate Simpson's Paradox:

- **Z → X**: The confounder (Z) influences which group individuals belong to or which treatment they receive (X).
- **Z → Y**: The confounder (Z) also directly influences the outcome (Y).
- **X → Y**: There is a true causal effect from X to Y that we are interested in measuring.

Key Insight: Failing to adjust for the confounder (Z) can lead to a reversal of the apparent effect of X on Y when data are aggregated, compared to when they are analyzed within subgroups.

Bottom line interpretation:

Simpson's paradox arises from the interplay of **confounding** and **uneven group sizes** or distributions influenced by that confounder.

Summary

1

Special Cases of Collider Bias

Birth-weight paradox, M-bias, and Berkson's paradox are well-documented examples where conditioning on colliders creates spurious associations.

2

Simpson's Paradox

A confounding phenomenon where associations reverse between aggregated and stratified data due to uneven group distributions.

Let's do unit 5 quiz: