HSS 8005

Quantitative analysis +

Week 5

Biases

Considerations for causal analysis

2024 Dr Chris Moreh

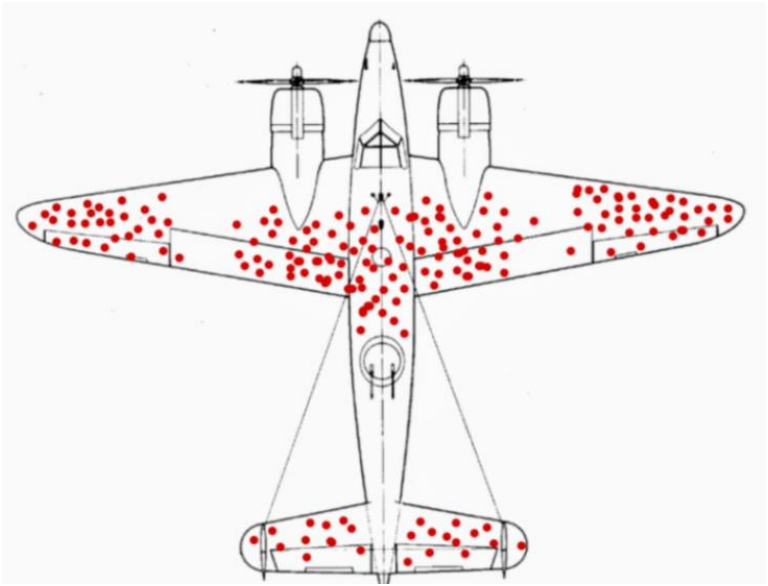
Press F to view full-screen

View on ReCap

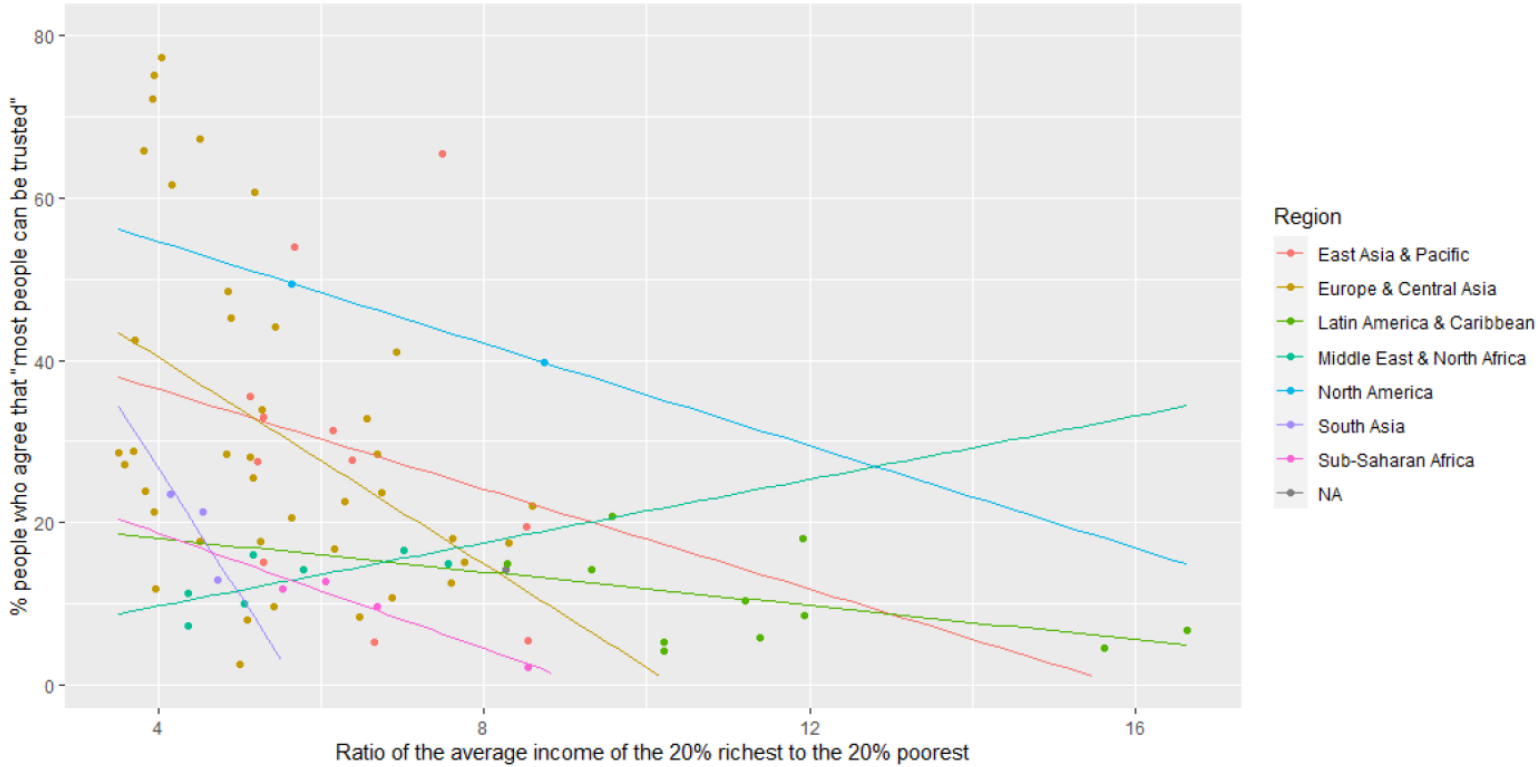
Biases

- Non-statistical biases
  - Data selection?
  - Interpretation?
  - Wrong research question?
- Statistical biases
  - “good” and “bad” controls?
- Causal thinking =/= statistical thinking

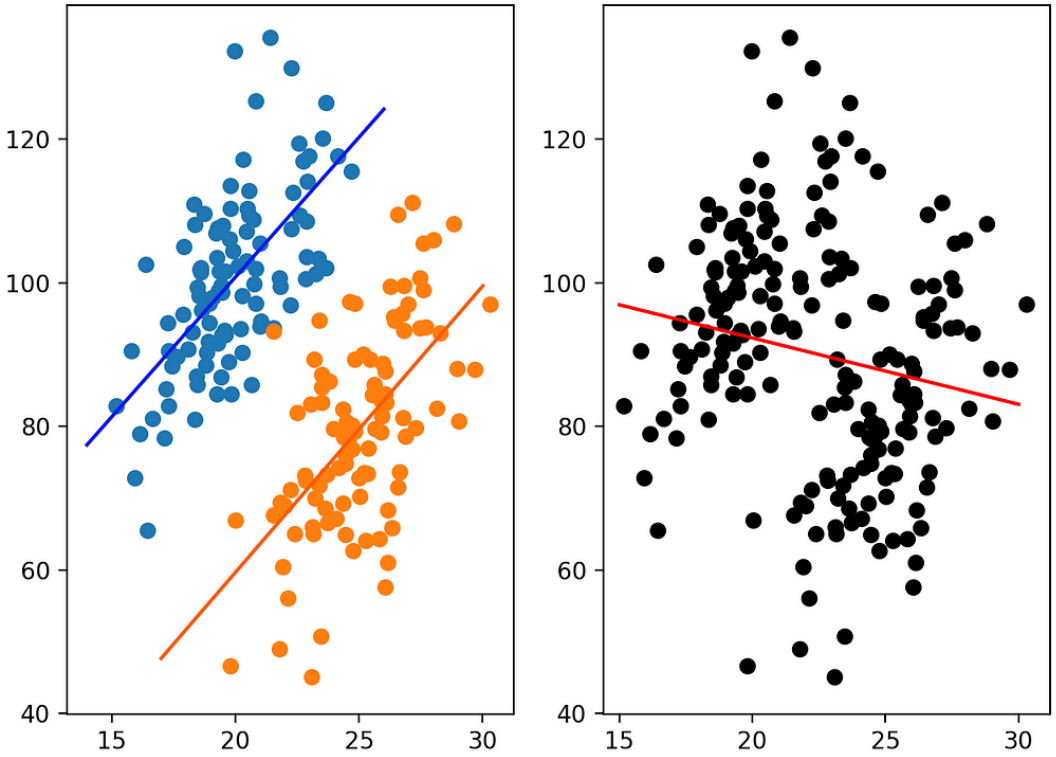
Biases



“Simpson’s” “paradox”



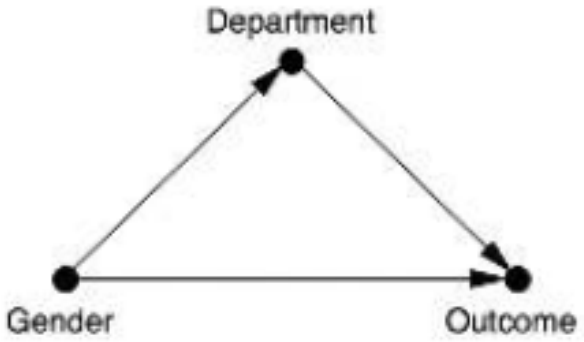
“Simpson’s” “paradox”

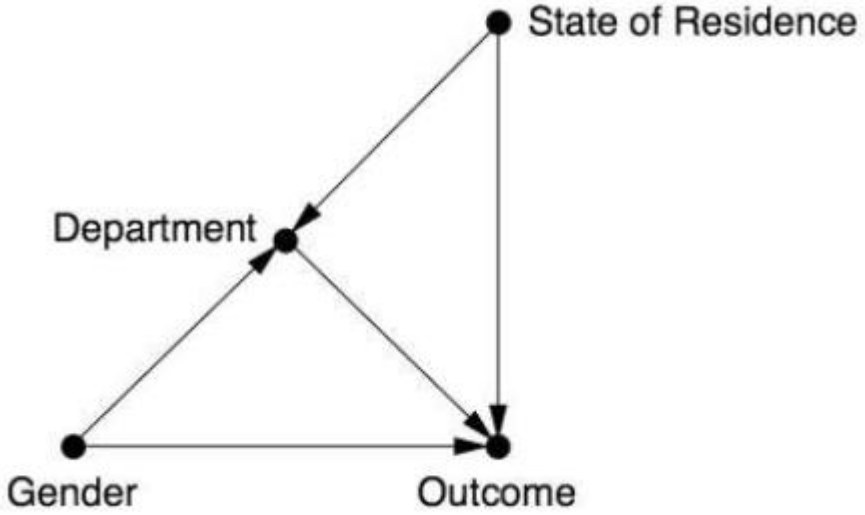


“Simpson’s” “paradox”

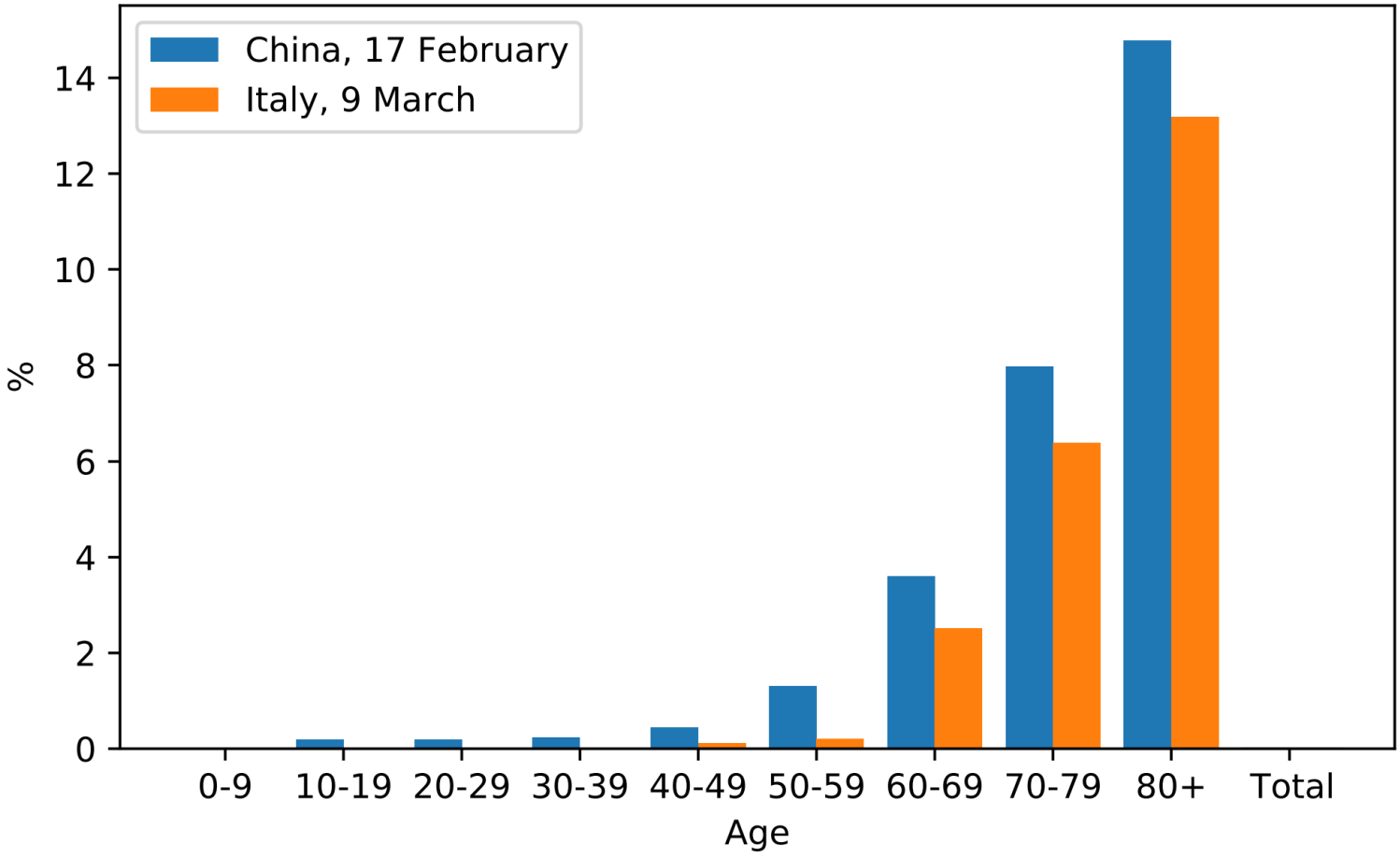
	Men		Women	
	# Applied	% Admitted	# Applied	% Admitted
Major A	825	62% admitted	108	82% admitted
Major B	560	63% admitted	25	68% admitted
Major C	325	37% admitted	593	34% admitted
Major D	417	33% admitted	375	35% admitted
Major E	191	28% admitted	393	24% admitted
Major F	373	6% admitted	341	7% admitted
Total:	2,690	44% admitted	1,835	34.5% admitted

“Simpson’s” “paradox” - causal answer?

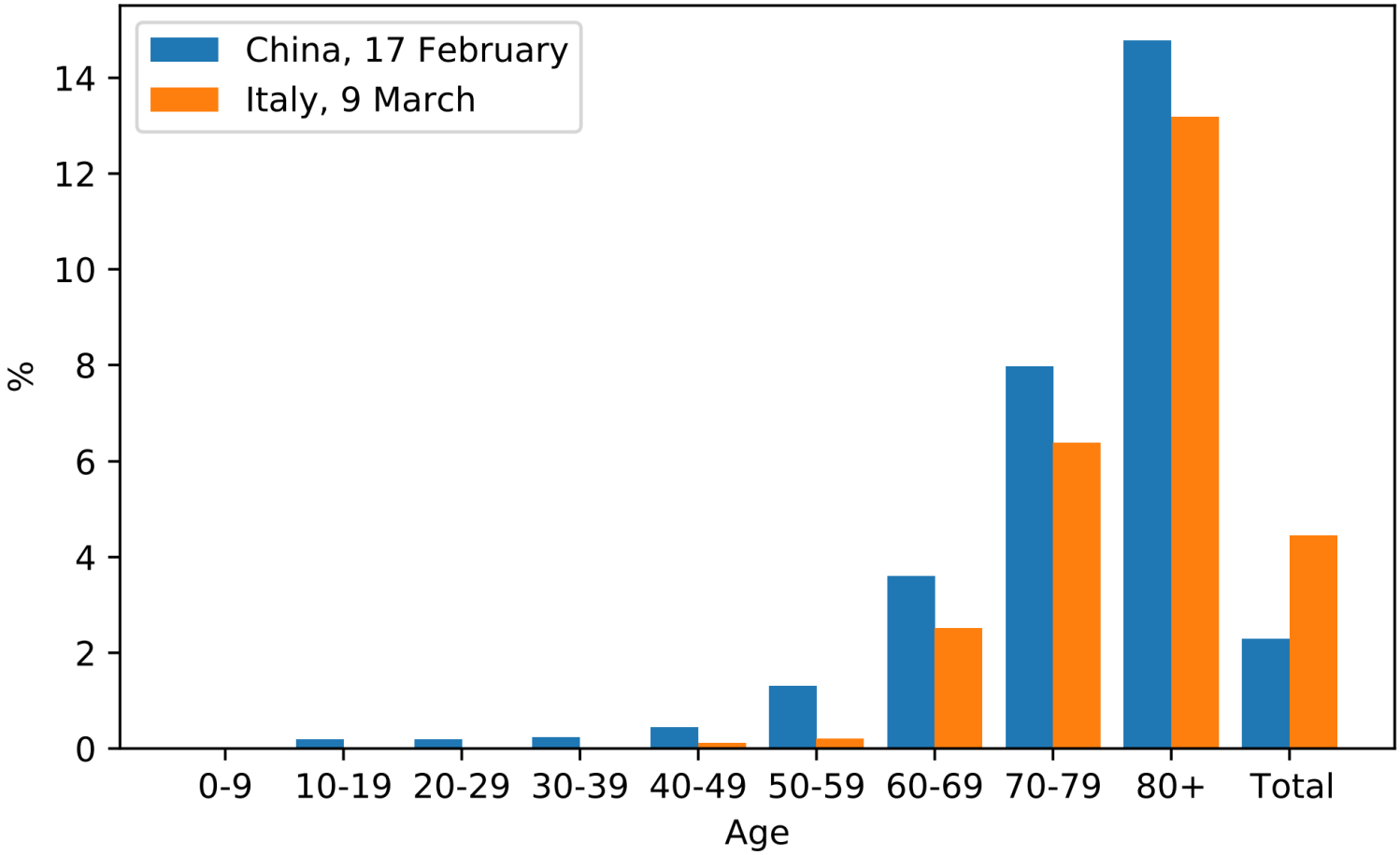




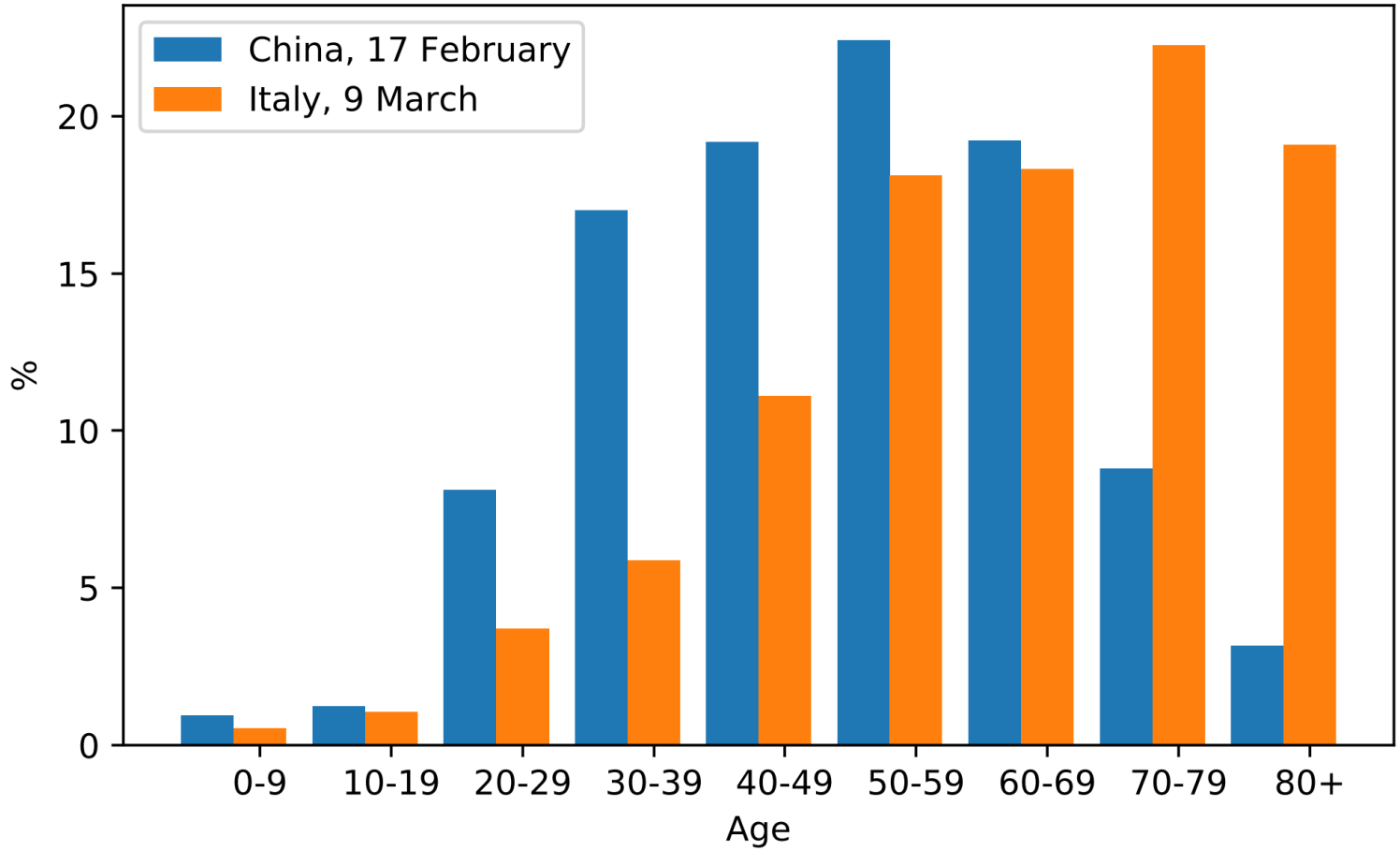
Covid-19 Case fatality rates (CFRs) by age group, China vs. Italy



Covid-19 Case fatality rates (CFRs) by age group, China vs. Italy



Proportion of confirmed Covid-19 cases by age group, China vs. Italy



Correlation and Causation

SCIENCE

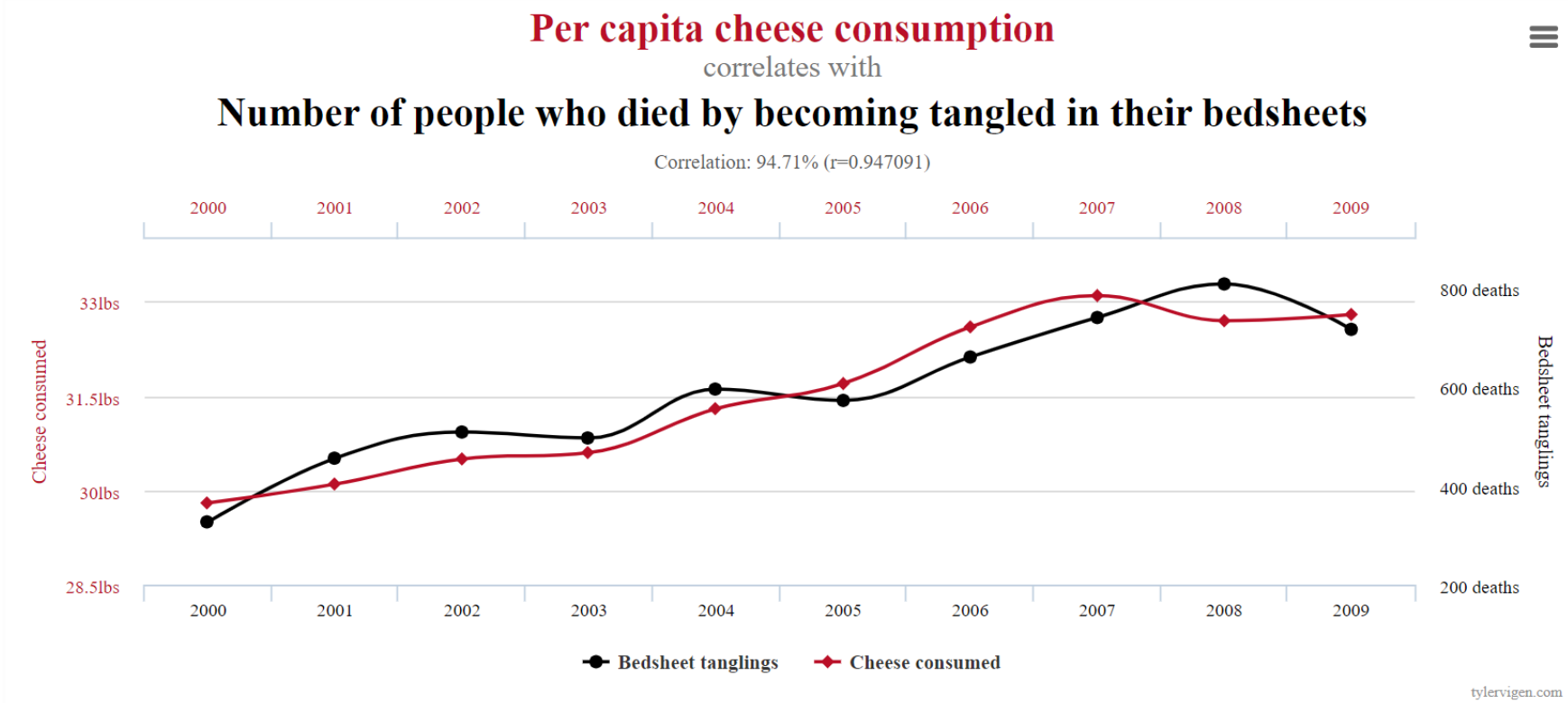
The Internet Blowhard's Favorite Phrase

Why do people love to say that correlation does not imply causation?

By DANIEL ENGBER

OCT 02, 2012 • 8:33 AM

Not everyone found the news believable. "Facepalm. Correlation doesn't imply causation," wrote one unhappy Internet user. "That's pretty much how I read this too... correlation is NOT causation," agreed a Huffington Post superuser, seemingly distraught. "I was surprised not to find a discussion of correlation vs. causation," cried someone at Hacker News. "Correlation does not mean causation," a reader moaned at Slashdot. "There are so many variables here that it isn't funny."



Data sources: U.S. Department of Agriculture and Centers for Disease Control & Prevention

Correlation and causation

- When we say X causes Y, we mean the connection has a one-way direction, going from X to Y (an **asymmetric** relationship)
- The term ‘reciprocal causality’ is used, to mean that X causes Y and Y causes X as well (an underlying causal mechanism has not been clearly identified)
- Causation commonly seen as a subset of correlation when the direction of a relationship is specified.

Correlation ~~and causation~~

"When we **vary the cause**, the phenomenon changes, but not always to the same extent; it changes, but has **variation in its change**.

The less the variation in that change the more nearly the cause defines the phenomena, the more closely we assert the **association** or the **correlation** to be.

It is this conception of **correlation** between two occurrences embracing all relationships from absolute independence to complete dependence, which **is the wider category by which we have to replace the old idea of causation**."

(Karl Pearson, *The Grammar of Science*, 1892)

Correlation and causation

Correlation and causation

Three criteria for inferring a causal relationship (Agresti and Finlay, 2008: 302):

(1) covariation between the presumed cause and outcome;

(2) temporal precedence of the cause; and

(3) exclusion of alternative explanations for cause-outcome connections.

Correlation and causation

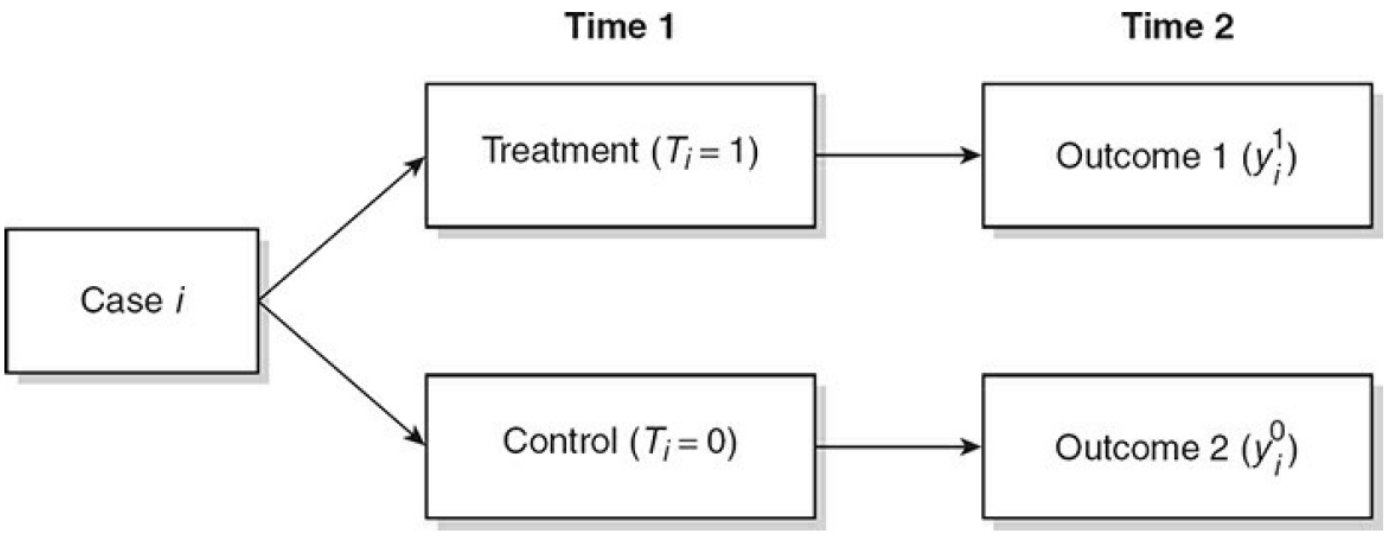
"A variable X is a cause of a variable Y if Y in any way **relies** on X for its value.… X is a cause of Y if Y **listens** to X and **decides** its value in response to what it hears"

(Pearl, Glymour, and Jewell 2016, 5–6)

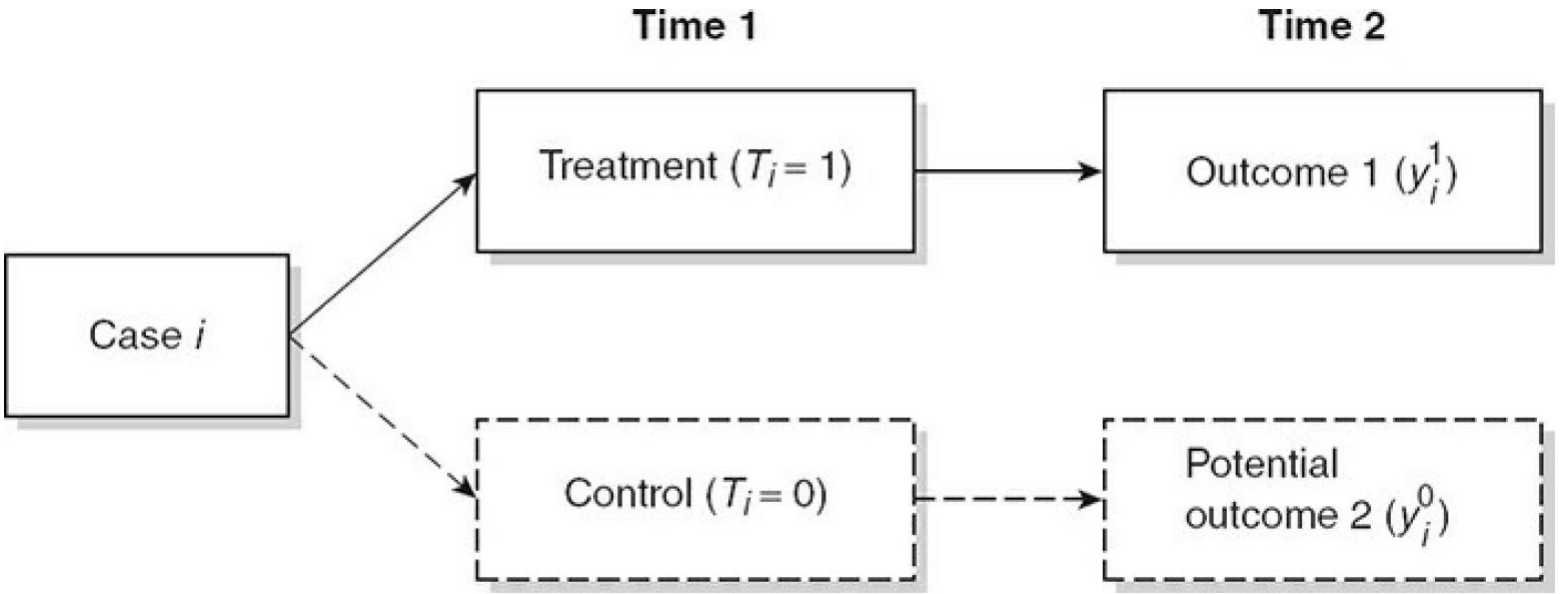
X → Y

		Outcome (Y)	
		Has happened	*Has not happened*
Cause (*X*)	Exists	Scenario 1	Scenario 2
	Does not exist	Scenario 3	Scenario 4

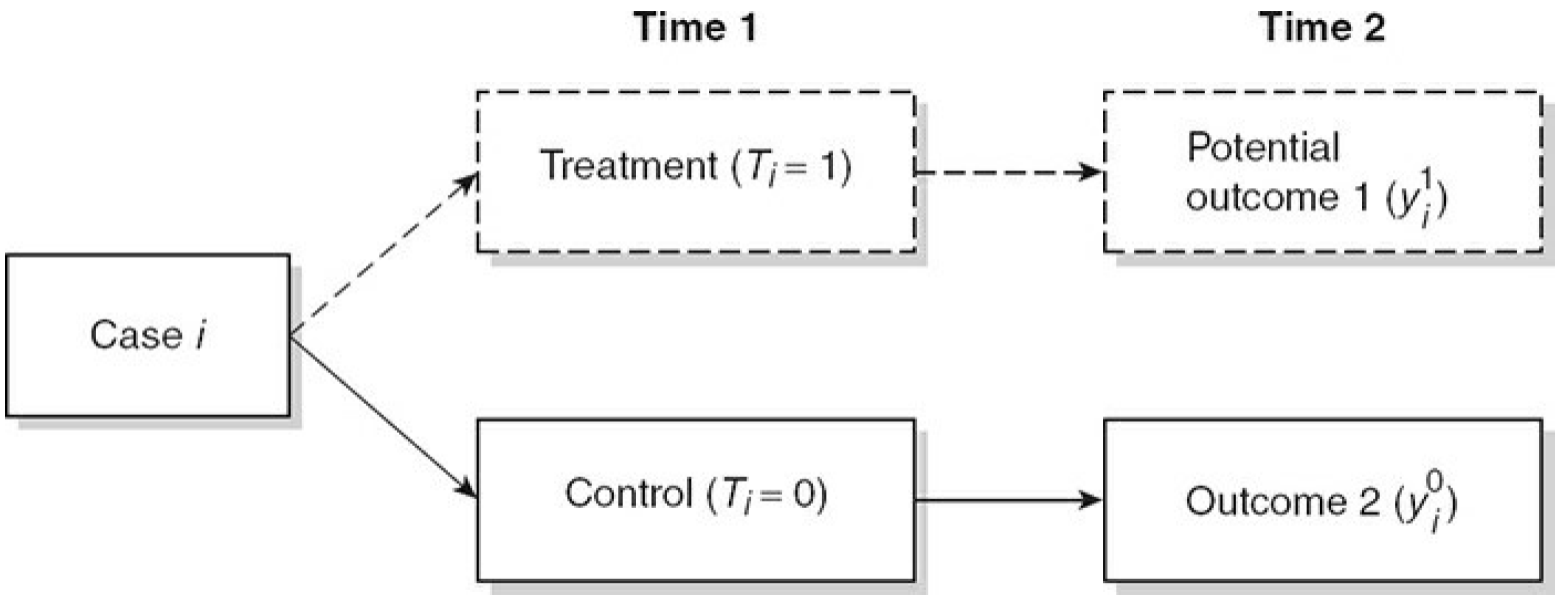
T1 → T2: the ideal situation



T1 → T2: common situations



T1 → T2: common situations



T1 → T2: common situations

- the fundamental difficulty for establishing causal relations *at the individual level* is the impossibility of collecting information on the **potential** outcomes
- we may then be able to explore causal relations *at the group level*. That is, we define the causal effect – more precisely, the mean of the causal effect – as the difference between the means of the two groups on the interested variable

T1 → T2: common situations

- In observational studies, we often only have a sample. We are therefore attempting to use the sample means (average) to estimate the means (average) in the population
- We must assume that *the treatment group and the control group are the same with regard to the mean of* Y. If we can somehow show that the two groups are equivalent or substitutable, then we can use the average **sample** causal effect as an acceptable estimator of the targeted **population** average causal effect

T1 → T2: common situations

- The best, albeit not certain, way of satisfying these assumptions is to **assign** the cases **randomly** into different causal conditions
- The idea is that randomization will virtually eliminate any possible systematic differences between the two groups (treatment and control) so that they can be deemed as substitutable
- **But in observational studies the cases assign themselves** to a particular condition

T1 → T2: common situations

- the challenge is *to ensure the comparability of the groups even though they are not formed through a random-assignment process*
- Standard **regression** approach: if we think that the cases have selected themselves for different causal conditions because of a particular attribute, then we can divide the cases into groups according to the levels of that attribute

Thinking causally (example)

Leahey, Erin. 2007. 'Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings'. *American Sociological Review* 72(4):533–61. doi: 10.1177/000312240707200403.

Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings

Erin Leahey
University of Arizona

The popular adage "publish or perish" has long defined individual career strategies as well as scholarly investigations of earnings inequality in academe, as researchers have relied heavily on research productivity to explain earnings inequality among faculty members. Academia, however, has changed dramatically in the last few decades: it has become larger and more demographically diverse, and fears of overspecialization prompt calls for interdisciplinary approaches. In this new environment, other factors, in addition to productivity, are likely relevant to our understanding of earnings differentials. In this article, I assess whether two additional factors—visibility and the extent of research specialization—contribute to men's earning advantage. Using probability samples of tenure-track academics in two disciplines, a variety of data sources, and innovative measures, I find that both factors are highly relevant to the process by which earnings are determined. Women earn less than men largely because they specialize less. Lower levels of specialization hinder productivity, productivity enhances visibility, and visibility has a direct, positive, and significant effect on salary. I discuss the practical implications of these findings and lay the foundation for a broader theory of the role of research specialization in work processes.

Thinking causally (example)

- Leahey (2007) has noted that, on average, female academics in many fields are paid less than men and also have lower research productivity
- Imagine if someone were to argue that the entire reason female academics are paid less than men is that they are less productive. We might then diagram the proposed relationship between gender (G), productivity (P), and earnings (E) like this:

G → P → E

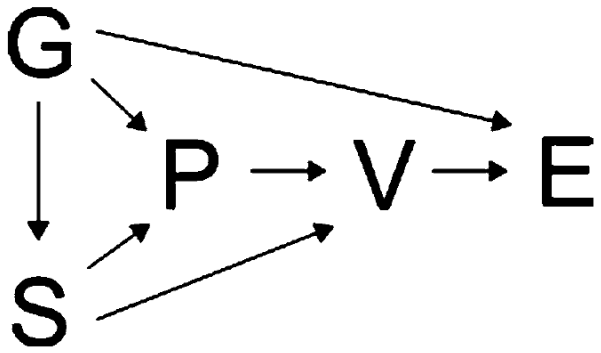
Thinking causally (example)

- Leahey proposes that gender causes differences in the degree of specialization by academics, and degree of specialization (*S*) enhances productivity.
- She also posits that productivity differences cause differences in visibility (*V*) among academics, and differences in visibility cause differences in earnings:

G → S → P → V → E

Thinking causally (example)

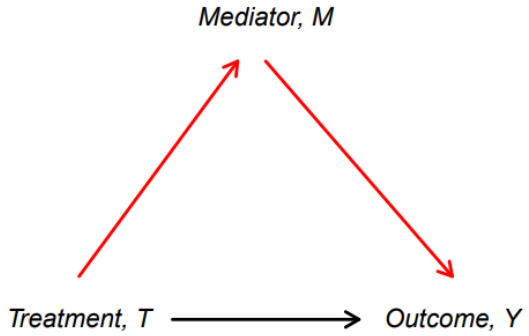
- More commonly, social science proceeds by identifying more proximate causes that **partially**, rather than strictly, **mediate** the relationship between a cause and an outcome.
- For example, the theoretical model that Leahey proposes for the relationship between gender and earnings is actually:



RCTs vs Observational studies, Causal mediation

- Randomized experiments as gold standard for causal inference
- But, experiments are a black box
- Can only tell whether the treatment causally affects the outcome
- Not how and why the treatment affects the outcome
- Qualitative research uses process tracing

Graphical representation



Goal is to decompose total effect into direct and indirect effects

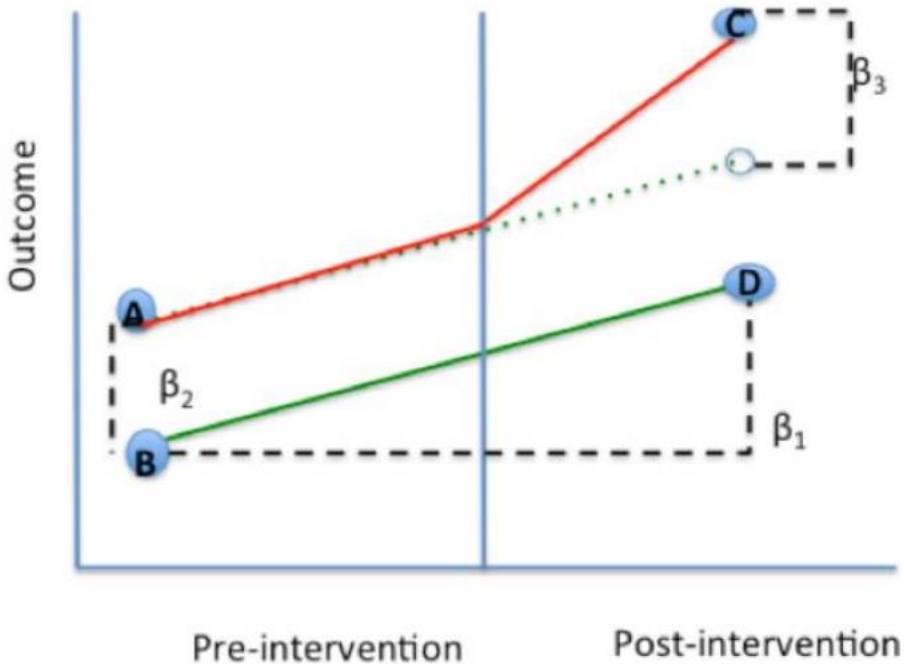
Alternative approach: decompose the treatment into different components

Causal mediation analysis as quantitative process tracing

Difference in differences

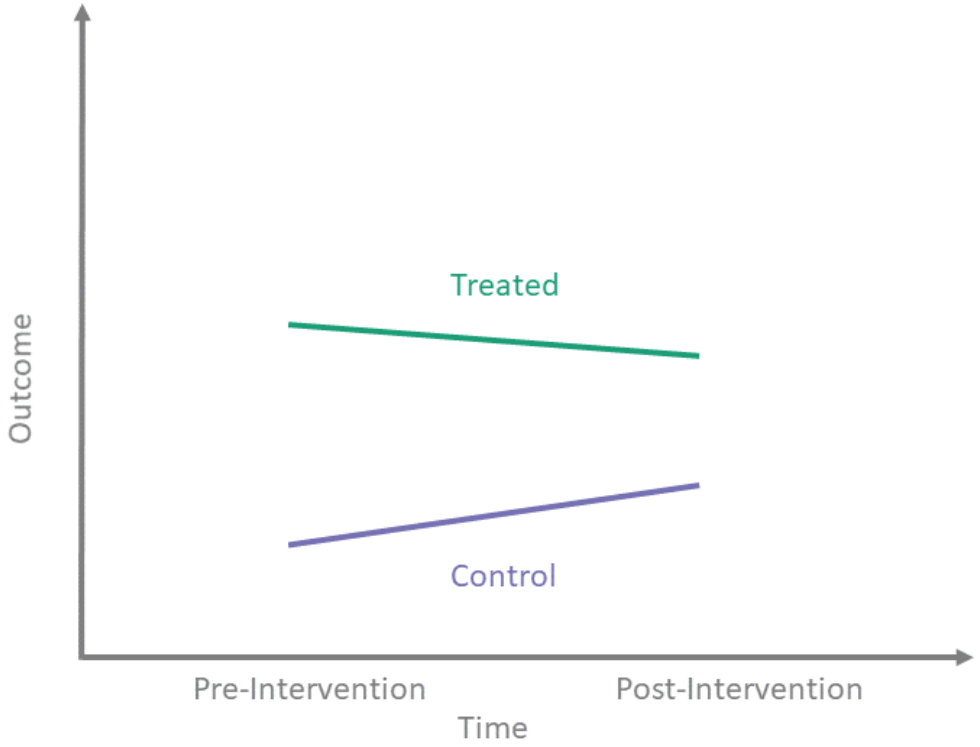
- Differences-in-Differences regression (DiD) is used to assess the causal effect of an event by comparing the set of units where the event happened (treatment group) in relation to units where the event did not happen (control group).
- The logic behind DiD is that if the event never happens, the differences between treatment and control groups should stay the same overtime, see graph next slide

Difference in differences

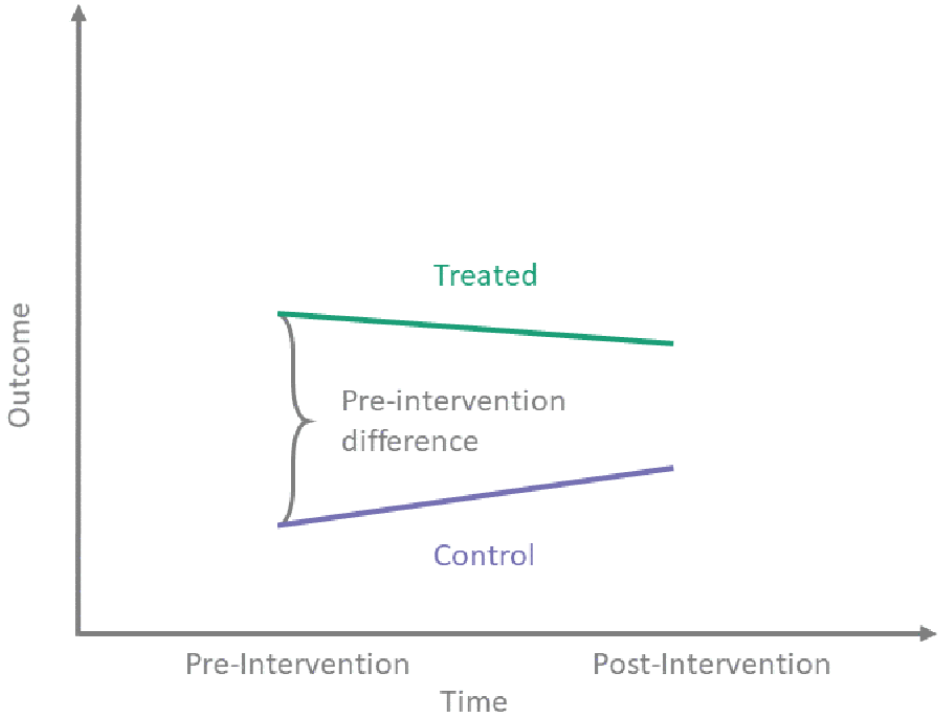


Coefficient	Calculation	Interpretation
β_0	B	Baseline average
β_1	D-B	Time trend in control group
β_2	A-B	Difference between two groups pre-intervention
β_3	(C-A)-(D-B)	Difference in changes over time

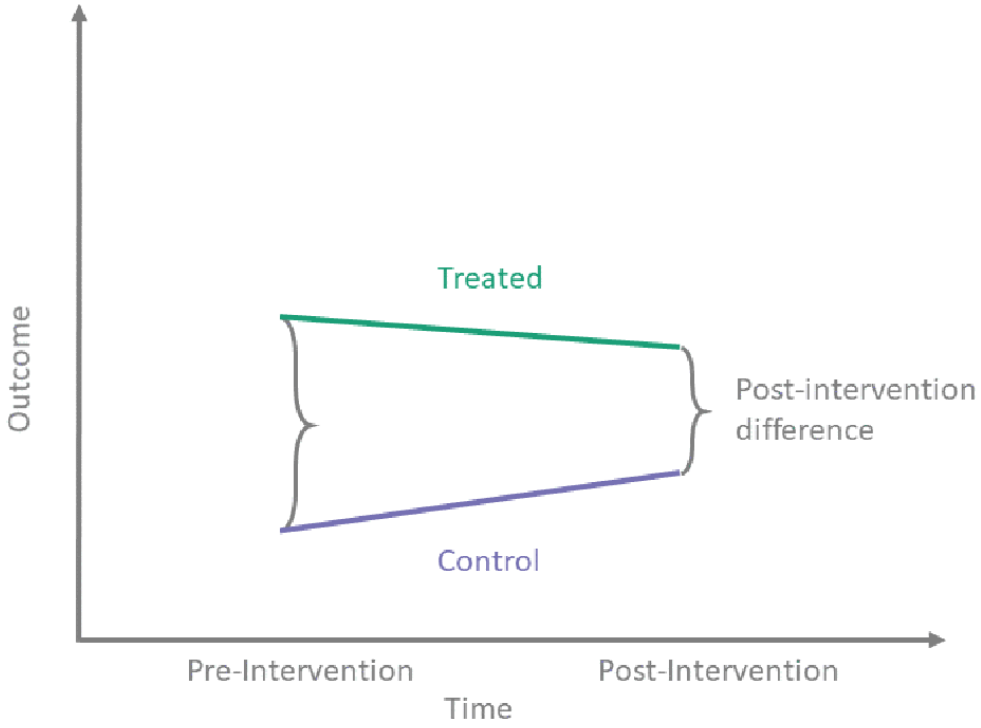
Difference in differences



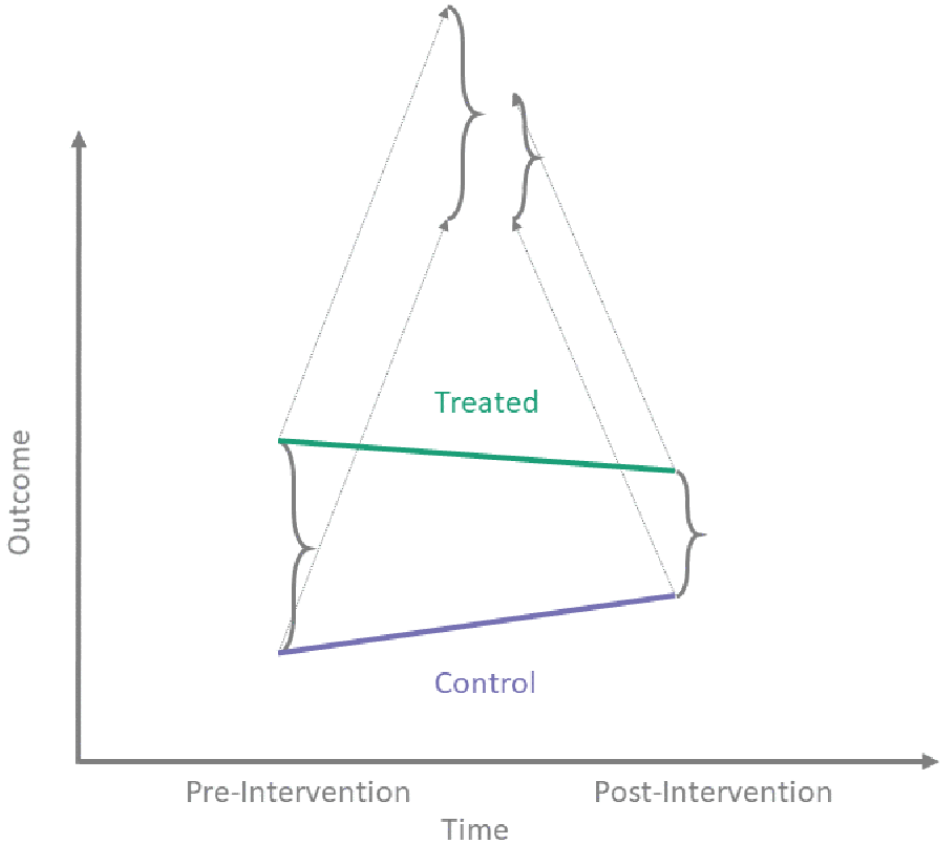
Difference in differences



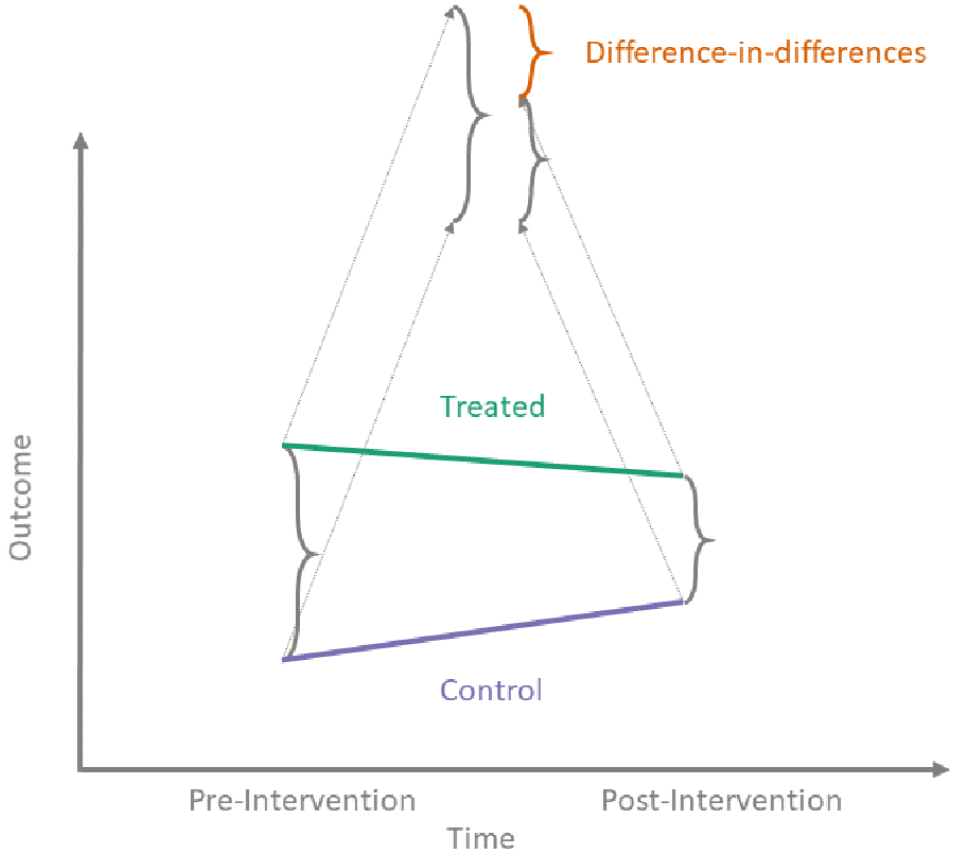
Difference in differences

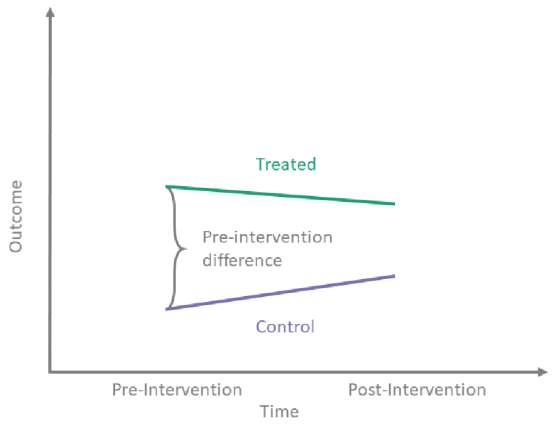


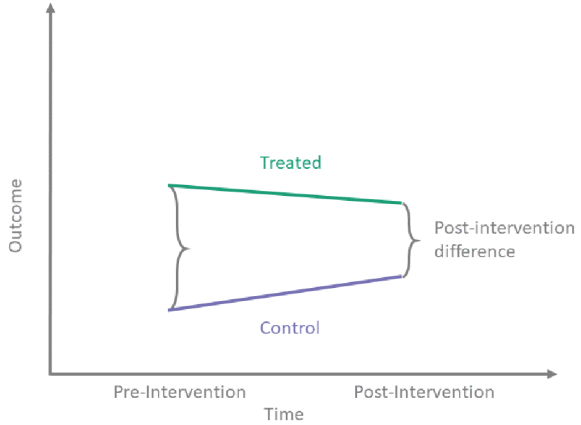
Difference in differences

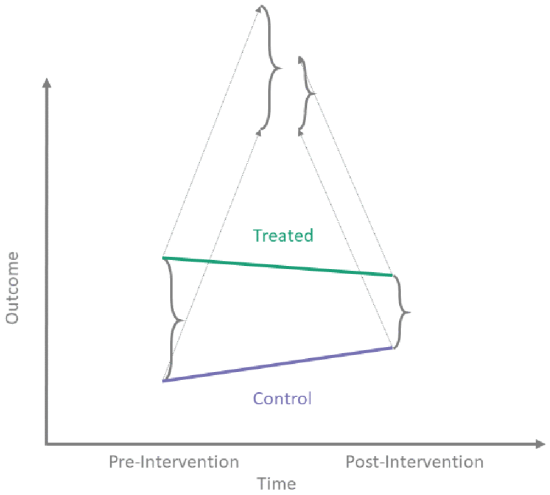


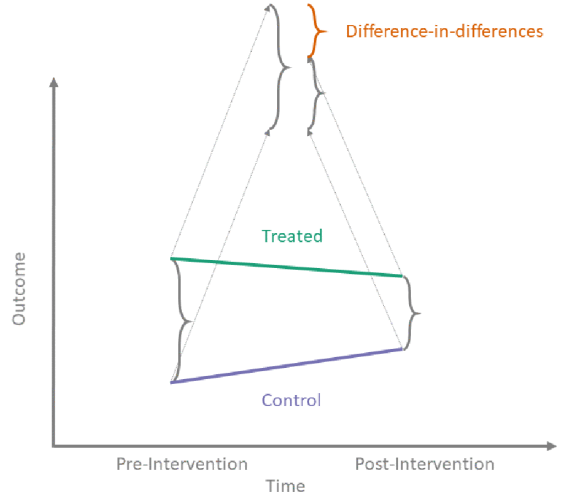
Difference in differences




Treated
Pre-intervention
difference
Control
Outcome
Pre-Intervention
Post-Intervention
Time


Treated
Post-intervention
difference
Control
Outcome
Pre-Intervention
Post-Intervention
Time


Treated
Control
Outcome
Pre-Intervention
Post-Intervention
Time


Difference-in-differences
Treated
Control
Outcome
Pre-Intervention
Post-Intervention
Time

Worksheet example

Exploiting a Rare Communication Shift to Document the Persuasive Power of the News Media

Jonathan McDonald Ladd Georgetown University
Gabriel S. Lenz Massachusetts Institute of Technology

Using panel data and matching techniques, we exploit a rare change in communication flows—the endorsement switch to the Labour Party by several prominent British newspapers before the 1997 United Kingdom general election—to study the persuasive power of the news media. These unusual endorsement switches provide an opportunity to test for news media persuasion while avoiding methodological pitfalls that have plagued previous studies. By comparing readers of newspapers that switched endorsements to similar individuals who did not read these newspapers, we estimate that these papers persuaded a considerable share of their readers to vote for Labour. Depending on the statistical approach, the point estimates vary from about 10% to as high as 25% of readers. These findings provide rare evidence that the news media exert a powerful influence on mass political behavior.