Simpson's Paradox & Causal Calculus Toolbox

Mark Goh
17/11/21

Contents

- Simpson’s Paradox
- Randomized Controlled Test
- Causal Calculus Toolkit
  - Causal Models
  - Causal Conditional Probabilities
  - d-separation

What is the Simpson's Paradox?

~~Why the show isn't cancelled~~

Explain via examples first

Clinical Drug Test

Based on the statistics below, does a drug treatment (C) leads to a better recovery rate (E) in patients?

Combined	E	¬E		Recovery Rate
Drug (C)	20	20	40	50%
No-Drug (¬C)	16	24	40	40%

What about now?

M	E	¬E		Recovery Rate
Drug (C)	18	12	30	60%
No-Drug (¬C)	7	3	40	70%

¬M	E	¬E		Recovery Rate
Drug (C)	2	8	10	20%
No-Drug (¬C)	9	21	30	30%

Clinical Drug Test

- Overall, recovery rates for patients receiving drug > control (treatment is preferred)
- Breaking down into males & females, recovery rate for control > treated patient (for both case)

Which partition do we choose?

Intuitively: take both

College Admissions

Based on the statistics below, is there discrimination in acceptance rates?

Combined	A	¬A		Acceptance Rate
M (C)	250	300	550	45.5%
F (¬C)	250	400	650	38.5%

What about now?

Fac A	A	¬A		Acceptance Rate
M (C)	200	200	400	50%
F (¬C)	100	100	200	50%

Fac B	A	¬A		Acceptance Rate
M (C)	50	100	150	33%
F (¬C)	150	300	450	33%

College Admissions

- Overall, acceptance rate for males > females (treatment is preferred)
- Breaking down into faculties, acceptance rate is the same (for all fac)

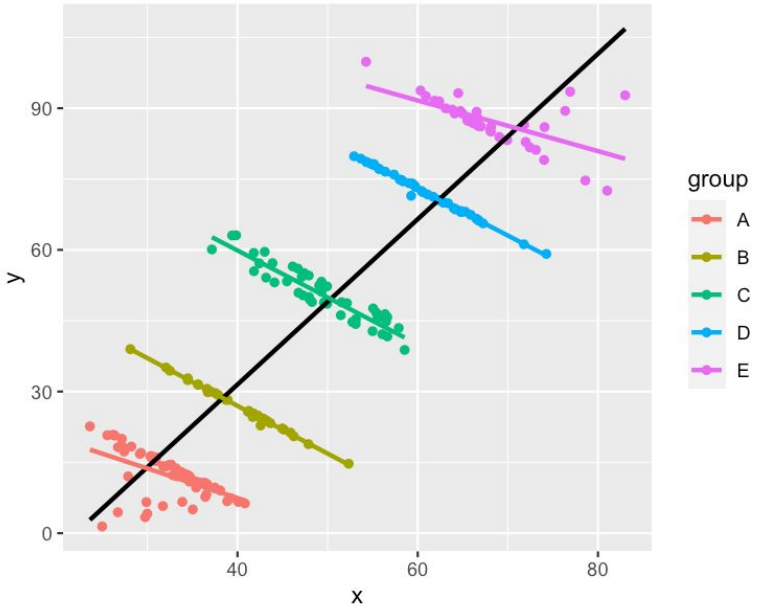
Which partition do we choose?

Intuitively: separate faculties

Simpson's "Paradox"

Phenomenon in probability/statistics where a trend appears in several groups of data but disappears or reverses when the groups are combined

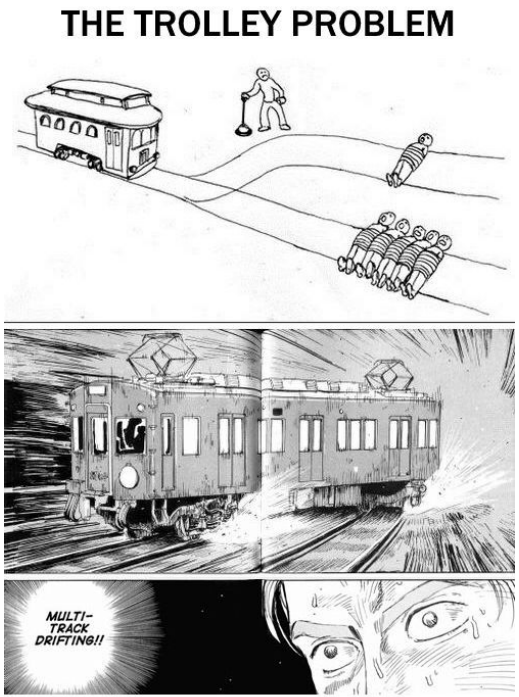
Not a real paradox, but a failure of reasoning with statistics, improper choice of group partition, and linking correlation and causation



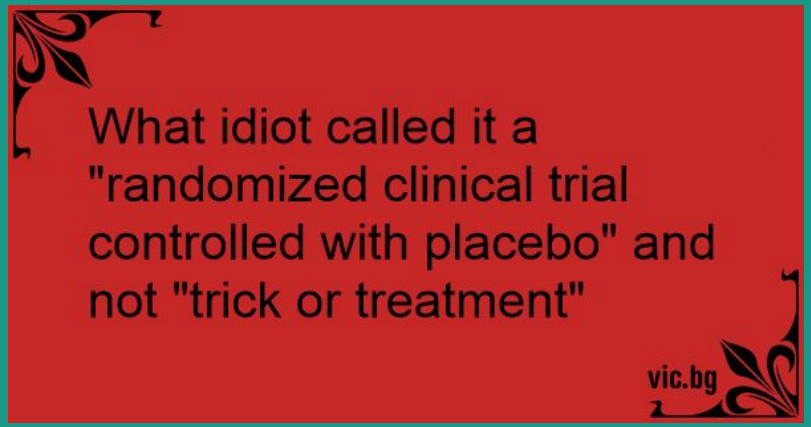
How do we choose?

Can’t leave it to intuition
(differs between people)

Context dependent: Need to identify causal relation. How do we do so?

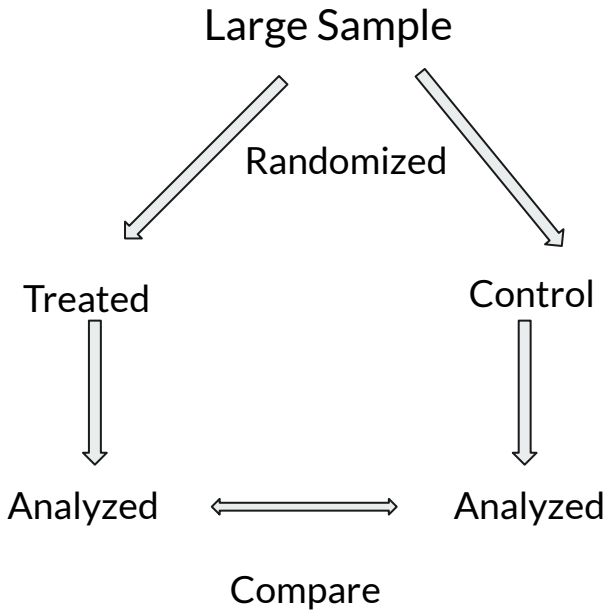


Randomised Controlled Trial



Randomised Controlled Trial (RCT)

- Gold standard for checking the causal effect between 2 variables



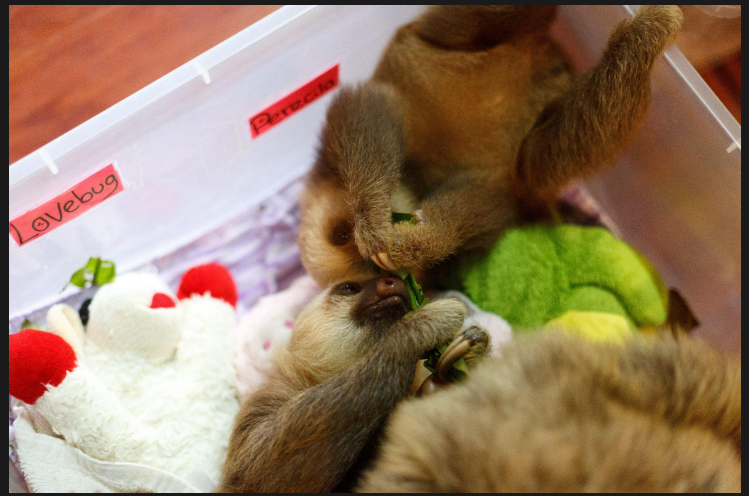
Problem

Idealistic

- Need large sample size
- Timely & Expensive
- Potentially Illegal and Unethical

Intermission: enjoy the sight/cute animal

Source: my/friend’s phone/camera



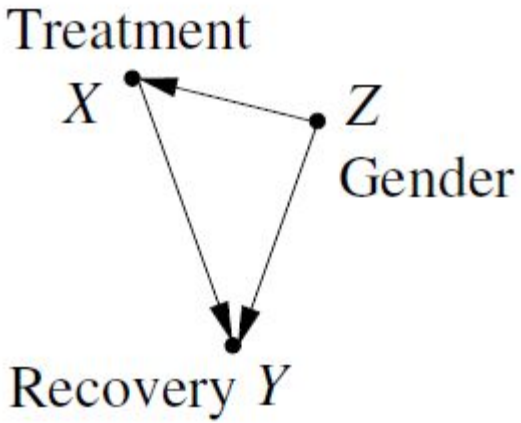
Causal Calculus Toolbox

Causal Models

Introduce causal diagram to identify cases where Simpson's paradox might occur

Directed Acyclic Graph (DAG)

Q: why no cyclic graphs?



Causal Probability

Define causal influence by requiring variable to be a function of its parents

$$X_j = f_j\left(X_{pa(j)}, Y_{j1}, Y_{j2}, \cdots\right)$$

Where Y_j

- Are independent of each other
- Independent of all $X_{k \neq j}$

Causal Conditional Probabilities

Hypothetical RCT

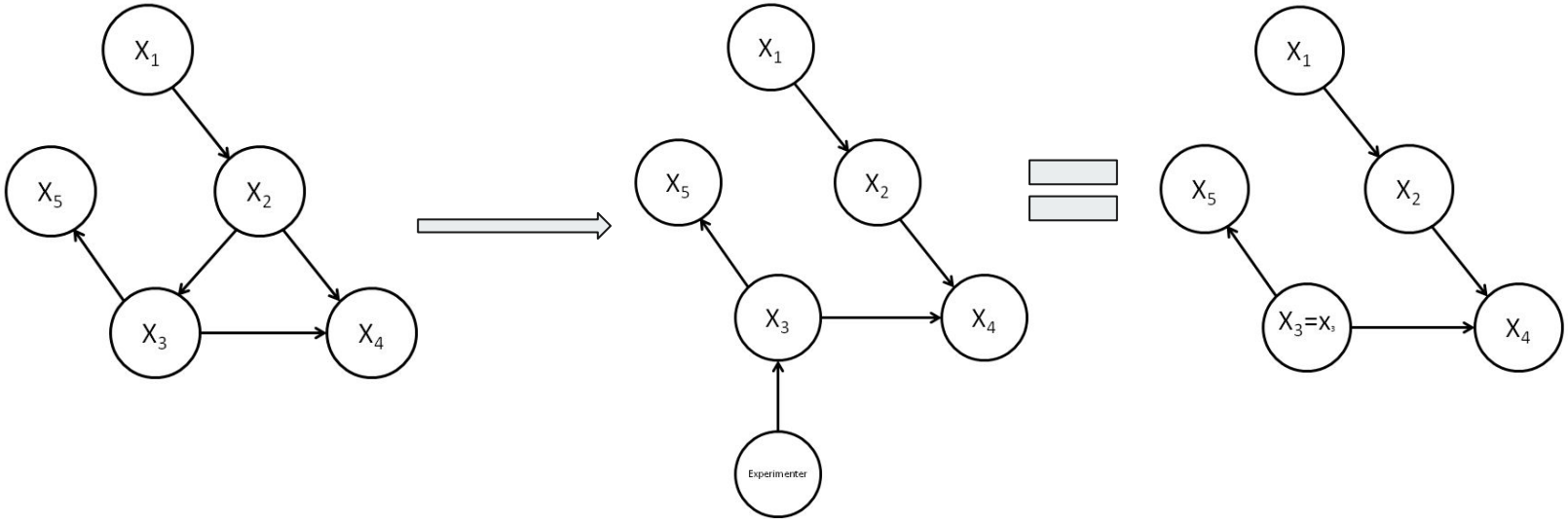
Consider hypothetical world where we can forced a person to do X or ¬X

Introduce a *conditional causal probability* in said hypothetical world

$$p\left(\text{Drunk}|\text{do}[\text{Karneval}]\right)$$

Read as: (Causal conditional) probability of getting drunk given that we “do” karneval (i.e. someone has been “forced” to go celebrate karneval in a “hypothetical” scenario)

Example (in terms of causal diagram)



What's the Point?

Normal conditional probability/calculus cannot identify causal structure, only correlation

In some cases, can link usual conditional probabilities to causal conditional probabilities

Summary

- What the Simpson's "paradox" is
- What is a Randomised Controlled Trial
- Basic toolkit for Causal Calculus

References

- Pearl, Judea. (2013). Understanding Simpson's Paradox. SSRN Electronic Journal. 68. 10.2139/ssrn.2343788.
- https://michaelnielsen.org/ddi/if-correlation-doesnt-imply-causation-then-what-does/
- Sprenger, Jan and Naftali Weinberger, "Simpson's Paradox", *The Stanford Encyclopedia of Philosophy* (Summer 2021 Edition), Edward N. Zalta (ed.),
  [WARNING FOR THIS SOURCE, to Quote my philosophy Prof ,, ah very Stanford, very detail but also unreadable ”

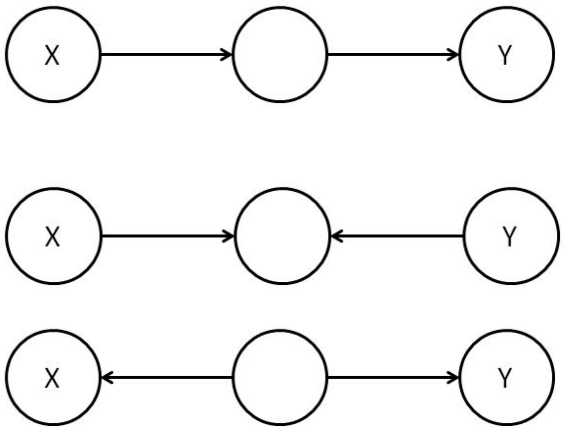
Definitions :'(

X and Y are **d-connected** when knowing X gives us some information about Y

X and Y are **d-seperated** when knowing X or Y doesn't give us new information about each other

A node is a **collider** if it has only incoming links

A node is a **fork** if it has only outgoing links

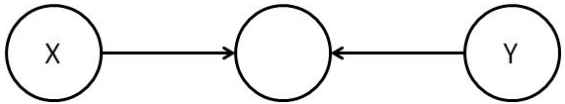


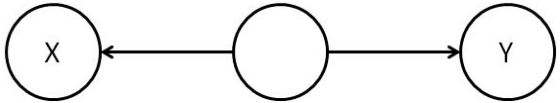
More Definitions T.T

A path from X to Y that has a collider is a **blocked path**, else, it is an **unblocked path**

X and Y are **d-connected** if there is an **unblocked path**. If no such path exists, they are **d-seperated**

Q: What about a fork?

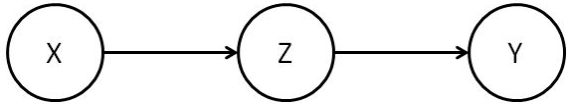




More complications (i.e. 3rd variable) 小三

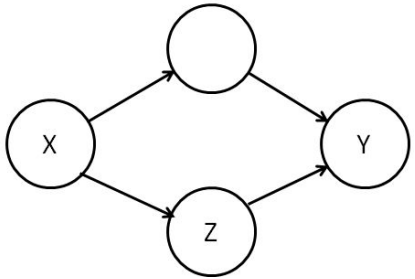
If we know Z, then knowing X doesn't give us any new info on Y

X and Y are **d-separated** given Z



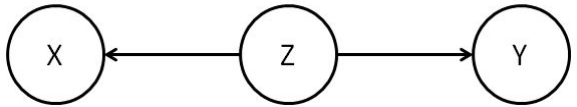
Even if we know Z, knowing X can still give us new info on Y

X and Y are **d-connected** given Z

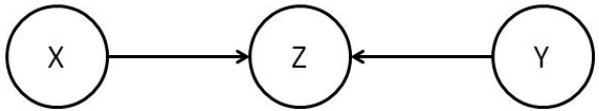


3rd variable

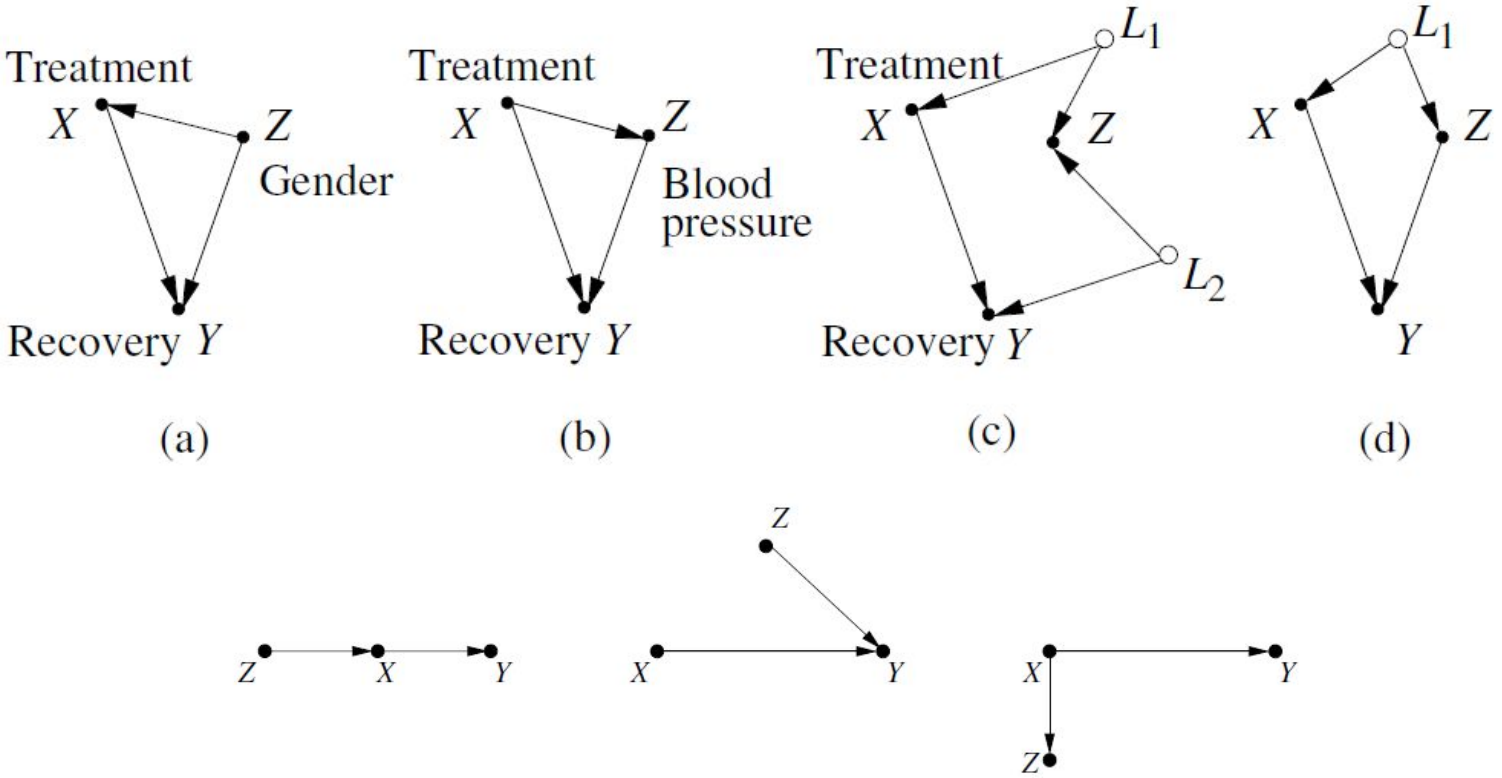
If we know Z, then knowing X doesn't give us new info on Y and vice versa



Q: What about a collider?



Examples


Treatment
X
Z
Gender
Recovery Y
(a)
Treatment
X
Z
Blood pressure
Recovery Y
(b)
Treatment
X
L_1
Z
L_2
Recovery Y
(c)
L_1
X
Z
Y
(d)
Z
X
Y
X
Y
Z
X
Y
Z