

Introduction to Causal Modeling

Session 1: Causal Inference

Oisín Ryan

Department of Methodology and Statistics
Faculty of Social Sciences
Utrecht University

November 1, 2021

Today: Causal Inference.

- Given our beliefs about the system we are studying (i.e., our causal model), can we estimate a causal effect from data? If so, how?

December: Causal Discovery

- Given data, but no strong beliefs about the causal system, can we learn the causal model?

- ① Why Causal Modeling?
- ② Potential Outcomes
 - Hands-on Part 1
- ③ Structural Causal Models
 - Hands-on Part 2
- ④ Discussion

Why Causal Modeling?

Causal questions dominate scientific research

- Does smoking cause cancer?
- Does the expression of gene X produce phenotype Y?
- What is the effect of social media use on adolescent well-being?
- What effect could we expect a sugar tax to have on rates of adult-onset diabetes in the general population?
- Which treatment type will be most effective in reducing symptoms for this type of individual?

Why Causal Modeling?

Causal questions dominate scientific research

- Does smoking cause cancer?
- Does the expression of gene X produce phenotype Y?
- What is the effect of social media use on adolescent well-being?
- What effect could we expect a sugar tax to have on rates of adult-onset diabetes in the general population?
- Which treatment type will be most effective in reducing symptoms for this type of individual?

Causal Modeling: When can we answer causal questions using data? And how should we go about doing this?

Why Causal Modeling?

Statistical modeling¹ gives us a rich language to describe uncertainty in the world we see around us

- The language of *co-occurrences*, *expected values*, (*joint, marginal and conditional*) *probabilities* and *statistical dependencies*.
- It helps us *describe* patterns and make (certain types of) *predictions*.

¹Insert your favourite term here: machine learning / data science / probability theory

Why Causal Modeling?

Statistical modeling¹ gives us a rich language to describe uncertainty in the world we see around us

- The language of *co-occurrences*, *expected values*, (*joint, marginal and conditional*) *probabilities* and *statistical dependencies*.
- It helps us *describe* patterns and make (certain types of) *predictions*.

But *by itself*, statistical models have very little to say about causal relations!

Causal Modeling involves using concepts and techniques from statistical modeling

- But causal models and causal information exist on a level **above** statistical information

¹Insert your favourite term here: machine learning / data science / probability theory

Example

Imagine we are a team of epidemiologists.

We take a blood sample from a random sample of the population and record:

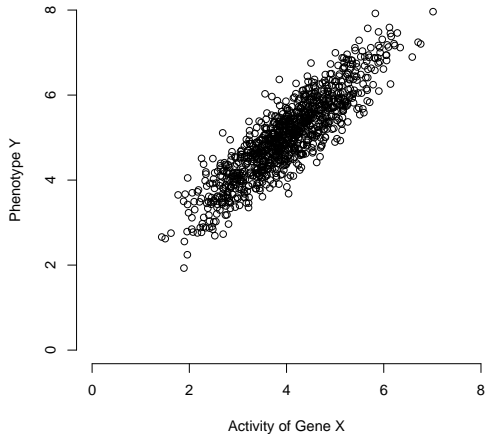
- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

Example

Imagine we are a team of epidemiologists.

We take a blood sample from a random sample of the population and record:

- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).

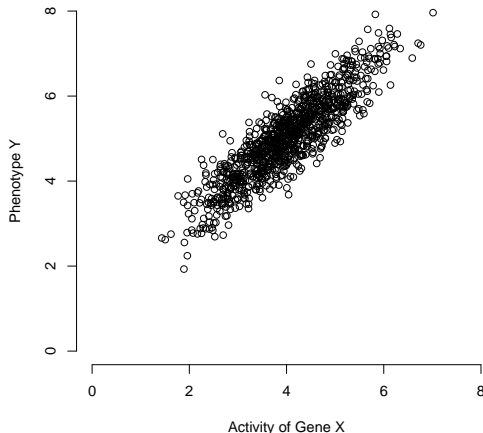


Example

Imagine we are a team of epidemiologists.

We take a blood sample from a random sample of the population and record:

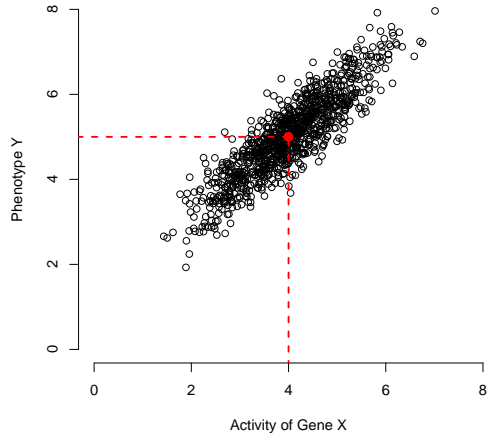
- The level of expression of a particular gene X
- The level of expression of a phenotype Y (e.g. blood insulin levels).



What kind of information can we extract from this data? What tasks can we perform, and what research questions can we answer using statistical techniques?

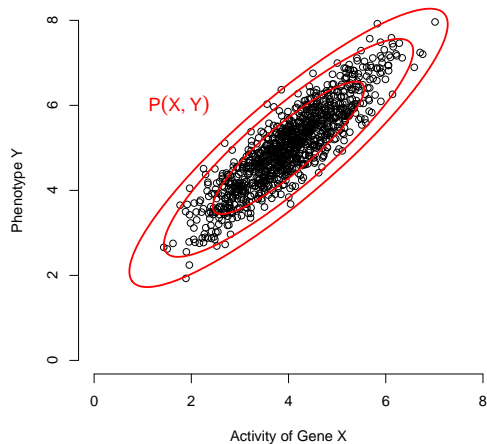
Example

- Description: What is the average level of gene expression in the population?



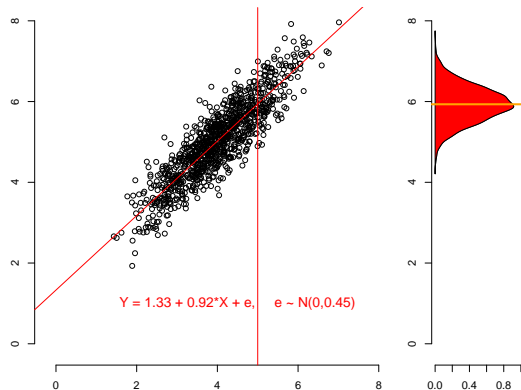
Example

- Description: What is the average level of gene expression in the population?
- Models for co-occurrence: What is the probability of observing high gene expression and high insulin levels?



Example

- Description: What is the average level of gene expression in the population?
- Co-occurrence: What is the probability of observing high gene expression and high insulin levels?
- Prediction: If I observe a gene expression score of 5, what is my best guess of what phenotype level that person has?



What kinds of questions can we **not answer**?

What kinds of questions can we **not answer**?

What if instead of just observing genes and phenotypes, I was to *manipulate* / intervene on / *change* the expression of that gene.

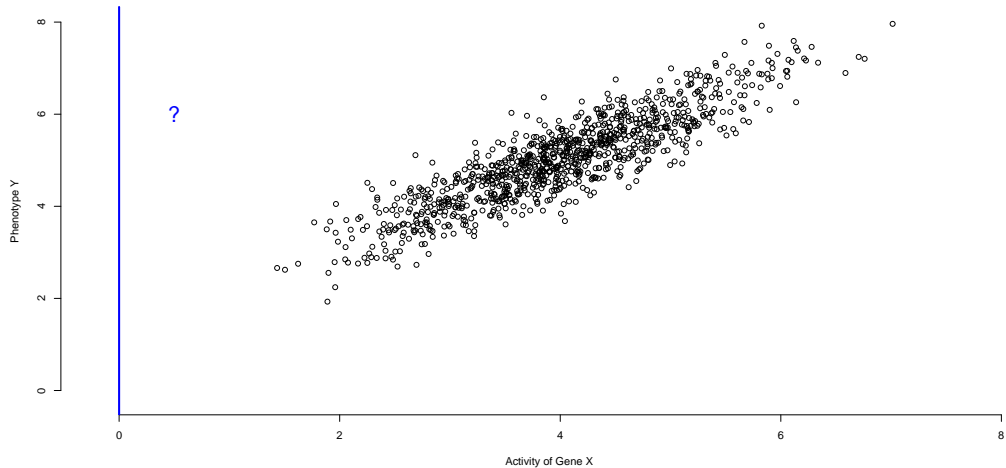
- E.g., deactivating or suppressing gene expression entirely.

What level of phenotype expression would I expect to see if I did that?

- Predicting phenotype from gene expression in a different setting: The intervention setting instead of the observation setting

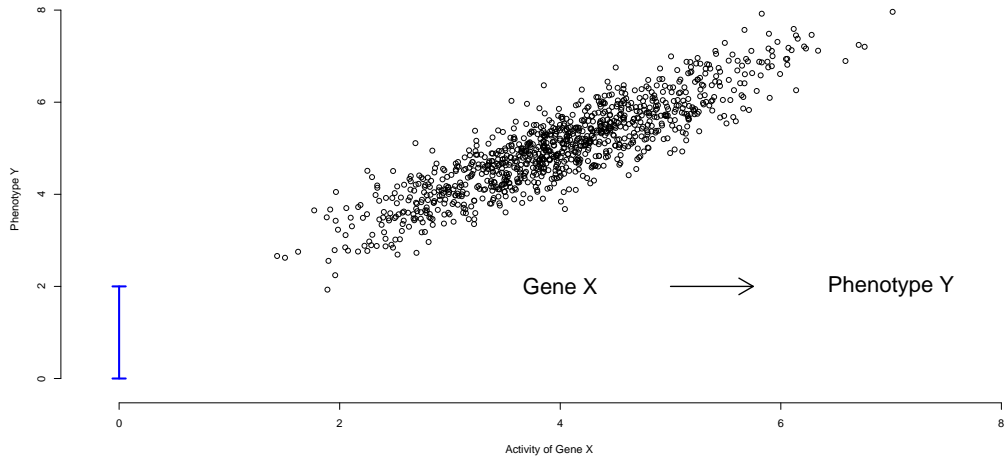
Example: Causal Reasoning^a

^apeters2017elements.



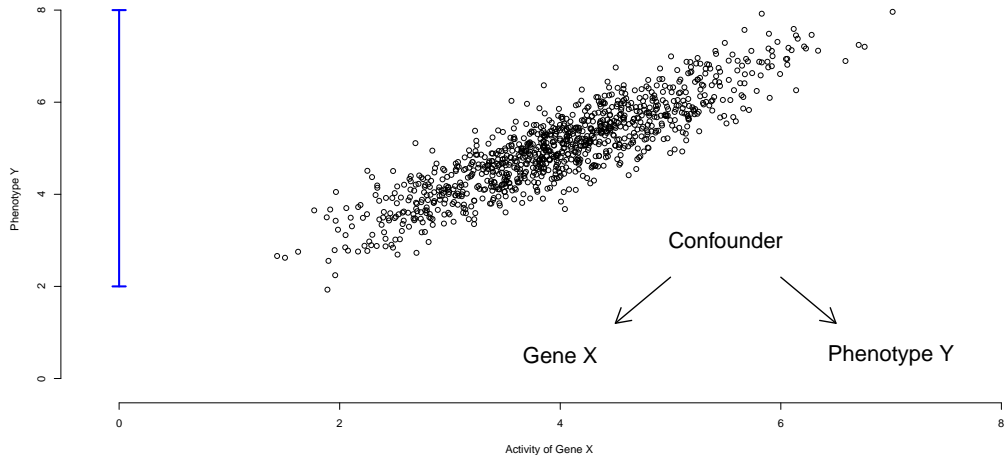
Example: Causal Reasoning^a

^apeters2017elements.



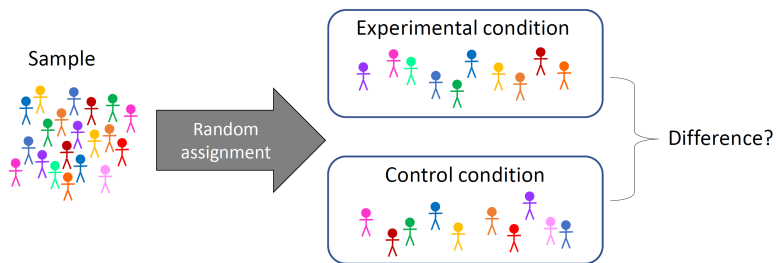
Example: Causal Reasoning^a

^apeters2017elements.



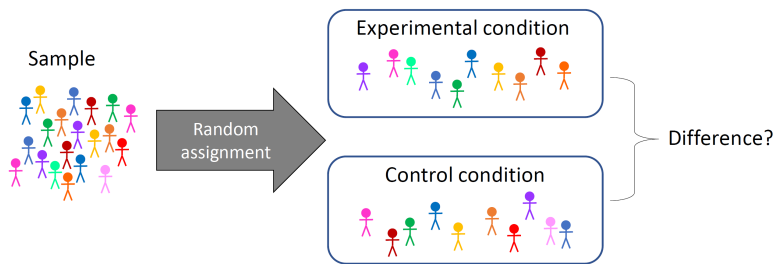
Randomized Control Trials

Randomized Control Trials (RCTs) are the gold standard for estimating causal effects. In principle, we could answer the gene-phenotype question with an RCT



Randomized Control Trials

Randomized Control Trials (RCTs) are the gold standard for estimating causal effects. In principle, we could answer the gene-phenotype question with an RCT





Great! But:

- What if the RCT doesn't work perfectly? What if I have non-compliance?
- What if I can't perform an RCT due to ethical or practical constraints?
Observational / non-experimental data?



Science: A 2x2 table

Target of Inference	Type of Data Available	
	RCT	Everything Else
Causal Relations		
Description / (Limited) Prediction		

Science: A 2x2 table



Target of Inference	Type of Data Available	
	RCT	Everything Else
Causal Relations	 (Very easy) Statistics	???
Description / (Limited) Prediction	?	 Statistics / Statistical Learning

Science: A 2x2 table

Target of Inference	Type of Data Available	
	RCT	Everything Else
Causal Relations	 (Very easy) Statistics	???
Description / (Limited) Prediction	?	 Statistics / Statistical Learning

 You
(most science)
are
here

Science: A 2x2 table

Target of Inference	Type of Data Available	
	RCT	Everything Else
Causal Relations	 (Very easy) Statistics	Causal Modeling
Description / (Limited) Prediction	?	 Statistics / Statistical Learning

 You
(most science)
are
here

Today: Causal Inference.

- Given our beliefs about the system we are studying (i.e., our causal model), can we estimate a causal effect from data? If so, how?

December: Causal Discovery

- Given data, but no strong beliefs about the causal system, can we learn the causal model?

Two Frameworks for Causal Inference

Potential Outcomes (Part 1).

- Developed by statistician Don Rubin (m)
- Imbens (l) & Angrist (r): Nobel Prize for Economics 2021



Structural Causal Models (Part 2).

- Developed by Judea Pearl, a computer scientist
- “Bayesian Networks”



- ① Why Causal Modeling?
- ② **Potential Outcomes**
 - Hands-on Part 1
- ③ Structural Causal Models
 - Hands-on Part 2
- ④ Discussion

Potential Outcomes

Headaches and Aspirin

- action: Aspirin ($X = 1$) or No Aspirin ($X = 0$)
- outcome: Headache gone ($Y = 1$) or Headache remains ($Y = 0$)

We want to know: Should I take an aspirin?

- I want to take aspirin if my headache level after taking aspirin is different than my headache levels if I don't take aspirin
- Two **potential versions of the outcome** for every person. Outcome if treated ($Y^{X=1}$) and outcome if not treated ($Y^{X=0}$)

A causal effect is defined as the **difference in potential outcomes**

Individual causal effect:

$$ICE_i = Y_i^{X=1} - Y_i^{X=0}$$

Individual causal effect:

$$ICE_i = Y_i^{X=1} - Y_i^{X=0}$$

The fundamental problem of causal inference (Holland, 1986): We can only observe one potential outcome per unit

Individual causal effect:

$$ICE_i = Y_i^{X=1} - Y_i^{X=0}$$

The fundamental problem of causal inference (Holland, 1986): We can only observe one potential outcome per unit

Instead we typically focus on trying to infer the **average causal effect**

Average causal effect:

$$ACE = E[Y_i^{X=1} - Y_i^{X=0}] = E[Y^1] - E[Y^0]$$

Example: Aspirin and Headaches

	Potential outcomes		ICE
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
George	0	0	0
Susan	1	0	1
Tracy	1	1	0
Ken	0	1	-1
Pete	1	0	1
Helen	1	0	1
Kate	0	0	0

Example: Aspirin and Headaches

	Potential outcomes		ICE
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$
Charles	1	1	0
George	0	0	0
Susan	1	0	1
Tracy	1	1	0
Ken	0	1	-1
Pete	1	0	1
Helen	1	0	1
Kate	0	0	0

$$\begin{aligned}ACE &= E[Y^1] - E[Y^0] \\ACE &= 5/8 - 3/8 = 0.25\end{aligned}$$

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

Expected value of recovery **aspirin takers** ($X = 1$): $(0+1+1+0)/4 = 0.5$

Expected value of recovery **aspirin avoiders** ($X = 0$): $(1+1+1+0)/4 = 0.75$

But we only observe one outcome per person

	Unobserved			Observed	
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i
Charles	1	1	0	0	1
George	0	0	0	1	0
Susan	1	0	1	1	1
Tracy	1	1	0	0	1
Ken	0	1	-1	0	1
Pete	1	0	1	1	1
Helen	1	0	1	0	0
Kate	0	0	0	1	0

Expected value of recovery **aspirin takers** ($X = 1$): $(0+1+1+0)/4 = 0.5$

Expected value of recovery **aspirin avoiders** ($X = 0$): $(1+1+1+0)/4 = 0.75$

$$E(Y|X = 1) - E(Y|X = 0) = -0.25$$

Naive conclusion: Aspirin decreases chances of headache relief.

What is the problem with observing?

Observing \neq intervening

$E(Y|X = 1) - E(Y|X = 0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X = 1) \neq E(Y|X = 0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y .

What is the problem with observing?

Observing \neq intervening

$E(Y|X = 1) - E(Y|X = 0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X = 1) \neq E(Y|X = 0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y .

In RCTs, we **can** use $E(Y|X = 1) - E(Y|X = 0)$ as an **estimate** of the *ACE*.

What is the problem with observing?

Observing \neq intervening

$E(Y|X = 1) - E(Y|X = 0)$ is **not the same** as $E(Y^1) - E(Y^0)$

Observing that $E(Y|X = 1) \neq E(Y|X = 0)$ (in words: the average value of headache levels for those who did and did not take aspirin are unequal), does not, in general, **imply a causal effect** of X on Y .

In RCTs, we **can** use $E(Y|X = 1) - E(Y|X = 0)$ as an **estimate** of the *ACE*.

Think about the RCT you would ideally want to conduct.

- Under what conditions (assumptions) can we make the same types of inferences from observational data as we would make from that RCT?

The Two “Tasks” of Causal Inference

Identification

Assuming I have **population-level statistical information** (given these variables but with an infinite sample size), can I infer the causal effect of interest?

What causal assumptions/conditions need to be met?

Estimation

Given that my causal effect is identified, how should I go about estimating this effect from sample data?

Statistical assumptions - functional form, distributions, etc.

Assumption 1: SUTVA

Stable unit treatment value assumption (SUTVA):

The potential outcomes for any unit do not vary with the treatments assigned to other units (i.e., **no interference**), and, for each unit, there are **not different versions of each treatment level** that lead to different potential outcomes.

SUTVA is important when moving from a single unit (with an *ICE*), to multiple units (when we consider the *ACE*).

If there are multiple ways to raise X from 0 to 1, this means:

- there are **multiple treatments**
- these may have **different effects**
- and hence the causal question is **ill-defined**

Examples:

- What is the effect of obesity on health?
- Does physical punishment compromise children's well-being?
- Does alcohol undermine cognitive performance in young adolescents?

To formulate better questions, we should define the **target trial**: The randomized controlled trial we would have done, if it had been possible.

Assumption 2: Exchangeability

At best, **half** of the potential outcomes are **observed**; hence, causal inference is at its core a **missing data problem**.

The critical question is: What is the **missing data mechanism**?

Or: What is the **assignment mechanism**?

If there is a relation between the **assignment mechanism** and the **potential outcomes**, this may bias the estimation of the causal effect.

Exchangeability:

The actual treatment received (X) and the potential outcome given treatment Y^X are independent: $Y^x \perp\!\!\!\perp X$ for all x

This is also known as **unconfoundedness**: The missing potential outcome is missing **completely at random**. Individuals across treatment groups are **exchangeable**

Treatment NOT independent of potential outcomes

In a **non-randomized study**, treatment may depend on person features that also relate to the potential outcomes.

	Unobserved			Observed		Confounder
	Y_i^1	Y_i^0	$Y_i^1 - Y_i^0$	X_i	Y_i	
Charles	1	1	0	0	1	3
George	0	0	0	1	0	9
Susan	1	0	1	1	1	8
Tracy	1	1	0	0	1	5
Ken	0	1	-1	0	1	4
Pete	1	0	1	1	1	2
Helen	1	0	1	0	0	5
Kate	0	0	0	1	0	4

Average headache levels are higher among those who took the aspirin. But, people who took aspirin also scored higher on the covariate **dehydration levels** Z_i .

Conditional Exchangeability

Luckily, we don't need full exchangeability for causal inference. We only need **conditional exchangeability**; conditional on a set of observed covariates, the potential outcomes are independent of treatment assignment.

Conditional exchangeability:

The actual treatment received (X) and the potential outcome given treatment Y^X are independent within certain levels of Z : $Y^x \perp\!\!\!\perp X|Z$

This implies that data are *missing at random* (rather than *missing completely at random*).

Estimation of the *ACE* can proceed **as long as we can properly account for (i.e. condition on) the confounder Z** . But to be able to do this, we need...

Assumption 3: Positivity

There must be **exposed and unexposed participants** at every combination of values of Z in the population under study.

In an **RCT**, positivity is **present by design**.

In a **non-experimental study**, **violations** can be **detected** by:

- making tables of each categorical covariate and treatment (should be no empty cells)
- categorize a continuous covariate and make table (but this depends on number and width of categories)
- considering all combinations of covariates (becomes impossible)

Three Conditions/Assumptions necessary for causal **identification**:

- 1 SUTVA
- 2 (Conditional) Exchangeability
- 3 Positivity

Causal Estimation:

Given our data and causal identification assumptions, how should we estimate the causal effect

Propensity Scores

Propensity Scores are a tool used in the PO framework for causal estimation

Propensity scores (assuming no unobserved confounding):

The probability of exposure/treatment given confounders Z

$$\pi_i = P[X_i = 1|Z_i] = \frac{\exp(Z_i'\phi)}{1+\exp(Z_i'\phi)}$$

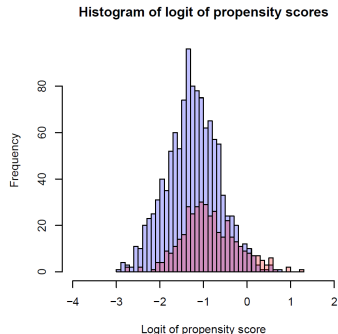
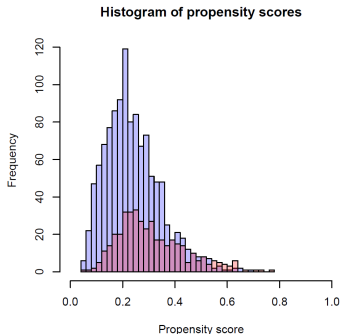
We estimate π_i using **logistic regression**

Propensity scores:

- Summarize information about the relationship between **pre-treatment** confounders (Z) and treatment (X)
- Are used to ensure *conditional exchangeability*

Get $Y^X \perp\!\!\!\perp X|\pi$ to replace $Y^X \perp\!\!\!\perp X|Z$

Overlap of propensity scores



The **distributions of $\text{logit}(\pi_i)$** for the treated and the untreated are typically different, but should fully (and properly) **overlap**:

- non-overlapping areas imply **violation of positivity assumption**
- non-overlapping areas **require extrapolation**
- areas with very few people in one groups imply there are **few matches**

Propensity Scores for Matching

Matching implies you **create pairs** that consist of a treated and a non-treated person, who have **identical propensity scores**.

Background: In an **RCT** we have: $P(Z|X = 1) = P(Z|X = 0)$

Balancing property:

$$P(Z|\pi = c, X = 1) = P(Z|\pi = c, X = 0)$$

If the propensity model is **correct**, then comparing treated and untreated **individuals with the same π** is a way of **mimicking an RCT**.

The **probability of received treatment** is:

- π_i for those who were **treated** ($X_i = 1$)
- $1 - \pi_i$ for those who were **NOT treated** ($X_i = 0$)

Among **treated individuals**, those with large π_i are **overrepresented** in comparison to those with small π_i .

Among **untreated individuals**, those with large $1 - \pi_i$ are **overrepresented** in comparison to those with small $1 - \pi_i$.

To account for this imbalance, we **create a pseudo-population** where each case is **weighted** by the **inverse probability of received treatment**:

- $\frac{1}{\hat{\pi}_i}$ for $X_i = 1$
- $\frac{1}{1 - \hat{\pi}_i}$ for $X_i = 0$

Causal Inference is a missing data problem

- When can I infer $E[Y^1] - E[Y^0]$ if i don't fully observe either?
- I can do this with an RCT, so how do I mimic an RCT with observational data?

Steps (broadly):

- Identify the target trial. Measure potential confounders
- Assess SUTVA, Exchangeability and Positivity
- If you can meet those conditions, use covariate-based techniques like propensity scores to create balanced groups of treated and not treated, mimicing an RCT
- Estimate ACE by adjusting for group differences on confounders (e.g., weighting, matching)

- ① Why Causal Modeling?
- ② Potential Outcomes
 - **Hands-on Part 1**
- ③ Structural Causal Models
 - Hands-on Part 2
- ④ Discussion

Structural Causal Models

Why Two Approaches?

Different but highly related approaches.

- Two different “hats”
- Rubin model: Individual observations \rightarrow individual causal effects \rightarrow average causal effect
- SCMs: Causal relations and manipulations of *variables*
- Graphical representations of causal structure
- Advantage of SCMs: Easier to deal with many different variables and many causal relations at the same time

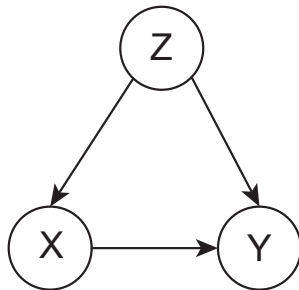
Causal Graphs: Draw Your Assumptions before your Conclusions

A causal graph is a diagram representing our beliefs about which variables share causal relations with each other

Causal Graphs: Draw Your Assumptions before your Conclusions

A causal graph is a diagram representing our beliefs about which variables share causal relations with each other

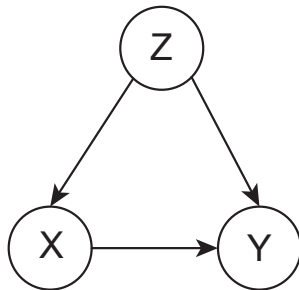
- The arrow $X \rightarrow Y$ represents our belief that X is a direct cause of Y
- We omit an arrow if expert knowledge tells us that one variable does not directly cause another. The *absence* of an arrow is a strong statement



Causal Graphs: Draw Your Assumptions before your Conclusions

A causal graph is a diagram representing our beliefs about which variables share causal relations with each other

- The arrow $X \rightarrow Y$ represents our belief that X is a direct cause of Y
- We omit an arrow if expert knowledge tells us that one variable does not directly cause another. The *absence* of an arrow is a strong statement



Graph known as a **Directed Acyclic Graph** (DAG) or Bayesian Network

We can formalize the idea that the arrows in the DAG represent our beliefs about causal relations by saying that the DAG visualizes a Structural Causal Model (SCM)

An SCM is a set of equations describing causal relations between variables, which are also influenced by independent noise terms N (typically not drawn in the graph).

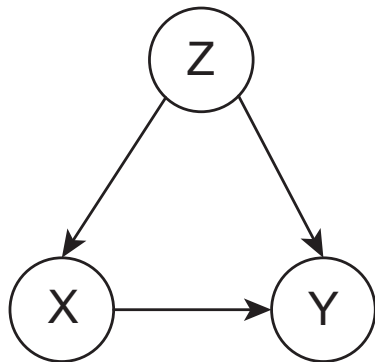
$$X := f_x(N_x)$$

$$Z := f_m(X, N_z)$$

$$Y := f_y(X, Z, N_y)$$

where

- N_i are jointly independent
- f represents potentially any function - any distribution, any type of functional form



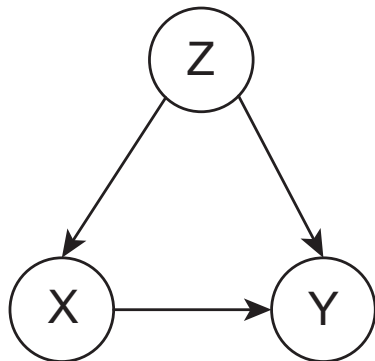
$$X := \epsilon_X$$

$$Z := 2X + \epsilon_Z$$

$$Y := 1X + 2Z + \epsilon_Y$$

where

- $\epsilon_X, \epsilon_Z, \epsilon_Y$ are iid, $\sim \mathcal{N}(0, 1)$



The do-operator $do(X = x)$ represents a “surgical intervention” to set the value of the variable X to a constant value x

The do-operator $do(X = x)$ represents a “surgical intervention” to set the value of the variable X to a constant value x

Often interested in the causal effect of a *do*- intervention on the **mean** of another variable

Average causal effect:

$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

The do-operator $do(X = x)$ represents a “surgical intervention” to set the value of the variable X to a constant value x

Often interested in the causal effect of a *do*- intervention on the **mean** of another variable

Average causal effect:

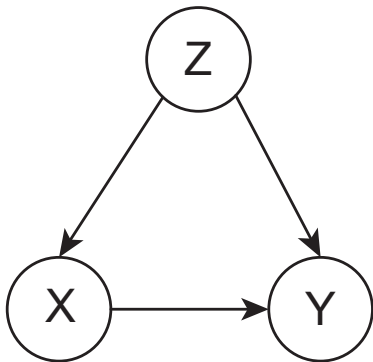
$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

In the graph, a *do*- operation on X cuts-off all incoming ties

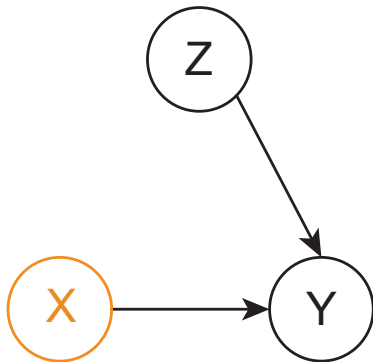
- Intervening makes X independent of other causes

Two versions of the causal system

Observing



Intervening



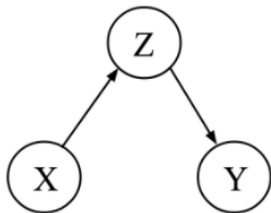
Let's say we have **observational** data. This allows us to learn statistical dependencies in the observational setting

But we want to use observational data to obtain dependencies in the **intervention** setting (i.e., we want to estimate a causal effect)

The structure of the DAG tells us how to do this!

3 fundamental graphical structures

Chain

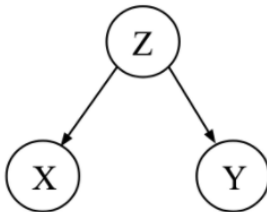


X: Smoking
Z: Tar
Y: Cancer

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

Fork

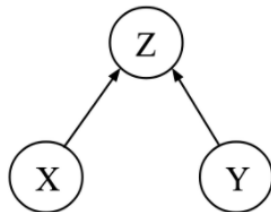


X: Storks
Z: Environment
Y: Babies

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

Collider



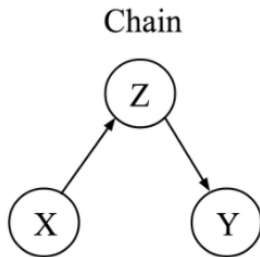
X: Attractiveness
Z: Being Single
Y: Intelligence

$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \mid Z$

Chains transmit causal associations

- X changes Z which in turn changes Y
- aka mediation
- Conditioning on Z (i.e. controlling for Z) blocks transmission of causal information



X: Smoking

Z: Tar

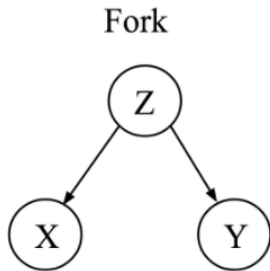
Y: Cancer

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

Forks transmit non-causal (statistical) information

- Z causes X and Y , which makes X and Y statistically dependent
- But intervening on X doesn't change Y
- aka confounding or common-cause variables
- This is known as a **backdoor path**
- Conditioning on Z (i.e. controlling for Z) blocks transmission of non-causal information



X: Storks
Z: Environment
Y: Babies

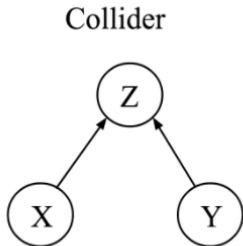
$$X \not\perp\!\!\!\perp Y$$
$$X \perp\!\!\!\perp Y \mid Z$$

Colliders

Colliders do not transmit any information

- X and Y are uncorrelated, but both cause Z
- aka a common effect
- But, conditioning on Z *introduces* a non-causal (spurious) association between X and Y
- This is known as collider bias

Implication: Estimating a causal effect by controlling for **everything** is a **terrible idea**

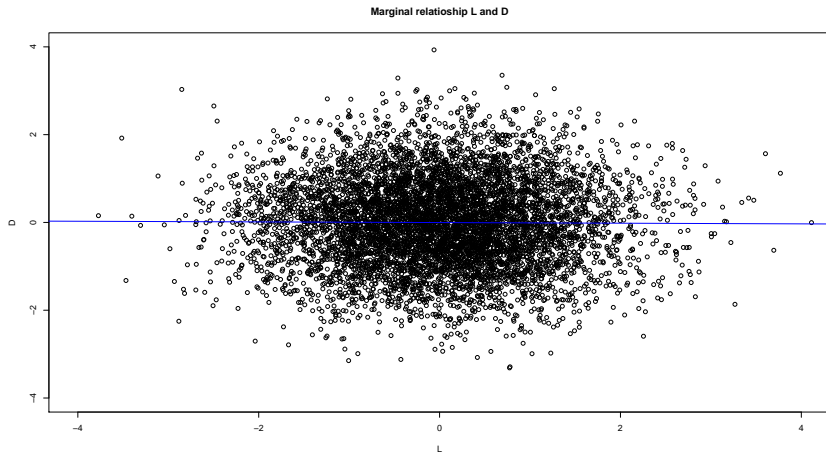


X: Attractiveness
Z: Being Single
Y: Intelligence

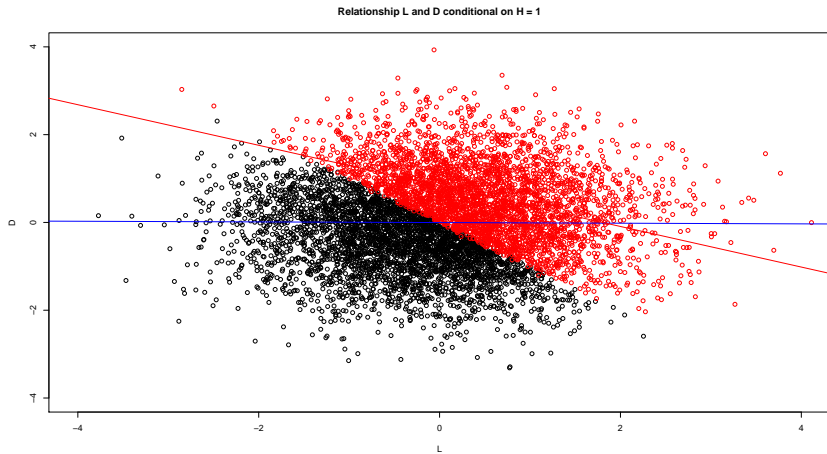
$$X \perp\!\!\!\perp Y$$

X	Y	Z
1	0	0
1	0	1
1	1	0
1	1	1
0	0	0
0	0	1
0	1	0
0	1	1

Collider Bias



Collider Bias



In a nutshell: To estimate the causal effect of X on Y we

- block *backdoor paths* by conditioning on confounders
- avoid conditioning on any colliders or mediators

In a nutshell: To estimate the causal effect of X on Y we

- block *backdoor paths* by conditioning on confounders
- avoid conditioning on any colliders or mediators

We want to estimate the causal effect of *Aspirin* (X) on *Headache Recovery* (Y). We have also measured whether participants were *Dehydrated* (Z) prior to taking aspirin.

Target of Inference:

$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

Target of Inference:

$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

Naive Estimate:

$$E[Y|X = 1] - E[Y|X = 0]$$

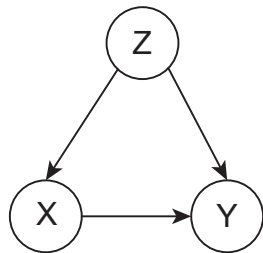
Example: Aspirin and Headaches

Target of Inference:

$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

Naive Estimate:

$$E[Y|X = 1] - E[Y|X = 0]$$



Example: Aspirin and Headaches

Target of Inference:

$$ACE = E[Y \mid do(X = 1)] - E[Y \mid do(X = 0)]$$

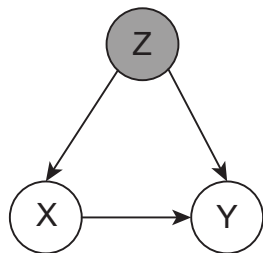
Naive Estimate:

$$E[Y \mid X = 1] - E[Y \mid X = 0]$$

Seeing \neq Doing: Unblocked backdoor through Z

Correct Estimate:

$$E[Y \mid X = 1, Z] - E[Y \mid X = 0, Z]$$



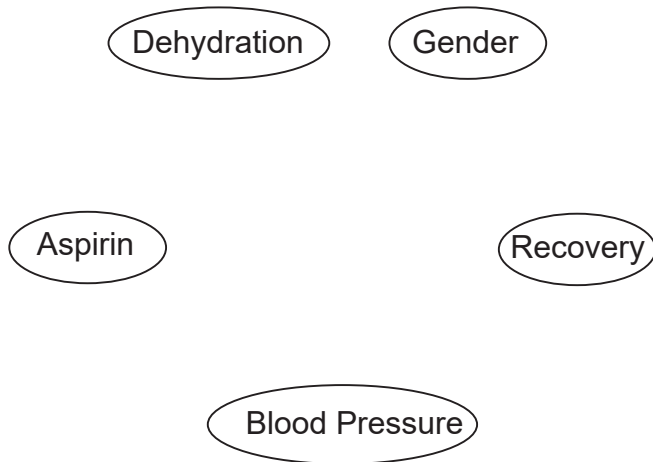
The DAG tells us which variables to condition on, and which variables not to condition on, to estimate a causal effect

Valid Adjustment Set

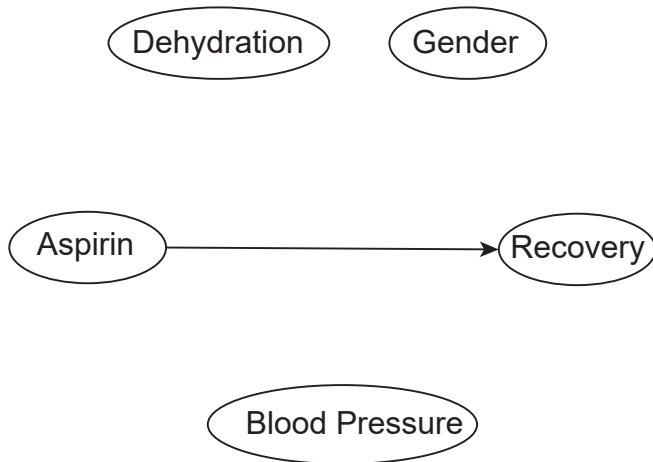
A set of variables that we can adjust for in order to estimate the effect of an intervention (e.g. from observational data)

By drawing bigger DAGs, and including observed and unobserved variables, we can assess if and how the causal effect of interest is **identified**

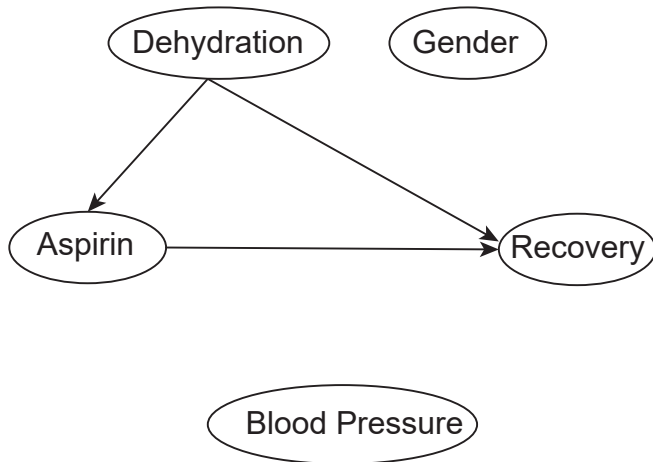
Benefits of a DAG approach



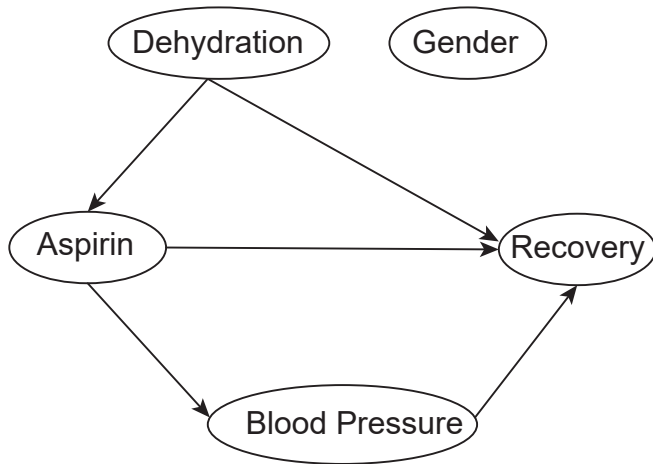
Benefits of a DAG approach



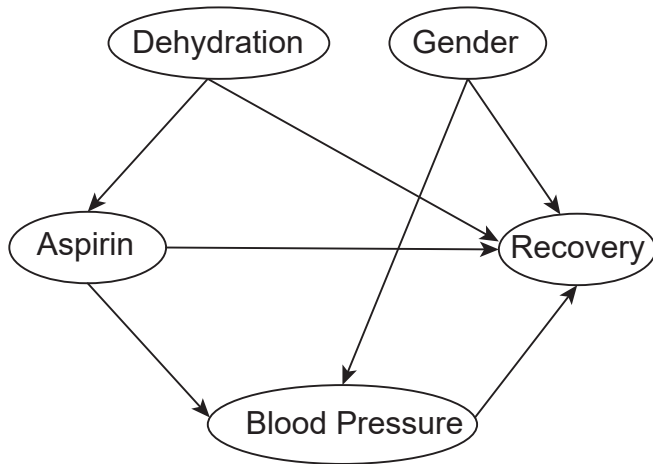
Benefits of a DAG approach

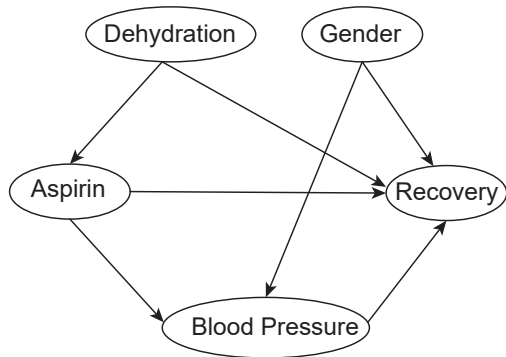


Benefits of a DAG approach



Benefits of a DAG approach

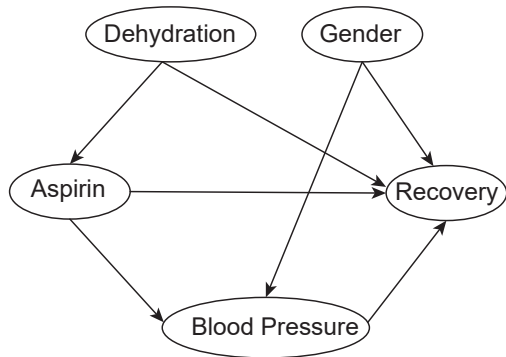




To estimate ACE of Aspirin on Recovery
Condition on *Dehydration* only

- Not necessary to condition on gender

Estimating Causal Effects

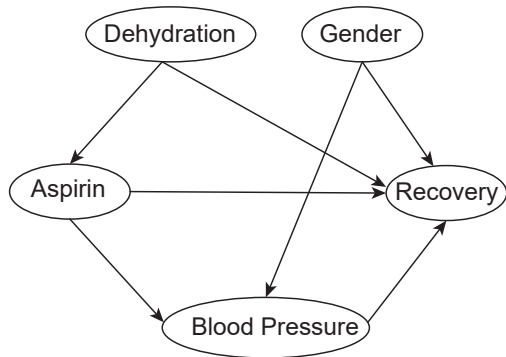


To estimate ACE of Aspirin on Recovery
Condition on *Dehydration* only

- Not necessary to condition on gender
- Conditioning on BP blocks a causal path, and opens a collider path

$$A - G \rightarrow R$$

Estimating Causal Effects

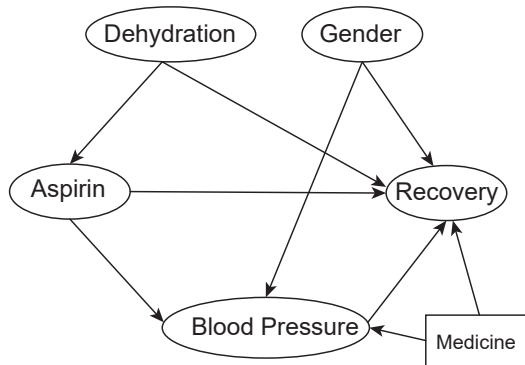


To estimate ACE of Aspirin on Recovery
Condition on *Dehydration* only

- Not necessary to condition on gender
- Conditioning on BP blocks a causal path, and opens a collider path
 $A - G \rightarrow R$

Unobserved confounders should also be included in your DAG to determine if a causal effect is identified

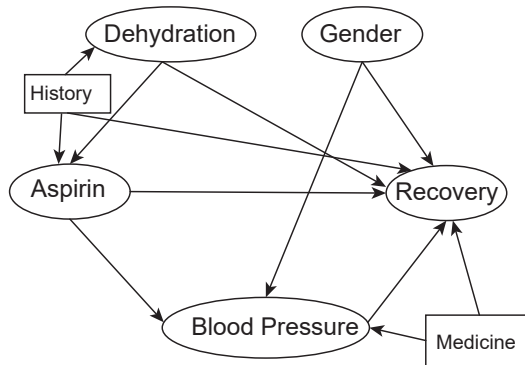
Unobserved Confounders



Anti-inflammatory *Medicine* reduces blood pressure and recovery

- But the ACE is still identified, even without observing this variable

Unobserved Confounders



If someone has a *History* of migraines, this may effect A and R

- The ACE is not identified from the observed data in this case
- But, we can perform *sensitivity analyses*: how strong would this effect need to lead to a different conclusion about the causal effect?

Drawing DAGs gives you practical guidelines about when causal inferences can be made, what variables to control for, and what variables not to control for

- Transparent way of representing *your/expert beliefs* about the causal system at hand
- These beliefs guide statistical analyses in a straightforward way

Benefits of DAGs: Conceptual Clarity

Drawing DAGs gives you practical guidelines about when causal inferences can be made, what variables to control for, and what variables not to control for

- Transparent way of representing *your/expert beliefs* about the causal system at hand
- These beliefs guide statistical analyses in a straightforward way

Emphasis here on *identification* rather than *estimation*: You still need to choose *how* to condition on variables!

Benefits of DAGs: Conceptual Clarity

Drawing DAGs gives you practical guidelines about when causal inferences can be made, what variables to control for, and what variables not to control for

- Transparent way of representing *your/expert beliefs* about the causal system at hand
- These beliefs guide statistical analyses in a straightforward way

Emphasis here on *identification* rather than *estimation*: You still need to choose *how* to condition on variables!

Many controversial and seemingly difficult problems are made easy by drawing DAGs

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Example (Pearl, Glymour & Jewell, 2016):

- 700 sick patients are given the choice to take a new drug: 350 choose to take it.
- We are interested in effects of a drug (D) on recovery (R). We also record the gender (G)
- Should we prescribe the drug?

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Simpsons Paradox

- Estrogen levels negatively affect recovery
- Women are more likely to take the drug than men
- We should condition on Gender - it blocks a backdoor path!

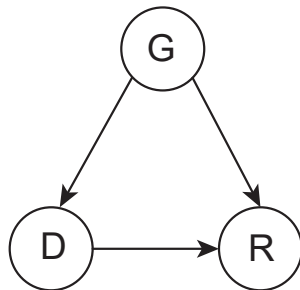


Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Simpsons Paradox

Suppose that we measure post-treatment blood pressure (B) instead

- Statistical information is exactly the same
- B cannot cause drug taking
- The drug works in part by decreasing blood pressure
- We should **not** condition on blood pressure

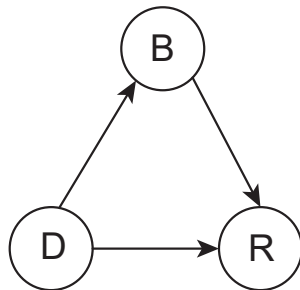
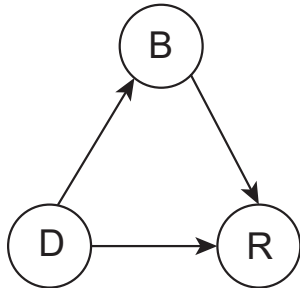
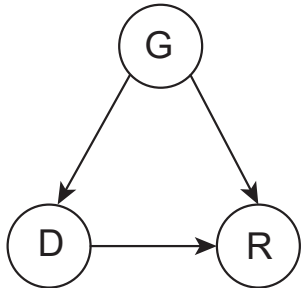


Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	Drug	No drug
Low BP	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High BP	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Simpsons Paradox

- Statistical information alone cannot provide the answer
- Two different DAGs can produce the exact same statistical dependencies in the observational setting
 - Observationally equivalent
- These DAGs imply different intervention effects, and different ways to estimate those effects from observational data!



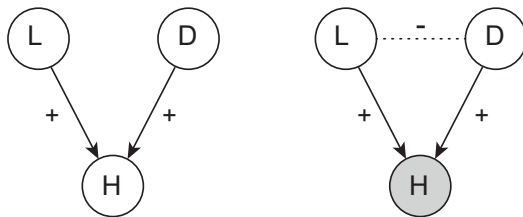
Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

Classic example: We are interested in the relationship between *Lung Cancer* (L) and *Diabetes* (D)

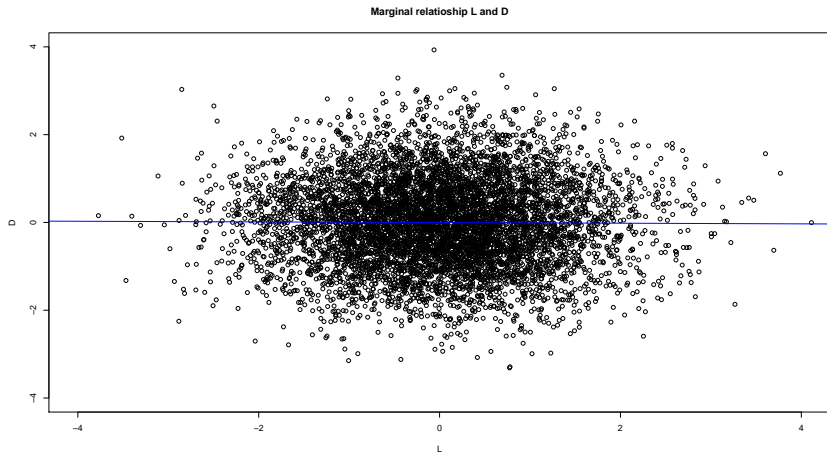
- General population, these two variables are independent.
- In a sample of *hospital patients*, there is a negative dependency - patients who don't have diabetes are *more likely* to have lung cancer.

Selection Bias

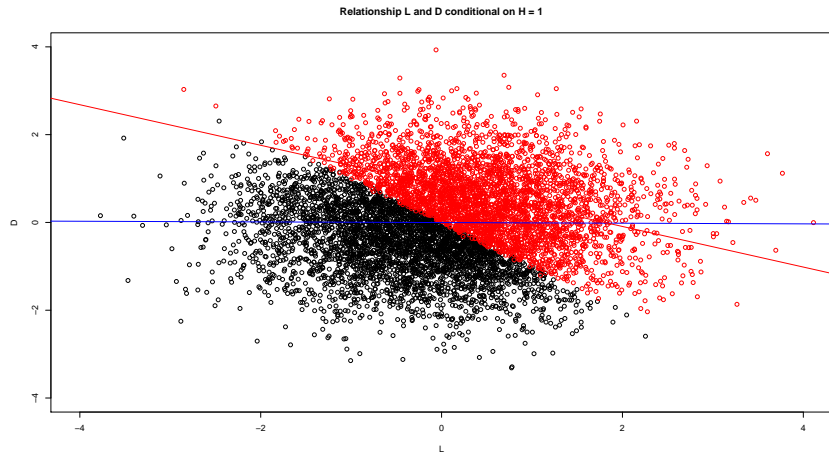


- Lung cancer L and diabetes D cause hospitalization H
- By taking participants from a hospital we *condition* on hospitalization ($H = 1$)
- If you are hospitalised, and you *don't* have diabetes, probably you do have lung cancer (Otherwise - why would you be in hospital?).
- $P(D|L = 1, H = 1) \neq P(D|L = 1) \neq P(D|do(L) = 1)$
- We have conditioned on a *collider*

Collider Bias



Collider Bias



Causal Inference using DAGs comes with assumptions:

Modularity and Localized Interventions:

We assume that it is possible to intervene on a variable without fundamentally changing how it relates to other variables

- We can change $p(X)$ without changing $p(Z | X)$
- We can change one cause-effect mechanism without changing the others

The better the idea we have about the DAG, the better our inferences become

- No free lunch!

Causal Inference from an SCM perspective

Causal Inference is the problem of making inferences about the **interventional** density of our variables using the **observational density**

Steps (broadly):

- Specify your causal target of inference
- Draw the DAG of your causal system. Include any observed OR unobserved variables that relate to *at least two* variables in your causal system
- Find *valid adjustment sets*: what variables do you need to condition on to block backdoor paths
- Decide if your causal effect is *identified*: Can you block all backdoor paths using only *observed variables*?
- Estimate causal effect by conditioning appropriately* on those variables

- ① Why Causal Modeling?
- ② Potential Outcomes
 - Hands-on Part 1
- ③ Structural Causal Models
 - **Hands-on Part 2**
- ④ Discussion

Potential Outcomes

- Causal effects as Target Trial
- Emulating RCT where $X - Y$ effect is only thing of interest
- View on covariates: Use only **pre-treatment**, throw everything into propensity score
- Emphasis on estimation tools

Structural Causal Models

- Causal effects as variable relationships
- Intervention density
- Much more detailed view of “covariates” - distinguishing multivariate systems of causal effects
- Emphasis on identification
- Things like mediation, direct/indirect effects can be defined more easily

Agree on most things, just a different perspective/emphasis/level of abstraction

- Non-parametric estimation: Use only causal assumptions, as little statistical assumptions as possible!
- Utilizing ML techniques for conditioning
- Sensitivity analysis
- Instrumental variables
- Transportability / Generalizability
- **Next Week:** Discovering the causal graph from data

Recommended Reading

Hernán, M. A. (2018). The C-word: scientific euphemisms do not improve causal inference from observational data. *American journal of public health*, 108(5), 616-619.

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 1(1), 27-42.

Schafer, J. L., Kang, J. (2008). Average causal effects from nonrandomized studies: a practical guide and simulated example. *Psychological methods*, 13(4), 279.

Hernán MA, Robins JM (2020). Causal Inference: What If. Boca Raton: Chapman Hall/CRC. Free copy: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2021/03/ciwhatif_hernanrobins_30mar21.pdf

Pearl, J., Glymour, M. & Jewell, N.P. (2016) Causal Inference in Statistics: A Primer

Materials of today in part adapted from the course **Social Behaviour Dynamics** from the ADS academic master

- With thanks to Ellen Hamaker!

See you in december!