Day 1: Introduction to causal inference

Chris Holmes and Andrew Yiu

Department of Statistics, University of Oxford

Leuven, September 2024

DEPARTMENT OF
STATISTICS

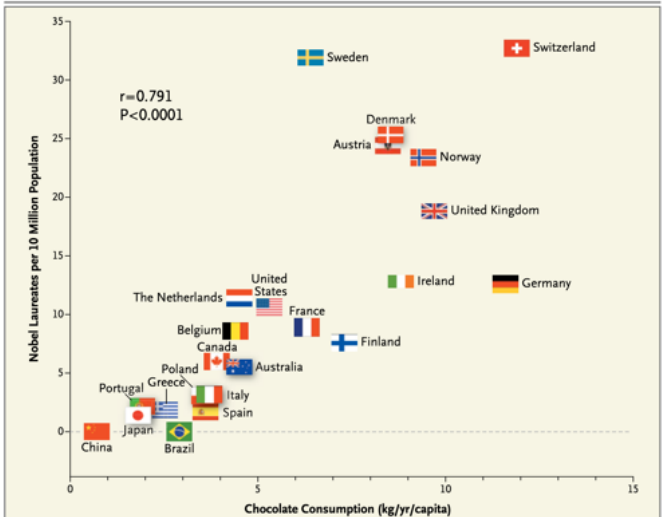


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Confectionery news.com

Eating chocolate produces Nobel prize winners, says study

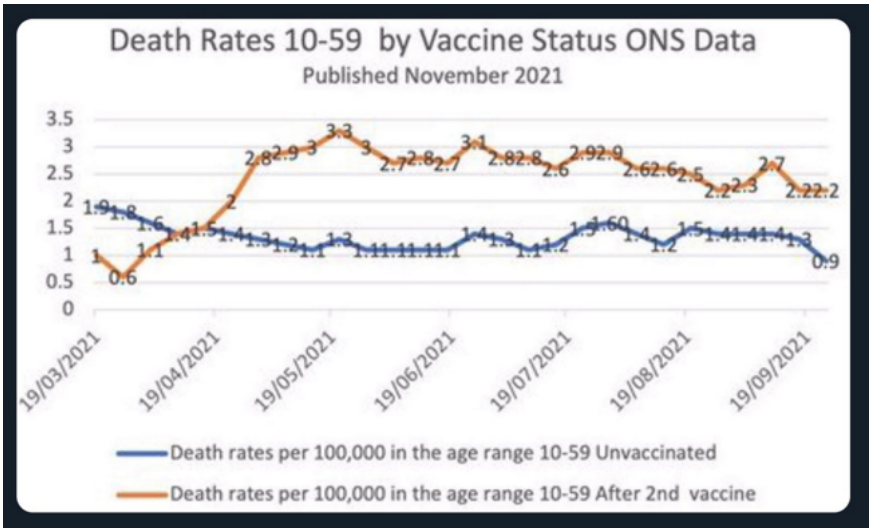
By Oliver Nieburg

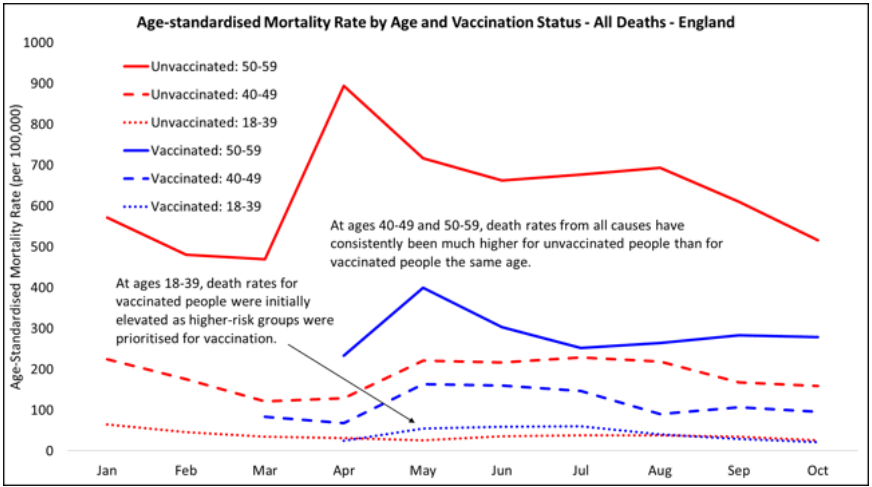
10-Oct-2012 - Last updated on 11-Oct-2012 at 11:51 GMT

"Association is not causation."

Examples of associational concepts: regression, classification, conditional independence, likelihood, p-value, confidence interval, root mean squared error, standard deviation, sufficient statistic, hazard ratio...

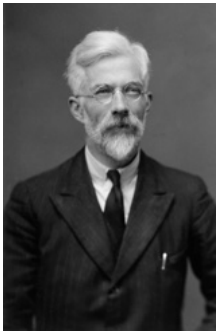
But are we ever interested in association without causation?


Death Rates 10-59 by Vaccine Status ONS Data
Published November 2021
3.5
3
2.5
2
1.5
1
0.5
0
19/03/2021
19/04/2021
19/05/2021
19/06/2021
19/07/2021
19/08/2021
19/09/2021
Death rates per 100,000 in the age range 10-59 Unvaccinated
Death rates per 100,000 in the age range 10-59 After 2nd vaccine


Age-standardised Mortality Rate by Age and Vaccination Status - All Deaths - England
Age-Standardised Mortality Rate (per 100,000)
1000
900
800
700
600
500
400
300
200
100
0
Unvaccinated: 50-59
Unvaccinated: 40-49
Unvaccinated: 18-39
Vaccinated: 50-59
Vaccinated: 40-49
Vaccinated: 18-39
At ages 40-49 and 50-59, death rates from all causes have consistently been much higher for unvaccinated people than for vaccinated people the same age.
At ages 18-39, death rates for vaccinated people were initially elevated as higher-risk groups were prioritised for vaccination.
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct

Randomized experiments

Fortunately, there's a method that guarantees association to be causation: conduct a randomized experiment.



Ronald Fisher

Rothamsted Research

The lady tasting tea

Muriel Bristol

At teabreak one day at Rothamsted, Fisher offered a cup of tea to a female colleague. She declined, stating that she preferred it when the milk had been poured first.

Fisher was incredulous, thinking that there was surely no difference in taste.

William Roach—a biochemist working in the antiseptics and insecticides lab—overheard this conversation and said: "Let's test her."

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

"The Design of Experiments" (1935) – R.A. Fisher

Fisher's null

> A subject without any faculty of discrimination would in fact divide the 8 cups correctly into two sets of 4 in one trial out of 70, or, more properly, with a frequency which would approach 1 in 70 more and more nearly the more often the test were repeated. Evidently this

Null hypothesis: the lady has no "faculty of discrimination" between the order of milk and tea. In other words, the treatment does not influence her choice of grouping.

There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ ways of choosing a group of 4 objects from 8. The randomization makes each of the 70 possible orderings equally likely. This defines our probability space.

Fisher's test

Fisher argued that we can treat the lady's choice of grouping as being non-random under the null, because the only source of randomness arises from the treatment assignment, and the treatment has no influence on her choice of grouping by definition of the null.

So the probability that she guesses all cups correctly is 1/70 under the null (i.e. the treatment assignment just happens to perfectly align with her grouping).

- Why did Fisher choose 8 cups? With 6 cups, there are $\binom{6}{3} = 20$ ways of choosing a group of 3 cups. Fisher wanted to set a significance level of 5%, and $(1/70) \approx 1.4\% < 5\%$.
- According to Fisher's biography, Muriel Bristol did guess all cups correctly. She also got married to William Roach.

The importance of randomization

"Whole forests have been destroyed to provide paper for disputes about Fisher's analysis." Stephen Senn

Some statisticians were sceptical of the idea of physically introducing randomization into the experiment. Fisher treated experimentation as a game played against an adversary (e.g. the laws of nature). If you randomize, you guarantee that your comparison groups are indeed comparable on average, no matter how the adversary tries to conspire against you.

"...having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity." Bradford Hill

Randomized experiments

So we can draw causal conclusions from randomized experiments. But randomized experiments are not always feasible:

- Ethical concerns
- Too expensive
- Impractical

To this day, many researchers believe that causal inference is impossible unless we have a randomized experiment.

"*Dear author: Your observational study cannot prove causation. Please replace all references to causal effects by references to associations.*" from "The C-Word" (2018) - Miguel Hernán

Smoking and lung cancer

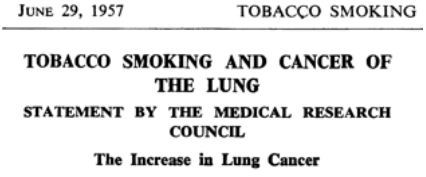
Between 1922 and 1947, the number of deaths attributed to lung cancer increased 15-fold across England and Wales.

In the 1950's, it was estimated that about 80% of adult men and 40% of adult women in the UK were smokers.

Smoking and lung cancer

In 1950, the epidemiologist Bradford Hill and the physician Richard Doll published a study in the British Medical Journal showing an extremely strong association between smoking and lung cancer. Doll gave up smoking two-thirds of the way through the study.

By 1957, the evidence linking smoking with lung cancer was overwhelming enough for the Medical Research Council to issue a statement. It contained: *"...the most reasonable interpretation of this evidence is that the relationship is one of direct cause and effect."*



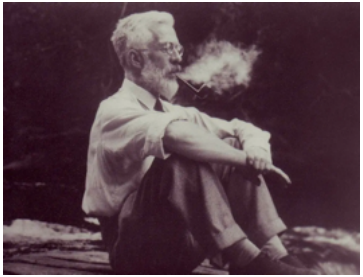
JUNE 29, 1957 TOBACCO SMOKING

TOBACCO SMOKING AND CANCER OF THE LUNG

STATEMENT BY THE MEDICAL RESEARCH COUNCIL

The Increase in Lung Cancer

Fisher's response



Fisher said this was **"terrorist propaganda exhorting people to stop smoking".**

In a series of letters to the British Medical Journal, he provided evidence that there could be a gene that had the potential to both increase the propensity to smoke and the propensity to develop lung cancer.

He also suggested that the development of acute lung cancer could be preceded by an undiagnosed "chronic inflammation" that led to mild discomfort, which could lead the subject to smoke cigarettes for relief.

Aftermath

We now "know" that smoking does indeed cause lung cancer. The UK adult smoking rate is below 15%, and the number of annual deaths from lung cancer in 2000 was half of what it was in 1965.

Fisher's logic was correct; indeed, the whole point of randomization is to eliminate possible confounding effects from our analysis. But the magnitude of those effects couldn't possibly explain away the observed associations between smoking and lung cancer.

This example illustrates that we cannot restrict to ourselves to randomized experiments for the purpose of doing causal inference.

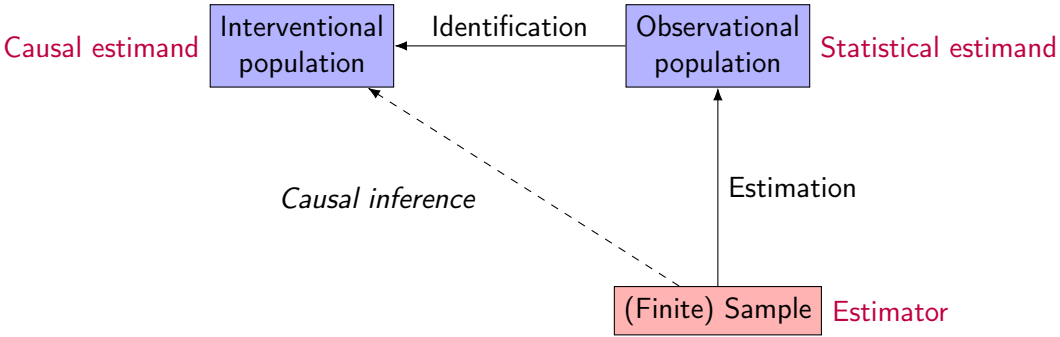
Causal inference with observational data

General strategy

1. Determine a **causal quantity** that will answer the scientific question of interest;
2. Check that the quantity is **identifiable** from the data you have and assumptions you are willing to make;
3. Perform **statistical inference** for the identified quantity based on your data and assumptions.

For much of today, we will focus on the second objective: identification. Once we move away from randomized experiments, it is no longer guaranteed that we can distinguish between causation and (just) association.

The "roadmap"



If we have randomized, experimental data, we can go straight from the data to the interventional population. But this is infeasible otherwise.

Frameworks for identification

There are many frameworks for doing causal identification. The ones we will cover are:

- Rubin's Causal Model (potential outcomes)
- Pearl's Structural Causal Model (including Pearl's causal graphs)
- Frameworks without counterfactuals (Dawid's decision-theoretic framework, target trial emulation)

As we will see, the motivations and utilities of these frameworks are very similar. Our main objectives:

- Develop some familiarity with the basic ideas and terminology of each framework.
- Gain an understanding of when one framework might be preferable to others.

The Rubin Causal Model

Donald Rubin

In a series of papers in the 1970's, Don Rubin introduced a causal framework for the analysis of observational data using potential outcomes.

His philosophy was that we should use observational data to approximate (or: "reconstruct" or "emulate") a true randomized experiment as closely as possible.

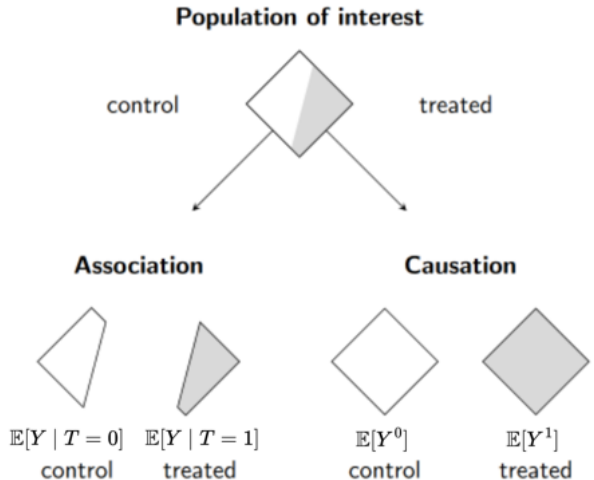
Set-up

Suppose we have data $Z_i = (Y_i, T_i, X_i)$ from an observational study:

- Y_i is the outcome variable of interest (e.g. all-cause mortality)
- T_i is a binary variable indicating treatment assignment (e.g. $T_i = 1$ if subject i receives vaccination).
- X_i is a vector of measured pre-treatment covariates (e.g. age). For expositional simplicity, we will assume throughout that the covariates are discrete.

For each unit, we also define a pair of variables (Y_i^1, Y_i^0): the potential outcome Y_i^t is the outcome that would be observed if—possibly contrary to fact—the subject receives treatment t.

Or in other words, Y_i^t is the outcome under an interventional regime where everybody takes treatment t (instead of being assigned treatment observationally). Everything aside from the treatment assignment is unchanged.



From "Causal Inference: What If." – Hernán & Robins (2020)

SUTVA

The potential outcome variables (Y_i^1, Y_i^0) are formally linked to the observed outcomes by the following:

$$T_i = t \implies Y_i = Y_i^t. \tag{1}$$

Implicit in this formulation is the stable unit treatment value assumption (SUTVA), which has two components:

1. There is no interference; that is, it is assumed that a unit's potential outcomes are unaffected by the treatments assigned to other units. (counterexample: vaccinations)
2. The second is that there are no hidden variations of treatments: $Y_i^{(t)}$ does not depend on how unit i received treatment t, e.g. the hospital that a patient visited.

In epidemiology, equation (1) is often called consistency (see, for example, Hernán & Robins (2020)).

The fundamental problem

Under SUTVA, we can frame causal inference as a missing data problem:

Unit	Y^0	Y^1	T	X
1	?	Y_1^1	1	X_1
2	Y_2^0	?	0	X_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	?	Y_n^1	1	X_n

From Rubin's perspective, both (Y_i^1, Y_i^0) exist; it's just that we observe at most one of them. This is often referred to as the "fundamental problem of causal inference".

He compares this to the Heisenberg Uncertainty Principle in quantum mechanics, in which both the position and momentum of a particular are well-defined but it is impossible to measure both precisely.

Potential outcomes

Potential outcomes were introduced by Jerzy Neyman in his 1923 PhD thesis (in the context of randomized experiments). So the Rubin causal model is sometimes called the Neyman-Rubin causal model instead.

Jerzy Neyman

Famously, Neyman and Fisher hated each other.

DISCUSSION ON DR. NEYMAN'S PAPER.

PROFESSOR R. A. FISHER, in opening the discussion, said he had hoped that Dr. Neyman's paper would be on a subject with which the author was fully acquainted, and on which he could speak with authority, as in the case of his address to the Society delivered last summer. Since seeing the paper, he had come to the conclusion that Dr. Neyman had been somewhat unwise in his choice of topics.

From *Discussion of "Statistical problems in agricultural experimentation"*, Royal Statistical Society Meeting (1935)

Fisher's test with potential outcomes

Let's try analysing Fisher's tea example using potential outcomes. Recall that the lady tastes 8 cups of tea, 4 of which are made with milk added first and the other 4 are made with tea added first.

	Cup							
	1	2	3	4	5	6	7	8
Y^1	1	?	?	1	?	1	1	?
Y^0	?	0	0	?	0	?	?	0
T	1	0	0	1	0	1	1	0

- $Y = 1$: she guesses milk first; $Y = 0$ otherwise.
- $T = 1$: the tea was made with milk first; $T = 0$ otherwise.

Fisher's test with potential outcomes

	Cup							
	1	2	3	4	5	6	7	8
Y^1	1	0	0	1	0	1	1	0
Y^0	1	0	0	1	0	1	1	0
T	1	0	0	1	0	1	1	0

Under the null hypothesis *"no faculty of discrimination"*

$$H_0 : Y_i^1 = Y_i^0 \text{ for } i = 1, \ldots 8,$$

we can fill in the missing data. The outcomes are treated as fixed; the only randomness comes from the randomization in T.

Fisher's test with potential outcomes

Under the null, what is the probability that she guesses all cups correctly?

Test statistic:

$$S = \#\text{correct guesses} = \sum_{i=1}^{8} 1(Y_i = T_i)$$

Probability of guessing all cups correctly:

$$\mathbb{P}(\text{all correct}) = \mathbb{P}(S = 8) = \frac{1}{\binom{8}{4}}.$$

Superpopulation inference

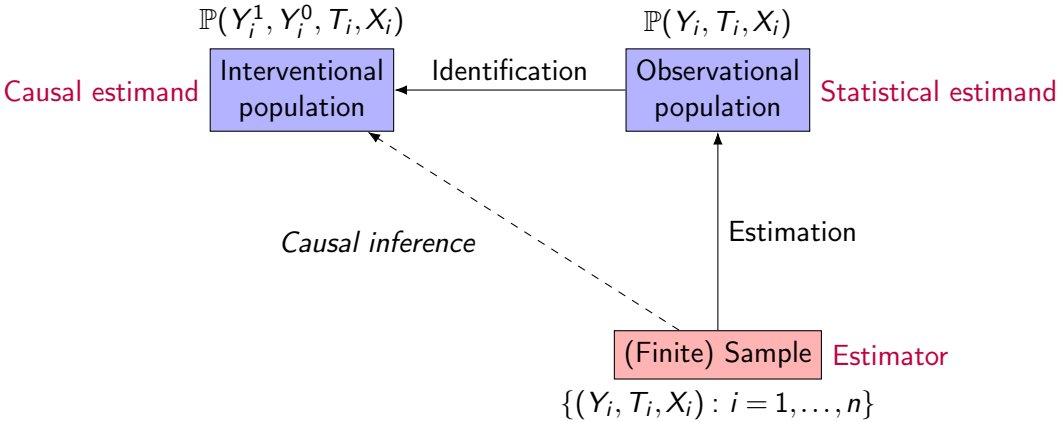
Let's now return to the observational setting. Fisher's test only focuses on the sample—it's difficult to generalize Fisher's test outside of randomized experiments.

So from here on, we will shift our focus to superpopulation inference. This means that we view our data $\{(Y_i, T_i, X_i) : i = 1, \ldots, n\}$ as being drawn from a practically infinite superpopulation.

Example: suppose we have a dataset of individual health records from a Belgian database. We might then view this dataset as being a random sample from, for example, the entire adult population of Belgium.

Because the size of superpopulation is assumed to be much larger than the size of the sample, it is often considered reasonable to model our data as being iid. Accordingly, when we work with expectations $\mathbb{E}$ and probabilities $\mathbb{P}$, we should interpret these as averages and frequencies across the superpopulation respectively.

Back to the "roadmap"



Causal estimands

Examples of common causal estimands:

- Average treatment effect

$$ATE = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$$

- Average treatment effect on the treated

$$ATT = \mathbb{E}[Y^1 \mid T = 1] - \mathbb{E}[Y^0 \mid T = 1]$$

- Causal risk ratio

$$CRR = \frac{\mathbb{P}(Y^1 = 1)}{\mathbb{P}(Y^0 = 1)} \quad \text{(if } Y \text{ is binary)}$$

- Conditional average treatment effect (CATE)

$$CATE(x) = \mathbb{E}[Y^1 \mid X = x] - \mathbb{E}[Y^0 \mid X = x]$$

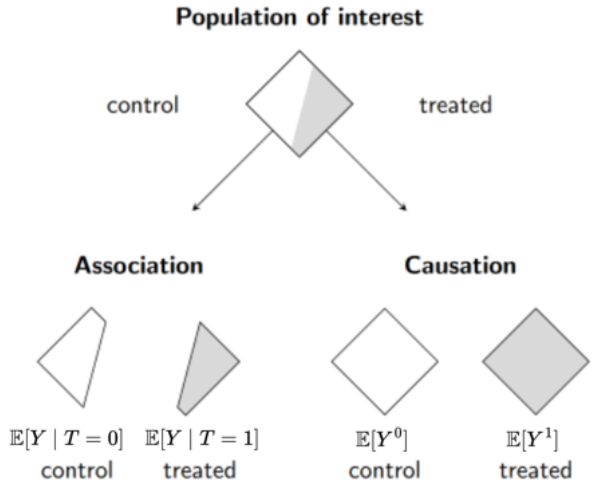
Confounding

In observational studies, the treatment assignment is not under the control of the experimenter/statistician. Without randomization, it is likely that

$$\mathbb{E}[Y^t] \neq \mathbb{E}[Y \mid T = t].$$

If this is indeed the case, we say that there is confounding.

Due to the nature of observational data, we won't know for sure whether it is possible to adjust for confounding and perform valid causal inference. We must make untestable assumptions to proceed.



From "Causal Inference: What If." – Hernán & Robins (2020)

Unconfoundedness

The most common assumption to deal with confounding is called unconfoundedness. For each unit i, we have

$$T_i \perp\!\!\!\perp (Y_i^1, Y_i^0) \mid X_i,$$

or equivalently, $\mathbb{P}(T_i \mid Y_i^1, Y_i^0, X_i) = \mathbb{P}(T_i \mid X_i)$. This is also known as *"ignorability"* or *"conditional exchangeability"*.

Important: this doesn't mean that the "treatment" is conditionally independent of the "outcomes"! Remember that T_i is the treatment **assignment** and (Y_i^1, Y_i^0) are the **potential** outcomes.

The way to interpret this is that we assume that there is randomization within levels of X, i.e. conditional on X, all other factors are balanced on average across both the treated and untreated groups.

Separating the science from the design

Under the unconfoundedness assumption, the data distribution factorizes as

$$\begin{aligned}\mathbb{P}(T_i, Y_i^1, Y_i^0, X_i) &= \mathbb{P}(T_i \mid Y_i^1, Y_i^0, X_i)\mathbb{P}(Y_i^1, Y_i^0, X_i) \\ &= \mathbb{P}(T_i \mid X_i)\mathbb{P}(Y_i^1, Y_i^0, X_i) \quad \text{(unconfoundedness)}\end{aligned}$$

Rubin calls (Y_i^1, Y_i^0, X_i) the "science"; this is what we're actually interested in learning about.

An important aspect of the Rubin Causal Model is that the "science" is separated from the "design" of the experiment (i.e. the treatment assignment mechanism). Intuitively, $\mathbb{P}(Y_i^1, Y_i^0, X_i)$ should be invariant to the choice of experiment used to uncover its properties.

Overlap

The final assumption is called overlap (aka "positivity"). This means that

$$0 < \mathbb{P}(T_i = 1 \mid X_i) < 1$$

with probability one for each unit i. In words, this requires the probability of receiving either treatment conditional on X to be strictly positive.

This would be violated, for example, if doctors always assign treatment to critically ill patients. In this case, it would be impossible to estimate causal effects among the critically ill without making heroic extrapolations from the data on other patients.

Identification

Given our assumptions (SUTVA, unconfoundedness and overlap), we can identify our causal estimand; that is, we can write it in terms of quantities that we can estimate from the observed data. For instance, let's suppose that we are interested in the average treatment effect $\chi = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$.

$$\begin{aligned}\mathbb{E}[Y^1] &= \mathbb{E}[\mathbb{E}[Y^1 \mid X]] \quad \text{(tower law)} \\ &= \mathbb{E}[\mathbb{E}[Y^1 \mid X, T = 1]] \quad \text{(unconfoundedness and overlap)} \\ &= \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] \quad \text{(SUTVA)}\end{aligned}$$

This is not the same as conditioning on $T = 1$!

$$\mathbb{E}[Y \mid T = 1] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1] \mid T = 1]$$

i.e. association is (generally) not causation!

Estimands as functionals

By doing the same for $\mathbb{E}[Y^0]$, the average treatment effect can be written as

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]].$$

The right-hand side is a nonparametric (or "model-free") formula. It describes the estimand directly as a mapping of the data-generating distribution $\chi = \chi(P)$ (often called a statistical functional).

The practice of defining estimands as model-free functionals is commonplace in causal inference, but it differs from the classical perspective of first choosing a statistical model (e.g. linear and logistic regression, the Cox model, additive hazards models) and then defining the estimand in terms of its parameters.

The average treatment effect on the treated

As a further example, let's also consider the average treatment effect on the treated from earlier:

$$\mathbb{E}[Y^1 - Y^0 \mid T = 1].$$

This is the average treatment effect across the subpopulation currently on treatment. Or in other words, *"What is the effect of withholding treatment on those being treated?"*

The treatment arm component is easy:

$$\mathbb{E}[Y^1 \mid T = 1] = \mathbb{E}[Y \mid T = 1] \quad \text{(SUTVA/consistency)}.$$

For the control arm:

$$\begin{aligned}
\mathbb{E}[Y^0 \mid T = 1] &= \mathbb{E}[\mathbb{E}\{Y^0 \mid T = 1, X\} \mid T = 1] \quad \text{(tower law)} \\
&= \mathbb{E}[\mathbb{E}\{Y^0 \mid T = 0, X\} \mid T = 1] \quad \text{(unconfoundedness)} \\
&= \mathbb{E}[\mathbb{E}\{Y \mid T = 0, X\} \mid T = 1] \quad \text{(SUTVA/consistency)}.
\end{aligned}$$

Adjustment sets

To summarize: if we have found a valid set of covariates X that satisfies the unconfoundedness and overlap assumptions, then we can identify average treatment effects:

$$ATE = \mathbb{E}[Y^1 - Y^0] = \mathbb{E}[\mathbb{E}[Y \mid X, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid X, T = 0]]$$
$$ATT = \mathbb{E}[Y^1 - Y^0 \mid T = 1] = \mathbb{E}[(Y - \mathbb{E}\{Y \mid T = 0, X\}) \mid T = 1].$$

We can then estimate the effects by fitting models for the particular components of the data-generating distribution.

But how do we decide on X?

- Should we adjust for all measured pre-treatment variables?
- What if there are unmeasured variables that we know are relevant to the model? Can we still estimate the causal effect?

Judea Pearl

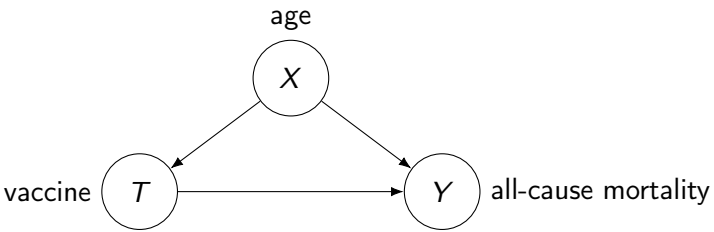
Judea Pearl

Judea Pearl—a computer scientist at UCLA—advocates a graphical approach to causality.

Among many other things, he invented a simple graphical test called the back-door criterion to choose a valid adjustment set for identification.

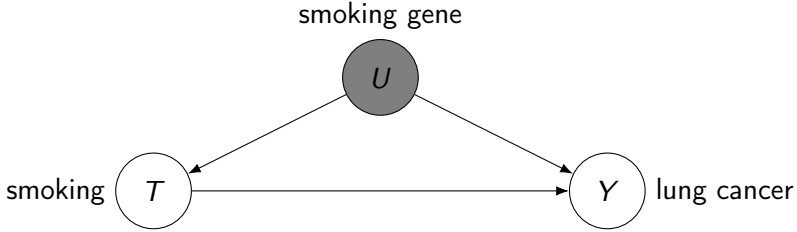
We will build up to this by first introducing some graphical terminology and concepts.

Example: vaccine efficacy



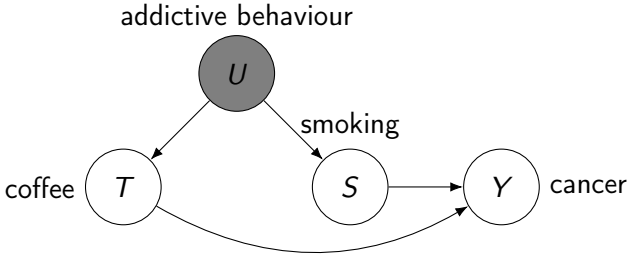
We will use graphs like the above to represent our causal models. Each node (or *vertex*) represents a random variable, and each arrow (or *directed edge*) indicates that we believe there might be a causal effect in that direction.

Example with unmeasured variables: smoking and lung cancer



This is the Fisher smoking example again. In this case, we have a potentially relevant unmeasured variable: smoking gene. We denote unmeasured variables with shaded vertices.

Example with missing edges: addictive behaviour

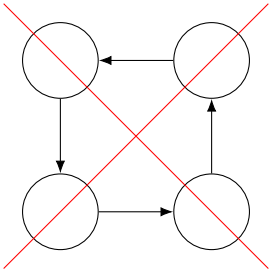


In this example, we are trying to investigate whether coffee has a causal effect on developing cancer.

Here there are some nodes without edges between them. For example, we are positing that there is no direct causal effect between drinking coffee and smoking. Instead, the graph indicates that addictive behaviour acts as a common cause, which might induce associations between coffee consumption and smoking.

Directed acyclic graphs

The graphs we use are called directed acyclic graphs (DAGs), also known as *Bayesian networks*. You may have used them before to represent Bayesian hierarchical models. As we will see, they are extremely useful for handling conditional independencies.



"Acyclic" means that we don't allow cycles, i.e. we can't allow moving along arrows and end up where we started.

DAG definitions

Definition: Paths

- A path is a sequence of distinct adjacent vertices.
- A directed path is a path where all the edges point towards the final vertex.

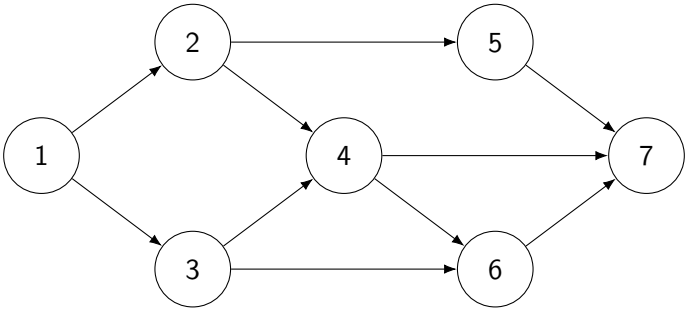
Definition: Parents and children, ancestors and descendants

- If $X \to Y$, then X is a parent of Y, and Y is a child of X.
- If $X \to \ldots \to Y$ or $X = Y$, then X is an ancestor of Y, and Y is a descendant of X.

Definition: Topological ordering

A topological ordering of vertices $V_1, \ldots, V_p$ is an ordering of $\{1, \ldots, p\}$ such that parents precede their children. By acyclicity, an ordering always exists, but it may not be unique.

DAG definition example



$\{2\}$ has parent $\{1\}$ and children $\{4,5\}$. The descendants of $\{2\}$ are $\{2,4,5,6,7\}$.

We can look for topological orderings recursively. Search for a vertex with no parents (e.g. start from any vertex and keep moving backwards): the only choice here is $\{1\}$. Then "remove" it from the graph (including edges going out of it) and search again for a vertex with no parents etc.

Blocking and d-separation

Definition: Blocking

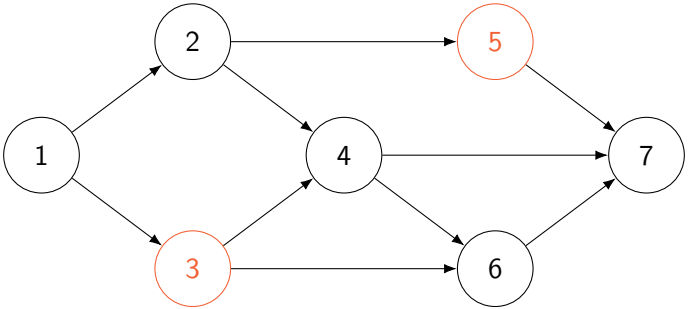
A path between V_1 and V_n is blocked by a set S (containing neither V_1 nor V_n) whenever there is a node V_k satisfying either

1. $V_k \in S$ and
   - $V_{k-1} \to V_k \to V_{k+1}$ (chain)
   - $V_{k-1} \leftarrow V_k \leftarrow V_{k+1}$ (chain)
   - $V_{k-1} \leftarrow V_k \to V_{k+1}$ (fork)
2. $V_{k-1} \to V_k \leftarrow V_{k+1}$ (collider) with neither V_k nor any of its descendants contained in S.

Definition: d-separation

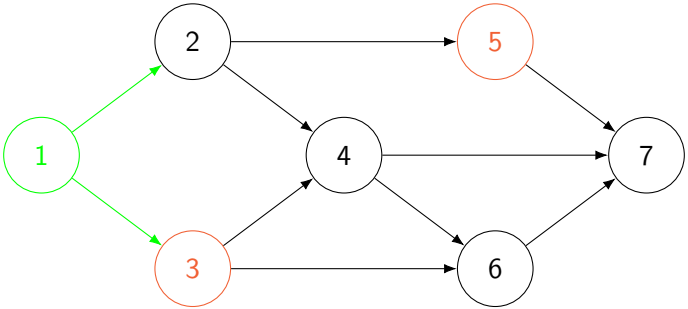
For disjoint subsets of vertices $\mathbf{A}, \mathbf{B}$ and $\mathbf{C}$, we say $\mathbf{A}$ and $\mathbf{B}$ are d-separated by $\mathbf{C}$ if all paths between $\mathbf{A}$ and $\mathbf{B}$ are blocked by $\mathbf{C}$.

Example



Claim: $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$.

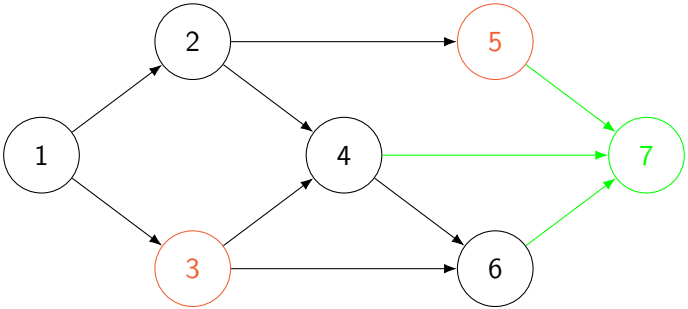
Example



Claim: $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$

First, any path that goes through $\{1\}$ is blocked.

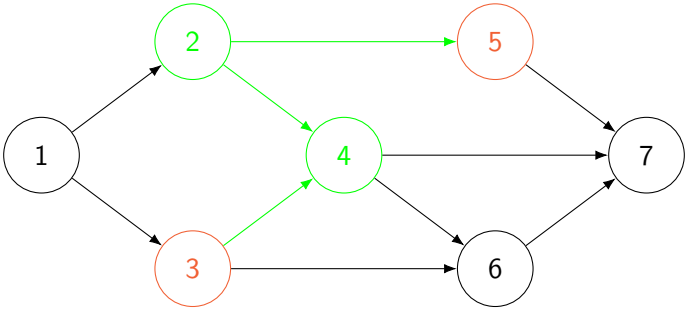
Example



Claim: $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$

Vertex $\{7\}$ has no children, so it must be a collider along any path that goes through it. Thus any path through $\{7\}$ is blocked by $\{1\}$.

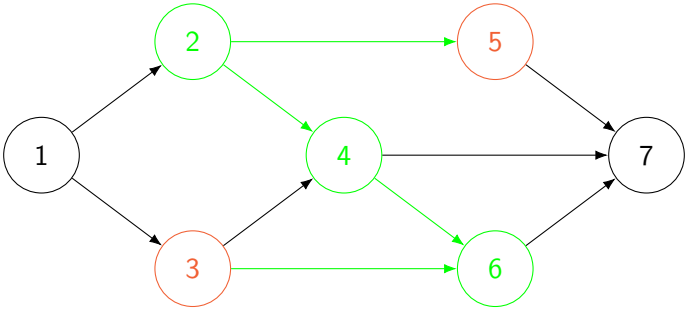
Example



Claim: $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$

$\{4\}$ is a collider and none of $\{4,6,7\}$ are contained in $\{1\}$.

Example



Claim: $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$
$\{6\}$ is a collider and $\{6\}$ and $\{7\}$ are not contained in $\{1\}$.

We have now checked that all the paths going from $\{3\}$ to $\{5\}$ are blocked, which proves the claim.

Markov factorization

Before proceeding to work with causal graphs, we introduce some useful probabilistic properties of DAGs. We want to use DAGs to encode the independence structure of our model.

Suppose $\mathscr{G}$ is a DAG with vertices $\mathscr{V} = \{V_1, \ldots, V_p\}$.

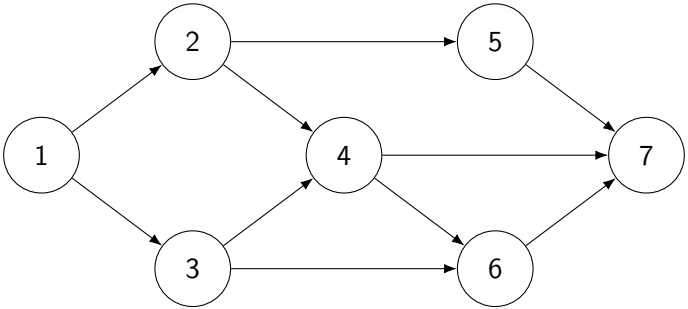
Definition

A density p on $\mathscr{V}$ factorizes according to $\mathscr{G}$ if

$$p(V_1, \ldots, V_p) = \prod_{i=1}^{p} p(V_i \mid \mathrm{pa}_{\mathscr{G}}(V_i)),$$

where $\mathrm{pa}_{\mathscr{G}}(V_i)$ is the set of parents of V_i in $\mathscr{G}$.

Example



If p factorizes according to the above, then

$$p(1,\ldots,7) = p(1)p(2 \mid 1)p(3 \mid 1)p(4 \mid 2,3)p(5 \mid 2)p(6 \mid 3,4)p(7 \mid 4,5,6).$$

Local and global Markov properties

Definition

A density p on $\mathscr{V}$ satisfies the local Markov property with respect to $\mathscr{G}$

$$V_i \perp\!\!\!\perp \underbrace{\mathrm{nd}_{\mathscr{G}}(V_i)}_{\text{"past"}} \mid \underbrace{\mathrm{pa}_{\mathscr{G}}(V_i)}_{\text{"direct causes"}}$$

for each V_i, where $\mathrm{nd}_{\mathscr{G}}(V_i)$ is the set of non-descendants of V_i in $\mathscr{G}$.

Definition

A density p on $\mathscr{V}$ satisfies the global Markov property with respect to $\mathscr{G}$ if

$$\mathbf{A} \text{ and } \mathbf{B} \text{ are } d\text{-separated by } \mathbf{C} \text{ in } \mathscr{G} \implies V_{\mathbf{A}} \perp\!\!\!\perp V_{\mathbf{B}} \mid V_{\mathbf{C}}$$

Markov properties

Theorem: Equivalence of Markov properties

The three Markov properties

- Factorization
- Local Markov property
- Global Markov property

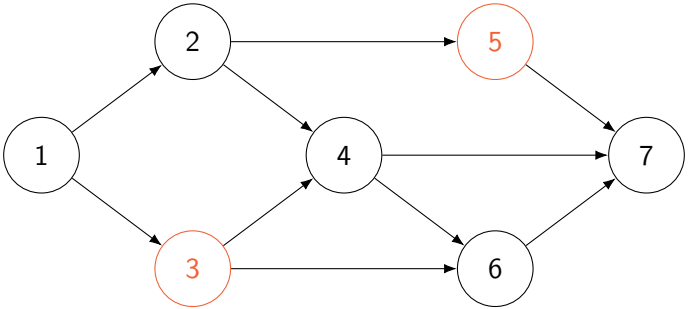
are equivalent.

Theorem: Completeness of d-separation

d-separation gives all the conditional independences implied by the DAG model.

e.g. if $\mathbf{A}$ and $\mathbf{B}$ are not d-separated by $\mathbf{C}$ in $\mathcal{G}$, then there exists a p that factorizes according to $\mathcal{G}$ such that $V_{\mathbf{A}} \not\perp\!\!\!\perp V_{\mathbf{B}} \mid V_{\mathbf{C}}$.

Example



Let's try giving an alternative argument that $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$ using the Markov properties. $\{5\}$ is a non-descendant of $\{3\}$ and $\{1\}$ is the sole parent of $\{3\}$. So the local Markov property implies $\{3\} \perp\!\!\!\perp \{5\} \mid \{1\}$.

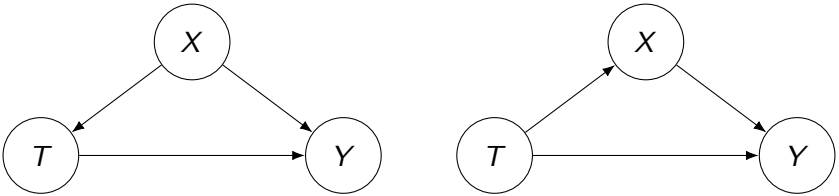
Completeness of d-separation $\implies$ $\{3\}$ and $\{5\}$ are d-separated by $\{1\}$.

Markov equivalence

Definition: Markov equivalence

We say that two DAGs $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent if they imply the same conditional independences.

By completeness, $\mathcal{G}_1$ and $\mathcal{G}_2$ are Markov equivalent iff they have the same d-separations.



So these two DAGs above are equivalent probabilistically, but they have very different causal interpretations!

Causal graphs

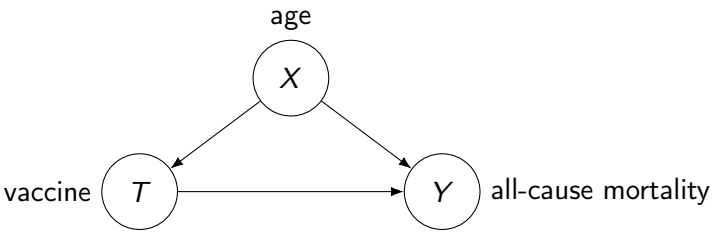
"An associational concept is any relationship that can be defined in terms of a joint distribution of observed variables, and a causal concept is any relationship that cannot be defined from the distribution alone." - Judea Pearl

Markov properties are purely associational concepts. Causal DAGs go beyond Markov properties by describing what happens under interventions, i.e. forcing variables to take particular values, rather than simply observing them.

Markov equivalence is the limit of what we can learn from data and statistics alone. Distinguishing between Markov equivalent DAGs must require user input beyond probability and stats.

The graphical methodology we are about to introduce will not conjure causal inference out of a vacuum—it is a logical framework that allows us to transform causal assumptions into causal conclusions *"Causal in—causal out"*.

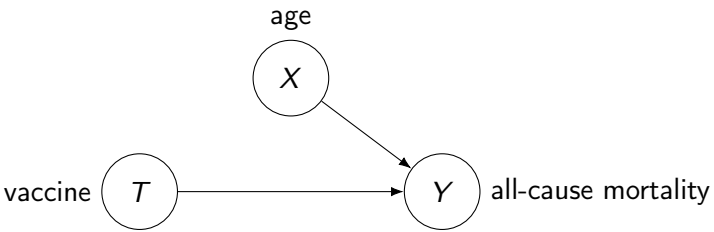
Example: vaccine efficacy



The DAG tells us the Markov structure in the observational regime:

$$p(Y, T, X) = p(Y \mid T, X)p(T \mid X)p(X).$$

Example: vaccine efficacy

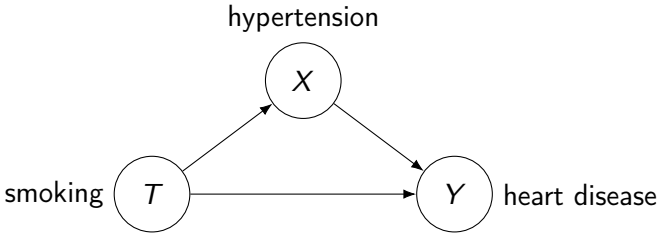


Now let's set T to the value t. Graphically, this corresponds to deleting the edge from X to T. Probabilistically, we have the truncated factorization/g-formula:

$$p(Y, X \mid do(T = t)) = p(Y \mid T = t, X)p(X).$$

This would be the model under a randomized trial where we randomized assignment to the vaccine.

Example: smoking and hypertension



This DAG has the same Markov structure in the observational regime, but in this case, the graph is unchanged if we intervene on T.

The truncated factorization is:

$$p(Y, X \mid do(T = t)) = p(Y \mid T = t, X)p(X \mid T = t).$$

Interventions

Suppose $\mathscr{G}$ is a DAG with vertices $\mathscr{V} = \{V_1, \ldots, V_p\}$, and let p be the observational density on $\mathscr{V}$.

If we intervene by setting $V_j = t$, then:

- *Graphically:* Delete all edges going into V_j in $\mathscr{G}$.
- *Probabilistically:* The interventional density has truncated factorization

$$p(\{V_i : i \neq j\} \mid do(V_j = t))) = \prod_{i \neq j} p(V_i \mid \mathrm{pa}_{\mathscr{G}}(V_i))|_{V_j = t}$$

The truncated factorization removes the intervened variable and tells us that the new interventional distribution factorizes according to the modified graph.

So we've solved identification once we've elicited a causal DAG and all variables are measured. But what is some variables are unmeasured?

The back-door criterion

Definition

A set of variables W satisfies the back-door criterion relative to (T, Y) if

- no node in W contains a descendant of T
- W blocks all paths from T to Y entering T through the back-door ($T \leftarrow \ldots$).

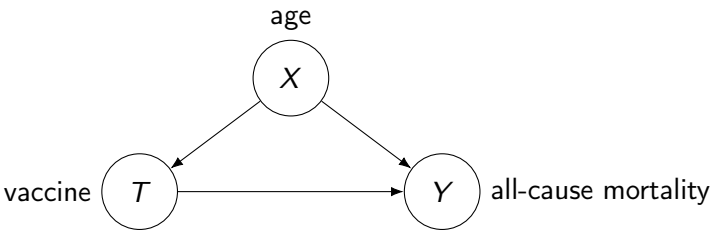
If a set of variables W satisfies the back-door criterion relative to (T, Y), then the average treatment effect of T on Y is given by the adjustment formula, i.e.

$$\mathbb{E}[Y^1] - \mathbb{E}[Y^0] = \mathbb{E}[\mathbb{E}[Y \mid W, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid W, T = 0]],$$

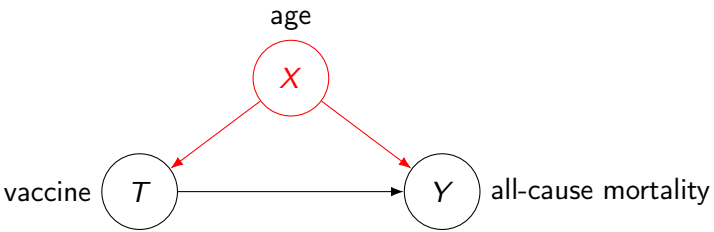
or in *do*-notation:

$$\mathbb{E}[Y \mid do(T = 1)] - \mathbb{E}[Y \mid do(T = 0)] = \mathbb{E}[\mathbb{E}[Y \mid W, T = 1]] - \mathbb{E}[\mathbb{E}[Y \mid W, T = 0]].$$

Example 1



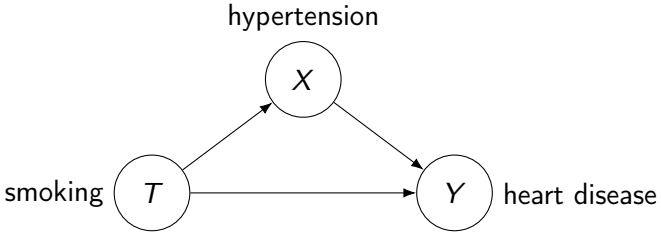
Example 1



- X is not a descendant of T
- X blocks the (only) back-door path $T \leftarrow X \rightarrow Y$.

So the back-door criterion tells us we should adjust for age!

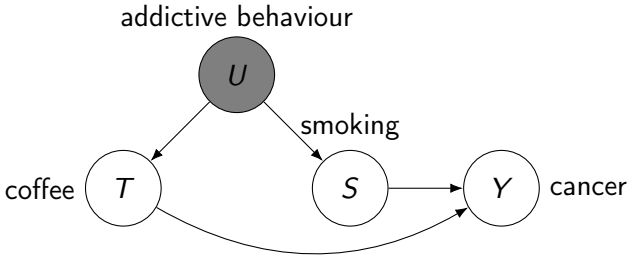
Example 2: mediation



Recall that this DAG is Markov equivalent to the previous one. So we cannot distinguish between the two graphs using data alone.

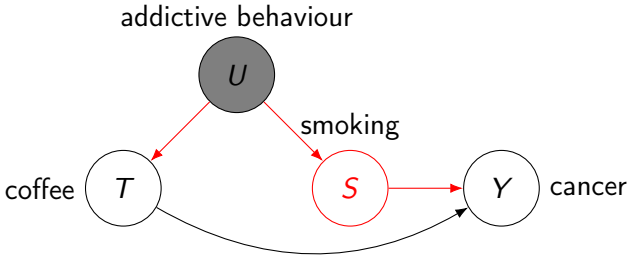
But in this case, the back-door criterion tells us **we shouldn't adjust for** X. The variable X is called a mediator.

Example 3



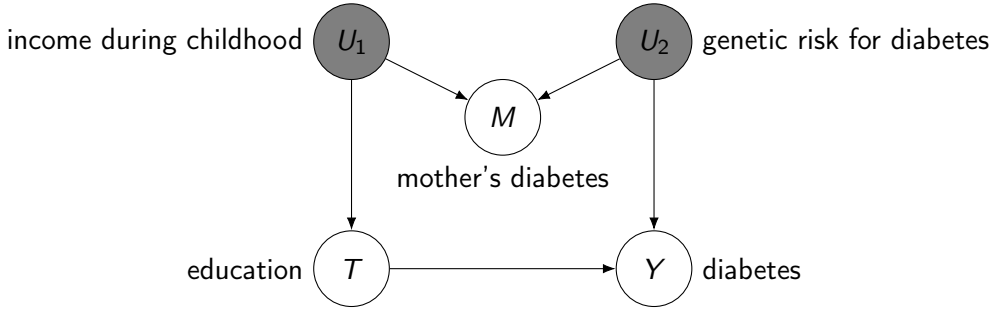
Addictive behaviour is unmeasured.

Example 3



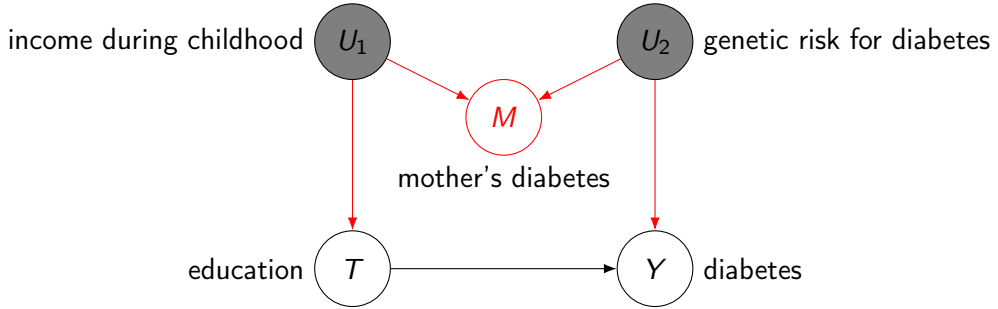
Addictive behaviour is unmeasured. We should adjust for smoking even though it doesn't have a causal effect on drinking coffee!

Example 4: M-bias



Should we adjust for M?

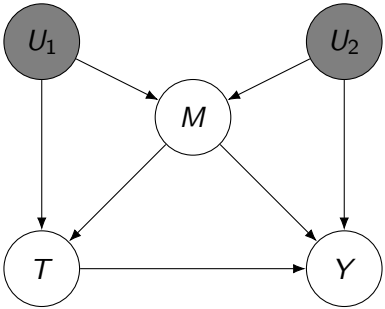
Example 4: M-bias



We shouldn't adjust for the mother's diabetes even though

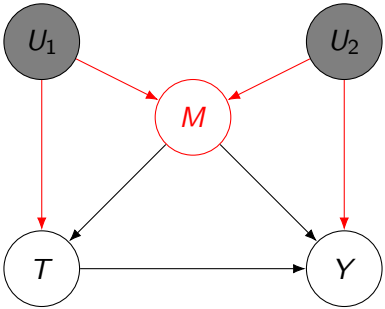
- it's a pre-treatment variable
- it's associated with both T and Y.

Example 5: butterfly bias



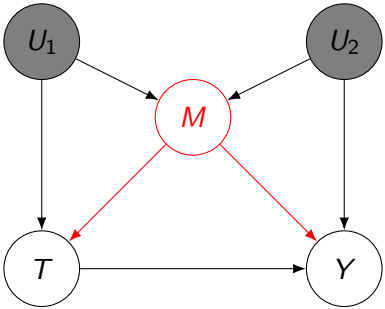
Should we adjust for M?

Example 5: butterfly bias



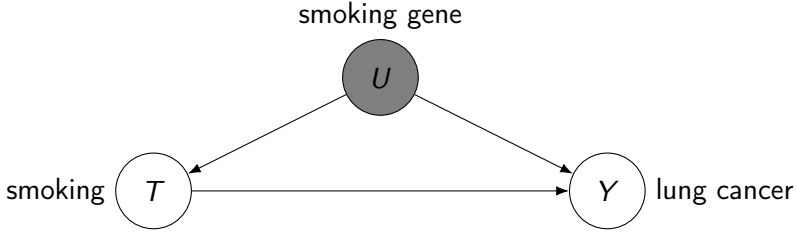
Similar to before, M is a collider on this path, which suggests we shouldn't adjust for it.

Example 5: butterfly bias

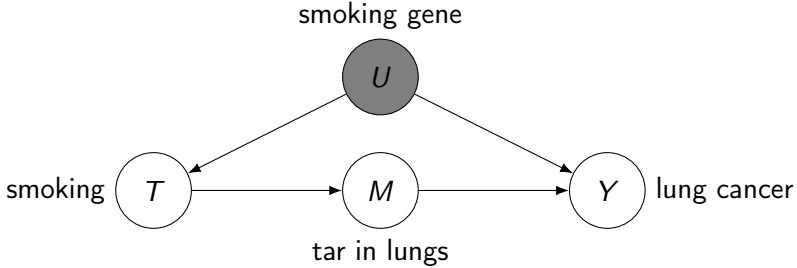


But at the same time, we need to adjust for M because it's a common cause of T and Y! In this case, the average treatment effect is unidentifiable.

Back to the Fisher smoking example



The front-door criterion



To resolve the Fisher smoking problem, Pearl suggested a possible measurable variable *"tar in lungs"* that lies on the causal pathway between smoking and lung cancer.

There is no back-door adjustment set, but remarkably, the treatment effect is identifiable by the so-called front-door criterion:

$$\mathbb{E}[Y^t] = \sum_{m,t'} \mathbb{E}[Y \mid M = m, T = t'] p(M = m \mid T = t) p(T = t').$$

Structural equation models

In Pearl's Structural Causal Model, causal graphs are abstractions of nonparametric structural equation models.

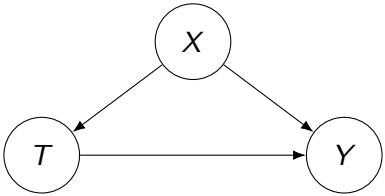
Each variable in a DAG $\mathcal{G}$ is assumed to be a deterministic function of its parents and some additional noise:

$$V_i = f_i(\mathrm{pa}_{\mathcal{G}}(V_i), U_i),$$

where $\{U_i\}$ are jointly independent noise variables (aka *exogenous* variables).

- "Nonparametric" refers to the fact that the deterministic functions f_i and the marginal distributions of the noise variables U_i have (so far) been left completely unspecified.
- "Structural" refers to the causal structure that is implied by the equations.

Structural equation models



This graph is associated with:

$$
\begin{aligned}
X &= U_X \\
T &= f_T(X, U_T) \\
Y &= f_Y(T, X, U_Y),
\end{aligned}
$$

where (U_X, U_T, U_Y) are jointly independent.

Linear structural equation models

Early use of structural equation models centred around linear SEM's:

$$
\begin{aligned}
X &= U_X \\
T &= \alpha X + U_T \\
Y &= \beta T + \gamma X + U_Y.
\end{aligned}
$$

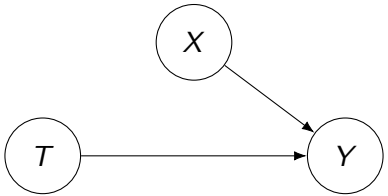
This looks a lot like linear regression! But additionally, these equations also tell us what happens if we intervene on treatment.

Structural equation models under intervention

Suppose we force treatment to take the value 1:

$$
\begin{aligned}
X &= U_X \\
T &= 1 \\
Y &= \beta \times 1 + \gamma X + U_Y.
\end{aligned}
$$

Note that X is unchanged; intuitively, it occurred "before" treatment.



Connections with potential outcomes

Potential outcomes are implicit in structural equations models! Define

$$Y^1 = f_Y(1, X, U_Y) \quad \text{(force treatment to be 1)}$$
$$Y^0 = f_Y(0, X, U_Y) \quad \text{(force treatment to be 0)}$$

So $T \perp\!\!\!\perp (Y^1, Y^0) \mid X \iff U_Y \perp\!\!\!\perp U_T$.

It's also possible to go in the opposite direction: we can use a potential outcome model to induce a structural equation model[1].

So potential outcome models and nonparametric structural equation models are mathematically equivalent!

[1] see `arxiv.org/abs/2008.06017`.

Graphs

Potential outcome models dominate *statistics* and *econometrics*, while graphical causal models are prevalent in *computer science*. Epidemiologists often use both together.

Pros of graphs:

- Graphs transparently represent the causal relationships assumed in the analysis.
- Very convenient for identification.
- Can help to guide the user in selecting variables to measure?

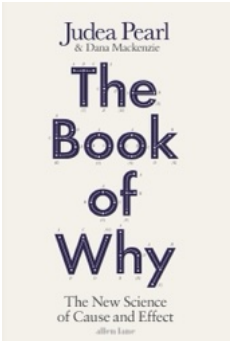
Cons of graphs:

- Might make the user overconfident in their causal findings?

More on identification

- Single world intervention graphs (SWIGs) combine potential outcomes and graphs. They allow us to read off assumptions like unconfoundedness/ignorability by applying *d*-separation to modified graphs under intervention. SWIGs are also based on a potential outcome model that only makes Markov assumptions verifiable by randomized experiments (see Chapter 7 of Hernán & Robins (2020)).
- The do-calculus: most examples in practice can be handled with the back-door and front-door criteria, but not all. The *"do-calculus"* is a set of three rules that is complete for identification; that is, by iterating the rules, it can always return the identification formula for any identifiable model.
- Confounder selection: We learned that the back-door criterion can help us to identify sufficient adjustment sets. But adjustment sets are possibly non-unique. There are many potential objectives to consider when choosing between adjustments sets, including robustness to model misspecification, efficiency, resources etc.[2].

[2] see `https://arxiv.org/abs/2208.13871` for a review.



Pearl (2018)

Hernán and Robins (2020)

- "The Book of Why" (2018) - Judea Pearl and Dana Mackenzie: popular science-type book with lots of history and fun examples.
- "Causal Inference: What If" (2021) - Miguel Hernán and James Robins: accessible textbook that provides a balanced overview of the different approaches to causal inference.

Summary

- It is important to think about our problems from a causal perspective in order to guide decision making.
- The cornerstone of causal inference is the randomized experiment. But there are many questions that can only be answered with the help of non-randomized data.
- Causal analyses separate identification from estimation.
- Graphs are a very helpful way of representing our causal assumptions and enables convenient tools for identification such as the backdoor criterion.