CF. Lecture of Introduction to Causal inference from a Machine Learning Perspective.

Book by Brady Neal.

① Motivations.
② Potential Outcomes
③ The flow of Association & Causation in graphs
④ Causal Models
  . do operator
  . Structural Causal models
⑤ Randomized experiments.
⑥
⋮
⑪ Causal discovery from observational data
⑫ Causal _____ interventional data. ~.

(non fini).

Chap 1 Motivation.

1 Simpson's doust.

Imagine a pandemic. 2 treatments.

Treatment (T)		Mild (C)	Severe (C)	Total
	A	45% (210/1400)	30% (30/100)	16%* (240/1500)
	B	10%* (5/50)	20%* (100/500)	19% (105/550)

□ % of deaths

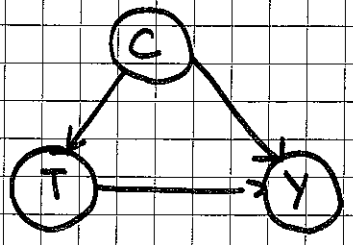
Y outcome (death).

⚠ Treatment A more efficient for Mild Condition
Severe Condition
Treatment B ——— for total population

What treatment should we use?

The answer comes from the causal structure we could come?
set on the data.

Scenario 1

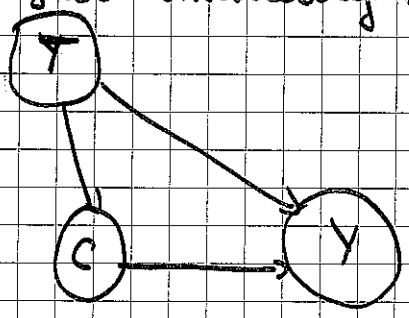


ex doctors keep treatment B (more expensive) for patients with severe conditions.

In this case, treatment B is associated to ↑ mortality because it is given to people more likely to die.
→ choose B because in each condition it works better

Scenario 2 Treatment is a cause of condition

For instance treatment B is rare (scarce) so people have to wait for it and it worsens there condition => higher mortality.



=> Prefer A.

Conclusion Setting a causal assumption will help to decide.

[=> We need to be able to get the causal graph to solve the problem.

② How to capture the causal relation?

⚠ correlation / association is not causality

2 curves. purely coincidence. spurious correlation

- see wiz

(although AI will find an explanation).

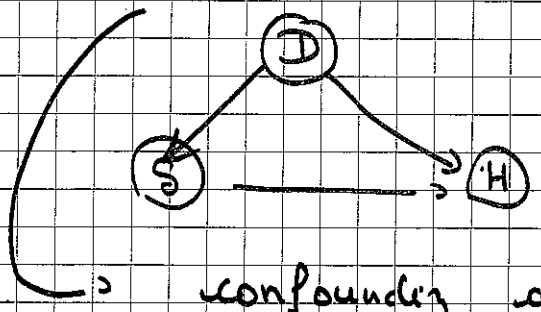
Why correlation replaced by association.

because correlation = linear.

Here association = statistical dependence

Ex people going to bed with shoes are more likely to have headaches.

⚠ Common cause. : drinking the night before.



⇒ confounding association.

If association = causality ⇒ standard statistics would be sufficient.

description of main themes ⇒ worldimage.

Chap 2 — Potential outcomes

Consider the situation / experiment where you get a treatment (T) and observe an outcome (Y) after treatment.

ex T Y

medic / no medic ↓ead or not.

dog / no dog being more happy or not.

- If getting a dog makes you happy how can you be sure that you r happy BECAUSE of the DOG.
- If get the medecine you are better. But what would have happened without the treatment? Maybe you would be or too?

We can not go back in time !!!

We define 4 variables.

T treatment

Y outcome of the experiment

$Y(t)$ $\begin{cases} Y(1) & \text{potential outcome with the treatment } 1 \\ Y(0) & \text{potential outcome with the treatment } 0. \end{cases}$

Assumptions 1 consistency

$Y = Y(T)$

if treatment T is given the the observed outcome Y is the potential outcome under T.

Individual treatment effect.

$$\tau = Y(1) - Y(0).$$

⚠ if population $i = 1 \dots n$ $\quad Y_i, X_i, T_i$ (X_i: covariates)

$$\tau_i = Y_i(1) - Y_i(0).$$

Pb observations

fundamental problem of causal inference

we only observe $Y_i(1)$ or $Y_i(0)$ never both.

		Y		
i	T	$Y(1)$	$Y(0)$	$Y(1) - Y(0)$.
1	0	?	0	?
2	1	1	?	?
3	1	0	?	?
4	0	?	1	?

The potential outcome that we do not observe are called **counterfactuels**

The counter factuels only exists once the outcome has been observed.

2.3 Getting around the fundamental problem.

We can't access the individual T.E

Can we get the **average treatment effect** (ATE) over a population?

$ATE = \tau = \mathbb{E}[Y(1) - Y(0)]$ where the average is over the individuals.

$= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ (linearity of expectation).

Is it equal to $\mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$?

- No in general.
- Yes if $(Y(1), Y(0)) \perp\!\!\!\perp T$ (A.1) Ignorability / Exchangeability

Indeed If this is true. $\Rightarrow$

$ATE = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] = \mathbb{E}[Y(1)|T=1] - \mathbb{E}[Y(0)|T=0]$

$= \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$

$\Rightarrow$ on peut estimer ATE en faisant des moyennes / colonnes ds le tableau.

Stat classiques.

Ignorability ⇔ T indépendant des potential outcome.
or T lié au fait que Y(1) ou (Y(0) est non observé.
le fait que c'est manquant ne dépend pas des outcome ⇒ on peut faire des moyennes en "ignorant" le pb des données manquantes.

Exchangeability -

$$\mathbb{E}[Y(1) \mid T=0] = \mathbb{E}[Y(1) \mid T=1]$$

ie. les groupes sont échangeables.

Traitement.

Group treatments are comparables.
the same in all relevant aspects other than treatment.

Linked to the concept of : "controlling for"

We used the assumption A1 to identify causal effet

def 1 A causal quantity (ex $\mathbb{E}[Y(t)]$) is identifiable if we can compute it from a purely statistical quantity of $\mathbb{E}[Y \mid T=t]$.

But A1 is unrealistic : surely there is confounding.
Ignorability un most data.

Solution : randomized experiences. (Chap 5).

- propose a less unrealistic assumption.

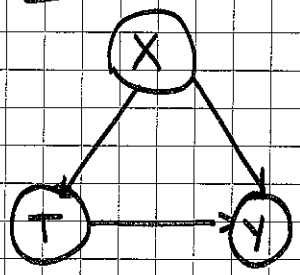
⇒ conditional exchangeability.

Assumption 2

Covariates X which describes the individuals.

$$(Y(1), Y(0)) \perp\!\!\!\perp T \mid X$$

Conditional Exchangeability or unconfoundness.



⇒ we are able to identify the causal effets with en levels of X

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

  $(A2)$ $= \mathbb{E}[Y(1) \mid X, T=1] - \mathbb{E}[Y(0) \mid X, T=0]$

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

- **Marginal effet.**

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

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

$$\boxed{\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}_X[\mathbb{E}(Y \mid X, T=1) - \mathbb{E}[Y \mid X, T=0]]}$$

↓ Δ positivity

demo were detailed after (*)

⚠ exchangeability ⇒ conditional exchangeability to be more realistic.

But it many exists unobserved confounders that are note in $\underline{X}$.

⇒ observation data (Chap 3 & 4)

⇒ nonrandomized trials → not a pb.

⚠ We are tempted to add as many covariates as possible.

But in fact we need an assumption of positivity

[A3] $\forall x \mid \quad 0 < \mathbb{P}(T=1 \mid X=x) < 1$ — Positivity / Overlap / Common support.

$P(X=x) > 0$.

Why is it important

$$\mathbb{E}_X \left[\mathbb{E}[Y \mid T=1, X] - \mathbb{E}[Y \mid T=0, X] \right]$$

$$= \sum_x \mathbb{P}(X=x) \left[\sum_y y\, \mathbb{P}(Y=y \mid T=1, X=x) - \sum_y y\, \mathbb{P}(Y=y \mid T=0, X=x) \right]$$

$$= \sum_x \mathbb{P}(X=x) \left[\sum_y y \frac{\mathbb{P}(Y=y, T=1, X=x)}{\mathbb{P}(T=1 \mid X=x)\,\mathbb{P}(X=x)} - \sum_y y \frac{\mathbb{P}(Y=y, T=0, X=x)}{\mathbb{P}(T=0 \mid X=x)\,\mathbb{P}(X=x)} \right]$$

$\mathbb{P}(T=0 \mid X=x) \neq 0$ ⚠

$$= \sum_x \sum_y y\, \mathbb{P}(Y=y \mid T=1, X=x) - \sum_x \sum_y y\, \mathbb{P}(Y=y \mid T=0, X=x)$$

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

Intuition: means you do observe some ~~observed~~ covariates with treatment or not.

Why "same support" $Supp(\mathbb{P}(X \mid T=0)) = Supp(\mathbb{P}(X \mid T=1))$.

Positivity - unconfoundness tradeoff of life

Other assumptions.

Hyp 4 No interference

$$Y_i(t_1, \dots t_{i-1}, t_{i+1}, \dots t_n) = Y_i(t_i).$$

Independance ??

Note 2.4 $\Rightarrow$ terms of statistics.

- estimand = quantity that we want to estimate
- statistical estimand: does not imply the potential outcome.

$$\underbrace{\mathbb{E}[Y(1) - Y(0)]}_{\text{causal estimand}} = \underbrace{\mathbb{E}_X[\mathbb{E}(Y \mid T=1, X) - \mathbb{E}[Y \mid T=0, X)]]}_{\text{statistical estimand}}$$

identification

3:30 2:30

6:00 × 5

How to estimate.

$\mathbb{E}_x\left[\mathbb{E}[Y \mid T=1, X]\right]$ ⇝ **model**-assisted estimators.

↳ $\frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}[Y \mid T=1, X=x_i]}_{\text{linear model}}$.

Empirical over data.

Continuous treatment

We care about $\mathbb{E}[Y(t)] = \mathbb{E}[Y \mid T=t]$: how it changes with t ⇒

if linear $\alpha t + \beta$: $\frac{\partial}{\partial t} \mathbb{E}[Y(t)] = \alpha$.

⇒ pb if not linear of

Chap 3 The flow of Associations & Causation in Graphs.

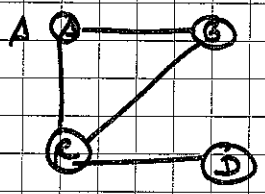
- Graphs have seemed to be convenient to represent aid intuition

⇒ utility of graphical models.

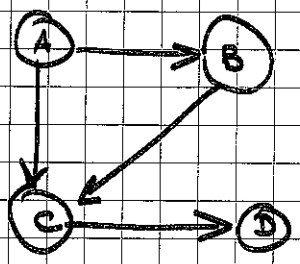
3.1 graph terminology

Definitions

- **Graph** = collection of edges & nodes.
- **undirected**



- **directed**

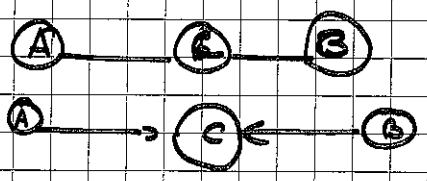


- The directed edge goes from **parent node** to **child node**.
- $\begin{cases} pa(X) = \text{ensemble of the parents of } X. \\ pa_i = pa(X_i) \end{cases}$
- 2 Nodes are **adjacent** if there ∃ an edge between them.

B & D non adjacent

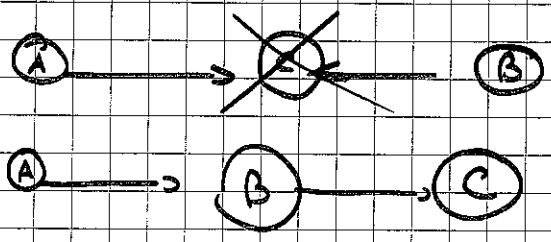
A & B adjacent

- **Path** = any sequence of adjacent nodes regardless the direction of edges.



} path.

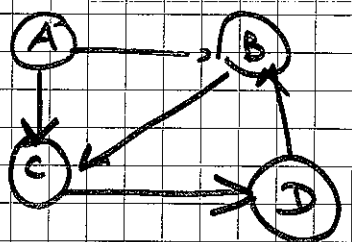
- **Directed path** : a path with directed edges all in the same direction



- If a directed path start in X and ends in Y. then

{ X = ancestor
{ Y = descendant

- **Cycle** if a path starts and end at the same node.



B → C → D → B

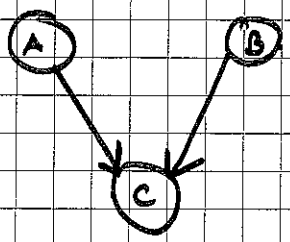
is a cycle

If there is no cycles in a directed graph.

graphe = DAG

directed acyclic graph.

Immorality : two parents and no edges between the parents



2. Bayesian Networks.

proba graphical models : broad class of models that uses graphs to represent the relationships/dependance between random variables.

→ directed graphs : ex Bayesian networks

non-directed graph : Markov field

Bayesian network the graph is a DAG

(is a proba graphical model with particular graph).

Factorisation : $p(X_1 \dots X_m) = \prod_{i=1}^{m} p(X_i \mid pa(X_i))$

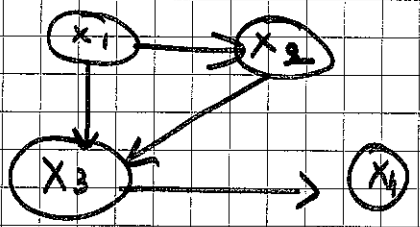
- Let $P(X_1 \dots X_n)$ be a proba distribution.
  and G a DAG with n nodes.
  with node i representing X_i.

Def P is Markov with respect to G if.
(locally)
any random variable X_i is independant of
all its non descendants in G given its parents

Conséquence **Take 4 nodes**

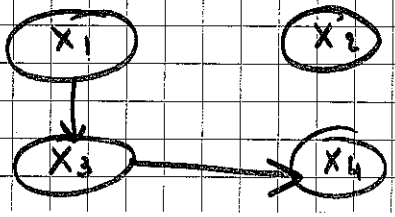
- $P(x_1\ x_2\ x_3\ x_4) = P(x_1)\ P(x_2 \mid x_1)\ P(x_3 \mid x_2\ x_1)\ P(x_4 \mid x_3\ x_2\ x_1)$
  | Always true (Chain rule of probality.)
- If we know that P is markov with repect to

G_1 :



$\Rightarrow p(x_4 \mid x_3 \overbrace{x_2\ x_1}^{\text{non descendants}}) = p(x_4 \mid x_3)$
$\equiv$ parent

G_2 :



$p(x_2 \mid x_1) = p(x_2)$
$p(x_3 \mid x_1\ x_2) = p(x_3 \mid x_1)$
$p(x_4 \mid x_3\ x_2\ x_1) = p(x_4 \mid x_3)$

Consequence

If P is locally Markov with respect to $\mathcal{D}$.

Then

$$P(x_1, \dots x_n) = \prod_{i=1}^{n} p(x_i \mid PA(x_i)).$$

|| Bayesian Network factorisation.
|| Chain rule for Bayesian network.
|| Markov compatibility.

① Je pense que c'est pas direct comme DEMO a que la DAG y joue un rôle. A revoir avec un vrai livre de graphes.

in fact it is an equivalence.

Remarque The local Markov assumption only gives us information on the independencies

It does not tell us noting about dependance between adjacent nodes in the DAG.

Pour assurer cela il faut en dire un peu plus.

A.S Minimality Assumption

P is locally Markov with respect to G & G is minimal if

- Given its parents a node is indep of its non descendant.
- Adjacent nodes in DAG are dependant

if P is minimal there is no additional independencies

=> if a ~~DAG~~ (X) ——— (Y).

if P is Markov with repect to P

=> $P(x,y) = P(y|x)\,p(x)$.

But maybe also $P(x,y) = p(y)\,p(x)$

if B is minimal

=> $P(x,y) \neq p(x)\,p(y)$.

We can't remove any other edges. if G is minimal

=> Any edge is <u>active</u>

3. Causal Graphs

<u>Def</u>- What is a cause.

A variable X is said to be a cause of Y if Y can change in response to changes in X.

<u>Causal Edge assumption</u>