CI = A DEFINITION OF CAUSAL EFFECT

× Notation

- $A = \begin{cases} 1 & \text{if Treated} \\ 0 & \text{o/w} \end{cases}$ = dichotomous treatment variable

(assuming here that there is a single version of the treatment; which is not always the same)

- $Y = \begin{cases} 1 & \text{if Survival} \\ 0 & \text{if death} \end{cases}$ = dichotomous outcome variable

- $Y^{a=1} = Y^1$ = outcome variable that would have been observed under the treatment value $a=1$.
- $Y^{a=0} = Y^0$ = —— " —— $a=0$.

Y^0 & Y^1 are referred to as POTENTIAL OUTCOMES.

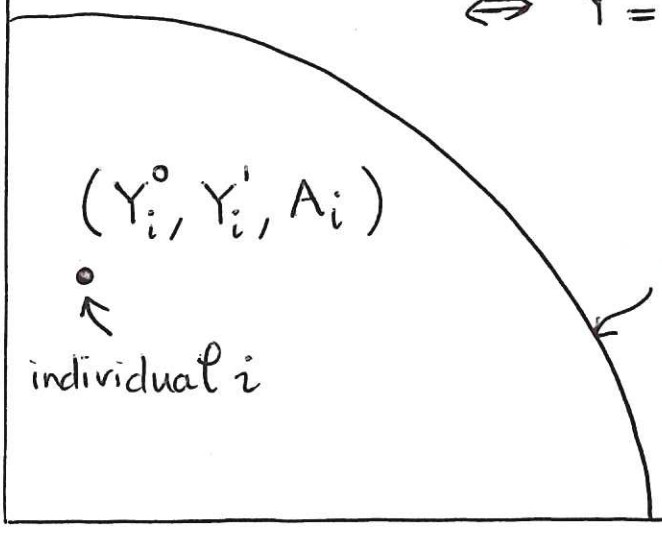
"either one of these may be potentially observed".

↳ one of them is actually factual i.e. observed.

If $A = a$, then $Y = Y^a$

"CONSISTENCY"

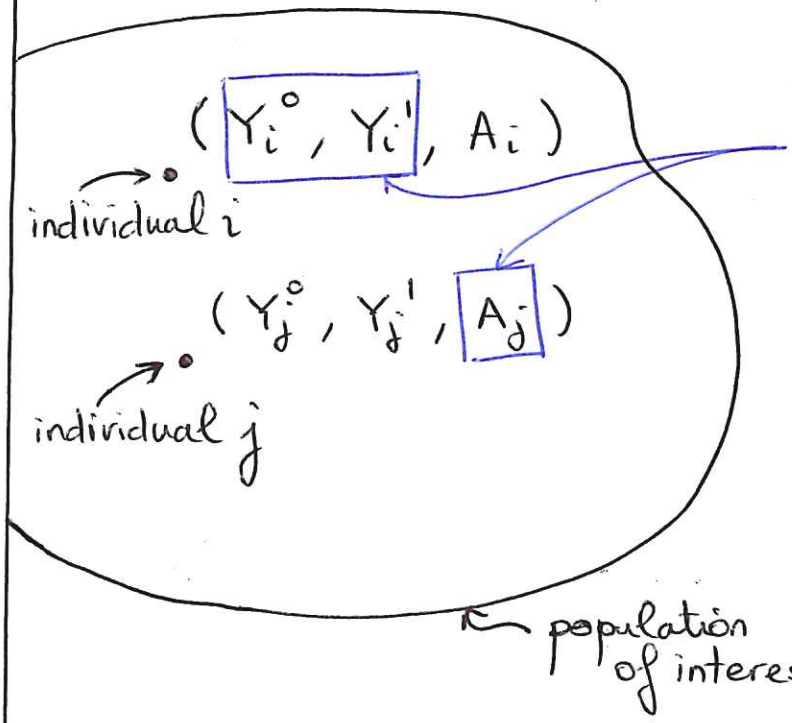
$\Leftrightarrow Y = A Y^1 + (1 - A) Y^0$



The causal effect for individual i is $Y_i^1 - Y_i^0$

Individual effects cannot be identified since only one potential outcome is factual.

↳ instead, we look into aggregated effects i.e. average causal effects in a population of individuals.



The causal effect of individual i is well defined provided it does not depend on the treatment value of individual j: i.e. when there is no interference.

SUTVA assumption

Rubin (1980)

Under SUTVA, we can compute / estimate $\mathbb{P}(Y^a = 1)$.

finite population, all observed.

infinite / large population, only a sample is observed

Def: An average causal effect of treatment A on outcome Y is present if $\mathbb{P}(Y^1 = 1) \neq \mathbb{P}(Y^0 = 1)$.

↖ An absence of average causal effect does not imply an absence of individual effects.

effect measures

$\mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1)$ [risk difference] (~ absolute numbers)

$\mathbb{P}(Y^1 = 1) / \mathbb{P}(Y^0 = 1)$ [risk ratio] (~ how much more likely)

x **Remark** = There are two sources of random error =

SAMPLING VARIABILITY + NONDETERMINISTIC P.O.

Causation vs Association.

For individual i, $Y_i = A_i Y_i^1 + (1 - A_i) Y_i^0$.

We observe the treatment A_i and the outcome Y_i.

We can compute / estimate $\mathbb{P}(Y=1 \mid A=a)$

A measure of association between A and Y.

If $\mathbb{P}(Y=1 \mid A=1) = \mathbb{P}(Y=1 \mid A=0)$,

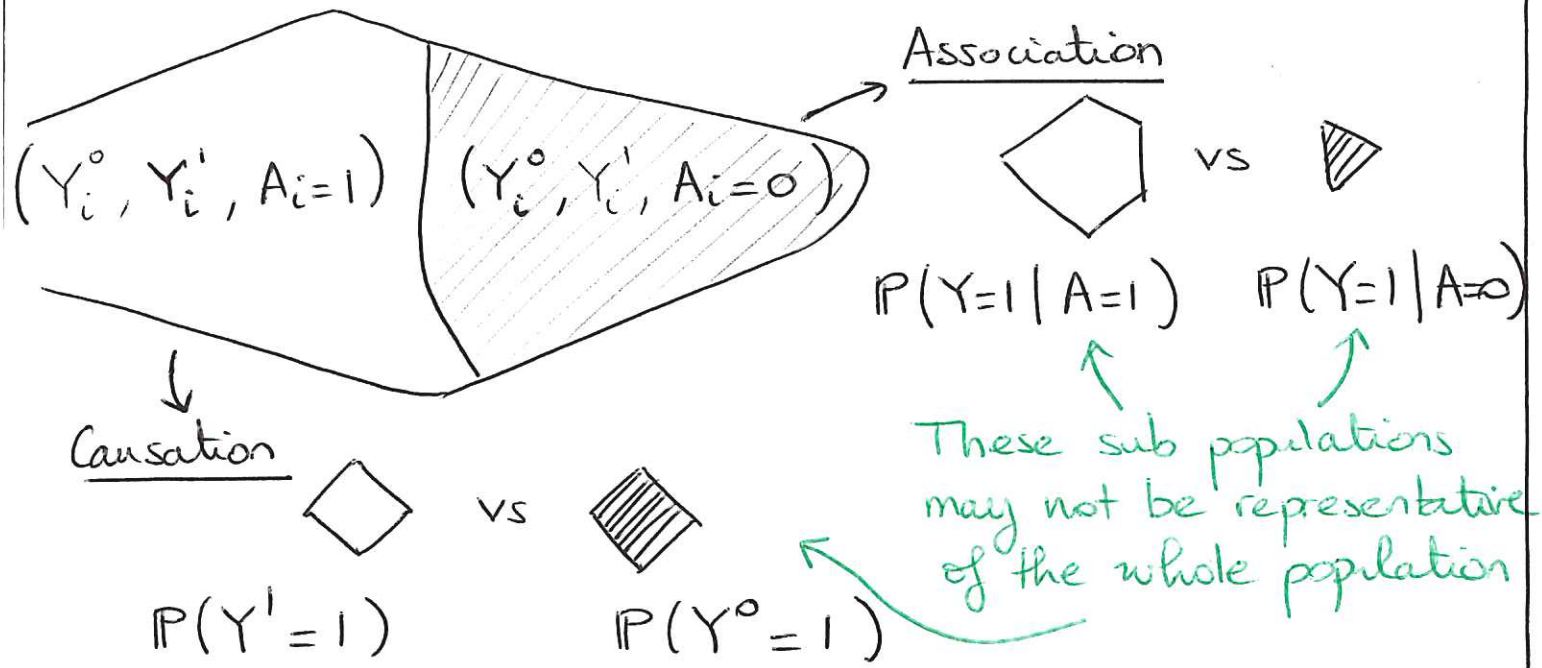
then $Y \perp A$ (independence)

Association Measures

$\mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=1 \mid A=0)$

$\frac{\mathbb{P}(Y=1 \mid A=1)}{\mathbb{P}(Y=1 \mid A=0)}$

[Not unusual to find an association when there is no effect]

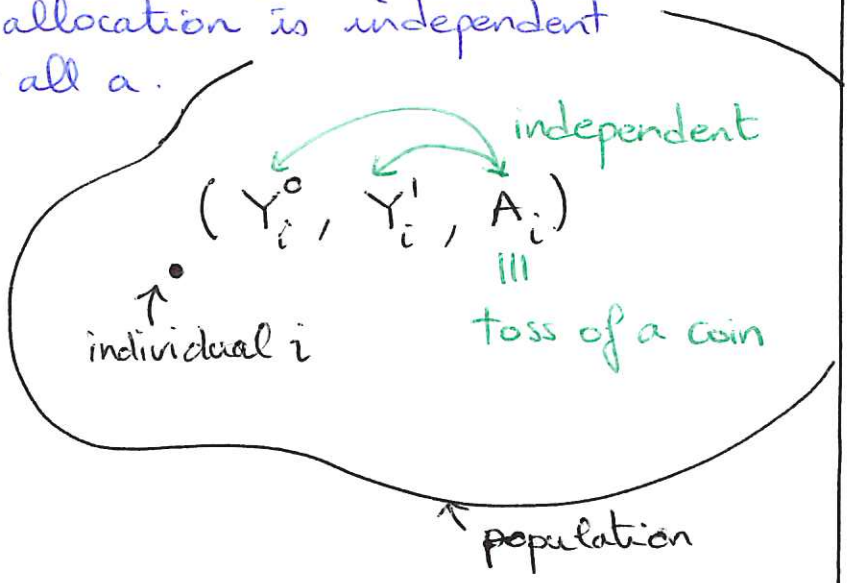


CI = RANDOMISED EXPERIMENTS

One of Y^0 or Y^1 is factual

↳ In a random experiment, which one of Y^0 or Y^1 is observed occurs just by chance.

⇒ Treatment allocation is independent of Y^a for all a.



This setup is referred to as **EXCHANGEABILITY**

$$Y^a \perp A \text{ for all } a$$

Consequence =

$$\mathbb{P}(Y^0=1) = \mathbb{P}(Y^0=1 \mid A=1) = \mathbb{P}(Y^0=1 \mid A=0)$$
$$\mathbb{P}(Y^1=1) = \mathbb{P}(Y^1=1 \mid A=1) = \mathbb{P}(Y^1=1 \mid A=0).$$

These are precisely the conditions of a Randomized Experiment.

- Remark: $Y^a \perp A$ vs $Y \perp A$

In cases where an average causal effect of treatment A on outcome Y exist, $\mathbb{P}(Y^1=1) \neq \mathbb{P}(Y^0=1)$

consistency ↳ $\mathbb{P}(Y^1=1 \mid A=1) = \mathbb{P}(Y=1 \mid A=1)$

consistency ↳ $\mathbb{P}(Y^0=1 \mid A=0) = \mathbb{P}(Y=1 \mid A=0)$

We conclude that $\mathbb{P}(Y=1 \mid A=1) \neq \mathbb{P}(Y=1 \mid A=0)$, i.e. $Y \not\perp A$.

$\Rightarrow$ Exchangeability condition is on the potential outcomes: ($Y^a \perp A$ for all a), not on the outcome variable Y.

• Conditional Randomization

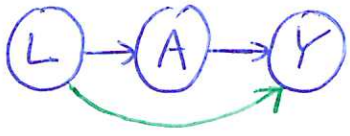
Consider an additional variable $L \in \{0, 1\}$

↗ marginally randomized experiment
(toss one coin and allocate individuals)

↘ conditionally randomized experiment
(toss one coin for individuals with $L=0$, and another coin for those with $L=1$)
↖ i.e. conditionally on L.

↓

equivalent to two separate marginally randomized experiments.



Thus $Y^a \perp A \mid L=0 \quad \forall a$

$Y^a \perp A \mid L=1 \quad \forall a$

holds on $\neq$ subsets / strata of the population.

In other words, $Y^a \perp A \mid L$

The average causal effect can be computed / estimated within each stratum, or across the whole population.

We discuss next two (equivalent) techniques to compute / estimate $\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$ in a conditionally randomized experiment = STANDARDIZATION & INVERSE PROBABILITY WEIGHTING.

• STANDARDIZATION

Use the law of Total Probability:

$$\mathbb{P}(Y^a=1) = \sum_\ell \mathbb{P}(Y^a=1, L=\ell)$$

$$= \sum_\ell \underbrace{\mathbb{P}(Y^a=1 \mid L=\ell)}\ \mathbb{P}(L=\ell)$$

conditionally on L, $(Y^a \perp A \mid L)$

$$= \mathbb{P}(Y^a=1 \mid L=\ell, A=a)$$

$$= \sum_\ell \mathbb{P}(Y^a=1 \mid L=\ell, A=a)\ \mathbb{P}(L=\ell)$$

consistency ↪

$$= \sum_\ell \mathbb{P}(Y=1 \mid L=\ell, A=a)\ \mathbb{P}(L=\ell)$$

(*) ↗

These can be computed / estimated from the data

• INVERSE PROBABILITY WEIGHTING

Note that $\mathbb{P}(Y=1 \mid L=\ell, A=a) = \dfrac{\mathbb{P}(L=\ell, A=a, Y=1)}{\mathbb{P}(L=\ell, A=a)}$

⇒ From (*) we get

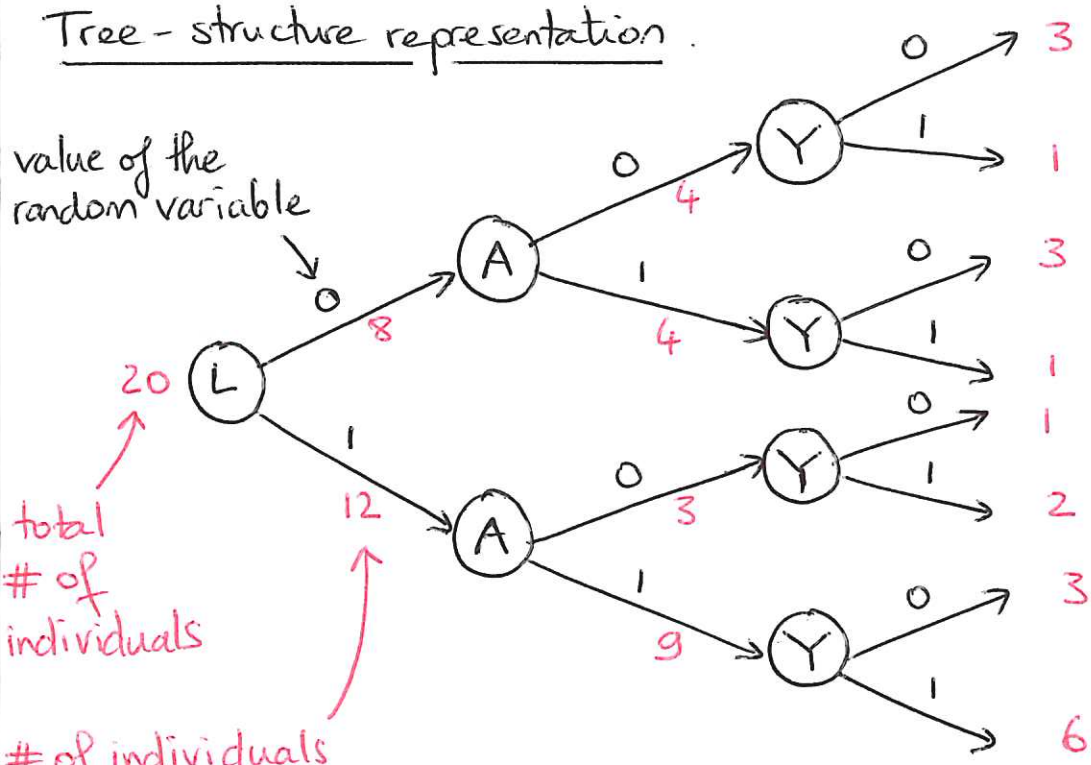
$$\mathbb{P}(Y^a=1) = \sum_\ell \boxed{\frac{1}{\mathbb{P}(A=a \mid L=\ell)}}\ \mathbb{P}(L=\ell, A=a, Y=1)$$

$\equiv w(a,\ell)$

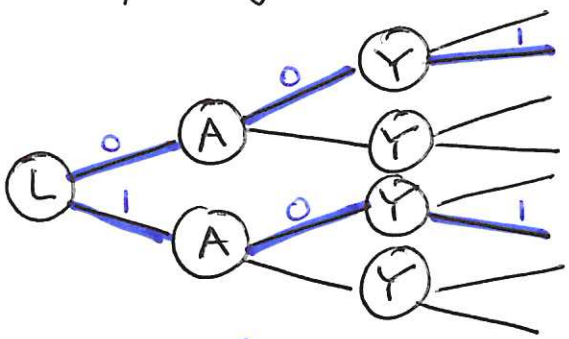
$$\boxed{\mathbb{P}(Y^a=1) = \sum_\ell w(a,\ell)\ \mathbb{P}(L=\ell)\mathbb{P}(A=a \mid L=\ell) \times \mathbb{P}(Y=1 \mid L=\ell, A=a)}$$

↑ We make sense of this expression on the next page, and in particular of the term $w(a,\ell)$.

Tree-structure representation

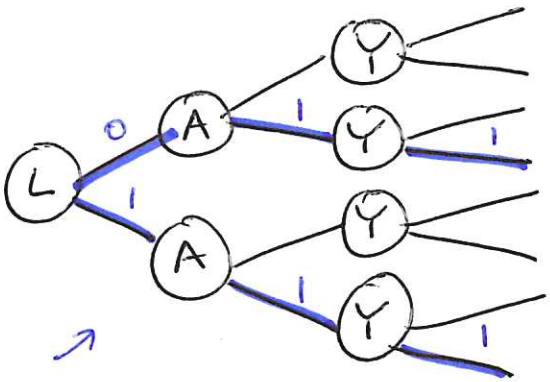


To compute $\mathbb{P}(Y^0=1)$, we need to look up branches corresponding to A=0 and Y=1



A situation where everyone (since counterfactual term) in the population remained untreated

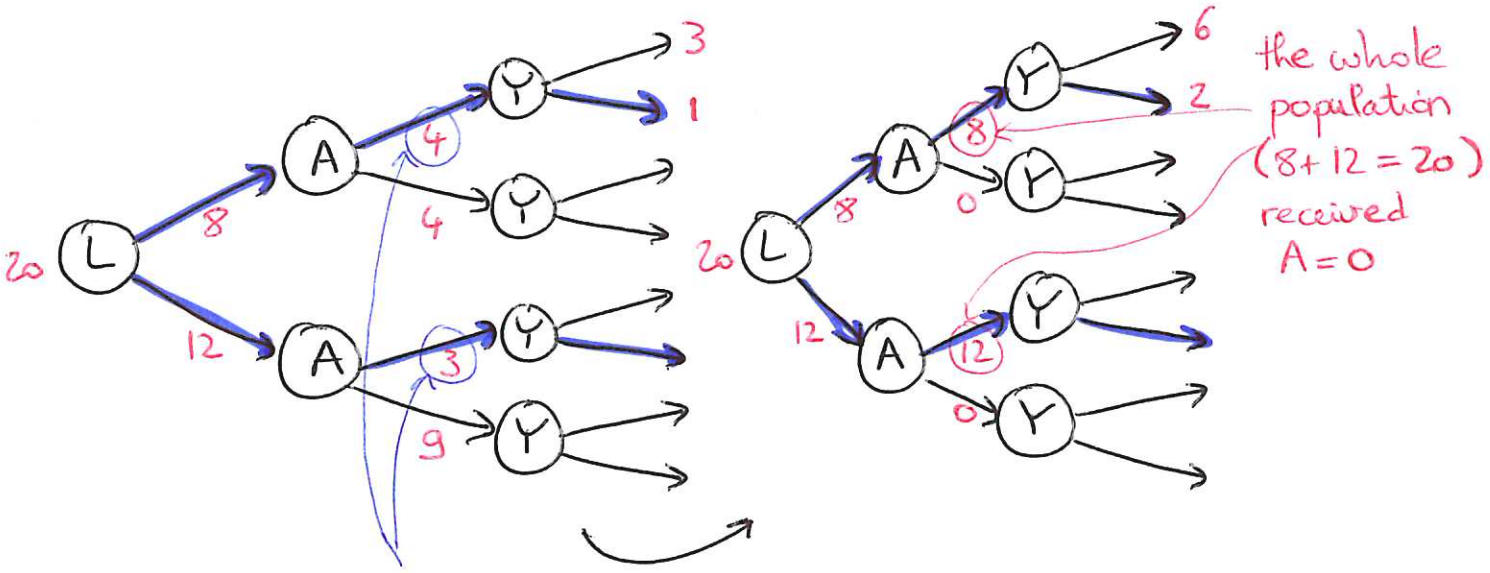
Similarly, to compute $\mathbb{P}(Y^1=1)$, we need to look up branches corresponding to A=1 and Y=1



a situation where everyone in the population had been treated

- Let's focus on $\mathbb{P}(Y^0 = 1)$.

We need to
- ↘ transfert the 4 individuals who received $A=1 \mid L=0$ to non treatment
- ↘ transfert the 9 individuals who received $A=1 \mid L=1$ to non treatment



The number of individuals is scaled up by $1/\mathbb{P}(A=0 \mid L=\ell) = w(0,\ell)$

$\Rightarrow$ Individuals within each branch $L=0$ and $L=1$ are swapped or exchanged, as if they were interchangeable. But this is precisely the conditional exchangeability condition $Y^a \perp A \mid L \ \forall a$; that allow us to draw causal conclusions from observed values (A, L, Y).

The scaling $w(0,\ell)$ translates all the way to the right hand side of the tree, where the number of individuals in each subbranch $(L=\ell, A=0, Y=1)$ is scaled up by $w(0,\ell)$.

$$\mathbb{P}(Y^0=1) = \sum_{\ell} \boxed{w(0,\ell)}\, \mathbb{P}(L=\ell, A=0, Y=1)$$

weight associated with obs $(L=\ell, A=0, Y=1)$

× Remark = Inverse Probability Estimator.

Consider a population of n individuals

$\mathcal{L}_n = \{(L_1, A_1, Y_1), \dots, (L_n, A_n, Y_n)\}$

- $\mathbb{P}(Y^1 = 1) = \sum_{\ell} w(1, \ell)\, \mathbb{P}(L = \ell, A = 1, Y = 1)$

$= \sum_{\ell, a, y} w(a, \ell)\, a\, y\, \mathbb{P}(L = \ell, A = a, Y = y)$

$= \mathbb{E}\{w(A, L)\, A\, Y\}$

$\simeq \frac{1}{n} \sum_{i=1}^{n} w(A_i, L_i)\, A_i Y_i$

each observation (L_i, A_i, Y_i) is weighted by $w(A_i, L_i)$.

- $\mathbb{P}(Y^0 = 1) = \sum_{\ell} w(0, \ell)\, \mathbb{P}(L = \ell, A = 0, Y = 1)$

$= \sum_{\ell, a, y} w(a, \ell)\, (1-a)\, y\, \mathbb{P}(L = \ell, A = a, Y = y)$

$= \mathbb{E}\{w(A, L)\, (1 - A)\, Y\}$

$\simeq \frac{1}{n} \sum_{i=1}^{n} w(A_i, L_i)\, (1 - A_i)\, Y_i$

$\Rightarrow \mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1)$ is estimated by

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{e(L_i)} - \frac{(1 - A_i) Y_i}{1 - e(L_i)} \right\}$$

, where $e(L_i) = \mathbb{P}(A_i = 1 \mid L_i)$

The Horvitz-Thompson estimator

Note that $\mathbb{E}\{\omega(A,L)A\} = \mathbb{E}\{\omega(A,L)(1-A)\} = 1$

$$\Rightarrow \mathbb{P}(Y^1=1) = \mathbb{E}\{\omega(A,L)AY\}$$

$$= \frac{\mathbb{E}\{\omega(A,L)AY\}}{\mathbb{E}\{\omega(A,L)A\}}$$

$$\approx \frac{\sum_{i=1}^{n} \omega(A_i,L_i)A_iY_i}{\sum_{i=1}^{n}\omega(A_i,L_i)A_i}$$

$$= \sum_{i\,|\,A_i=1} \left\{ \frac{\omega(A_i,L_i)}{\sum_{j\,|\,A_j=1}\omega(A_j,L_j)} \right\} Y_i$$

$= \omega'(A_i,L_i)$
(sum to 1)
(normalized weight)

& similarly for $\mathbb{P}(Y^0=1)$

$\Rightarrow$ Leads to the modified HT estimator of the ATE (Robins '98)

$$\widehat{ATE} = \sum_{i\,|\,A_i=1} \omega'(A_i,L_i)Y_i - \sum_{i\,|\,A_i=0} \omega'(A_i,L_i)Y_i .$$

This estimator can be derived as the solution of a weighted LS problem.

Consider the (saturated) model $Y = \beta_0 + \beta_1 A + \varepsilon$, $A \in \{0,1\}$ where ε is zero mean.

When weighting obs (A_i, Y_i, L_i) by $\omega(A_i,L_i)$, we remove confounding by L, and $\beta_1 = ATE$.

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \omega(A_i, L_i)(Y_i - \beta_0 - \beta_1 A_i)^2$$

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (X^t W X)^{-1} X^t W Y,$$

where $W = \underset{(n \times n)}{\operatorname{diag}}\{\omega(A_i, L_i)\}$

$$\underset{(n \times 2)}{X} = \begin{pmatrix} 1 & A_1 \\ \vdots & \vdots \\ 1 & A_n \end{pmatrix} \qquad \underset{(n \times 1)}{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

After calculations, $X^t W X = \begin{pmatrix} \sum \omega_i & \sum \omega_i A_i \\ \sum \omega_i A_i & \sum \omega_i A_i \end{pmatrix}$

$\omega_i = \omega(A_i, L_i)$.

$$X^t W Y = \begin{pmatrix} \sum_i \omega_i Y_i \\ \sum_i \omega_i A_i Y_i \end{pmatrix}$$

$$\Rightarrow \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \dfrac{\sum \omega_i Y_i (1 - A_i)}{\sum \omega_i (1 - A_i)} \\ \dfrac{(\sum \omega_i)(\sum \omega_i A_i Y_i) - (\sum \omega_i A_i)(\sum \omega_i Y_i)}{(\sum \omega_i A_i)(\sum \omega_i (1 - A_i))} \end{pmatrix}$$

It follows that $\hat{\beta}_0 = \dfrac{\sum_{i | A_i = 0} \omega(A_i, L_i) Y_i}{\sum_{j | A_j = 0} \omega(A_j, L_j)}$

$$\hat{\beta}_0 + \hat{\beta}_1 = \frac{\sum_{i | A_i = 1} \omega(A_i, L_i) Y_i}{\sum_{j | A_j = 1} \omega(A_j, L_j)}$$

$\Rightarrow \hat{\beta}_1$ is the modified HT estimator (variance usually smaller than HT)

Normalized weights are often preferred over original weights, especially when $w(a,\ell)^{-1} = \mathbb{P}(A=a \mid L=\ell)$ is close to 0 or 1, which typically increase the variance of the HT estimator.

→ Alternatively, truncate / trim weights to their 5% and 95% percentiles.

→ Another approach is to <u>stabilize</u> weights $w(a,\ell)$ by multiplying them by a quantity of the same order of magnitude as the denominator $\mathbb{P}(A=a \mid L=\ell)$. A commonly used value is $\mathbb{P}(A=a)$, leading to the <u>STABILIZED WEIGHTS</u>

$$\tilde{w}(a,\ell) := \frac{\mathbb{P}(A=a)}{\mathbb{P}(A=a \mid L=\ell)}$$

Note however that for a binary treatment $A \in \{0,1\}$ and saturated model $Y = \beta_0 + \beta_1 A + \varepsilon$, the weighted LS problem using weights $\tilde{w}$ instead of w leads to exactly the same solution $(\tilde{\beta}_0, \tilde{\beta}_1) = (\hat{\beta}_0, \hat{\beta}_1)$

since

$$\tilde{\beta}_0 = \frac{\sum_{i \mid A_i=0} \tilde{w}_i(A_i, L_i) Y_i}{\sum_{j \mid A_j=0} \tilde{w}_i(A_j, L_j)}$$

$$= \frac{\sum w_i(A_i, L_i) Y_i}{\sum w_i}$$

$$= \hat{\beta}_0$$

Since the extra term $\mathbb{P}(A=a)$ is common to all observations & cancels out

Stabilized weights must be used instead in more complex settings such as continuous or time-varying treatments.

× Remark: Instead of a saturated model $Y = \beta_0 + \beta_1 A + \varepsilon$ fit using weights $w(A, L)$, we may consider equivalently the same linear model for counterfactual statements $Y^a = \beta_0 + \beta_1 a + \varepsilon$. Models involving Y^a quantities are commonly referred to as Marginal Structural Models.

× Remark: Back to standardization. Recall that

$$\mathbb{E}\, Y^a = \sum_{\ell} \underbrace{\mathbb{E}(Y \mid A=a, L=\ell)}\, \mathbb{P}(L=\ell)$$

fit a regression / classification model to estimate the (mean of the) distribution of $Y \mid A=a, L=\ell$.
Once an estimate $\hat{\mathbb{E}}(Y \mid A=a, L=\ell)$ is available, we directly get an estimator of $\mathbb{E}\, Y^a$:

$$\mathbb{E}\, Y^a \approx \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbb{E}}(Y \mid A=a, L=L_i)$$

Summary: . IP weighting requires the modeling of $A \mid L$
. standardization requires —"— $Y \mid A, L$.

↑ Parametric IP weighting & standardization are expected to differ. They can only agree asymptotically either (i) when non-parametric or (ii) when the parametric modeling for IP weighting & standardization are both correct. Model misspecification will introduce some bias.

× Revisiting conditional exchangeability with g-estimation.

Under the assumption $Y^a \perp A \mid L$,

$$\mathbb{P}(A=1 \mid L=1) = \mathbb{P}(A=1 \mid L, Y^0) \quad \text{(logistic regression model)}$$

$$= \sigma^{-1}(\alpha_0 + \alpha_1 L + \alpha_2 Y^0) \qquad (\diamond)$$

(α_2 must be zero.)

↓

Cannot fit this model using data since Y^0 is not observed.

However, we can make the following additional assumption $\mathbb{E}(Y^a - Y^0) = \beta_1 a$ (*)

$$\Leftrightarrow \mathbb{E}Y^0 = \mathbb{E}Y^a - \beta_1 a \quad \text{(consistency)}$$

$$\Leftarrow Y^0 = Y - \beta_1 A + \varepsilon \;, \; \mathbb{E}\varepsilon = 0.$$

The expression $Y^0 \approx Y - \beta_1 A$ can be used to reconstruct the counterfactual Y^0 under assumption (*).

→ Perform a grid search over β such that the reconstructed Y^0 yield a zero α_2 coefficient in $(\diamond)$.

× Remark: Condition (*) can be relaxed / weakened / generalized to include effect modifiers,

Ex: $\mathbb{E}(Y^a - Y^0 \mid L) = \beta_1 a + \beta_2 aL$, (aka Structural Nested Model (SNM))

which can in turn be used together with $(\diamond)$ to recover the causal effect within each stratum $\{L = l\}$.

CI = OBSERVATIONAL STUDIES

In conditional randomized experiments, we showed (p.6) that

$$\mathbb{P}(Y^a=1) = \sum_{\ell} w(a,\ell)\,\mathbb{P}(L=\ell)\,\mathbb{P}(A=a \mid L=\ell) \times \mathbb{P}(Y=1 \mid L=\ell, A=a).$$

To derive this expression, we made use of the following.

a) **CONSISTENCY** If $A=a$, then $Y=Y^a$

$$\mathbb{P}(Y^a=1 \mid L=\ell, A=a) = \mathbb{P}(Y=1 \mid L=\ell, A=a)$$

b) **CONDITIONAL EXCHANGEABILITY** $Y^a \perp A \mid L$

$$\mathbb{P}(Y^a=1 \mid L=\ell) = \mathbb{P}(Y^a=1 \mid L=\ell, A=a)$$

c) **POSITIVITY** $\mathbb{P}(A=a \mid L=\ell) > 0 \quad \forall \ell \mid \mathbb{P}(L=\ell) > 0$

since $w(a,\ell) = 1/\mathbb{P}(A=a \mid L=\ell)$

b) and c) only are referred to as **IGNORABILITY** (weak & strong)

↓

Need full exchangeability

$(Y^0, Y^1) \perp A \mid L$

Rosenbaum & Rubin (1983)

Refferred to as **IDENTIFIABILITY** conditions. They hold by design.

Since "Causal Effects are identifiable".

⇒ To analyse observational data as if treatment had been randomly assigned conditionally on measured covariates L,

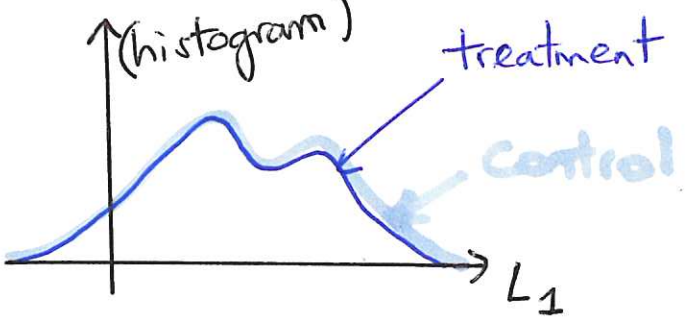
we need these three properties to hold.

↳ these become <u>assumptions</u> and it is our responsability to assess to which extent they hold / break down.

↘ <u>conditional exchangeability</u>.

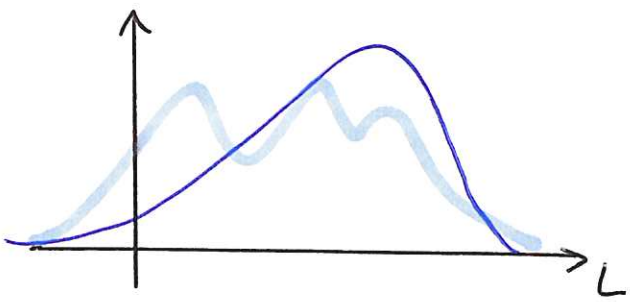
In a randomized experiment, $Y^a \perp A$, and all (measurable) features you can think of are balanced in the treatment and control group. This follows from the fact that the two subpopulations are representative samples of the whole population.

Practically, denoting $L_1, L_2, \ldots$ these covariates, a histogram of L_i conditionally on $A=0$ and $A=1$ should "look alike".



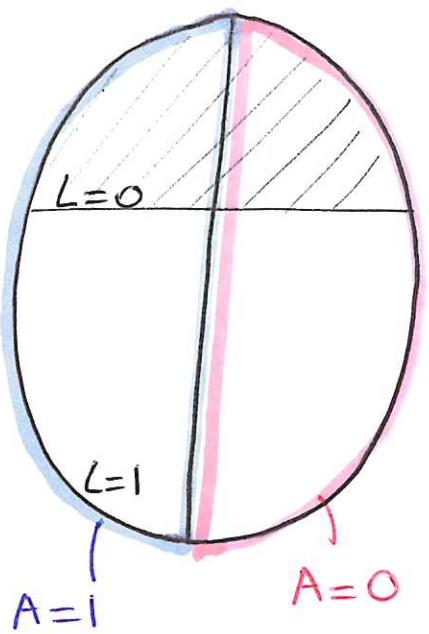
The same ideas apply in a conditional randomized experiment.

1) It is expected by design that L is unbalanced between the control and the treated groups, since $Y^a \perp A \mid L$.

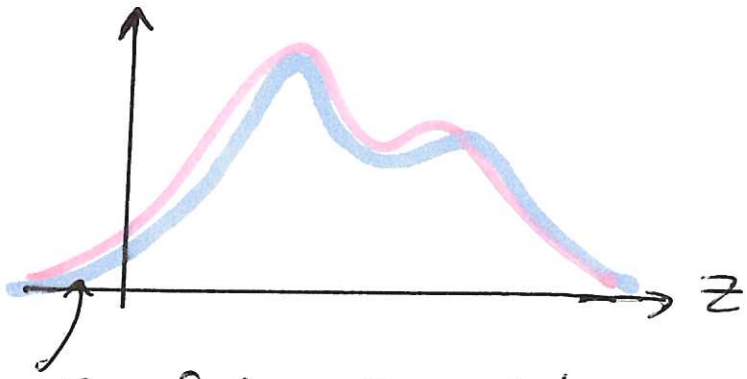


2) Within each stratum $L=l$, all other (measurable)

features Z must be balanced.



In each subpopulation $L=l$, plot the distribution (histogram) of another measured feature Z, conditionally on $A=a$



Conditionally on $L=l$, all other features Z must be balanced for $Y^a \perp A \mid L$ to hold.
If the histograms do not superpose, this is an indication that $Y^a \not\perp A \mid L$.

However...

- you may not be able to verify this in practice since you may not have collected the required data
- the set of features is unlimited & cannot be exhaustive in practice.
- some features are not measurable.

Conclusions:

1) Expert knowledge is needed.
2) We can easily gather information that conditional exchangeability breaks down, but one can never be fully sure it holds ⇒ risky task.

↘ positivity

In a randomized experiment, to compute the average causal effect $\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$ using $\mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=1 \mid A=0)$, we need to collect data in both the treatment & the control groups.

The same idea applies in conditional randomized experiments. In order to estimate the average causal effect, the learning sample must be rich enough / contain enough information about the two groups. The mathematical way of translating this is that

$$\forall \ell \text{ s.t } \mathbb{P}(L=\ell) > 0 \Rightarrow \mathbb{P}(A=a \mid L=\ell) > 0$$

↑ the positivity assumption / condition ↑

Let $\mathcal{A} = \{0, 1\}$ (binary treatment)

$\mathcal{L} = \{0, 1\}$ (binary feature)

$$\mathcal{L}(a) := \{\ell \in \mathcal{L} \mid \mathbb{P}(L=\ell) > 0 \ \& \ \mathbb{P}(A=a \mid L=\ell) > 0\}$$

When positivity holds, $\mathcal{L}(0) = \mathcal{L}(1) = \mathcal{L}$

When positivity does not hold, $\mathcal{L}(0) \neq \mathcal{L}(1)$.

Recall the Horvitz-Thompson estimator (p. 9)

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{A_i Y_i}{e(L_i)} - \frac{(1-A_i) Y_i}{1-e(L_i)} \right\}$$

$$\approx \mathbb{E}\{w(A,L) A Y\} - \mathbb{E}\{w(A,L)(1-A) Y\}$$

$e(\ell) = \mathbb{P}(A=1 \mid L=\ell)$

$w(a, \ell) = 1 / \mathbb{P}(A=a \mid L=\ell)$

We compute next $\mathbb{E}\{w(A,L)AY\}$ and $\mathbb{E}\{w(A,L)(1-A)Y\}$ when positivity holds & does not hold.

$$\mathbb{E}\{w(A,L)AY\} = \sum_{y\in\{0,1\}} \sum_{a\in A} \sum_{\ell\in\mathcal{L}(a)} w(a,\ell)\, a\, y \times \mathbb{P}(L=\ell, A=a, Y=y)$$

the support of (L,A,Y)

$$= \sum_{\ell\in\mathcal{L}(1)} w(1,\ell)\, \mathbb{P}(L=\ell, A=1, Y=1)$$

terms vanish for $y=0$ and $a=0$

$$= \sum_{\ell\in\mathcal{L}(1)} \mathbb{P}(Y=1 \mid L=\ell, A=1)\, \mathbb{P}(L=\ell)$$

same as p. 6

$$= \sum_{\ell\in\mathcal{L}(1)} \mathbb{P}(Y^1=1, L=\ell)$$

$$= \sum_{\ell\in\mathcal{L}} \mathbb{P}(Y^1=1, L=\ell, L\in\mathcal{L}(1))$$

$$= \sum_{\ell\in\mathcal{L}} \mathbb{P}(Y^1=1, L=\ell \mid L\in\mathcal{L}(1))\, \mathbb{P}(L\in\mathcal{L}(1))$$

LTP

$$= \mathbb{P}(L\in\mathcal{L}(1))\, \mathbb{P}(Y^1=1 \mid L\in\mathcal{L}(1)),$$

and similarly for $\mathbb{E}\{w(A,L)(1-A)Y\}$.

⇒ In general, the Horvitz-Thompson estimator is an appr. of

$$\mathbb{P}(L\in\mathcal{L}(1))\, \mathbb{P}(Y^1=1 \mid L\in\mathcal{L}(1)) - \mathbb{P}(L\in\mathcal{L}(0))\, \mathbb{P}(Y^0=1 \mid L\in\mathcal{L}(0))$$

$\neq \mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$ unless $\mathcal{L}(0)=\mathcal{L}(1)=\mathcal{L}$, i.e. unless the positivity assumption holds.

↘ consistency

If $A = a$, then $Y_i = Y_i^a$.

↑

First, note that the treatment $A = a$ must be well defined (e.g. in experimental studies, the experimenter has control of the treatment, which must be administered the same way to all participants)

↘ usuall ok in medical studies

↘ less ok in biological & social sciences.

↓ In obs studies, there may be different versions of the way a treatment is administered.

Let r = version of a treatment.

We collect/analyse observations associated with all patients that received ≠ versions of treatment a.

Then the consistency property holds provided the following assumption holds

If $A = a$, then $Y_i = Y_i^{a,r}$ $\forall r$

↑

Potential Outcome associated with version r of treatment a

↑

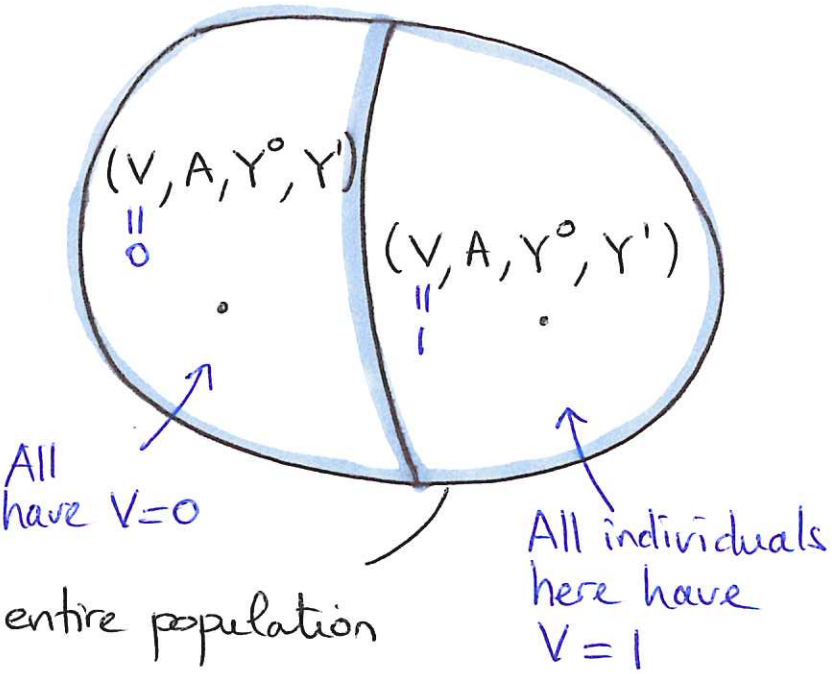
In other words, $Y_i^{a,r} = Y_i^{a,r'}$ $\forall r, r'$.

i.e. it is assumed that all versions of the treatment a produce the same causal effect.

CI = EFFECT MODIFICATION

We are interested here in the average causal effect about subsets of a population.

Let $V \in \{0, 1\}$ = binary indicator splitting an entire population into two subpopulations (e.g. Male vs Female)



We may be interested in comparing

$$\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$$

with

$$\mathbb{P}(Y^1=1 \mid V=1) - \mathbb{P}(Y^0=1 \mid V=1)$$

and

$$\mathbb{P}(Y^1=1 \mid V=0) - \mathbb{P}(Y^0=1 \mid V=0)$$

Def: V is a **MODIFIER** of the effect of A on Y when the average causal effect of A on Y varies across levels of V.

↑ There can be a multiplicative but not additive effect modification by V.

To identify effect modification, a stratified analysis is a natural way to go:

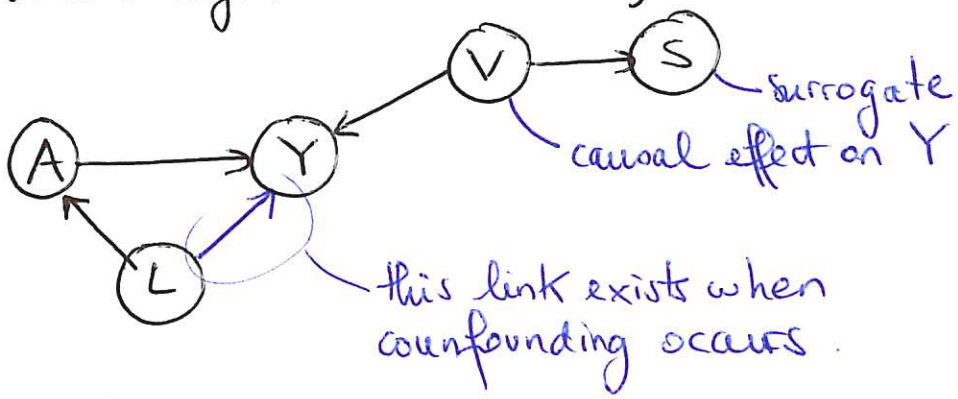
- In an RCT:
  1) Partition the population into strata $V = v$
  2) For each strata, estimate the average causal effect $\mathbb{P}(Y^1 = 1)_{V=v} - \mathbb{P}(Y^0 = 1)_{V=v}$ using $\mathbb{P}(Y=1 \mid A=1)_{V=v} - \mathbb{P}(Y=1 \mid A=0)_{V=v}$.

- In a conditional randomized experiment,
  1) —"—
  2) Use standardization / IP weighing under the exchangeability assumption $Y^a \perp A \mid L$

Remarks:
- No claims are being made about the causal mechanisms. The variable V may be a SURROGATE effect modifier, while the CAUSAL effect modifier is unknown.

  Eg: V = carying a lighter
  Y = lung cancer

- graphical representation (difference between L, V, and a surrogate variable S)



- If V modifies the effect of treatment A on the outcome Y, then the average causal effect will differ btw populations with ≠ prevalence of V. (→ question of TRANSPORTABILITY of causal

inferences across populations.)

- Different forms of adjustments.
  - ↳ stratification
  - ↳ standardization / IP weighting
  - ↳ matching

We review/introduce/compare these three techniques next.

Suppose we observe $\mathcal{L}_n = \{(L_i, A_i, Y_i)\}_{i=1:n}$, where each $(L_i, A_i, Y_i) \overset{d}{\sim} (L, A, Y)$; $L \in \{0,1\}$, $A \in \{0,1\}$, $Y \in \{0,1\}$

(i-th unit)

a) STRATIFICATION splits the data into two subpopulations, according to the strata $L=0$ and $L=1$.

Within each stratum $L=l$, compute

$$\mathbb{P}(Y=1 \mid A=a, L=l)$$

consistency ↳ $= \mathbb{P}(Y^a=1 \mid A=a, L=l)$

under the assumption of conditional exchangeability ↳ $= \mathbb{P}(Y^a=1 \mid L=l)$; $Y^a \perp A \mid L=l$

Under these two conditions,

$$\mathbb{P}(Y^1=1 \mid L=l) - \mathbb{P}(Y^0=1 \mid L=l)$$
$$= \mathbb{P}(Y=1 \mid A=1, L=l) - \mathbb{P}(Y=1 \mid A=0, L=l)$$

↑ One average causal estimate per stratum is computed.
(generalizes well to non-binary L)

b) STANDARDIZATION / IP WEIGHTING.

Adjust for L using the techniques detailed on page 6.

Under identifiability conditions, these provide an estimate of the ATE $\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$

Only one estimate; for the entire population ($\neq$ stratification; one per stratum).

× Remarks: (i) The risk difference and risk ratio for the entire population lies in the convex hull of the causal estimates computed within each stratum.

proof: ↘ $\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$

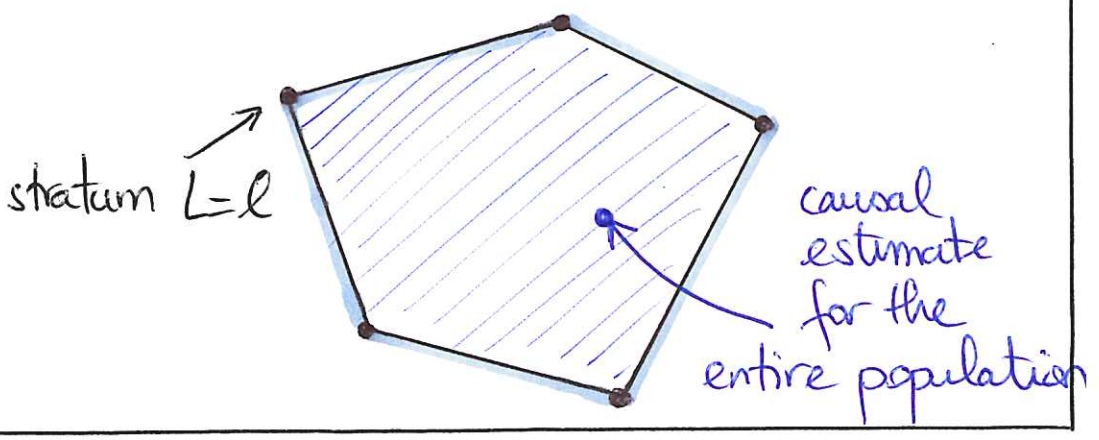
$$= \sum_\ell \left\{ \mathbb{P}(Y^1=1 \mid L=\ell) - \mathbb{P}(Y^0=1 \mid L=\ell) \right\} w_\ell$$

where $w_\ell = \mathbb{P}(L=\ell) \geq 0$; $\sum_\ell w_\ell = 1$.

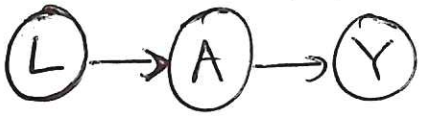
↘
$$\frac{\mathbb{P}(Y^1=1)}{\mathbb{P}(Y^0=1)} = \sum_\ell \left\{ \frac{\mathbb{P}(Y^1=1 \mid L=\ell)}{\mathbb{P}(Y^0=1 \mid L=\ell)} \right\} w_\ell ,$$

where

$$w_\ell = \frac{\mathbb{P}(Y^0=1 \mid L=\ell)\,\mathbb{P}(L=\ell)}{\mathbb{P}(Y^0=1)} \geq 0 \; ; \; \sum_\ell w_\ell = 1$$



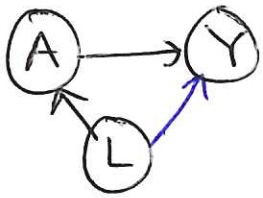
(ii). If L is not an effect modifier, then all average causal effects (within each stratum + at the entire population level) are equal.



$L \perp (Y^0, Y^1)$ ← def p. 16

Then $\mathbb{P}(Y^a = 1 \mid L = l) = \mathbb{P}(Y^a = 1)$; and the result follows from the expressions on page 19.

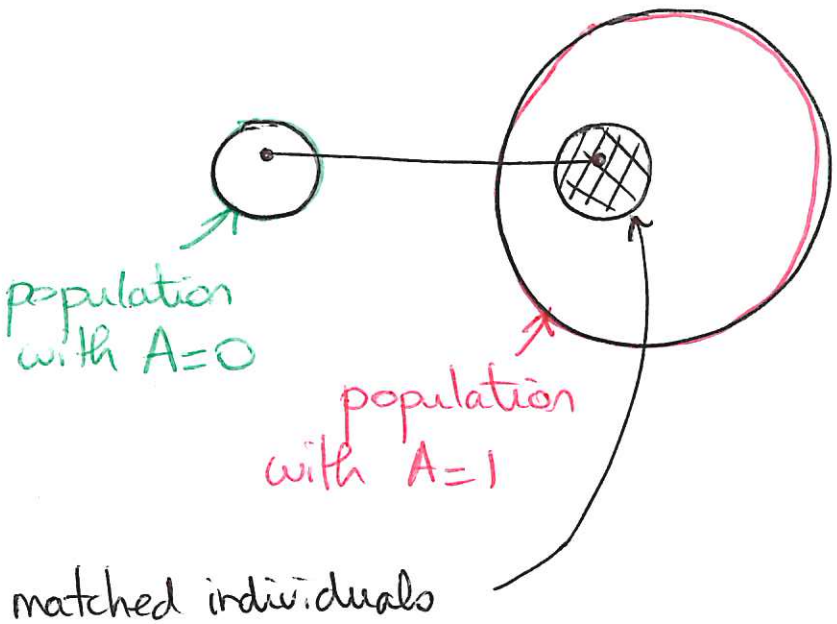
- If L is an effect modifier, then the average causal effects will be different



c) MATCHING.

The idea is to rebalance the distribution of the variable L within the two groups (to make them comparable)

- Suppose that we take the untreated (A=0) group as the reference group. Then, for each individual in A=0 with L=l, pick one / match it with an individual in A=1 with the same L=l.



By doing so, the two groups ◯ and ▨ are comparable (same distribution of L)

Provided conditional exchangeability holds in the original population $Y^a \perp A \mid L = l$, then

unconditional exchangeability holds for ◯ & ▨.

Since the reference population is $A=0$, a direct comparison of the two populations ◯ & ▨ yields an estimate of $\mathbb{P}(Y^1=1 \mid A=0) - \mathbb{P}(Y^0=1 \mid A=0)$

$$= \mathbb{E}(Y^1 - Y^0 \mid A=0)$$

$$= \text{ATNT}$$

Average Treatment on the Non-Treated

The matching can be made by considering the group $A=1$ as the reference group; in which case we estimate the $\text{ATT} = \mathbb{E}(Y^1 - Y^0 \mid A=1)$.

× Remark: ATT and ATNT require the computation of quantities like $\mathbb{P}(Y^a=1 \mid A=a')$ $a \neq a'$. It can be done under the usual consistency & conditional exch. assumptions.

$$\mathbb{P}(Y^a=1 \mid A=a') = \sum_\ell \mathbb{P}(Y^a=1 \mid A=a', L=\ell)\, \mathbb{P}(L=\ell \mid A=a')$$

consistency ↳

$$= \sum_\ell \mathbb{P}(Y^a=1 \mid A=a, L=\ell)\, \mathbb{P}(L=\ell \mid A=a')$$

$Y^a \perp A \mid L$ ↳

$$= \sum_\ell \mathbb{P}(Y=1 \mid A=a, L=\ell)\, \mathbb{P}(L=\ell \mid A=a')$$

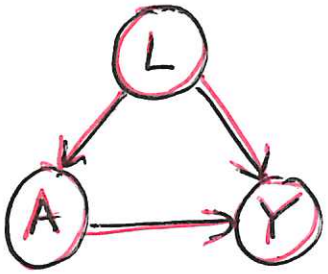
can be computed from data.

▨

CI = GRAPHICAL REPRESENTATION OF CAUSAL EFFECTS

A Structural Causal Model (SCM) $\mathcal{C} = (\underset{\parallel}{X}, \mathbb{P}_\varepsilon)$ consists of a collection of d $(X_1, \ldots, X_d)$ assignments $X_j = f_j(Pa_j, \varepsilon_j)$, where $Pa_j \subseteq \{X_1, \ldots, X_d\}$ are the parents of X_j & a joint distribution $\mathbb{P}_\varepsilon$ over the noise variables $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_d)$. (unobserved)

- Ex: Conditional RCT



$X_1 = L$; $Pa_1 = \emptyset$
$X_2 = A$; $Pa_2 = \{L\}$
$X_3 = Y$; $Pa_3 = \{A, L\}$

↑ A Directed Acyclic Graph (DAG) with 3 nodes & 3 edges

[Together with structural equations $\begin{cases} X_1 = f_1(Pa_1, \varepsilon_1) \\ X_2 = f_2(Pa_2, \varepsilon_2) \\ X_3 = f_3(Pa_3, \varepsilon_3), \end{cases}$

it constitutes a causal DAG]

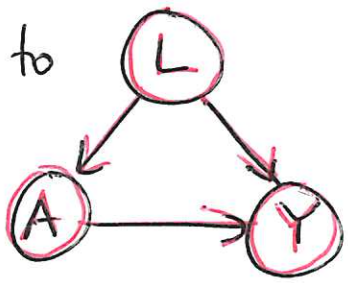
↑ These allow us to link the DAG to obs & attach a probabilistic structure to it.

The structural equations show that time flows from ⟶ to i.e. presence of a direct causal effect.

↘ No cycle = a variable cannot cause itself, either directly or thrugh another variable

↘ Conditionally on its direct causes (Pa), a variable is independent of any other variable for which it is not the cause.

Back to



$$L = X_1 = f_1(\varepsilon_1) = \varepsilon_1$$
$$A = X_2 = f_2(L, \varepsilon_2) = \alpha A + \varepsilon_2$$
$$Y = X_3 = f_3(A, L, \varepsilon_3) = \beta A + \gamma L + \varepsilon_3$$

Assuming linear structural equations.

The joint distribution $\mathbb{P}_\varepsilon$ over $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3)$ induces a joint distribution $\mathbb{P} = \mathbb{P}_X$ over the vector $X = (X_1, X_2, X_3)$.

- DAGs & Potential Outcomes.

This DAG representation does not naturally include the underlying conterfactual variables on the graph : the link between Y and (Y^0, Y^1) remain hidden.

Richardson & Robin (2013) found a graphical representation that explicitely include conterfactuals in the graph (Single World Intervention Graph SWIG); see next chapter.

However, the DAG representation can include potential outcomes

provided we intervene : any variable X_j on the DAG can be intervened on & set to a particular value x_j. (assumption)

$$\begin{cases} L = X_1 = f_1(\varepsilon_1) \\ A = X_2 = f_2(L, \varepsilon_2) \\ Y = X_3 = f_3(A, L, \varepsilon_3) \end{cases} \xrightarrow{\text{intervention}} \begin{cases} L^a = X_1^a = f_1(\varepsilon_1) \\ A = X_2 = a \\ Y^a = X_3^a = f_3(a, L, \varepsilon_3) \end{cases}$$

↑ ↑

superscript indicating we intervened & set A to the value a ; irrespectively of the Structural Graph.

The intervention modifies the joint distribution over (X_1, X_2, X_3)

Pearl's do calculus "do $(A=a)$"

↓

Joint distribution $\mathbb{P} = \mathbb{P}_X$

(observational studies)

Intervention distribution

$\mathbb{P}_{do(A=a)}$

(randomized experiments)

x Remark = some invariants

$$\begin{cases} \mathbb{P}(L=\ell) = \mathbb{P}_{do(A=a)}(L^a=\ell) \\ \mathbb{P}(Y=1 \mid A=a, L=\ell) = \mathbb{P}_{do(A=a)}(Y^a=1 \mid A=a, L=\ell) \end{cases}$$

These invariants can be used to recover average causal effects :

$$\mathbb{P}_{do(A=a)}(Y^a=1) = \sum_{\ell} \mathbb{P}_{do(A=a)}(Y^a=1, A=a, L^a=\ell)$$

↑ (what we want)

$$= \sum_l \mathbb{P}_{do(A=a)}(Y^a=1 \mid A=a, L=l) \times \mathbb{P}_{do(A=a)}(A=a, L=l)$$

$$\mathbb{P}(Y=1 \mid A=a, L=l) \qquad \text{(invariants)} \qquad \mathbb{P}(L=l)$$

$$= \sum_l \mathbb{P}(Y=1 \mid A=a, L=l)\, \mathbb{P}(L=l)$$

↑ the STANDARDIZATION expression.

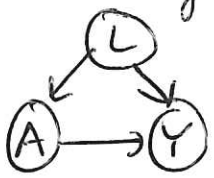
⇒ The techniques seen before have a graphical equivalence. Assumptions of positivity, consistency & conditional exchangeability must also be interpreted using DAGs.

↳ positivity = arrow from L to A is not deterministic

↳ consistency = arrow from A to Y corresponds to a possibly hypothetical but unambiguous intervention.

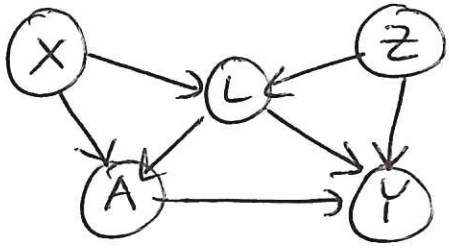
↳ cond. exch = see next chapter. ■

- We recovered the average causal effect in a simple example where



(equivalent to standardization, see above)

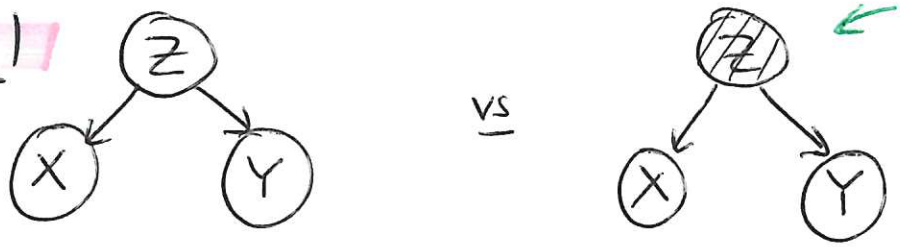
For example, how should we generalize the expression above when the SCM looks something like



To answer this question, we must study indpce & conditional indpce statements in DAGs for some simple structures. We identify 3 of them.

• Conditional Independence & Graphs.

× Example 1



← an observed node is darkened

$$p(x,y,z) = p(x|z)p(y|z)p(z)$$

— Are X and Y independent? —

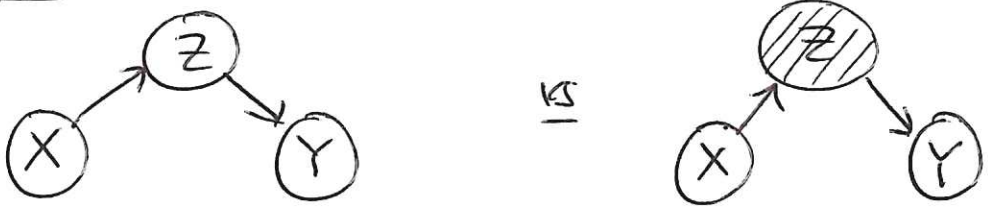
$$p(x,y) = \sum_z p(x|z)p(y|z)p(z) \neq p(x)p(y) \text{ in general}$$

$$p(x,y|z) = p(x|z)p(y|z)$$

$X \not\perp Y \mid \emptyset$ $\qquad$ $X \perp Y \mid Z$

We say that Z blocks the path btwn X and Y.

× Example 2



$$p(x,y,z) = p(y|z)p(z|x)p(x)$$

— Are X and Y independent? —

$$p(x,y) = \sum_z p(y|z)p(z|x)p(x)$$
$$= p(x)\,p(y|x)$$
$$\neq p(x)p(y) \text{ in general}$$

$$p(x,y|z) = \frac{p(x,y,z)}{p(z)} = p(x|z)\,p(y|z)$$

$X \perp Y \mid \emptyset$ $\qquad$ $X \perp Y \mid Z$

Z blocks the path btw X & Y.

× Example 3

$$p(x,y,z) = p(x)p(y)p(z|x,y)$$

— Are X and Y independent ? —

$$p(x,y) = \sum_z p(x)p(y)p(z|x,y) = p(x)p(y) \checkmark$$

$$p(x,y|z) = \frac{p(x)p(y)p(z|x,y)}{p(z)} \neq p(x|z)p(y|z)$$

$X \perp Y \mid \emptyset$ $\quad$ $X \not\perp Y \mid Z$

We say that Z is a collider

> Z blocks the path between X and Y
> $\Leftrightarrow$ Z is tail-to-tail or head-to-tail wrt that path

This leads us to the following definition

• Def d-separation

A path is blocked by a set of nodes Z iff

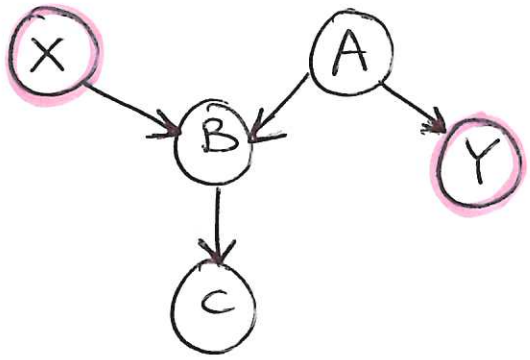
(i) The path contains a chain of nodes $A \to B \to C$ or a fork $A \leftarrow B \to C$ such that the middle node B is in Z (B is conditioned on)

(ii) The path contains a collider $A \to B \leftarrow C$ such that the collisionnode B is not in Z ; and no descendent of B is in Z.

↳ If Z blocks every path between two nodes X & Y,

then X & Y are d-separated by Z, and thus are independent conditionally on Z.

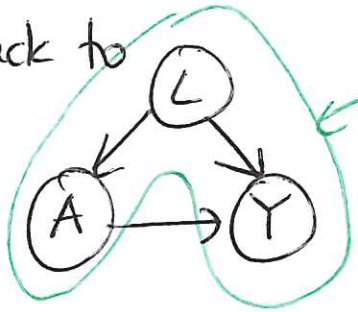
Ex:



- Is $X \perp Y \mid A$?
  Conditioning on A blocks the path $X \rightarrow B \leftarrow A \rightarrow Y$. In addition, B (a collider) & its descendent C remain unobserved $\Rightarrow$ X and Y are d-separated by $\{A\}$.
- Is $X \perp Y \mid C$?
  B is a collider and C is a descendent of B $\Rightarrow$ X and Y are *not* d-separated by $\{C\}$. ∎

Conditional independence for these three building blocks will allow us to identify which set of variables in a DAG we need to condition on (≡ adjust for) in order to compute/identify the average causal effect of A on Y.

Back to

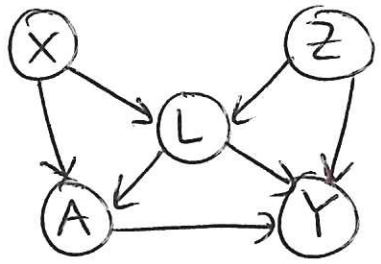


We are interested in the path $A \leftarrow L \rightarrow Y$

We know that in this path, $A \not\perp Y$; while $A \perp Y \mid L$

⇒ Provided L is observed, A and Y are independent along the path $A \leftarrow L \rightarrow Y$ (& this is the reason why in the standardization formula, we sum (& observe) over $L = l$

We now have the tools to answer the question at the bottom of page 25 :



There are 4 paths to consider:

$A \leftarrow X \rightarrow L \leftarrow Z \rightarrow Y$ ——— (i)

$A \leftarrow L \rightarrow Y$ ——— (ii)

$A \leftarrow L \leftarrow Z \rightarrow Y$ ——— (iii)

$A \leftarrow X \rightarrow L \rightarrow Y$ ——— (iv)

We see from (i) that L is a collider so the set $\{L\}$ does not block those paths.

However, $\{L, X\}$ blocks all the 4 paths

Also $\{L, Z\}$ and $\{L, Z, X\}$ ■

We get back to this more extensively in subsequent chapters, where we discuss various types of **bias** preventing us to compute the average causal effect.

We say that there is a systematic bias whenever

$$\mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1) \neq \mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=1 \mid A=0);$$

which happens under a lack of exchangeability.

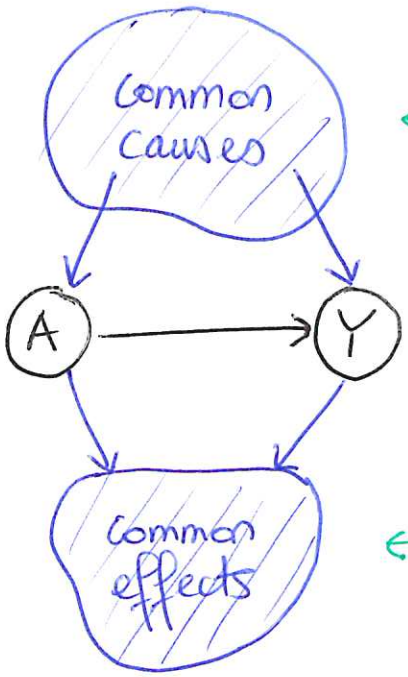
We will consider

- confounding (due to common causes)
- selection bias (cond. on common effects)
- measurement error.

CI = CONFOUNDING

- Task = estimate the causal effect from A to Y

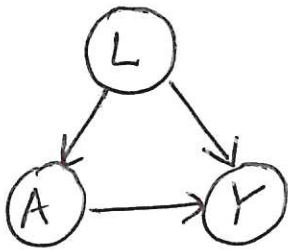
DAG = + Markov IP on top of it

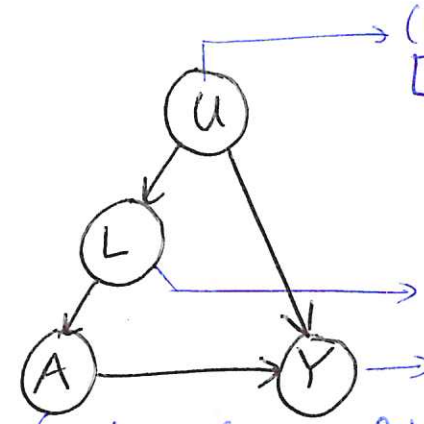


← In this chapter, we are interested in the structure of confounding i.e. bias due to common causes of treatment & outcome

← Assume here there are no common effects.

- Examples



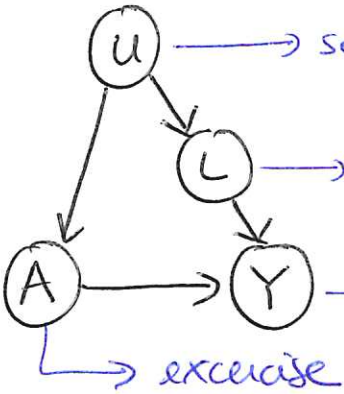


→ (unmeasured) atherosclerosis [build up of fat + cholesterol on artery walls; potentially causing arteries to narrow & blocking blood flow]

→ heart disease

→ stroke

→ drug (more likely to be prescribed to individuals with a certain condition)

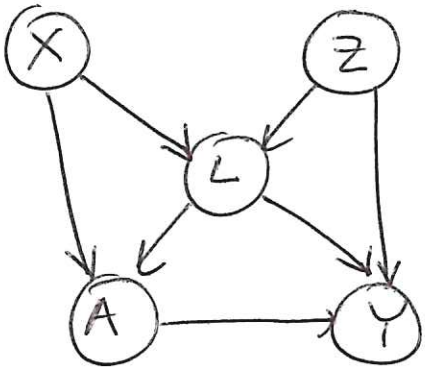


→ social factors

→ cigarette / smoking

→ death

→ exercise



Theorem (backdoor criterion)

A set of nodes L satisfy the backdoor criterion relative to $A \to Y$ if

(i) no node in L is a descendent of A
(no common causes)

(ii) L blocks every path between A and Y that contain an arrow into A.

If a set of nodes L satisfy the backdoor criteria for $A \to Y$, the causal effect from A to Y is $\mathbb{P}(Y^1 = 1) - \mathbb{P}(Y^0 = 1)$, where

$$\mathbb{P}(Y^a = 1) = \sum_{\ell} \mathbb{P}(Y = y \mid L = \ell, A = a)\, \mathbb{P}(L = \ell) \qquad (*)$$

↑ $\Leftrightarrow$ all backdoor paths between A and Y are d-separated by the set of nodes L

× Sketch of proof / intuition.

Expression (*) requires conditional exchangeability $Y^a \perp A \mid L$; see page 6. It remains to prove that

L satisfies the backdoor criterion for $A \to Y$ $\Rightarrow$ conditional exch. holds $Y^a \perp A \mid L$

↘ follows from the following lemma =

Lemma = The following statements are equivalent:

(a) Markov factorization property

$$p(x) = \prod_{j=1}^{d} p(x_j \mid pa_j)$$

(b) global Markov property

X_i & X_j are d-separated by L $\Rightarrow$ $X_i \perp X_j \mid L$

(c) local Markov property

Each variable is independent of its non-descendents given its parents.

$\Rightarrow$ Intervening on the graph & removing all incoming arrows into A does not change the distribution of (A, Y) given L, whether $A = a$ or $do(A = a)$.

$\Rightarrow$ Causal effects can be identified & computed from observational conditional probabilities via the backdoor formula (*)

- Remark The equivalence

L satisfies the backdoor criterion for $A \to Y$ $\Leftrightarrow$ cond. exchangeability holds $Y^a \perp A \mid L$

holds under the additional assumption of faithfulness

Faithfulness

X_i and X_j are d-sep by L $\Leftarrow$ $X_i \perp X_j \mid L$

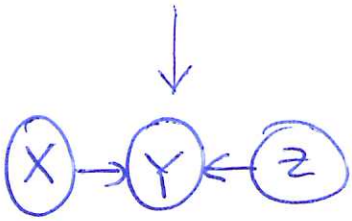
↑ Reverse direction of the global Markov property (top of page)

A nice assumption since under Markov + Faithfulness, we have a 1-1 relation between statements about graphs & probability distributions (↳ enables Causal discovery)

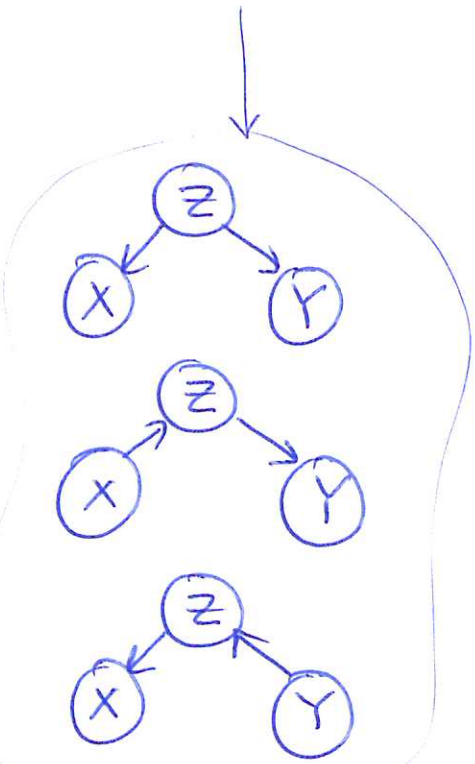
- Ex = Assume that $\mathbb{P}$ is Markov & Faithful w.r.t. its DAG.
Recover all DAGs compatible with the following complete list of conditional indpce statements.

(i) $X \perp Z$

(X, Y, Z)



(ii) $X \perp Y \mid Z$



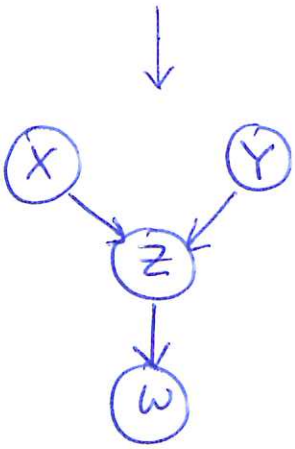
(iii) $X \perp Y$

$X \perp W \mid Z$

$Y \perp W \mid Z$

$X \perp W \mid Z, Y$

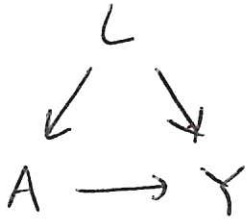
$Y \perp W \mid Z, X$



Encode the same set of d-separation. DAGs are called Markov Equivalent.

- Remark = The backdoor criterion has a nice visual translation using Single-World Intervention Graphs (SWIG) that include counterfactuals as nodes
↳ The graph representing counterfactual worlds is created by a single intervention.

Ex: original DAG (no counterfactual notation)



SWIG

$A \backslash a \longrightarrow Y^a$ — counterfactual notation since the outcome variable Y is a descendent of A

— we intervene on A and do(A=a)

— two separate nodes

- a inherits all nodes coming out of A in the original DAG.
- A inherits all nodes incoming into A in the original DAG.

In this representation, the backdoor path $A \leftarrow L \rightarrow Y^a$ is d-separated by L $\Rightarrow Y^a \perp A \mid L$. ∎

To summarize, there are two classes of methods to consider in the presence of confounders L:

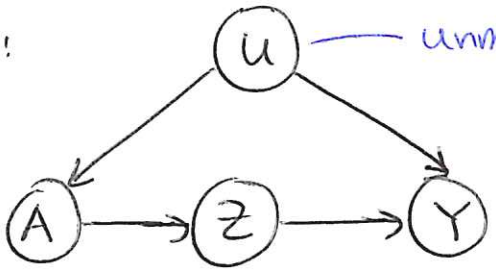
(a) Methods requiring conditional exchangeability
- G-methods (IP weighting / stand / backdoor)
- stratification based methods / matching

(b) Methods not requiring cond. exchangeability
- front-door criterion
- instrumental variables
- difference-in-difference (DiD)

We introduce these next.

- **Front-door-criterion.**
  - an alternative to the back-door criterion.
  - typically when an additional variable Z fully mediates the effect of A on Y & that share no unmeasured causes with either A or Y.

Ex:



(•) Since $A \leftarrow U \rightarrow Y \leftarrow Z$ is blocked by Y, (Y is a collider)

A and Z are independent

$$\Rightarrow \mathbb{P}_{do(A=a)}(Z=z) = \mathbb{P}(Z=z \mid A=a)$$

(••) Since $Z \leftarrow A \leftarrow U \rightarrow Y$ is blocked cond. on A, the backdoor criterion relative to $Z \rightarrow Y$ applies:

$$\mathbb{P}_{do(Z=z)}(Y=y) = \sum_a \mathbb{P}(Y=y \mid Z=z, A=a)\,\mathbb{P}(A=a$$

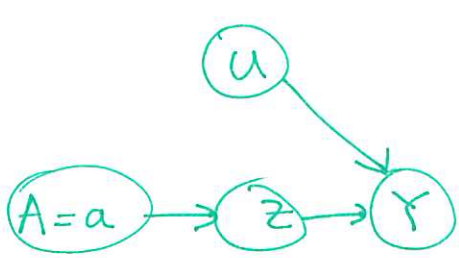
The law of total probability

$$\mathbb{P}_{do(A=a)}(Y=y) = \sum_z \mathbb{P}_{do(A=a)}(Y=y, Z=z)$$

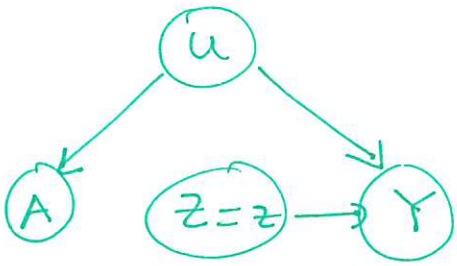
$$= \sum_z \mathbb{P}_{do(A=a)}(Y=y \mid Z=z)\,\mathbb{P}_{do(A=a)}(Z=z)$$

$$= \mathbb{P}_{do(Z=z)}(Y=y)$$

since $Y = f(Z, U, N_Y)$ and ... ↪



vs



$$\begin{cases} U = f(N_U) \\ A = a \\ Z = f(A, N_Z) \\ Y = f(U, Z, N_Y) \end{cases}$$

int. prob $\mathbb{P}_{do(A=a)}(\cdot)$

$$\begin{cases} U = f(N_U) \\ A = f(U, N_A) \\ Z = z \\ Y = f(U, Z, N_Y) \end{cases}$$

int. prob $\mathbb{P}_{do(Z=z)}(\cdot)$

Expression of U and N_Y unchanged
+
whether we $do(Z=z)$ or observe $Z = f(a, N_Z) = z$, this has no influence on Y.

$$\Rightarrow \mathbb{P}_{do(A=a)}(Y=y \mid Z=z) = \mathbb{P}_{do(Z=z)}(Y=y)$$

$$\Downarrow$$

$$\mathbb{P}_{do(A=a)}(Y=y) = \sum_z \underbrace{\mathbb{P}_{do(Z=z)}(Y=y)}_{(\cdot\cdot)} \underbrace{\mathbb{P}_{do(A=a)}(Z=z)}_{(\cdot)}$$

Front-door formula:

$$\mathbb{P}(Y^a = y) = \mathbb{P}_{do(A=a)}(Y=y)$$

$$= \sum_z \mathbb{P}(Z=z \mid A=a) \sum_{a'} \mathbb{P}(Y=y \mid Z=z, A=a') \times \mathbb{P}(A=a')$$

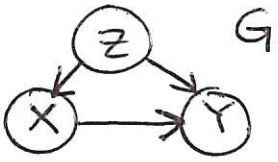
The backdoor criterion is not necessary to identify causal effects: if the frontdoor criterion is satisfied, we also have identifiability. However, the frontdoor criterion is itself a sufficient criterion. This naturally leads us to the following question: can we identify causal estimates when the DA G satisfies neither the backdoor nor the frontdoor criterion?

↳ In other words, is there a general procedure to express intervention probabilities $\mathbb{P}_{do(X=x)}(Y=y)$ using only observational proba $\mathbb{P}(Y=y \mid X=x, \dots)$?

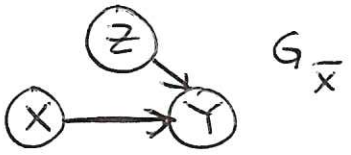
Pearl's do-calculus gives the answer to this question. The do-calculus provides tools to identify causal effects using only the structure of the causal graph.

- Notation:

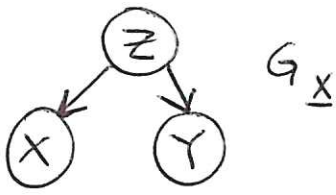
G = causal graph



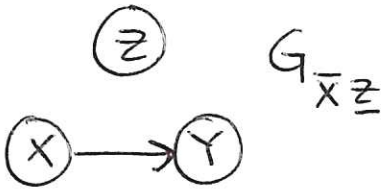
$G_{\bar{X}}$ = graph that we get if we take G and remove all incoming arrows into X



$G_{\underline{X}}$ = graph that we get if we take G and remove all outgoing edges from X



$G_{\bar{X}\underline{Z}}$ = graph that we get if we take G, and remove all incoming edges into X and remove all outgoing edges from Z

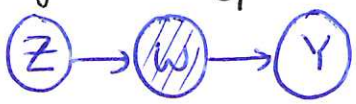


Thm: Rules of do-calculus

Simplified Rule

- **Rule 1** = adding / deleting an observation

$$p(y \mid z, w) = p(y \mid w) \quad \text{if } Y \perp_G Z \mid W$$



this rule is generalized to intervention probabilities

General Rule (Pearl)

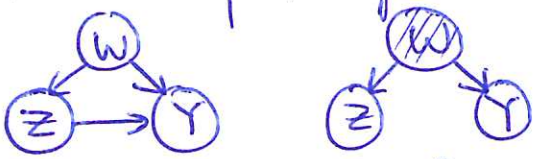
$$P_{do(x)}(y \mid z, w) = P_{do(x)}(y \mid w) \quad \text{if } Y \perp_{G_{\bar{X}}} Z \mid X, W$$

We need to consider $G_{\bar{X}}$ instead of G since we are interested in $P_{do(x)}(\cdot) \Rightarrow$ edges incoming to X must be removed. Also, $\{X = x\}$ is observed & thus conditioned on.

- **Rule 2** = exchanging observation $\leftrightarrow$ action

$$P_{do(z)}(y \mid w) = p(y \mid z, w) \quad \text{if } Y \perp_{G_{\underline{Z}}} Z \mid W$$

In other words, W blocks all backdoor paths from Z to Y



$$P_{\substack{do(x) \\ do(z)}}(y \mid w) = P_{do(x)}(y \mid z, w) \quad \text{if } Y \perp_{G_{\bar{X}\underline{Z}}} Z \mid X, W$$

generalize by adding an extra "do" operator.

- **Rule 3** = adding / deleting an action

$$P_{do(z)}(y) = p(y) \quad \text{if } Y \perp_{G_{\bar{Z}}} Z$$

(i) $$P_{do(z)}(y \mid w) = p(y \mid w) \quad \text{if } Y \perp_{G_{\overline{Z(W)}}} Z \mid W$$

where $Z(W)$ denotes the set of nodes of Z that are not ancestors of any node of W

(ii) $P_{do(x)\,do(z)}(y|w) = P_{do(z)}(y|w)$

if $Y \perp_{G_{\overline{X}\,\overline{Z(W)}}} Z \mid X, W$

Thm Shpitser & Pearl (2006)
The do-calculus is complete.

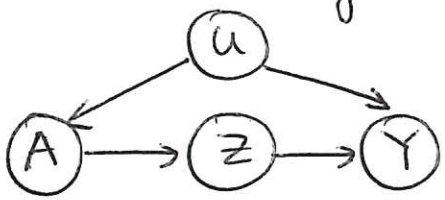
↑ In otherwords, any identifiable quantity can be deduced from these 3 rules of do-calculus:
The 3 rules are sufficient to derive all identifiable causal effects.
(& their proof is constructive → identification can be obtained in polynomial time)

× **Remark**: All the above is about non-parametric identification
The do-calculus tells us if we can identify causal effects using only the structure of the causal graph.
Under additional parametric assumptions, one can recover more causal effects.

↘ cf instrumental variables p.65

Ex: Derivation of the frontdoor adjustment using the 3 rules of the do-calculus

G:



$$p(y \mid do(a)) = \sum_z P_{do(a)}(y, z) \quad \text{(Law of Total Proba.)}$$

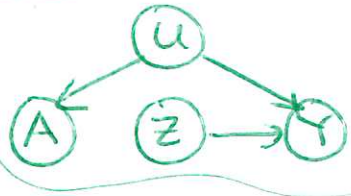
$$= \sum_z P_{do(a)}(y \mid z) \, P_{do(a)}(z)$$

$$= \sum_z P_{do(a)}(y \mid z) \, p(z \mid a)$$

Rule 2

$Z \perp_{G_{\underline{A}}} A$

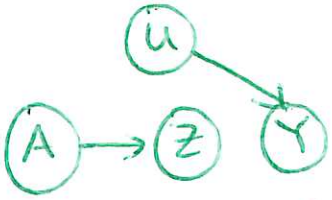
$G_{\underline{A}}$:



Rule 2

$Y \perp_{G_{\bar{A}\underline{Z}}} Z \mid A$

$G_{\bar{A}\underline{Z}}$



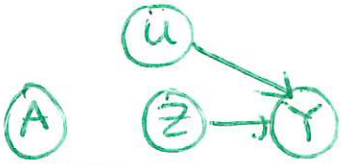
$$= \sum_z P_{do(a),do(z)}(y) \, p(z \mid a)$$

$$= \sum_z P_{do(z)}(y) \, p(z \mid a)$$

Rule 3

$Y \perp_{G_{\bar{Z}\bar{A}}} A \mid Z$

$G_{\bar{Z}\bar{A}}$:



$$P_{do(z)}(y) = \sum_a P_{do(z)}(y \mid a) \, P_{do(z)}(a)$$

$$= \sum_a p(y \mid a, z) \, p(a)$$

Rule 2

$Y \perp_{G_{\underline{Z}}} Z \mid A$

$G_{\underline{Z}}$

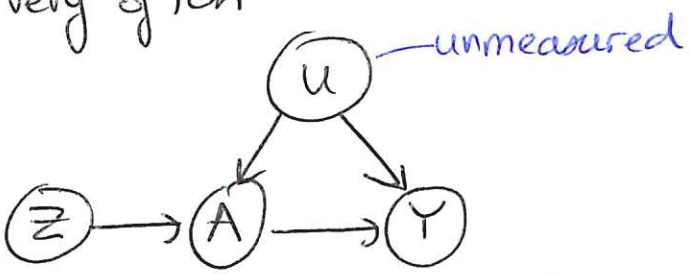
Rule 3

$A \perp_{G_{\bar{Z}}} Z$

• Instrumental variables

∠ used in economics very often

∠ typical situation

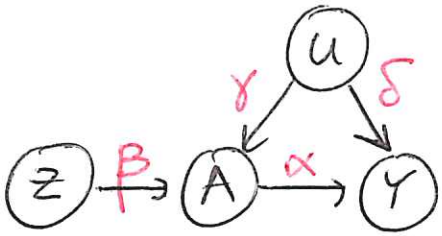


— Z is known as an instrumental variable —

To identify the causal effect of A on Y, we typically make further assumption on the structural equations.

Ex: Linear structural equations (continuous case)

$$\begin{cases} A = \beta Z + \gamma U + N_A \\ Y = \alpha A + \delta U + N_Y \end{cases}$$



↘ Since

$$Y = \alpha\boxed{(\beta Z)} + \boxed{(\alpha\gamma+\delta)U + \alpha N_A + N_Y},$$

new variable — noise

linear model

we can make use of the following procedure

(a) Regress A on Z & get an unbiased estimate $\hat{\beta}$ of β.

(b) Compute the fitted values $\hat{\beta} Z$

(c) Regress Y on the fitted values $\hat{\beta} Z$ & get an unbiased estimate $\hat{\alpha}$ of α; the causal effect of A on Y.

• difference-in-difference (DiD).

Recall: under the unconfoundedness assumption $(Y^0, Y^1) \perp A$, we can identify the ATE $= \mathbb{P}(Y^1=1) - \mathbb{P}(Y^0=1)$

$(Y^0, Y^1) \perp A$ ↳ $= \mathbb{P}(Y^1=1 \mid A=1) - \mathbb{P}(Y^0=1 \mid A=0)$

consistency ↳ $= \mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=0 \mid A=0)$

To recover the ATT, we only need to assume that $Y^0 \perp A$ since

$$\begin{aligned} \text{ATT} &= \mathbb{P}(Y^1=1 \mid A=1) - \mathbb{P}(Y^0=1 \mid A=1) \\ &= \mathbb{P}(Y^1=1 \mid A=1) - \mathbb{P}(Y^0=1 \mid A=0) \\ &= \mathbb{P}(Y=1 \mid A=1) - \mathbb{P}(Y=1 \mid A=0) \end{aligned}$$

ATE	↔	$(Y^0, Y^1) \perp A$
ATT	↔	$Y^0 \perp A$

When treatment and outcome are confounded by L, we need extra steps to recover the ATE & ATT. In addition, when the confounder is unmeasured, we cannot use the backdoor criterion. "Adjustment" can be made if a front-door criterion is applicable, or in the presence of an instrumental variable. Alternatively, we may recover the ATT (& potentially the ATE) provided pre-treatment measurements are available, and under similar/related

independence assumptions on the treatment variable and the potential outcomes.

Let. τ = time indicator $\in \{$ pre-test $(=0)$, test $(=1)\}$

. $Y^a(\tau)$ = potential outcome under treatment a at time τ.

We are interested in estimating the ATT

$$\boxed{ATT = \mathbb{E}(Y^1(1) - Y^0(1) \mid A=1)}$$

We will see that DiD is commonly used to identify the ATT; but additional assumptions the ATE can be identified as well.

We make the following assumptions.

(i) Consistency: $\forall \tau,\ A=a \Rightarrow Y(\tau) = Y^a(\tau)$.

(ii) Parallel trend: (the "essence" of DiD)

$$\boxed{Y^0(1) - Y^0(0) \perp A}$$

Compare with the traditional $Y^0 \perp A$ assumpt.
* unconfoundedness is about the difference *

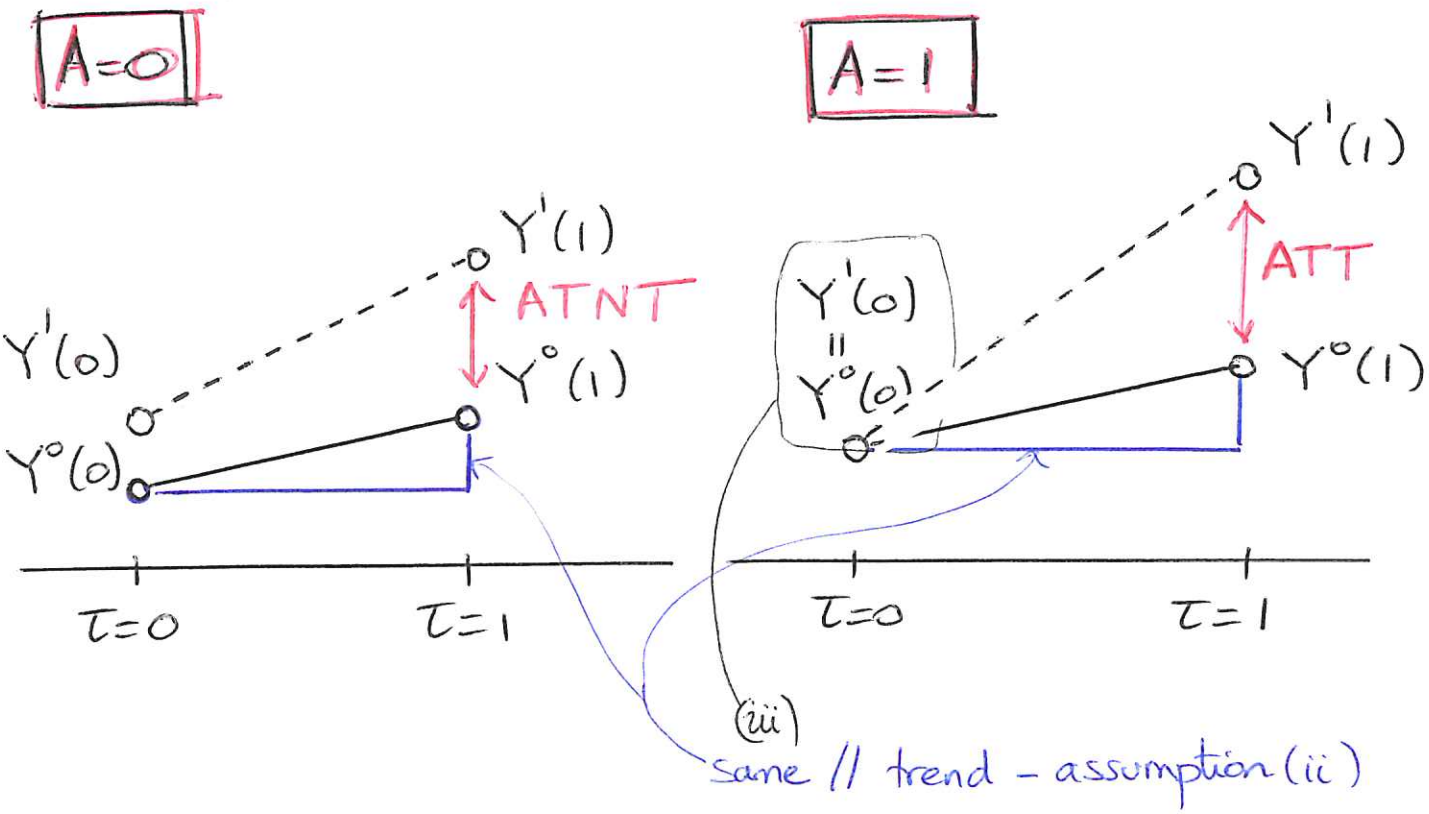
csq: $\mathbb{E}(Y^0(1) - Y^0(0) \mid A=1)$
$= \mathbb{E}(Y^0(1) - Y^0(0) \mid A=0)$.

(iii) no-pre-treatment effect.

$$\mathbb{E}(Y^1(0) - Y^0(0) \mid A=1) = 0$$

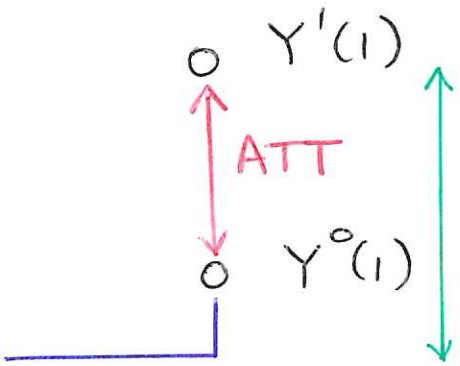
usually ok, unless patients anticipate the treatment.

- Visually



Not all quantities are observed : we only observe
↳ $Y^\circ(0)$ and $Y^\circ(1)$ in the group $A=0$
↳ $Y'(0)$ and $Y'(1)$ in the group $A=1$.
(solid lines)

⇒ We immediately visualize that the parallel trend assumption allow us to identify the ATT. In addition,



$$ATT = \mathbb{E}(Y(1) - Y(0) \mid A=1) - \mathbb{E}(Y(1) - Y(0) \mid A=0)$$

- Formal proof = $ATT = \mathbb{E}(Y'(1) - Y^\circ(1) \mid A=1)$

$$= \mathbb{E}(Y'(1) \mid A=1) - \underline{\mathbb{E}(Y^0(1) \mid A=1)}$$

(ii)
$$= \mathbb{E}(Y^0(0) \mid A=1) + \mathbb{E}(Y^0(1) \mid A=0) - \mathbb{E}(Y^0(0) \mid A=0)$$

$$= \left\{ \mathbb{E}(Y'(1) \mid A=1) - \boxed{\mathbb{E}(Y^0(0) \mid A=1)} \right\} - \left\{ \mathbb{E}(Y^0(1) \mid A=0) - \mathbb{E}(Y^0(0) \mid A=0) \right\}$$

$$= \mathbb{E}(Y'(0) \mid A=1) \quad \text{(iii)}$$

$$= \mathbb{E}(Y'(1) - Y'(0) \mid A=1) - \mathbb{E}(Y^0(1) - Y^0(0) \mid A=0)$$

$$= \mathbb{E}(Y(1) - Y(0) \mid A=1) - \mathbb{E}(Y(1) - Y(0) \mid A=0) \quad \text{(i)}$$

∎

× Remark : The ATNT $= \mathbb{E}(Y'(1) - Y^0(1) \mid A=0)$ can be identified replacing assumption (ii) and (iii) with

(ii)' $\underbrace{Y'(1) - Y'(0)}_{\text{on } Y' \text{ instead of } Y^0} \perp A$ (parallel trend)

(iii)' $\mathbb{E}(Y'(0) - Y^0(0) \mid A=0) = 0$ (no pre-tr. effect)

↓ Under (i), (ii)', (iii)', we get that

$$\text{ATNT} = \mathbb{E}(Y(1) - Y(0) \mid A=1) - \mathbb{E}(Y(1) - Y(0) \mid A=0)$$

same expression as before

Finally, under (i), (ii), (iii), (ii)', (iii)',

$$ATE = \mathbb{E}(Y(1) - Y(0) \mid A=1) - \mathbb{E}(Y(1) - Y(0) \mid A=0)$$

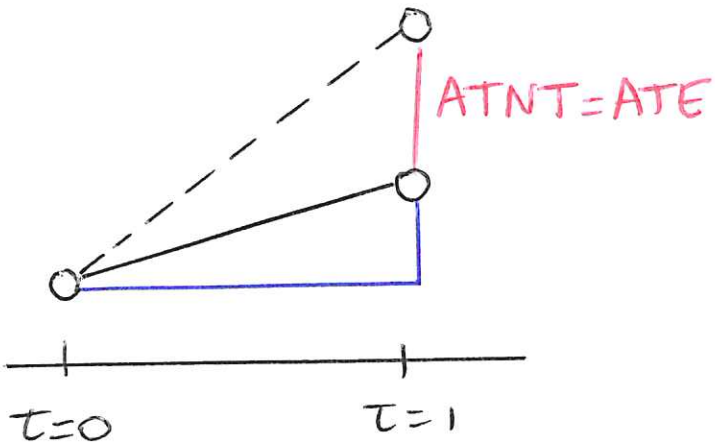
since

$$ATE = ATT \times \mathbb{P}(A=1) + ATNT \times \mathbb{P}(A=0) .$$

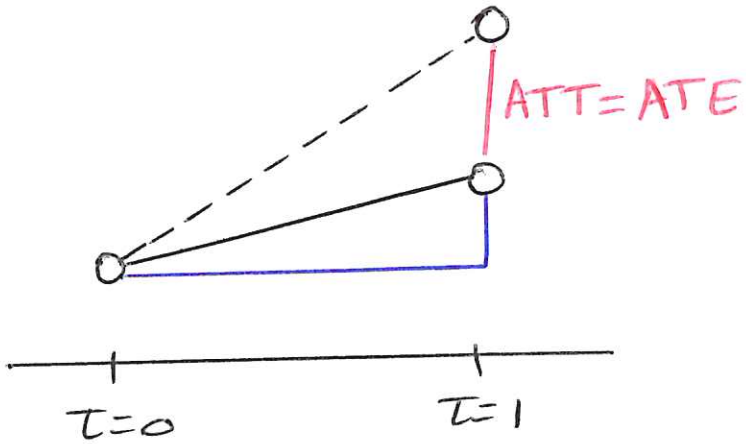
(ATT: same expression; ATNT: ↑)

- Visually, under (i), (ii), (iii), (ii)' and (iii)', all quantities are symmetrical

A=0

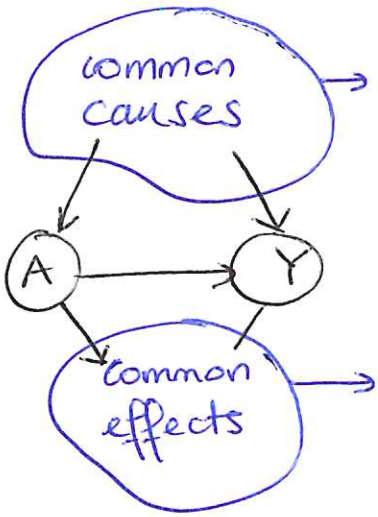


A=1



CI = SELECTION BIAS

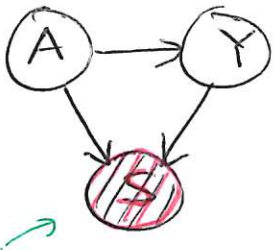
When measuring the causal effect of A on Y, selection bias arises when conditioning on a common effect of both A an Y



→ leads to confounding. Adjust for variables satisfying the backdoor criterion

→ potentially leads to selection bias. We discuss here procedures generalising / extending the backdoor criterion to adjust for selection bias.

Ex :



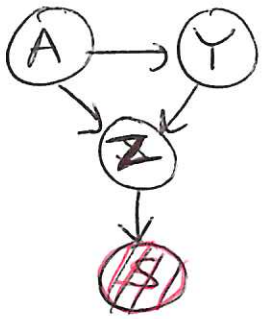
indicates that S is conditioned on

A = folic acid supplement intakes
Y = cardiac malformation during the first 2 months of pregnancy
S = death before birth.

↳ The study selects individuals who survived until birth.

Note that without explicitly including the sampling process, the study, assuming carried under an RCT, would look like (A) → (Y) ⇒ this DAG is oblivious to selection bias ⇒ highlights the importance to include a node (S) encoding the sampling procedure.

Alternatively, we may consider



A & Y as before
Z = death before birth
S = parental grief.

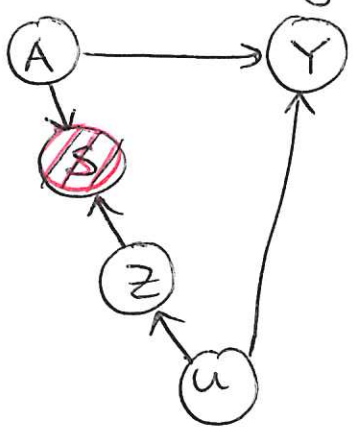
Selection bias remains when conditioning on a descendent of a common cause Z of both A and Y.

If the sampling process was completely random, S would be independent of all other variables in the analysis (no arrow coming in or out of S). When samples are collected preferentially, the causal effect must be recovered from $P(\underline{X} = \underline{x} \mid S = 1)$ instead of $P(\underline{X} = \underline{x})$

↳ encodes selected individuals

↳ $\underline{X} = (X_1, \ldots, X_d)$ = variables appearing in the DAG, including A and Y.

Ex: HIV study



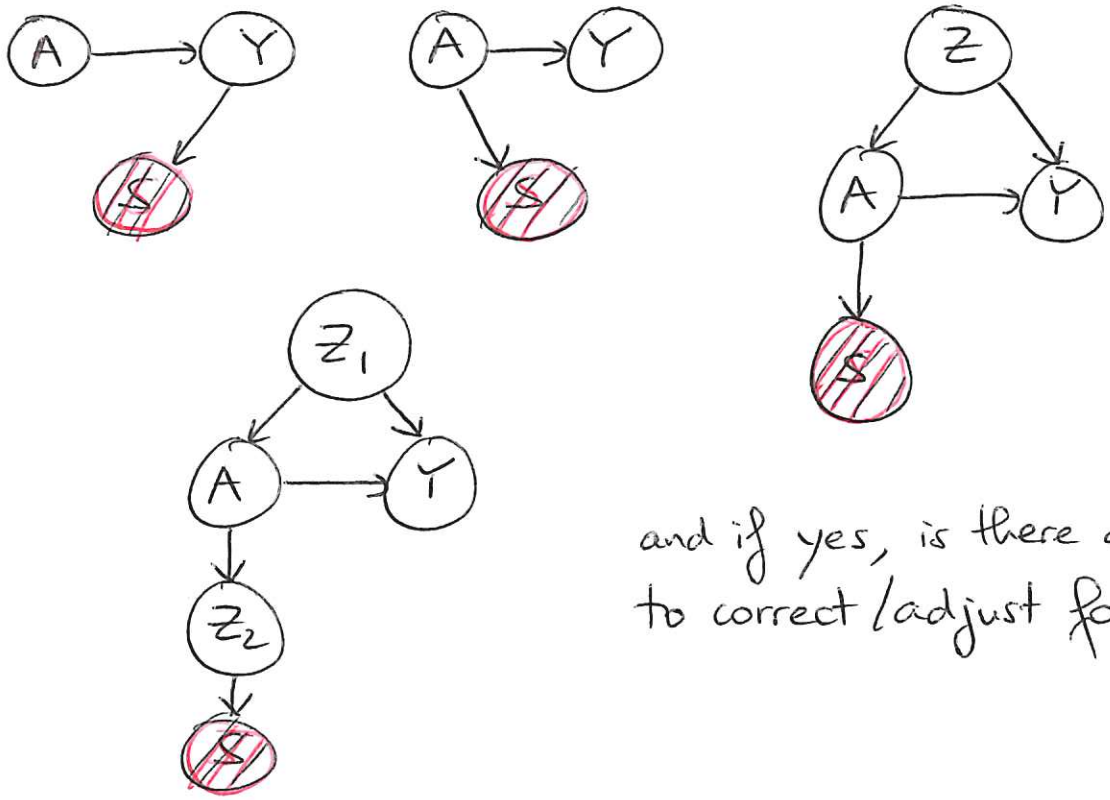
A = antiretroviral drug intake
Y = 3-year death
U = 1(high level of immunosuppression)
Z = presence of symptoms (fever, weight loss)
CD4 counts
viral load in plasma
S = 1(individual remain in the study)

Individuals with U=1 are more likely to be censored (S=0) due to the severity of the disease. Individuals

receiving the treatment are more likely to drop out due to the presence of heavy side effects.

The causal effect of A on Y suffers from the sampling process, restricted to individuals not dropping out of the study.

- Other examples: do we have selection bias in the following cases?



and if yes, is there a procedure to correct/adjust for it?

x Remark: Selection bias occurs in observational studies as well as in RCT.
Randomization protects against confounding, but not against selection bias.

no selection bias = sampling process at random
no confounding = treatment allocation at random

Consequences of selection bias are worse than with confounding = we are not even sure to recover conditional probabilities $\mathbb{P}(Y=y \mid A=a)$ from the dataset collected under selection bias $\mathbb{P}(\underline{X}=\underline{x} \mid S=1)$; let alone intervention probabilities $\mathbb{P}(Y^a=y) = \mathbb{P}(Y=y \mid do(A=a))$

↑ This is well understood intuitively: conditional probabilities are related to the frequency of events in a population. When some individuals are preferred, their characteristics may not be representative of the whole population.

Following Bareinboim, Tian & Pearl (2014), we consider 3 problems

(a) Selection without external data.

dataset $\sim \mathbb{P}(\underline{X}=\underline{x} \mid S=1)$

Q: Under which conditions is $\mathbb{P}(Y=y \mid A=a)$ recoverable?

(b) Selection with external data

dataset $\sim \mathbb{P}(\underline{X}=\underline{x} \mid S=1)$ + unbiased samples $\mathbb{P}(\underline{T}=\underline{t})$ for $\underline{T} \subset \underline{X}$

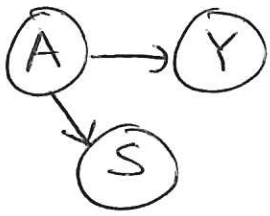
Q: Under which conditions is $\mathbb{P}(Y=y \mid A=a)$ recoverable?

(c) Selection in causal inference

dataset $\sim \mathbb{P}(\underline{X}=\underline{x} \mid S=1)$ + unbiased samples $\mathbb{P}(\underline{T}=\underline{t})$ for $\underline{T} \subset \underline{X}$

Q: Under which conditions is $\mathbb{P}(Y^a=y) = \mathbb{P}(Y=y \mid do(A=a))$ recoverable?

× Example 1:



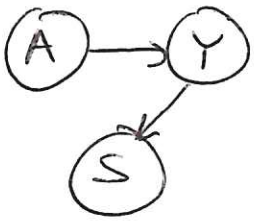
Since $Y \perp S \mid A$, $\mathbb{P}(Y=y \mid A=a, S=1)$ (*)

$= \mathbb{P}(Y=y \mid A=a)$ ← no confounding

$= \mathbb{P}(Y^a=y)$

⇒ No need to resort to additional data external to the biased study to recover conditional effects (& causal effects).

× Example 2:

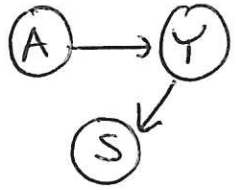


$Y \not\perp S \mid A$ so (*) is not true in general.

In fact, we show next that no matter how many samples $(A, Y, S=1)$ are drawn $\sim \mathbb{P}(A=a, Y=y \mid S=1)$, or how sophisticated the estimation technique is, the estimator of $\mathbb{P}(Y=y \mid A=a)$ will never converge to its true value.

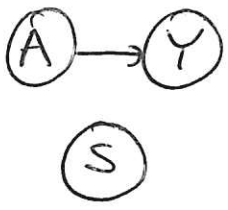
Consider $\mathbb{P}_1$ and $\mathbb{P}_2$ such that

↘ $\mathbb{P}_1$ is compatible with $G_1 =$



(preferred sampling)

↘ $\mathbb{P}_2$ is compatible with $G_2 \subset G_1$ =



(sampling at random)

Task = Construct P_1 and P_2 such that

$$P_1(A=a, Y=y \mid S=1) = P_2(A=a, Y=y \mid S=1)$$

they agree on the biased sample

but for which

$$P_1(Y=y \mid A=a) \neq P_2(Y=y \mid A=a)$$

↑ The assumptions embedded in the causal model are not sufficient to express conditional probabilities in terms of the distribution under selected bias.

Put $p_i(a, y, s) = P_i(A=a, Y=y, S=s)$, $i=1,2$.

- P_1 compatible with $G_1 \Rightarrow p_1(a,y,s) = p_1(s|y)p_1(y|a)p_1(a)$
- P_2 —— " —— $G_2 \Rightarrow p_2(a,y,s) = p_2(s)p_2(y|a)p_2(a)$

↑ we need to define explicitely all these terms so that the task above is fulfilled.

$$p_1(a, y \mid S=1) = p_2(a, y \mid S=1) = p_2(a, y)$$

↑ enforced by our task; ↑ $A, Y \perp S$ in G_2

Take $p_1(a) = p_2(a) =$ something arbitrary, $\forall a$

↓ dividing all members by $p_1 = p_2$ yields

$$p_1(y \mid a, S=1) = p_2(y \mid a, S=1) = p_2(y \mid a)$$

Thus

$$p_2(y|a) = p_1(y|a, S=1) = \frac{p_1(y, a, S=1)}{p_1(a, S=1)}$$

$$= \frac{p_1(S=1|y)\, p_1(y|a)\, p_1(a)}{p_1(S=1|a)\, p_1(a)}$$

$$= \frac{p_1(S=1|y)\, p_1(y|a)}{p_1(S=1|y,a)\, p(y|a) + p_1(S=1|1-y,a) \times p_1(1-y|a)}$$

Put • $\alpha = p_1(S=1|y)$

• $\beta = p_1(S=1|1-y)$; $\alpha \neq \beta$

• $p_1(y|a) = p_1(1-y|a) = 1/2$

(all terms p_1 are now defined)

$$\Downarrow$$

$$p_2(y|a) = \frac{\alpha}{\alpha+\beta} \neq \frac{1}{2} p_1(y|a), \text{ as required } \blacksquare$$

Definition: S-RECOVERABILITY.

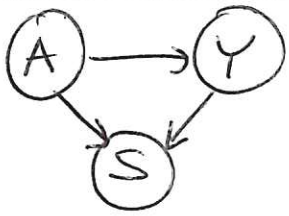
The distribution $\mathbb{P}(Y=y \mid A=a)$ is said to be s-recoverable from the selection biased data if $\forall$ two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ compatible with the DAG G,

$$\mathbb{P}_1(\underline{X}=\underline{x} \mid S=1) = \mathbb{P}_2(\underline{X}=x \mid S=1) > 0$$

$$\Downarrow$$

$$\mathbb{P}_1(Y=y|A=a) = \mathbb{P}_2(Y=y|A=a)$$

x Example 3



$\mathbb{P}(Y=y \mid A=a)$ is not s-recoverable (corollary from example 2).

Theorem Bareinboim, Tian & Pearl (2014)

$\mathbb{P}(Y=y \mid A=a)$ is s-recoverable from its DAG

$\Updownarrow$

$S \perp Y \mid A$.

$\Uparrow$ Sufficiency is immediate

$\Downarrow$ Necessity is tedious: need to show that for all DAGs in which $S \not\perp Y \mid A$ we can construct a counterexample showing agreement on $\mathbb{P}(\underline{X}=\underline{x} \mid S=1)$ & disagreement on $\mathbb{P}(Y=y \mid A=a)$.

We turn our attention to the case where external additional measurements are available at the population level (e.g. from census data).

↳ The definition of s-recoverability needs to be redefined to account for external data.

Suppose we collect

- data under selection bias $\mathbb{P}_i(\underline{M}=\underline{m} \mid S=1)$
- external data $\mathbb{P}_i(\underline{T}=\underline{t})$

where $\underline{M} \subset \{X_1, \ldots, X_d\}$; $\underline{T} \subset \{\underline{X}_1, \ldots, \underline{X}_d\}$

The distribution $\mathbb{P}(Y=y \mid A=a)$ is said to be s-recoverable from selection bias if for any two proba distrib $\mathbb{P}_1$ and $\mathbb{P}_2$ compatible with the DAG,

$$\mathbb{P}_1(\underline{M}=\underline{m} \mid S=1) = \mathbb{P}_2(\underline{M}=\underline{m} \mid S=1) > 0$$

& $$\mathbb{P}_1(\underline{T}=\underline{t}) = \mathbb{P}_2(\underline{T}=\underline{t}) > 0$$

$$\Rightarrow \mathbb{P}_1(Y=y \mid A=a) = \mathbb{P}_2(Y=y \mid A=a)$$

We identify conditions under which s-recoverability holds:

$$\mathbb{P}(Y=y \mid A=a) = \sum_z \mathbb{P}(Y=y, Z=z \mid A=a)$$

For some variables Z that are measured both in the biased study with A and Y, and at the population level together with A.

$$= \sum_z \mathbb{P}(Y=y \mid Z=z, A=a)\, \mathbb{P}(Z=z \mid A=a)$$

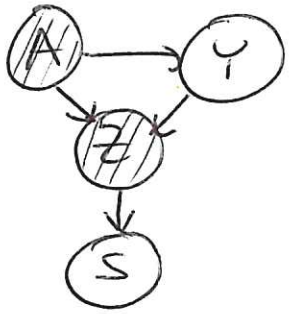
if $S \perp Y \mid Z, A$

$$= \sum_z \mathbb{P}(Y=y \mid Z=z, A=a, S=1) \times \mathbb{P}(Z=z \mid A=a)$$

This term can be estimated from selection biased data

Estimated from external data (population level)

Ex:



$S \perp Y \mid A, Z$ so we can recover the cond. proba $\mathbb{P}(Y=y \mid A=a)$ using the expression above.

Summary = If there is a set Z that is measured in the biased study together with $\{A, Y\}$ and at the population level with A such that $S \perp Y \mid Z, A$, then $P(Y=y \mid A=a)$ is s-recoverable and

$$P(Y=y \mid A=a) = \sum_z P(Y=y \mid Z=z, A=a, S=1) \times P(Z=z \mid A=a)$$

- Recovering causal effects $P(Y^a=y)$

Notation: Z^* = ~~all~~ set of variables $= Z^+ \cup Z^-$
Z^+ contains all non-descendents of A
Z^- —" —— descendents of A

× Case I = No confounding ($Z^+ = \emptyset$)

$$P(Y^a=y) = P(Y=y \mid do(A=a)) = P(Y=y \mid A=a) = \sum_{z^-} P(Y=y \mid A=a)\, P(Z^-=z^-)$$

(no confounding)

if $Z^- \perp Y \mid A$:

$$P(Y=y \mid A=a) = P(Y=y \mid A=a, Z^-=z^-)$$

if $S \perp Y \mid A, Z^-$:

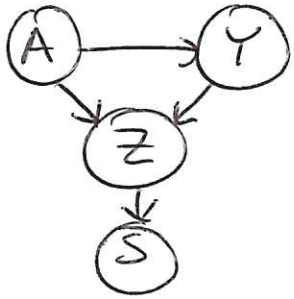
$$P(Y=y \mid A=a, Z^-=z^-) = P(Y=y \mid A=a, Z^-=z^-, S=1)$$

$$\boxed{P(Y^a=y) = \sum_{z^-} P(Y=y \mid A=a, Z^-=z^-, S=1)\, P(Z^-=z^-)}$$

↑ Expression obtained under $S \perp Y \mid A, Z^-$ (≡ s-recov.) + additional assumption $Z^- \perp Y \mid A$; the price

to pay for no external measurement of both A and Z at the population level (compare the weighting by $\mathbb{P}(Z^- = z^-)$ instead of $\mathbb{P}(Z^- = z^- \mid A = a)$).

- Ex (continued)



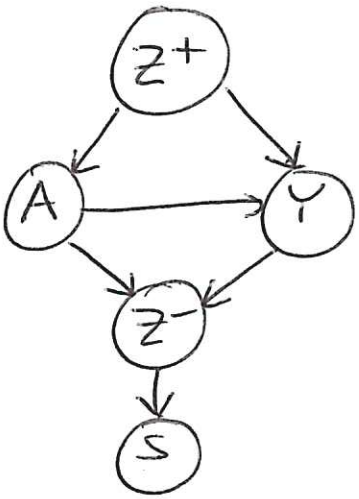
$Z \not\perp Y \mid A$

$S \perp Y \mid A, Z$

$$\mathbb{P}(Y^a = y) = \sum_z \mathbb{P}(Y = y \mid A = a, Z = z, S = 1)\,\mathbb{P}(Z = z \mid A = a)$$

$$\neq \sum_z \text{"}\mathbb{P}(Y = y \mid A = a, Z = z, S = 1)\text{"} \times \mathbb{P}(Z = z)$$

- Case II = general case.



$$\mathbb{P}(Y = y \mid do(A = a))$$

$$= \sum_{z^+} \mathbb{P}(Y = y \mid A = a, Z^+ = z^+)\,\mathbb{P}(Z^+ = z^+)$$

(backdoor criterion adjusting for Z^+)

$$= \sum_{z^+, z^-} \mathbb{P}(Y = y \mid A = a, Z^+ = z^+) \times \mathbb{P}(Z^+ = z^+, Z^- = z^-)$$

$$= \mathbb{P}(Y = y \mid A = a, Z^+ = z^+, Z^- = z^-)$$

provided $Z^- \perp Y \mid A, Z^+$

$$= \mathbb{P}(Y = y \mid A = a, Z = z)$$

$$= \mathbb{P}(Y = y \mid A = a, Z = z, S = 1)$$

provided $S \perp Y \mid A, Z$

$$\mathbb{P}(Y=y \mid do(A=a)) = \sum_{z} \mathbb{P}(Y=y \mid A=a, Z=z, S=1)\, \mathbb{P}(Z=z).$$

↑ Under (i) Z^+ blocks all backdoor paths from A to Y.

(ii) $Z^- \perp Y \mid A, Z^+$

(iii) $S \perp Y \mid A, Z$

(iv) $Z \cup \{A, Y\}$ ~ selection biased

Z ~ population level.

These 4 conditions are called selection-backdoor criterion.

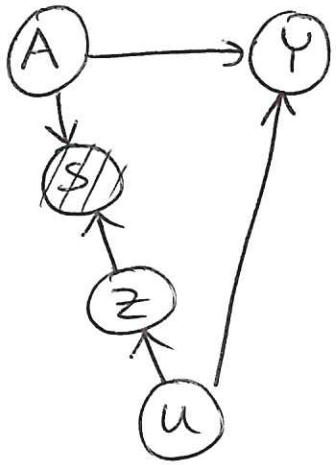
& the expression above the selection-backdoor adjustment.

× Remark = The selection-backdoor adjustment coincide with the backdoor formula derived in the presence of confounding. Note however that we are adjusting for $Z = Z^+ \cup Z^-$, where Z^- contains descendents of A, which was forbidden in the backdoor criterion.

There are no contradictions. The reason is that the backdoor criterion is a sufficient condition. There is extra room to adjust for descendents under additional assumptions, and this is exactly what the selection-backdoor criterion is.

Note that the selection-backdoor criterion is a sufficient condition to recover the causal effect, as illustrated in the next example.

• Ex = HIV study



No confounding $Z_+ = \emptyset$
Take $Z_- = Z$.

(a) $S \perp Y \mid A, Z$ since

$S \leftarrow A \rightarrow Y$ ✓

$S \leftarrow Z \leftarrow U \rightarrow Y$ ✓

(b) $Z \not\perp Y \mid A$ since

$Z \rightarrow S \leftarrow A \rightarrow Y$ ✓

$Z \leftarrow U \rightarrow Y$ ✗

⇒ Condition (ii) of the selection-backdoor criterion fails to hold. However, we can derive the same adjustment formula from scratch, making use of alternative independence structure emerging from the graph.

$$\mathbb{P}(Y=y \mid do(A=a)) = \mathbb{P}(Y=y \mid A=a)$$

$$= \sum_z \mathbb{P}(Y=y \mid Z=z, A=a)\, \mathbb{P}(Z=z \mid A=a)$$

since $Z \perp A$: $\mathbb{P}(Z=z \mid A=a) = \mathbb{P}(Z=z)$

$Z \rightarrow S \leftarrow A$ ⟶ S = collider

$Z \leftarrow U \rightarrow Y \leftarrow A \rightarrow Y$ = collider

since $S \perp Y \mid A, Z$

$$= \sum_z \mathbb{P}(Y=y \mid Z=z, A=a, S=1)\, \mathbb{P}(Z=z)$$

■

× Remark : Selection-backdoor adjustment and exchangeability

Assuming that the selection-backdoor aiterion holds,

$$\mathbb{P}(Y^a = y) = \sum_z \mathbb{P}(Y=y \mid A=a, Z=z, S=1)\mathbb{P}(Z=z)$$

$$= \sum_z \frac{\mathbb{P}(Y=y, A=a, Z=z, S=1)}{\mathbb{P}(A=a, Z=z, S=1)} \mathbb{P}(Z=z)$$

$$= \sum_z \frac{1}{\mathbb{P}(A=a, S=1 \mid Z=z)} \mathbb{P}(Y=y, A=a, Z=z, S=1)$$

$$= \sum_z \frac{1}{\mathbb{P}(A=a \mid Z=z)} \frac{1}{\mathbb{P}(S=1 \mid Z=z, A=a)} \mathbb{P}(-\text{joint}-$$

adjusts for confounding ↑ adjusts for selection bias ↑

$$= \sum_z \underbrace{w(a,z)\, w^s(a,z)}_{=: w^c(a,z)} \mathbb{P}(-\text{joint} -)$$

⇒ Each observation $(A_i, Z_i, Y_i, S_i = 1)$ receives a weight $w^c(A_i, Z_i)$

[We may consider as well stabilized weights

$\tilde{w}^c(a,z) := \tilde{w}(a,z)\, \tilde{w}^s(a,z)$, where

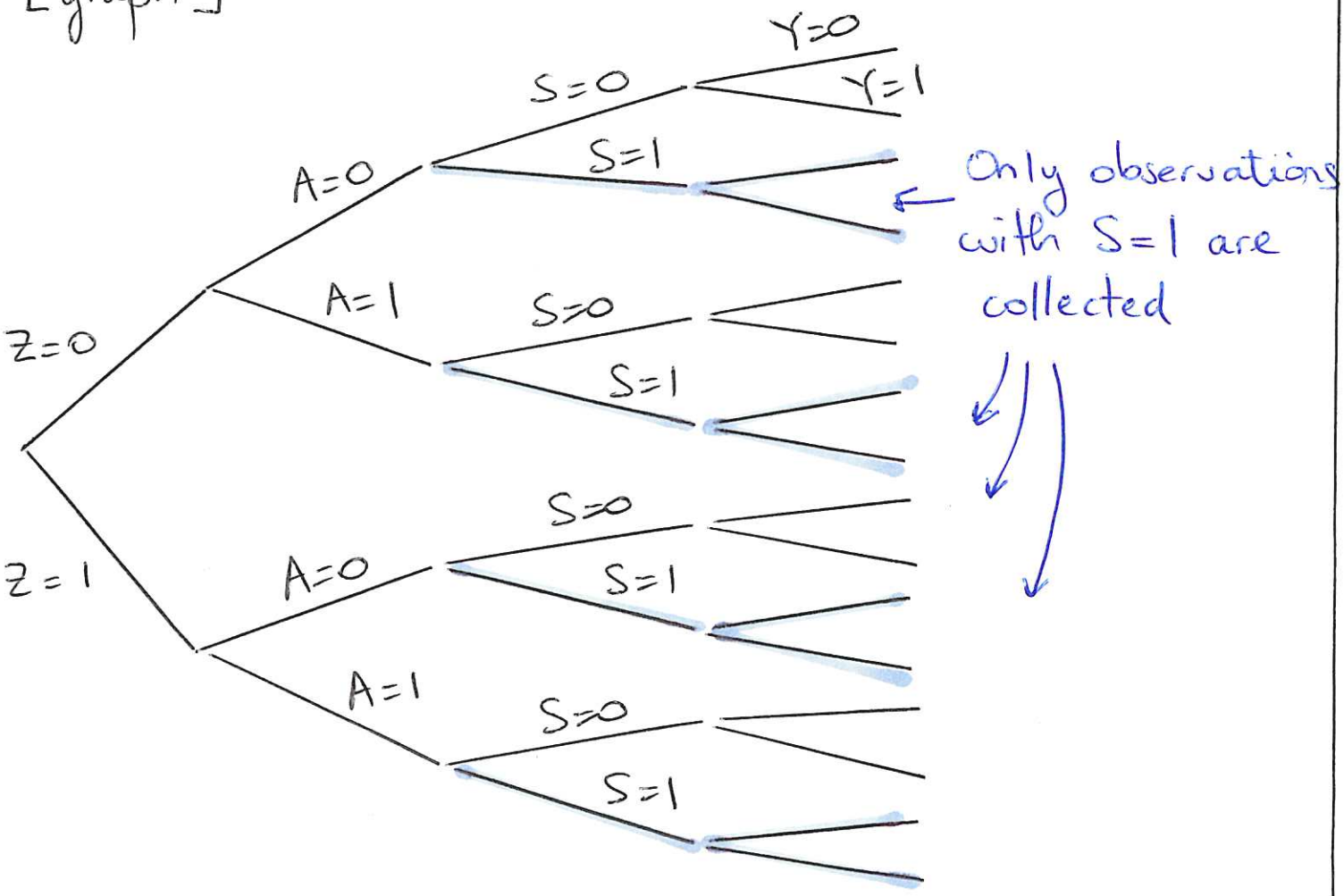
$\tilde{w}(a,z) := \mathbb{P}(A=a) / \mathbb{P}(A=a \mid Z=z)$ (p. 9c)

$\tilde{w}^s(a,z) := \mathbb{P}(S=1 \mid A=a) / \mathbb{P}(S=1 \mid A=a, Z=z)$]

↑ The numerator must be independent of Z

The modified HT estimator of the ATE is obtained by fitting the (saturated) linear model $Y \mid A, S=1 = \beta_0 + \beta_1 A + \varepsilon$ using a weighted LS approach, where each observation $(A_i, Z_i, Y_i, S_i = 1)$ receives weight $w^c(A_i, Z_i)$.

[graph]

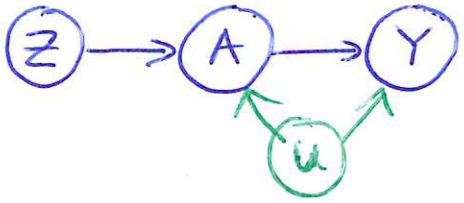


When interested in estimating $\mathbb{P}(Y^a = 1)$, we must reallocate all users with $A=0$ to the branch $A=1$ [that's the job of $w(a,z)$] & reallocate users with $S=0$ to the branch $S=1$ [the job of $w^s(a,z)$]

The appropriate scalings are $1/\mathbb{P}(A=a \mid Z=z)$ and $1/\mathbb{P}(S=1 \mid A=a, Z=z)$, respectively. (see explanation p.718 with no censoring)

CI = MEASUREMENT BIAS

- The main contribution of this chapter is to introduce the Intent-To-Treat (ITT) effect under partial compliance to treatment assignment.



Z = treatment allocation $\in \{0,1\}$
A = treatment status $\in \{0,1\}$
Y = outcome $\in \{0,1\}$
U = unmeasured confounders.

↑

In a RCT with partial compliance, the treatment assignment is independent of any confounder U, so that

$$\mathbb{P}(Y^{z=1} = 1) = \mathbb{P}(Y=1 \mid Z=1)$$
$$\& \ \mathbb{P}(Y^{z=0} = 1) = \mathbb{P}(Y=1 \mid Z=0).$$

↑ The difference between these two quantities is precisely the Intent-To-Treat (ITT):

$$ITT = \mathbb{E}(Y^{z=1} - Y^{z=0}).$$

- The notation $Y^{z=z}$ stands for Y under the intervention $do(Z=z)$. Similarly, we denote $A^{z=1}$ and $A^{z=0}$ the potential outcomes of A under the interventions $do(Z=1)$ and $do(Z=0)$, so that $A = ZA^{z=1} + (1-Z)A^{z=0}$

- In addition, we introduce $Y^{a=1}$ and $Y^{a=1}$ to denote Y under the intervention $do(A=1)$ and $do(A=0)$.

Then, $Y = AY^{a=1} + (1-A)Y^{a=0}$.

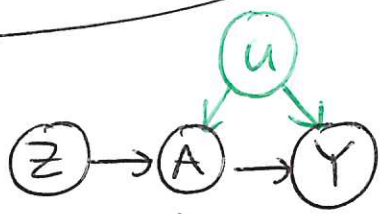
individual i — indpt of Z

$(Z, A^{z=0}, A^{z=1}, Y^{a=0}, Y^{a=1})$

$A = ZA^{z=1} + (1-Z)A^{z=0}$

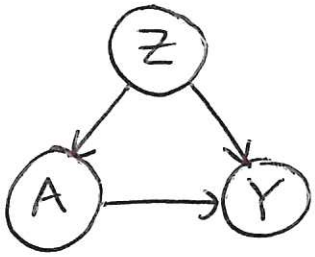
$Y = AY^{a=1} + (1-A)Y^{a=0}$

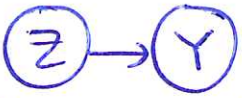
↳ observe (Z_i, A_i, Y_i) for each individual i.

× Remark = The DAG 

implicitly assumes the exclusion principle: no direct arrow from Z to Y.

Consider the more general scenario:



Ex: 

that encodes the awareness of the assigned treatment might lead to changes in the behavior of study participants; having an effect on the outcome.

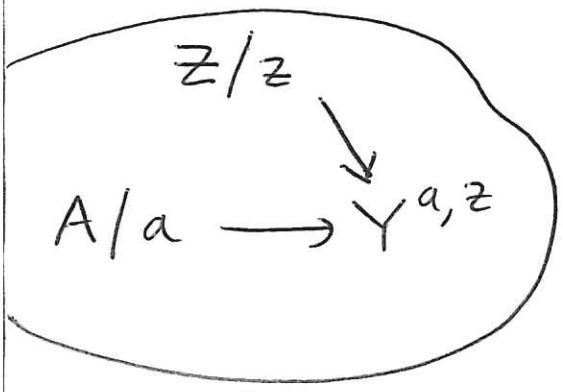
$Z \to A$: degree of adherence to the assigned treatment

$A \to Y$: effect of the treatment

$Z \to Y$: concurrent behavioral change

↖ to break this, perform a double-blind RCT.

In the general case, we need to consider potential outcomes $Y^{a,z}$, $a \in \{0,1\}$, $z \in \{0,1\}$ under the double intervention $do(Z=z)$ & $do(A=a)$.



Observe

$$Y = AZY^{1,1} + (1-A)Y^{0,1} + A(1-Z)Y^{1,0} + (1-A)(1-Z)Y^{0,0}$$

When the exclusion principle holds (no arrow from Z to Y),

$$Y^{a=0,z=0} = Y^{a=0,z=1} = Y^{a=0}$$
$$Y^{a=1,z=0} = Y^{a=1,z=1} = Y^{a=1}$$

and Y simplifies to $Y = AY^{a=1} + (1-A)Y^{a=0}$, as required. ∎

Assumption 1: $A^{z=0} \sim B(\pi_0)$, $A^{z=1} \sim B(\pi_1)$

If allocated to control, the proba to receive the treatmt is π_0.

If allocated to treatment, the proba to receive the treatment is π_0.

The total population can be partitioned into 4 groups, based on their values of $A^{z=0}$ and $A^{z=1}$.

$A^{z=0}$ \ $A^{z=1}$	0	1
0	Never Takers	Compliers
1	Defiers	Always Takers

Assumption 2 = $\pi_0 = 0$

Subjects in the control group will not take the treatment

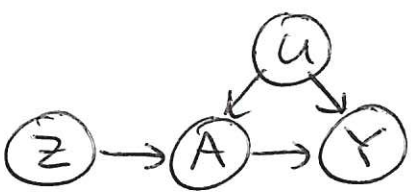
$\Leftrightarrow A^{z=0} = 0$ a.s.

Under assumption 2, the population is split between never takers & compliers:

$$\text{Never Takers} = \{\omega \in \Omega \mid A^{z=1}(\omega) = 0\}$$
$$\text{Compliers} = \{\omega \in \Omega \mid A^{z=1}(\omega) = 1\}.$$

- **Summary**: Under assumptions 1 and 2, and under the exclusion principle,

$$A^{z=1} \sim B(\pi_1)$$
$$A = z\,A^{z=1}$$
$$Y = A\,Y^{a=1} + (1-A)\,Y^{a=0}$$



individual i

$$(Z_i, A_i^{z=1}, Y_i^{a=0}, Y_i^{a=1})$$

Note that under the intervention $do(Z=0)$ and $do(Z=1)$,

- When we $do(Z=0)$, then $A = A^{Z=0} \equiv 0$ (assumption 2), and $Y^{Z=0} = Y^{a=0}$.
- When we $do(Z=1)$, then $A = A^{Z=1}$, and $Y^{Z=1} = A^{Z=1} Y^{a=1} + (1 - A^{Z=1}) Y^{a=0}$.

$$
\begin{aligned}
ITT &= \mathbb{E}(Y^{Z=1} - Y^{Z=0}) \\
&= \mathbb{E}(A^{Z=1} Y^{a=1} + (1 - A^{Z=1}) Y^{a=0} - Y^{a=0}) \\
&= \mathbb{E}(\text{---}''\text{---} \mid A^{Z=1} = 1)\, \mathbb{P}(A^{Z=1} = 1) \\
&\quad + \boxed{\mathbb{E}(\text{---}''\text{---} \mid A^{Z=0} = 0)\, \mathbb{P}(A^{Z=1} = 0)} \\
&\qquad \hookrightarrow = 0
\end{aligned}
$$

$$
\boxed{ITT = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{Z=1} = 1)\, \mathbb{P}(A^{Z=1} = 1)}
$$

- $\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{Z=1} = 1)$: treatment effect of the compliers, aka the Local Average Treatment Effect (LATE) or the Complier Average Causal Effect (CACE)
- $\mathbb{P}(A^{Z=1} = 1) = \pi_1$ = proportion of compliers.

* Remark: When $\pi_1 = 1$, then $ITT = \mathbb{E}(Y^{a=1} - Y^{a=0}) = ATE$.

× Remark = Identification of the LATE under monotonicity (Imbens & Angrist '94)

$$\begin{aligned} ITT &= \mathbb{E}(Y|Z=1) - \mathbb{E}(Y|Z=0) \\ &= \mathbb{E}\left(A^{Z=1}Y^{a=1} + (1-A^{Z=1})Y^{a=0} \mid Z=1\right) \\ &\quad - \mathbb{E}\left(A^{Z=1}Y^{a=1} + (1-A^{Z=1})Y^{a=0} \mid Z=0\right) \end{aligned}$$

Assuming $Z \perp (A^{Z=0}, A^{Z=1}, Y^{a=0}, Y^{a=1})$ (p.59)

$$\begin{aligned} &= \mathbb{E}\left[(A^{Z=1} - A^{Z=0})(Y^{a=1} - Y^{a=0})\right] \\ &= \mathbb{E}\left(Y^{a=1} - Y^{a=0} \mid A^{Z=1} - A^{Z=0} = 1\right)\mathbb{P}(A^{Z=1} - A^{Z=0} = 1) \\ &\quad + \mathbb{E}(\text{—}'' \text{—} \mid \text{—}'' \text{—} = -1)\mathbb{P}(\text{—}'' \text{—} = -1) \end{aligned}$$

Without further assumptions, we may construct proba on $(A^{Z=0}, A^{Z=1}, Y^{a=0}, Y^{a=1})$ such that the ITT is zero, while the causal effect of A on Y is >0.

However, assuming $A^{Z=1} \geq A^{Z=0}$ (monotonicity), the second term in the RHS vanishes and

$$ITT = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{Z=1} \neq A^{Z=0})\,\mathbb{P}(A^{Z=1} \neq A^{Z=0})$$

Population made of [redacted] compliers. Under the previous/earlier assumption that $A^{Z=0} \equiv 0$, we recover conditioning on compliers as well.

Ex: $Z \in \{0,1\}$ $A = \mathbb{1}(\gamma_0 + \gamma_1 Z + \varepsilon > 0)$ $Y = \beta_0 + \beta_1 A + \eta$ "linear latent model"

Re-arranging the terms,

$$\mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} \neq A^{z=0}) = \frac{E(Y|z=1) - E(Y|z=0)}{\mathbb{P}(A^{z=1} \neq A^{z=0})}$$

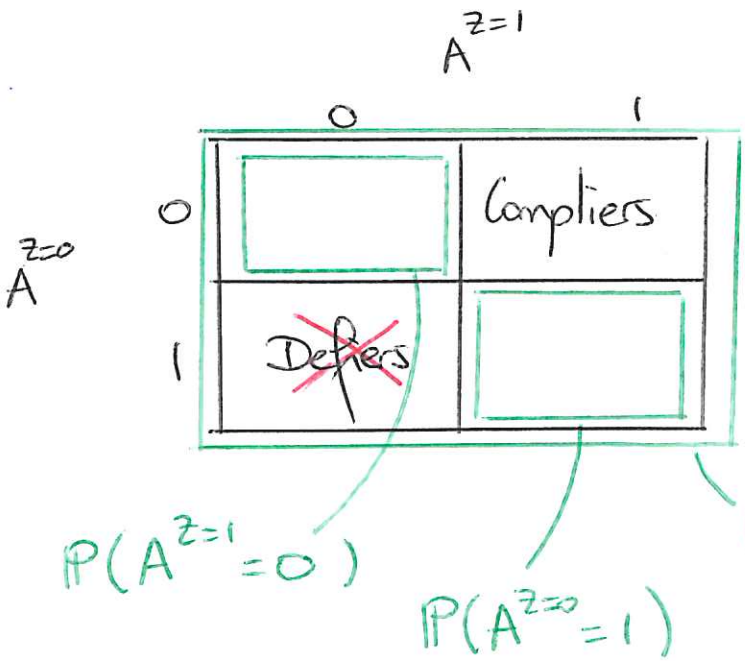
↑ This is the CATE i.e. the ATE within compliers. Under monotonicity, we have indeed that

$$\{A^{z=1} \neq A^{z=0}\} = \{A^{z=0}=0, A^{z=1}=1\}$$
$$= \{\text{compliers}\}$$

↑ dénominateur = $\mathbb{P}(A^{z=0}=0, A^{z=1}=1)$, which can be re-expressed as

$$1 - \mathbb{P}(A^{z=1}=0) - \mathbb{P}(A^{z=0}=1)$$

since there are no defiers.



$$= 1 - \mathbb{P}(A=1 \mid z=0) - \mathbb{P}(A=0 \mid z=1)$$
$$= \mathbb{P}(A=1 \mid z=1) - \mathbb{P}(A=1 \mid z=0)$$
$$= \mathbb{E}(A|z=1) - \mathbb{E}(A|z=0)$$

Summary :

$$\text{CATE} = \mathbb{E}(Y^{a=1} - Y^{a=0} \mid A^{z=1} \neq A^{z=0}) = \frac{\mathbb{E}(Y|z=1) - \mathbb{E}(Y|z=0)}{\mathbb{E}(A|z=1) - \mathbb{E}(A|z=0)}$$

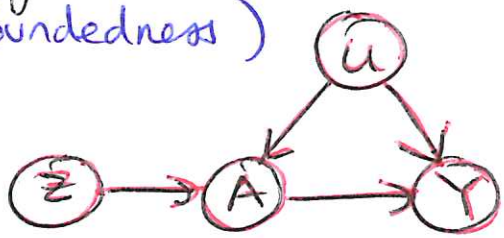
↑ Non-parametric identification of the CATE under monotonicity assumptions.

When there is partial compliance to treatment assignment, the variable Z is called an INSTRUMENT. It satisfies 3 key properties:

(i) Z has a causal effect on A (relevance)

(ii) Z has a causal effect on Y that is fully mediated by A (exclusion restriction)

(iii) There are no backdoor paths from Z to Y (instrumental unconfoundedness)



Previously, we established a non-parametric expression of the CATE. It turns out that there are no non-parametric expressions for the ATE, since the effect of A on Y is confounded by U.

⇒ With instrument variables, we must make assumptions about the parametric form to identify causal effects.

"do" probabilities can be expressed as functions of observation probabilities (eg. backdoor adjustment, frontdoor adjustment)

Ex 1: Binary linear $A, Z \in \{0, 1\}$

$Y = \delta A + \alpha U$ $\quad \delta = \text{ATE}$

$$E(Y|Z=1) - E(Y|Z=0)$$
$$= E(\delta A + \alpha U | Z=1) - E(\delta A + \alpha U | Z=0)$$
$$= \delta [E(A|Z=1) - E(A|Z=0)]$$
$$+ \alpha [\underbrace{E(U|Z=1) - E(U|Z=0)}_{=0 \text{ since } U \perp Z}]$$

$$\Rightarrow \delta = ATE = \frac{E(Y|Z=1) - E(Y|Z=0)}{E(A|Z=1) - E(A|Z=0)}$$

= same expression as the CATE under monotonicity. ∎

Ex 2: Continuous Linear $A, Z \in \mathbb{R}$

Consider instead $Cov(Y,Z) = E(YZ) - E(Y)E(Z)$
$$= E(\delta A + \alpha U)Z - E(\delta A + \alpha U)EZ$$
$$= \delta \{E(AZ) - (EA)(EZ)\} + \alpha \{E(UZ) - (EU)(EZ)\}$$
$$= \delta Cov(A,Z) + \alpha \underbrace{Cov(U,Z)}_{=0}$$

$$\Rightarrow \delta = ATE = \frac{Cov(Y,Z)}{Cov(A,Z)}.$$ ∎

× Remark: When there is a weak association between A and Z, the denominators $E(A|Z=1) - E(A|Z=0)$ & $Cov(A,Z)$ are close to 0, inflating the numerator = ITT [potential for explosive bias].

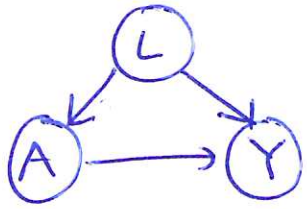
CI = PROPENSITY SCORES

When L satisfies the backdoor criterion ($\equiv$ conditional exchangeability), so that $(Y^0, Y^1) \perp A \mid L$, we may wonder if it is necessary to condition on the whole vector L, especially when L is high dimensional.

Rosenbaum & Rubin (1983) showed that it is enough to condition on $\mathbb{P}(A=1 \mid L)$. The quantity $e(L) := \mathbb{P}(A=1 \mid L)$ is commonly referred to as the PROPENSITY SCORE (PS) since it represents the propensity (probability) of receiving the treatment given L.

Thm $(Y^0, Y^1) \perp A \mid L \Rightarrow (Y^0, Y^1) \perp A \mid e(L)$

proof: In the simple setting where



,

the arrow from L to A symbolizes the probabilistic relationship between L and A i.e. $\mathbb{P}(A=1 \mid L)$ for a binary treatment variable, which is precisely the PS.

More formally, we show that $(Y^0, Y^1) \perp A \mid e(L)$ by showing that $\mathbb{P}(A=1 \mid Y^a, e(L))$ does not depend on Y^a, where $a=0$ or $a=1$.

$$\mathbb{P}(A=1 \mid Y^a, e(L)) = \mathbb{E}(A \mid Y^a, e(L))$$

$$= \mathbb{E}\Big[\underbrace{\mathbb{E}\{A \mid Y^a, \cancel{e(L)}, L\}}_{\text{average on something finer}} \mid Y^a, e(L)\Big]$$

$$= \mathbb{E}\Big[\underbrace{\mathbb{E}\{A \mid Y^a, L\}}_{\parallel} \mid Y^a, e(L)\Big]$$

$$\mathbb{E}\{A \mid L\} \text{ since } Y^a \perp A \mid L$$

$$\parallel$$

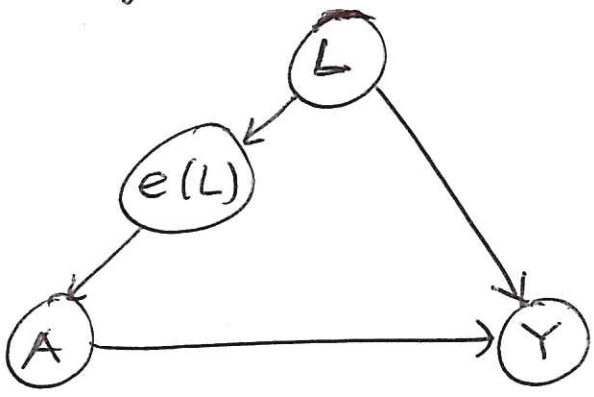
$$\mathbb{P}(A=1 \mid L) = e(L)$$

$$= \mathbb{E}\big[e(L) \mid Y^a, e(L)\big]$$

$$= e(L)$$

↖ independent of Y^a.

Graphically, this means that $e(L)$ is a full mediator of the effect of L on A



↗ And we see that $A \perp L \mid e(L)$ i.e. the PS balances the covariates between the treated & the untreated.

vs Randomization which balances measured & unmeasured covariates [& thus preferred over PS]

Csq: We can adjust for $e(L)$ in place of L.

*Remark: balancing scores

Rosenbaum & Rubin (1983) introduced the concept of a balancing score $b(L)$.

Def: $b(L)$ is a balancing score $\Leftrightarrow$ $A \perp L \mid b(L)$

↳ the PS is the simplest example of a balancing score.

They showed the following result

Thm: $A \perp L \mid b(L)$ $\Leftrightarrow$ $e(L) = f(b(L))$ for some function f

Ex: $b(L) = L$ is the finest balancing score
$b(L) = e(L)$ is the coarsest.

proof ⇑ Suppose $\exists f$ s.t. $e(L) = f(b(L))$
We want to show that $A \perp L \mid b(L)$.
It is sufficient to show that
$\mathbb{E}(A \mid b(L)) = e(L)$
since $e(L) = \mathbb{E}(A \mid L)$
$= \mathbb{E}(A \mid L, b(L))$.
This would then imply that
$\mathbb{E}(A \mid b(L)) = \mathbb{E}(A \mid L, b(L))$
i.e. $A \perp L \mid b(L)$.

$$\mathbb{E}(e(L) \mid b(L)) = \mathbb{E}\big(\mathbb{E}(A \mid L) \mid b(L)\big)$$
$$= \mathbb{E}\big(\underbrace{\mathbb{E}(A \mid L, b(L))}_{\text{finer averaging}} \mid b(L)\big)$$
$$= \mathbb{E}(A \mid b(L))$$

On the other hand,

$$\mathbb{E}(e(L) \mid b(L)) = \mathbb{E}(f(b(L)) \mid b(L))$$
$$= f(b(L)) = e(L),$$

so that $\mathbb{E}(A \mid b(L)) = e(L)$ indeed.

⇓ Suppose that $L \perp A \mid b(L)$.

By contradiction, suppose that $\exists \ell_1, \ell_2$ s.t.

$e(\ell_1) \neq e(\ell_2)$

but $b(\ell_1) = b(\ell_2)$

By definition of e,

$$\mathbb{P}(A=1 \mid L=\ell_1) \neq \mathbb{P}(A=1 \mid L=\ell_2)$$

$$\| \qquad\qquad\qquad\qquad \|$$

$$\mathbb{P}(A=1 \mid L=\ell_1, b(L)=b(\ell_1)) \qquad \mathbb{P}(A=1 \mid L=\ell_2, b(L)=b(\ell_2))$$

$\Rightarrow L \not\perp A \mid b(L)$; a contradiction. ∎

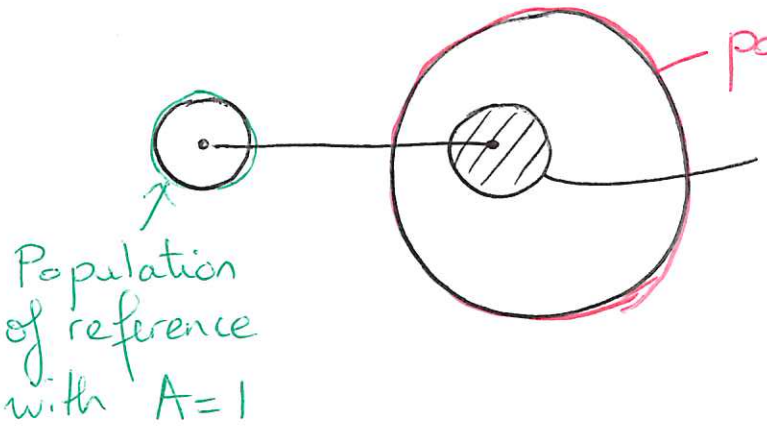
- Propensity Standardization = Adjust for $e(L)$ instead of L.

$$\mathbb{E}\,Y^a = \sum_e \mathbb{P}(Y=y \mid A=a, e(L)=e)\,\mathbb{P}(e(L)=e)$$

(if discrete)

- **Propensity Matching.**

Recall that matching rebalances the distribution of the variable L within the two groups. Depending on how the matching is performed, the quantity of interest is the ATT or ATNT (see p. 20)



In this case, we estimate the $ATT = \mathbb{E}(Y^1 - Y^0 \mid A=1)$

Instead of matching based on L, we may match based on $e(L)$

$$ATT = \mathbb{E}(Y^1 \mid A=1) - \mathbb{E}(Y^0 \mid A=1) \quad \text{(consistency)}$$

$$= \underbrace{\mathbb{E}(Y \mid A=1)}_{\text{easily estimated}} - \underbrace{\mathbb{E}(Y^0 \mid A=1)}_{\text{counterfactual term}}$$

$$\downarrow \qquad\qquad \downarrow$$

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \mathbb{1}(A_i=1) \qquad \frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i^0 \mathbb{1}(A_i=1)$$

i = i-th individual

where $\hat{Y}_i^0$ is an estimate of the potential outcome for individual i.

$$\hat{Y}_i^0 = \frac{1}{K}\sum_{k \in K_i} Y_k,$$

where K_i contains the K closest individuals in the control group based on their $e(L)$ value.

- **Stratification.**

↳ After computing the PS $e(L)$, the sample is divided into strata / blocks such that within each block, the PS is approximatively constant.

Consider J intervals $(b_{j-1}, b_j]$. We define

$$s(L) = j \quad \text{if} \quad \log\left\{\frac{e(L)}{1-e(L)}\right\} \in (b_{j-1}, b_j]$$

Rubin (2001)

$$\begin{aligned}
ATT &= \mathbb{E}(Y^1 - Y^0 \mid A=1) \\
&= \mathbb{E}\Big(\underbrace{\mathbb{E}(Y^1 - Y^0 \mid A=1, s(L))}_{\| \; ATT(s(L))} \mid A=1\Big) \\
&= \mathbb{E}\big(ATT(s(L)) \mid A=1\big) \\
&= \sum_s \underbrace{ATT(s)}_{\text{estimated using}} \underbrace{\mathbb{P}(S(L)=s \mid A=1)}_{\approx \text{ number of individuals in the treated group with } s(L)=s \text{ divided by the number of individuals in the treated group.}}
\end{aligned}$$

estimated using

$$\frac{1}{n_{1s}} \sum_{i=1}^{n} Y_i \mathbb{1}(A_i=1)\mathbb{1}(s(L_i)=s)$$

$$-\frac{1}{n_{0s}} \sum_{i=1}^{n} Y_i \mathbb{1}(A_i=0)\mathbb{1}(s(L_i)=s)$$

↑ direct comparison of the two groups since the PS are almost the same, rebalancing the confounders L.

$$n_{as} = \sum_{i=1}^{n} \mathbb{1}(A_i=a)\mathbb{1}(s(L_i)=s) = \#\text{ observations with treatment allocation } a \text{ within stratum } s.$$

x Summary of part I

Interested in identifying the causal effect of A on Y

$$ATE = \mathbb{P}(Y^{a=1}=1) - \mathbb{P}(Y^{a=0}=1).$$

We may also be interested in the average causal effect amongst the treated individuals

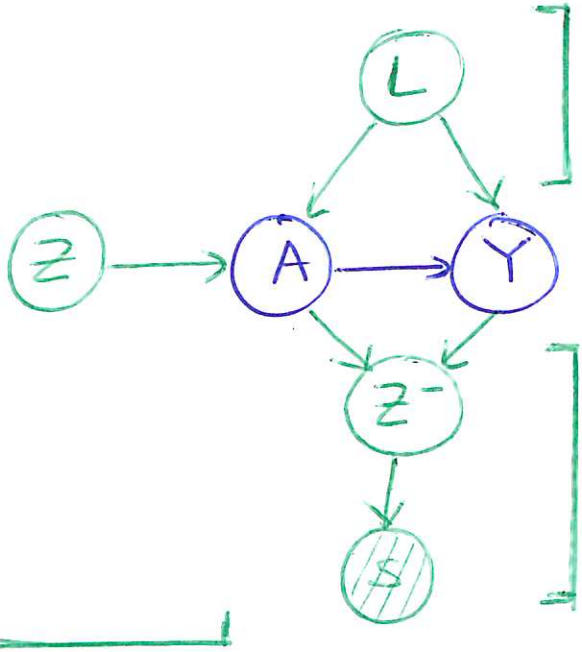
$$ATT = \mathbb{P}(Y^{a=1}=1 \mid A=1) - \mathbb{P}(Y^{a=0}=1 \mid A=1)$$

In presence of confounders, adjust via the backdoor formula

$\mathbb{P}(Y^a=y \mid A=1)$ (p. 6/21)

$= \sum_\ell \mathbb{P}(Y=y \mid A=a, L=\ell)\mathbb{P}(L=\ell \mid A=1)$

if L blocks all backdoor paths between A and Y. (p. 31)



In presence of selection bias, causal effects may be recovered with or without external data. When external data are available, and when the selection backdoor criterion is satisfied (p. 54),

$\mathbb{P}(Y^a=y \mid A=1)$

$= \sum_{\text{▨}} \mathbb{P}(Y=y \mid A=a, \text{▨}, S=1) \times \mathbb{P}(\text{▨} \mid A=1)$

Under partial complier, consider instead the ITT (p. 58) or the LATE/CACE (p. 62)