THE UNIVERSITY of EDINBURGH

Methods for Causal Inference
Lecture 19: Revision

Ava Khamseh

School of Informatics
2025-2026

For the exam

1. Candidates may consult up to THREE A4 pages (6 sides) of notes.

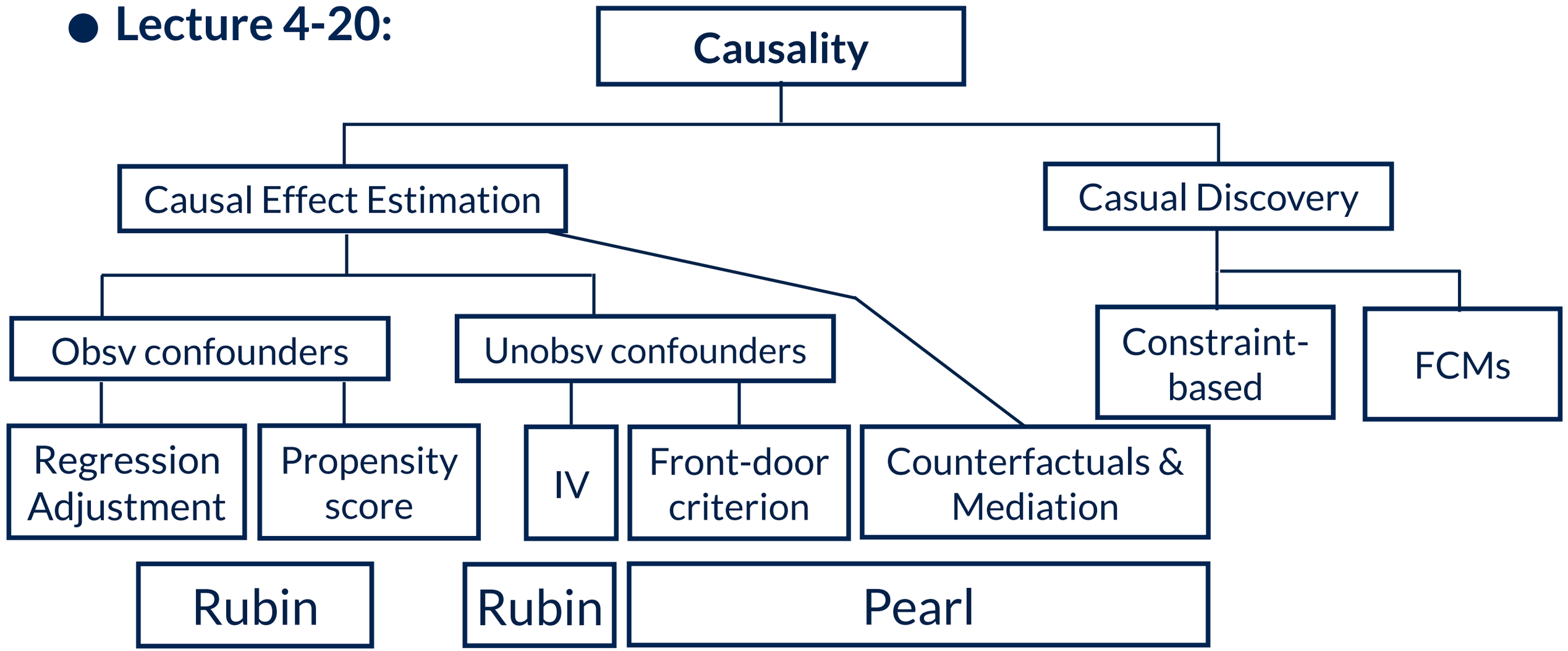
2. CALCULATORS MAY BE USED IN THIS EXAMINATION, please bring your own calculators (they will **not** be provided).

Learning Outcomes

1. Explain the difference between **causal** and **associational** estimation and justify why causal inference techniques are necessary to derive meaning from observational data
2. Explain the difference between **randomised** trials vs **observational** studies related to public health and other types of data more generally
3. **Learn and apply foundational causal identification & estimation techniques** using two major frameworks: (i) **Rubin**'s Potential Outcomes and (ii) **Pearl**'s Structural (graphical) causal models to simulated examples and real world data, in the presence of observed and unobserved variables
4. Explain different types of causal discovery algorithm, learn their underlying assumptions and short-comings, and be able to apply them to data using available software.
5. Be able to modify/repurpose a current technique in order to apply it to a particular problem of interest.

Overview of the course

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**



Causal Inference

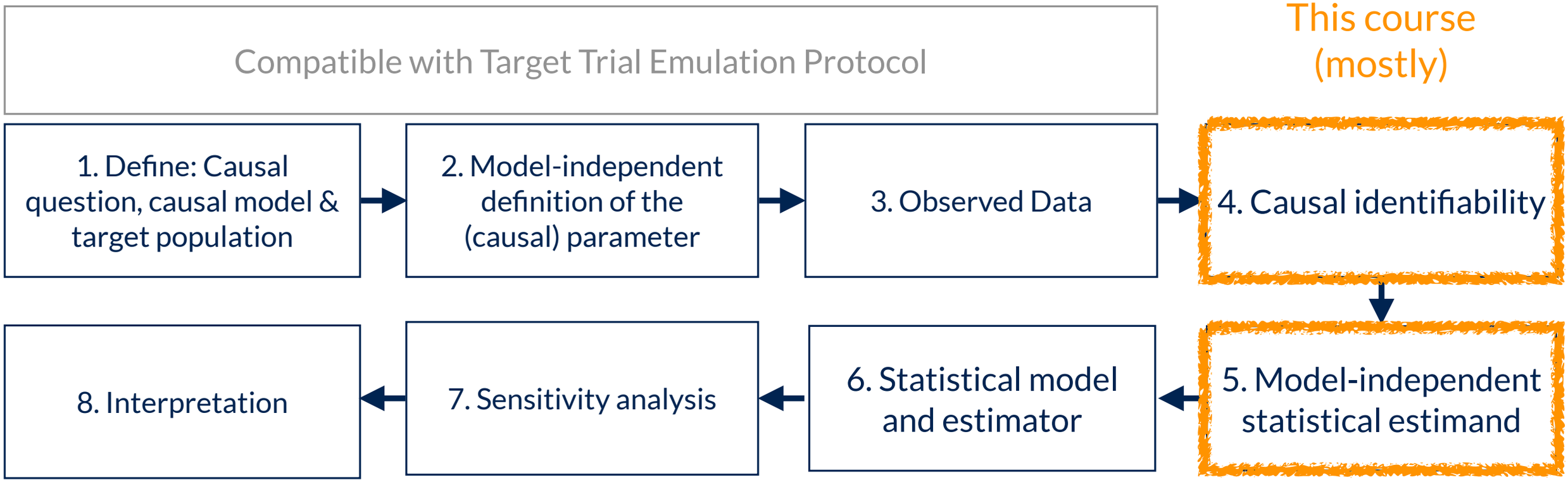
- **Model** a causal inference problem with assumptions manifest in Causal Graphical Models [**Pearl]**

- **Identify** an expression for the causal effect under these assumptions ("causal estimand"), [**Pearl**]

- **Estimate** the expression using statistical methods such as matching or instrumental variables, [**Rubin**'s Potential Outcomes]

- **Verify** the validity of the estimate using a variety of robustness checks.

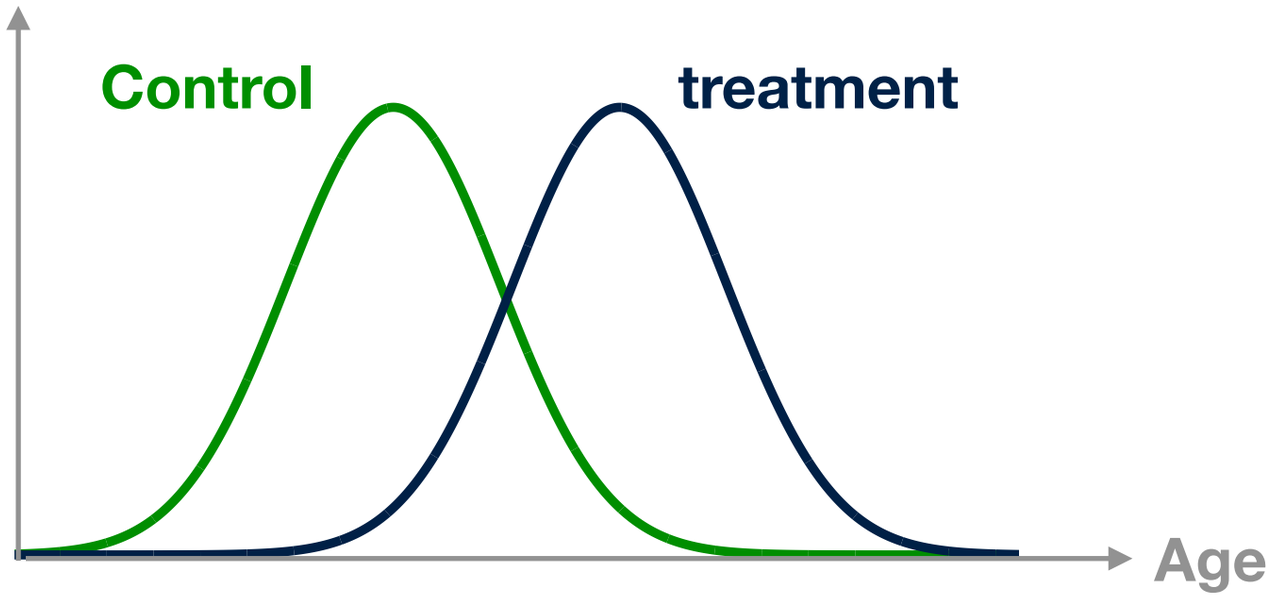
The Causal Roadmap



A causal roadmap for generating high-quality real-world evidence
https://doi.org/10.1017/cts.2023.635

Observational data: What goes wrong?

$$p(x|t=1) \neq p(x|t=0)$$



$$\left(\int y_1(x)p(x|t=1)dx \; - \int y_0(x)p(x|t=0)dx\right) \neq \int \big(y_1(x) - y_0(x)\big)p(x)dx$$

In contrast to randomised control trial.

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

$y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes
$t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1

Observed outcomes: $y_0^{(i)}$ **OR** $y_1^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)} = \begin{cases} y_0^{(i)} & \text{if } t^{(i)} = 0 \\ y_1^{(i)} & \text{if } t^{(i)} = 1 \end{cases}$$

Potential Outcomes Framework (Rubin-Neyman)

Definition: Given treatment, t, and outcome, y, the potential outcome of instance/individual i is denoted by $y_t^{(i)}$ is the value y *would have* taken if individual i had been under treatment t.

$y_0^{(i)}$ and $y_1^{(i)}$ are not **observed**, but **potential** outcomes
$t^{(i)}$ is the observed treatment applied to individual (i), 0 or 1

Observed outcomes: $y_0^{(i)}$ **OR** $y_1^{(i)}$ depend on treatment (**fundamental problem of causal inference**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)} = \begin{cases} y_0^{(i)} & \text{if } \; t^{(i)} = 0 \\ y_1^{(i)} & \text{if } \; t^{(i)} = 1 \end{cases}$$

Counterfactual (missing) outcome "what would have happened if …"

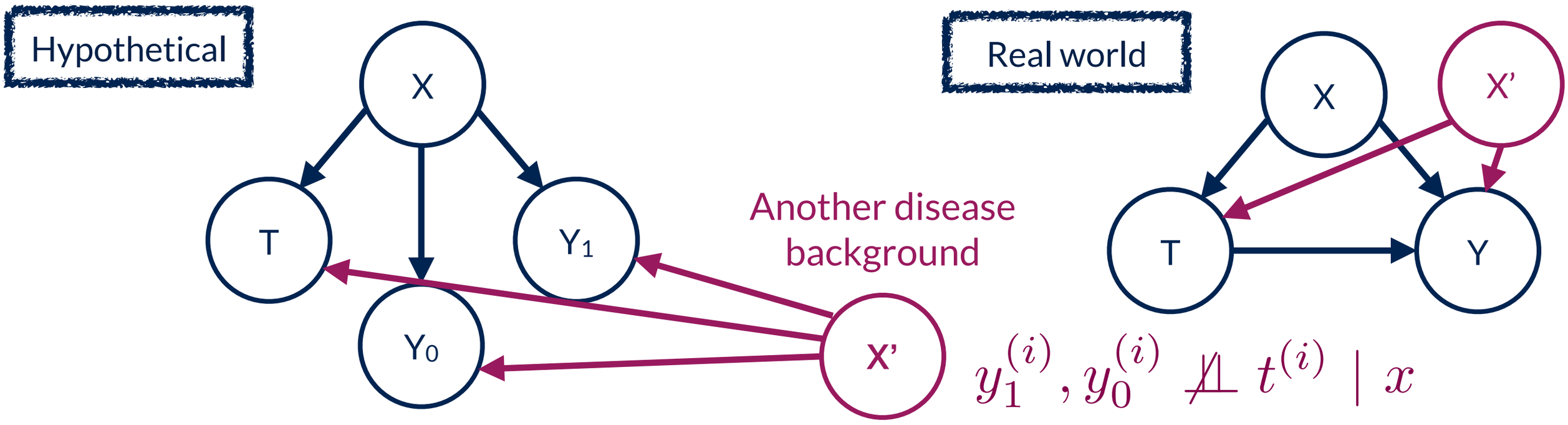
$$y_{CF}^{(i)} = (1 - t^{(i)}) y_1^{(i)} + t^{(i)} y_0^{(i)} = \begin{cases} y_1^{(i)} & \text{if } t^{(i)} = 0 \\ y_0^{(i)} & \text{if } t^{(i)} = 1 \end{cases}$$

Potential Outcomes Framework: Assumptions

SUTVA

Positivity $0 < P(T = 1|X = x) < 1$

Unconfoundedness $y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$



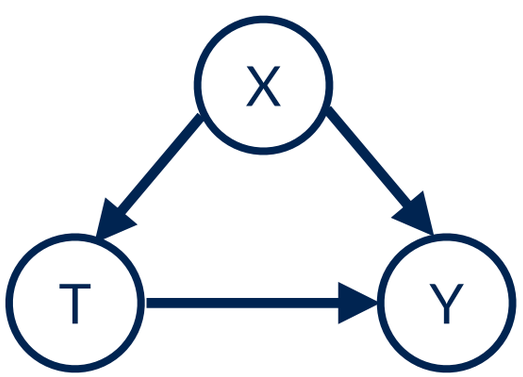
Observed Confounders: Regression Adjustment

Fit a model for $Q(T, X) = \mathbb{E}[Y|T, X]$

(we substituted T=1 and T=0 into individual treatment effect = $Q(1, x^{(i)}) - Q(0, x^{(i)})$, then took average over all individuals i, via linear regression). Under the linearity assumption:

$$\mathbb{E}[Y|T, X] = \alpha_0 + \beta_x X + \beta_t T + \epsilon \ , \ \mathbb{E}[\epsilon] = 0$$

$$\begin{aligned} ATE &= \mathbb{E}_X \Big[\mathbb{E}[Y|T=1, X] - \mathbb{E}[Y|T=0, X] \Big] \\ &= \Big(\alpha_0 + \beta_x \mathbb{E}[X] + \beta_t \Big) - \Big(\alpha_0 + \beta_x \mathbb{E}[X] \Big) \\ &= \beta_t \end{aligned}$$

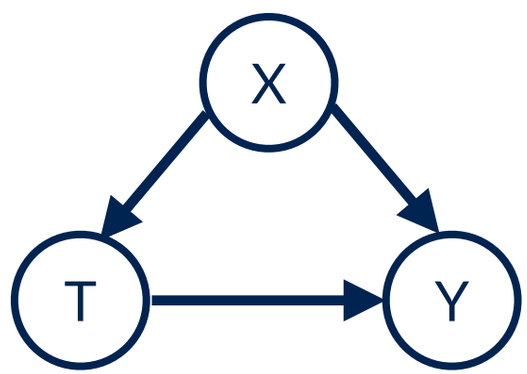


Important remarks about the previous form:

1) Depends on the structure of the causal graph of interest

2) Data need not be linear

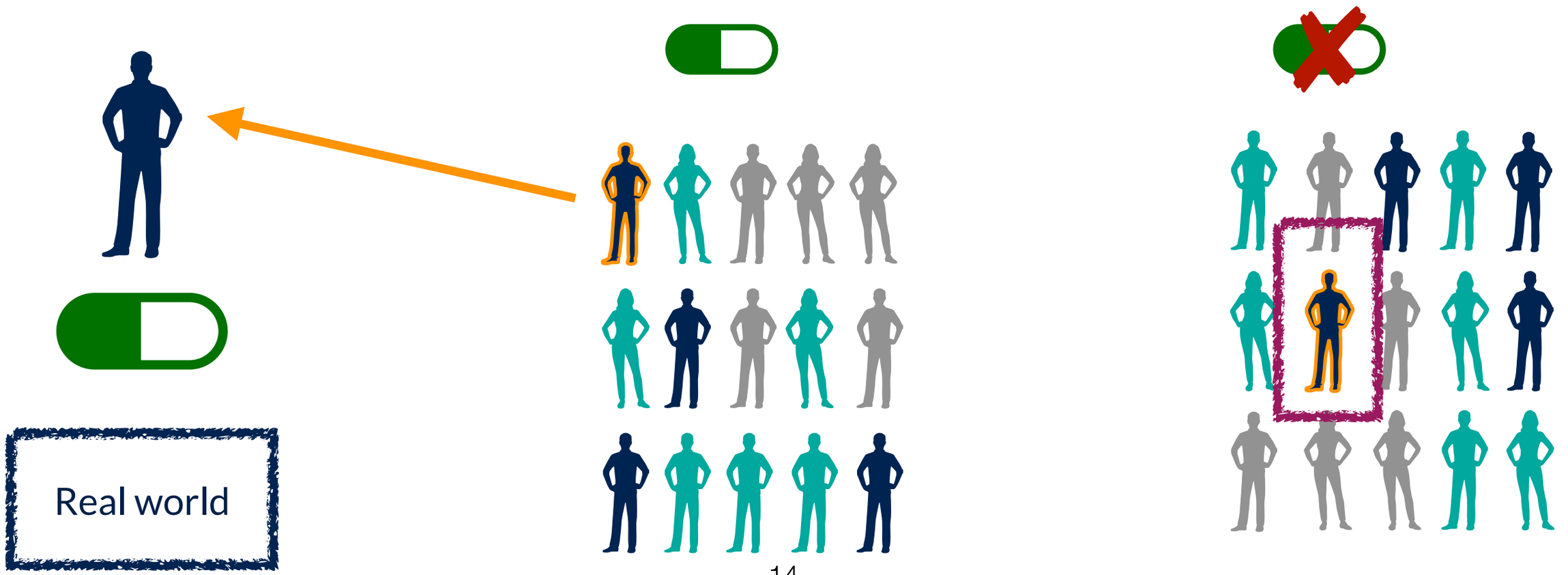
model-misspecification -> statistical bias



Observed Confounders: Propensity score

Propensity score matching $\boxed{e(x) = p(t = 1|x)}$

Need to know the notation of balancing score,
why propensity is the coarsest balancing score,
how to estimate and match propensity scores in principle



Unobserved Confounders (Part 1): IV

Instrumental variable approach

Assumptions and why they are necessary:

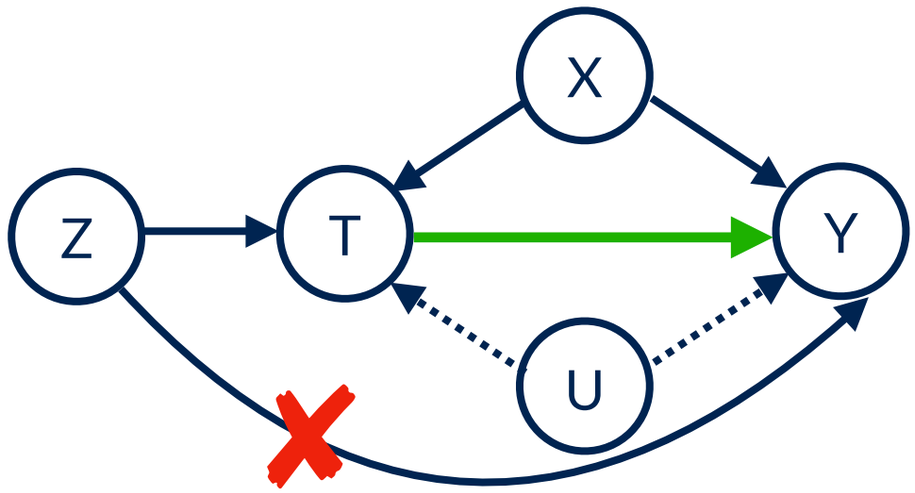
- SUTVA

- Exclusion restriction

- Non-zero average (Z to T association)

- Monotonicity

$$\tau = \frac{\mathbb{E}\left[(Y|z=1) - (Y|z=0)\right]}{\mathbb{E}\left[(T|z=1) - (T|z=0)\right]}$$



Unobserved Confounders (Part 1): IV

Instrumental variable approach

Assumptions and why they are necessary:

- SUTVA

- Exclusion restriction

- Non-zero average (Z to T association)

- Monotonicity

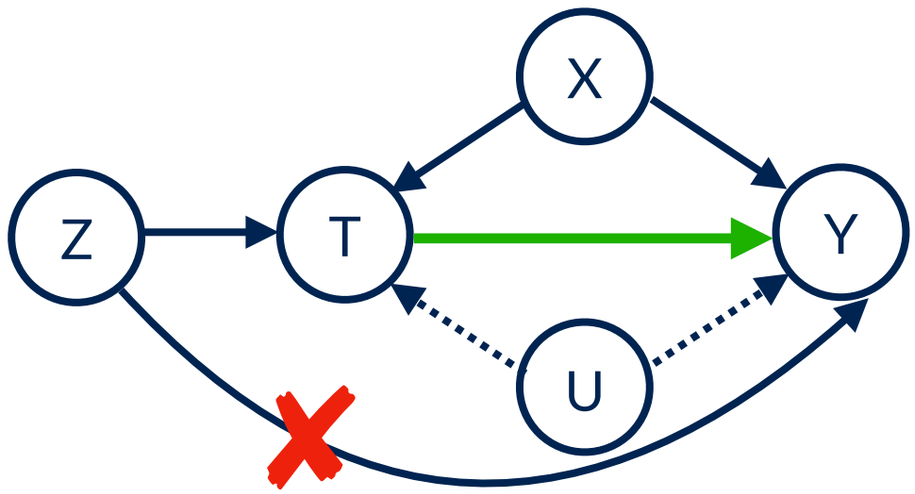
$$\tau = \frac{\mathbb{E}\left[(Y|z=1) - (Y|z=0)\right]}{\mathbb{E}\left[(T|z=1) - (T|z=0)\right]}$$

<u>Estimation:</u>

Binary case

Continuous case

- Ratio of Covs
- 2-Step regression



Other Causal estimators

ATE: $\mathbb{E}[Y_1 - Y_0]$

ATT: $\mathbb{E}[Y_1 - Y_0 | T = 1]$

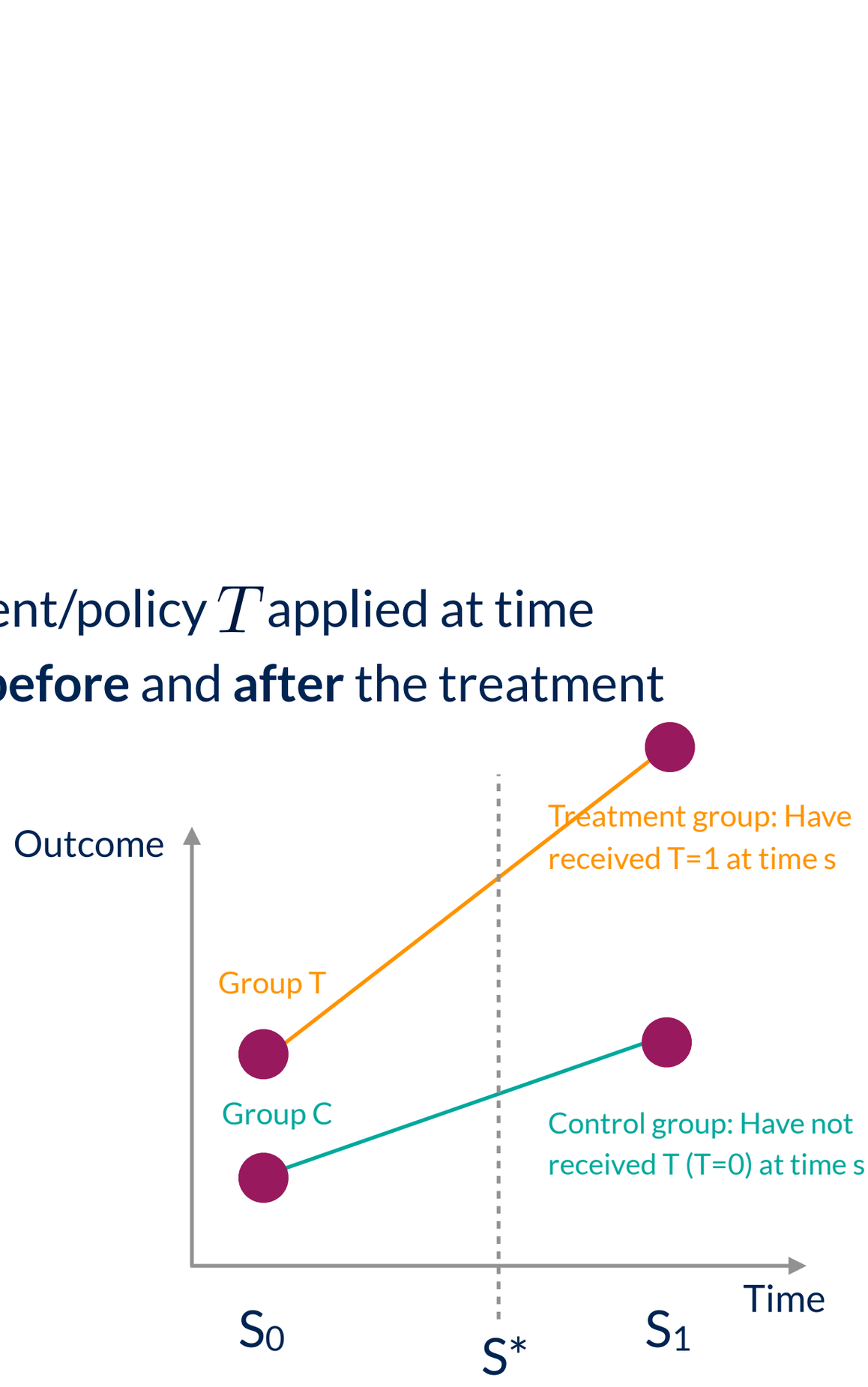
Other Causal estimators

ATE: $\mathbb{E}[Y_1 - Y_0]$

ATT: $\mathbb{E}[Y_1 - Y_0 | T = 1]$

Difference in difference:

We wish to estimate the impact of a treatment/policy T applied at time on some outcome Y by using information **before** and **after** the treatment



= measured (line is for visualisation only!)

Other Causal estimators

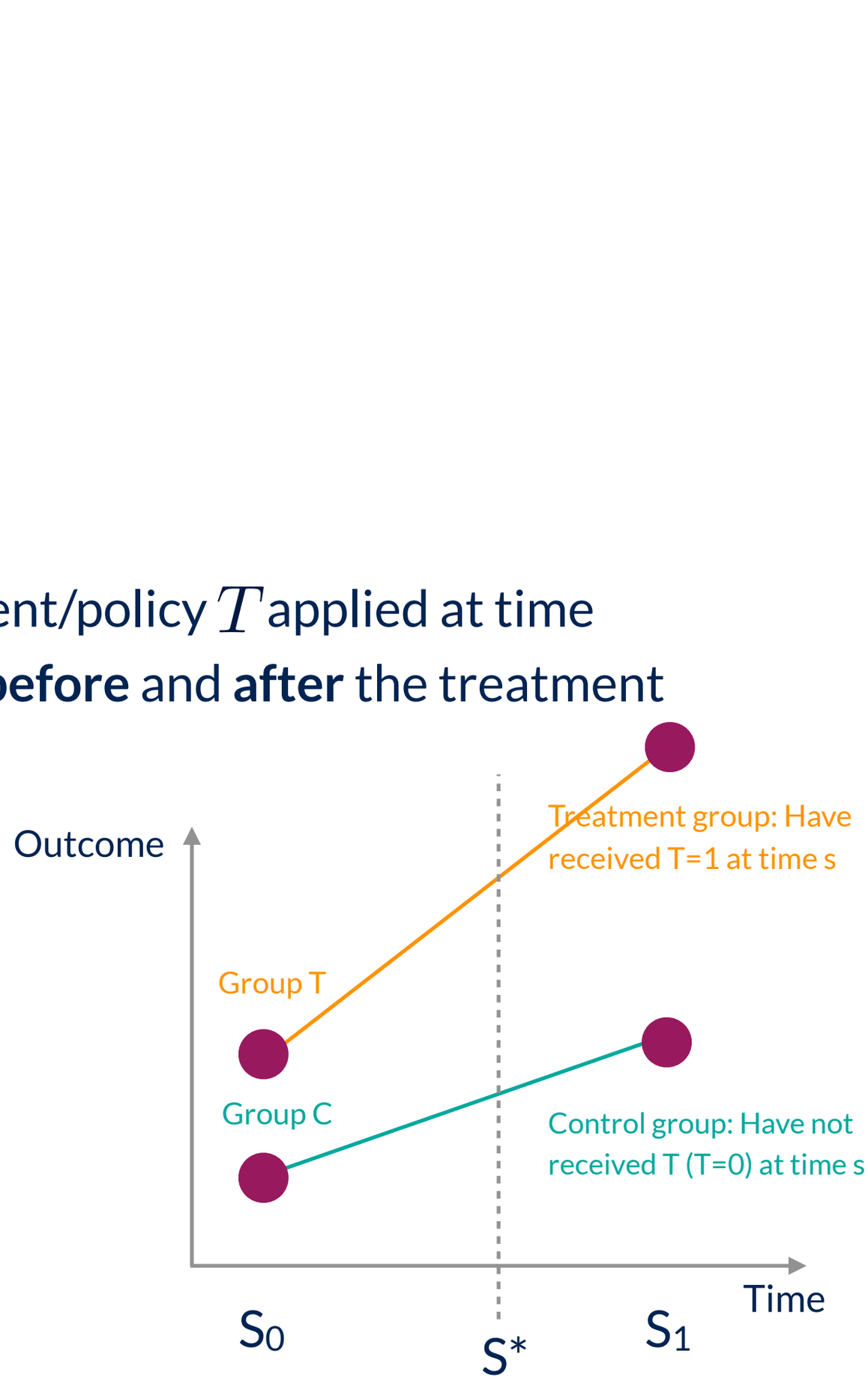
ATE: $\mathbb{E}[Y_1 - Y_0]$

ATT: $\mathbb{E}[Y_1 - Y_0|T = 1]$

Difference in difference:

We wish to estimate the impact of a treatment/policy T applied at time on some outcome Y by using information **before** and **after** the treatment

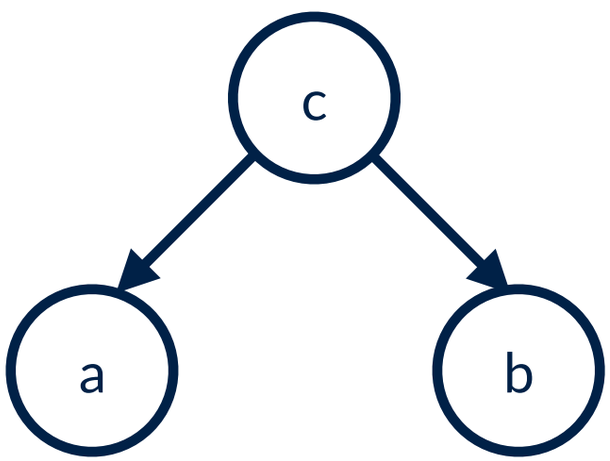
Regression discontinuity



= measured (line is for visualisation only!)

The notion of d-separation

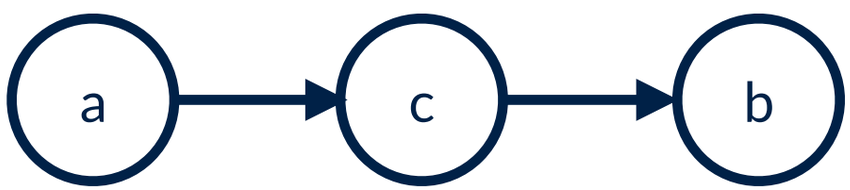
- Conditional independence via graphs and **D-separation**
- 3 main graph structures:



Fork

$a \not\perp\!\!\!\perp b|\emptyset$

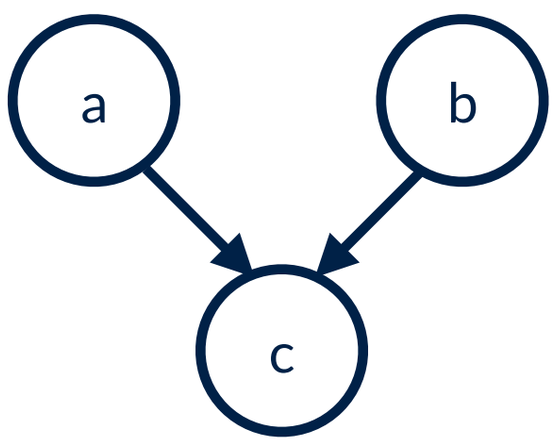
$a \perp\!\!\!\perp b|c$



Chain

$a \not\perp\!\!\!\perp b|\emptyset$

$a \perp\!\!\!\perp b|c$



Collider

$a \perp\!\!\!\perp b|\emptyset$

$a \not\perp\!\!\!\perp b|c$

Graphical models and do-calculus

Observation vs intervention vs counterfactual

(Conditioning vs applying the do-operation)

The adjustment formula (revisited)

How to get from:

$$p(Y = 1|do(T = 1)) - p(Y = 1|do(T = 0))$$

To

$$\boxed{p(Y = y|do(T = t)) = \sum_x p(Y = y|T = t, X = x)p(X = x)}$$

(Using the modified graph as a tool)

Graphical models and do-calculus

Observation vs intervention vs counterfactual

(Conditioning vs applying the do-operation)

The adjustment formula (revisited)

How to get from:

$$p(Y = 1|do(T = 1)) - p(Y = 1|do(T = 0))$$

To

$$p(Y = y|do(T = t)) = \sum_x p(Y = y|T = t, X = x)p(X = x)$$

(Using the modified graph as a tool)

The Backdoor Criterion

Under what conditions does a causal model permit computing the causal effect of one variable on another, from **data** obtained from **passive observations**, with **no intervention**? i.e.,
Under what conditions is the structure of a causal graph sufficient of computing a causal effect from a given data set? **Identifiability**

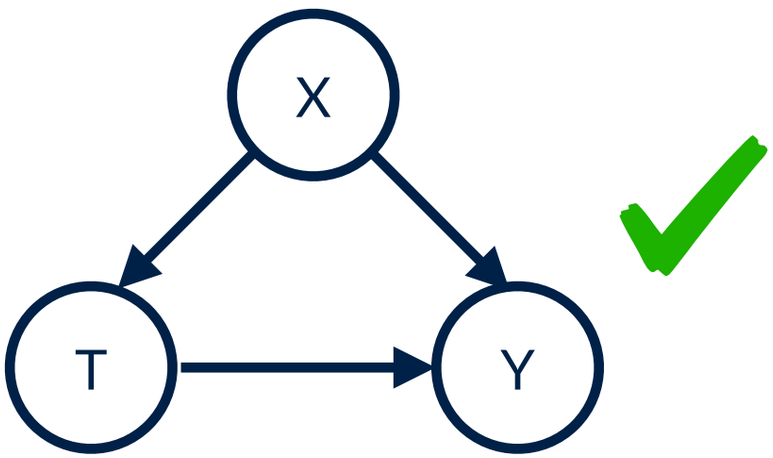
Backdoor Criterion: Given an ordered pair of variables (T,Y) in a DAG G, a set of variables X satisfies the backdoor criterion relative to (T,Y) if:

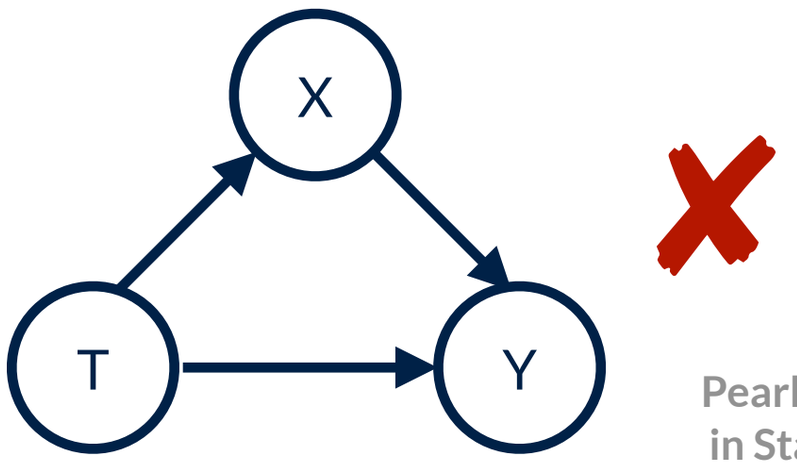
(i) no node in X is a descendent of T

(ii) X block every path between T and Y that contains an arrow into T

If X satisfies the backdoor criterion then the causal effect of T on Y is given by:

$$p(Y = y|do(T = t)) = \sum_x p(Y = y|T = t, X = x)p(X = x)$$





Pearl, Causal Inference in Statistics (2016)

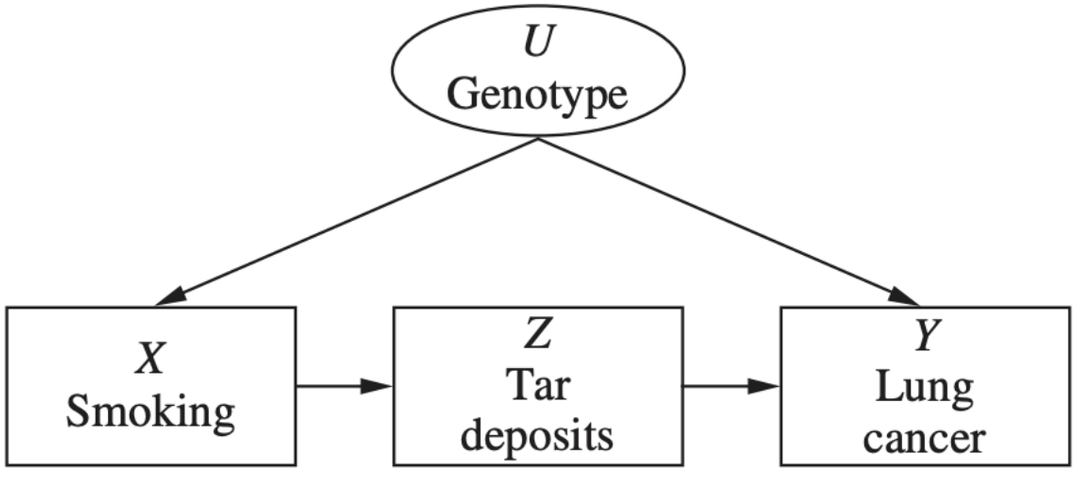
Unobserved Confounders (Part 2): Front-door

$$p(Y = y|do(X = x)) = \sum_z \sum_{x'} p(Y = y|Z = z, X = x')p(X = x')p(Z = z|X = x)$$

To compute ATE:

$$p(Y = 1|do(X = 1)) - p(Y = 1|do(X = 0))$$

(Understand derivation and example, why useful)



Pearl, Causal Inference in Statistics (2016)

Do-Calculus Rules

Let X, Y, Z, W be arbitrary disjoint sets of nodes in a DAG G

Rule 1 (insertion/deletion of observations):

$$p(Y|do(X=x),Z,W) = p(Y|do(X=x),W) \text{ if } (Y \perp\!\!\!\perp Z)|X,W \text{ in } G_{\overline{X}}$$

Rule 2 (Action/observation exchange):

$$p(Y|do(X=x),do(Z=z),W) = p(Y|do(X=x),z,W) \text{ if } (Y \perp\!\!\!\perp Z)|X,W \text{ in } G_{\overline{X}\underline{Z}}$$

Rule 3 (Insertion/deletion of actions):

$$p(Y|do(X=x),do(Z=z),W) = p(Y|do(X=x),W) \text{ if } (Y \perp\!\!\!\perp Z)|X,W \text{ in } G_{\overline{X}\overline{Z(W)}}$$

Provides conditions for introducing/deleting an external intervention without affecting the conditional probability of Y.

Proof in Pearl 1995a

Counterfactual and Mediation

Counterfactuals

Attribution

Probability of necessity

Mediation

Mediation: CDE

Controlled Direct Effect (CDE):

$$p(Y = y|do(T = t), do(X = x)) - p(Y = y|do(T = t'), do(X = x))$$

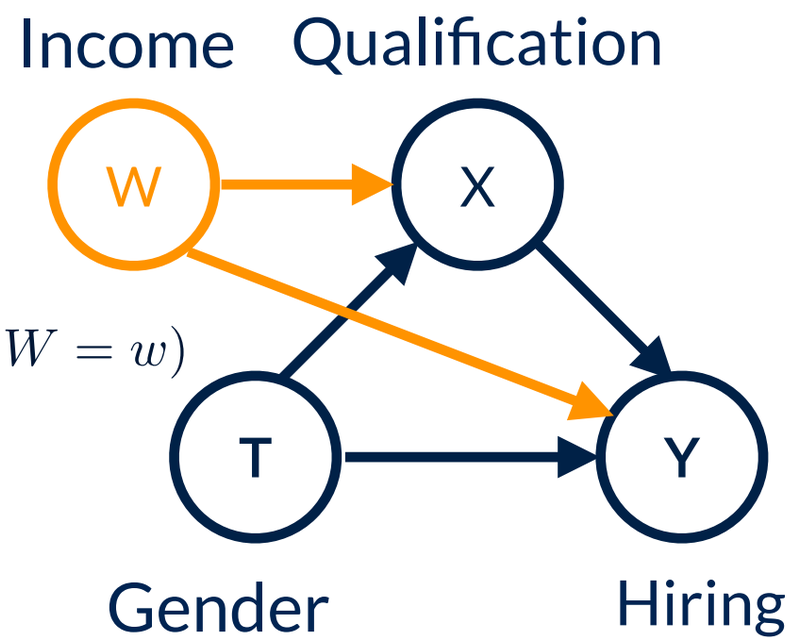
There are no backdoor paths from T to Y, hence the above is equal to:

$$p(Y = y|T = t, do(X = x)) - p(Y = y|T = t', do(X = x))$$

There are 2 back-door paths from X to Y in the original graph:

1) through gender T, which is blocked by T
2) Through income W, so we condition on W

$$\sum_w \Big(p(Y = y|T = t, X = x, W = w) - p(Y = y|T = t', X = x, W = w)\Big)p(W = w)$$



Mediation and Path-disabling Interventions

Example 4.4.5 *A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training. The former concerns the "direct effect" of gender on hiring, whereas the latter concerns the "indirect effect," or the effect* mediated *via job qualification.*

Aim: Which of the two causal effects is greater (i) the direct effect (gender on hiring), or (ii) the indirect effect (education on job qualification on hiring)?
—> Could inform policy where to invest resources to address disparity

This concerns enabling/disabling processes (e.g., educational reforms) rather than lowering/raising values of specific variables. Thus, the do-operator and the controlled direct effect (CDE) seen earlier do not suffice …

… as before, we phrase the problem mathematically via counterfactuals!

Pearl's Primer book Chapter 4, page 114

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

The result is $\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q|X = 0)$

Male applicants have similar chances, but averaging over $p(Q = q|X = 1)$

Mediation and Path-disabling Interventions

Since Q varies over the population, we average this quantity according to the distribution of the qualification of female applicants, $p(Q = q|X = 0)$

The result is $\sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] p(Q = q|X = 0)$

Male applicants have similar chances, but averaging over $p(Q = q|X = 1)$

Subtracting the two quantities yields the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q:

$$\text{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right] \left(p(Q = q|X = 0) - p(Q = q|X = 1)\right)$$

Allow Q to vary naturally between applicants, as opposed to the CDE. Here we disable the capacity of Y to respond to X but leave its response to Q unaltered.

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

$$\mathrm{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

The following result is known as Pearl's ***Mediation formula***

Pearl's Primer book Chapter 4, page 115

Mediation and Path-disabling Interventions

It remains to identify the *Natural Indirect Effect (NIE)* of gender on hiring, mediated by the level of qualification Q, in order to allow estimation:

$$\mathrm{NIE} = \sum_q \mathbb{E}\left[Y_{X=1,Q=q}\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

The following result is known as Pearl's ***Mediation formula***

Theorem (Pearl, 2001)

In the absence of confounding, the NIE can be identified as follows

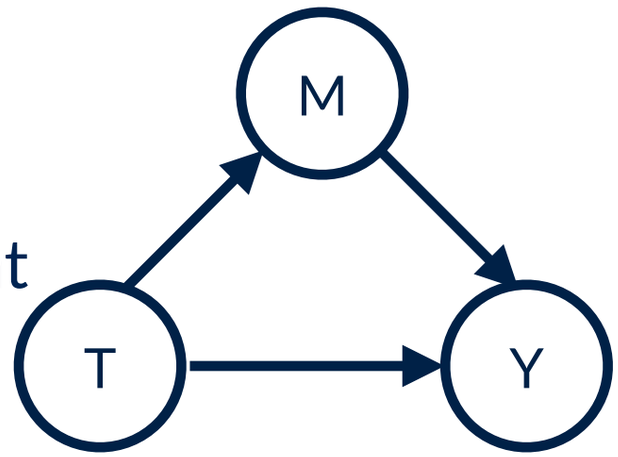
$$\mathrm{NIE} = \sum_q \mathbb{E}\left[Y|X=1,Q=q\right]\left(p(Q=q|X=0) - p(Q=q|X=1)\right)$$

In words: It measures the extent to which the effect of X on Y is *explained* by its effect on the mediator Q. In the NIE we "freeze" the direct effect of X on Y, yet allow the mediator Q of each unit to react to X in a natural "unfrozen" way.

Pearl's Primer book Chapter 4, page 115

Mathematical toolkit: Mediation

Four types of effects when we go from T=0 to T=1:



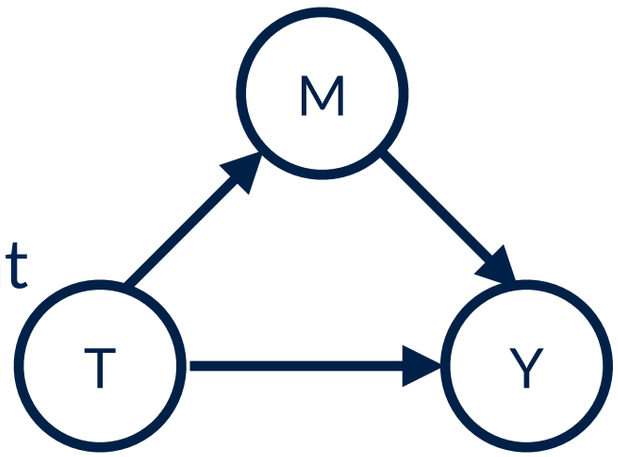
1. **Total effect (TE):** Measures the increase in Y as treatment changes from T=0 to T=1 while mediator M changes freely as per the structural function f_M

$$\begin{aligned}\mathrm{TE} &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[Y|\mathrm{do}(T=1)] - \mathbb{E}[Y|\mathrm{do}(T=0)]\end{aligned}$$

Pearl's Primer book Section 4.5.2

Mathematical toolkit: Mediation

Four types of effects when we go from T=0 to T=1:



1. **Total effect (TE):** Measures the increase in Y as treatment changes from T=0 to T=1 while mediator M changes freely as per the structural function f_M

$$\begin{aligned}\mathrm{TE} &= \mathbb{E}[Y_1 - Y_0] \\ &= \mathbb{E}[Y|\mathrm{do}(T=1)] - \mathbb{E}[Y|\mathrm{do}(T=0)]\end{aligned}$$

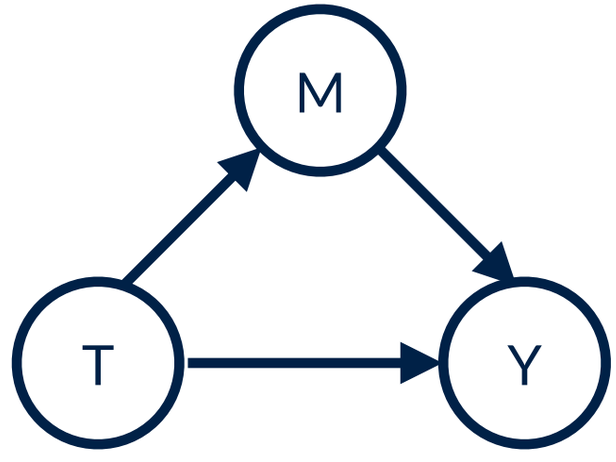
2. **Controlled direct effect (CDE(m)):** Measures the expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to M = m uniformly

$$\begin{aligned}\mathrm{CDE} &= \mathbb{E}[Y_{1,m} - Y_{0,m}] \\ &= \mathbb{E}[Y|\mathrm{do}(T=1, M=m)] - \mathbb{E}[Y|\mathrm{do}(T=0, M=m)]\end{aligned}$$

Pearl's Primer book Section 4.5.2

Mathematical toolkit: Mediation

Four types of effects when we go from T=0 to T=1:



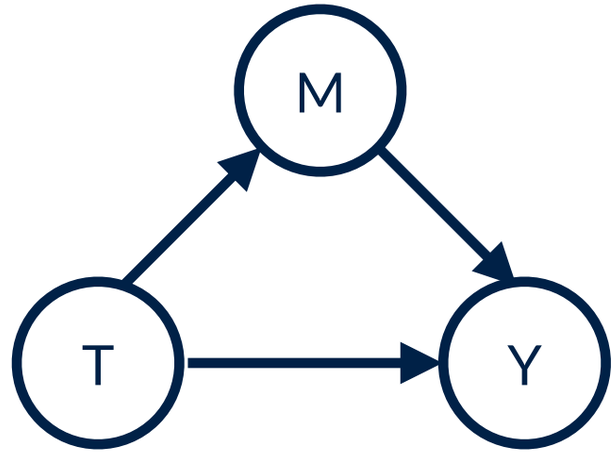
3. **Natural direct effect (NDE):** Measures expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to whatever value it *would have attained* (for each individual) prior to change, that is, under T = 0.

$$\text{NDE} = \mathbb{E}\left[Y_{1,M_0} - Y_{0,M_0}\right]$$

Pearl's Primer book Section 4.5.2

Mathematical toolkit: Mediation

Four types of effects when we go from T=0 to T=1:



3. **Natural direct effect (NDE):** Measures expected increase in Y as treatment changes from T=0 to T=1 while mediator is set to whatever value it *would have attained* (for each individual) prior to change, that is, under T = 0.

$$\mathrm{NDE} = \mathbb{E}\left[Y_{1,M_0} - Y_{0,M_0}\right]$$

4. **Natural indirect effect (NIE):** Measures the expected increase in Y when the treatment is held constant at T=0 and the mediator M changes to whatever value it *would have attained* (for each individual) under T=1

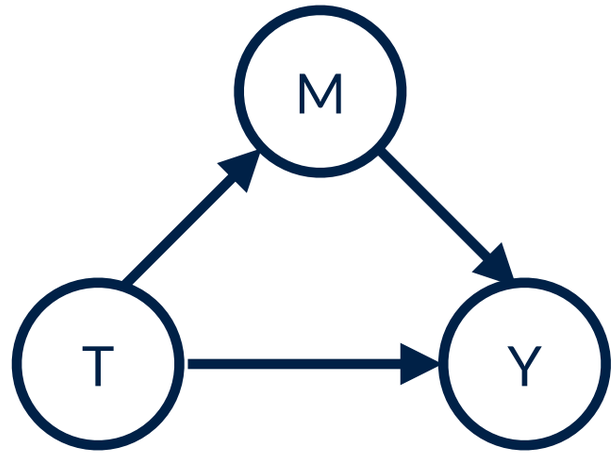
$$\mathrm{NIE} = \mathbb{E}\left[Y_{0,M_1} - Y_{0,M_0}\right]$$

It captures the portion of the effect that can be explained by mediation alone, while disabling (or "freezing") the capacity of Y to respond to T

Pearl's Primer book Section 4.5.2

Mathematical toolkit: Mediation

Some remarks on these four types of effects



1. TE and CDE(m) are *do*-expressions so can be estimated from experimental data or observational studies using the backdoor and front-door criteria

2. NDE and NIE are **not** *do*-expressions, so their causal identifiability will require a new set of results and, possibly, further assumptions

Pearl's Primer book Section 4.5.2

Mathematical toolkit: Mediation

Under certain conditions A, B, C, D, and E, described in the lecture:

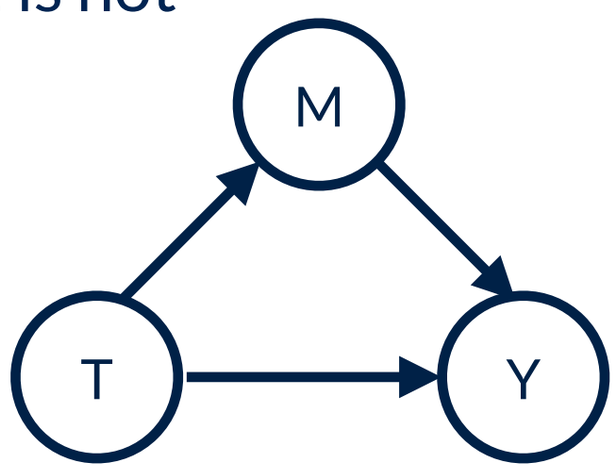
$$\mathrm{NDE} = \sum_m \big[\mathbb{E}[Y|T=1, M=m] - \mathbb{E}[Y|T=0, M=m]\big] p(M=m|T=0)$$

and, similarly,

$$\mathrm{NIE} = \sum_m \mathbb{E}[Y|T=0, M=m] \big[p(M=m|T=1) - p(M=m|T=0)\big]$$

These two expressions are known as the *mediation formulas*

Note that NDE is a weighted average of CDE(m), whereas NIE is not



Pearl's Primer book Section 4.5.2

Causal discovery

D-separation

PC algorithm:

- Markov condition
- Causal sufficiency
- Faithfulness

Notion of Markov Equivalence Classes and Markov Blanket

Overview of the course

- **Lecture 1**: Introduction & Motivation, why do we care about causality? Why deriving causality from observational data is non-trivial.
- **Lecture 2:** Recap of probability theory, variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lecture 4-20:**