Methods for Causal Inference
Lecture 19: Revision

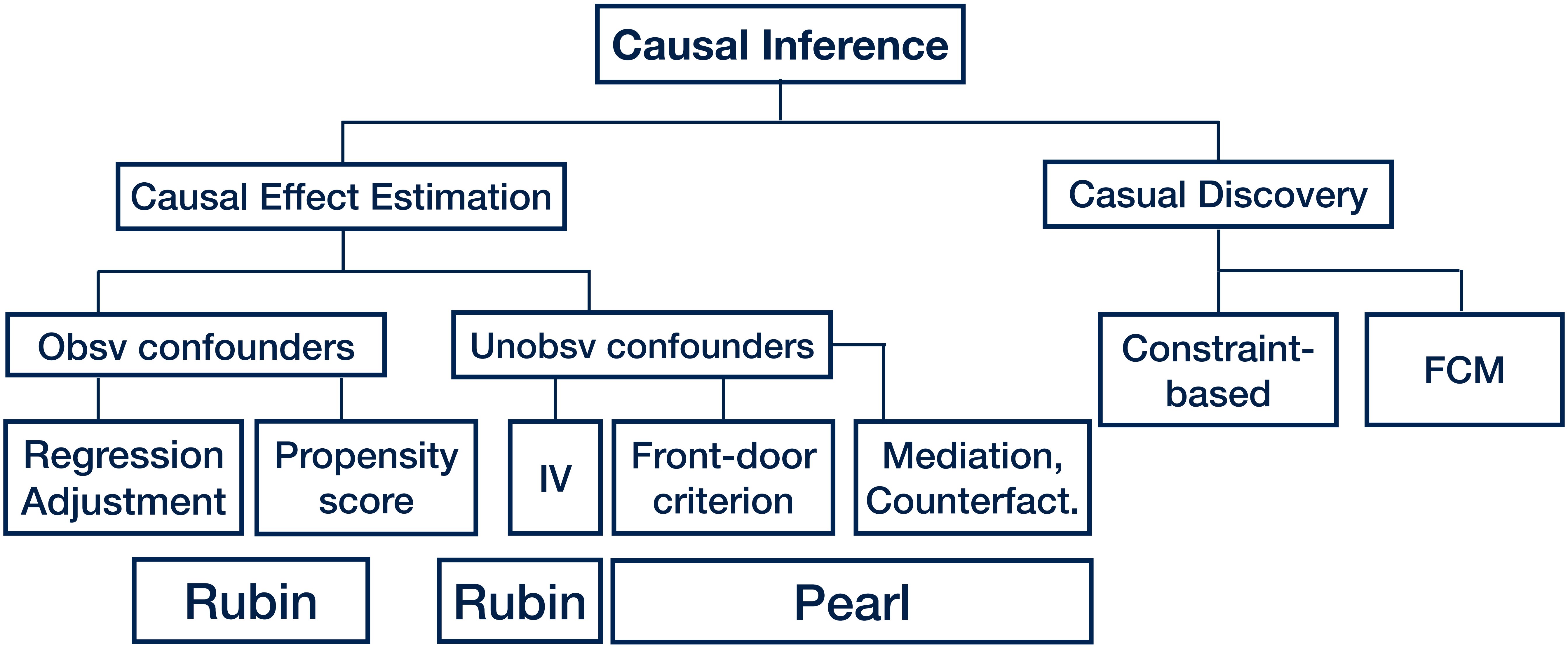
Ava Khamseh
School of Informatics

2021-2022

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 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.

- **Lecture 1**: Introduction & motivation, why do we care about causality?
- **Lecture 2**: Recap of probability theory, e.g., variables, events, conditional probabilities, independence, law of total probability, Bayes' rule
- **Lecture 3**: Recap of regression, multiple regression, graphs, SCM
- **Lectures 4-20**:

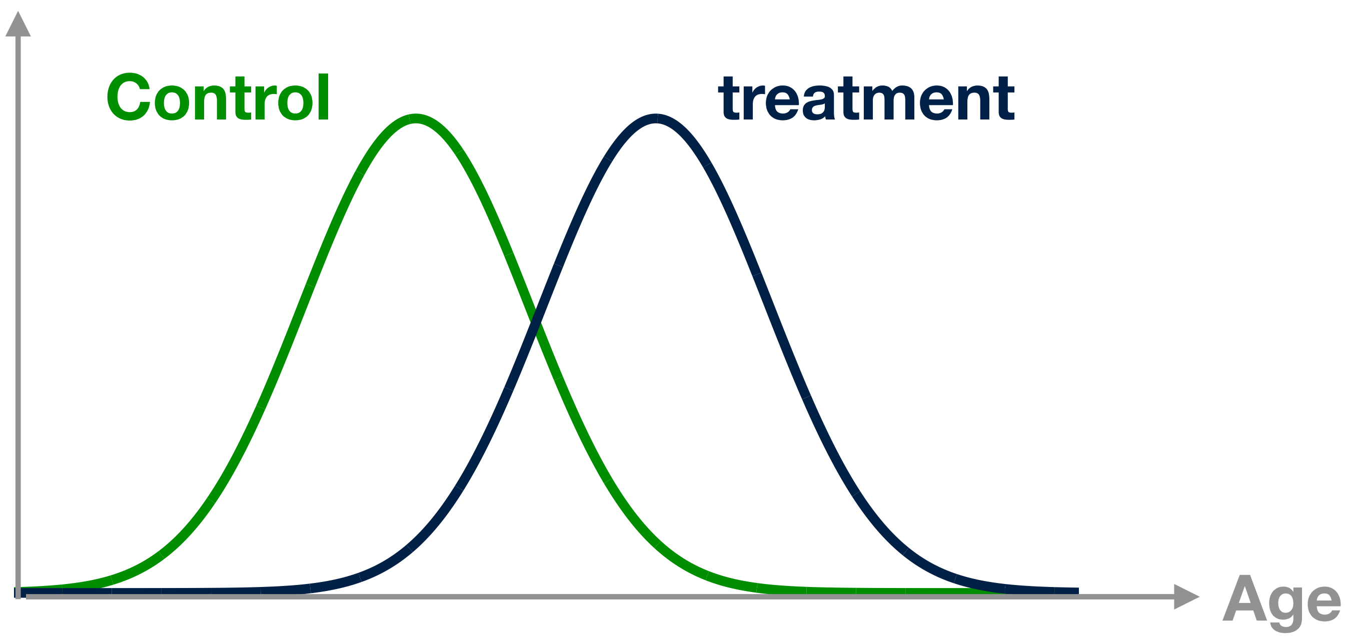


Causal Inference

- **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 regression adjustment, matching, instrumental variables, front-door, …
- **Verify** the validity of the estimate using a variety of robustness checks. (sensitivity analysis)

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: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ 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: $y_{\underline{0}}^{(i)}$ **OR** $y_{\underline{1}}^{(i)}$ 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 …"

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

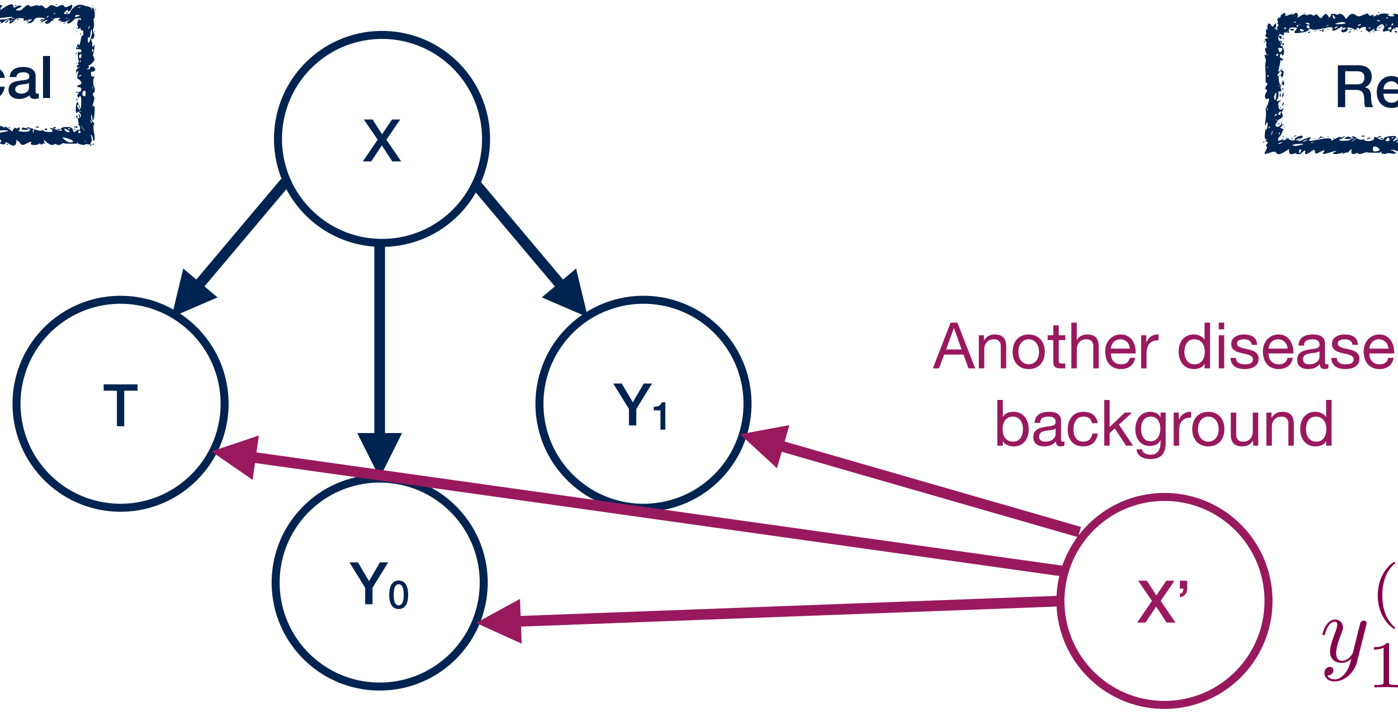
Potential Outcomes Framework: Assumptions

SUTVA

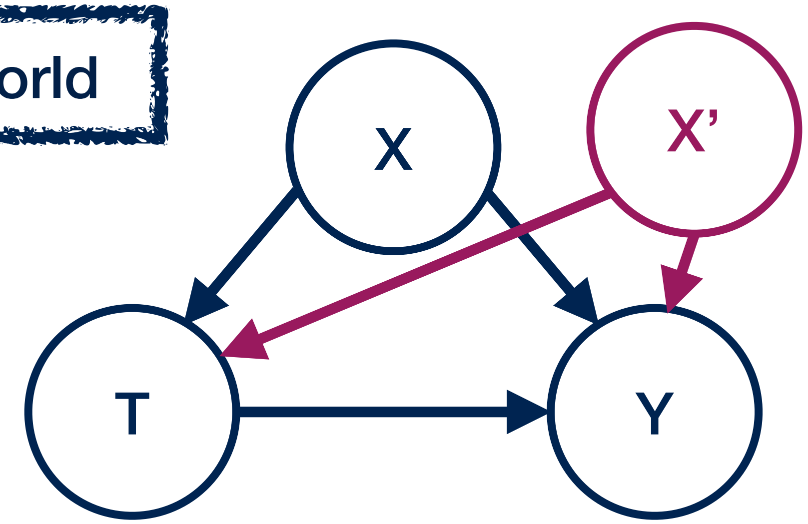
Positivity $0 < P(T = 1|X = x) < 1$

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

Hypothetical



Real world



$y_1^{(i)}, y_0^{(i)} \not\!\perp\!\!\!\perp t^{(i)} \mid x$

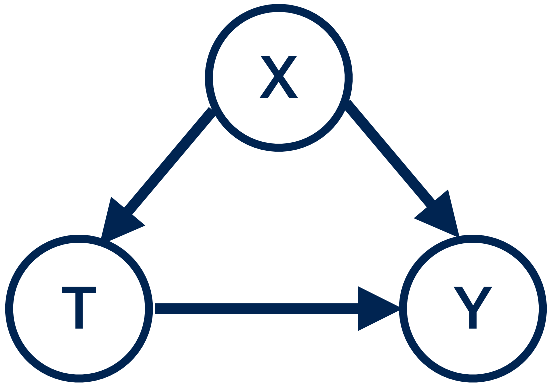
Observed Confounders: Regression Adjustment

Regression/Covariant adjustment

(Linear model example:)

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

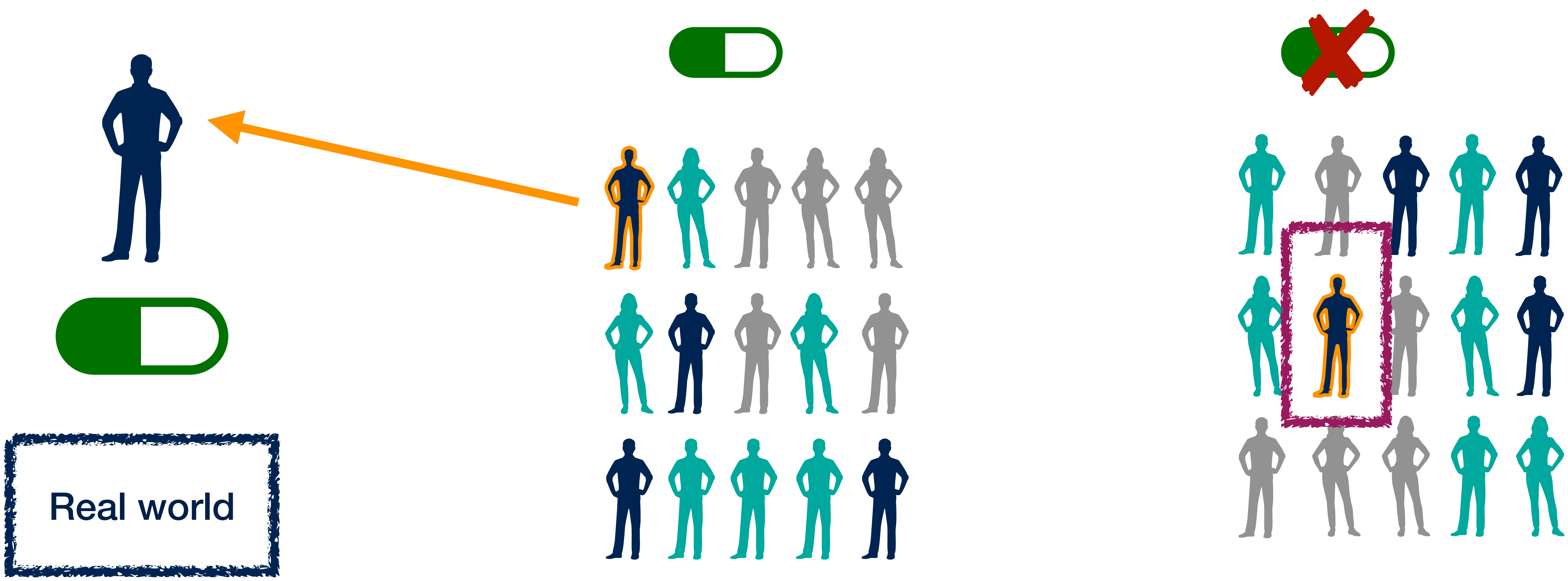
$$\begin{aligned}\boxed{ATE = \mathbb{E}_X\Big[\mathbb{E}[Y|T=1,X] - \mathbb{E}[Y|T=0,X]\Big]} \\ &= \Big(\cancel{\alpha_0 + \beta_x \mathbb{E}[X]} + \beta_t\Big) - \Big(\cancel{\alpha_0 + \beta_x \mathbb{E}[X]}\Big) \\ &= \beta_t\end{aligned}$$



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:

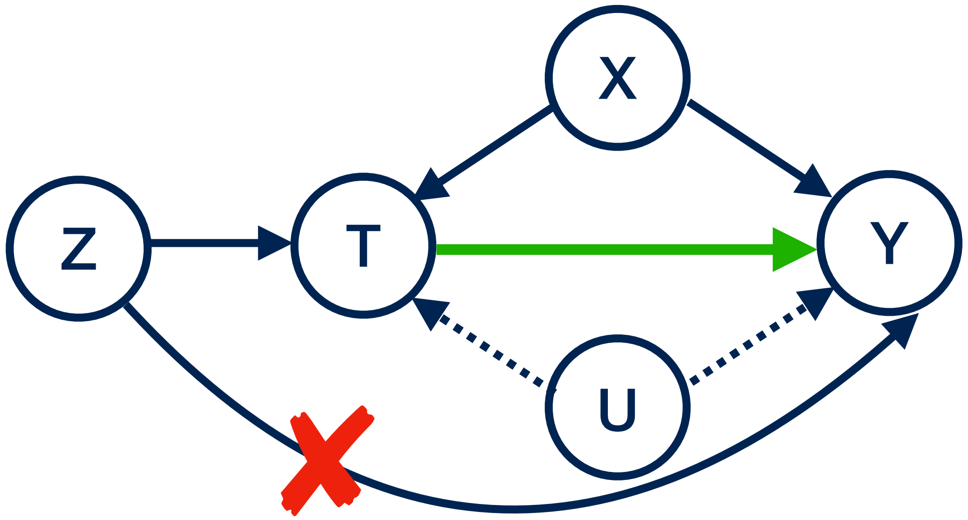
- SUTVA

- Exclusion restriction

- Non-zero average

- Monotonicity

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



Unobserved Confounders (Part 1): IV

Instrumental variable approach

Assumptions and why they are necessary:

- SUTVA

- Exclusion restriction

- Non-zero average

- Monotonicity

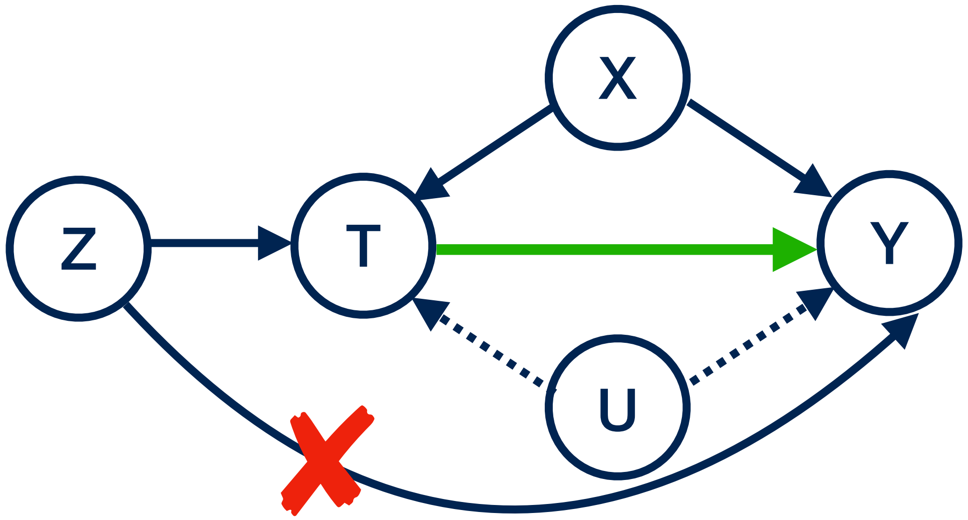
$$\boxed{\tau = \frac{\mathbb{E}\left[(Y|z=1) - (Y|z=0)\right]}{\mathbb{E}\left[(T|z=1) - (T|z=0)\right]}}$$

Estimation:

Binary case

Continuous case

- Ratio of Covs
- 2-Step regression



Other Causal estimators

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

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

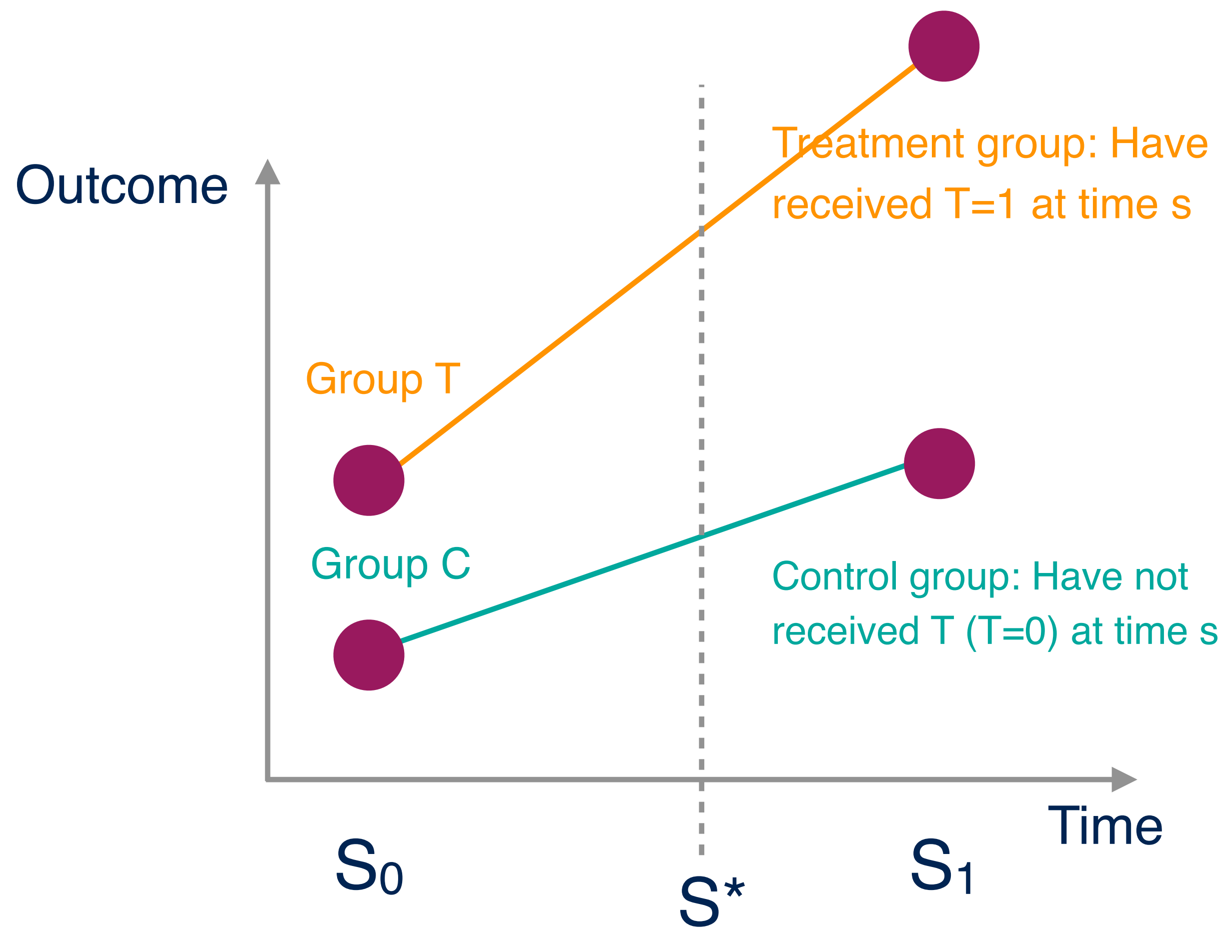
Other Causal estimators

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

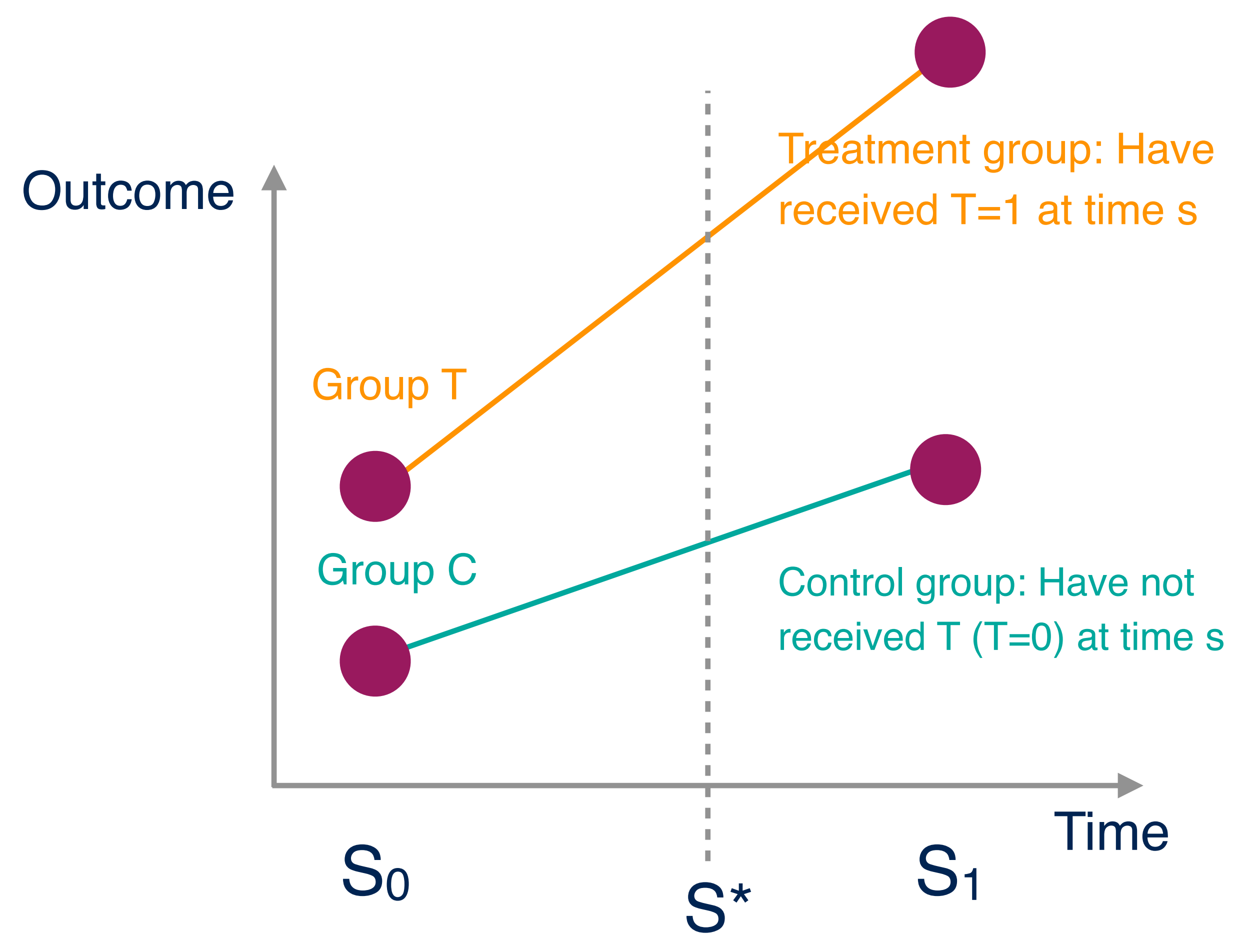
ATE: $\mathbb{E}[Y_1 - Y_0]$

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

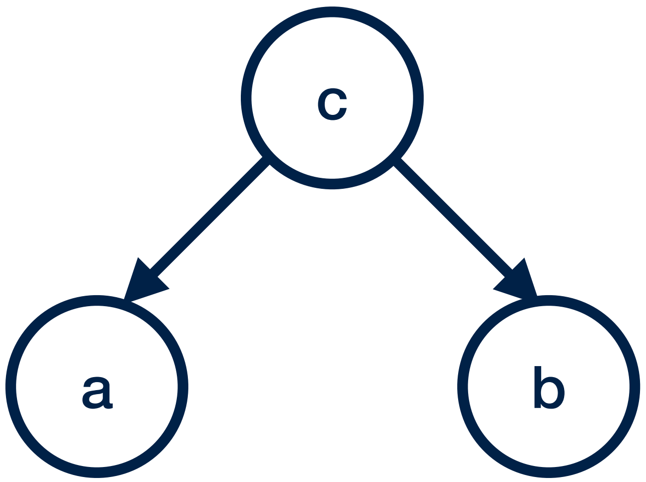
Regression discontinuity



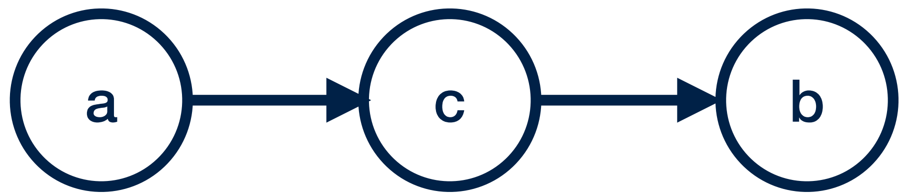
= measured (line is for visualisation only!)

The notion of **d-separation**

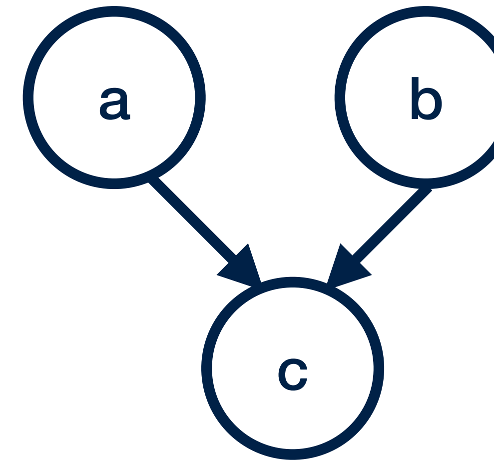
- Conditional independence via graphs and **D-separation**
- 3 main graph structures:



Fork



Chain



Collider

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)

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.,

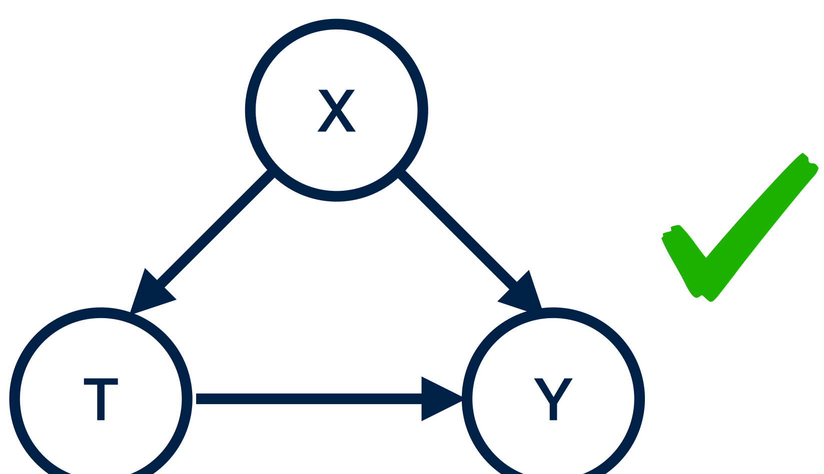
Under what conditions is the structure of a causal graph sufficient of computing a causal effect from a given data set? **Identifiability**

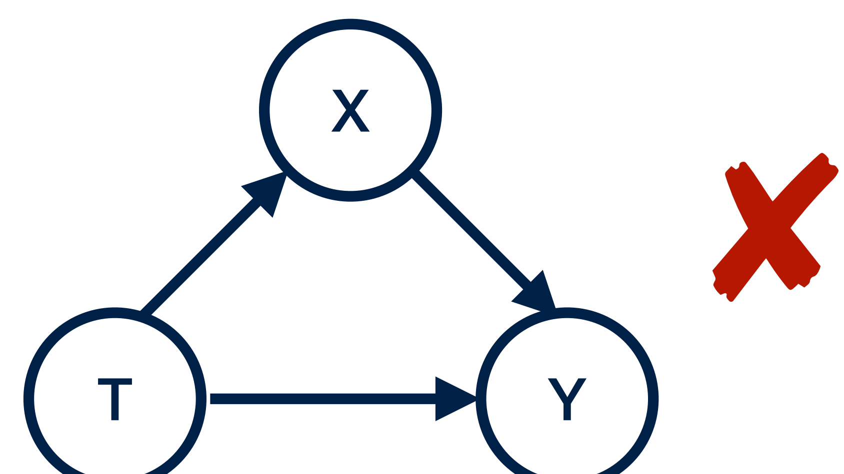
Backdoor Criterion: Given an ordered pair of variables (T,Y) in a DAG G, a set of variables X satisfies the backdoor criterion relative to (T,Y) if:

- (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)

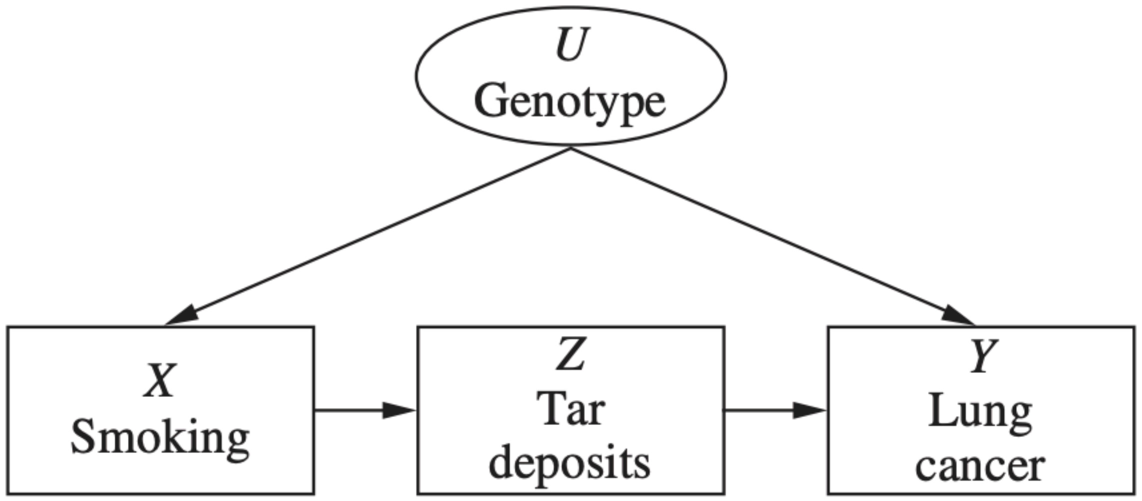
Unobserved Confounders (Part 2): Front-door

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

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

Both within Pearl’s framework

Different types of mediation and counterfactual questions.

Causal discovery

PC algorithm:

- Markov condition
- Causal sufficiency
- Faithfulness

Notion of Markov Equivalence Classes (and Markov Blanket)

Other types of approaches to causal inference:
- Functional causal models and utilising asymmetry in data
- LiNGAMs and additive noise models

Methods for Causal Inference
Lecture 19: Revision

Ava Khamseh
School of Informatics

2021-2022