Methods for Causal Inference
Lecture 4

Ava Khamseh
School of Informatics

2021-2022

Notations and conventions

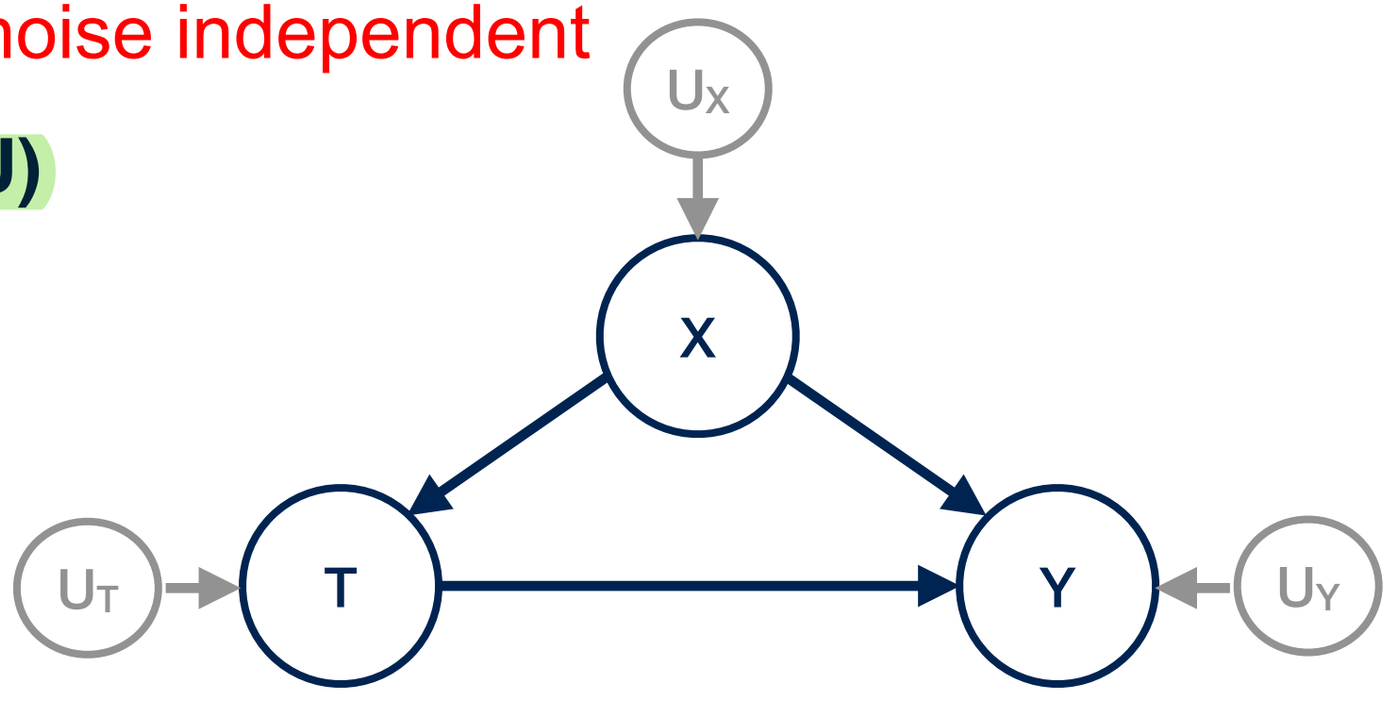
- Variable to be manipulated: **treatment (T)**, e.g. drug

- Variable we observe as response: **outcome (Y)**, e.g. success/ failure of drug

- Other observable variables that can affect treatment and outcome causally and we wish to correct for: **confounders (X)**, e.g. age, gender, …

- Unobservable confounder **(U)**

noise independent



For simplicity drop U_i's from graphs if:

$$U_T \perp\!\!\!\perp U_X \perp\!\!\!\perp U_Y$$

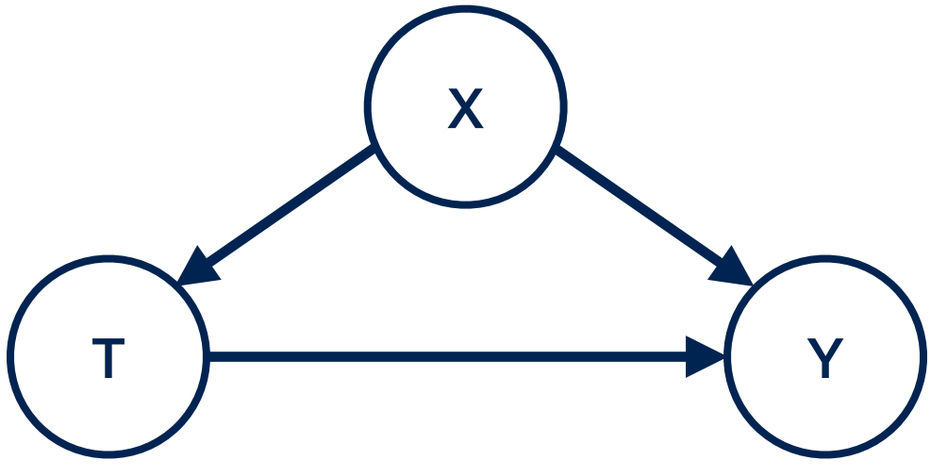
Two main Frameworks for causal estimation/discovery

- **Potential outcomes (Rubin):**

  - Requires a given treatment-outcome pair (known directionality)
  - Mainly applies to causal estimation (learning effects)
  - More familiar to biologists

- **Structural causal models (Pearl):**

  - Causal graph
  - Structural equations
  - Algorithmic: Causal Discovery



$$x = f_x(\epsilon_x),\ t = f_t(x, \epsilon_t),\ y = f_y(x, t, \epsilon_y)$$

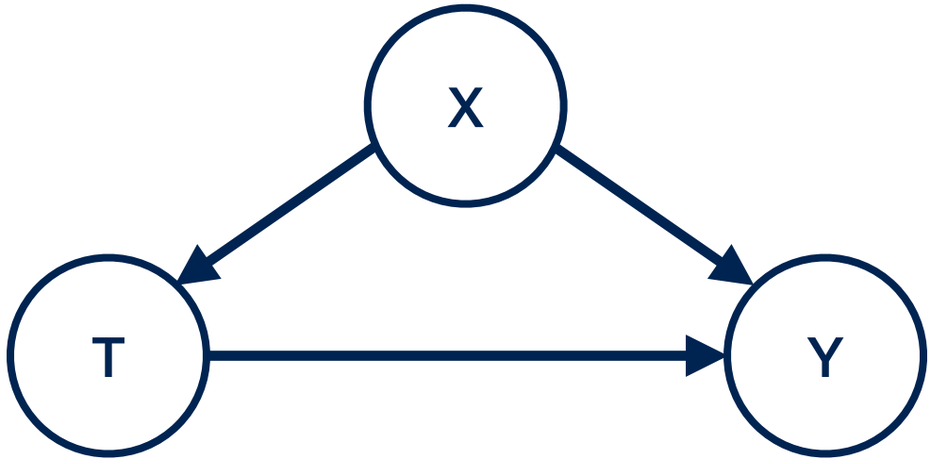
Extend the language of probability theory: do-calculus

Assumption: Independent noise terms: $\epsilon_x \perp\!\!\!\perp \epsilon_t \perp\!\!\!\perp \epsilon_y$

Two main Frameworks for causal estimation/discovery

- **Potential outcomes (Rubin):**
  - Requires a given treatment-outcome pair (known directionality)
  - Mainly applies to causal estimation (learning effects)
  - More familiar to biologists

- **Structural causal models (Pearl):**
  - Causal graph
  - Structural equations
  - Algorithmic: Causal Discovery



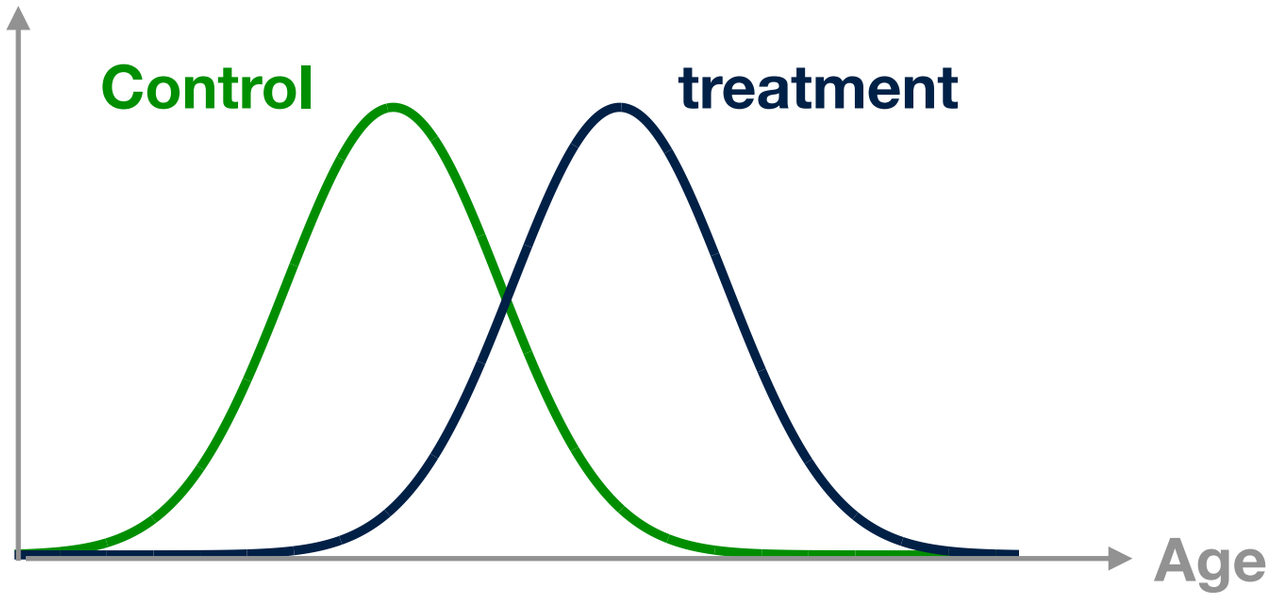
$$x = f_x(\epsilon_x),\ t = f_t(x, \epsilon_t),\ y = f_y(x, t, \epsilon_y)$$

Extend the language of probability theory:
do-calculus

Assumption: Independent noise terms: $\epsilon_x \perp\!\!\!\perp \epsilon_t \perp\!\!\!\perp \epsilon_y$

Recall: **Observational data,** what goes wrong?

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



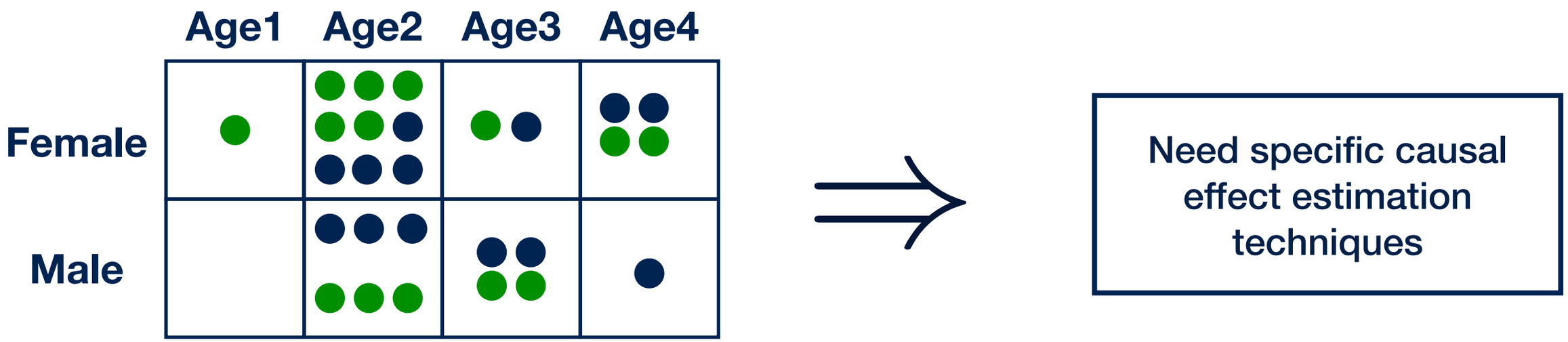
$$\left(\int y_1(x) p(x|t=1) dx \; - \int y_0(x) p(x|t=0) dx \right) \neq \int \big(y_1(x) - y_0(x) \big) p(x) dx$$

Observational data: Stratification

- Measure outcome (success/failure), **within** each of the young/old groups **separately**
- Take weighted average by the probability of being young/old

$$\mathbb{E}(\text{Healed}|t=1) = \mathbb{E}(\text{Healed}|t=1,\text{young})p(\text{young}) + \mathbb{E}(\text{Healed}|t=1,\text{old})p(\text{old})$$

- Disadvantages:
  - All possible confounders need to be observed
  - Assumes overlap between the two distributions (if there is no overlap, sample is not representative, e.g. performing the experiment only for old people)
  - Bad estimates as confounder dimensionality increases



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 (Rubin-Neyman)

Inverting previous relations, equivalently:

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

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

Knowing the potential outcomes is equivalent to knowing the observed and counterfactual outcomes

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**):

$$y_{obs}^{(i)} = t^{(i)} y_1^{(i)} + (1 - t^{(i)}) y_0^{(i)}$$

Individual treatment effect (causal): $\tau^{(i)} = y_1^{(i)} - y_0^{(i)}$

Average treatment effect (causal): $\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N} \sum_{i=0}^{N} \left(y_1^{(i)} - y_0^{(i)} \right)$

Example (Missing data interpretation)

	treatment	outcome	Y_0	Y_1	Y_1-Y_0
0	0.0	-10.039205	-10.039205	?	?
1	0.0	-10.671335	-10.671335	?	?
2	1.0	-9.216676	?	-9.216676	?
3	0.0	-6.952074	-6.952074	?	?
4	1.0	-9.842891	?	-9.842891	?
...	...	...	...	...	...
995	0.0	-6.344171	-6.344171	?	?
996	1.0	-9.563686	?	-9.563686	?
997	1.0	-8.414478	?	-8.414478	?
998	0.0	-9.731127	-9.731127	?	?
999	1.0	-8.097447	?	-8.097447	?

conterfounder missing here

Example (Missing data interpretation)

	treatment	outcome	Y_0	Y_1	Y_1-Y_0
0	0.0	-10.039205	-10.039205	?	?
1	0.0	-10.671335	-10.671335	?	?
2	1.0	-9.216676	?	-9.216676	?
3	0.0	-6.952074	-6.952074	?	?
4	1.0	-9.842891	?	-9.842891	?
...	...	...	...	...	...
995	0.0	-6.344171	-6.344171	?	?
996	1.0	-9.563686	?	-9.563686	?
997	1.0	-8.414478	?	-8.414478	?
998	0.0	-9.731127	-9.731127	?	?
999	1.0	-8.097447	?	-8.097447	?

What about the naive observational estimator?

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

-9.70

Example (Missing data interpretation)

	treatment	outcome	Y_0	Y_1	Y_1-Y_0
0	0.0	-10.039205	-10.039205	?	?
1	0.0	-10.671335	-10.671335	?	?
2	1.0	-9.216676	?	-9.216676	?
3	0.0	-6.952074	-6.952074	?	?
4	1.0	-9.842891	?	-9.842891	?
...	...	...	...	...	...
995	0.0	-6.344171	-6.344171	?	?
996	1.0	-9.563686	?	-9.563686	?
997	1.0	-8.414478	?	-8.414478	?
998	0.0	-9.731127	-9.731127	?	?
999	1.0	-8.097447	?	-8.097447	?

What about the naive observational estimator?

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

-9.70 -8.55

= -1.14

wrong number

Example (counterfactuals)

	treatment	outcome	treatment_CF	outcome_CF
0	0.0	-10.039205	1.0	-8.807301
1	0.0	-10.671335	1.0	-8.687408
2	1.0	-9.216676	0.0	-10.466275
3	0.0	-6.952074	1.0	-6.769770
4	1.0	-9.842891	0.0	-10.214971
...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128
996	1.0	-9.563686	0.0	-10.027234
997	1.0	-8.414478	0.0	-9.372274
998	0.0	-9.731127	1.0	-8.558852
999	1.0	-8.097447	0.0	-8.706807

Individual treatment effect:

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

Example (counterfactuals)

	treatment	outcome	treatment_CF	outcome_CF
0	0.0	-10.039205	1.0	-8.807301
1	0.0	-10.671335	1.0	-8.687408
2	1.0	-9.216676	0.0	-10.466275
3	0.0	-6.952074	1.0	-6.769770
4	1.0	-9.842891	0.0	-10.214971
...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128
996	1.0	-9.563686	0.0	-10.027234
997	1.0	-8.414478	0.0	-9.372274
998	0.0	-9.731127	1.0	-8.558852
999	1.0	-8.097447	0.0	-8.706807

Individual treatment effect:

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

Estimated as:

$$\frac{1}{N}\sum_{i=0}^{N}\left(y_1^{(i)} - y_0^{(i)}\right)$$

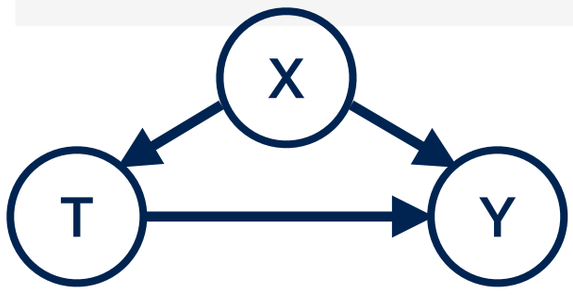
Example (individual treatment effect)

	treatment	outcome	treatment_CF	outcome_CF	Y_1-Y_0
0	0.0	-10.039205	1.0	-8.807301	1.231904
1	0.0	-10.671335	1.0	-8.687408	1.983927
2	1.0	-9.216676	0.0	-10.466275	1.249599
3	0.0	-6.952074	1.0	-6.769770	0.182305
4	1.0	-9.842891	0.0	-10.214971	0.372080
...	...	...	...	...	...
995	0.0	-6.344171	1.0	-6.584128	-0.239957
996	1.0	-9.563686	0.0	-10.027234	0.463548
997	1.0	-8.414478	0.0	-9.372274	0.957795
998	0.0	-9.731127	1.0	-8.558852	1.172276
999	1.0	-8.097447	0.0	-8.706807	0.609360

1.00 $\neq$ -1.14

Example (individual treatment effect)

	treatment	confounder	outcome	treatment_CF	outcome_CF	Y_1-Y_0
0	0.0	3.935767	-10.039205	1.0	-8.807301	1.231904
1	0.0	3.895803	-10.671335	1.0	-8.687408	1.983927
2	1.0	4.155425	-9.216676	0.0	-10.466275	1.249599
3	0.0	3.256590	-6.952074	1.0	-6.769770	0.182305
4	1.0	4.071657	-9.842891	0.0	-10.214971	0.372080
...	...	...	...	...	...	...
995	0.0	3.194709	-6.344171	1.0	-6.584128	-0.239957
996	1.0	4.009078	-9.563686	0.0	-10.027234	0.463548
997	1.0	3.790758	-8.414478	0.0	-9.372274	0.957795
998	0.0	3.852951	-9.731127	1.0	-8.558852	1.172276
999	1.0	3.568936	-8.097447	0.0	-8.706807	0.609360



Potential Outcomes: Assumptions

- **SUTVA:** Stable Unit Treatment Value Assumption
  - **Consistency:** Well-defined treatment (no different versions) observed outcome is independent of how the treatment is assigned
  - **No interference**: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

1. no matter how the treatment is given, the effect of treatment is same
2. people in trial do not affect each other

Potential Outcomes: Assumptions

- **SUTVA:** Stable Unit Treatment Value Assumption
  - **Consistency:** Well-defined treatment (no different versions) observed outcome is independent of how the treatment is assigned
  - **No interference**: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

- **Positivity:** Every individual has a non-zero chance of receiving the treatment/control: $p(t = 1|x) \in (0, 1) \text{ if } P(x) > 0$

Potential Outcomes: Assumptions

- **SUTVA:** Stable Unit Treatment Value Assumption
  - **Consistency:** Well-defined treatment (no different versions) observed outcome is independent of how the treatment is assigned
  - **No interference**: Different individuals (units) within a population do not influence each other (e.g. does not work in social behavioural studies, care must be taken for time series data when defining the units)

- **Positivity:** Every individual has a non-zero chance of receiving the treatment/control: $p(t = 1|x) \in (0, 1) \text{ if } P(x) > 0$

  e.g. young don't take vaccine

- **Unconfoundedness (ignorability/exchangeability):** Treatment assignment is random, given confounding features X

Unconfoundedness

- **Unconfoundedness**: Treatment assignment is random, given X:

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

x: confounder

Unconfoundedness

- **Unconfoundedness**: Treatment assignment is random, given X:

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

- given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)

Unconfoundedness

- **Unconfoundedness**: Treatment assignment is random, given X:

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

- given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person B

Unconfoundedness

- **Unconfoundedness**: Treatment assignment is random, given X:

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

- given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person B
- There may be difference in sample size between case and control:

  $p(t = 1|x)$ not necessarily = $p(t = 0|x)$

Unconfoundedness

- **Unconfoundedness**: Treatment assignment is random, given X:

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

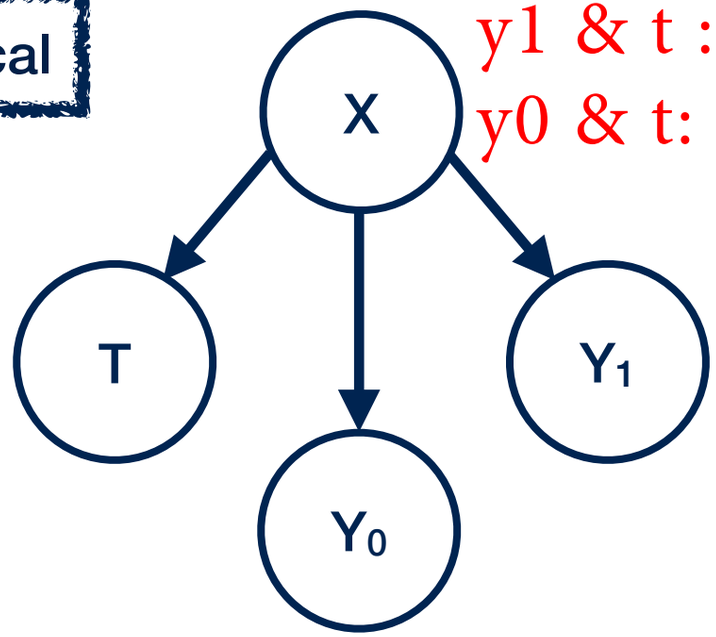
- given X, there is no preference for individual (i) to get assigned the treatment as compared to individual (j) (i.e. randomised)
- e.g., restricting to the old group, person A has the same probability of receiving the treatment as person B
- There may be difference in sample size between case and control:
  $p(t = 1|x)$ not necessarily = $p(t = 0|x)$
- However, if we do not restrict to the old group, there is a clear preference: older individuals are more likely to receive the drug
- **No unobserved confounders**
  (see later: unverifiable in observational data)

Unconfoundedness: A graphical representation

- **Unconfoundedness**: Treatment assignment is random, given X:

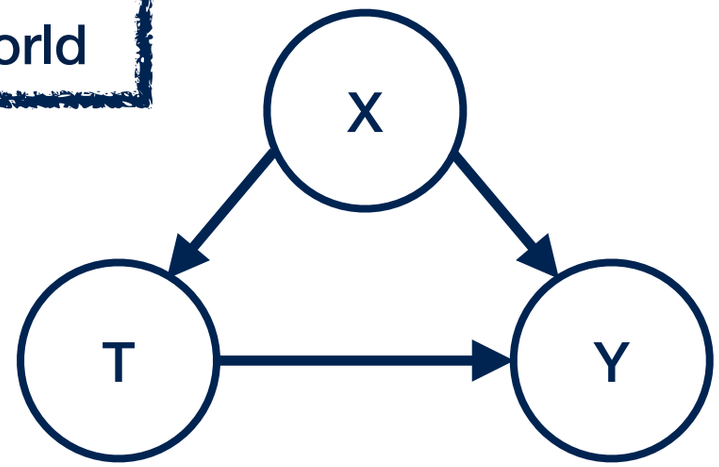
$$y_1^{(i)}, y_0^{(i)} \perp\!\!\!\perp t^{(i)} \mid x$$

Hypothetical



y1 & t : independent
y0 & t: independent

Real world



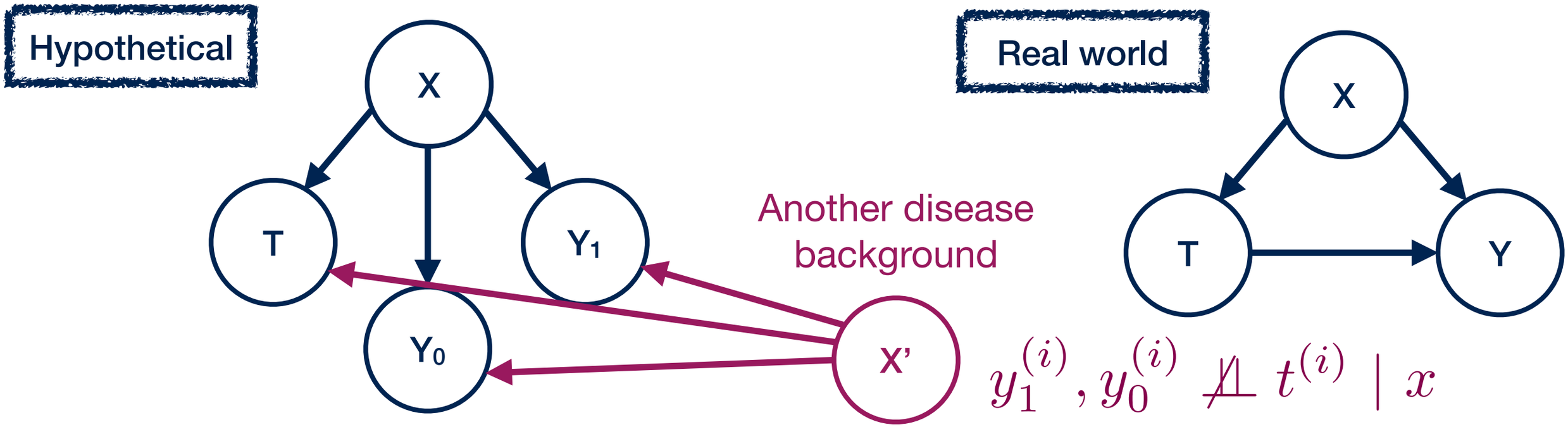
If everyone receive the treatment: Y_1
If everyone is prevented from receiving the treatment: Y_0

Then the hypothetical outcomes are entirely determined by the set of features X of the individuals.

Unconfoundedness: A graphical representation

- **Unconfoundedness**: Treatment assignment is random, given X:

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



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

If everyone receive the treatment: Y_1

If everyone is prevented from receiving the treatment: Y_0

Then the hypothetical outcomes are entirely determined by the set of features X of the individuals.

Positivity

For existing values of covariates in the population, i.e., $P(X = x) > 0$ (binary T)

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

Intuitively: If everyone was given the treatment, i.e., there is not control group, we have no idea if/how the outcomes observed are due to the treatment itself (because we have no background to compare it to!)

All same is bad

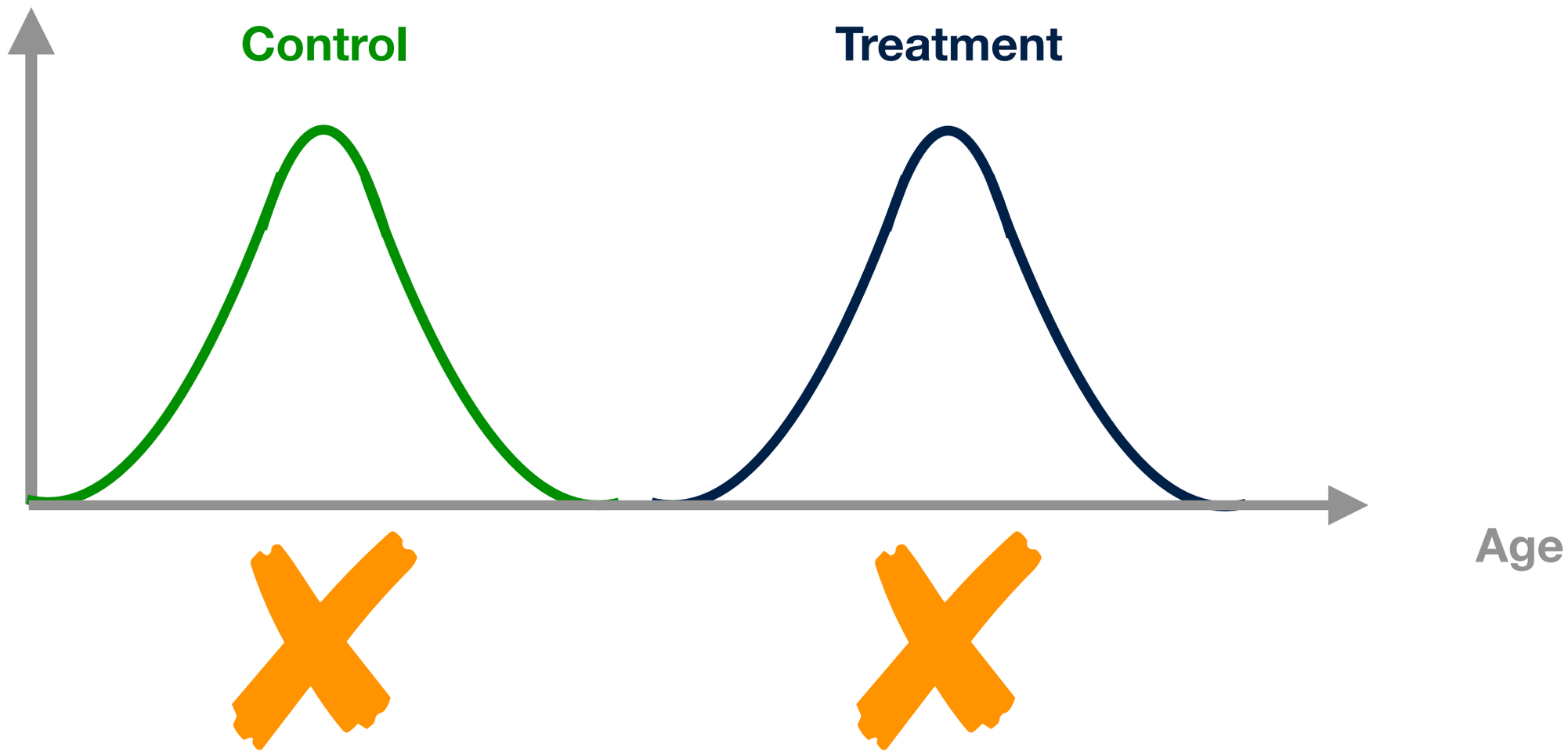
Similarly, when everyone is in the control group: Then we will not have tested the treatment.

Tutorial question: See why this condition is essential (**mathematically**)

Positivity (common support/overlap)

Control: T = 0

Treatment T=1



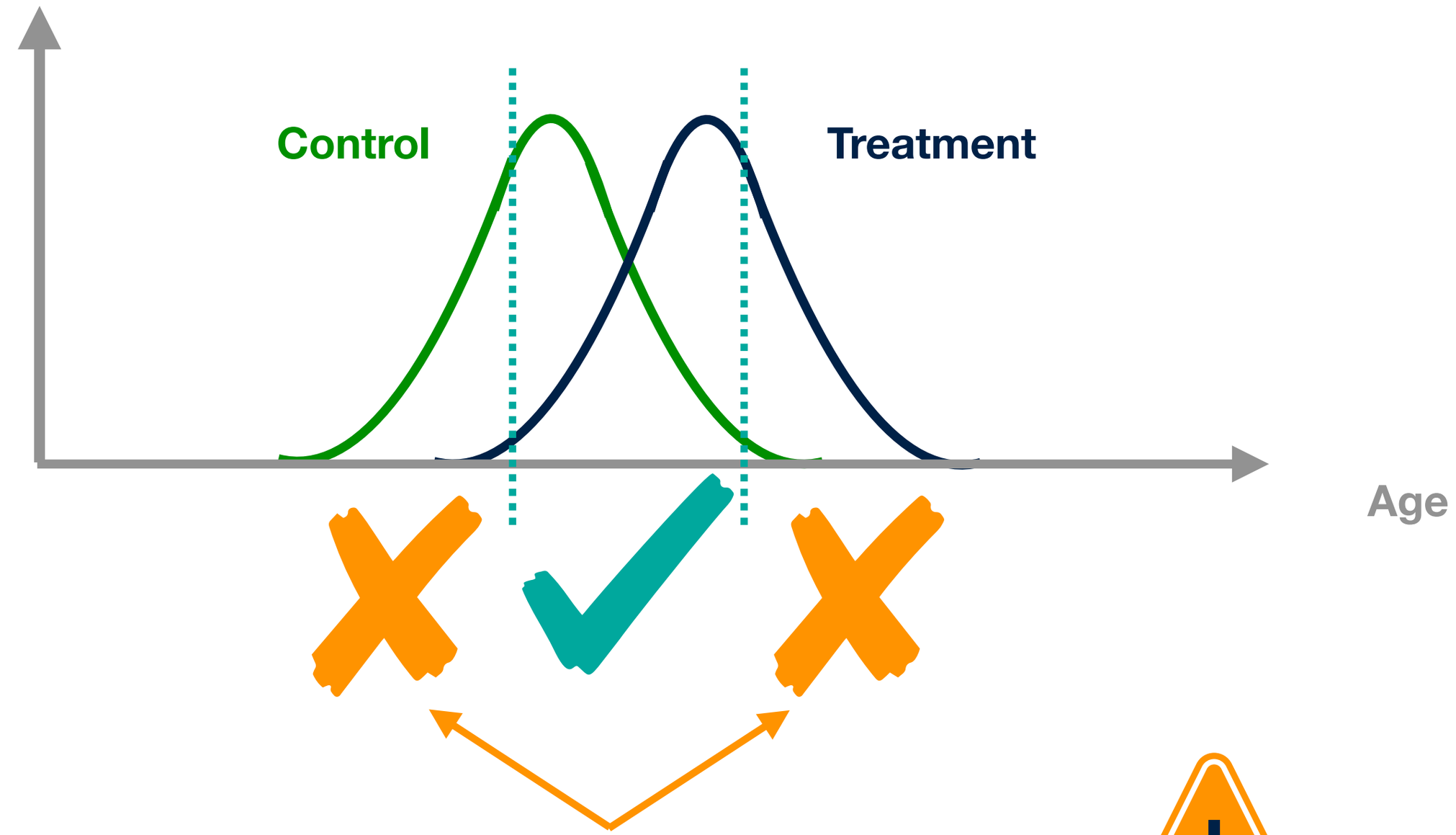
No overlap

Complete violation of positivity

Positivity (common support/overlap)

Control: T = 0

Treatment T=1

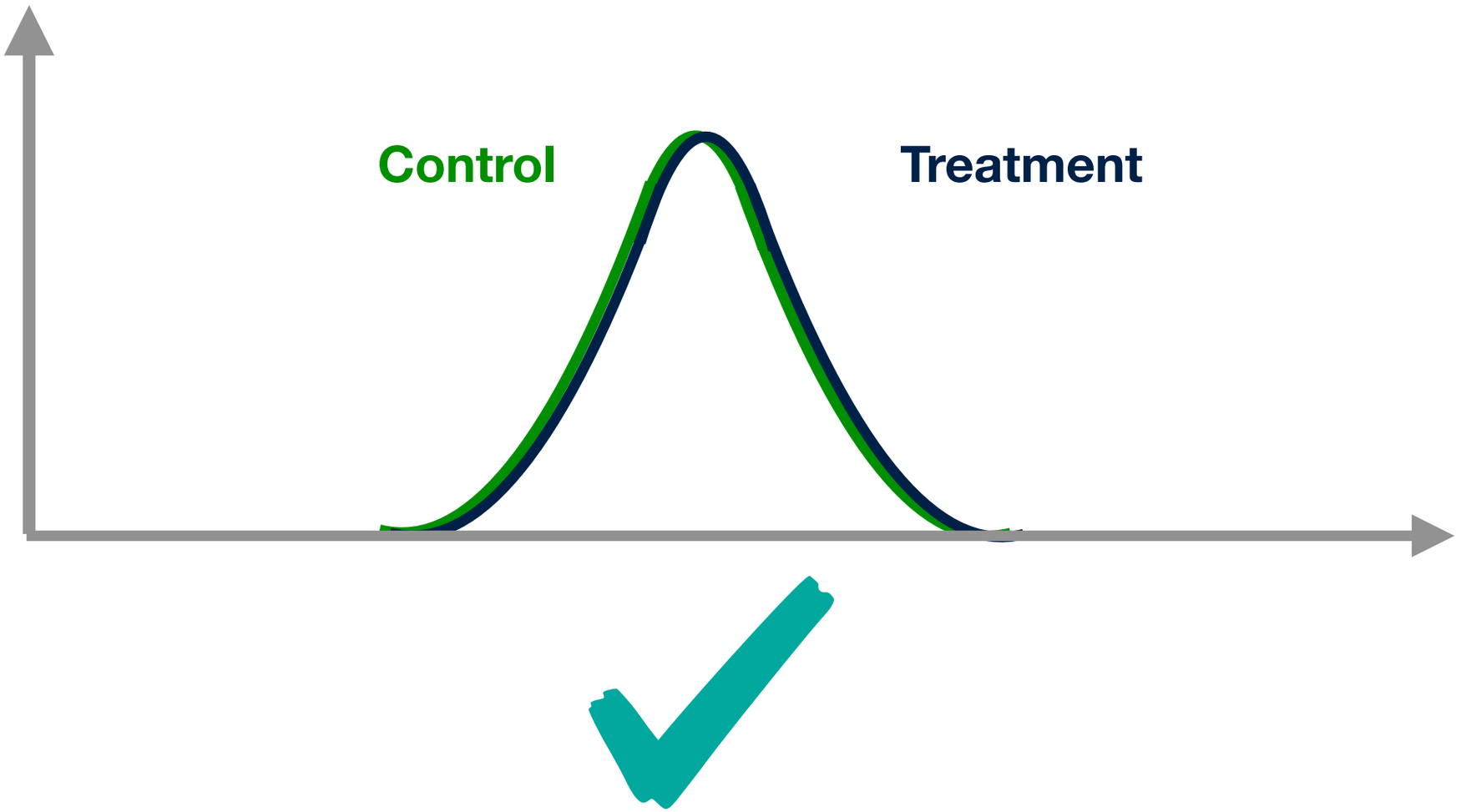


Some violation of positivity

Positivity (common support/overlap)

Control: T = 0

Treatment T=1



Complete overlap: No positivity violation

Positivity vs unconfoundedness

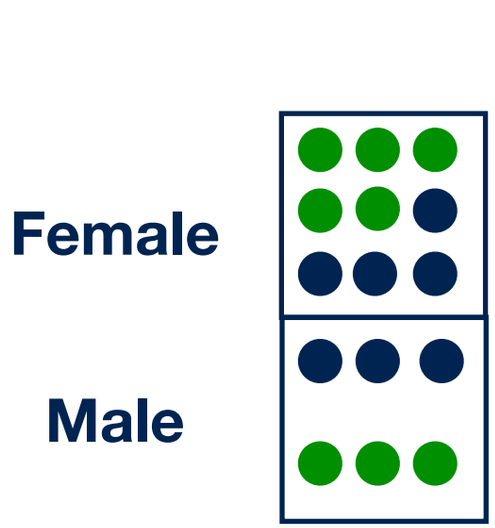
Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

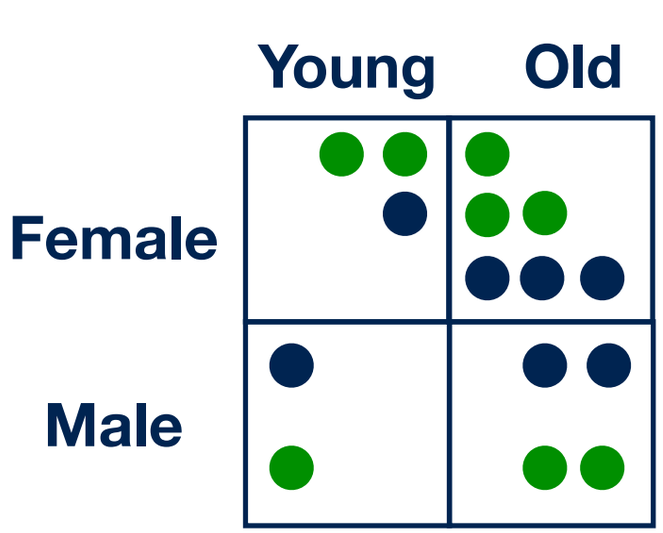
Positivity vs unconfoundedness

Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

But the more we condition on, the harder it is to satisfy positivity

Example:





Easy to check for binary/categorical variable X:

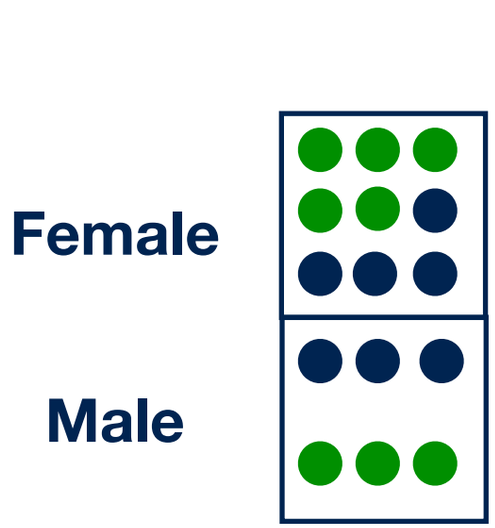
$$0 < P(T = 1|X = x) < 1$$

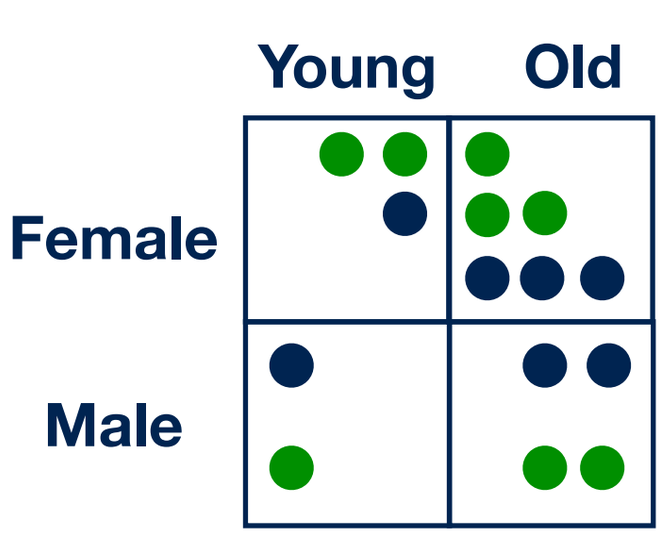
Positivity vs unconfoundedness

Issue: We potentially wish to condition on many variables to make it more likely for unconfoundedness to be satisfied …

But the more we condition on, the harder it is to satisfy positivity

Example:

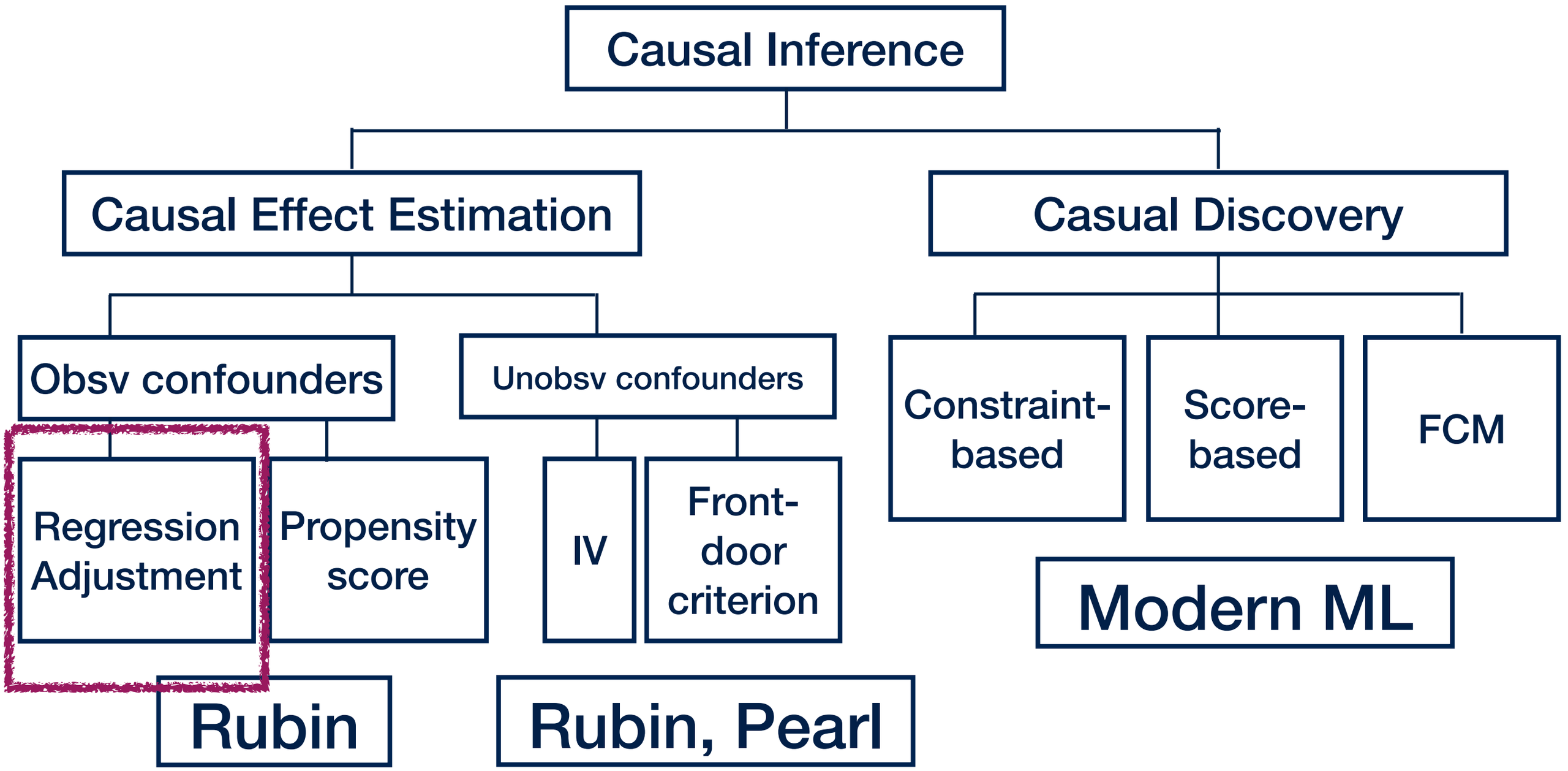




Tutorial question: Discuss the problem of no support, extrapolation and model-misspecification

Overview of the course

- Estimating causal effects
- Randomised trial vs observational data

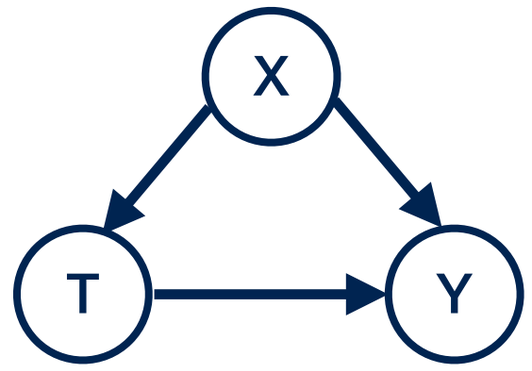


Regression Adjustment

- X is a sufficient set of confounders if conditioning on X, there would be no confounding bias

- For individual (i) there is only one **observed** outcome: $y^{(i)}_{\underline{t}_i}$
- Would like to estimate (infer) **counterfactual**: $\hat{y}^{(i)}_{1-t_i} = \hat{\mathbb{E}}\left[y^{(i)}|1-t_i, x^{(i)}\right]$

- Using a design matrix, fit: $Y = \beta_X X + \beta_T T + \epsilon$



Ctrl **Drug**

$$T = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ .. & .. \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

Young **Old**

$$X = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ .. & .. \\ 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ .. \\ y^{(N-1)} \\ y^{(N)} \end{pmatrix} = \begin{pmatrix} \beta_{t=0} + \beta_{x=\text{young}} \\ \beta_{t=0} + \beta_{x=\text{old}} \\ .. \\ \beta_{t=1} + \beta_{x=\text{young}} \\ \beta_{t=1} + \beta_{x=\text{old}} \end{pmatrix}$$

- Assumptions: Overlap and additivity

additivity: no XT terms

$$\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N}\sum_{i=0}^{N}\left(y_1^{(i)} - y_0^{(i)}\right)$$

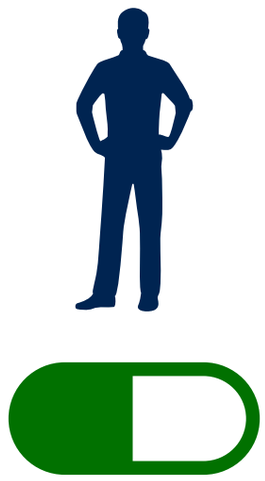
Matching

Idea: Create a ‘clone/twin’ for each individual (in terms of X)

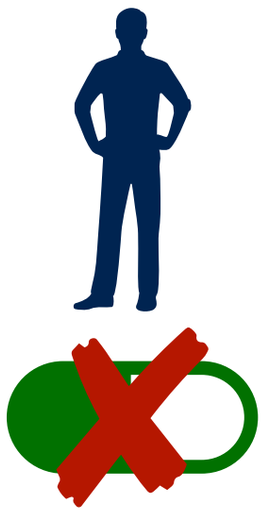
i.e. if individual 1 has t = 1, then their ‘clone/twin’ has t = 0.

Blind ourselves to the outcomes, try to get as similar to a randomised experiment as possible (‘correct for confounding’)

Example:



Real world



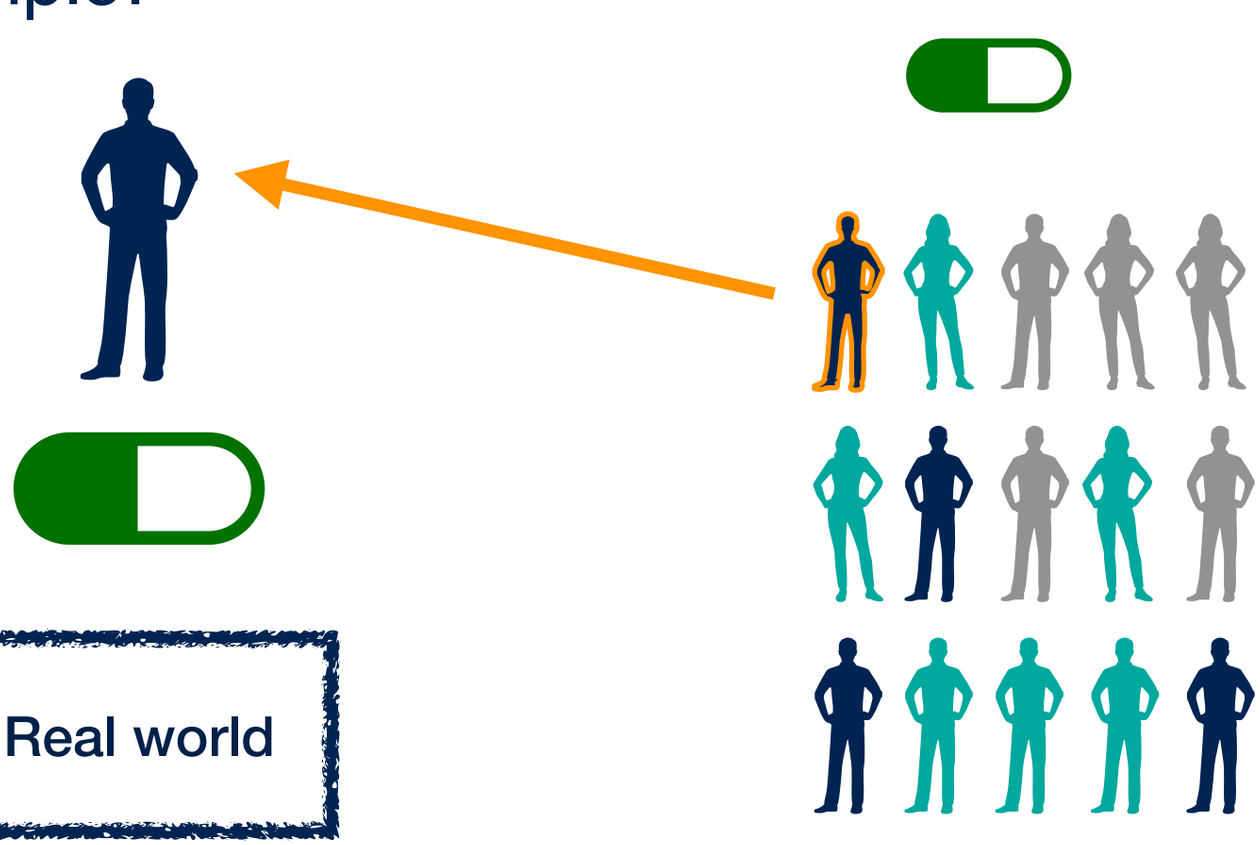
Hypothetical (counterfactual)

Matching

Idea: Create a ‘clone/twin’ for each individual (in terms of X)
i.e. if individual 1 has t = 1, then their ‘clone/twin’ has t = 0.

Blind ourselves to the outcomes, try to get as similar to a randomised experiment as possible (‘correct for confounding’)

Example:

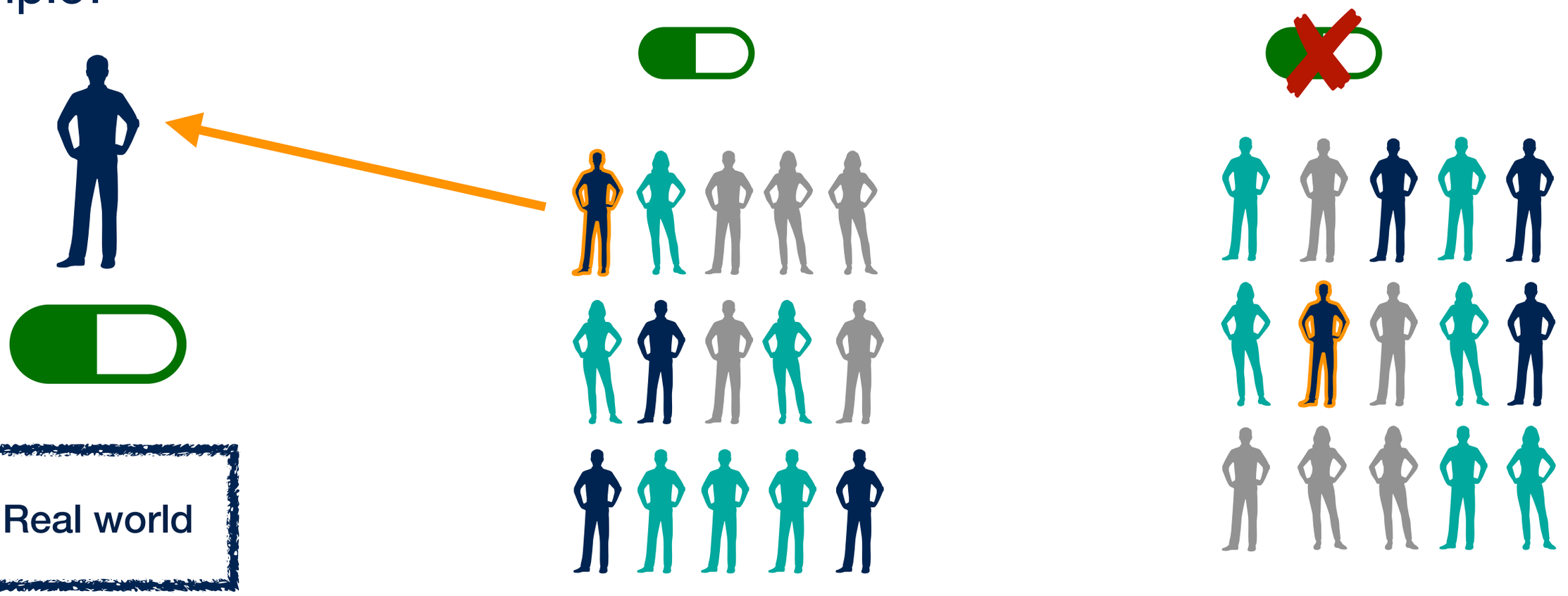


Matching

Idea: Create a ‘clone/twin’ for each individual (in terms of X)
i.e. if individual 1 has t = 1, then their ‘clone/twin’ has t = 0.

Blind ourselves to the outcomes, try to get as similar to a randomised experiment as possible (‘correct for confounding’)

Example:

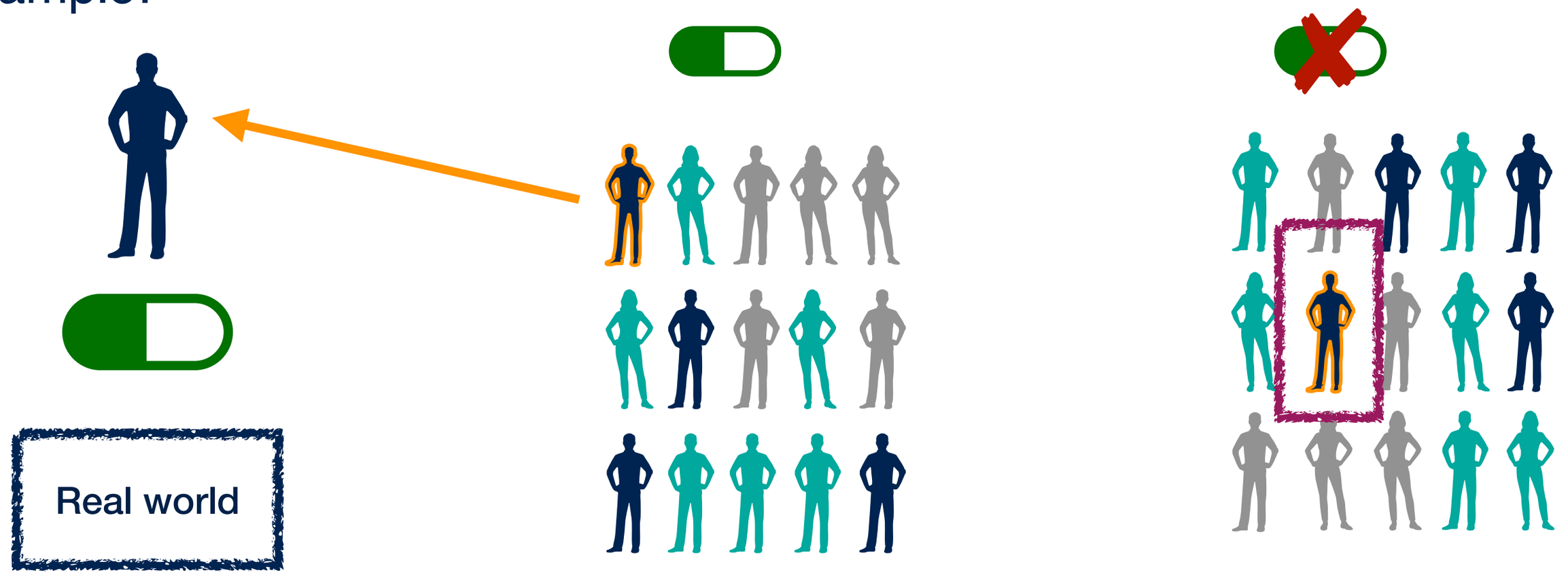


Matching

Idea: Create a ‘clone/twin’ for each individual (in terms of X)
i.e. if individual 1 has t = 1, then their ‘clone/twin’ has t = 0.

Blind ourselves to the outcomes, try to get as similar to a randomised experiment as possible (‘correct for confounding’)

Example:



Matching

- Reveals **lack of overlap** in treatment vs control distributions: individuals in the treatment group that have no chance of having an '**equivalent**' in control group, ie, parts of the distribution with: $p(t=1|x)=0,\ p(t=0|x)=0$

- **Mahalanobis distance**: Difference scaled by variance

$$D(x^{(i)}, x^{(j)}) = \sqrt{\left(x^{(i)} - x^{(j)}\right)^T S^{-1} \left(x^{(i)} - x^{(j)}\right)},\ S = \mathrm{Cov}(X)$$

- Issues: Outliers. Use a calliper: maximum acceptable distance, to avoid violating the positivity assumption. But the populations becomes harder to define.

Propensity Score

- In a perfect **randomised** trial: p(t=1|x)=p(t=1)=0.5

- In an **observational study**, p(t=1|x) can be **estimated**, since it involves **observational data** at a t and x (hence identifiable).

- A **balancing score** is any function b(x) such that:

$$x \perp\!\!\!\perp t | b(x)$$

i.e., distribution of confounders is independent of treatment given b(x):

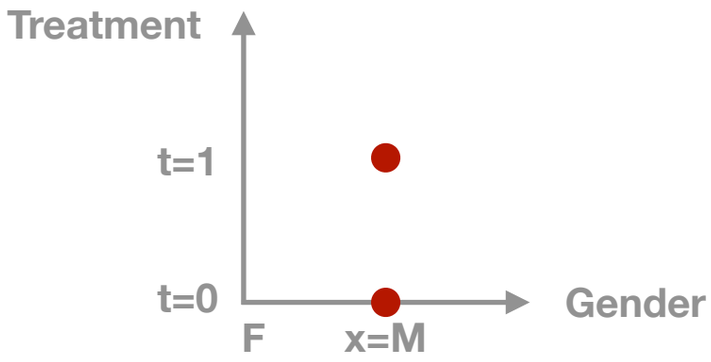
$$\boxed{p(X = x|b(x), t = 1) = p(X = x|b(x), t = 0)}$$

Rosenbaum, Rubin 1983

Propensity Score

- Candidate b(x) = x, trivially satisfies:

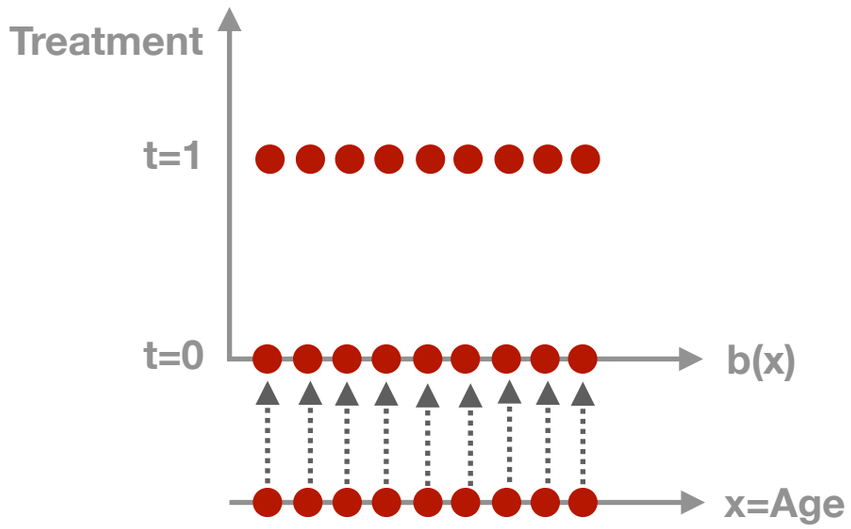
$$p(X = x|x, t = 1) = p(X = x|x, t = 0) = 1$$

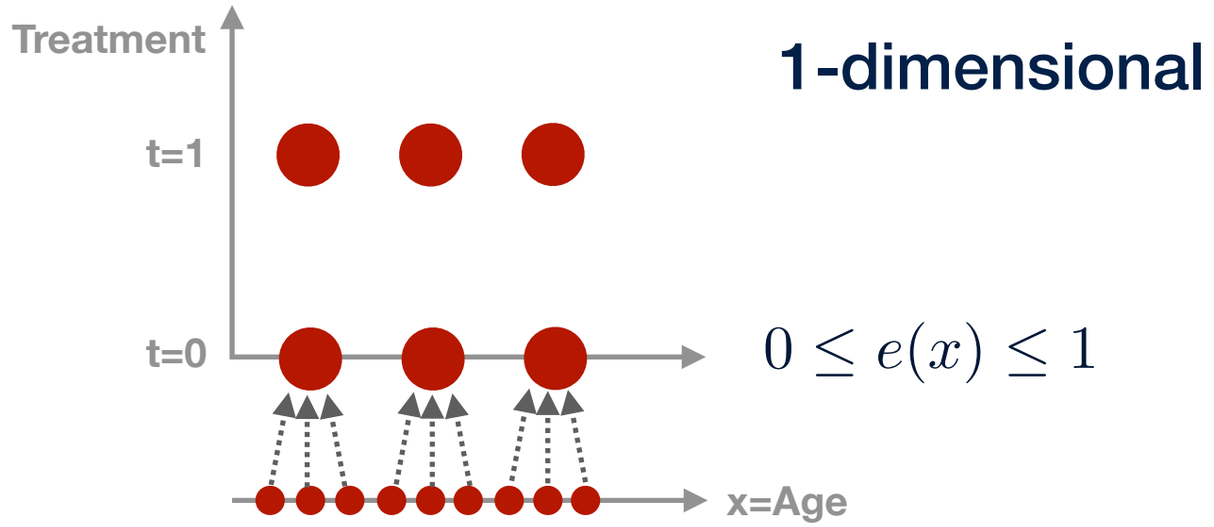


- b(x) = x is the **finest** such function: OK for e.g. binary confounders, but only gives point estimates for (almost) continuous confounders!

- **Propensity score** is the **coarsest** such function (i.e. more data points, leading to better estimates):

$$e(x) = p(t = 1|x)$$





Propensity Score Matching

- Let the distribution of covariates follow an exponential family of distributions ($P_{t^*}(x)$ polynomial of degree k):

$$p(x|t = t^*) = h(X) \exp(P_{t^*}(x)) \ , \ \text{for } t = 0 \text{ or } 1$$

- Estimate propensity score e(x)=p(t=1|x):

$$\log\left(\frac{e(x)}{1-e(x)}\right) = \log\left(\frac{p(t=1|x)}{p(t=0|x)}\right) = \log\left(\frac{p(x|t=1)p(t=1)}{p(x|t=0)p(t=0)}\right) = \log\left(\frac{p(t=1)}{p(t=0)}\right) + P_1(x) - P_0(x)$$

- If we consider k=1, linear exponential family (e.g. Bernoulli),

$$\log\left(\frac{e(x)}{1-e(x)}\right) = wx + w_0 \ \Rightarrow \ e(x) = \frac{1}{1+e^{-wx-w_0}}$$

- Fit parameters by maximising log-likelihood: $LL = \frac{1}{N}\sum_{i=0}^{N} \log p(t^{(i)}|x^{(i)})$

Rosenbaum, Rubin 1983

Propensity Score Matching Algorithms

- Match control and treatment individuals based on their propensity score
- Greedy matching:
  - Randomly order list of control and treated.
  - Start with the first individual from e.g. treated and match to control with the smallest distance (i.e. obtains the **local** minimum)
  - Remove individuals from control and matched treated
  - Move to the next treated subject

Treatment	Control
40	50
65	25

Propensity Score Matching Algorithms

- Match control and treatment individuals based on their propensity score
- Greedy matching:
  - Randomly order list of control and treated.
  - Start with the first individual from e.g. treated and match to control with the smallest distance (i.e. obtains the **local** minimum)
  - Remove individuals from control and matched treated
  - Move to the next treated subject

Treatment		Control
40	→	50
65	→	25

Propensity Score Matching Algorithms

- Match control and treatment individuals based on their propensity score
- Greedy matching:
  - Randomly order list of control and treated.
  - Start with the first individual from e.g. treated and match to control with the smallest distance (i.e. obtains the **local** minimum)
  - Remove individuals from control and matched treated
  - Move to the next treated subject

Total diff: 50

Total diff: 30

Propensity Score Matching Algorithms

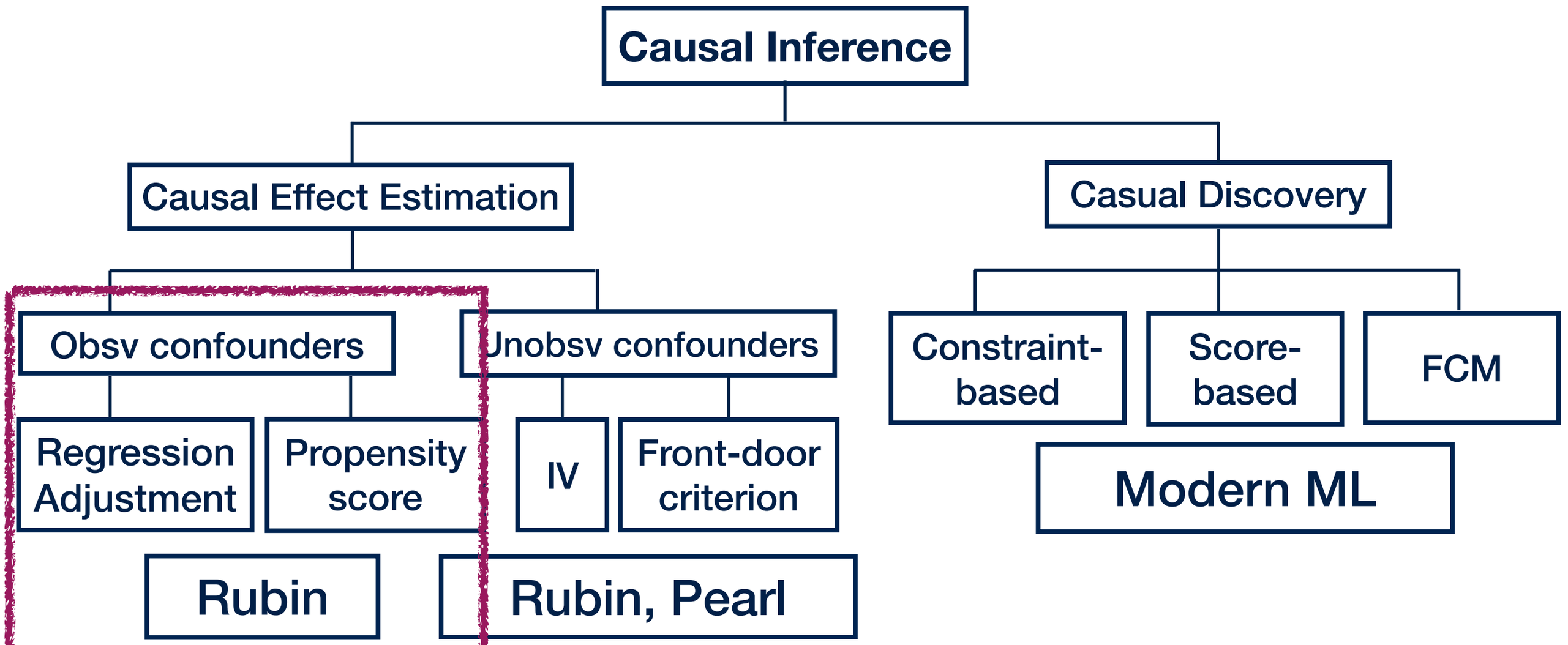
- Match control and treatment individuals based on their propensity score
- Greedy matching:
  - Randomly order list of control and treated.
  - Start with the first individual from e.g. treated and match to control with the smallest distance (i.e. obtains the **local** minimum)
  - Remove individuals from control and matched treated
  - Move to the next treated subject

- Optimal matching: Minimises the **global** distance, computationally demanding

- **ATE**: $\tau = \hat{\mathbb{E}}[\tau^{(i)}] = \hat{\mathbb{E}}[y_1^{(i)} - y_0^{(i)}] = \frac{1}{N}\sum_{i=0}^{N}\left(y_1^{(i)} - y_0^{(i)}\right)$

Overview of the course

- **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**:



Methods for Causal Inference
Lecture 4

Ava Khamseh
School of Informatics

2021-2022