CSDS 452 Causality and Machine Learning

Lecture 2: Potential Outcome Framework

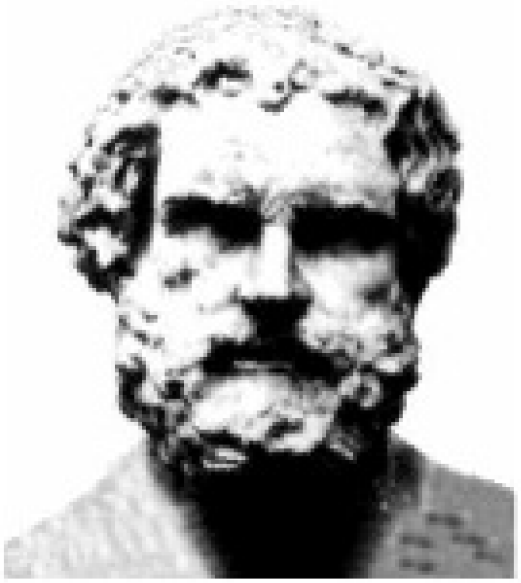
Instructor: Jing Ma

Fall 2024, CDS@CWRU

Outline

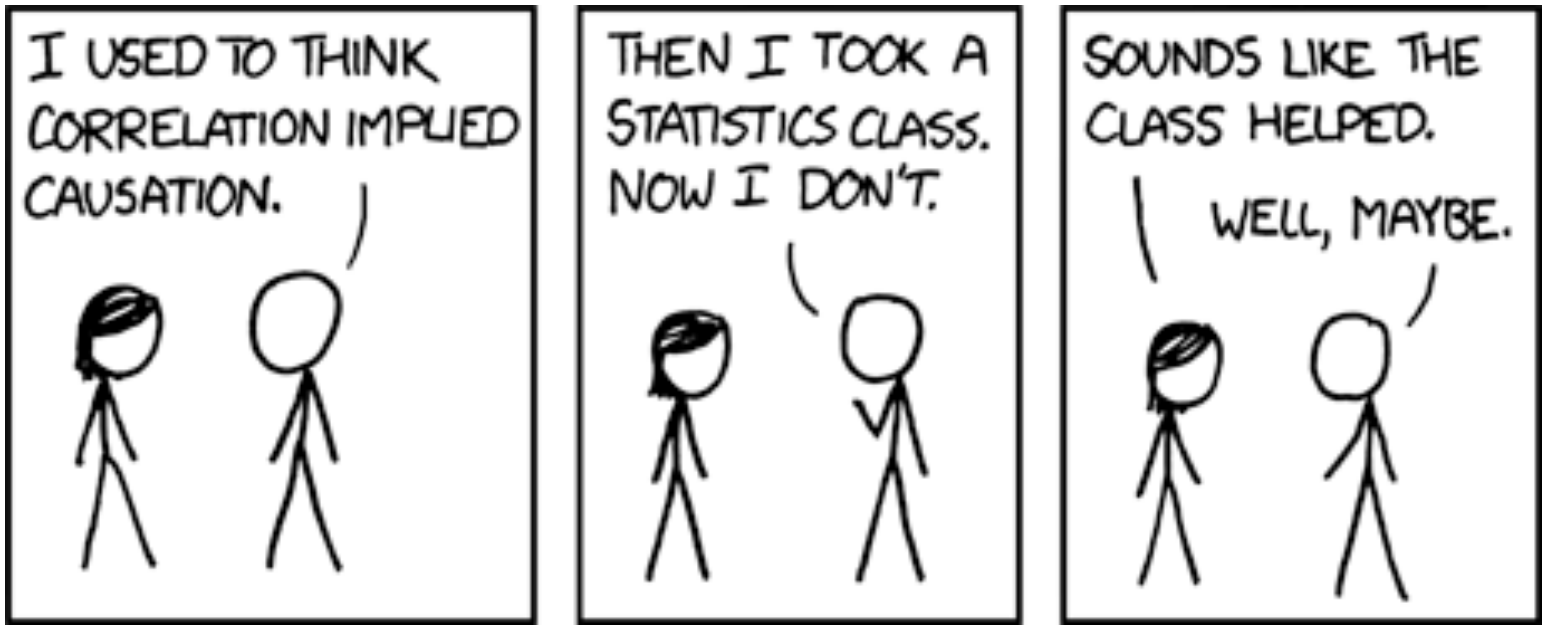
- Potential outcome
- Definition & concepts
- Assumptions in observational studies
- Example: Causal effect estimation

Causality Learning has a Long History



- Democritus (460-390 BC), philosopher
- “I would rather discover a single causal relationship than be king of Persia”

Recap: Causality is Important



http://imgs.xkcd.com/comics/correlation.png

Causality → dependence !

X and Y are associated iff $\exists x_1 \neq x_2$
$\mathrm{P}(\mathrm{Y}|X = x_1) \neq \mathrm{P}(\mathrm{Y}|X = x_2)$

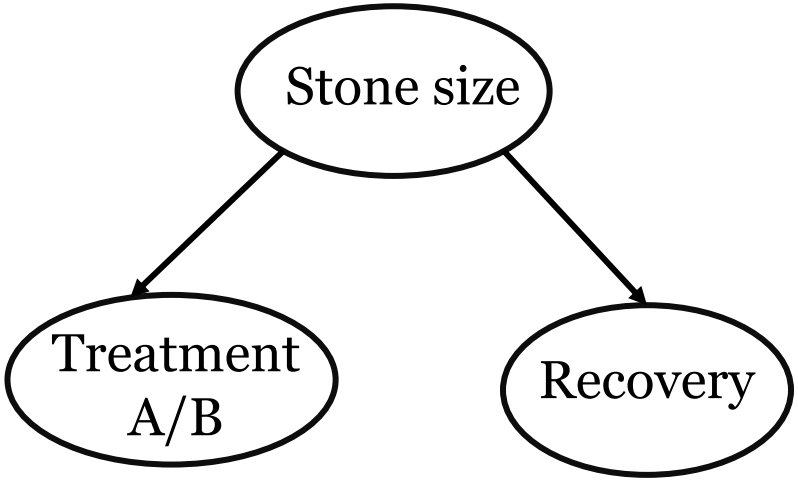
Dependence → causality ?

X is a cause of Y iff $\exists x_1 \neq x_2$
$\mathrm{P}(\mathrm{Y}|do(X = x_1)) \neq \mathrm{P}(\mathrm{Y}|do(X = x_2))$

Recap: Simpson's paradox

A trend appears in several groups of data but **disappears or reverses** when the groups are combined.

	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**

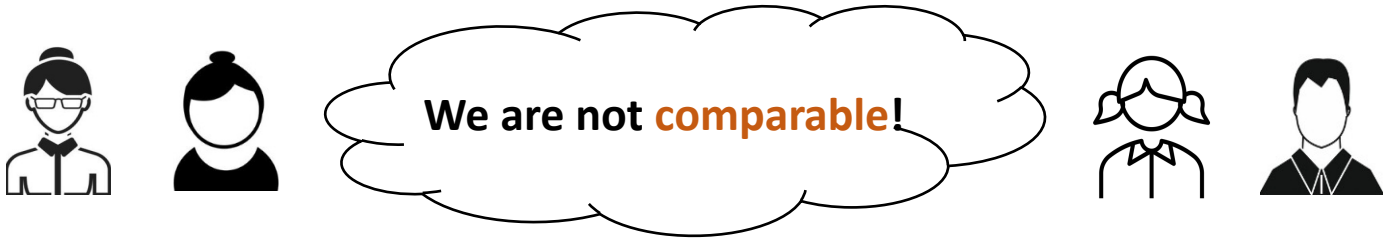


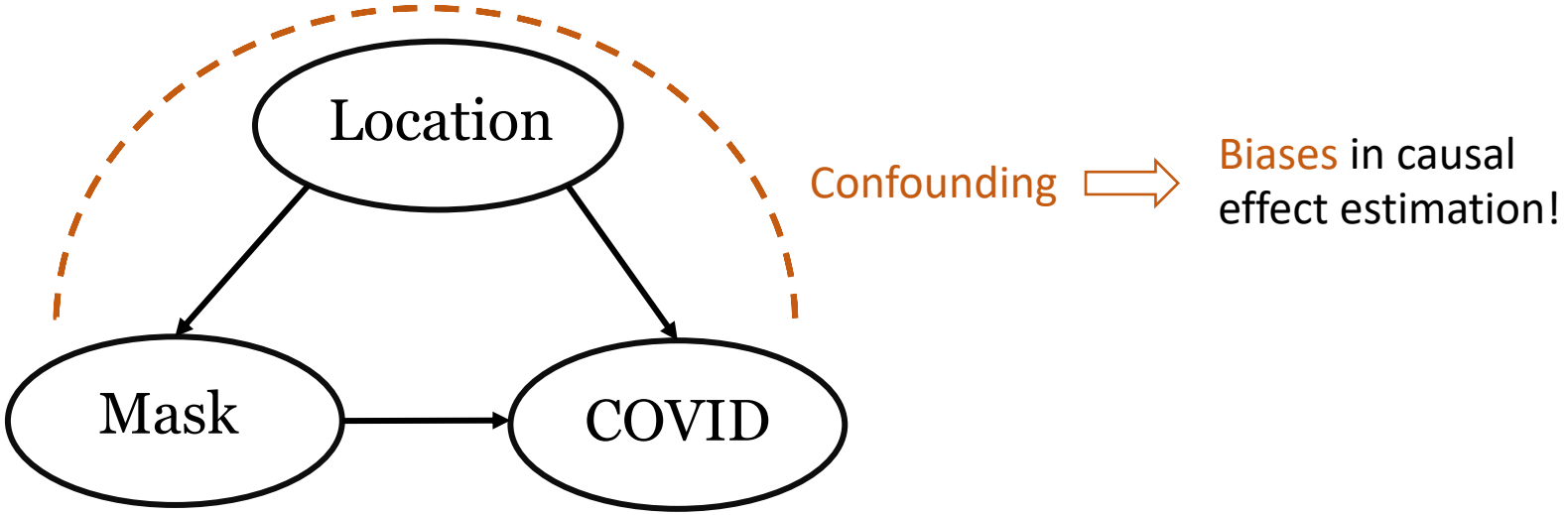
- Doctors tend to give patients with large stones treatment A, and the patients with small stones treatment B
- Larger stones has less recovery rate

"Confounding effect"

Dependency ≠ Causation

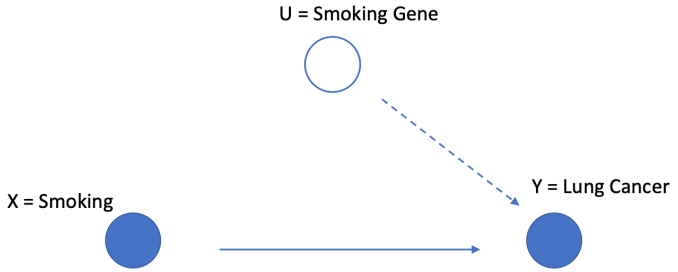
- $E[Y(1) - Y(0)] \neq E[Y|T = 1] - E[Y|T = 0]$





Frameworks in Causal Inference

- Structural Causal Model
  - Based on graphical models
  - Causal graph + structural equations



Judea Pearl

Reference books:

- Pearl J. Causality[M]. Cambridge university press, 2009.
- Pearl J, Mackenzie D. The book of why: the new science of cause and effect[M]. Basic books, 2018.

Frameworks in Causal Inference

- Potential Outcome Framework (Neyman–Rubin causal model)
  - An approach to the statistical analysis of cause and effect based on the framework of potential outcomes

Jerzy Neyman

Donald B. Rubin

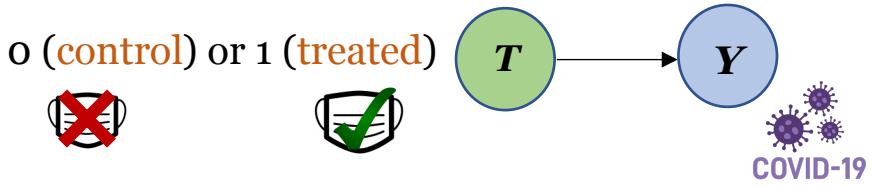
Reference Book:
Guido Imbens & Donald Rubin (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge: Cambridge University Press.

What is Potential Outcome?

Start with a COVID Example

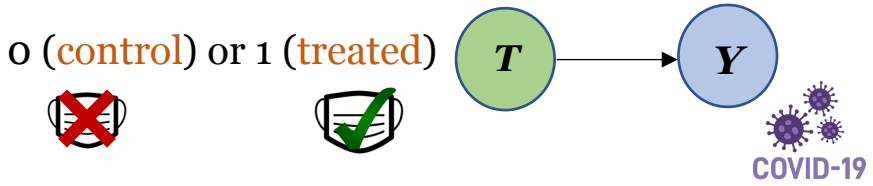
What is Potential Outcome?

- Consider we are interested in the causal relation between a cause (i.e., treatment) on an outcome

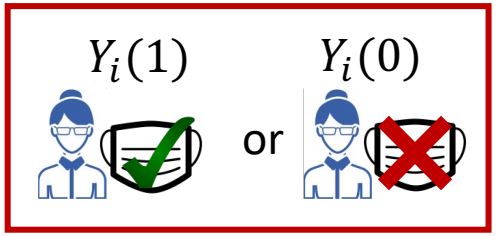


What is Potential Outcome?

- Consider we are interested in the causal relation between a cause (i.e., treatment) on an outcome



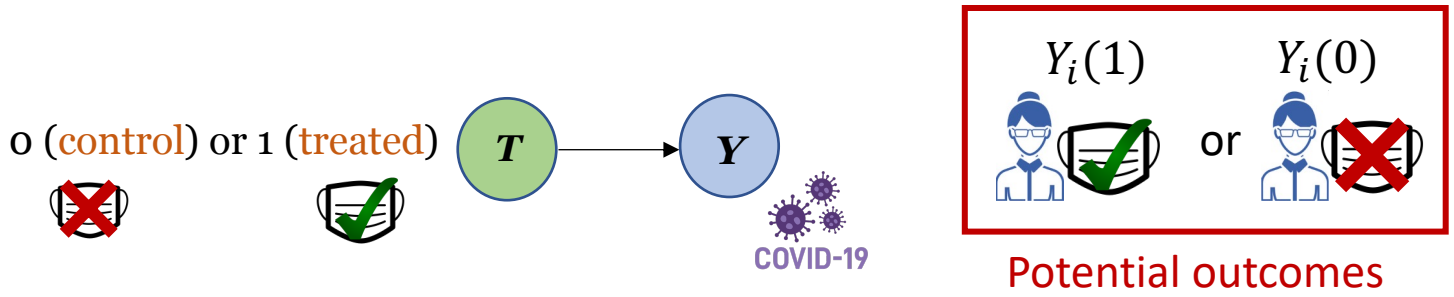
- **Potential outcome** refers to the possible value of an outcome variable that a unit could have if they were exposed to a particular treatment



Potential outcomes

Only one of them can be observed!

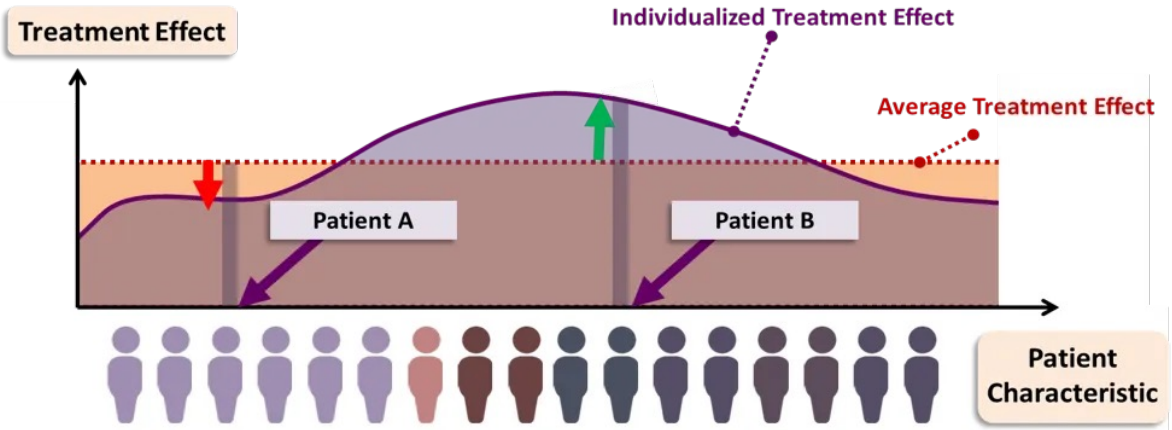
Notations



- i: unit index, each unit can be an individual/instance
- T: observed treatment
- Y: observed outcome
- $Y_i(1)$: potential outcome under $T = 1$
- $Y_i(0)$: potential outcome under $T = 0$

Recap: Treatment Effect

- Individual treatment effect (ITE):
  - ITE $= Y_i(1) - Y_i(0)$
- Average treatment effect (ATE):
  - ATE $= E[Y_i(1) - Y_i(0)]$
- Conditional average treatment effect (CATE):
  - CATE = $E[Y_i(1) - Y_i(0)|X = x]$

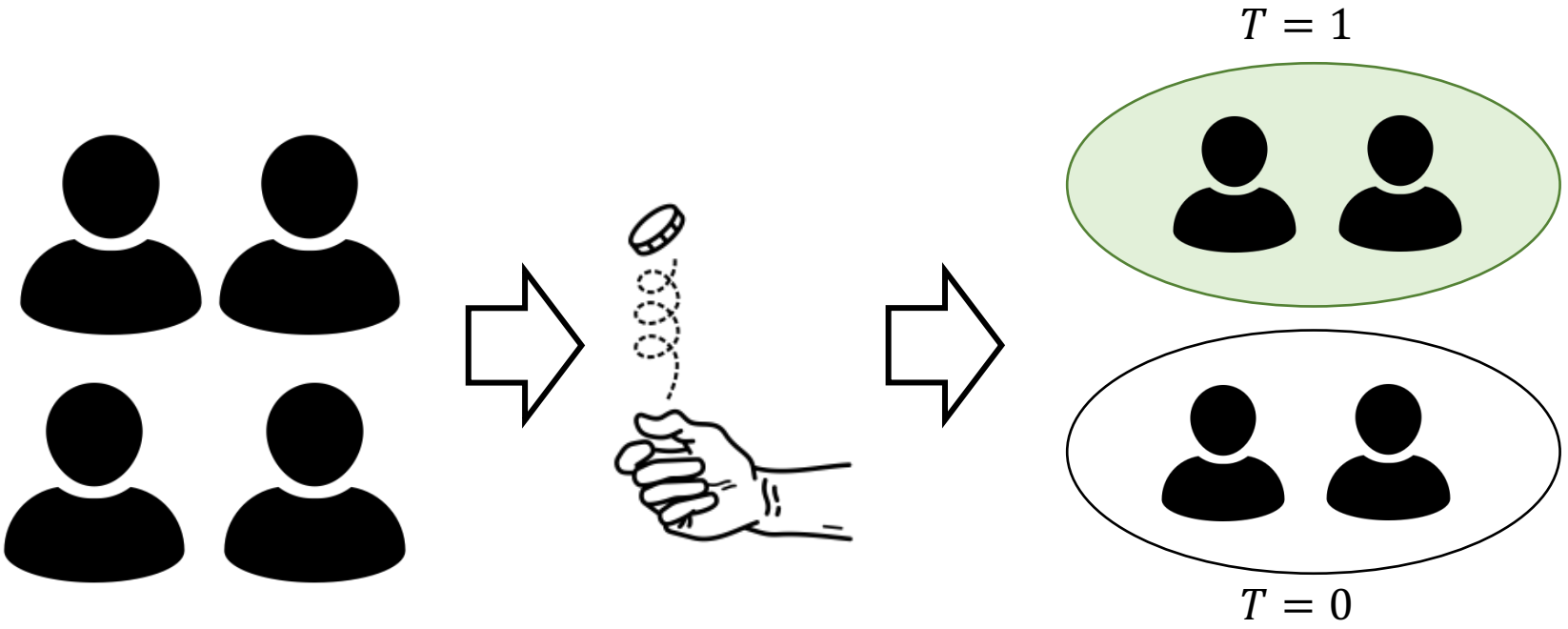


https://www.vanderschaar-lab.com/individualized-treatment-effect-inference/

ITE and ATE

- ITE is unobservable (due to missing of counterfactual outcome)
- ATE = Diff. in averages
- ATE can be estimated relying on less daring assumptions (Holland 1986, 948f)
  - e.g., (proper) random assignment (should) partly satisfy them (“statistical solution”)

Recap: Randomized Controlled Trial (A/B Testing)



Difference-in-means Estimator

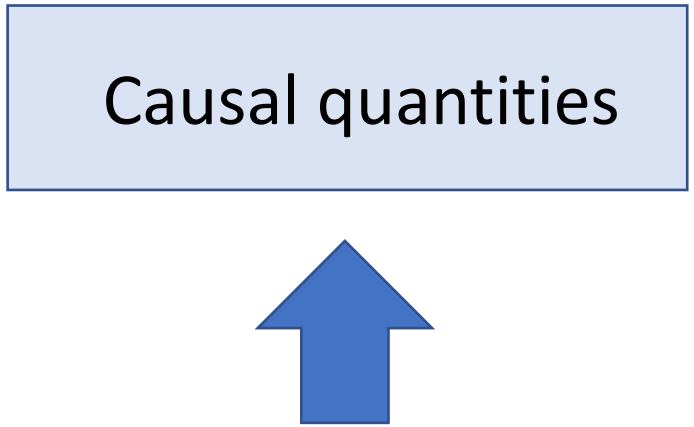
- $\hat{\tau}_{DM} = \frac{1}{n_1}\sum_{T_i=1} Y_i - \frac{1}{n_0}\sum_{T_i=0} Y_i$
- Under RCT, $\hat{\tau}_{DM}$ is unbiased

Experimental Data from RCT

- **Strengths**
  - Gold standard to assess the causal effect.
  - The allocation of the treatment is under control. The distribution of the covariates for treated and control patients is balanced.
- **Weaknesses**
  - Ethical issues
  - Expensive, take a long time to set,
  - Small sample size, due to either recruitment difficulties or restrictive inclusion/exclusion criteria.
  - Narrowly-defined trial sample that is different from the population potentially eligible for the treatment

  Lack of feasibility and generalizability (external validity) to a target population. Study in one company/hospital/state/country could fail to generalize to others

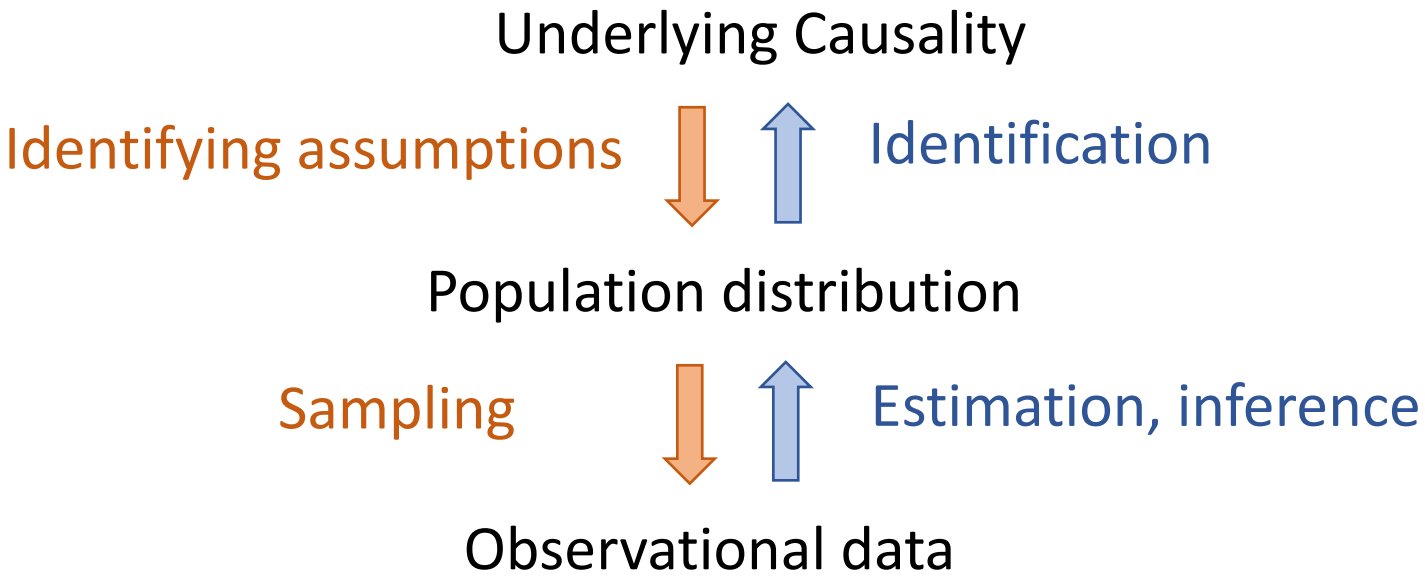
Identifiability



Identifiability: if a causal quantity can be expressed as function of the distribution (i.e., probabilities) of the observed data, we say that the it is identified or identifiable; otherwise, it is unidentified or not identifiable.

Identification and Estimation

- Two components in learning causality
  - (1) Identification
  - (2) Estimation, inference



Identification and Estimation

- Identification:
  - Learning about underlying structures (e.g. a causal effect) from a population distribution (e.g. an expectation)
  - What could one learn from “ideal” data? (a.k.a., if we have an infinitely large sample/the population data/if we know the distribution)

Identification and Estimation

- Identification:
  - To "identify"
    - (1) Take a causal quantity (e.g., a causal effect) → can one write it as a function of the statistical quantities / moments (e.g. expectation, variance) of the distribution of the data?
      - What these moments can identify depends on model's assumptions and other identifying assumptions
    - (2) how do we back out parameters of a structural object (e.g., a model parameter) given knowledge of the population joint distribution of observable variables?

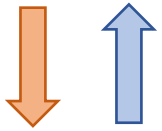
Identification and Estimation

- Estimation, inference:
  - Learning about a population distribution from a finite number of observations.

Identification and Estimation

More formally,

Causal Model $\{F_\theta : \theta \in \Theta\}$

Identifying assumptions  Identification of θ

Population distribution of observable variables $D \sim F$

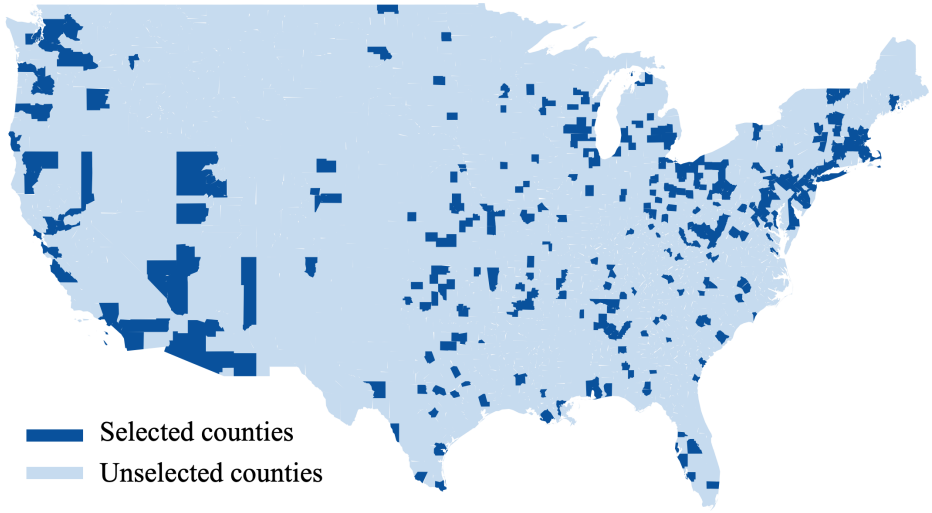
Based on observing D, θ is **point-identified** if the mapping $\theta \rightarrow F(\theta)$ is one-to-one.

Observational Study

- Learning causal effects from observational data
- Scenarios: disease analysis, epidemiological studies, biobanks/ data routinely collected via EHR, insurance claims, administrative data
- Less costly, representative of the target populations
- However, the quality of observational data is often far from perfect

Example: Covid Data

- Covid data in 2020
- 391 counties
- Covariates:
  - Age
  - Gender
  - Medicine
  - …
- Treatment: Over 60+ Covid-19 related policies
- Outcome: Covid infection cases



What issues can exist in observational data for causal effect estimation?

Example: Covid Data

- High-dimensional features: hundreds/thousands of features
- Unbalanced data:
  - Feature distribution may be unaligned or has obvious distribution shift in different locations / hospitals
    - E.g., “medicine record” may include very different choices of medicines in different hospitals
    - Feature values may vary a lot in different locations (e.g., patients in child health hospital only include individuals at young age)
  - Treatment assignment is unbalanced
    - Covid-related policies are quite different in different areas
    - E.g., State A has only 0.03% individuals wearing face mask, while State B has over 99.5% individuals wearing face mask.

Example: Covid Data

- Unmeasured confounders
  - E.g., People's personality, culture background, …
- Interference among units
  - E.g., will the mask-wearing practice of other people influence my own infection risk?

What assumptions should we have?

Assumptions in Observational Studies

Exchangeablibity

- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

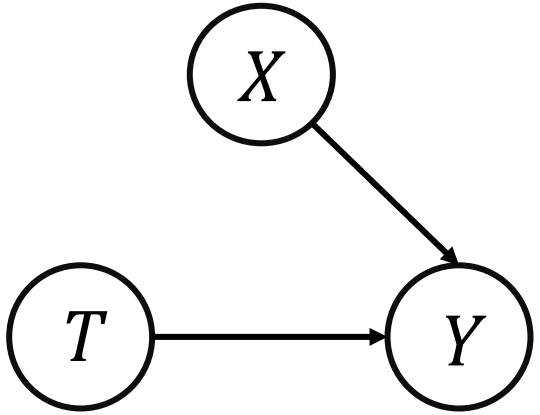
$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$

Exchangeablibity

- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$

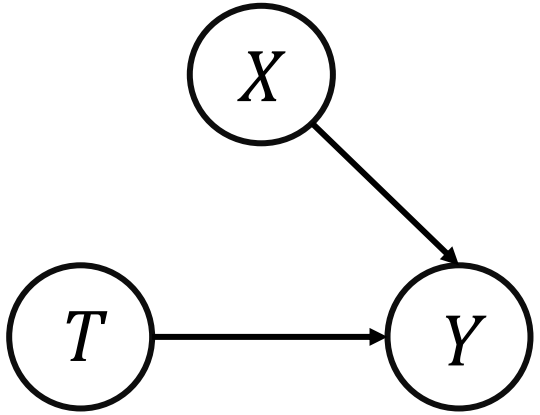


Exchangeablibity

- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$



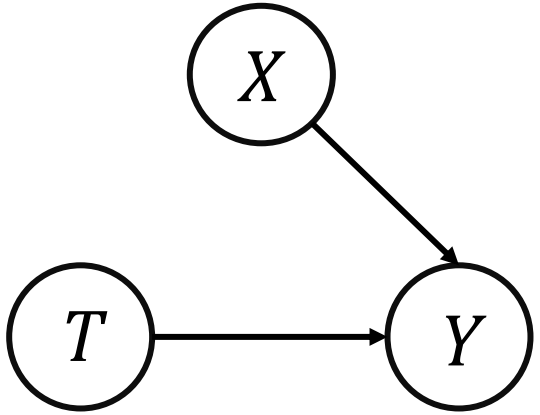
Exchangeability means that the potential outcomes in the treatment group would have the same as the potential outcomes in the control group had individuals in the control group received the treatment.

Exchangeablibity

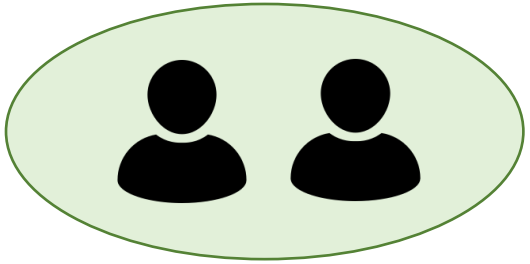
- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

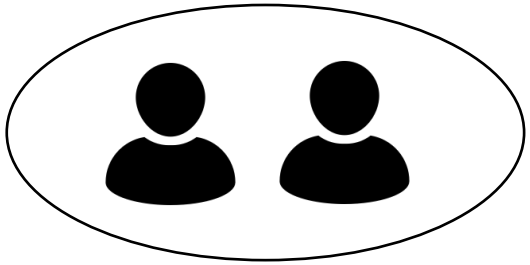
$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$



Treatment group $T = 1$



Control group $T = 0$



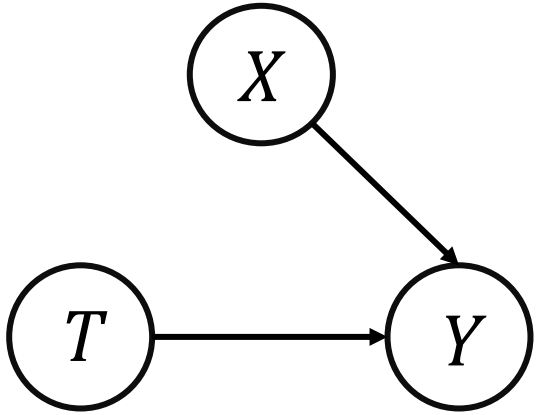
$$E[Y(1)] = E[Y(1)|T = 1] = E[Y(1)|T = 0]$$
$$E[Y(0)] = E[Y(0)|T = 1] = E[Y(0)|T = 0]$$

Exchangeablibity

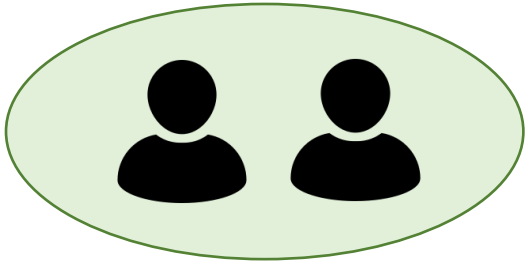
- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

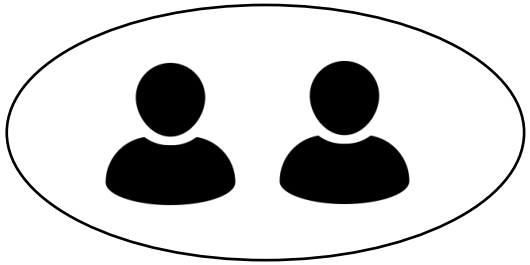
$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$



Treatment group $T = 1$



Control group $T = 0$



$$E[Y(1)] = E[Y(1)|T = 1] = E[Y(1)|T = 0]$$
$$E[Y(0)] = E[Y(0)|T = 1] = E[Y(0)|T = 0]$$

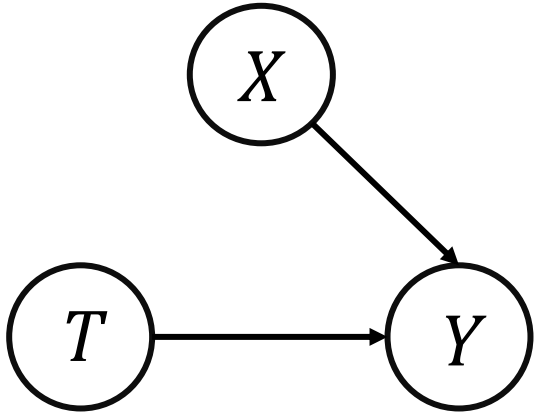
Treatment group and control group are comparable (“exchangeable”)

Exchangeablibity

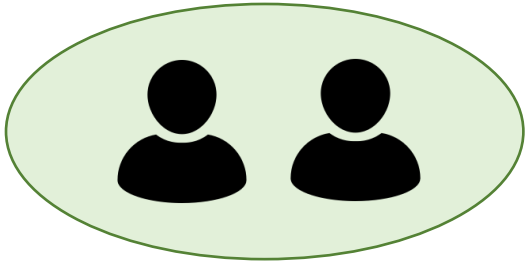
- $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

Caution!

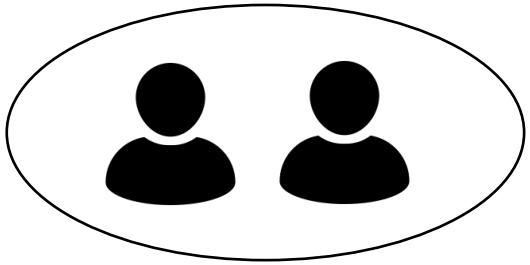
$\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$ is different from $Y \perp\!\!\!\perp T$



Treatment group $T = 1$



Control group $T = 0$



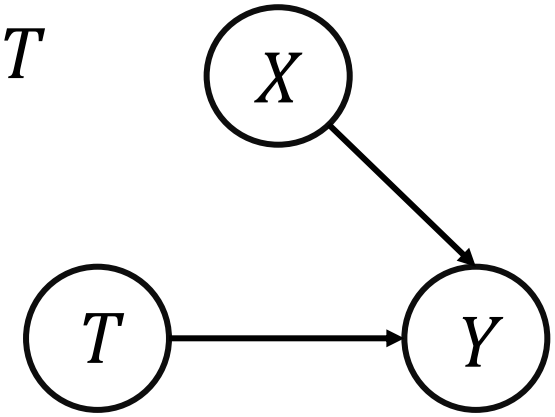
$$E[Y(1)] = E[Y(1)|T = 1] = E[Y(1)|T = 0]$$
$$E[Y(0)] = E[Y(0)|T = 1] = E[Y(0)|T = 0]$$

ATE: $E[Y(1) - Y(0)] = E[Y(1)|T = 1] - E[Y(0)|T = 0]$

Conditional Exchangeability

- Exchangeability: $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T$

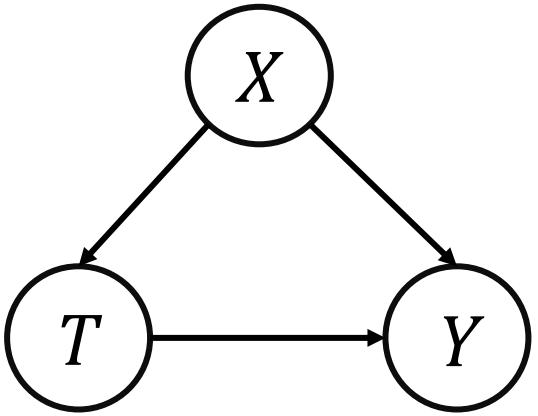
No confounders



- Conditional exchangeability: $\big(Y(1), Y(0)\big) \perp\!\!\!\perp T \mid X$

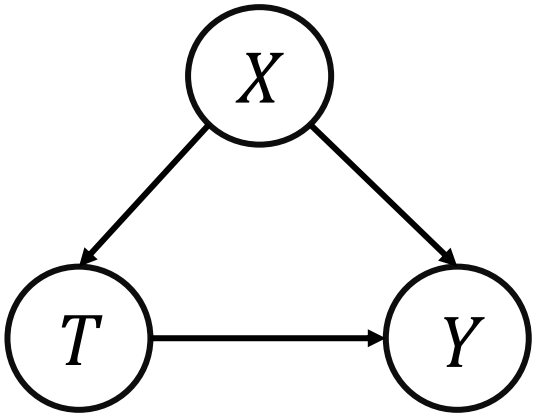
No unmeasured confounders

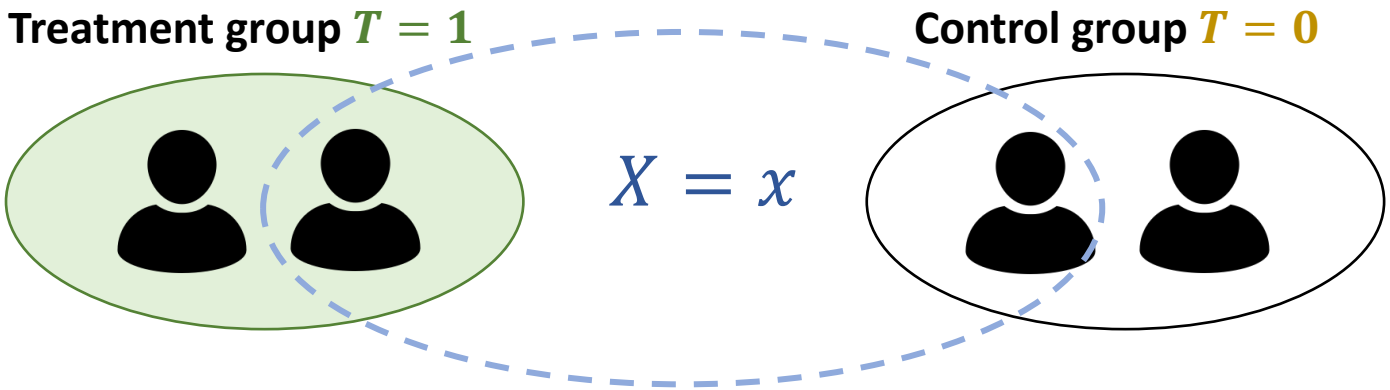
"Unconfoundedness/ignorability"



Conditional Exchangeability

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





CATE: $E[Y(1) - Y(0)|X] = E[Y(1)|X] - E[Y(0)|X] = E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]$

ATE: $E[Y(1) - Y(0)] = E_X[E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]]$

Positivity / Overlap

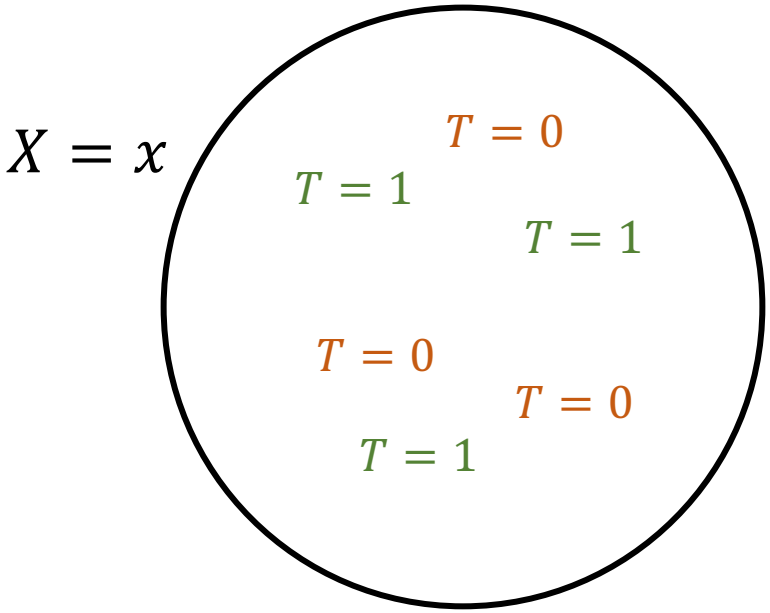
- For all values of $X = x$ with $P(X = x) > 0$ in the population of interest:

$$P(T = t|X = x) > 0$$

Positivity / Overlap

- For all values of $X = x$ with $P(X = x) > 0$ in the population of interest:

$$P(T = t|X = x) > 0$$



Tradeoff between Positivity and Unconfoundedness

- Conditioning on more covariates
  - higher chance of satisfying unconfoundedness
  - higher chance of violating positivity
- Example:
  - Conditioning on 1 dimension – 50% overlap
  - Conditioning on 2 dimension – 25% overlap
  - …

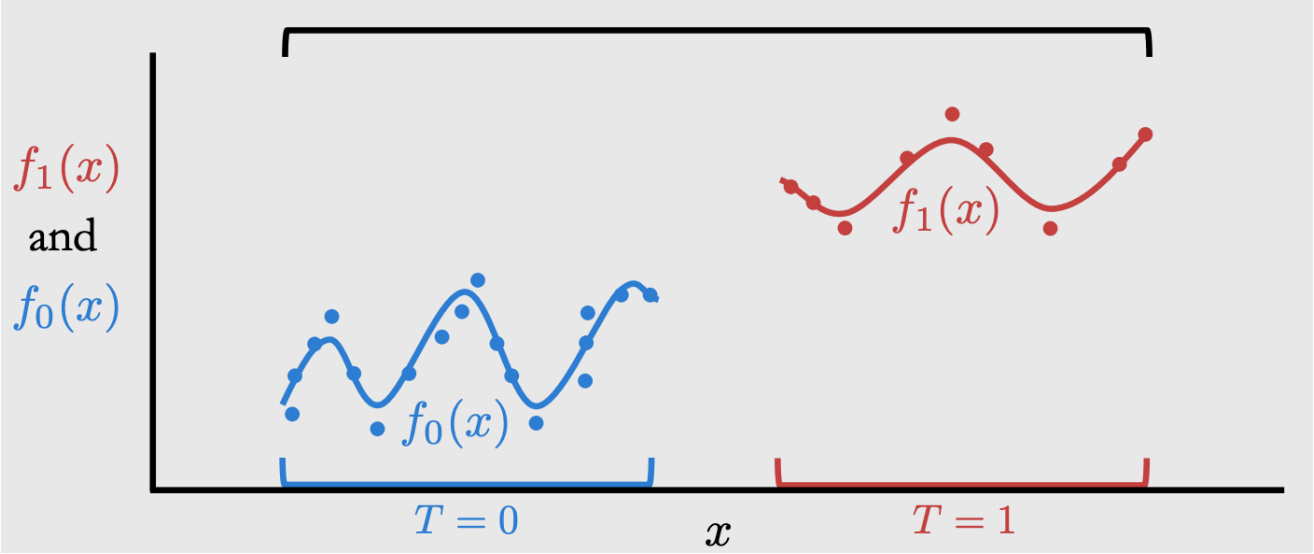
  Related to the Curse of dimensionality

Extrapolation and Overlap

- Violations of the positivity assumption can lead to demanding too much from models and getting very bad behavior in return.
- Adjustment formula:

Model with $f_1(x)$ Model with $f_0(x)$

$$E_X[E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]]$$



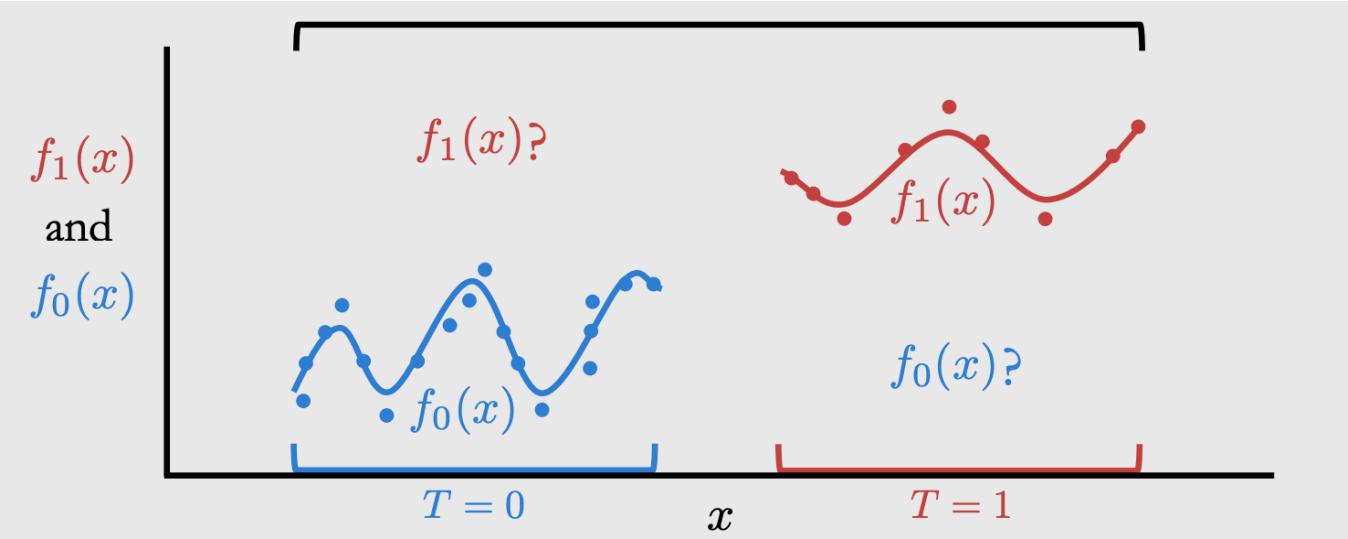
Brady Neal. Introduction to Causal Inference from a Machine Learning Perspective

Extrapolation and Overlap

- Violations of the positivity assumption can lead to demanding too much from models and getting very bad behavior in return.
- Adjustment formula:

Model with $f_1(x)$ Model with $f_0(x)$

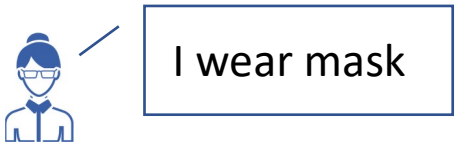
$$E_X[E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]]$$



Brady Neal. Introduction to Causal Inference from a Machine Learning Perspective

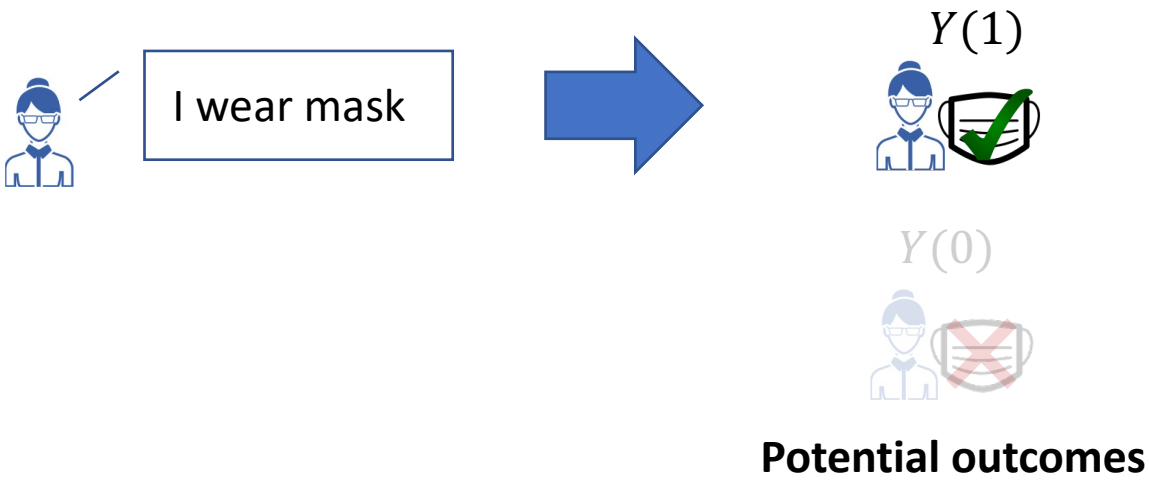
Consistency

- $Y = Y(t)$ when $T = t$



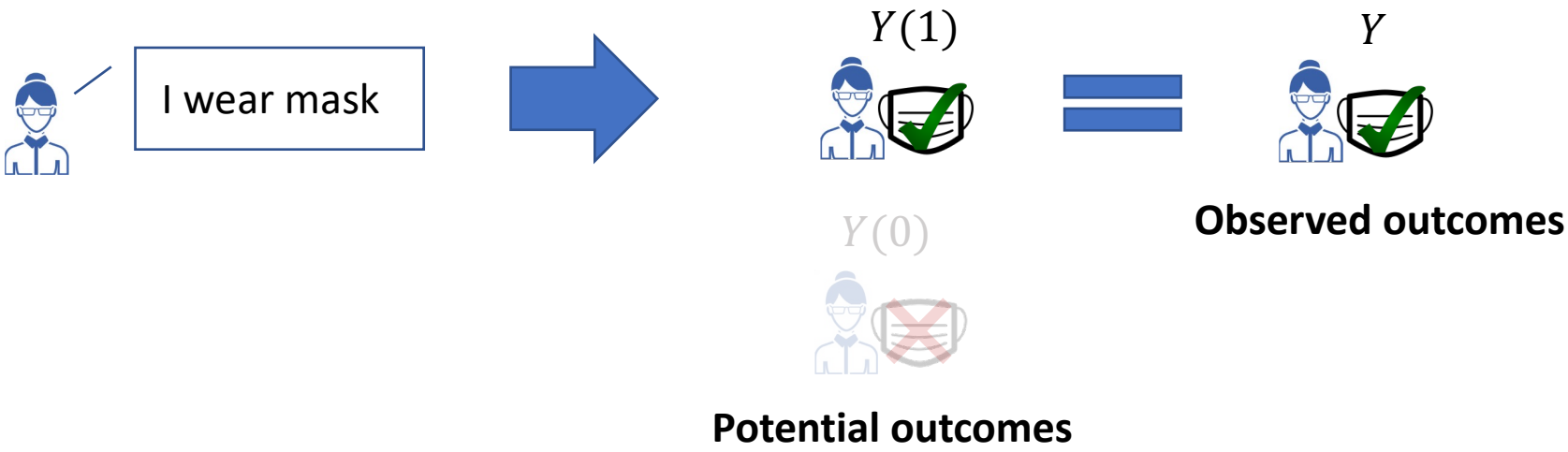
Consistency

- $Y = Y(t)$ when $T = t$



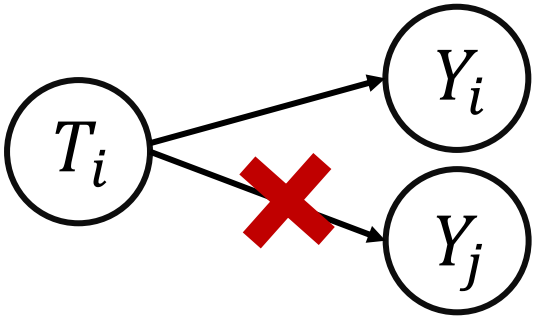
Consistency

- $Y = Y(t)$ when $T = t$



Stable Unit Treatment Value Assumption (SUTVA)

- The potential outcomes for any unit do not vary with the treatments assigned to other units.
  - No **interference**



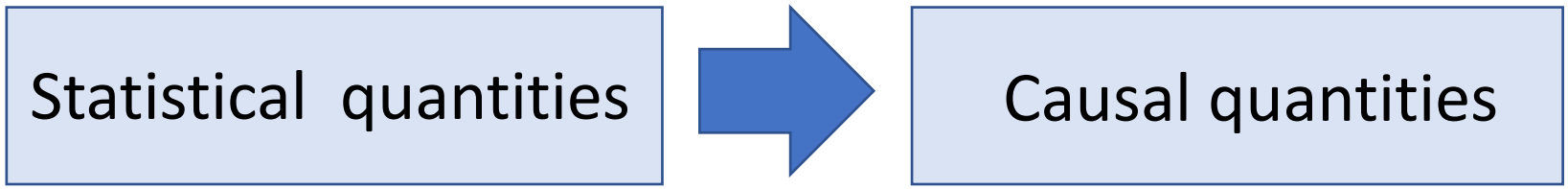
- For each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.
  - E.g., when treatment is “take a surgery”, this surgery is operated by the same surgeon with the same procedure

Go Back to Identifiability

- ATE:

$$E[Y(1) - Y(0)]$$

$$= E[Y(1)] - E[Y(0)]$$



Go Back to Identifiability

- ATE:

$$E[Y(1) - Y(0)]$$

$$= E[Y(1)] - E[Y(0)]$$

$$= E_X[E[Y(1)|X] - E[Y(0)|X]]$$ Law of total expectation

Statistical quantities → Causal quantities

Go Back to Identifiability

- ATE:

$$E[Y(1) - Y(0)]$$

$$= E[Y(1)] - E[Y(0)]$$

$$= E_X[E[Y(1)|X] - E[Y(0)|X]]$$ Law of total expectation

$$= E_X[E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]]$$ Unconfoundedness & positivity

Statistical quantities → Causal quantities

Go Back to Identifiability

- ATE:

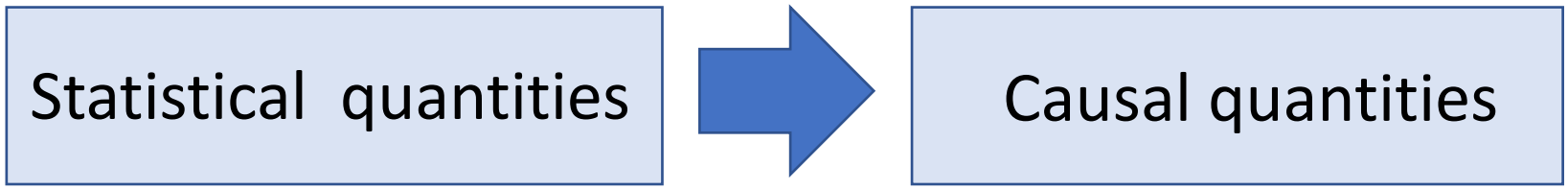
$$E[Y(1) - Y(0)]$$

$$= E[Y(1)] - E[Y(0)]$$

$$= E_X[E[Y(1)|X] - E[Y(0)|X]]$$ Law of total expectation

$$= E_X[E[Y(1)|X, T = 1] - E[Y(0)|X, T = 0]]$$ Unconfoundedness & positivity

$$= E_X[E[Y|X, T = 1] - E[Y|X, T = 0]]$$ consistency



Weaker Assumptions?

- In real world, these assumptions may not be fully satisfied
- Many efforts have been made to estimate causal effects under weaker assumptions

An Example: Causal Effect Estimation

Problem: effect of sodium intake on blood pressure

- Motivation: 46% of Americans have high blood pressure and high blood pressure is associated with increased risk of mortality [1]

Luque-Fernandez et al. (2018), 'Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application'

Problem: effect of sodium intake on blood pressure

- Motivation: 46% of Americans have high blood pressure and high blood pressure is associated with increased risk of mortality [1]
- Data:
  - Outcome Y: (systolic) blood pressure (continuous)
  - Treatment T: sodium intake (1 if above 3.5 mg and 0 if below)
  - Covariates X: age and amount of protein excreted in urine
  - Simulation: we know the true ATE is 1.05

Luque-Fernandez et al. (2018), 'Educational Note: Paradoxical collider effect in the analysis of non-communicable disease epidemiological data: a reproducible illustration and web application'

ATE Estimation

- True ATE: $E[Y(1) - Y(0)] = 1.05$

ATE Estimation

- True ATE: $E[Y(1) - Y(0)] = 1.05$
- Identification:
  $E[Y(1) - Y(0)] = E_X[E[Y|T = 1, X] - E[Y|T = 0, X]]$

ATE Estimation

- True ATE: $E[Y(1) - Y(0)] = 1.05$
- Identification:
  $E[Y(1) - Y(0)] = E_X[E[Y|T = 1, X] - E[Y|T = 0, X]]$
- Estimation:

$$\frac{1}{n}\sum_x [E[Y|T = 1, x] - E[Y|T = 0, x]]$$

Model (linear regression)

ATE Estimation

```
1  import numpy as np
2  import pandas as pd
3  from sklearn.linear_model import LinearRegression
4
5  Xt = df[['sodium', 'age', 'proteinuria']]
6  y = df['blood_pressure']
7  model = LinearRegression()
8  model.fit(Xt, y)
9
10 Xt1 = pd.DataFrame.copy(Xt)
11 Xt1['sodium'] = 1
12 Xt0 = pd.DataFrame.copy(Xt)
13 Xt0['sodium'] = 0
14 ate_est = np.mean(model.predict(Xt1) - model.predict(Xt0))
15 print('ATE estimate:', ate_est)
```

ATE Estimation

- True ATE: $E[Y(1) - Y(0)] = 1.05$
- Identification:
  $E[Y(1) - Y(0)] = E_X[E[Y|T = 1, X] - E[Y|T = 0, X]]$
- Estimation:

$$\frac{1}{n}\sum_x [E[Y|T = 1, x] - E[Y|T = 0, x]]$$

Model (linear regression)

Estimate: 0.85

Bias: $\frac{|0.85-1.05|}{1.05} \times 100\% = 19\%$

ATE Estimation

- True ATE: $E[Y(1) - Y(0)] = 1.05$
- Identification:
  $E[Y(1) - Y(0)] = E_X[E[Y|T = 1, X] - E[Y|T = 0, X]]$
- Estimation:

$$\frac{1}{n}\sum_x [E[Y|T = 1, x] - E[Y|T = 0, x]]$$

Model (linear regression)

Estimate: 0.85

Bias: $\frac{|0.85-1.05|}{1.05} \times 100\% = 19\%$

- Naïve estimator: $E[Y|T = 1] - E[Y|T = 0]$
  - Naïve estimate: 5.33

Bias: $\frac{|5.33-1.05|}{1.05} \times 100\% = 407\%$

Using coefficient of linear regression

- Assumption: linear parametric form

$$Y = \alpha T + \beta X$$

- Estimation: linear regression:

$$Y = \hat{\alpha} T + \hat{\beta} X$$

Using coefficient of linear regression

- Assumption: linear parametric form

$$Y = \alpha T + \beta X$$

- Estimation: linear regression:

$$Y = \hat{\alpha} T + \hat{\beta} X$$

- limitations: the causal effect is the same for all individuals

$$Y(1) - Y(0) = \alpha \times 1 + \beta X \text{ - } (\alpha \times 0 + \beta X) = \alpha$$

Heterogeneous Treatment Effect

- In some literature, treatment effect is assumed to be the homogeneous (the same for all units)
- For different units, treatment effect may be different (heterogeneous)
  - E.g., people at different ages may have different responses for a medicine

- **Effect modifier**: factors that change the treatment effect

Reading after Class

- Miguel Hernan, Jamie Robins. “Causal Inference: What If” -- Book Chapter 3 “Observational studies”

Thank you!
Questions?