Statistical Models for Causal Analysis

Teppei Yamamoto

Keio University

Introduction to Causal Inference
Spring 2016

Three Modes of Statistical Inference

1. Descriptive Inference: summarizing and exploring data
   - Inferring ideal points from roll-call votes
   - Inferring topics from texts and speeches
   - Inferring social networks from surveys
2. Predictive Inference: forecasting out-of-sample data points
   - Inferring future state failures from past failures
   - Inferring population average turnout from a sample of voters
   - Inferring individual level behavior from aggregate data
3. Causal Inference: predicting *counterfactuals*
   - Inferring the effects of ethnic minority rule on civil war onset
   - Inferring why incumbency status affects election outcomes
   - Inferring whether the lack of war among democracies can be attributed to regime types

Three Modes of Statistical Inference

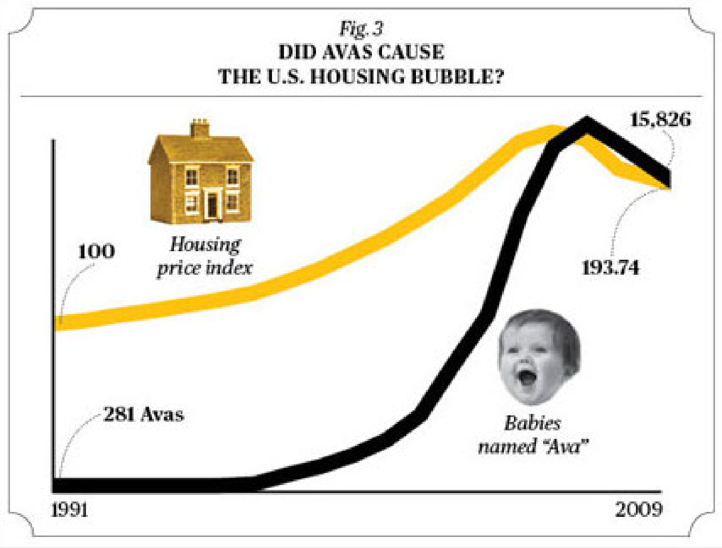
1. Descriptive Inference: summarizing and exploring data
   - Inferring ideal points from roll-call votes
   - Inferring topics from texts and speeches
   - Inferring social networks from surveys

2. Predictive Inference: forecasting out-of-sample data points
   - Inferring future state failures from past failures
   - Inferring population average turnout from a sample of voters
   - Inferring individual level behavior from aggregate data

3. Causal Inference: predicting *counterfactuals*
   - Inferring the effects of ethnic minority rule on civil war onset
   - Inferring why incumbency status affects election outcomes
   - Inferring whether the lack of war among democracies can be attributed to regime types

Causal inference is the most difficult of the three.

Correlation ≠ Causation



What is Causal Inference?

Causal inference $=$ inference about counterfactuals

What is Causal Inference?

Causal inference $=$ inference about counterfactuals

Examples:

- Incumbency advantage:
  What *would* have been the election outcome if the candidate had not been an incumbent?
- Democratic peace:
  Would the two countries have fought each other if they had been both autocratic?
- Policy intervention:
  How many more disadvantaged youths *would* get employed under the new job training program?

What is Causal Inference?

Causal inference $=$ inference about counterfactuals

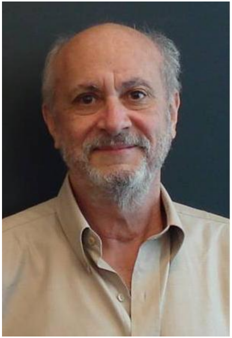
Examples:

- Incumbency advantage:
  What *would* have been the election outcome if the candidate had not been an incumbent?
- Democratic peace:
  Would the two countries have fought each other if they had been both autocratic?
- Policy intervention:
  How many more disadvantaged youths *would* get employed under the new job training program?

We need a statistical model that can explicitly distinguish factuals and counterfactuals.

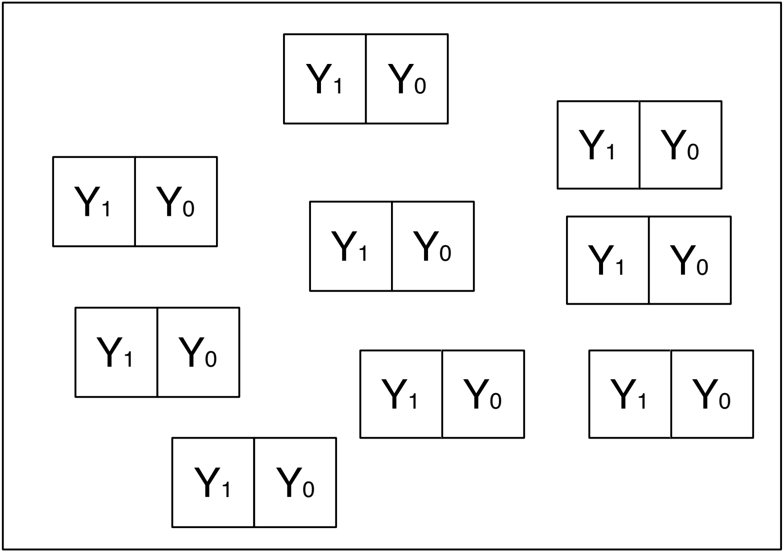
Neyman-Rubin Potential Outcomes Model

Jerzy Neyman (1894–1981)

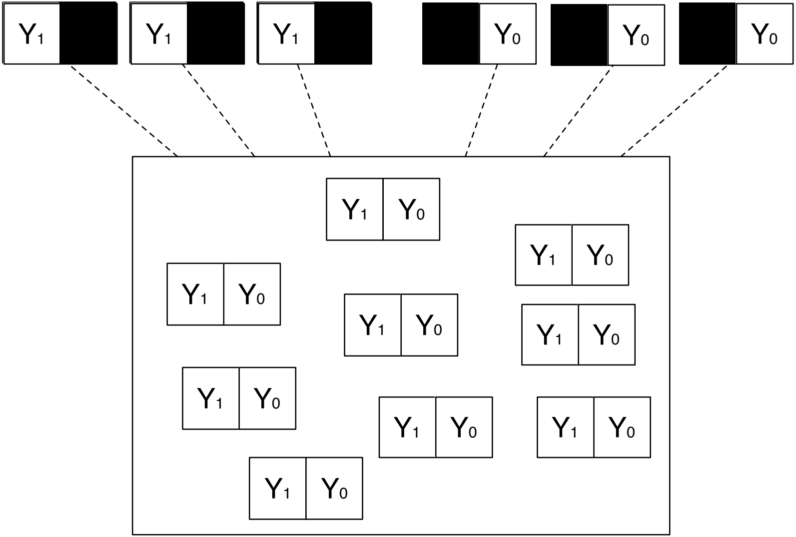


Donald Rubin (1943–)

Neyman Urn Model



Neyman Urn Model



Causality with Potential Outcomes

Definition (Treatment)

D_i: Indicator of treatment intake for unit i, where $i = 1, ..., N$

$$D_i = \begin{cases} 1 & \text{if unit } i \text{ received the treatment} \\ 0 & \text{otherwise} \end{cases}$$

Definition (Observed Outcome)

Y_i: Variable of interest whose value may be affected by the treatment

Definition (Potential Outcomes)

Y_{di}: Value of the outcome that *would* be realized if unit i received the treatment d, where $d = 0$ or 1

$$Y_{di} = \begin{cases} Y_{1i} & \text{Potential outcome for unit } i \text{ with treatment} \\ Y_{0i} & \text{Potential outcome for unit } i \text{ without treatment} \end{cases}$$

Alternative notation: $Y_i(d)$, Y_i^d, etc.

Causality with Potential Outcomes

Definition (Causal Effect, or Unit Treatment Effect)

Causal effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

Causality with Potential Outcomes

Definition (Causal Effect, or Unit Treatment Effect)

Causal effect of the treatment on the outcome for unit *i* is the difference between its two potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

Note that observed outcomes are realized from potential outcomes as

$$Y_i = Y_{D_i i} = D_i Y_{1i} + (1 - D_i) Y_{0i} \quad \text{so} \quad Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

Causality with Potential Outcomes

Definition (Causal Effect, or Unit Treatment Effect)

Causal effect of the treatment on the outcome for unit i is the difference between its two potential outcomes:

$$\tau_i = Y_{1i} - Y_{0i}$$

Note that observed outcomes are realized from potential outcomes as

$$Y_i = Y_{D_i i} = D_i Y_{1i} + (1 - D_i) Y_{0i} \quad \text{so} \quad Y_i = \begin{cases} Y_{1i} & \text{if } D_i = 1 \\ Y_{0i} & \text{if } D_i = 0 \end{cases}$$

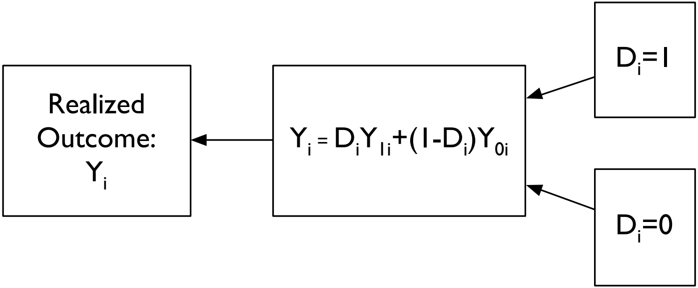
Fundamental Problem of Causal Inference (Holland 1986):

We can never observe both Y_{1i} and Y_{0i} for the same i

This makes τ_i unidentifiable without further assumptions.

Causal Inference as a Missing Data Problem

Problem: Causal Inference is difficult because it involves missing data.
How can we calculate $\tau_i = Y_{1i} - Y_{0i}$?



Causal Inference as a Missing Data Problem

Problem: Causal Inference is difficult because it involves missing data. How can we calculate $\tau_i = Y_{1i} - Y_{0i}$?

- One "solution": Assume unit homogeneity

$$\tau_i = \tau \quad \text{for all } i$$

  - If Y_{1i} and Y_{0i} are constant across individual units, then cross-sectional comparisons will recover $\tau = \tau_i$
  - If Y_{1i} and Y_{0i} are constant across time, then before-and-after comparisons will recover $\tau = \tau_i$

Causal Inference as a Missing Data Problem

Problem: Causal Inference is difficult because it involves missing data.
How can we calculate $\tau_i = Y_{1i} - Y_{0i}$?

- One "solution": Assume unit homogeneity

$$\tau_i = \tau \quad \text{for all } i$$

  - If Y_{1i} and Y_{0i} are constant across individual units, then cross-sectional comparisons will recover $\tau = \tau_i$
  - If Y_{1i} and Y_{0i} are constant across time, then before-and-after comparisons will recover $\tau = \tau_i$

- This may be sometimes plausible in physical sciences

- Unfortunately, rarely true in social sciences

Stable Unit Treatment Value Assumption (SUTVA)

- Recall: $Y_i = Y_{D_i i}$, or equivalently $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$
- This notation implicitly makes the following assumption:

Stable Unit Treatment Value Assumption (SUTVA)

- Recall: $Y_i = Y_{D_i i}$, or equivalently $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$
- This notation implicitly makes the following assumption:

Assumption (SUTVA)

$$Y_{(D_1, D_2, \ldots, D_N)i} \;=\; Y_{(D'_1, D'_2, \ldots, D'_N)i} \quad \textit{if} \quad D_i = D'_i$$

Stable Unit Treatment Value Assumption (SUTVA)

- Recall: $Y_i = Y_{D_i i}$, or equivalently $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$
- This notation implicitly makes the following assumption:

Assumption (SUTVA)

$$Y_{(D_1, D_2, \ldots, D_N)i} = Y_{(D'_1, D'_2, \ldots, D'_N)i} \quad \textit{if} \quad D_i = D'_i$$

SUTVA consists of two sub-assumptions:

1. No interference between units
   - Potential outcomes for a unit must not be affected by treatment for any other units
   - Violations: spill-over effects, contagion, dilution

Stable Unit Treatment Value Assumption (SUTVA)

- Recall: $Y_i = Y_{D_i i}$, or equivalently $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$
- This notation implicitly makes the following assumption:

Assumption (SUTVA)

$$Y_{(D_1, D_2, \ldots, D_N) i} = Y_{(D'_1, D'_2, \ldots, D'_N) i} \quad \textit{if} \quad D_i = D'_i$$

SUTVA consists of two sub-assumptions:

1. No interference between units
   - Potential outcomes for a unit must not be affected by treatment for any other units
   - Violations: spill-over effects, contagion, dilution
2. No different versions of treatment (a.k.a. stability, consistency)
   - Nominally identical treatments are in fact identical
   - Violations: variable levels of treatment, technical errors

Stable Unit Treatment Value Assumption (SUTVA)

- Recall: $Y_i = Y_{D_i i}$, or equivalently $Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$
- This notation implicitly makes the following assumption:

Assumption (SUTVA)

$$Y_{(D_1, D_2, \ldots, D_N)i} = Y_{(D'_1, D'_2, \ldots, D'_N)i} \quad \text{if} \quad D_i = D'_i$$

SUTVA consists of two sub-assumptions:

1. No interference between units
   - Potential outcomes for a unit must not be affected by treatment for any other units
   - Violations: spill-over effects, contagion, dilution
2. No different versions of treatment (a.k.a. stability, consistency)
   - Nominally identical treatments are in fact identical
   - Violations: variable levels of treatment, technical errors

Violation examples: Vaccination, fertilizer on plot yield, communication

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

$$(D_1, D_2) \ = \ (0, 0) \text{ or } (1, 0) \text{ or } (0, 1) \text{ or } (1, 1)$$

How many potential outcomes for unit 1?

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

$$(D_1, D_2) \;=\; (0,0) \text{ or } (1,0) \text{ or } (0,1) \text{ or } (1,1)$$

How many potential outcomes for unit 1?

$$Y_{(0,0)1},\;\; Y_{(1,0)1},\;\; Y_{(0,1)1}\;\; Y_{(1,1)1}.$$

How many causal effects for unit 1?

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

$$(D_1, D_2) \ = \ (0,0) \text{ or } (1,0) \text{ or } (0,1) \text{ or } (1,1)$$

How many potential outcomes for unit 1?

$$Y_{(0,0)1}, \ Y_{(1,0)1}, \ Y_{(0,1)1} \ Y_{(1,1)1}.$$

How many causal effects for unit 1?

$$\begin{array}{ll} Y_{(1,1)1} - Y_{(0,0)1}, & Y_{(1,1)1} - Y_{(0,1)1}, \\ Y_{(1,0)1} - Y_{(0,0)1}, & Y_{(1,0)1} - Y_{(0,1)1}, \\ Y_{(1,1)1} - Y_{(1,0)1}, & Y_{(0,1)1} - Y_{(0,0)1}. \end{array}$$

How many observed outcomes for unit 1?

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

$$(D_1, D_2) \ = \ (0,0) \text{ or } (1,0) \text{ or } (0,1) \text{ or } (1,1)$$

How many potential outcomes for unit 1?

$$Y_{(0,0)1}, \ Y_{(1,0)1}, \ Y_{(0,1)1} \ Y_{(1,1)1}.$$

How many causal effects for unit 1?

$$\begin{array}{ll} Y_{(1,1)1} - Y_{(0,0)1}, & Y_{(1,1)1} - Y_{(0,1)1}, \\ Y_{(1,0)1} - Y_{(0,0)1}, & Y_{(1,0)1} - Y_{(0,1)1}, \\ Y_{(1,1)1} - Y_{(1,0)1}, & Y_{(0,1)1} - Y_{(0,0)1}. \end{array}$$

How many observed outcomes for unit 1? Only one: $Y_1 = Y_{(D_1,D_2)1}$

Causal Inference without SUTVA

Let $\mathbf{D} = (D_1, D_2)$ be a vector of binary treatments for $N = 2$.

How many different values can $\mathbf{D}$ possibly take?

$$(D_1, D_2) \ = \ (0,0) \text{ or } (1,0) \text{ or } (0,1) \text{ or } (1,1)$$

How many potential outcomes for unit 1?

$$Y_{(0,0)1}, \ Y_{(1,0)1}, \ Y_{(0,1)1} \ Y_{(1,1)1}.$$

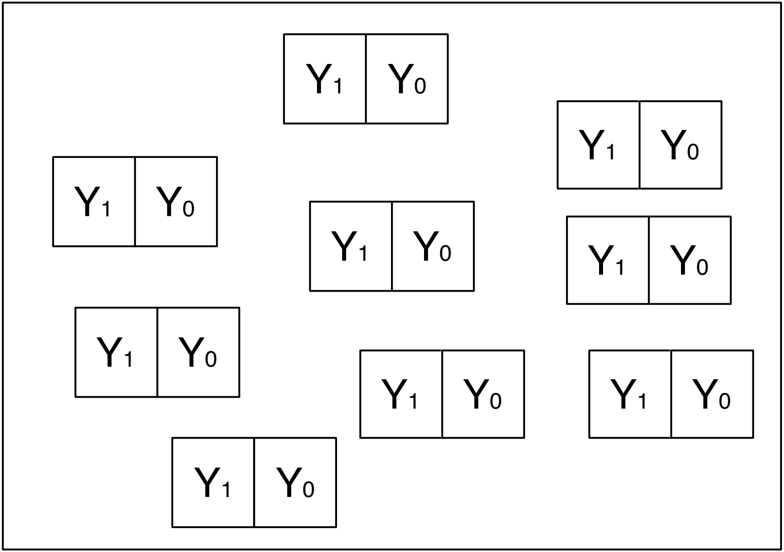
How many causal effects for unit 1?

$$\begin{array}{ll} Y_{(1,1)1} - Y_{(0,0)1}, & Y_{(1,1)1} - Y_{(0,1)1}, \\ Y_{(1,0)1} - Y_{(0,0)1}, & Y_{(1,0)1} - Y_{(0,1)1}, \\ Y_{(1,1)1} - Y_{(1,0)1}, & Y_{(0,1)1} - Y_{(0,0)1}. \end{array}$$

How many observed outcomes for unit 1? Only one: $Y_1 = Y_{(D_1,D_2)1}$

Without SUTVA, causal inference becomes exponentially more difficult as N increases.

Back to the Neyman Urn Model



Causal Quantities of Interest, or Estimands

- Unit-level causal effects are fundamentally unobservable
- We instead focus on *averages* in most situations
- For now, assume we observe all units in the population of interest, i.e. $N =$ size of population

Causal Quantities of Interest, or Estimands

- Unit-level causal effects are fundamentally unobservable
- We instead focus on *averages* in most situations
- For now, assume we observe all units in the population of interest, i.e. $N =$ size of population

Definition (Average treatment effect, ATE)

$$\tau_{ATE} = \frac{1}{N}\sum_{i=1}^{N}\{Y_{1i} - Y_{0i}\}$$

or equivalently

$$\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$$

- Note that τ_{ATE} is still unidentified
- In the rest of this course, we will consider various assumptions under which τ_{ATE} can be identified from observed information

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1}\sum_{i=1}^{N} D_i \{Y_{1i} - Y_{0i}\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \{Y_{1i} - Y_{0i}\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} \;=\; \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals the number of treated units.

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \{Y_{1i} - Y_{0i}\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} \; = \; \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals the number of treated units.
- When would $\tau_{ATT} \neq \tau_{ATE}$?

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \{Y_{1i} - Y_{0i}\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} \;=\; \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals the number of treated units.
- When would $\tau_{ATT} \neq \tau_{ATE}$? When D_i and Y_{di} are associated.
- Exercise: Define τ_{ATC}, ATE on the untreated (control) units.

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \{Y_{1i} - Y_{0i}\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals the number of treated units.
- When would $\tau_{ATT} \neq \tau_{ATE}$? When D_i and Y_{di} are associated.
- Exercise: Define τ_{ATC}, ATE on the untreated (control) units.

Definition (Conditional average treatment effects)

$$\tau_{CATE}(x) = \mathbb{E}[Y_{1i} - Y_{0i} | X_i = x]$$

where X_i is a pre-treatment covariate for unit i

- In words, $\tau_{CATE}(x)$ is

Other Causal Estimands

Definition (Average treatment effect on the treated, ATT)

$$\tau_{ATT} = \frac{1}{N_1} \sum_{i=1}^{N} D_i \left\{ Y_{1i} - Y_{0i} \right\} \quad \text{where} \quad N_1 = \sum_{i=1}^{N} D_i$$

or equivalently $\tau_{ATT} \; = \; \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$

- In words, N_1 equals the number of treated units.
- When would $\tau_{ATT} \neq \tau_{ATE}$? When D_i and Y_{di} are associated.
- Exercise: Define τ_{ATC}, ATE on the untreated (control) units.

Definition (Conditional average treatment effects)

$$\tau_{CATE}(x) = \mathbb{E}[Y_{1i} - Y_{0i} | X_i = x]$$

where X_i is a pre-treatment covariate for unit i

- In words, $\tau_{CATE}(x)$ is a subgroup effect, treatment effect on units who have particular characteristics x.

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i
1	1	3
2	1	1
3	0	0
4	0	1

What is $\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$?

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i
1	1	3
2	1	1
3	0	0
4	0	1
$\mathbb{E}[Y_i \mid D_i = 1]$		2
$\mathbb{E}[Y_i \mid D_i = 0]$		0.5
$\mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0]$		1.5

What is $\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$?

Naïve estimator:

$$\begin{aligned} \tilde{\tau} &= \mathbb{E}[Y_i \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 0] \quad \text{(observed difference in means)} \\ &= \frac{3+1}{2} - \frac{0+1}{2} = 1.5 \quad \text{Could this be wrong?} \end{aligned}$$

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	0	1	?	1	?

What is $\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}]$? We need potential outcomes that we do not observe!

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	?
2	1	1	1	1	?
3	0	0	1	0	?
4	0	1	1	1	?

Suppose hypothetically: $Y_{01} = 0, Y_{02} = Y_{13} = Y_{14} = 1$.

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	0	1	1	1	0
$\mathbb{E}[Y_{1i}]$			1.5		
$\mathbb{E}[Y_{0i}]$				0.5	
$\mathbb{E}[Y_{1i} - Y_{0i}]$					1

$$\tau_{ATE} \;=\; \mathbb{E}[Y_{1i} - Y_{0i}] \;=\; \mathbb{E}[\tau_i] \;=\; \frac{3+0+1+0}{4} \;=\; 1.$$

Illustration: Average Treatment Effect

Suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	0	1	1	1	0
$\mathbb{E}[Y_{1i}]$			1.5		
$\mathbb{E}[Y_{0i}]$				0.5	
$\mathbb{E}[Y_{1i} - Y_{0i}]$					1

$$\tau_{ATE} = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[\tau_i] = \frac{3+0+1+0}{4} = 1.$$

Why $\tau_{ATE} \neq \tilde{\tau}$? When would they be equal?

Illustration: Average Treatment Effect on the Treated

Again suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	?	?
2	1	1	1	?	?
3	0	0	?	0	?
4	0	1	?	1	?

What is $\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$?

Illustration: Average Treatment Effect on the Treated

Again suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	0	1	1	1	0

What is $\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$?

Illustration: Average Treatment Effect on the Treated

Again suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	0	1	1	1	0
$\mathbb{E}[Y_{1i} \mid D_i = 1]$			2		
$\mathbb{E}[Y_{0i} \mid D_i = 1]$				0.5	
$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$					1.5

$$\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \mathbb{E}[\tau_i \mid D_i = 1] = \frac{3+0}{2} = 1.5.$$

Illustration: Average Treatment Effect on the Treated

Again suppose we observe a population of 4 units:

i	D_i	Y_i	Y_{1i}	Y_{0i}	τ_i
1	1	3	3	0	3
2	1	1	1	1	0
3	0	0	1	0	1
4	0	1	1	1	0
$\mathbb{E}[Y_{1i} \mid D_i = 1]$			2		
$\mathbb{E}[Y_{0i} \mid D_i = 1]$				0.5	
$\mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1]$					1.5

$$\tau_{ATT} = \mathbb{E}[Y_{1i} - Y_{0i} \mid D_i = 1] = \mathbb{E}[\tau_i \mid D_i = 1] = \frac{3+0}{2} = 1.5.$$

Why does $\tau_{ATT} \neq \tau_{ATE}$?

Selection Bias

- Comparisons of observed outcomes for the treated and the untreated do not usually give the right answer:

$$
\begin{aligned}
\tilde{\tau} &= \mathbb{E}[Y_i | D_i = 1] - \mathbb{E}[Y_i | D_i = 0] \\
&= \mathbb{E}[Y_{1i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0] \\
&= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]}_{\tau_{ATT}} + \underbrace{\mathbb{E}[Y_{0i} | D_i = 1] - \mathbb{E}[Y_{0i} | D_i = 0]}_{\text{Bias}}
\end{aligned}
$$

- Bias term $\neq 0$ if selection into treatment is associated with potential outcomes

Selection Bias

- Comparisons of observed outcomes for the treated and the untreated do not usually give the right answer:

$$\begin{aligned}
\tilde{\tau} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\
&= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\
&= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]}_{\tau_{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]}_{\text{Bias}}
\end{aligned}$$

- Bias term $\neq 0$ if selection into treatment is associated with potential outcomes

Example: Church attendance and turnout

- churchgoers differ from individuals who do not attend church in many ways (e.g. civic duty)
- turnout for churchgoers would be higher than for non-churchgoers even if churchgoers never attended church ($E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] > 0$)

Selection Bias

- Comparisons of observed outcomes for the treated and the untreated do not usually give the right answer:

$$\begin{aligned}\tilde{\tau} &= \mathbb{E}[Y_i|D_i = 1] - \mathbb{E}[Y_i|D_i = 0] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]}_{\tau_{ATT}} + \underbrace{\mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0]}_{\text{Bias}}\end{aligned}$$

- Bias term $\neq 0$ if selection into treatment is associated with potential outcomes

Example: Job training program for the disadvantaged

- participants are self-selected from a subpopulation of individuals in difficult labor situations
- post-training period earnings for participants would be lower than those for nonparticipants in the absence of the program ($E[Y_{0i}|D_i = 1] - E[Y_{0i}|D_i = 0] < 0$)

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

1. Treatment is randomized by the researcher ☺☺☺
    - Laboratory experiments
    - Survey experiments
    - Field experiments

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

1. Treatment is randomized by the researcher ☺☺☺
    - Laboratory experiments
    - Survey experiments
    - Field experiments
2. Treatment is haphazard (natural experiment) ☺☺
    - Birthdays
    - Weather
    - Close elections
    - Arbitrary administrative rules

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

1. Treatment is randomized by the researcher ☺☺☺
    - Laboratory experiments
    - Survey experiments
    - Field experiments
2. Treatment is haphazard (natural experiment) ☺☺
    - Birthdays
    - Weather
    - Close elections
    - Arbitrary administrative rules
3. Treatment is "as-if" random after statistical control ☺
    - Regression
    - Matching

Essential Role of Research Design

- Causal inference requires good identification strategy
- Treatment assignment mechanism determines whether average causal effects are identifiable

1. Treatment is randomized by the researcher ☺☺☺
   - Laboratory experiments
   - Survey experiments
   - Field experiments
2. Treatment is haphazard (natural experiment) ☺☺
   - Birthdays
   - Weather
   - Close elections
   - Arbitrary administrative rules
3. Treatment is "as-if" random after statistical control ☺
   - Regression
   - Matching
4. Treatment is self-selected and no plausible control is available ☹

An Alternative Causal Model: Causal Graphs

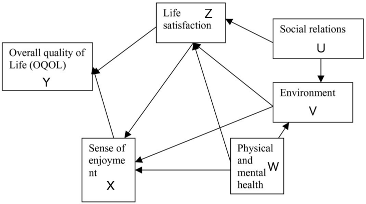
- Did social scientists not do causal inference before Rubin?

An Alternative Causal Model: Causal Graphs

- Did social scientists not do causal inference before Rubin? No!

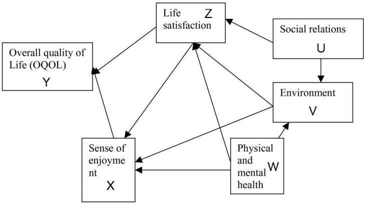
An Alternative Causal Model: Causal Graphs

- Did social scientists not do causal inference before Rubin? No!
- The old paradigm: structural equation modeling (SEM) and path analysis
  1. Postulate a causal mechanism and draw a corresponding path diagram



An Alternative Causal Model: Causal Graphs

- Did social scientists not do causal inference before Rubin? No!
- The old paradigm: structural equation modeling (SEM) and path analysis
  1. Postulate a causal mechanism and draw a corresponding path diagram



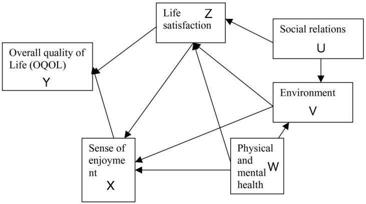
  2. Translate it into a (typically linear) system of equations:

$$
\begin{aligned}
Y &= \alpha_0 + \alpha_1 X + \alpha_2 Z + \varepsilon_\alpha \\
X &= \beta_0 + \beta_1 Z + \beta_2 W + \beta_3 V + \varepsilon_\beta \quad \cdots
\end{aligned}
$$

  3. Estimate α, β, etc. typically assuming normality and exogeneity

An Alternative Causal Model: Causal Graphs

- Did social scientists not do causal inference before Rubin? No!
- The old paradigm: structural equation modeling (SEM) and path analysis
  1. Postulate a causal mechanism and draw a corresponding path diagram



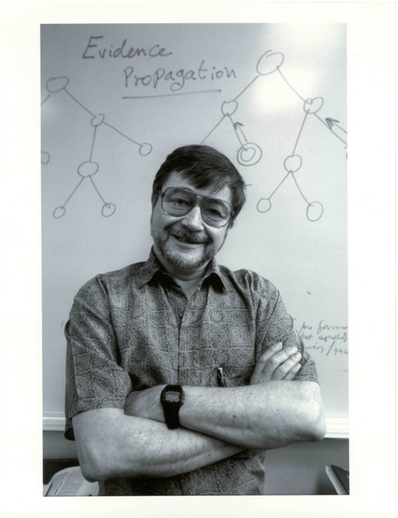
  2. Translate it into a (typically linear) system of equations:

$$
\begin{aligned}
Y &= \alpha_0 + \alpha_1 X + \alpha_2 Z + \varepsilon_\alpha \\
X &= \beta_0 + \beta_1 Z + \beta_2 W + \beta_3 V + \varepsilon_\beta \quad \cdots
\end{aligned}
$$

  3. Estimate α, β, etc. typically assuming normality and exogeneity

- Went out of fashion:
  - Strong distributional/functional form assumptions
  - No language to distinguish causation from association

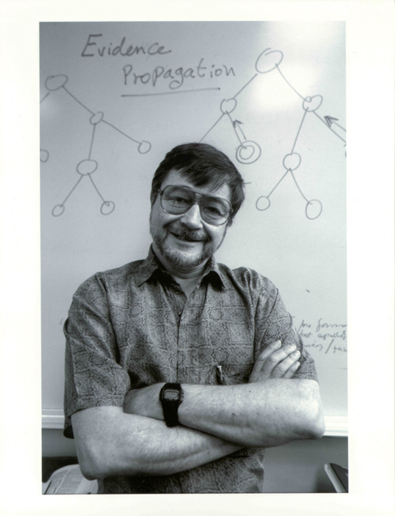
Pearl's Attack



Judea Pearl (1936–) proposed a new causal inference framework based on nonparametric structural equation modeling (NPSEM)

- Originally a computer scientist
- Previous important work on artificial intelligence
- *Causality* (2000, Cambridge UP)
- Won the Turing Award in 2011 for his causal work

Pearl's Attack

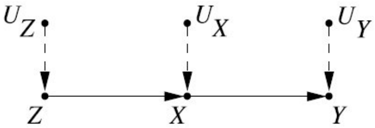


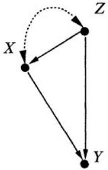
Judea Pearl (1936–) proposed a new causal inference framework based on nonparametric structural equation modeling (NPSEM)

- Originally a computer scientist
- Previous important work on artificial intelligence
- *Causality* (2000, Cambridge UP)
- Won the Turing Award in 2011 for his causal work

Pearl's framework builds on SEMs and revives it as a formal language of causality.

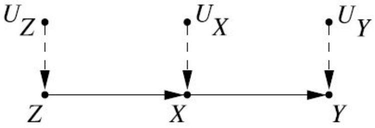
Causal Diagram

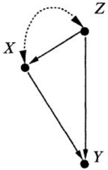




- A causal diagram is a directed acyclic graph (DAG) composed of:
  - Nodes (representing variables in the causal model)
  - Directed edges or arrows (representing *possible* causal effect)
  - Bidirected arcs (representing *possible* existence of confounding)

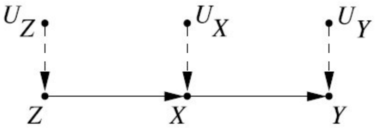
Causal Diagram

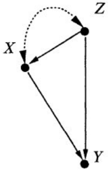




- A causal diagram is a directed acyclic graph (DAG) composed of:
  - Nodes (representing variables in the causal model)
  - Directed edges or arrows (representing *possible* causal effect)
  - Bidirected arcs (representing *possible* existence of confounding)
- Relationships involving unobserved variables are often represented by dashed/dotted lines
- Exogenous variables that is not explicitly modelled (“errors”) can be omitted from a graph

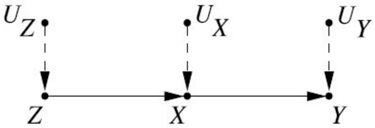
Causal Diagram

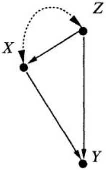




- A causal diagram is a directed acyclic graph (DAG) composed of:
  - Nodes (representing variables in the causal model)
  - Directed edges or arrows (representing *possible* causal effect)
  - Bidirected arcs (representing *possible* existence of confounding)
- Relationships involving unobserved variables are often represented by dashed/dotted lines
- Exogenous variables that is not explicitly modelled ("errors") can be omitted from a graph
- Note that it is *missing* edges that encode causal assumptions
  - Missing arrows encode exclusion restrictions
  - Missing dashed arcs encode independencies between error terms

NPSEM and Treatments

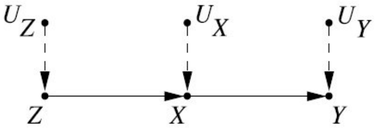


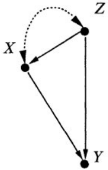


- A causal DAG has a one-to-one relationship with an NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

NPSEM and Treatments



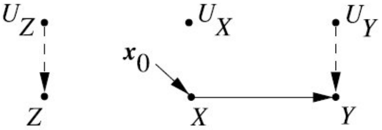


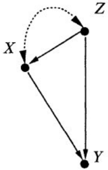
- A causal DAG has a one-to-one relationship with an NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

- These are *structural* equations (as opposed to algebraic) and represent causation — the equal signs are thus *directional* (i.e. no moving around)

NPSEM and Treatments





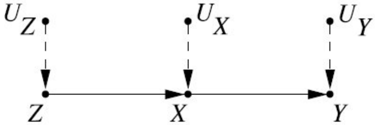
- A causal DAG has a one-to-one relationship with an NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

- These are *structural* equations (as opposed to algebraic) and represent causation — the equal signs are thus *directional* (i.e. no moving around)
- Treatments (interventions) are represented by the *do*() operator
- For example, $do(x_0)$ holds X at x_0 exogenously and thus replaces the structural equation for X with this value:

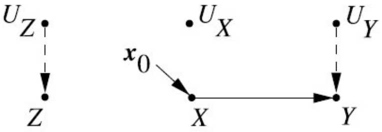
$$z = f_Z(u_Z), \quad x = x_0, \quad y = f_Y(x, u_Y)$$

Causal Effects and Identification



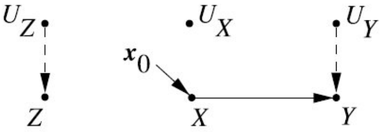
- The pre-intervention distribution: $P(x, y, z)$

Causal Effects and Identification



- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$

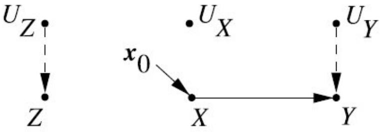
Causal Effects and Identification



- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$
- The ATE of X on Y can be defined as the average difference in Y between two intervention regimes, i.e.,

$$\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$$

Causal Effects and Identification

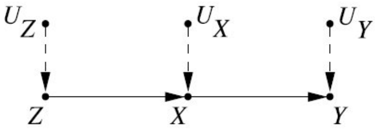


- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$
- The ATE of X on Y can be defined as the average difference in Y between two intervention regimes, i.e.,

$$\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$$

- Identification: Can $P(y \mid do(x))$ be estimated from data governed by the pre-intervention distribution $P(x, y, z)$?

Causal Effects and Identification



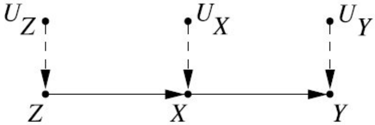
- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$
- The ATE of X on Y can be defined as the average difference in Y between two intervention regimes, i.e.,

$$\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$$

- Identification: Can $P(y \mid do(x))$ be estimated from data governed by the pre-intervention distribution $P(x, y, z)$?
- Example:

$$\mathbb{E}[Y \mid X = x_0] \quad =$$

Causal Effects and Identification



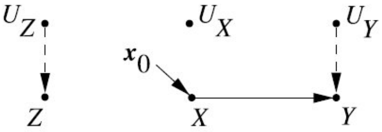
- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$
- The ATE of X on Y can be defined as the average difference in Y between two intervention regimes, i.e.,

$$\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$$

- Identification: Can $P(y \mid do(x))$ be estimated from data governed by the pre-intervention distribution $P(x, y, z)$?
- Example:

$$\mathbb{E}[Y \mid X = x_0] \quad = \quad \mathbb{E}[f_Y(x, u_Y) \mid X = x_0] \; = \; \mathbb{E}[f_Y(x_0, u_Y)] \; =$$

Causal Effects and Identification



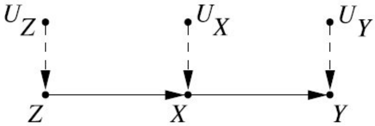
- The pre-intervention distribution: $P(x, y, z)$
- The post-intervention distribution: $P(y, z \mid do(x_0))$
- The ATE of X on Y can be defined as the average difference in Y between two intervention regimes, i.e.,

$$\mathbb{E}[Y \mid do(x_1)] - \mathbb{E}[Y \mid do(x_0)]$$

- Identification: Can $P(y \mid do(x))$ be estimated from data governed by the pre-intervention distribution $P(x, y, z)$?
- Example:

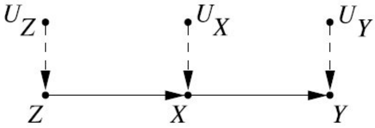
$$\mathbb{E}[Y \mid X = x_0] \quad = \quad \mathbb{E}[f_Y(x, u_Y) \mid X = x_0] \;=\; \mathbb{E}[f_Y(x_0, u_Y)] \;=\; \mathbb{E}[Y \mid do(x_0)]$$

Translation into Potential Outcomes



- A causal model represented as a graph can be translated into potential outcomes.

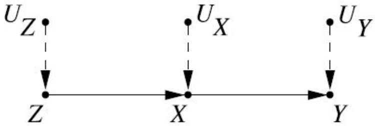
Translation into Potential Outcomes



- A causal model represented as a graph can be translated into potential outcomes.
- First, write it down in NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

Translation into Potential Outcomes



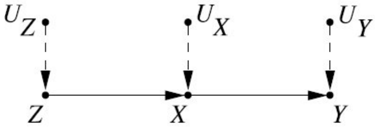
- A causal model represented as a graph can be translated into potential outcomes.
- First, write it down in NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

- This directly corresponds to potential outcomes:

$$Z_i, \quad X_{zi}, \quad Y_{xi}$$

Translation into Potential Outcomes



- A causal model represented as a graph can be translated into potential outcomes.
- First, write it down in NPSEM:

$$z = f_Z(u_Z), \quad x = f_X(z, u_X), \quad y = f_Y(x, u_Y)$$

- This directly corresponds to potential outcomes:

$$Z_i, \quad X_{zi}, \quad Y_{xi}$$

- Because of this fundamental equivalence, we will mostly work with potential outcomes, currently the standard framework in social sciences.
- Graphs are useful for expressing and visualizing a causal model in empirical research.

Aside: Modern Schools of Statistical Causal Inference

1. Potential Outcomes (Rubin, Rosenbaum, Imbens)
    - Most widely used in applied sciences, especially economics and political science
    - Causal inference as missing data problem
2. NPSEM (Pearl)
    - Uses mathematical theory of graphs
    - Borrows concepts from Bayes net and neural networks
3. Sufficient Component Causes (Rothman, VanderWeele)
    - Originates in epidemiology; "causal pies"
    - Resembles the qualitative comparative analysis (QCA)
4. Decision-theoretic causality (Dawid)
    - Does not assume existence of counterfactuals

- Causal inference as an independent field of science: *Journal of Causal Inference*

Summing Up

- Potential outcomes framework (Neyman-Rubin model) as a dominant framework for causal inference
- Causal quantities are defined by potential outcomes (counterfactuals), not by realized (observed) outcomes
- No assumption of unit homogeneity; causal effects allowed to vary unit by unit
- Observed association is neither necessary nor sufficient for causality
- Estimation of causal effects often starts with studying the treatment assignment mechanism