Nonparametric Causal Identification

UCLA SOCIOL 212B
Winter 2025

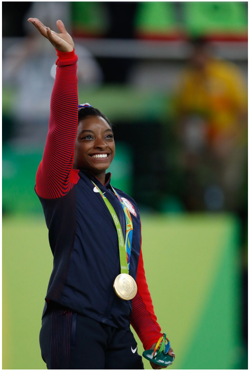
19 Feb 2025

Learning goals for today

By the end of class, you will be able to

- define causal effects
- identify average causal effects by
  - exchangeability
  - conditional exchangeability
- select a sufficient adjustment set using a Directed Acyclic Graph

Causal claims hinge on arguments, not on data



Left photo: By Fernando Frazão/Agência Brasil - `http://agenciabrasil.ebc.com.br/sites/_agenciabrasil2013/files/fotos/1035034-_mg_0802_04.08.16.jpg,CCBY3.0br,https://commons.wikimedia.org/w/index.php?curid=50548410`
Right photo: By Agencia Brasil Fotografias - EUA levam ouro na ginástica artística feminina; Brasil fica em 8 lugar, CC BY 2.0, `https://commons.wikimedia.org/w/index.php?curid=50584648`

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

	Do you win gold if you: Swing	Do not swing	Causal effect of swinging
Simone Biles	Yes (1)	?	?
Ian	?	No (0)	?

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

	Do you win gold if you:		Causal effect
	Swing	Do not swing	of swinging
Simone Biles	Yes (1)	No (0)	?
Ian	?	No (0)	?

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

	Do you win gold if you: Swing	Do not swing	Causal effect of swinging
Simone Biles	Yes (1)	No (0)	$+1$
Ian	?	No (0)	?

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

	Do you win gold if you: Swing	Do not swing	Causal effect of swinging
Simone Biles	Yes (1)	No (0)	$+1$
Ian	No (0)	No (0)	?

Causal claims hinge on arguments, not on data

1. Statistical evidence
   - Simone Biles swung on the uneven bars. She won a gold medal.
   - I did not swing on the uneven bars. I did not win a gold medal.
2. Possible causal claim
   - Swinging on the uneven bars causes a person to win a gold medal.

	Do you win gold if you: Swing	Do not swing	Causal effect of swinging
Simone Biles	Yes (1)	No (0)	$+1$
Ian	No (0)	No (0)	0

Fundamental problem of causal inference

Holland 1986

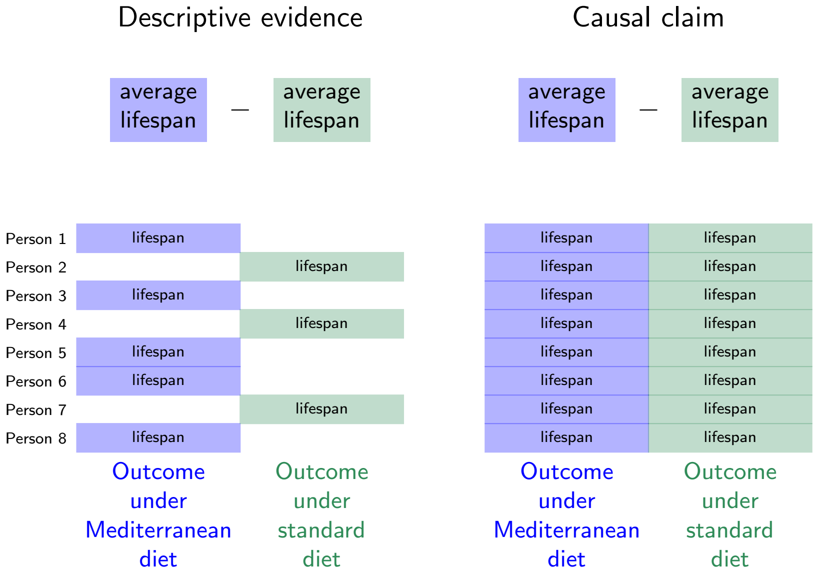
Descriptive evidence

average lifespan − average lifespan

	Outcome under Mediterranean diet	Outcome under standard diet
Person 1	lifespan	
Person 2		lifespan
Person 3	lifespan	
Person 4		lifespan
Person 5	lifespan	
Person 6	lifespan	
Person 7		lifespan
Person 8	lifespan	

Fundamental problem of causal inference

Holland 1986



Fundamental problem of causal inference

Holland 1986

Descriptive evidence

average lifespan − average lifespan

Person 1	lifespan	missing
Person 2	missing	lifespan
Person 3	lifespan	missing
Person 4	missing	lifespan
Person 5	lifespan	missing
Person 6	lifespan	missing
Person 7	missing	lifespan
Person 8	lifespan	missing
	Outcome under Mediterranean diet	Outcome under standard diet

Causal claim

average lifespan − average lifespan

lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
Outcome under Mediterranean diet	Outcome under standard diet

Fundamental problem of causal inference

Holland 1986

Descriptive evidence

average lifespan – average lifespan

Causal claim

average lifespan – average lifespan

Causal inference is a **missing data** problem

	Outcome under Mediterranean diet	Outcome under standard diet
Person 1	lifespan	missing
Person 2	missing	lifespan
Person 3	lifespan	missing
Person 4	missing	lifespan
Person 5	lifespan	missing
Person 6	lifespan	missing
Person 7	missing	lifespan
Person 8	lifespan	missing

Outcome under Mediterranean diet	Outcome under standard diet
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan
lifespan	lifespan

Mathematical notation: Potential outcomes

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Examples:

$Y_{\text{Ian}} = \texttt{survived}$ Ian survived

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Examples:

$Y_{\text{Ian}} = \texttt{survived}$	Ian survived
$A_{\text{Ian}} = \texttt{MediterraneanDiet}$	Ian ate a Mediterranean diet

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Examples:

$Y_{\text{Ian}} = \texttt{survived}$	Ian survived
$A_{\text{Ian}} = \texttt{MediterraneanDiet}$	Ian ate a Mediterranean diet
$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$	Ian would survive on a Mediterranean diet

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Examples:

$Y_{\text{Ian}} = \texttt{survived}$	Ian survived
$A_{\text{Ian}} = \texttt{MediterraneanDiet}$	Ian ate a Mediterranean diet
$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$	Ian would survive on a Mediterranean diet
$Y_{\text{Ian}}^{\text{StandardDiet}} = \texttt{died}$	Ian would die on a standard diet

Mathematical notation: Potential outcomes

Y_i	Outcome	Whether person i survived
A_i	Treatment	Whether person i ate a Mediterranean diet
Y_i^a	Potential Outcome	Outcome person i would realize if assigned to treatment value a

Examples:

$Y_{\text{Ian}} = \texttt{survived}$	Ian survived
$A_{\text{Ian}} = \texttt{MediterraneanDiet}$	Ian ate a Mediterranean diet
$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$	Ian would survive on a Mediterranean diet
$Y_{\text{Ian}}^{\text{StandardDiet}} = \texttt{died}$	Ian would die on a standard diet

Discuss.

Which potential outcome is observed?

Which is counterfactual?

The consistency assumption

The consistency assumption

$Y_i^{\text{MediterraneanDiet}}$

$Y_i^{\text{StandardDiet}}$

Potential Outcomes

The consistency assumption

$Y_i^{\text{MediterraneanDiet}}$

$Y_i^{\text{StandardDiet}}$

Potential Outcomes

Y_i

Factual Outcomes

The consistency assumption

Consistency Assumption

$$Y_i^{A_i} = Y_i$$

$Y_i^{\text{MediterraneanDiet}}$

$Y_i^{\text{StandardDiet}}$

Potential Outcomes

Y_i

Factual Outcomes

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$$

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\mathsf{Ian}}^{\mathsf{MediterraneanDiet}} = \texttt{survived}$$

The outcome for a random person is a **random variable**

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\mathsf{Ian}}^{\mathsf{MediterraneanDiet}} = \texttt{survived}$$

The outcome for a random person is a **random variable**

- Draw a random person from the population

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$$

The outcome for a random person is a **random variable**

- Draw a random person from the population
- Assign them a Mediterranean diet

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$$

The outcome for a random person is a **random variable**

- Draw a random person from the population
- Assign them a Mediterranean diet
- The outcome $Y^{\text{MediterraneanDiet}}$ is a random variable:
    - takes the value `survived` if we randomly sample some people
    - takes the value `died` if we randomly sample others

Mathematical notation: Potential outcomes are fixed

A person's potential outcome is a **fixed quantity**

$$Y_{\text{Ian}}^{\text{MediterraneanDiet}} = \texttt{survived}$$

The outcome for a random person is a **random variable**

- Draw a random person from the population
- Assign them a Mediterranean diet
- The outcome $Y^{\text{MediterraneanDiet}}$ is a random variable:
  - takes the value `survived` if we randomly sample some people
  - takes the value `died` if we randomly sample others

Check for understanding:
Does it make sense to write $\text{V}(Y_i^a)$? How about $\text{V}(Y^a)$

Notation: Expectation operator

The **expectation operator** E() denotes the population mean

$$\mathrm{E}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} Y_i^a$$

The quantity Y^a inside the expectation must be a random variable

Notation: Expectation operator

The **expectation operator** E() denotes the population mean

$$\mathrm{E}(Y^a) = \frac{1}{n}\sum_{i=1}^{n} Y_i^a$$

The quantity Y^a inside the expectation must be a random variable

A **conditional expectation** is denoted with a vertical bar

$$\mathrm{E}(Y \mid A = a) = \frac{1}{n_a}\sum_{i:A_i=a} Y_i$$

Practice: How would you say this in English?

We might wonder how a person's earnings relate to whether they hold a college degree

1. $\text{E}(\text{Earnings} \mid \text{Degree} = \texttt{TRUE}) > \text{E}(\text{Earnings} \mid \text{Degree} = \texttt{FALSE})$

2. $\text{E}(\text{Earnings}^{\text{Degree}=\texttt{TRUE}}) > \text{E}(\text{Earnings}^{\text{Degree}=\texttt{FALSE}})$

Practice: How would you say this in English?

We might wonder how a person's earnings relate to whether they hold a college degree

1. $\text{E}(\text{Earnings} \mid \text{Degree} = \texttt{TRUE}) > \text{E}(\text{Earnings} \mid \text{Degree} = \texttt{FALSE})$
   - Average earnings are higher among those with college degrees

2. $\text{E}(\text{Earnings}^{\text{Degree}=\texttt{TRUE}}) > \text{E}(\text{Earnings}^{\text{Degree}=\texttt{FALSE}})$

Practice: How would you say this in English?

We might wonder how a person's earnings relate to whether they hold a college degree

1. $\text{E}(\text{Earnings} \mid \text{Degree} = \texttt{TRUE}) > \text{E}(\text{Earnings} \mid \text{Degree} = \texttt{FALSE})$
   - Average earnings are higher among those with college degrees
2. $\text{E}(\text{Earnings}^{\text{Degree}=\texttt{TRUE}}) > \text{E}(\text{Earnings}^{\text{Degree}=\texttt{FALSE}})$
   - On average, a degree causes higher earnings

Practice: How would you write this in math?

1. On average, students who do the homework learn more than those who don't

2. On average, doing the homework causes more learning

Practice: How would you write this in math?

1. On average, students who do the homework learn more than those who don't

$$\text{E}(\text{Learning} \mid \text{HW} = \texttt{TRUE}) > \text{E}(\text{Learning} \mid \text{HW} = \texttt{FALSE})$$

2. On average, doing the homework causes more learning

Practice: How would you write this in math?

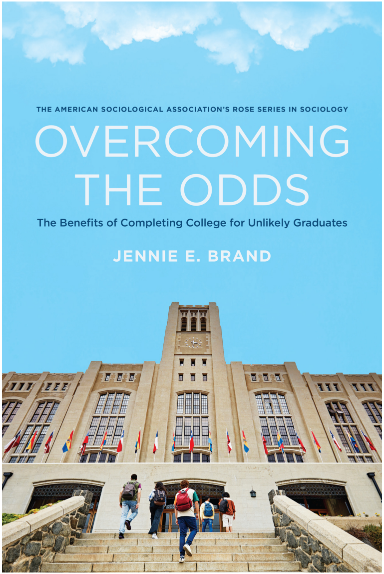
1. On average, students who do the homework learn more than those who don't

$$\text{E}(\text{Learning} \mid \text{HW} = \texttt{TRUE}) > \text{E}(\text{Learning} \mid \text{HW} = \texttt{FALSE})$$

2. On average, doing the homework causes more learning

$$\text{E}(\text{Learning}^{\text{HW}=\texttt{TRUE}}) > \text{E}(\text{Learning}^{\text{HW}=\texttt{FALSE}})$$

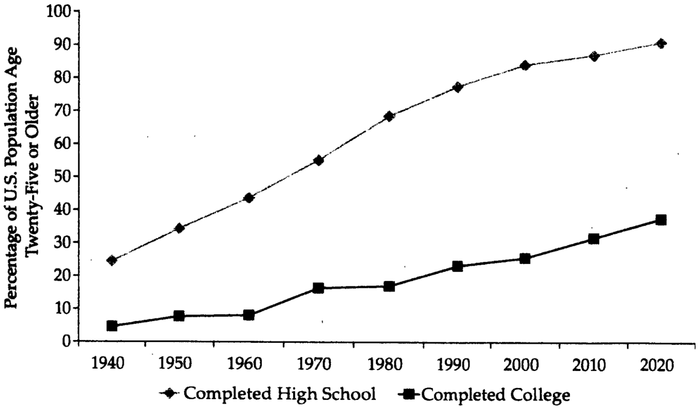
An example about inequality



Americans' education in 1900 (Brand 2023 p. 6)

- 6% graduated from high school
- 3% graduated from college

Figure 1.1 High School and Four-Year College Completion Rates, 1940–2020



Source: U.S. Census Bureau, March Current Population Survey and Annual Social and Economic Supplement to the Current Population Survey.

(Brand 2023)

We would like to know whether **college pays off**:
does it have positive effects on desired outcomes?

Mathematical notation for two types of claims

Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings

Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings

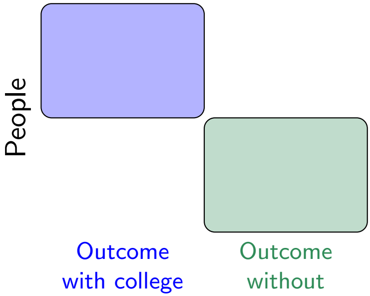
Two sets of people
Two treatments

Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings

Two sets of people
Two treatments



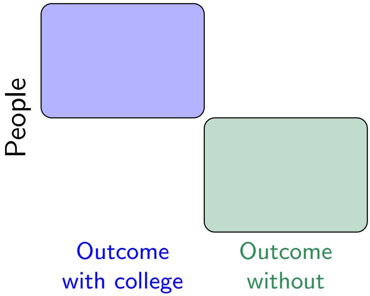
Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings

Two sets of people
Two treatments

Same people
Two treatments

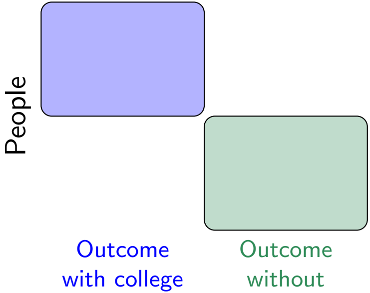


Mathematical notation for two types of claims

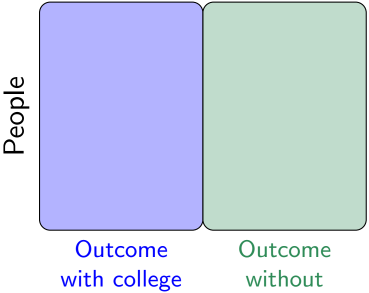
People with
college degrees
earn more

A college degree
causes
higher earnings

Two sets of people
Two treatments



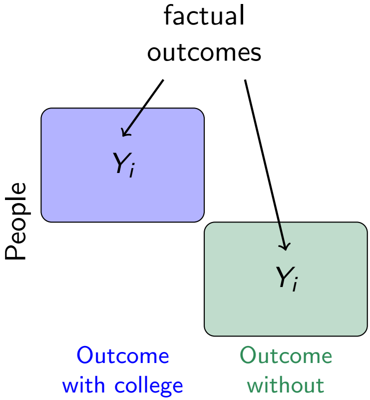
Same people
Two treatments

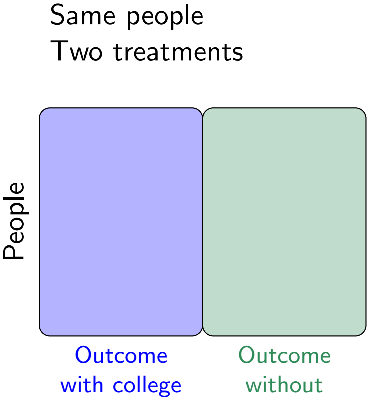


Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings

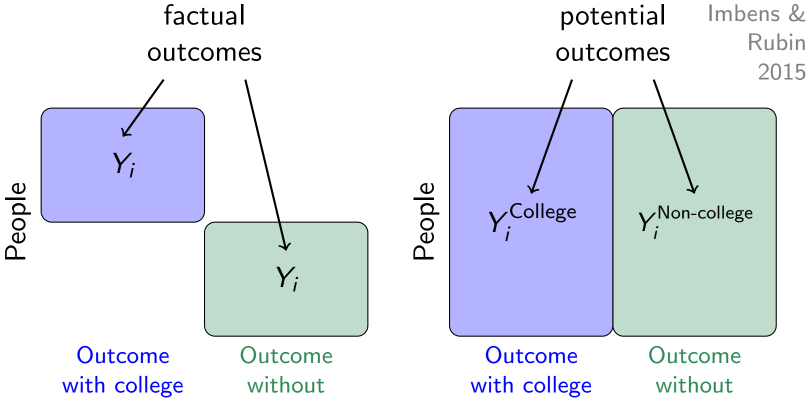




Mathematical notation for two types of claims

People with
college degrees
earn more

A college degree
causes
higher earnings



The fundamental problem of causal inference

The data

Each Row is a Person

Outcome under treatment	Outcome under control
$Y_{\text{Nick}}^{\text{College}}$	
$Y_{\text{William}}^{\text{College}}$	
	$Y_{\text{Rich}}^{\text{Non-college}}$
$Y_{\text{Diego}}^{\text{College}}$	
	$Y_{\text{Javier}}^{\text{Non-college}}$
	$Y_{\text{Jesús}}^{\text{Non-college}}$

Holland 1986

The fundamental problem of causal inference

Each Row is a Person

The data

Outcome under treatment	Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	
$Y^{\text{College}}_{\text{William}}$	
	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	
	$Y^{\text{Non-college}}_{\text{Javier}}$
	$Y^{\text{Non-college}}_{\text{Jesús}}$

The claim

Outcome under treatment		Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$

Holland 1986

The fundamental problem of causal inference

Each Row is a Person

The data

Outcome under treatment	Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	?
$Y^{\text{College}}_{\text{William}}$	?
?	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	?
?	$Y^{\text{Non-college}}_{\text{Javier}}$
?	$Y^{\text{Non-college}}_{\text{Jesús}}$

The claim

Outcome under treatment		Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$

Counterfactuals are **not observed**

Holland 1986

Preview: Solving the problem by assumptions

Each Row is a Person

The data

Outcome under treatment	Outcome under control
$Y_{\text{Nick}}^{\text{College}}$	?
$Y_{\text{William}}^{\text{College}}$	?
?	$Y_{\text{Rich}}^{\text{Non-college}}$
$Y_{\text{Diego}}^{\text{College}}$	?
?	$Y_{\text{Javier}}^{\text{Non-college}}$
?	$Y_{\text{Jesús}}^{\text{Non-college}}$

The claim

Outcome under treatment		Outcome under control
$Y_{\text{Nick}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{Nick}}^{\text{Non-college}}$
$Y_{\text{William}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{William}}^{\text{Non-college}}$
$Y_{\text{Rich}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{Rich}}^{\text{Non-college}}$
$Y_{\text{Diego}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{Diego}}^{\text{Non-college}}$
$Y_{\text{Javier}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{Javier}}^{\text{Non-college}}$
$Y_{\text{Jesús}}^{\text{College}}$	$\longleftrightarrow$	$Y_{\text{Jesús}}^{\text{Non-college}}$

Preview: Solving the problem by assumptions

The data

Each Row is a Person

Outcome under treatment	Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	?
$Y^{\text{College}}_{\text{William}}$	?
?	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	?
?	$Y^{\text{Non-college}}_{\text{Javier}}$
?	$Y^{\text{Non-college}}_{\text{Jesús}}$

The claim

Outcome under treatment		Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$

Preview: Solving the problem by assumptions

The data

Each Row is a Person

$Y^{\text{College}}_{\text{Nick}}$	?
$Y^{\text{College}}_{\text{William}}$	?
?	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	?
?	$Y^{\text{Non-college}}_{\text{Javier}}$
?	$Y^{\text{Non-college}}_{\text{Jesús}}$
Outcome under treatment	Outcome under control

The claim

$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$
Outcome under treatment		Outcome under control

Preview: Solving the problem by assumptions

The data

Each Row is a Person

Outcome under treatment	Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	?
$Y^{\text{College}}_{\text{William}}$	?
?	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	?
?	$Y^{\text{Non-college}}_{\text{Javier}}$
?	$Y^{\text{Non-college}}_{\text{Jesús}}$

The claim

Outcome under treatment		Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$

Preview: Solving the problem by assumptions

The data

Each Row is a Person

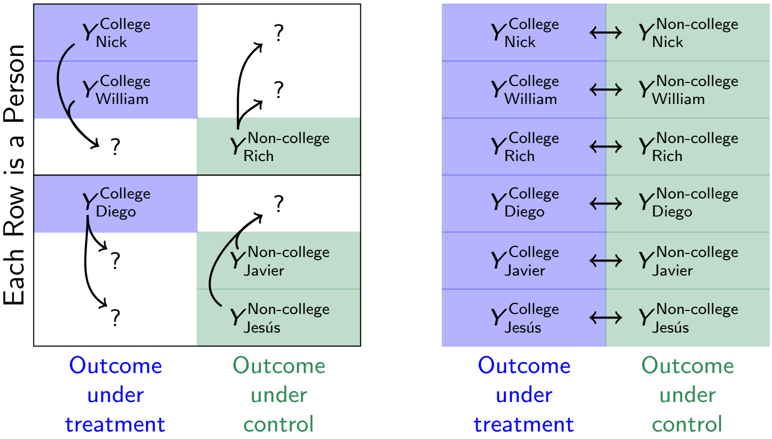
Outcome under treatment	Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	?
$Y^{\text{College}}_{\text{William}}$	?
?	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	?
?	$Y^{\text{Non-college}}_{\text{Javier}}$
?	$Y^{\text{Non-college}}_{\text{Jesús}}$

The claim

Outcome under treatment		Outcome under control
$Y^{\text{College}}_{\text{Nick}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Nick}}$
$Y^{\text{College}}_{\text{William}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{William}}$
$Y^{\text{College}}_{\text{Rich}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Rich}}$
$Y^{\text{College}}_{\text{Diego}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Diego}}$
$Y^{\text{College}}_{\text{Javier}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Javier}}$
$Y^{\text{College}}_{\text{Jesús}}$	$\longleftrightarrow$	$Y^{\text{Non-college}}_{\text{Jesús}}$

Preview: Solving the problem by assumptions

The data **The claim**



Quick review

Quick review

1. causal claims involve potential outcomes: Y^a
2. not all potential outcomes are observed
3. causal inference is a missing data problem

Exchangeable sampling from a population

Exchangeable sampling from a population

Population Outcomes

$Y_{\textbf{Maria}}$

$Y_{\textbf{William}}$

$Y_{\textbf{Rich}}$

$Y_{\textbf{Sarah}}$

$Y_{\textbf{Alondra}}$

$Y_{\textbf{Jesús}}$

Exchangeable sampling from a population

Population Outcomes	Randomized Sampling
$Y_{\textbf{Maria}}$	$S_{\textbf{Maria}} = 1$
$Y_{\textbf{William}}$	$S_{\textbf{William}} = 0$
$Y_{\textbf{Rich}}$	$S_{\textbf{Rich}} = 0$
$Y_{\textbf{Sarah}}$	$S_{\textbf{Sarah}} = 1$
$Y_{\textbf{Alondra}}$	$S_{\textbf{Alondra}} = 0$
$Y_{\textbf{Jesús}}$	$S_{\textbf{Jesús}} = 1$

Exchangeable sampling from a population

Population Outcomes	**Randomized Sampling**	**Sampled Outcomes**
Y_{Maria}	$S_{\text{Maria}} = 1$	Y_{Maria}
Y_{William}	$S_{\text{William}} = 0$	
Y_{Rich}	$S_{\text{Rich}} = 0$	
Y_{Sarah}	$S_{\text{Sarah}} = 1$	Y_{Sarah}
Y_{Alondra}	$S_{\text{Alondra}} = 0$	
$Y_{\text{Jesús}}$	$S_{\text{Jesús}} = 1$	$Y_{\text{Jesús}}$

Exchangeable sampling from a population

Population Outcomes	Randomized Sampling	Sampled Outcomes
Y_{Maria}	$S_{\text{Maria}} = 1$	Y_{Maria}
Y_{William}	$S_{\text{William}} = 0$	
Y_{Rich}	$S_{\text{Rich}} = 0$	
Y_{Sarah}	$S_{\text{Sarah}} = 1$	Y_{Sarah}
Y_{Alondra}	$S_{\text{Alondra}} = 0$	
$Y_{\text{Jesús}}$	$S_{\text{Jesús}} = 1$	$Y_{\text{Jesús}}$

Estimator:
Estimate the population mean by the sample mean

Key assumption:
Sampled and unsampled units are **exchangeable** due to random sampling

$Y \perp\!\!\!\perp S$

Now suppose our population all participate
in a randomized experiment with
treatment ($A = 1$) and control ($A = 0$) conditions

Exchangeable treatment assignment

Population Potential Outcomes

Y^1_{Maria}

Y^1_{William}

Y^1_{Rich}

Y^1_{Sarah}

Y^1_{Alondra}

$Y^1_{\text{Jesús}}$

Exchangeable treatment assignment

Population Potential Outcomes	**Randomized Treatment**
Y^1_{Maria}	$A_{\text{Maria}} = 1$
Y^1_{William}	$A_{\text{William}} = 0$
Y^1_{Rich}	$A_{\text{Rich}} = 0$
Y^1_{Sarah}	$A_{\text{Sarah}} = 1$
Y^1_{Alondra}	$A_{\text{Alondra}} = 0$
$Y^1_{\text{Jesús}}$	$A_{\text{Jesús}} = 1$

Exchangeable treatment assignment

Population Potential Outcomes	Randomized Treatment	Observed Outcomes
$Y^1_{\textbf{Maria}}$	$A_{\textbf{Maria}} = 1$	$Y^1_{\textbf{Maria}}$
$Y^1_{\textbf{William}}$	$A_{\textbf{William}} = 0$	
$Y^1_{\textbf{Rich}}$	$A_{\textbf{Rich}} = 0$	
$Y^1_{\textbf{Sarah}}$	$A_{\textbf{Sarah}} = 1$	$Y^1_{\textbf{Sarah}}$
$Y^1_{\textbf{Alondra}}$	$A_{\textbf{Alondra}} = 0$	
$Y^1_{\textbf{Jesús}}$	$A_{\textbf{Jesús}} = 1$	$Y^1_{\textbf{Jesús}}$

Exchangeable treatment assignment

Population Potential Outcomes	Randomized Treatment	Observed Outcomes
Y^1_{Maria}	$A_{\text{Maria}} = 1$	Y^1_{Maria}
Y^1_{William}	$A_{\text{William}} = 0$	
Y^1_{Rich}	$A_{\text{Rich}} = 0$	
Y^1_{Sarah}	$A_{\text{Sarah}} = 1$	Y^1_{Sarah}
Y^1_{Alondra}	$A_{\text{Alondra}} = 0$	
$Y^1_{\text{Jesús}}$	$A_{\text{Jesús}} = 1$	$Y^1_{\text{Jesús}}$

Estimator:
Estimate the population mean $\text{E}(Y^1)$ by the untreated sample mean

Key assumption:
Treated and untreated units are **exchangeable** due to random treatment assignment

$$Y^1 \perp\!\!\!\perp A$$

Exchangeable treatment assignment

Population Potential Outcomes	Randomized Treatment	Observed Outcomes
Y^0_{Maria}	$A_{\text{Maria}} = 1$	
Y^0_{William}	$A_{\text{William}} = 0$	Y^0_{William}
Y^0_{Rich}	$A_{\text{Rich}} = 0$	Y^0_{Rich}
Y^0_{Sarah}	$A_{\text{Sarah}} = 1$	
Y^0_{Alondra}	$A_{\text{Alondra}} = 0$	Y^0_{Alondra}
$Y^0_{\text{Jesús}}$	$A_{\text{Jesús}} = 1$	

Estimator:
Estimate the population mean $\text{E}(Y^0)$ by the untreated sample mean

Key assumption:
Treated and untreated units are **exchangeable** due to random treatment assignment

$$Y^0 \perp\!\!\!\perp A$$

Exchangeable treatment assignment

Population Potential Outcomes		Randomized Treatment	Observed Outcomes	
Y^1_{Maria}	Y^0_{Maria}	$A_{\text{Maria}} = 1$	Y^1_{Maria}	
Y^1_{William}	Y^0_{William}	$A_{\text{William}} = 0$		Y^0_{William}
Y^1_{Rich}	Y^0_{Rich}	$A_{\text{Rich}} = 0$		Y^0_{Rich}
Y^1_{Sarah}	Y^0_{Sarah}	$A_{\text{Sarah}} = 1$	Y^1_{Sarah}	
Y^1_{Alondra}	Y^0_{Alondra}	$A_{\text{Alondra}} = 0$		Y^0_{Alondra}
$Y^1_{\text{Jesús}}$	$Y^0_{\text{Jesús}}$	$A_{\text{Jesús}} = 1$	$Y^1_{\text{Jesús}}$	

Exchangeable treatment assignment

Causal Estimand:
(expected outcome if assigned to treatment)
− (expected outcome if assigned to control)

$$\mathrm{E}\left(Y^1\right) - \mathrm{E}\left(Y^0\right)$$

Exchangeability Assumption:
Potential outcomes are independent of treatment assignment

$$\{Y^0, Y^1\} \perp\!\!\!\perp A$$

Empirical Estimand:
(expected outcome among the treated)
− (expected outcome among the untreated)

$$\mathrm{E}(Y \mid A = 1) - \mathrm{E}(Y \mid A = 0)$$

Exchangeable treatment assignment: Proof

$$
\begin{aligned}
&\mathrm{E}\left(Y^1\right) - \mathrm{E}\left(Y^0\right) \\
&= \mathrm{E}\left(Y^1 \mid A = 1\right) - \mathrm{E}\left(Y^0 \mid A = 0\right) \\
&= \mathrm{E}(Y \mid A = 1) - \mathrm{E}(Y \mid A = 0)
\end{aligned}
$$

Exchangeable treatment assignment: Proof

$$\begin{aligned}
&\mathsf{E}\left(Y^1\right) - \mathsf{E}\left(Y^0\right) \\
&= \mathsf{E}\left(Y^1 \mid A = 1\right) - \mathsf{E}\left(Y^0 \mid A = 0\right) \\
&= \mathsf{E}\left(Y \mid A = 1\right) - \mathsf{E}\left(Y \mid A = 0\right) && \text{by consistency}
\end{aligned}$$

Exchangeable treatment assignment: Proof

$$
\begin{aligned}
&\mathsf{E}\left(Y^1\right) - \mathsf{E}\left(Y^0\right) \\
&= \mathsf{E}\left(Y^1 \mid A = 1\right) - \mathsf{E}\left(Y^0 \mid A = 0\right) && \text{by exchangeability} \\
&= \mathsf{E}\left(Y \mid A = 1\right) - \mathsf{E}\left(Y \mid A = 0\right) && \text{by consistency}
\end{aligned}
$$

Exchangeable treatment assignment: Proof

$$
\begin{aligned}
& \mathrm{E}\left(Y^1\right) - \mathrm{E}\left(Y^0\right) \\
& = \mathrm{E}\left(Y^1 \mid A = 1\right) - \mathrm{E}\left(Y^0 \mid A = 0\right) \quad \text{by exchangeability} \\
& = \mathrm{E}\left(Y \mid A = 1\right) - \mathrm{E}\left(Y \mid A = 0\right) \qquad\qquad \text{by consistency}
\end{aligned}
$$

This is an example of **causal identification**:
using assumptions to arrive at an empirical quantity
(involving only factual random variables, no potential outcomes)
that equals our causal estimand if the assumptions hold

The causal estimand $\mathrm{E}(Y^1) - \mathrm{E}(Y^0)$ is **identified** by the empirical estimand $\mathrm{E}(Y \mid A = 1) - \mathrm{E}(Y \mid A = 0)$

Potential outcome exercise: Covid vaccines

Thanks to Sam Wang at Cornell for slides on this vaccine example!

Potential outcome exercise: Covid vaccines

Thanks to Sam Wang at Cornell for slides on this vaccine example!

Suppose we know the following pieces of information:

- Martha was vaccinated against Covid.
  Martha tested negative 6 months later.
- Ezra was not vaccinated against Covid.
  Ezra tested positive 6 months later.

Potential outcome exercise: Covid vaccines

Thanks to Sam Wang at Cornell for slides on this vaccine example!

Suppose we know the following pieces of information:

- Martha was vaccinated against Covid.
  Martha tested negative 6 months later.
- Ezra was not vaccinated against Covid.
  Ezra tested positive 6 months later.

Which cells have known values? What are the values?

	A_i	Y_i	$Y_i^{\text{Vaccinated}}$	$Y_i^{\text{Unvaccinated}}$
Martha				
Ezra				

Experiments vs observational studies

Suppose we gathered data by surveying individuals in Fall of 2021

- Vaccinated for covid ($A_i = 1$) or not ($A_i = 0$)
- Tested positive for Covid in 2021: yes ($Y_i = 1$) or no ($Y_i = 0$)

Experiments vs observational studies

We observe evidence

- Of the individuals who are **vaccinated** ($A_i = 1$), 50% had a positive Covid test ($Y_i = 1$)
- Of the individuals who are **not vaccinated** ($A_i = 0$), 70% had a positive Covid test ($Y_i = 1$)

Experiments vs observational studies

We observe evidence

- Of the individuals who are **vaccinated** ($A_i = 1$), 50% had a positive Covid test ($Y_i = 1$)
- Of the individuals who are **not vaccinated** ($A_i = 0$), 70% had a positive Covid test ($Y_i = 1$)

How might a vaccine skeptic explain the data?

Experiments vs observational studies

Experiment designed by Pfizer **randomly assign** each individual (43,000 total) into two groups[1]:

- Two doses of BNT162b2 vaccine 21 days apart
- Two doses of placebo 21 days apart
- Whether a positive Covid test was recorded between 7 days and 14 weeks after the injection

[1] Polack et. al. NEJM 2020

Experiments vs observational studies

Experiment designed by Pfizer **randomly assign** each individual (43,000 total) into two groups[1]:

- Two doses of BNT162b2 vaccine 21 days apart
- Two doses of placebo 21 days apart
- Whether a positive Covid test was recorded between 7 days and 14 weeks after the injection

- Of the individuals who were given the vaccine ($A_i = 1$), 0.04% had a positive Covid test ($Y_i = 1$)
- Of the individuals who were given the placebo ($A_i = 0$), 0.9% had a positive Covid test ($Y_i = 1$)
- Individuals who received the placebo were ≈ 20 times more likely to get Covid

[1] Polack et. al. NEJM 2020

Experiments vs observational studies

Experiment designed by Pfizer **randomly assign** each individual (43,000 total) into two groups[1]:

- Two doses of BNT162b2 vaccine 21 days apart
- Two doses of placebo 21 days apart
- Whether a positive Covid test was recorded between 7 days and 14 weeks after the injection

- Of the individuals who were given the vaccine ($A_i = 1$), 0.04% had a positive Covid test ($Y_i = 1$)
- Of the individuals who were given the placebo ($A_i = 0$), 0.9% had a positive Covid test ($Y_i = 1$)
- Individuals who received the placebo were ≈ 20 times more likely to get Covid

Do the skeptic's objections still hold?

[1] Polack et. al. NEJM 2020

Why experiments "work": Exchangeability

Table 1. Demographic Characteristics of the Participants in the Main Safety Population.*

Characteristic	BNT162b2 (N=18,860)	Placebo (N=18,846)	Total (N=37,706)
Sex — no. (%)			
Male	9,639 (51.1)	9,436 (50.1)	19,075 (50.6)
Female	9,221 (48.9)	9,410 (49.9)	18,631 (49.4)
Race or ethnic group — no. (%)†			
White	15,636 (82.9)	15,630 (82.9)	31,266 (82.9)
Black or African American	1,729 (9.2)	1,763 (9.4)	3,492 (9.3)
Asian	801 (4.2)	807 (4.3)	1,608 (4.3)
Native American or Alaska Native	102 (0.5)	99 (0.5)	201 (0.5)
Native Hawaiian or other Pacific Islander	50 (0.3)	26 (0.1)	76 (0.2)
Multiracial	449 (2.4)	406 (2.2)	855 (2.3)
Not reported	93 (0.5)	115 (0.6)	208 (0.6)
Hispanic or Latinx	5,266 (27.9)	5,277 (28.0)	10,543 (28.0)
Country — no. (%)			
Argentina	2,883 (15.3)	2,881 (15.3)	5,764 (15.3)
Brazil	1,145 (6.1)	1,139 (6.0)	2,284 (6.1)
South Africa	372 (2.0)	372 (2.0)	744 (2.0)
United States	14,460 (76.7)	14,454 (76.7)	28,914 (76.7)
Age group — no. (%)			
16–55 yr	10,889 (57.7)	10,896 (57.8)	21,785 (57.8)
>55 yr	7,971 (42.3)	7,950 (42.2)	15,921 (42.2)
Age at vaccination — yr			
Median	52.0	52.0	52.0
Range	16–89	16–91	16–91
Body-mass index‡			
≥30.0: obese	6,556 (34.8)	6,662 (35.3)	13,218 (35.1)

* Percentages may not total 100 because of rounding.
† Race or ethnic group was reported by the participants.
‡ The body-mass index is the weight in kilograms divided by the square of the height in meters.

Why experiments "work": Exchangeability

In random experiments, the distribution of **potential outcomes** for those who are treated and those who are not treated (control group) are identically distributed!

$$\{Y^1, Y^0\} \perp\!\!\!\perp A$$

This is **exchangeability**

Question. Does exchangeability imply $Y \perp\!\!\!\perp A$?

Why experiments "work": Exchangeability

Exchangeability is about **potential** rather than **observed** outcomes

$$\{Y^0, Y^1\} \perp\!\!\!\perp A \qquad \text{rather than} \qquad Y \not\!\perp\!\!\!\perp A$$

Why experiments "work": Exchangeability

Exchangeability is about **potential** rather than **observed** outcomes

$$\{Y^0, Y^1\} \perp\!\!\!\perp A \qquad \text{rather than} \qquad Y \not\!\perp\!\!\!\perp A$$

- Potential outcomes are independent of treatment $\{Y^0, Y^1\} \perp\!\!\!\perp A$
  - Example: Risk of covid under no vaccine (Y^0) is the same for those with and without a vaccine

Why experiments "work": Exchangeability

Exchangeability is about **potential** rather than **observed** outcomes

$$\{Y^0, Y^1\} \perp\!\!\!\perp A \qquad \text{rather than} \qquad Y \not\!\perp\!\!\!\perp A$$

- Potential outcomes are independent of treatment $\{Y^0, Y^1\} \perp\!\!\!\perp A$
  - Example: Risk of covid under no vaccine (Y^0) is the same for those with and without a vaccine
- Observed outcomes are not independent of treatment $Y \not\!\perp\!\!\!\perp A$
  - Example: Risk of covid is lower for those with the vaccine
  - Why? Because for them $Y = Y^1$. For others, $Y = Y^0$.
  - If A affects Y, then $Y \not\!\perp\!\!\!\perp A$

Why experiments "work": Exchangeability

Exchangeability is about **potential** rather than **observed** outcomes

$$\{Y^0, Y^1\} \perp\!\!\!\perp A \qquad \text{rather than} \qquad Y \not\perp\!\!\!\perp A$$

- Potential outcomes are independent of treatment $\{Y^0, Y^1\} \perp\!\!\!\perp A$
  - Example: Risk of covid under no vaccine (Y^0) is the same for those with and without a vaccine
- Observed outcomes are not independent of treatment $Y \not\perp\!\!\!\perp A$
  - Example: Risk of covid is lower for those with the vaccine
  - Why? Because for them $Y = Y^1$. For others, $Y = Y^0$.
  - If A affects Y, then $Y \not\perp\!\!\!\perp A$

Under exchangeability, the only reason $Y \not\perp\!\!\!\perp A$ is if A causes Y.

Review of exchangeability

Exchangeable sampling from a population

Population Outcomes	Randomized Sampling	Sampled Outcomes
Y_{Maria}	$S_{\text{Maria}} = 1$	Y_{Maria}
Y_{William}	$S_{\text{William}} = 0$	
Y_{Rich}	$S_{\text{Rich}} = 0$	
Y_{Sarah}	$S_{\text{Sarah}} = 1$	Y_{Sarah}
Y_{Alondra}	$S_{\text{Alondra}} = 0$	
$Y_{\text{Jesús}}$	$S_{\text{Jesús}} = 1$	$Y_{\text{Jesús}}$

Estimator:
Estimate the population mean by the sample mean

Key assumption:
Sampled and unsampled units are **exchangeable** due to random sampling

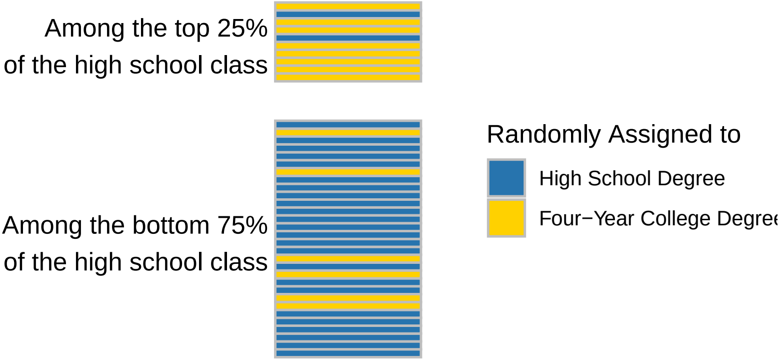
$Y \perp\!\!\!\perp S$

Exchangeable treatment assignment

Population Potential Outcomes		Randomized Treatment	Observed Outcomes	
$Y^1_{\textbf{Maria}}$	$Y^0_{\textbf{Maria}}$	$A_{\textbf{Maria}} = 1$	$Y^1_{\textbf{Maria}}$	
$Y^1_{\textbf{William}}$	$Y^0_{\textbf{William}}$	$A_{\textbf{William}} = 0$		$Y^0_{\textbf{William}}$
$Y^1_{\textbf{Rich}}$	$Y^0_{\textbf{Rich}}$	$A_{\textbf{Rich}} = 0$		$Y^0_{\textbf{Rich}}$
$Y^1_{\textbf{Sarah}}$	$Y^0_{\textbf{Sarah}}$	$A_{\textbf{Sarah}} = 1$	$Y^1_{\textbf{Sarah}}$	
$Y^1_{\textbf{Alondra}}$	$Y^0_{\textbf{Alondra}}$	$A_{\textbf{Alondra}} = 0$		$Y^0_{\textbf{Alondra}}$
$Y^1_{\textbf{Jesús}}$	$Y^0_{\textbf{Jesús}}$	$A_{\textbf{Jesús}} = 1$	$Y^1_{\textbf{Jesús}}$	

A **conditionally** randomized experiment

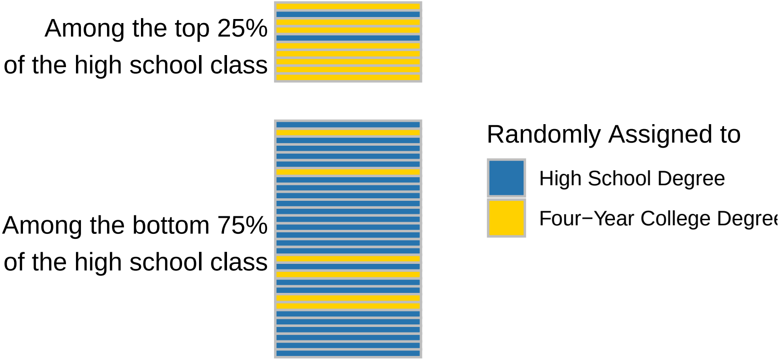
A hypothetical experiment:
Conditional randomization



Outcome: Employed at age 40

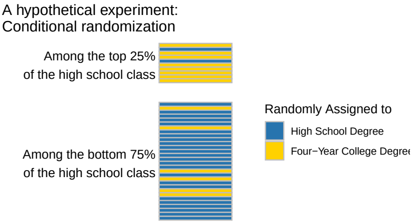
Does exchangeability hold? How would you analyze?

A hypothetical experiment:
Conditional randomization



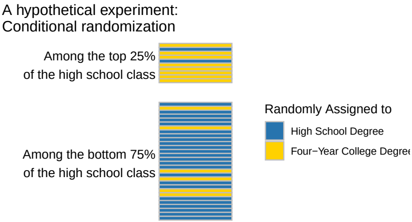
Outcome: Employed at age 40

Conditional randomization: Exchangeability does not hold



Conditional randomization: Exchangeability does not hold

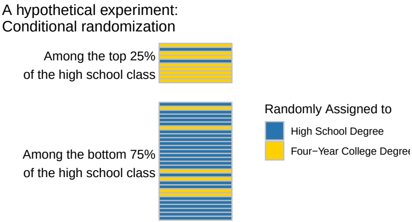
Treated units are more likely to have done well in high school



Conditional randomization: Exchangeability does not hold

Treated units are more likely to have done well in high school

Those who do well in high school are more likely to be employed at age 40 even without college

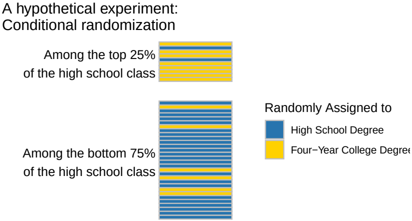


Conditional randomization: Exchangeability does not hold

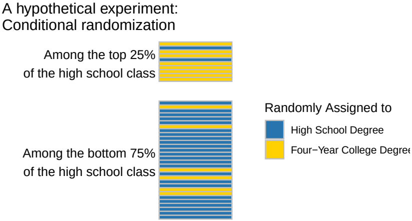
Treated units are more likely to have done well in high school

Those who do well in high school are more likely to be employed at age 40 even without college

$$\{Y^1, Y^0\} \not\perp A$$

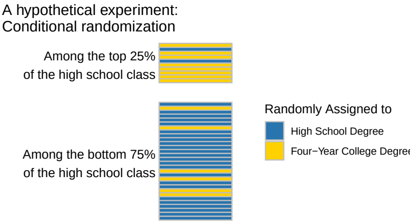


Conditional randomization: Analyze within subgroups



Conditional randomization: Analyze within subgroups

Among top 25%, simple random experiment.
Among bottom 75%, simple random experiment.

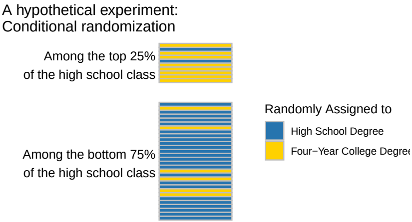


Conditional randomization: Analyze within subgroups

Among top 25%, simple random experiment.
Among bottom 75%, simple random experiment.

Conditional exchangeability:

$$\underbrace{\{Y^1, Y^0\}}_{\text{Potential Outcomes}} \quad \underbrace{\perp\!\!\!\perp}_{\text{Are Independent of}} \quad \underbrace{A}_{\text{Treatment}} \quad \underbrace{|}_{\text{Within Subgroups}} \quad \underbrace{X}_{\text{of X}}$$



Conditional average treatment effects

We get two estimates. Average effect of college on employment

- among those in the top 25% of their high school class
- among those in the bottom 75% of their high school class

These are **conditional average treatment effects**

$$\underbrace{\tau(x)}_{\substack{\text{Conditional}\\\text{Average}\\\text{Treatment}\\\text{Effect}\\\text{(CATE)}}} = \underbrace{\text{E}}_{\substack{\text{Expected}\\\text{value of}}} \left(\underbrace{Y^1 - Y^0}_{\text{treatment effect}} \underbrace{\mid}_{\substack{\text{within the}\\\text{subgroup}\\\text{for whom}}} \underbrace{\vec{X} = \vec{x}}_{\substack{\text{the predictors } \vec{X}\\\text{take the value } \vec{x}}} \right)$$

Effect heterogeneity: CATEs differ across subgroups

Why might the effect of college on future employment

- be larger for those from the top 25% of the high school class?
- be larger for those from the bottom 75% of the high school class?

Effect heterogeneity and policy

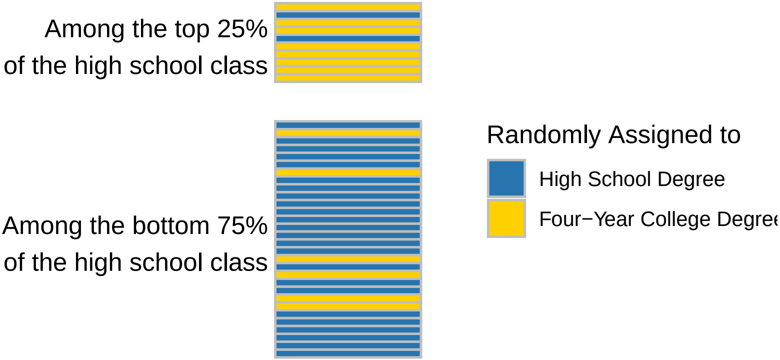
Suppose we study (college $\rightarrow$ employment) in two subgroups

- Advantaged subgroup
  - Both parents finished college
  - Top quartile of family income at age 14
  - Took college prep courses
- Disadvantaged subgroup
  - Neither parent finished college
  - Bottom quartile of family income at age 14
  - Took college prep courses

Discuss:

1. Whose CATE would be larger?
2. How might the difference inform policy?

A hypothetical experiment: Conditional randomization



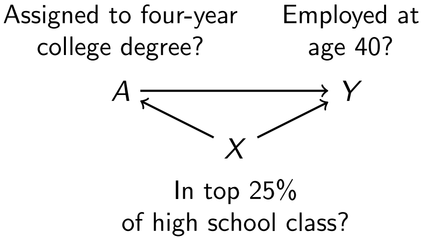
Outcome: Employed at age 40

Elements of a Directed Acyclic Graph (DAG)

$$
\begin{tikzcd}
\text{Assigned to four-year college degree?} & & \text{Employed at age 40?} \\
A \arrow[rr] & & Y \\
& X \arrow[ul] \arrow[ur] & \\
& \text{In top 25\% of high school class?} &
\end{tikzcd}
$$

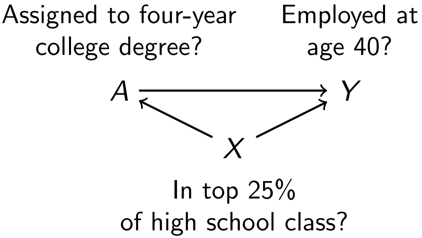
- **Nodes** (X, A, Y) are random variables
- **Edges** $(\rightarrow)$ are causal relationships.
  - X has a causal effect on A
  - X has a causal effect on Y
  - A has a causal effect on Y

Elements of a Directed Acyclic Graph (DAG)



A **path** is a sequence of edges connecting two nodes.

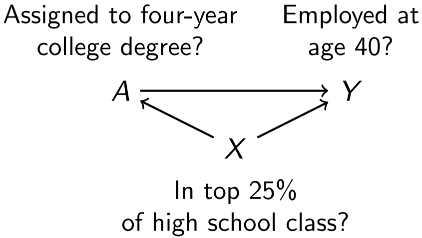
Elements of a Directed Acyclic Graph (DAG)



A **path** is a sequence of edges connecting two nodes.

Between A and Y, what are the two paths?

Elements of a Directed Acyclic Graph (DAG)



A **path** is a sequence of edges connecting two nodes.

Between A and Y, what are the two paths?

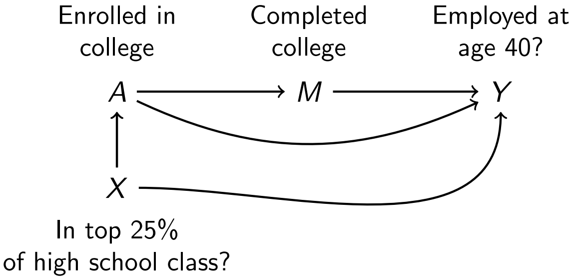
- $A \rightarrow Y$
- $A \leftarrow X \rightarrow Y$

Causal path: A path with arrows pointing one way

$\bullet \to \bullet \to \bullet$

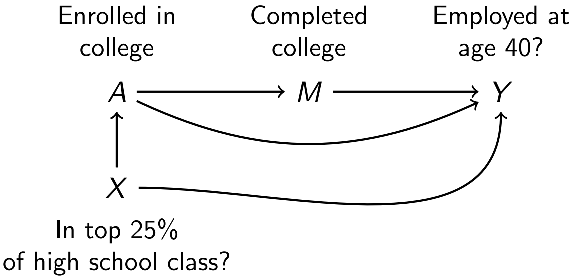
Causal path: A path with arrows pointing one way

$\bullet \rightarrow \bullet \rightarrow \bullet$



Causal path: A path with arrows pointing one way

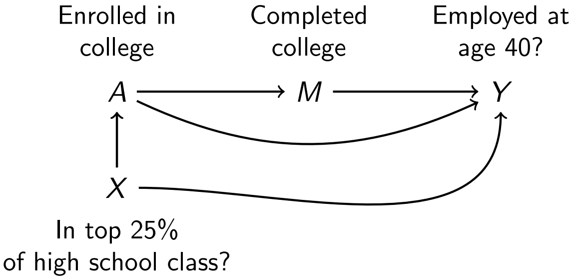
$\bullet \rightarrow \bullet \rightarrow \bullet$



What three paths connect A and Y?
Which two are causal paths?

Causal path: A path with arrows pointing one way

$\bullet \rightarrow \bullet \rightarrow \bullet$

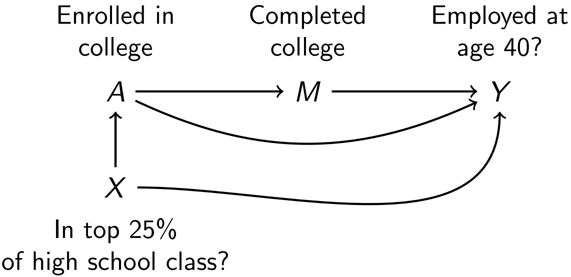


What three paths connect A and Y?
Which two are causal paths?

$A \rightarrow Y$
$A \rightarrow M \rightarrow Y$
$A \leftarrow X \rightarrow Y$

Causal path: A path with arrows pointing one way

$\bullet \rightarrow \bullet \rightarrow \bullet$

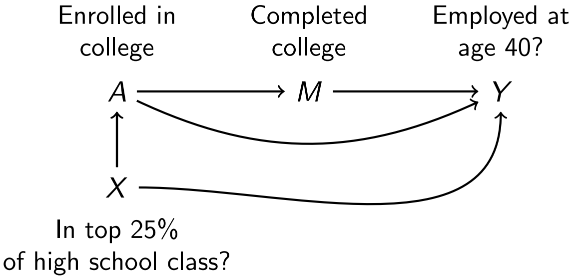


What three paths connect A and Y?
Which two are causal paths?

$A \rightarrow Y$ causal path
$A \rightarrow M \rightarrow Y$
$A \leftarrow X \rightarrow Y$

Causal path: A path with arrows pointing one way

$\bullet \to \bullet \to \bullet$

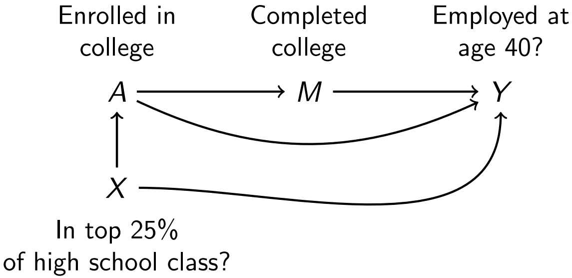


What three paths connect A and Y?
Which two are causal paths?

$A \to Y$	causal path
$A \to M \to Y$	causal path
$A \leftarrow X \to Y$	

Causal path: A path with arrows pointing one way

$\bullet \rightarrow \bullet \rightarrow \bullet$



What three paths connect A and Y?
Which two are causal paths?

$A \rightarrow Y$	causal path
$A \rightarrow M \rightarrow Y$	causal path
$A \leftarrow X \rightarrow Y$	not a causal path

Causal path: Marginal dependence

$\bullet \rightarrow \bullet \rightarrow \bullet$

A causal path $A \rightarrow \cdots \rightarrow B$ will make the variables A and B statistically dependent

Example:

(visits grocery store) $\rightarrow$ (buys ice cream) $\rightarrow$ (eats ice cream)

Causal path: Marginal dependence

$\bullet \to \bullet \to \bullet$

A causal path $A \to \cdots \to B$ will make the variables A and B statistically dependent

Example:

(visits grocery store) $\to$ (buys ice cream) $\to$ (eats ice cream)

What if we condition:
filter to those with (buys ice cream $=$ `FALSE`)?

Causal path: Conditional independence

$\bullet \rightarrow \bullet \rightarrow \bullet$

A causal path $A \rightarrow \cdots \rightarrow B$ will not make the variables A and B statistically dependent if we condition on a variable along the path

Example:

$$\text{(visits grocery store)} \rightarrow \boxed{\text{(buys ice cream)}} \rightarrow \text{(eats ice cream)}$$

Causal path: Conditional independence

$\bullet \rightarrow \bullet \rightarrow \bullet$

A causal path $A \rightarrow \cdots \rightarrow B$ will not make the variables A and B statistically dependent if we condition on a variable along the path

Example:

(visits grocery store) $\rightarrow$ $\boxed{\text{(buys ice cream)}}$ $\rightarrow$ (eats ice cream)

Among people who didn't buy ice cream today,
those who went to the store and didn't
are equally likely to be eating ice cream.

Causal path: Conditional independence

$\bullet \rightarrow \bullet \rightarrow \bullet$

A causal path $A \rightarrow \cdots \rightarrow B$ will not make the variables A and B statistically dependent if we condition on a variable along the path

Example:

$$\text{(visits grocery store)} \rightarrow \boxed{\text{(buys ice cream)}} \rightarrow \text{(eats ice cream)}$$

Among people who didn't buy ice cream today,
those who went to the store and didn't
are equally likely to be eating ice cream.

Conditioning on (buys ice cream = `FALSE`) **blocks** this path.

Fork structure

$\bullet \leftarrow \bullet \rightarrow \bullet$

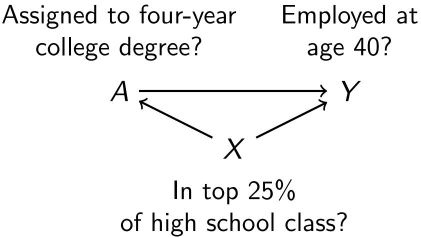
A sequence of edges within a path in which two variables are both caused by a third variable: $A \leftarrow C \rightarrow B$

Fork structure

$\bullet \leftarrow \bullet \rightarrow \bullet$

A sequence of edges within a path in which two variables are both caused by a third variable: $A \leftarrow C \rightarrow B$

In our initial graph, what path contains a fork structure?



Recall that there are two paths:

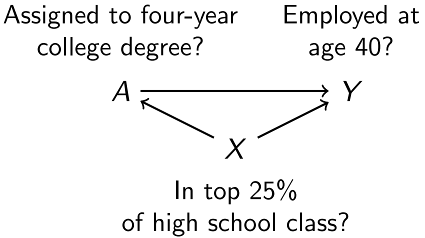
1. $A \rightarrow Y$
2. $A \leftarrow X \rightarrow Y$

Fork structure

$\bullet \leftarrow \bullet \rightarrow \bullet$

A sequence of edges within a path in which two variables are both caused by a third variable: $A \leftarrow C \rightarrow B$

In our initial graph, what path contains a fork structure?



Recall that there are two paths:

1. $A \rightarrow Y$
2. $A \leftarrow X \rightarrow Y$ (this path contains a fork structure)

Fork structure: Marginal dependence

$\bullet \leftarrow \bullet \rightarrow \bullet$

A fork structure $A \leftarrow C \rightarrow B$ will make A and B statistically dependent (because C causes both).

Example:

(completed college) $\leftarrow$ (top 25% of high school) $\rightarrow$ (employed at 40)

Fork structure: Marginal dependence

$\bullet \leftarrow \bullet \rightarrow \bullet$

A fork structure $A \leftarrow C \rightarrow B$ will make A and B statistically dependent (because C causes both).

Example:

$$\text{(lifeguard rescues)} \leftarrow \text{(temperature)} \rightarrow \text{(ice cream sales)}$$

Fork structure: Marginal dependence

$\bullet \leftarrow \bullet \rightarrow \bullet$

A fork structure $A \leftarrow C \rightarrow B$ will make A and B statistically dependent (because C causes both).

Example:

$$\text{(lifeguard rescues)} \leftarrow \text{(temperature)} \rightarrow \text{(ice cream sales)}$$

On days with many lifeguard rescues,
there are also many ice cream sales.
Warm temperature causes both.

Fork structure: Marginal dependence

$\bullet \leftarrow \bullet \rightarrow \bullet$

A fork structure $A \leftarrow C \rightarrow B$ will make A and B statistically dependent (because C causes both).

Example:

$$\text{(lifeguard rescues)} \leftarrow \text{(temperature)} \rightarrow \text{(ice cream sales)}$$

On days with many lifeguard rescues,
there are also many ice cream sales.
Warm temperature causes both.

What if we look only at days with a given temperature?

Fork structure: Conditional independence

$\bullet \leftarrow \bullet \rightarrow \bullet$

A fork structure $A \leftarrow \boxed{C} \rightarrow B$ does not make A and B statistically dependent if we condition on C.

Example:

$$\text{(lifeguard rescues)} \leftarrow \boxed{\text{(temperature)}} \rightarrow \text{(ice cream sales)}$$

Among days with a given temperature,
lifeguard rescues and ice cream sales are unrelated.

Conditioning on (temperature) blocks this path.

Collider structure

$\bullet \rightarrow \bullet \leftarrow \bullet$

A sequence of edges within a path in which two variables both cause a third variable: $A \rightarrow C \leftarrow B$

Collider structure

$\bullet \rightarrow \bullet \leftarrow \bullet$

A sequence of edges within a path in which two variables both cause a third variable: $A \rightarrow C \leftarrow B$

Example:

- sprinklers on a timer
- rain on random days
- either one can make the grass wet

$$\text{(sprinklers on)} \rightarrow \text{(grass wet)} \leftarrow \text{(raining)}$$

Collider structure

$\bullet \rightarrow \bullet \leftarrow \bullet$

A sequence of edges within a path in which two variables both cause a third variable: $A \rightarrow C \leftarrow B$

Example:

- sprinklers on a timer
- rain on random days
- either one can make the grass wet

$$\text{(sprinklers on)} \rightarrow \text{(grass wet)} \leftarrow \text{(raining)}$$

Are (sprinklers on) and (raining) statistically related?

Collider structure: Marginal independence

$\bullet \rightarrow \bullet \leftarrow \bullet$

In a collider structure $A \rightarrow C \leftarrow B$,
A and B are marginally independent.

$$\text{(sprinklers on)} \rightarrow \text{(grass wet)} \leftarrow \text{(raining)}$$

Knowing (sprinklers on $=$ `TRUE`) tells me nothing about whether (raining $=$ `TRUE`)

Collider structure: Marginal independence

$\bullet \rightarrow \bullet \leftarrow \bullet$

In a collider structure $A \rightarrow C \leftarrow B$,
A and B are marginally independent.

$$\text{(sprinklers on)} \rightarrow \text{(grass wet)} \leftarrow \text{(raining)}$$

Knowing (sprinklers on $=$ `TRUE`) tells me nothing about whether (raining $=$ `TRUE`)

What if I condition: look only at days when the grass is wet?

Collider structure: Conditional dependence

$\bullet \rightarrow \bullet \leftarrow \bullet$

(sprinklers on) $\rightarrow$ $\boxed{\text{(grass wet)}}$ $\leftarrow$ (raining)

Collider structure: Conditional dependence

$\bullet \rightarrow \bullet \leftarrow \bullet$

$$\text{(sprinklers on)} \rightarrow \boxed{\text{(grass wet)}} \leftarrow \text{(raining)}$$

Among days when (grass wet $=$ `TRUE`),
if (sprinklers on $=$ `FALSE`)
then it must be (raining $=$ `TRUE`)

(grass had to get wet somehow!)

Collider structure: Conditional dependence

$\bullet \rightarrow \bullet \leftarrow \bullet$

$$\text{(sprinklers on)} \rightarrow \boxed{\text{(grass wet)}} \leftarrow \text{(raining)}$$

Among days when (grass wet $=$ `TRUE`),
if (sprinklers on $=$ `FALSE`)
then it must be (raining $=$ `TRUE`)

(grass had to get wet somehow!)

In a collider structure $A \rightarrow \boxed{C} \leftarrow B$,
A and B are conditionally dependent.

Review: Three structures

Name	Structure	A and B marginally dependent?	A and B conditionally dependent given C?
Causal path	$A \rightarrow C \rightarrow B$	Yes	No
Fork	$A \leftarrow C \rightarrow B$	Yes	No
Collider	$A \rightarrow C \leftarrow B$	No	Yes

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works
sprinklers on
grass wet
raining

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
sprinklers on	
grass wet	
raining	

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
sprinklers on	yes
grass wet	
raining	

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
sprinklers on	yes
grass wet	yes
raining	

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
sprinklers on	yes
grass wet	yes
raining	no

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ (sprinklers on) $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
sprinklers on	yes
grass wet	yes
raining	no

We just learned: One collider can block an entire path

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works
grass wet
raining

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
grass wet	
raining	

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
grass wet	no
raining	

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
grass wet	no
raining	no

A path can involve forks, colliders, and causal paths

(timer displays clock) $\leftarrow$ (timer works) $\rightarrow$ $\boxed{\text{(sprinklers on)}}$ $\rightarrow$ (grass wet) $\leftarrow$ (raining)

(timer displays clock) is statistically related to which variables?

timer works	yes
grass wet	no
raining	no

We just learned: One conditioned non-collider can block an entire path

Rules for whether paths are open or blocked

- If a path contains an unconditioned collider, it is blocked
- If a path contains a conditioned non-collider, it is blocked
- Otherwise, the path is open

Open paths create statistical dependence.
Blocked paths do not.

Determining statistical dependence: A procedure

How do you know if two nodes (e.g., A and B are dependent?

Determining statistical dependence: A procedure

How do you know if two nodes (e.g., A and B are dependent?

$$\begin{tikzcd}[column sep=small, row sep=small]
 & C \arrow[dl] \arrow[dr] & \\
A \arrow[dr] & & B \arrow[dl] \\
 & D &
\end{tikzcd}$$

Determining statistical dependence: A procedure

How do you know if two nodes (e.g., A and B are dependent?

$$\begin{tikzcd}[row sep=small, column sep=small] & C \arrow[dl] \arrow[dr] & \\ A \arrow[dr] & & B \arrow[dl] \\ & D & \end{tikzcd}$$

1. List all paths between the two nodes
   - $A \leftarrow C \rightarrow B$
   - $A \rightarrow D \leftarrow B$

Determining statistical dependence: A procedure

How do you know if two nodes (e.g., A and B are dependent?

$$\begin{tikzcd} & C \arrow[dl] \arrow[dr] & \\ A \arrow[dr] & & B \arrow[dl] \\ & D & \end{tikzcd}$$

1. List all paths between the two nodes
   - $A \leftarrow C \rightarrow B$
   - $A \rightarrow D \leftarrow B$
2. Cross out any blocked paths that are blocked
   - $A \leftarrow C \rightarrow B$
   - $\cancel{A \rightarrow D \leftarrow B}$

Determining statistical dependence: A procedure

How do you know if two nodes (e.g., A and B are dependent?

$$
\begin{tikzcd}
 & C \arrow[dl] \arrow[dr] & \\
A \arrow[dr] & & B \arrow[dl] \\
 & D &
\end{tikzcd}
$$

1. List all paths between the two nodes
   - $A \leftarrow C \rightarrow B$
   - $A \rightarrow D \leftarrow B$
2. Cross out any blocked paths that are blocked
   - $A \leftarrow C \rightarrow B$
   - ~~$A \rightarrow D \leftarrow B$~~
3. If any paths remain, the two nodes are dependent
   - Dependent!

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

Power outage
throughout town X

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

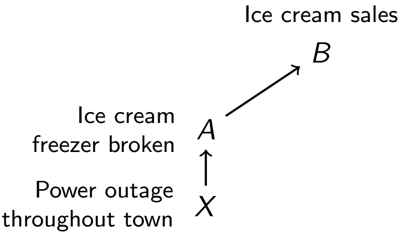
Ice cream freezer broken A

$\uparrow$

Power outage throughout town X

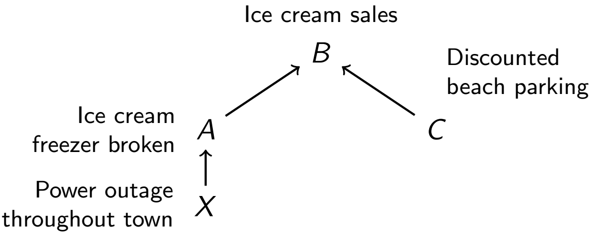
Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



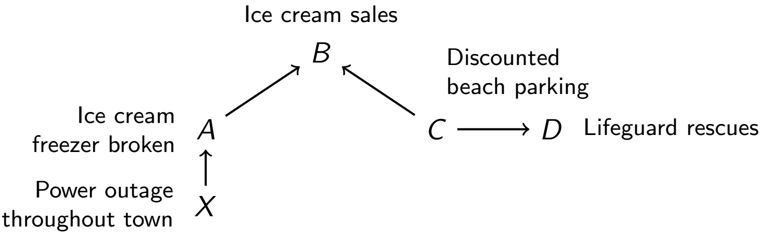
Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



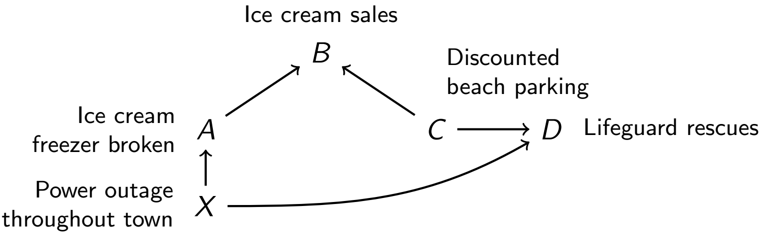
Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



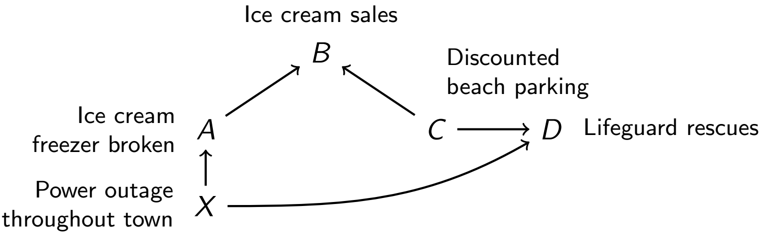
Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Determining statistical dependence: A procedure

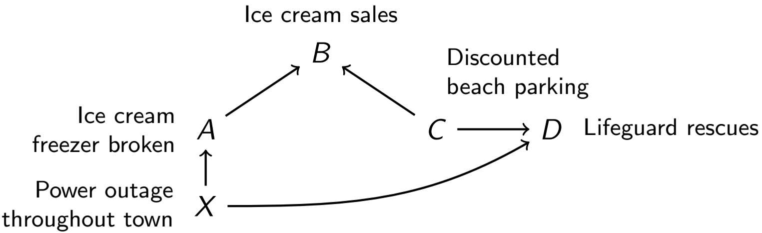
1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Are A and C statistically independent or dependent?

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

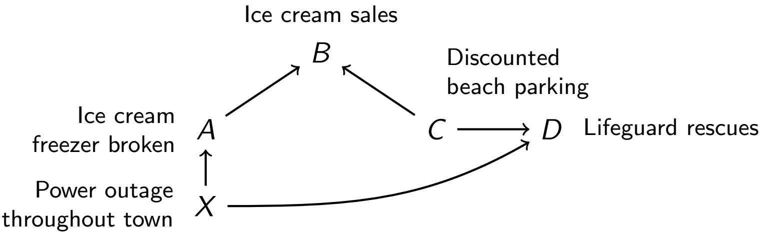


Are A and C statistically independent or dependent?

- $A \rightarrow B \leftarrow C$
- $A \leftarrow X \rightarrow D \leftarrow C$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

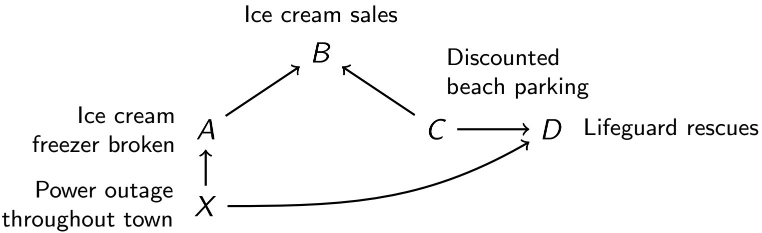


Are A and C statistically independent or dependent?

- $\cancel{A \rightarrow B \leftarrow C}$
- $A \leftarrow X \rightarrow D \leftarrow C$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

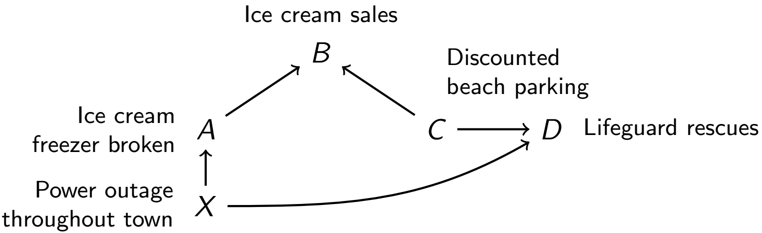


Are A and C statistically independent or dependent?

- $\cancel{A \rightarrow B \leftarrow C}$
- $\cancel{A \leftarrow X \rightarrow D \leftarrow C}$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



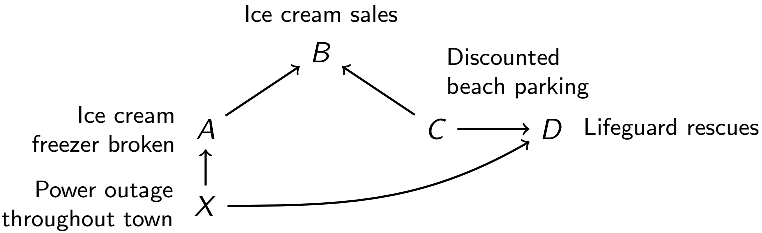
Are A and C statistically independent or dependent?

- $\cancel{A \rightarrow B \leftarrow C}$
- $\cancel{A \leftarrow X \rightarrow D \leftarrow C}$

No unblocked paths. Independent!

Determining statistical dependence: A procedure

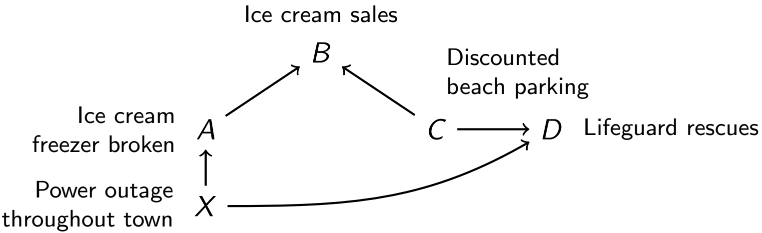
1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Are A and D statistically independent or dependent?

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

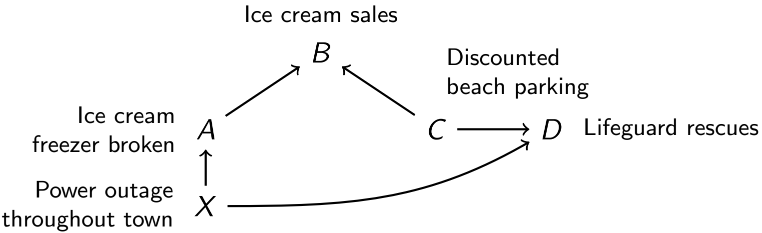


Are A and D statistically independent or dependent?

- $A \rightarrow B \leftarrow C \rightarrow D$
- $A \leftarrow X \rightarrow D$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

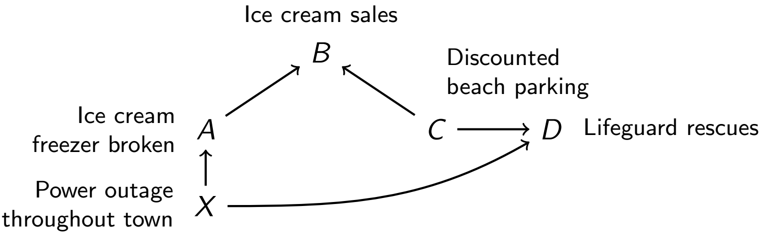


Are A and D statistically independent or dependent?

- $\cancel{A \rightarrow B \leftarrow C \rightarrow D}$
- $A \leftarrow X \rightarrow D$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Are A and D statistically independent or dependent?

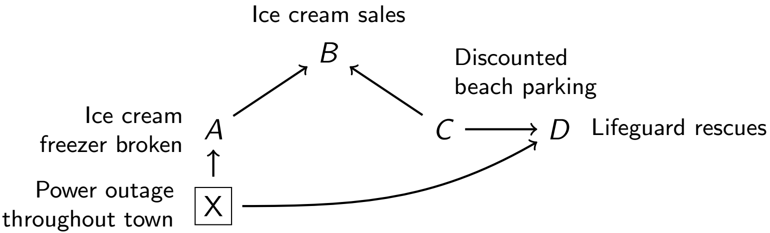
- ~~$A \rightarrow B \leftarrow C \rightarrow D$~~
- $A \leftarrow X \rightarrow D$

A path remains unblocked. Dependent!

Practice with **conditional** dependence (holding something constant)

Determining statistical dependence: A procedure

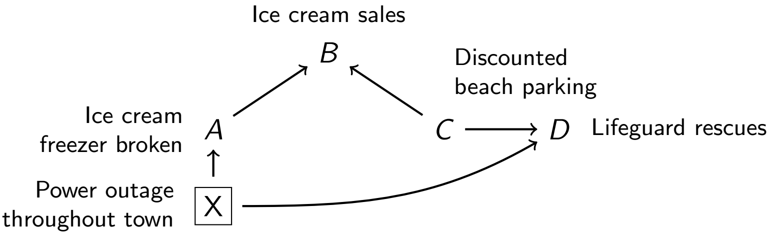
1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$?

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

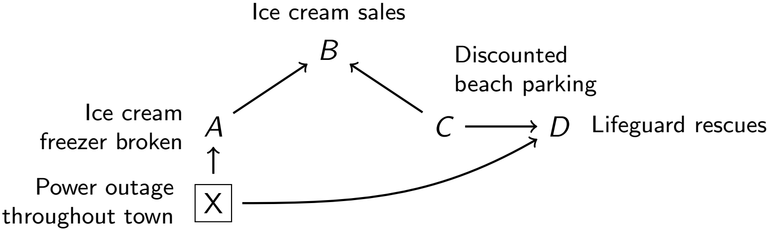


Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$?

- $A \rightarrow B \leftarrow C \rightarrow D$
- $A \leftarrow \boxed{X} \rightarrow D$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

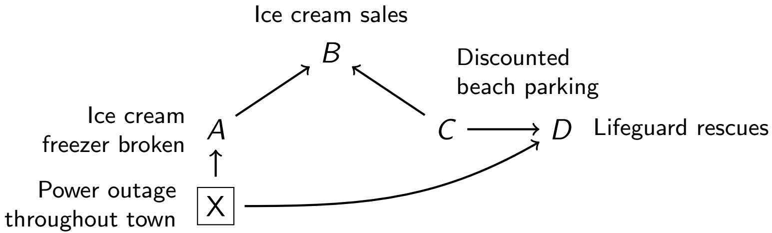


Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$?

- $\cancel{A \rightarrow B \leftarrow C \rightarrow D}$
- $A \leftarrow \boxed{X} \rightarrow D$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

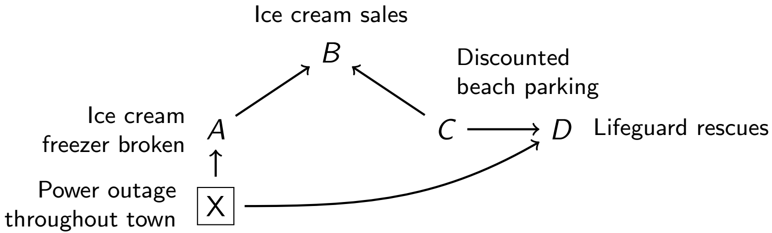


Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$?

- $\cancel{A \rightarrow B \leftarrow C \rightarrow D}$
- $\cancel{A \leftarrow \boxed{X} \rightarrow D}$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



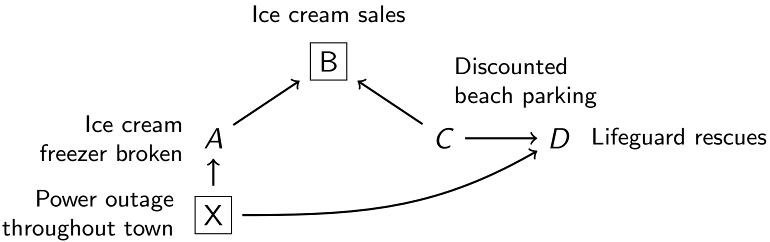
Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$?

- $\cancel{A \rightarrow B \leftarrow C \rightarrow D}$
- $\cancel{A \leftarrow \boxed{X} \rightarrow D}$

No unblocked paths. Independent!

Determining statistical dependence: A procedure

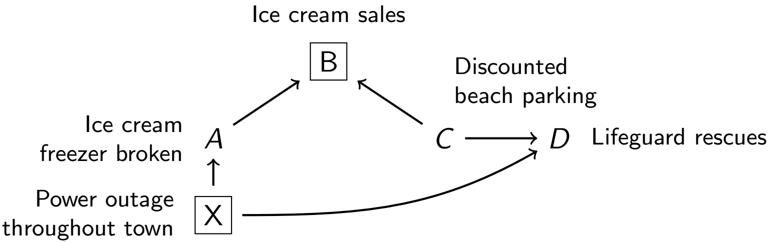
1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$ and $B = 0$?

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

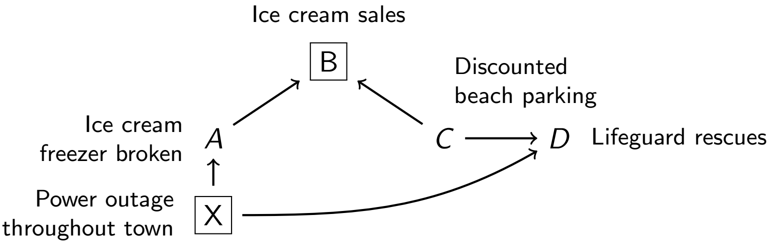


Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$ and $B = 0$?

- $A \to \boxed{B} \leftarrow C \to D$
- $A \leftarrow \boxed{X} \to D$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.

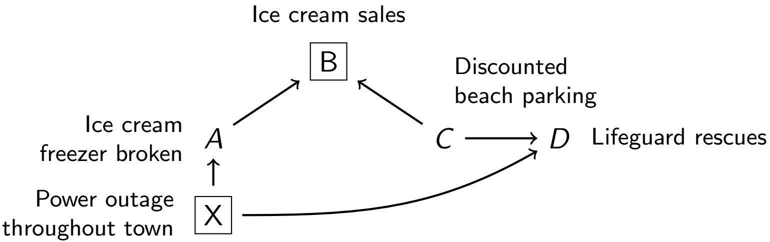


Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$ and $B = 0$?

- $A \rightarrow \boxed{B} \leftarrow C \rightarrow D$
- $\cancel{A \leftarrow \boxed{X} \rightarrow D}$

Determining statistical dependence: A procedure

1. List all paths. 2. Cross out blocked paths. 3. Dependent if any paths remain.



Practice: Are A and D statistically independent or dependent, conditional on $X = \texttt{FALSE}$ and $B = 0$?

- $A \rightarrow \boxed{B} \leftarrow C \rightarrow D$
- $\cancel{A \leftarrow \boxed{X} \rightarrow D}$

A path remains. Dependent!

DAGs and conditional exchangeability

When studying the effect of A on Y, conditional exchangeability holds if the only unblocked paths between A and Y are causal paths from A to Y.

DAGs and conditional exchangeability

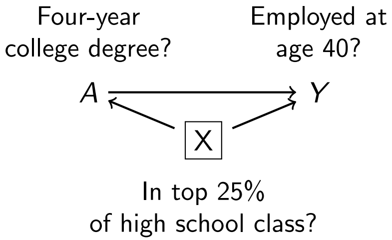
When studying the effect of A on Y, conditional exchangeability holds if the only unblocked paths between A and Y are causal paths from A to Y.

- Why? Because then any association between A and Y must be due to the causal effect

DAGs and conditional exchangeability

When studying the effect of A on Y, conditional exchangeability holds if the only unblocked paths between A and Y are causal paths from A to Y.

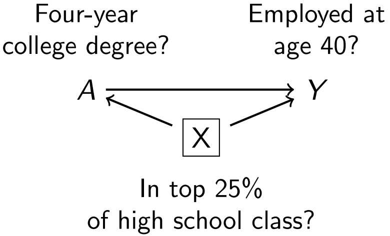
- Why? Because then any association between A and Y must be due to the causal effect



DAGs and conditional exchangeability

When studying the effect of A on Y, conditional exchangeability holds if the only unblocked paths between A and Y are causal paths from A to Y.

- Why? Because then any association between A and Y must be due to the causal effect



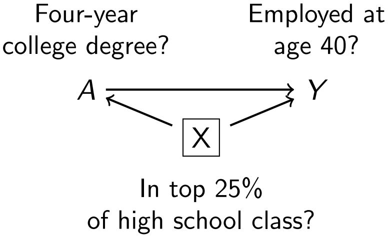
- $A \to Y$
- $A \leftarrow \boxed{X} \to Y$ (blocked by conditioning on X)

DAGs and conditional exchangeability

When studying the effect of A on Y, conditional exchangeability holds if the only unblocked paths between A and Y are causal paths from A to Y.

- Why? Because then any association between A and Y must be due to the causal effect

X is a **sufficient adjustment set** to identify the effect of A on Y

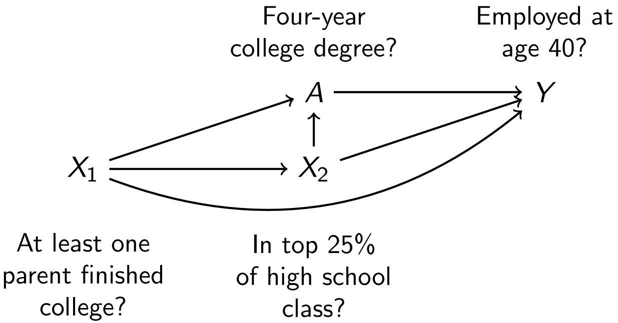


- $A \to Y$
- $A \leftarrow \boxed{X} \to Y$ (blocked by conditioning on X)

DAGs and conditional exchangeability: Practice

1. List all paths. 2. Choose adjustment set. 3. Only causal paths remain unblocked.

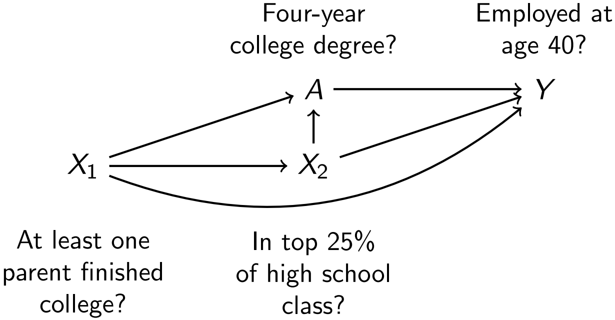
Find a sufficient adjustment set to identify the effect of A on Y.



DAGs and conditional exchangeability: Practice

1. List all paths. 2. Choose adjustment set. 3. Only causal paths remain unblocked.

Find a sufficient adjustment set to identify the effect of A on Y.

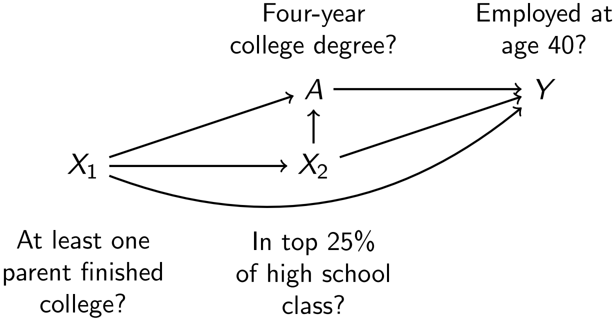


Paths: $(A \to Y)$, $(A \leftarrow X_2 \to Y)$, $(A \leftarrow X_1 \to Y)$,
$(A \leftarrow X_1 \to X_2 \to Y)$,$(A \leftarrow X_2 \leftarrow X_1 \to Y)$

DAGs and conditional exchangeability: Practice

1. List all paths. 2. Choose adjustment set. 3. Only causal paths remain unblocked.

Find a sufficient adjustment set to identify the effect of A on Y.



Paths: $(A \rightarrow Y)$, $(A \leftarrow X_2 \rightarrow Y)$, $(A \leftarrow X_1 \rightarrow Y)$,
$(A \leftarrow X_1 \rightarrow X_2 \rightarrow Y)$,$(A \leftarrow X_2 \leftarrow X_1 \rightarrow Y)$
Adjust for $\{X_1, X_2\}$

How to draw a DAG

1. Begin with treatment A and outcome Y
2. Add any variable that affects both
3. Add any variable that affects any two variables in the DAG.

Assumptions are about nodes and edges that you omit.

Exercise: Draw a DAG

Treatment is college degree. Outcome is employment at age 40. Identify a sufficient adjustment set under your DAG.

Learning goals for today

By the end of class, you will be able to

- define causal effects
- identify average causal effects by
  - exchangeability
  - conditional exchangeability
- select a sufficient adjustment set using a Directed Acyclic Graph