Data 102 Lecture 14:

Causal inference II

What is causality?

We mean different things when we say “cause”

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

I am who I am today because of my mum.

Gravity causes objects to fall to Earth.

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

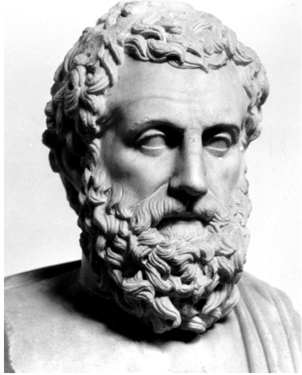
Smoking causes lung cancer.

All of form

A causes B

But meanings slightly different

Aristotle’s definition

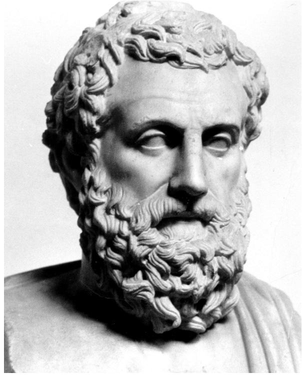


A cause is an answer to a “why” question

Four types of causes

- The **material** cause: “that out of which”, e.g., the bronze of a statue.
- The **formal** cause: “the form”, “the account of what-it-is-to-be”, e.g., the shape of a statue.
- The **efficient** cause: “the primary source of the change or rest”, e.g., the artisan, the art of bronze-casting the statue, the man who gives advice.
- The **final** cause: “the end, that for the sake of which a thing is done”, e.g., health is the end of walking, losing weight, purging, drugs, and surgical tools.

Modern scientific use focuses on **efficient** cause

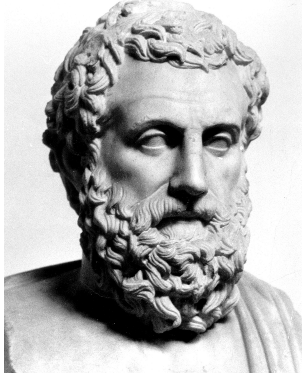


A cause is an answer to a “why” question

Four types of causes

- The **material** cause: “that out of which”, e.g., the bronze of a statue.
- The **formal** cause: “the form”, “the account of what-it-is-to-be”, e.g., the shape of a statue.
- The **efficient** cause: “the primary source of the change or rest”, e.g., the artisan, the art of bronze-casting the statue, the man who gives advice.
- The **final** cause: “the end, that for the sake of which a thing is done”, e.g., health is the end of walking, losing weight, purging, drugs, and surgical tools.

Modern scientific use focuses on **efficient** cause



A cause is an answer to a “why” question

Four types of causes

- The **material** cause: “that out of which”, e.g., the bronze of a statue.
- The **formal** cause: “the form”, “the account of what-it-is-to-be”, e.g., the shape of a statue.
- The **efficient** cause: “the primary source of the change or rest”, e.g., the artisan, the art of bronze-casting the statue, the man who gives advice.
- The **final** cause: “the end, that for the sake of which a thing is done”, e.g., health is the end of walking, losing weight, purging, drugs, and surgical tools.

We mean different things when we say “cause”

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

I am who I am today because of my mum.

Gravity causes objects to fall to Earth.

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

We mean different things when we say “cause”

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

~~I am who I am today because of my mum.~~

~~Gravity causes objects to fall to Earth.~~

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

Other differences in meaning...

Deterministic cause vs probabilistic cause

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

~~I am who I am today because of my mum.~~

~~Gravity causes objects to fall to Earth.~~

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

Instance vs class

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

~~I am who I am today because of my mum.~~

~~Gravity causes objects to fall to Earth.~~

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

Single cause vs multiple causes

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn’t spending enough time with her.

~~I am who I am today because of my mum.~~

~~Gravity causes objects to fall to Earth.~~

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

Simple action/event vs compound action/event

The soccer ball moved because I kicked it.

My girlfriend broke up with me because I wasn't spending enough time with her.

~~I am who I am today because of my mum.~~

~~Gravity causes objects to fall to Earth.~~

Down Syndrome is caused by an extra copy of chromosome 21.

Humans are causing climate change.

Barbarian invasions caused the Roman Empire to fall.

Smoking causes lung cancer.

Each domain of knowledge has their own conventions / nuances

Death certificate

CAUSE OF DEATH (See instructions and examples)

32. **PART I.** Enter the chain of events--diseases, injuries, or complications--that directly caused the death. DO NOT enter terminal events such as cardiac arrest, respiratory arrest, or ventricular fibrillation without showing the etiology. DO NOT ABBREVIATE. Enter only one cause on a line. Add additional lines if necessary.

		Approximate interval: Onset to death
IMMEDIATE CAUSE (Final disease or condition ---------> resulting in death)	a. Acute respiratory distress syndrome Due to (or as a consequence of):	2 days
Sequentially list conditions, if any, leading to the cause listed on line a. Enter the **UNDERLYING CAUSE** (disease or injury that initiated the events resulting in death) **LAST**	b. Pneumonia Due to (or as a consequence of):	10 days
	c. COVID-19 Due to (or as a consequence of):	10 days
	d.	

PART II. Enter other significant conditions contributing to death but not resulting in the underlying cause given in PART I.

33. WAS AN AUTOPSY PERFORMED?
☐ Yes [x] No

34. WERE AUTOPSY FINDINGS AVAILABLE TO COMPLETE THE CAUSE OF DEATH? ☐ Yes ☐ No

35. DID TOBACCO USE CONTRIBUTE TO DEATH?
☐ Yes ☐ Probably
[x] No ☐ Unknown

36. IF FEMALE:
[x] Not pregnant within past year
☐ Pregnant at time of death
☐ Not pregnant, but pregnant within 42 days of death
☐ Not pregnant, but pregnant 43 days to 1 year before death
☐ Unknown if pregnant within the past year

37. MANNER OF DEATH
[x] Natural ☐ Homicide
☐ Accident ☐ Pending Investigation
☐ Suicide ☐ Could not be determined

Takeaways

We have an intuitive sense of what causality means, but it’s hard to make it precise.

In everyday language, our use of the word “cause” has different meanings

Different domains has its own nuances and conventions about how they define causes

A lot of ongoing philosophical debate

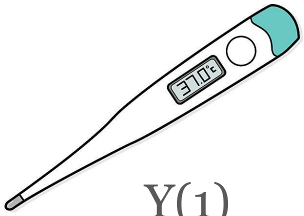
Lecture 14 overview

- Formalizing causality using counterfactuals
- Randomized experiments
  - Assumptions
  - Examples
  - What is the randomness
- Inference for randomized experiments
- Complications with randomized experiments

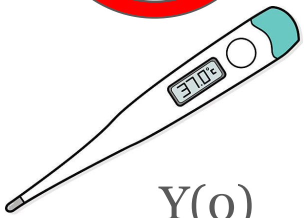
We define causality using counterfactuals

Universe 1

Universe 0



Y(1)



Y(0)

Y(1), Y(0) are called potential outcomes

Individual treatment effect is Y(1) - Y(0)

Fundamental problem of causal inference:

We only see one of the two potential outcomes, so the individual treatment effect is unidentified

Ronald Fisher's insight

Do a randomized experiment!

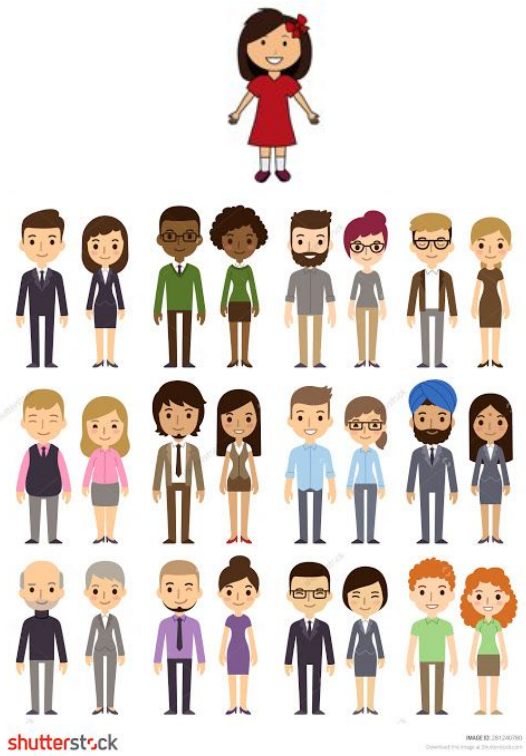
- Consider multiple individuals (units) at the same time
- Use randomization as a basis for inference

Fisher, R. The Design of Experiments (1935)

The Science Table of a randomized experiment

i	$Y_i(1)$	$Y_i(0)$
1	$Y_1(1)$	$Y_1(0)$
2	$Y_2(1)$	$Y_2(0)$
$\vdots$	$\vdots$	$\vdots$
n	$Y_n(1)$	$Y_n(0)$

The Science Table of a randomized experiment



i	$Y_i(1)$	$Y_i(0)$
1	?	$Y_1(0)$
2	$Y_2(1)$	?
$\vdots$	$\vdots$	$\vdots$
n	?	$Y_n(0)$

Some formalism / definitions

Z_i is the treatment assignment for unit i

- In the basic setting, treatment is binary, but can consider treatments with more levels

$Y_i(1)$ and $Y_i(0)$ are the potential outcomes of unit i

We observe $Y_{i,obs} = Y_i(Z_i)$

In order for Science Table to make sense...

Need the **Stable Unit Treatment Value Assumption (SUTVA)**

1. No interference: Potential outcomes for unit i do not depend on treatment assigned to other units
2. There are no other version of the treatment

Examples of experiments

1. We are interested in the effect of participating in a job training program or not on employment and wage. The intervention is participating in a job training program.
2. Gerber et al. (2008) were interested in the effect of different get-out-to-vote (GOTV) messages on the voting behavior. The intervention is different GOTV messages.
3. 2008 Oregon Health Insurance Experiment: Investigated the effect of health insurance on health and economic outcomes. The intervention was lottery Medicaid enrollment via lottery.
4. Technology companies are typically interested in how different versions of an app affects user engagement.

SUTVA not always reasonable...

1. We are interested in the effect of participating in a job training program or not on employment and wage. The intervention is participating in a job training program.
2. Gerber et al. (2008) were interested in the effect of different get-out-to-vote (GOTV) messages on the voting behavior. The intervention is different GOTV messages. People in same household will receive the same message.
3. 2008 Oregon Health Insurance Experiment: Investigated the effect of health insurance on health and economic outcomes. The intervention was lottery Medicaid enrollment via lottery.
4. Technology companies are typically interested in how different versions of an app affects user engagement. If User A gets affected by treatment, then he/she will interact more/less with their friends.

The randomization scheme

Determined by the people running the experiment.

Typically, do **complete randomization**.

$$\mathbb{P}(Z = z) = \frac{1}{\binom{n}{n_1}}$$

Where $\boldsymbol{z} = (z_1, \ldots, z_n)$ satisfies $\sum_{i=1}^n z_i = n_1$ and $\sum_{i=1}^n (1 - z_i) = n_0$

I.e. we choose n_1 units uniformly without replacement to form our treatment group

Inference for randomized experiments

The Average Treatment Effect (ATE)

$$\tau = \frac{1}{n} \sum_{i=1}^{n} (Y_i(1) - Y_i(0))$$

This is the average of individual treatment effects

This is a fixed quantity. The potential outcomes are assumed to be fixed.

It is unidentifiable, can only estimate it.

The Neyman estimator (difference-in-means)

$$\hat{\tau} = \frac{1}{n_1} \sum_{Z_i=1} Y_{i,obs} - \frac{1}{n_0} \sum_{Z_i=0} Y_{i,obs}$$

Mean and variance for Neyman estimator

$$\mathbb{E}[\hat{\tau}] = \tau$$

$$\mathrm{Var}(\hat{\tau}) \leq \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

Where for k = 0, 1: $S_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} \left(Y_i(k) - \bar{Y}(k)\right)^2$

Confidence intervals for ATE

See whiteboard notes

Hypothesis testing for randomized experiments

Neyman's weak null H_{0N}: ATE = 0

Fisher's strong null H_{0F}: ITE $Y_i(1) - Y_i(0) = 0$ for each unit i

Much debate about which null hypothesis is more meaningful.

See whiteboard notes for more details

Fisher’s exact test

See whiteboard notes

Data example: Racial discrimination in the labor market

Bertrand, M. and Mullainathan, S. (2004) wanted to study whether there was racial discrimination in the labor market

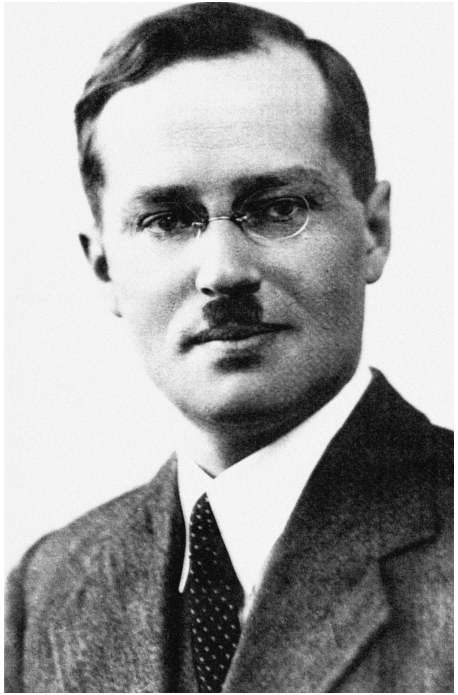
Experiment: Create 4870 fake resumes, half of them with stereotypical white names, and half of them with stereotypical black names, sent them to potential employers

Treatment: Race

Outcome: Whether or not the resume received a call back.

Bertrand, M. and Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? a field experiment on labor market discrimination. American Economic Review, 94(4):991–1013.

Background: Neyman-Rubin framework



Jerzy Neyman
(1894-1981)

Donald Rubin
(1943-)

Summary about inference under Neyman-Rubin framework

Does not require modeling assumptions on the potential outcomes

Not completely assumption-free: SUTVA

This is a frequentist framework

Randomness comes from the experiment design, not from sampling

Causality is formulated in terms of instances (what happens to me), but is estimated in terms of a class (what happens on average to everyone)

Heterogeneity is unaccounted for (for now…)

Complications with randomized experiments

Compliance

Units may not comply with their assigned treatment

E.g. 1: In medical trials, patients sometimes decline their assignment treatment (stop taking their medication, or decline surgical procedure.)

E.g. 2. In the Oregon Health Insurance experiment, only 45% of lottery winners enrolled in Medicaid. In addition, 18% of lottery losers enrolled.

External validity

The ATE is defined with respect to the experiment population. This may not be representative of the larger population of interest.

E.g. Most polls in the 2016 US presidential election overestimated the support for Hillary Clinton. This polling error was due to the polled population being different from the voting population (for instance, Trump voters were less likely to answer the phone)

- This example is not exactly about causal inference, but this is the crux of the external validity problem

Both problems are topics of current research