STA 235H - Potential Outcomes

Fall 2023

McCombs School of Business, UT Austin

How? Potential Outcomes Framework

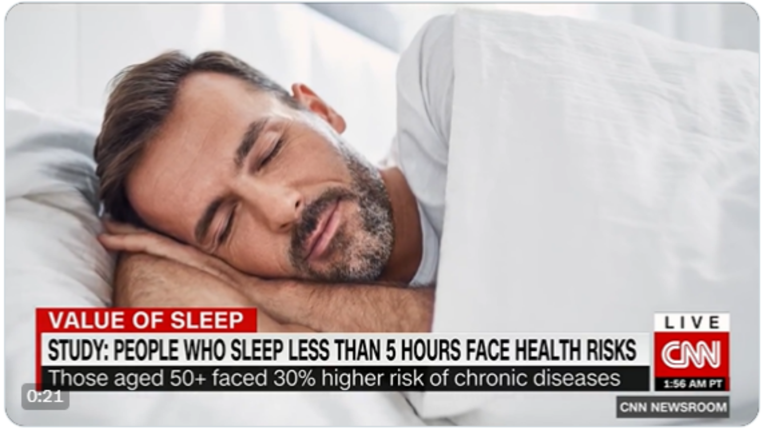
What? Causal Estimands

Why? Causal Questions and Study Design

The "How": Potential outcomes framework

CNN
@CNN

A large new study suggests that people aged 50 and older who sleep five hours or less at night have a greater risk of developing multiple chronic diseases. This is what sleep experts recommend for a better night's rest 👇 cnn.it/3VDs39u



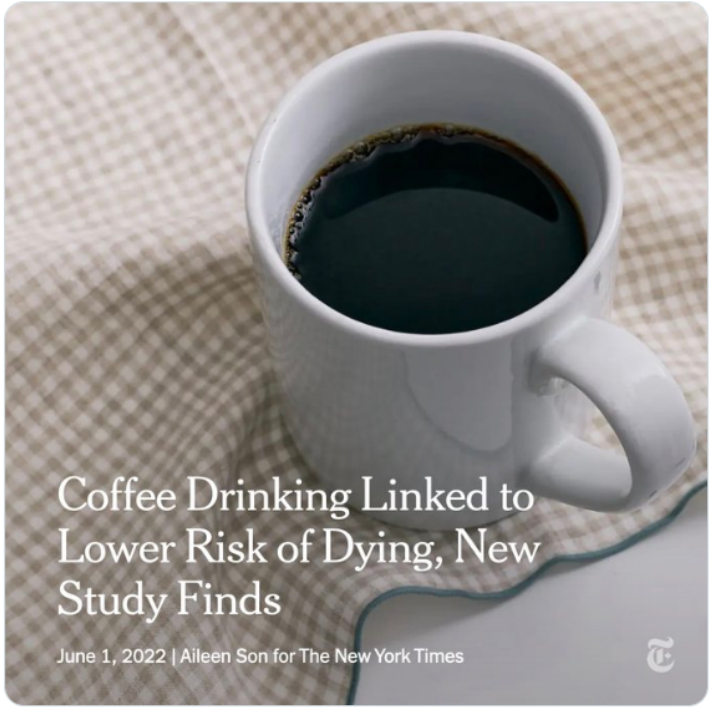
7:15 AM · Oct 19, 2022

141 Reposts 18 Quotes 529 Likes 59 Bookmarks

What do you think are the biggest issues here?

The New York Times
@nytimes

People who drank moderate amounts of coffee, 1.5 to 3.5 cups per day, were up to 30% less likely to die during a multiple year study period than those who didn't drink coffee, new research found. nyti.ms/3NaUTcK

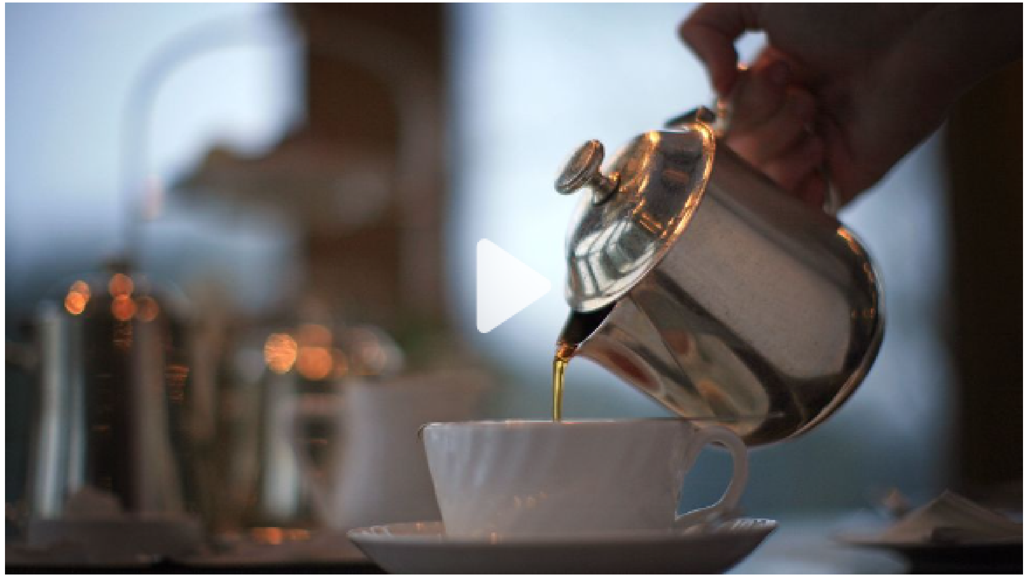


6:00 PM · Jun 1, 2022

1,676 Reposts **2,811** Quotes **6,967** Likes **381** Bookmarks

New research reveals how coffee and tea can affect risk of early death for adults with diabetes

By Sandee LaMotte, CNN
Updated 7:01 PM EDT, Wed April 19, 2023



Video Ad Feedback

The health benefits of tea

01:10 - Source: CNN

Before we start...

Be clear about your language

Be clear about your data

Be clear about your assumptions

What is Causal Inference?

Inferring the effect of one thing on another thing

- "My headache went away because I took an aspirin".
- "The new marketing campaign increased our sales by 20%"
- "Providing students support when filling out FAFSA forms improves college access and completion."

A world of potential (outcomes)

- Under a binary treatment or intervention, there are **two potential worlds**:

- **World 1**: You take the pill
- **World 2**: You don't take the pill

A world of potential (outcomes)

- A **potential outcome** is the outcome under each of these scenarios or "worlds".
  - *There will be one for each path!*
- A priori, each of these scenarios has a *potential outcome*
- A posteriori, I can only observe **at most one of the potential outcomes**

Fundamental Problem of Causal Inference

What are the potential outcomes for our previous example?

Potential Outcomes Examples

- "My headache went away because I took an aspirin".

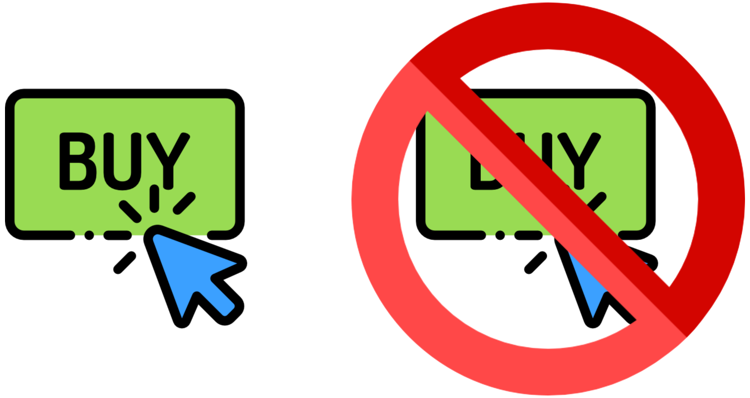
Headache status if I take an aspirin/ Headache status if I don't take an aspirin

- "The new marketing campaign increased our sales by 20%"
- "Providing students support when filling out FAFSA forms improves college access and completion."

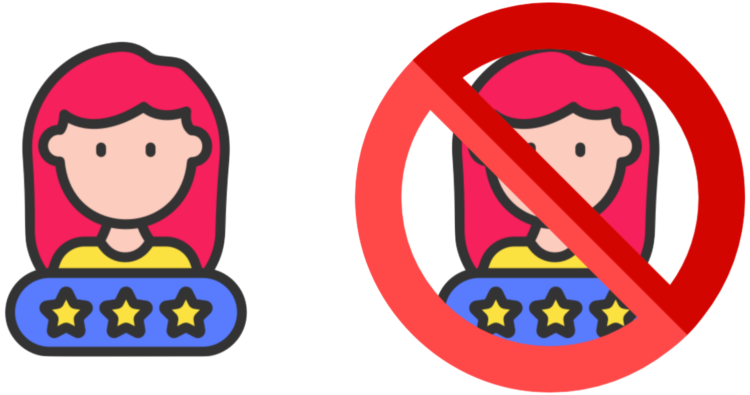
Let's see a specific example

- You work at a retail company and you are debating on whether to send out an **email campaign** to boost your sales:
- You are interested in **two specific outcomes**:

Sales: Whether a customer makes a purchase or not.



Churn: Whether a customer unsubscribes for your mailing list or not.



Potential Outcomes Framework

Let's introduce some notation:

- Let Y_i be the observed outcome for unit i (e.g. whether a person makes a purchase or not).
- Let Z_i be the treatment or intervention (e.g. receiving a promotional email (1) or not (0)).
- Let $Y_i(z)$ be the potential outcome under treatment $Z = z$. (e.g. whether the person would make a purchase or not *if* they received treatment z).

Then, if a person is *treated*, $Z_i = 1$, then their *observed outcome* Y_i will be the same as their *potential outcome under treatment*, $Y_i(1)$

$$Y_i|(Z_i = 1) \triangleq Y_i(1)$$

In the same fashion, if a person is not *treated*, $Z_i = 0$, then their *observed outcome* Y_i will be the same as their *potential outcome under control*, $Y_i(0)$

$$Y_i|(Z_i = 0) \triangleq Y_i(0)$$

Potential Outcomes Framework

This means that we can write the observed outcome as a function of the *potential outcomes*:

$$\rightarrow Y_i = Z_i \cdot Y_i(1) + (1 - Z_i) \cdot Y_i(0)$$

- This definition will be useful because we can see this as a missing data problem.

Causal Effects

Individual Causal Effect

$$ICE_i = Y_i(1) - Y_i(0)$$

Causal Effects

Individual Causal Effect

$$ICE_i = Y_i(1) - Y_i(0)$$

Can we ever observe individual causal effects?

Causal Effects

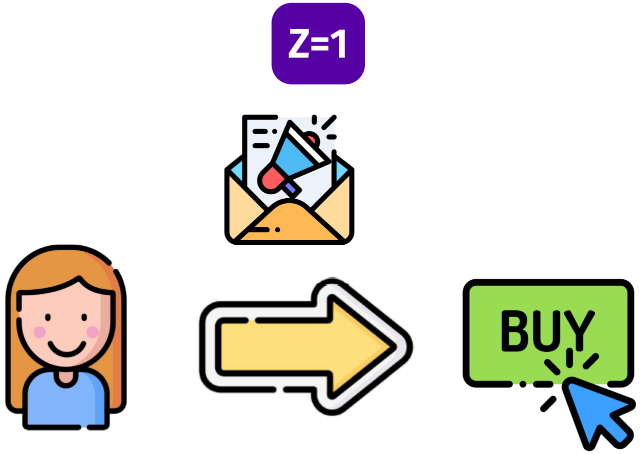
Individual Causal Effect

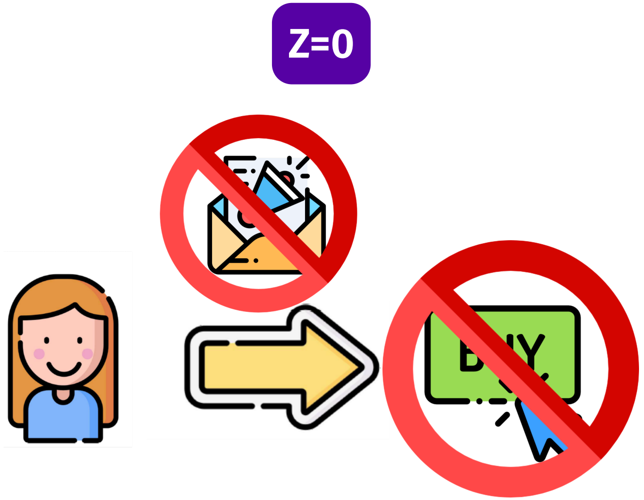
$$ICE_i = Y_i(1) - Y_i(0)$$

Can we ever observe individual causal effects?

No!*

Only one realization





The "What": Causal estimands, estimates, and estimators

Estimands vs Estimates vs Estimators

Estimand

A quantity we want to estimate

Estimator

A rule for calculating an estimate based on data

Estimate

The result of an estimation

Estimands vs Estimates vs Estimators

Estimand

A quantity we want to estimate

E.g.: Population mean

μ

Estimator

A rule for calculating an estimate based on data

E.g.: Sample mean

$$\frac{1}{n}\sum_i Y_i$$

Estimate

The result of an estimation

E.g.: Result of the sample mean for a given sample S

$\hat{\mu}$

Estimands vs Estimates vs Estimators

Source: Deng, 2022

Estimands vs Estimates vs Estimators

- Some important **estimands** that we need to keep in mind:

Average Treatment Effect (ATE)

Average Treatment Effect on the Treated (ATT)

Conditional Average Treatment Effect (CATE)

Estimands vs Estimates vs Estimators

- Some important estimands that we need to keep in mind:

ATE: E.g. Average Treatment Effect for all customers

ATT: E.g. Average Treatment Effect for customers that received the email

CATE: E.g. Average Treatmenf Effect for customer under 25 years old

Estimands vs Estimates vs Estimators

- Some important **estimands** that we need to keep in mind:

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

$$ATT = E[Y(1) - Y(0)|Z = 1]$$

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

Getting around the fundamental problem of causal inference

- Let's go back to our original example: **Does an email campaign increase sales?**

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0	?	0	?
2	1	0	0	?	?
3	1	1	1	?	?
4	0	1	?	1	?
5	0	0	?	0	?
6	1	1	1	?	?

Getting around the fundamental problem of causal inference

- We have a missing data problem

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0	?	0	?
2	1	0	0	?	?
3	1	1	1	?	?
4	0	1	?	1	?
5	0	0	?	0	?
6	1	1	1	?	?

Getting around the fundamental problem of causal inference

- Compare those who received the email to the ones did not received the email.

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0	?	0	?
2	**1**	**0**	**0**	**?**	**?**
3	**1**	**1**	**1**	**?**	**?**
4	0	1	?	1	?
5	0	0	?	0	?
6	**1**	**1**	**1**	**?**	**?**

Getting around the fundamental problem of causal inference

- Compare those who received the email to the ones did not received the email.

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0	?	0	?
2	1	0	0	?	?
3	1	1	1	?	?
4	0	1	?	1	?
5	0	0	?	0	?
6	1	1	1	?	?

Getting around the fundamental problem of causal inference

- Compare those who received the email to the ones did not received the email.

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0	?	0	?
2	1	0	0	?	?
3	1	1	1	?	?
4	0	1	?	1	?
5	0	0	?	0	?
6	1	1	1	?	?

$$\hat{\tau} = \frac{1}{3}\sum_{i \in Z=1} Y_i - \frac{1}{3}\sum_{i \in Z=0} Y_i = 0.333$$

Getting around the fundamental problem of causal inference

I we had more data, we could do the same with a simple regression:

$$Purchase = \beta_0 + \beta_1 Email + \varepsilon$$

Imagine you get the following results:

$$Purchase = 0.4 + 0.33 Email + \varepsilon$$

- Interpret the coefficient for *Email*:

What could be the problem with comparing the sample means?

Let's do a little exercise

Look at your green piece of paper and go to the following website

https://sta235h.click/week4

Would you go to a physician/urgent care?

The "Why": Causal questions and study designs

Under what assumptions is our estimate causal?

We are using:

$$\hat{\tau} = \frac{1}{3}\sum_{i \in Z=1} Y_i - \frac{1}{3}\sum_{i \in Z=0} Y_i)$$

to estimate:

$$\tau = E[Y_i(1) - Y_i(0)]$$

Under what assumptions is our estimate causal?

We are using:

$$\hat{\tau} = \frac{1}{3} \sum_{i \in Z=1} Y_i - \frac{1}{3} \sum_{i \in Z=0} Y_i)$$

to estimate:

$$\tau = E[Y_i(1) - Y_i(0)]$$

Let's do some math

Under what assumptions is our estimate causal?

$$\tau = E[Y_i(1) - Y_i(0)]$$
$$= E[Y_i(1)] - E[Y_i(0)]$$

Key assumption:

Ignorability

Ignorability means that the potential outcomes $Y(0)$ and $Y(1)$ are independent of the treatment, e.g. $(Y(0), Y(1)) \perp\!\!\!\perp Z$.

$$E[Y_i(1)|Z=0] = E[Y_i(1)|Z=1] = E[Y_i(1)]$$

and

$$E[Y_i(0)|Z=0] = E[Y_i(0)|Z=1] = E[Y_i(0)]$$

Under what assumptions is our estimate causal?

$$\tau = E[Y_i(1) - Y_i(0)]$$
$$= E[Y_i(1)] - E[Y_i(0)]$$

- Under ignorability (see previous slide), $E[Y_i(1)] = E[Y_i(1)|Z = 1] = E[Y_i|Z = 1]$ and $E[Y_i(0)] = E[Y_i(0)|Z = 0] = E[Y_i|Z = 0]$, then:

$$\tau = E[Y_i(1)] - E[Y_i(0)] = \underbrace{E[Y_i(1)|Z = 1]}_{\text{Obs. Outcome for T}} - \overbrace{E[Y_i(0)|Z = 0]}^{\text{Obs. Outcome for C}}$$

Ignorability Assumption

We can just "ignore" the missing data problem:

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0		0	
2	1	0	0		
3	1	1	1		
4	0	1		1	
5	0	0		0	
6	1	1	1		

Ignorability Assumption

We can just "ignore" the missing data problem:

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0		0	
2	1	0	0		
3	1	1	1		
4	0	1		1	
5	0	0		0	
6	1	1	1		

Ignorability Assumption

We can just "ignore" the missing data problem:

i	Z	Y	Y(1)	Y(0)	Y(1)-Y(0)
1	0	0		0	
2	1	0	0		
3	1	1	1		
4	0	1		1	
5	0	0		0	
6	1	1	1		
			2/3	1/3	

Main takeaway points

Causal Inference is hard

- Think about the causal problem
- Check validity of assumptions (*Is ignorability plausible? Am I controlling for the right covariates?*)
- Most of this chapter will be spent on looking for exogeneous variation to make the ignorability assumption happen.

Next week

- Randomized Controlled Trials:
  - Pros and Cons
  - Concept of validity
  - A/B Testing

References

- Angrist, J. & S. Pischke. (2015). "Mastering Metrics". *Chapter 1*.
- Cunningham, S. (2021). "Causal Inference: The Mixtape". *Chapter 4: Potential Outcomes Causal Model*.
- Neil, B. (2020). "Introduction to Causal Inference". *Fall 2020 Course*