Leids Universitair Medisch Centrum

Causal inference I

Week 1: Asking causal questions

Nan van Geloven

Outline

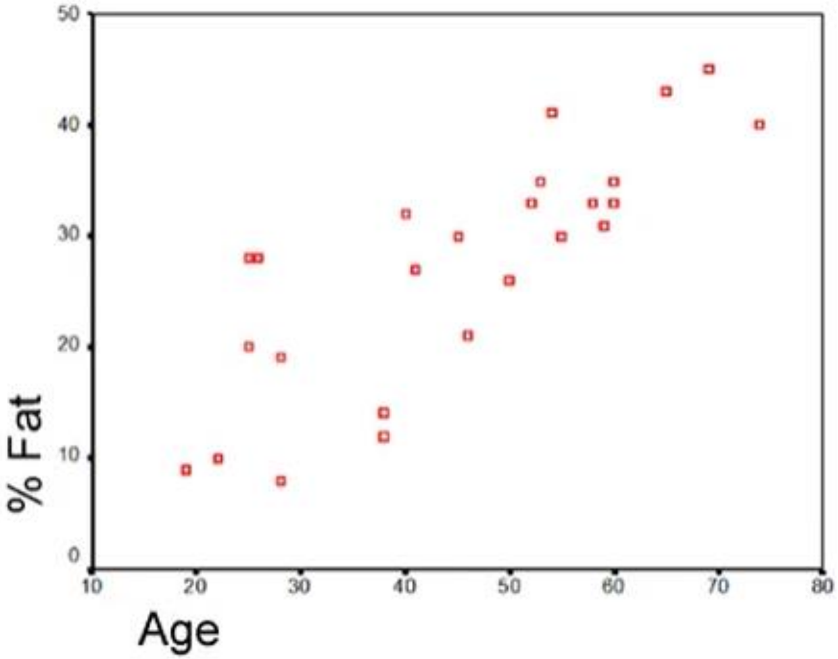
1. Recognizing causal questions
2. Formulating causal questions using the potential outcomes framework
3. Assumptions for answering causal questions with data
4. Practical exercises and assignment

Relates to Chapters 1 and 3 of the e-book r-causal.org

PART 1

Recognizing causal questions

How would you describe this relation between age and fat percentage?

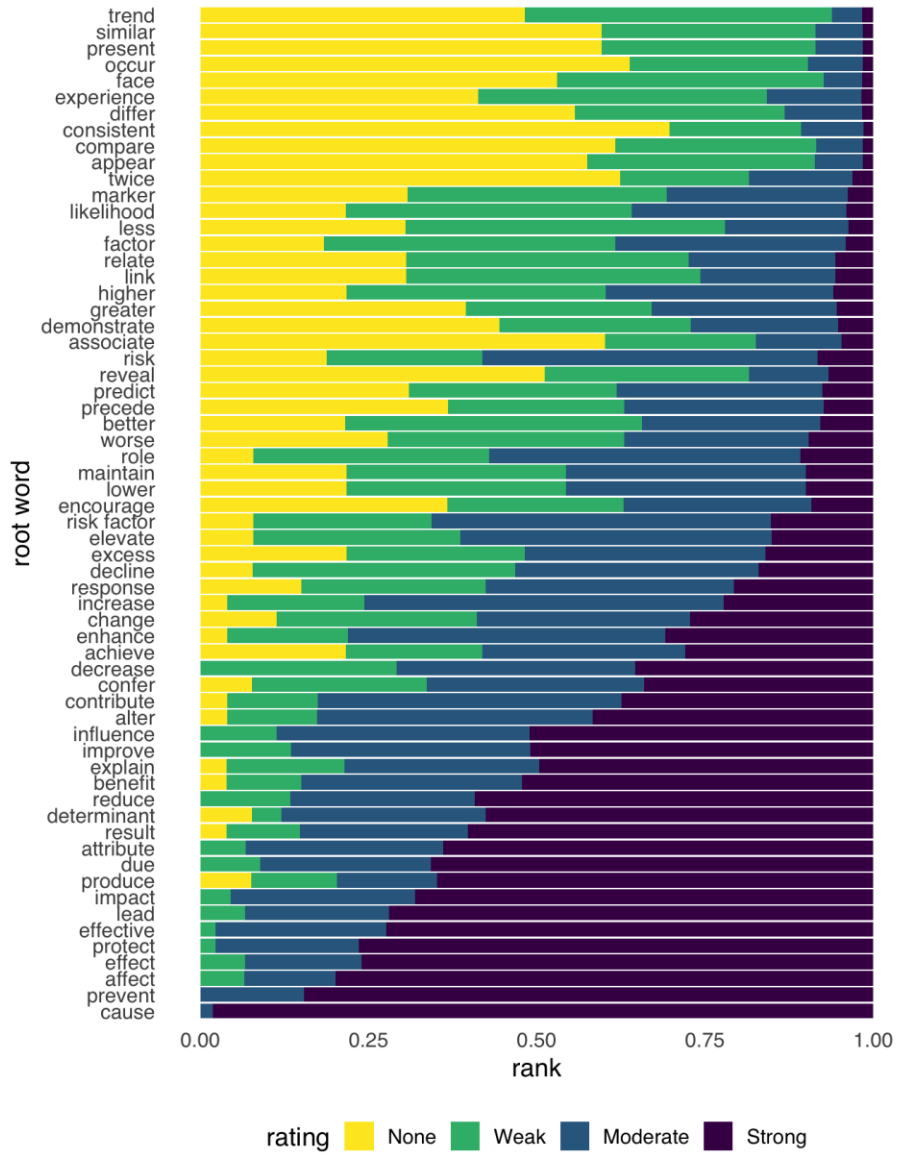


Potential answers

- %fat increases when age increases
- %fat and age are positively correlated
- Higher %fat with higher age
- Age is a predictor of %fat
- Age is a risk factor of higher %fat
- Higher age results in higher %fat
- Young age protects against high %fat
- In older persons, the %fat is expected to be higher
- There is a trend towards higher %fat with higher age
- Older persons are fatter
- etc.

- Do these wordings suggest that aging is *a cause* of increased %fat?

Do the following words suggest a causal effect?



Vagueness ("casual inference") gives confusion and noise

Haber et al 2022

Three fundamental tasks of data science

Description

Summarize trends in the data structure / who, where and when

e.g. revenues in past six months, covid infections over regions, visualize social networks; unsupervised learning

Prediction

Predict future (unseen) observations / bias-variance tradeoff, overfitting and optimism

e.g. projecting expected revenues, medical prognosis, chatGPT; supervised learning

Causation / causal explanation / causal inference

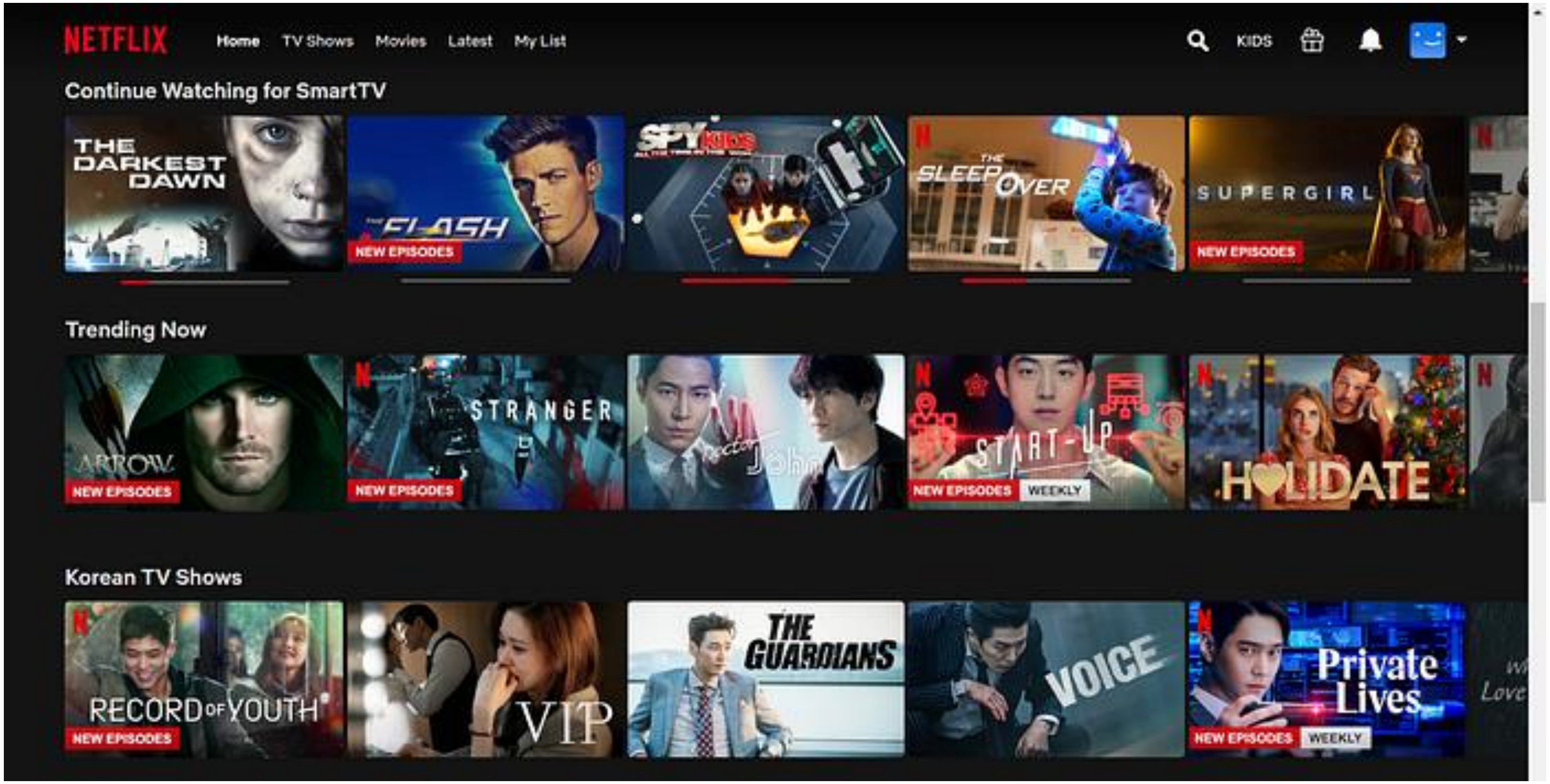
Predict what would happen **if** the world would be different

Explaining what causes what: exposure causes outcome

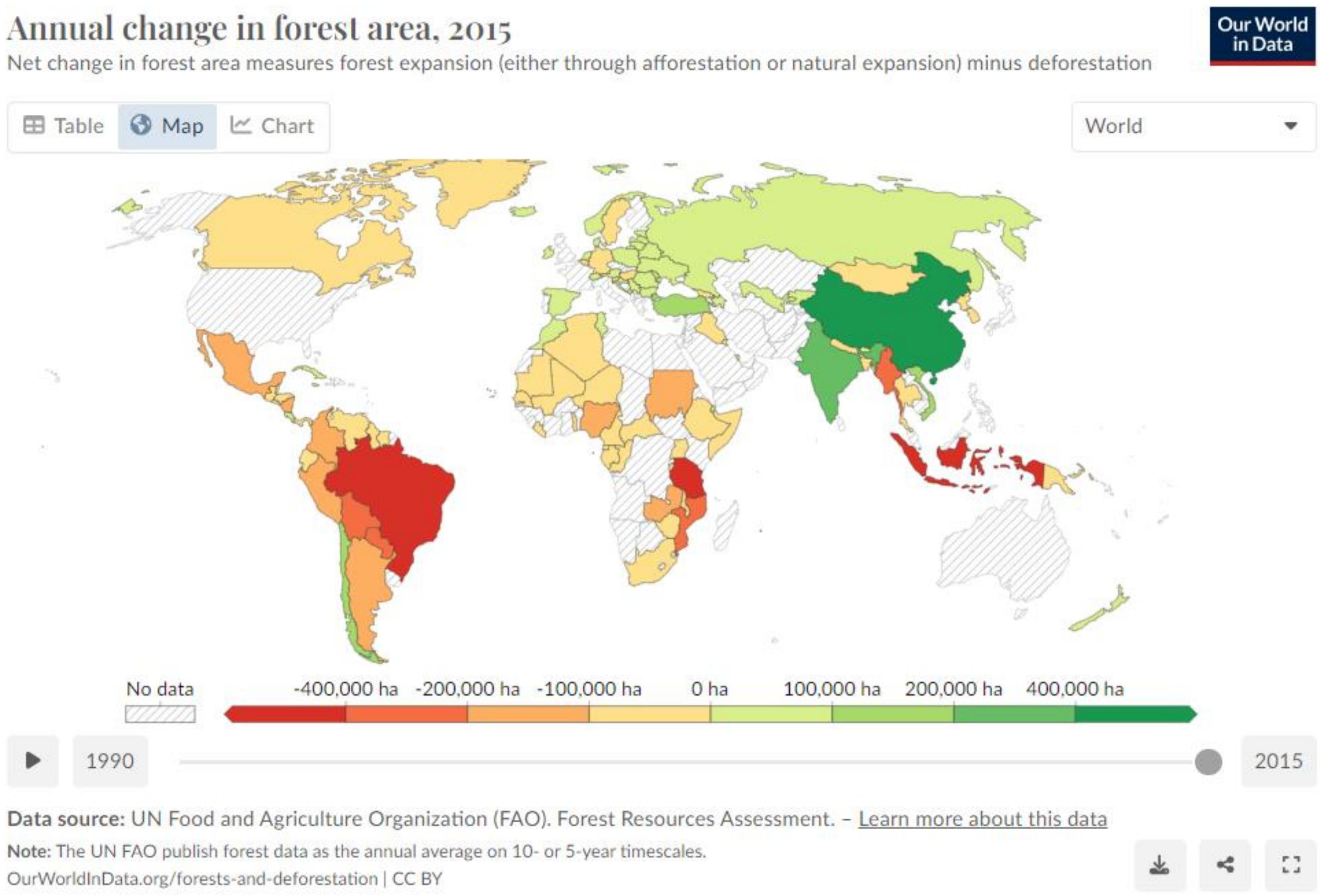
e.g. effect of a medical treatment, of a marketing strategy, of climate interventions etc.

Description, prediction or causation?

- Netflix suggest that you might like watching these series tonight



Description, prediction or causation?



Description, prediction or causation?

US districts that continued masking saw a drastically lower number of covid-19 infections than those that didn’t; almost 12,000 additional cases occurred due to the policy change

Isn't the right causal model just the best prediction model?

No!

Causal effects needn't predict particularly well, and good predictors needn't be causal

Data cannot tell us the difference!

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and prediction is common, yet the distinction must be understood for progressing scientific knowledge. While this distinction has been recognized in the philosophy of science, the statistical literature lacks a thorough discussion of the many differences that arise in the process of modeling for an explanatory versus a predictive goal. The purpose of this article is to clarify the distinction between explanatory and predictive modeling, to discuss its sources, and to reveal the practical implications of the distinction to each step in the modeling process.

Key words and phrases: Explanatory modeling, causality, predictive modeling, predictive power, statistical strategy, data mining, scientific research.

Shmueli, Statistical Science 2010, DOI: 10.1214/10-STS330

What is a causal effect?

Suppose you wake up in the morning with a headache….

… you take two tablets of paracetamol…

… after two hours, your headache has disappeared.

Was this due to the paracetamol?

What is a causal effect?

Suppose you visit your doctor …

… she measures your blood pressure, which is too high …

… she proposes to start antihypertensive medication …

… six weeks later, you return to the doctor’s office …

… and your blood pressure has normalized.

Was this due to the antihypertensive medication?

What is a causal effect?

Suppose there is a large-scale pandemic…

… with a highly infectious virus …

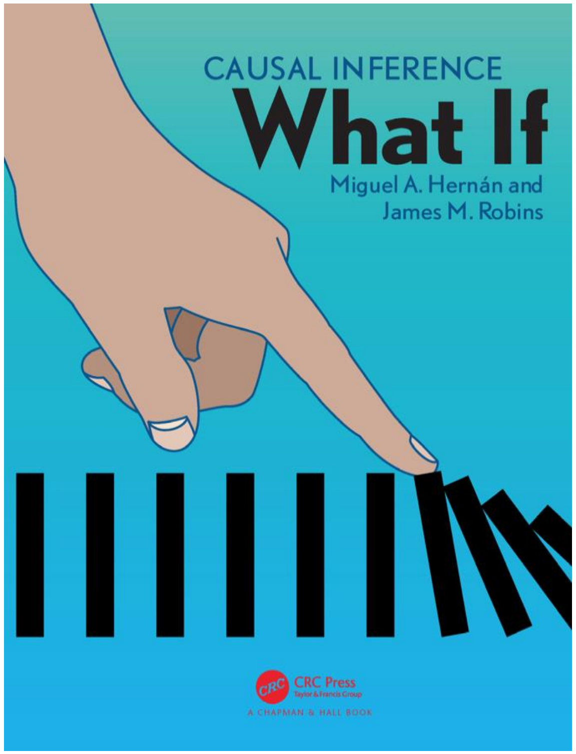
… and governments decide to lockdown schools and universities …

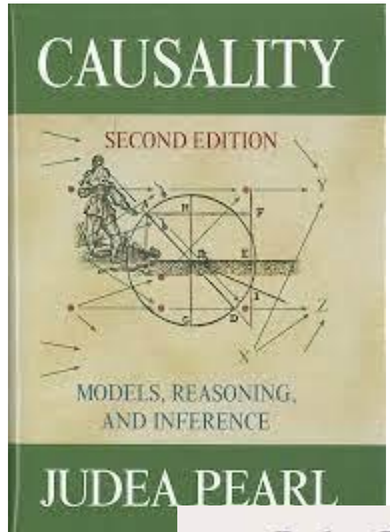
… after which they see observe that the number of patients admitted to the hospital gradually goes down …

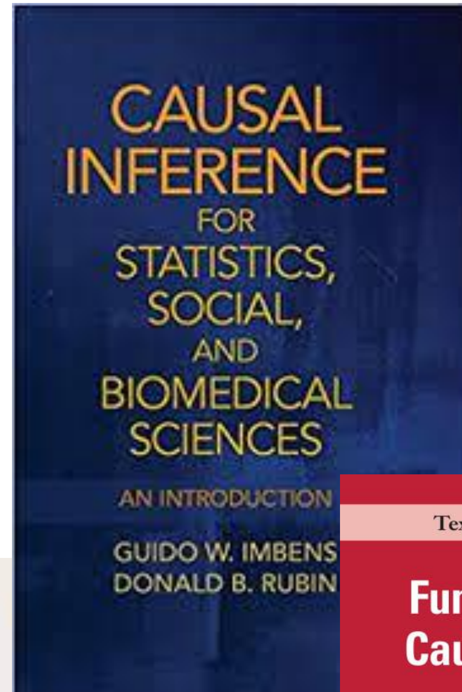
Was this due to the lockdown?

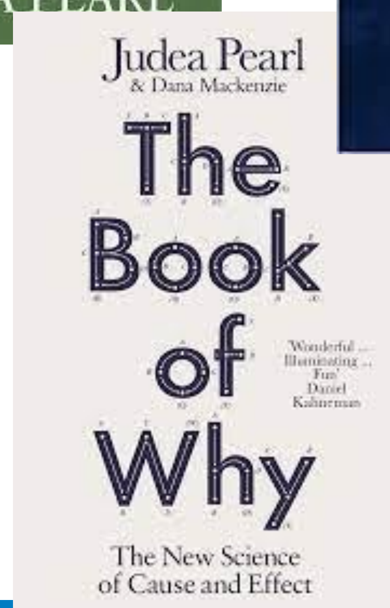
PART 2

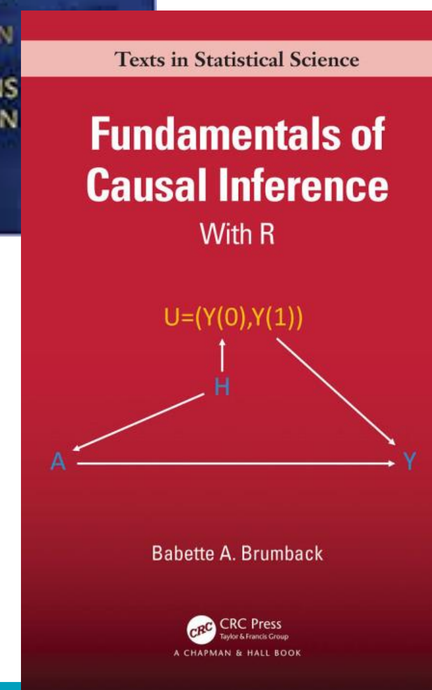
Formulating causal questions using Potential outcomes framework


CAUSAL INFERENCE
What If
Miguel A. Hernán and
James M. Robins
CRC Press
A CHAPMAN & HALL BOOK


CAUSALITY
SECOND EDITION
MODELS, REASONING,
AND INFERENCE
JUDEA PEARL


CAUSAL
INFERENCE
FOR
STATISTICS,
SOCIAL,
AND
BIOMEDICAL
SCIENCES
AN INTRODUCTION
GUIDO W. IMBENS
DONALD B. RUBIN


Judea Pearl
& Dana Mackenzie
The
Book
of
Why
'Wonderful ... Illuminating ... Fun'
Daniel Kahneman
The New Science
of Cause and Effect


Texts in Statistical Science
Fundamentals of
Causal Inference
With R
U=(Y(0),Y(1))
H
A
Y
Babette A. Brumback
CRC Press
A CHAPMAN & HALL BOOK

Notation (r-causal.org)

Y	observed outcome (dependent variable), e.g. Y=1
X	observed exposure, e.g., X=1 exposed, X=0 not exposed
Y(1),Y(0)	potential outcomes:

Y(1) outcome that would have been observed under the exposure value 1

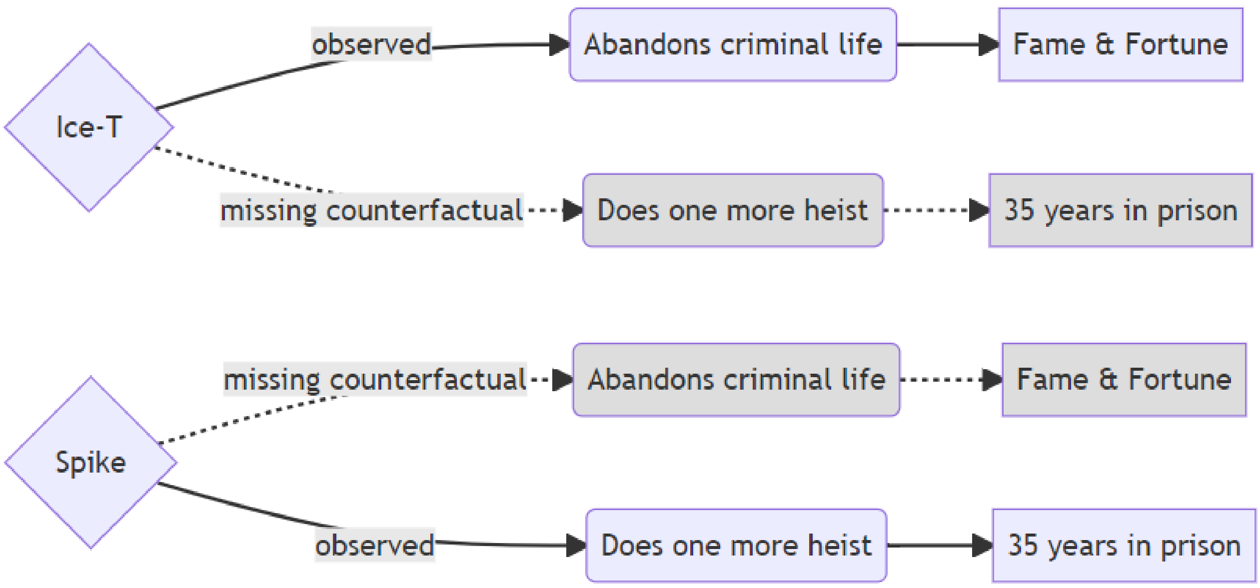
Y(0) outcome that would have been observed under the exposure value 0

Note: in real data we only observe an individual’s outcome under one exposure!

The other potential outcome is “missing” (or counterfactual)

Note: no consistent use of notation in throughout literature, always check definitions at start

‘Almost counterfactuals’ (r-causal.org)



Ice-T and Spike Causal Map

Causal effects can be expressed using potential outcomes

Causal effect exists for an individual if: $Y(1) \neq Y(0)$

"Fundamental problem of causal inference"

Only one potential outcome can be observed -> **individual** causal effects **cannot be identified**, i.e., they cannot be expressed as a function of the observed data

-> focus on *average* causal effects

In a defined population Y(1), Y(0) are random variables:

Average causal effect exists in the population if: $E\big(Y(1)\big) \neq \mathrm{E}(Y(0))$

Average causal effects **can sometimes be identified** from data

Expressing the causal contrast of interest (estimand)

For binary outcome common quantifications are:

Pr is ?

Risk difference: $E\big(Y(1)\big) - E\big(Y(0)\big) = \Pr(Y(1) = 1) - \Pr(Y(0) = 1)$

Risk ratio: $E\big(Y(1)\big)/E\big(Y(0)\big) = \Pr(Y(1) = 1)\,/\,\Pr(Y(0) = 1)$

Note that the observed risk for persons exposed to a certain level may be different than the counterfactual risk if all would be observed to that level:

$\Pr(Y(1) = 1)$ is in general not equal to $\Pr(Y = 1|X = 1)$

Eg: 1-year mortality risk if all Dutch inhabitants would get open heart surgery versus the observed risk for patients actually receiving such a surgery last year

From question to potential outcomes

Casual question: “Should I do a warming-up before starting exercising?”

More specific: ”Does performing a warming-up (at least 10 minutes cardio workout) decrease the risk of getting a sports injury during exercising, evaluated in students following the course Causal Inference I?”

Notation

X: the observed exposure: did person first do warming up?

Y: the outcome: sports injury 1=yes, 0=no

Estimand

$$\Pr(Y(1) = 1) - \Pr(Y(0) = 1)$$

“the difference in risk of having a sports injury if all students would do a warming up versus if they would all not do a warming up”

Estimands, estimators, and estimates (1)

被估计量

Estimand: the target quantity (expressed in potential outcomes)

估计量

Estimator: a function of the observed data

估计值

Estimate: a value of an estimator

Estimands, estimators, and estimates (2)

ESTIMAND	ESTIMATOR	ESTIMATE
What you seek	How you will get there	What you get

PART 3

Assumptions for answering causal questions with data

Identifiability

What do we need to estimate causal estimands with data?

(Strict definition: The distribution of the observed variables is compatible with exactly one value of the estimand.)

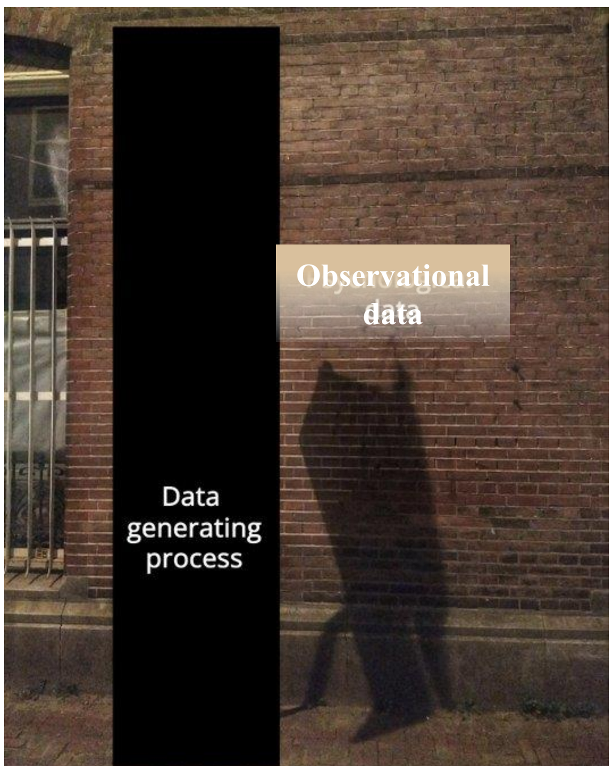


Photo by Eiko Fried

可识别性
Identifiability conditions

Conditions[1] that guarantee identifiability.

Consistency: If $X = x \rightarrow Y(x) = Y$, for all x
equivalent for binary X*:* $Y = (X)Y(1) + (1 - X)Y(0)$

Exchangeability: X is independent of $Y(x)$, for all x

Positivity: $\Pr(X = x) > 0$, for all x

Under these conditions $\Pr(Y(x) = 1) = \Pr(Y = 1 \mid X = x)$, for all x.

We will first discuss each condition and then show the equality
Note: in later weeks we will extend the exchangeability and positivity assumptions to conditional versions

[1] commonly used, but other sets of conditions also exist

Consistency: If $X = x \rightarrow Y(x) = Y$, for all x

The outcome you observe is exactly equal to the potential outcome under the exposure you received. This requires:

1. **Well defined exposure**: for each value of the exposure, there is no difference between subjects in the delivery of that exposure, i.e., multiple versions/flavors of the treatment do not exist

2. **No interference**: the outcome for any subject does not depend on another subject's exposure

"one needs to be able to explain how a certain level of exposure could be hypothetically assigned to a person exposed to a different level" *(Cole, Epidemiology 2009)*

Well-defined interventions

International Journal of Obesity (2008) 32, S8–S14

0307-0565/08 $30.00
www.nature.com/ijo

ORIGINAL ARTICLE

Does obesity shorten life? The importance of well-defined interventions to answer causal questions

MA Hernán[1,2] and SL Taubman[1,3]

过度肥胖 死亡

Can we estimate the effect of obesity on mortality?

Think of 3 ideal trials, sample size 1000.000, 30 years

Trial 1: ‘1 hour exercise/day’ vs ‘activity as usual’ 100,000 deaths prevented

Trial 2: ‘stringent diet’ vs ‘diet as usual’ 50,000 deaths prevented

Trial 3 ‘30 min exercise + dietary advise’ vs as usual 120,000 deaths prevented

3 trials had same BMI distribution during entire follow up

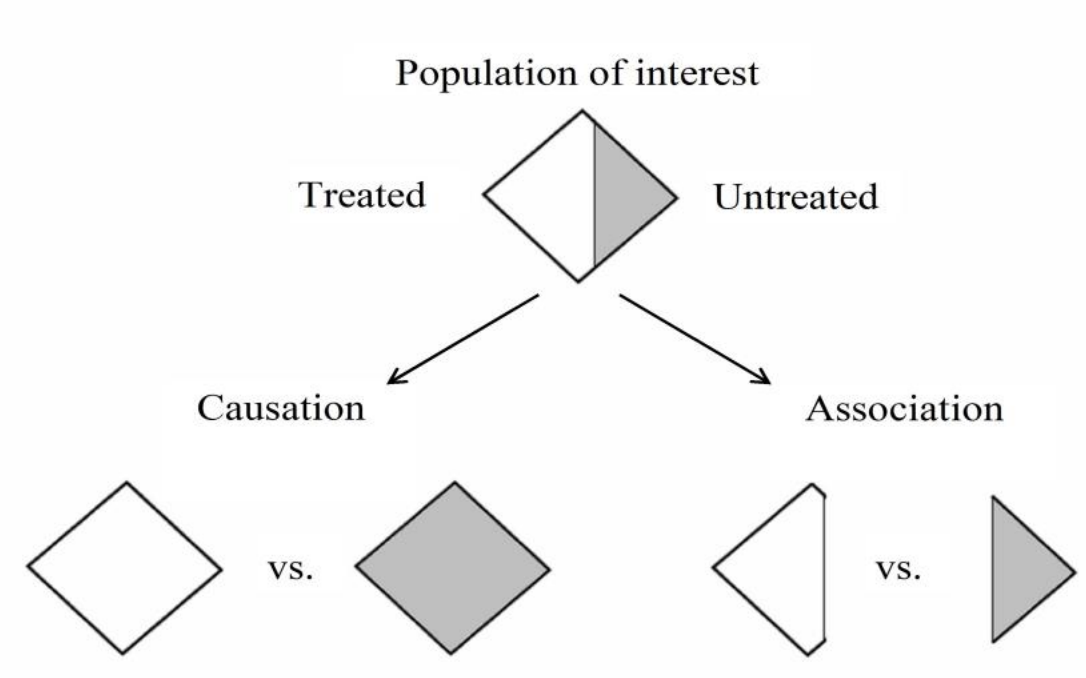
‘no causation without manipulation’

How about effect of ethnicity/race on health outcomes?

Exchangeability: : X is independent of $Y(x)$

"this means that the exposed and the unexposed would have experienced the same risk of death if they had received the same exposure level"

Also: $Y(x \mid X = 1) = Y(x \mid X = 0) = Y(x)$, for $x = 1$ and $x = 0$



Positivity

all exposure in dataset

Positivity: $\Pr(X = x) > 0$ for all x

- Why do we need positivity?
- What if there isn't positivity?
- Random non-positivity vs. structural non-positivity

Estimand & estimator

Identifiability conditions needed to link estimand to estimator:

estimand

$$\boxed{\Pr(Y(1))} = \Pr(Y(1) \mid X = 1) \quad \text{by exchangeability}$$

$$= \Pr(Y \mid X = 1) \quad \text{by consistency}$$

$$\boxed{= \frac{\sum (Y * 1_{\{X=1\}})}{\sum 1_{\{X=1\}}}} \quad \text{by positivity}$$

estimator

PART 4

Practical exercises and assignment