

WE'VE DESIGNED A DOUBLE-BLIND TRIAL TO TEST THE EFFECT OF SEXUAL ACTIVITY ON CARDIOVASCULAR HEALTH. BOTH GROUPS WILL *THINK* THEY'RE HAVING LOTS OF SEX, BUT ONE GROUP WILL ACTUALLY BE GETTING SUGAR PILLS.



THE LIMITATIONS OF BLIND TRIALS

## Randomization and a Philosophy of Causation

# Today

- Discussion of the reading
- Philosophical and historical perspectives on causality
- Notation of counterfactual random variables
- Synthetic data example
- Rough schedule for the course

# Questions from the reading

- Can you be sure your data is causal and not associational?
- Can we only approximate a causation-proving experiment?
- How do we determine all the confounding variables?
- Why do studies claim to find links or associations that seem to imply causation without using the word? Are academic papers and editors publishing and pushing more sensationalizable papers for media exposure and more citations?

## CampusWire question

<i>% who recover from migraines</i>	<b>Drug C</b>	<b>Drug D</b>
<b>Younger patients</b>	93% (81/87)	87% (234/270)
<b>Older patients</b>	73% (192/263)	69% (55/80)
<b>Both</b>	78% (273/350)	83% (289/350)

Why here do we focus only on the individual data for this table and not the combined percentages as well? Like would we always look at either the individual data or the combined due to Simpson's paradox and if so do we focus on individual data for confounding variables and combined data mediator variables?

# Historical perspectives

- David Hume, 1748
  - Regularity theory of causation
  - “We may define a cause to be an object, followed by another, and where all objects similar to the first, are followed by objects similar to the second.”

# Historical perspectives

- Hans Reichenbach, 1956
  - “Probability-raising Theories of Causation”
  - A binary variable A is a cause of binary variable Y if

$$p(Y=1 \mid A=1) > p(Y=1 \mid A=0)$$

- But this implies:

$$p(A=1 \mid Y=1) > p(A=1 \mid Y=0)$$

# Historical perspectives

- David Lewis, 1973:
  - “We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it. Had it been absent, its effects – some of them, at least, and usually all – would have been absent as well”

# Historical perspectives

- Cook and Campbell, 1979:
  - Three requirements for determining cause and effect
    1. Changes in the presumed cause must be related to changes in the presumed effect
    2. The presumed cause must occur before the presumed effect
    3. The presumed cause must be the only reasonable explanation for changes in the outcome measures.

Cook, T.D. and Campbell, D.T. (1979). Quasi-Experimentation: Design and Analysis for Field Settings

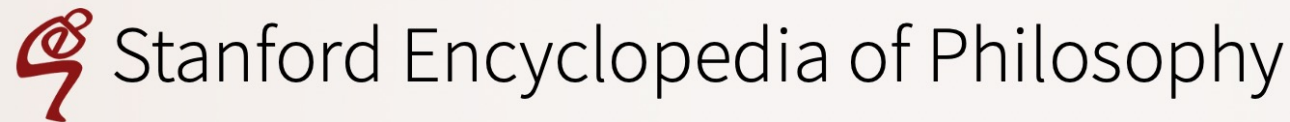





# A working definition

- A causes Y if, when we randomly assign A,

$$E[Y \mid A=1] \neq E[Y \mid A=0]$$

# More historical perspectives



 Browse  About  Support SEP

Search SEP



Entry Contents

Bibliography

Academic Tools

## Counterfactual Theories of Causation

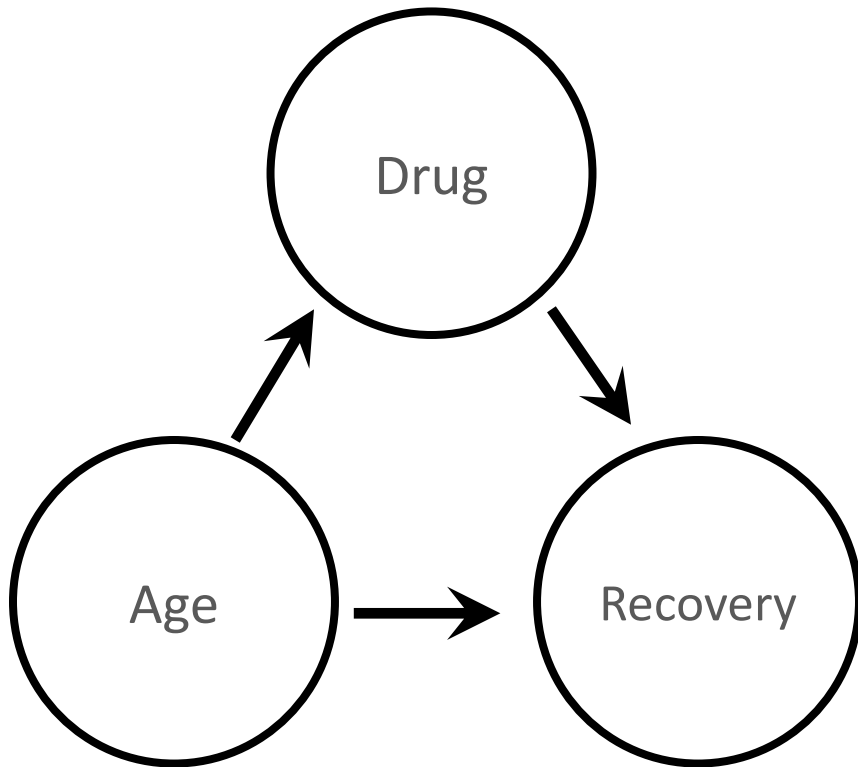
*First published Wed Jan 10, 2001; substantive revision Tue Oct 29, 2019*

## Probabilistic Causation

*First published Fri Jul 11, 1997; substantive revision Fri Mar 9, 2018*

<https://plato.stanford.edu/>

## Causal assumptions



<i>% who recover from migraines</i>	<b>Drug C</b>	<b>Drug D</b>
<b>Younger patients</b>	93% (81/87)	87% (234/270)
<b>Older patients</b>	73% (192/263)	69% (55/80)
<b>Both</b>	78% (273/350)	83% (289/350)

# Counterfactual random variables

ID	Age	Drug	Recover (C)	Recover (D)
1	Old	C	Yes	<i>No</i>
2	Young	C	Yes	<i>No</i>
3	Young	C	No	<i>No</i>
4	Young	D	<i>Yes</i>	Yes
5	Old	D	<i>No</i>	No

# Counterfactual random variables

Suppose we have three random variables

- Treatment **A**, Confounder **C**, Outcome **Y**
- In observational data, nature assigns:
  - **C** marginally,
  - **A** conditional on **C**,
  - **Y** conditional on **A** and **C**.
- In randomized data, the experimenter assigns **A**
- *Just as before*, nature assigns:
  - **C** marginally
  - **Y** conditional on **A** and **C**.

Is  $E[Y|A]$  the same?  $E[Y | A, C]$ ?

# Counterfactual random variables

- Suppose we could see the future, and let  $Y^{A=a}$  be the random variable, “what would  $Y$  have been had we randomly assigned  $A$  to value  $a$ ”
- Then under randomization,  $E[Y \mid A] = E[Y^A \mid A]$

# Counterfactual random variables

ID	C	A	$Y^{A=1}$	$Y^{A=0}$
1	1	1	1	0
2	0	1	1	0
3	0	1	0	0
4	0	0	1	1
5	1	0	0	0

# Synthetic data example



**C**: result of a  $k$ -sided die roll

Flip  $1 + k - C$  coins. Then

**A**: 1 if at least one head  
0 otherwise

Flip  $C + A$  coins. Then **Y** is the total number of heads



# Synthetic data example

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as smf

np.random.seed(42)
```

# Synthetic data example

```
1 def observed(n=100, c_dim=7, ols="y ~ a") :  
2     c = np.random.randint(1, c_dim, n)  
3     a = np.random.binomial(n=c_dim - c, p=0.5, size=n)  
4     a = (a > 0).astype(np.int32)  
5     y = np.random.binomial(n=a + c, p=0.5)  
6     df = pd.DataFrame(data=dict(c=c, a=a, y=y))  
7     print(smf.ols(ols, data=df).fit().params['a'])
```

# Synthetic data example

```
1 def randomized(n=100, c_dim=7, ols="y ~ a") :  
2     c = np.random.randint(1, c_dim, n)  
3     a = np.random.binomial(n=1, p=0.5, size=n)  
4     y = np.random.binomial(n=a + c, p=0.5)  
5     df = pd.DataFrame(data=dict(c=c, a=a, y=y))  
6     print(smf.ols(ols, data=df).fit().params['a'])
```

# Tentative schedule: Weeks 1-3

1. Motivating causal inference
  - Simpson's paradox
  - Counterfactuals
  - Randomized experiments
2. Review of fundamentals
  - Probability and statistics
  - Graphical models and conditional independence
3. Basic methods in causal inference
  - Connecting potential outcomes to observational data
  - Simple confounding and identification
  - **Project proposals due**

# Tentative schedule: Weeks 4-6

4. Estimators of causal effects
  - Outcome models
  - Propensity score models
5. Unmeasured confounding and identification
6. Structure learning
  - Testing for conditional independences
  - PC and GES Algorithms
  - **Project update due**

## Tentative schedule: Weeks 7-10

7. Missing data
8. Measurement error and proxies
  - **Project presentations**
9. Selection bias and case-control studies
10. Evaluating causal claims in scientific literature
  - **Presentation feedback due**
11. Project report due on day of final

# Final projects

1. Pick a dataset
  - What is the causal question you're interested in?
  - Does this dataset contain enough data to answer it?
2. Pick a graphical model that describes the data
  - Use domain knowledge and data-driven methods
3. Identification
  - Theoretical justification for why you can answer your question
4. Estimation
  - Use a method to compute the effect from data
5. Additional methods: missing data, measurement error, etc
6. Analysis of the results

# For Monday

- Read Chapters 1 and 2 of Hernan and Robins “What If”
  - Including Fine point 1.2
  - Skipping Fine points 1.1, 1.3, Technical points 1.1, 1.2
- Review basics of probability (as much as needed)
  - Chapter 2 of Murphy’s “Machine Learning: Probabilistic Perspective” posted to Canvas
  - Chapter 1 of Wasserman’s “All of Statistics”