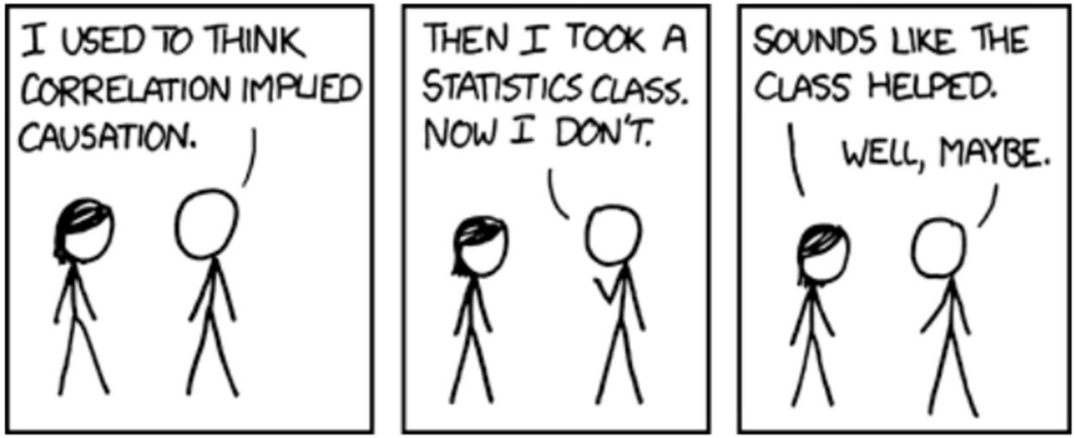
CAUSAL (NOT CASUAL) INFERENCE

An Introduction

BDSI 2022; University of Michigan

TODAY'S GOALS

- Learn how to formalize causal effects using **potential outcomes language**
- Understand key assumptions to infer average treatment effect using randomized trial and/or observational data
- Know where to look for more information on causal inference



WHAT IS CAUSAL INFERENCE?

- Causal inference ≈ Causal language/model + Statistical inference
- Inferring effects of any treatment/policy/intervention/etc.
- Suppose we infer statistically significant correlation between an outcome and a treatment
  - Key question: is this a causal relation?
  - Without causal language/model, we cannot perform causal inference
- Causal inference allows us to say when correlation **implies** causation

EX. 1: SMOKING AND LUNG CANCER

- By 1940s, lung cancer cases had tripled over the past 3 decades
- The cause for increase was unclear and a source of debate
  - Aging population
  - Clinical awareness
  - Air quality (automobile introduction)
  - Smoking
- A series of observational studies in the 1940s to assess potential link between cancer and smoking

EX. 1: SMOKING AND LUNG CANCER

- One study observed 36,975 pairs of heavy smoker and nonsmokers matched by age, race, nativity, rural versus urban residence, occupational exposures to dust and fumes, religion, education, marital status, …
- Of the 36,975 pairs, there were 1222 discordant pairs in which exactly one person died of lung cancer
- Among them:
  - 12 pairs in which nonsmoker died of lung cancer;
  - 110 pairs in which smoker died of lung cancer.
- Implies strong association between smoking and lung cancer.

EX. 1: SMOKING AND LUNG CANCER

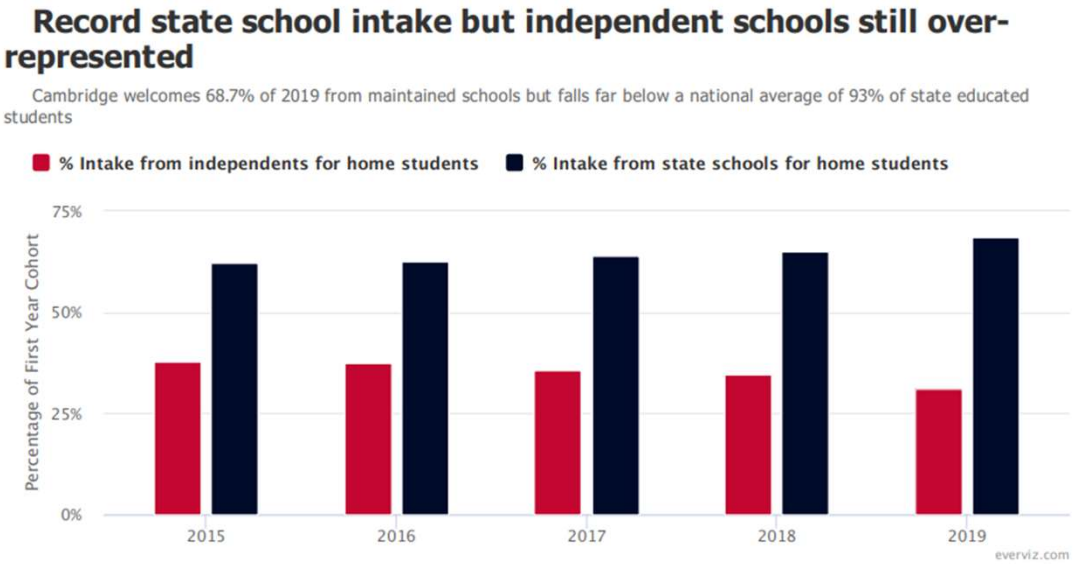
- But `nature is tricky' so can we be sure the association is causal?
- Most famous critic of this series of work was R.A. Fisher (the founder of modern statistics)

Such results suggest that an error has been made of an old kind, in arguing from correlation to causation.... Such differences in genetic make-up between those classes would naturally be associated with differences of disease incidence without the disease being causally connected with smoking.

- In 1959, Cornfield demonstrated that Fisher's genetic link would need to be biologically implausibly strong to account for the difference in risk between smokers and nonsmokers.
- Cornfield's argument has blossomed into an important aspect of causal inference called sensitivity analysis

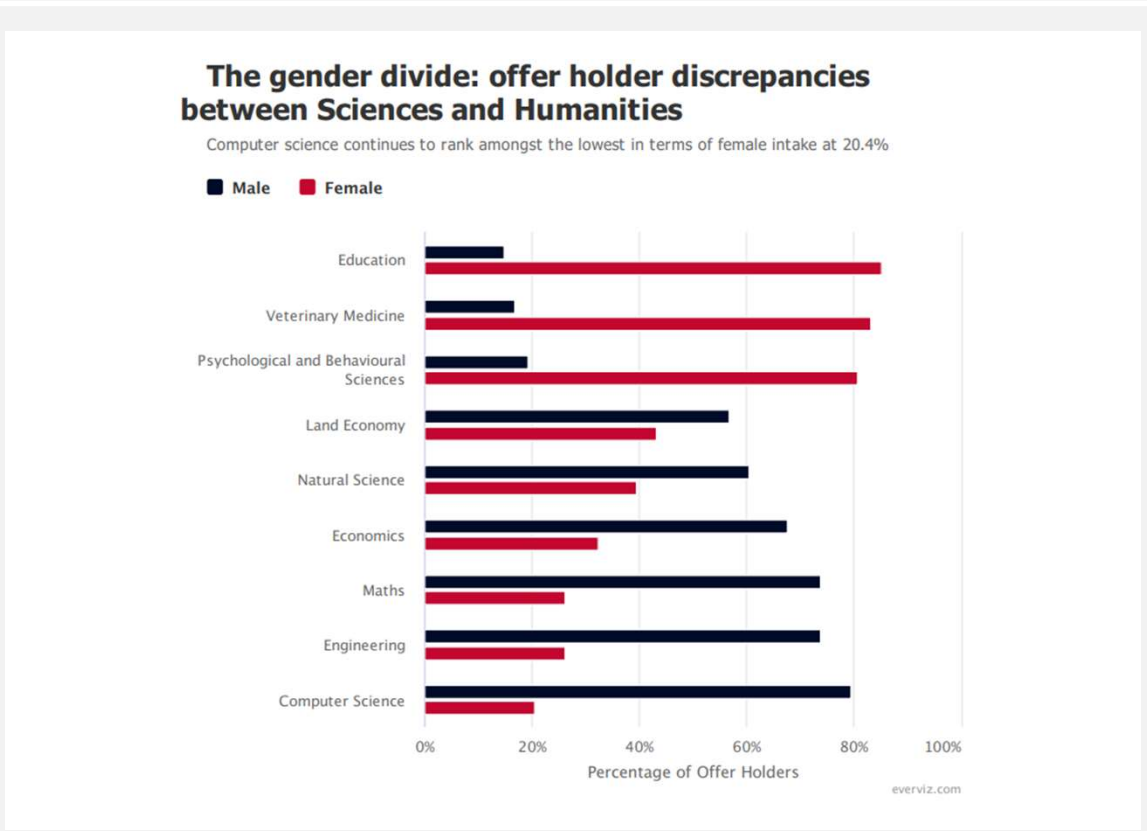
EX II: CAMBRIDGE ADMISSION DATA

- 93% of pupils in England are taught in state schools
- 90% of university students on average hailing from state schools across the country
- Does this mean Cambridge's admission is biased against state schools?
- Not necessarily. For example, applicants from independent schools may have better A-level results.
- Causal inference can be used to understand fairness in decisions made by human and computer algorithms



EX III: RACIAL DISPARITIES IN ACCEPTANCE RATES

- 1973 admission figures showed that men applying were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance
- However, when examining the individual departments, it appeared that six out of 85 departments were significantly biased against men, whereas four were significantly biased against women.
- Bickel et al. concluded that women tended to apply to more competitive departments with low rates of admission, even among qualified applicants (English department), whereas men tended to apply to less competitive departments with high rates of admission (Engineering department)



LANGUAGES OF CAUSALITY

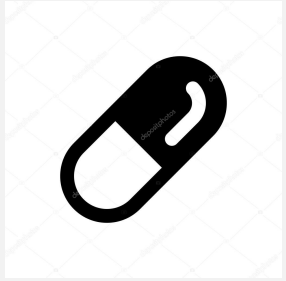
- Causal inference ≈ Causal language/model + Statistical inference
- Three typical *languages* to perform causal inference
- Using **potential outcomes/counterfactuals**
  - 1923 Neyman (statistics); 1973 Lewis (philosophy); 1974 Rubin (statistics); 1986 Robins (epidemiology);
- Using **graphs**
  - 1921 Wright (genetics); 1988 Pearl (computer science "AI"); 1993 Spirtes, Glymour, Scheines (philosophy).
- Using **structural equations**
  - 1921 Wright (genetics); 1943 Haavelmo (econometrics); 1975 Duncan (social sciences); 2000 Pearl (computer science).

EQUIVALENCE OF LANGUAGES

- **Counterfactuals:**
  - Easy to incorporate additional assumptions;
  - Elucidation of the meaning of statistical inference;
  - Not as convenient if system is complex.
- **Graphs:**
  - Easy to visualise the causal assumptions;
  - Difficult for statistical inference because model is nonparametric.
- **Structural equations:**
  - Bridge between graphs and counterfactuals;
  - Easy to operationalise;
  - Danger to be confused with regressions.

LANGUAGE 1: POTENTIAL OUTCOMES

Joe



LANGUAGE 1: POTENTIAL OUTCOMES

- Consider the ith patient out of a set of patients (i.e., statistical units)
- Assume a binary treatment per patient:
  - $\mathrm{A_i} = 1$ if the participant takes the pill
  - $\mathrm{A_i} = 0$ if the participant does not take the pill
- The patient's **potential outcomes** are the pair of outcomes one would observe on that patient if assigned to each of the two treatment options
  - $\mathrm{Y_i}(1)$: potential outcome if the participant takes the pill
  - $\mathrm{Y_i}(0)$: potential outcome score if the participant **does not** take the pill
- **Individual causal effect:** $Y_i(1) - Y_i(0)$

POTENTIAL OUTCOMES

Subject	$Y(1)$	$Y(0)$
Joe	0	1
Mary	1	0
Sally	0	0
Bob	1	1
Wendy	1	0
Sal	0	1

- The proportion of individuals that would have improved their outcome had all population individuals received treatment $a = 1$ is $\Pr(Y(1) = 1) = 3/6$
- The proportion of individuals that would have improved their outcome had all population individuals received treatment $a = 0$ is $\Pr(Y(0) = 1) = 3/6$

FACTUAL AND COUNTERFACTUAL

$do(A_i = 1)$

Factual

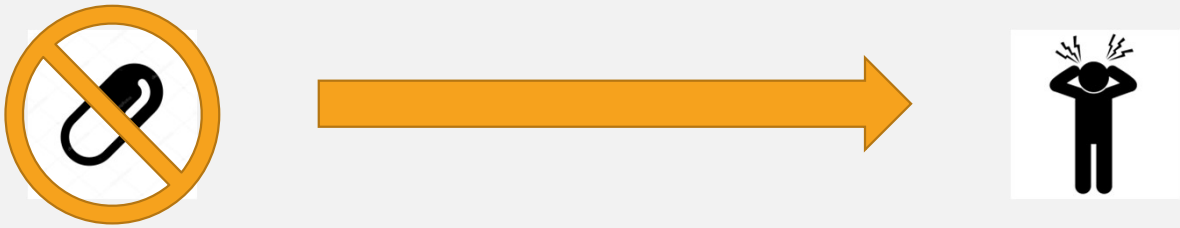
Counterfactual

FACTUAL AND COUNTERFACTUAL

Counterfactual

$do(A_i = 0)$

Factual

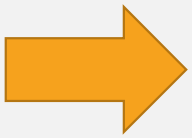


FUNDAMENTAL PROBLEM OF CAUSAL INFERENCE

- We only can observe each patient under treatment **OR** under control
- If individual treatment effects are not estimable, is there anything we can say?
- **Solution:** Consider the **average effect** across the population

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

- In our simple example, we have 6 individuals in our population and
  - $E[Y(1)] = \frac{1}{6}(0 + 1 + 0 + 1 + 1 + 0) = \frac{3}{6}$
  - $E[Y(0)] = \frac{1}{6}(1 + 0 + 0 + 1 + 0 + 1) = \frac{3}{6}$



$$ATE = \frac{3}{6} - \frac{3}{6} = 0$$

MISSING DATA PERSPECTIVE

Subject	A	Y	$Y(1)$	$Y(0)$
Joe	1	0	0	?
Mary	0	0	?	0
Sally	1	0	0	?
Bob	1	1	1	?
Wendy	0	0	?	0
Sal	0	1	?	1

RANDOMIZED CONTROLLED TRIAL

- Suppose there are **n** units in the experiment (ex. N = 6)
- Flip a coin for each patient to determine what treatment they receive
- Observe the outcome based on assigned treatment
- **If** the assigned treatment is taken (compliance) **and** the observed outcome is equal to the potential outcome under the assignment (consistency), then we can estimate the ATE
  - Joe assigned $\mathrm{a} = 1$ by coin flip then we observe $Y := Y(1)$ under these assumptions

RANDOMIZED CONTROLLED TRIAL

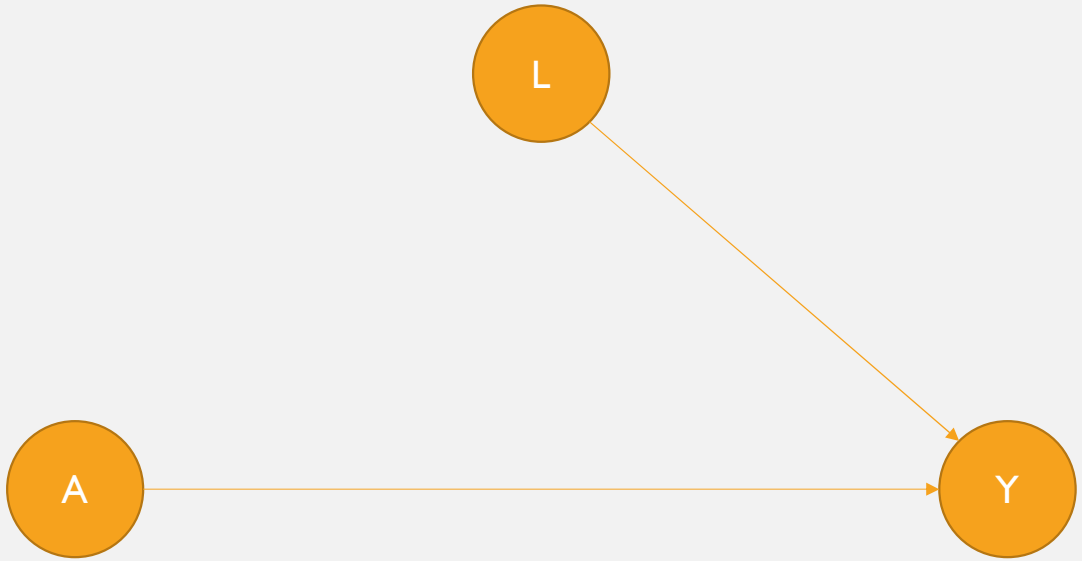
- **Stable unit treatment value assignment (SUTVA)**
  - No interference between units: treatment assignment of one unit does not affect potential outcomes of another unit.
- Under these assumptions, for a simple RCT, we have $E[Y(1)] = E[Y|A = 1]$
- So ATE can be estimated by difference-in-means estimator (Neyman):

$$\widehat{ATE} = \bar{Y}_1 - \bar{Y}_0 = \frac{\sum_i A_i Y_i}{\sum_i A_i} - \frac{\sum_i (1 - A_i) Y_i}{\sum_i 1 - A_i}$$

RCTS ARE MAGIC

- Simple randomization ensures the treatment assignment is independent of the potential outcome distribution
- In more complex studies, the randomization may be covariate dependent
- Comparability and covariate balance
  - Treatment and control groups are the same in all aspects except treatment
  - The distribution of covariates X is the same across treatment groups
- Exchangeable
- No backdoor paths

LANGUAGE 2: CAUSAL GRAPHS



A SLIGHTLY MORE COMPLEX SETTING

Subject	L	A	Y	$Y(1)$	$Y(0)$
Joe	1	1	0	0	?
Mary	1	0	0	?	0
Sally	1	1	0	0	?
Bob	0	1	1	1	?
Wendy	0	0	0	?	0
Sal	0	0	1	?	1

A SLIGHTLY MORE COMPLEX SETTING

- Then this is like having 2 simple coin flip trials that we must combine

$$E[Y(1) - Y(0)] = E\big[E[Y(1) - Y(0)|L]\big] = E\big[E[Y(1) - Y(0)|L]\big]$$

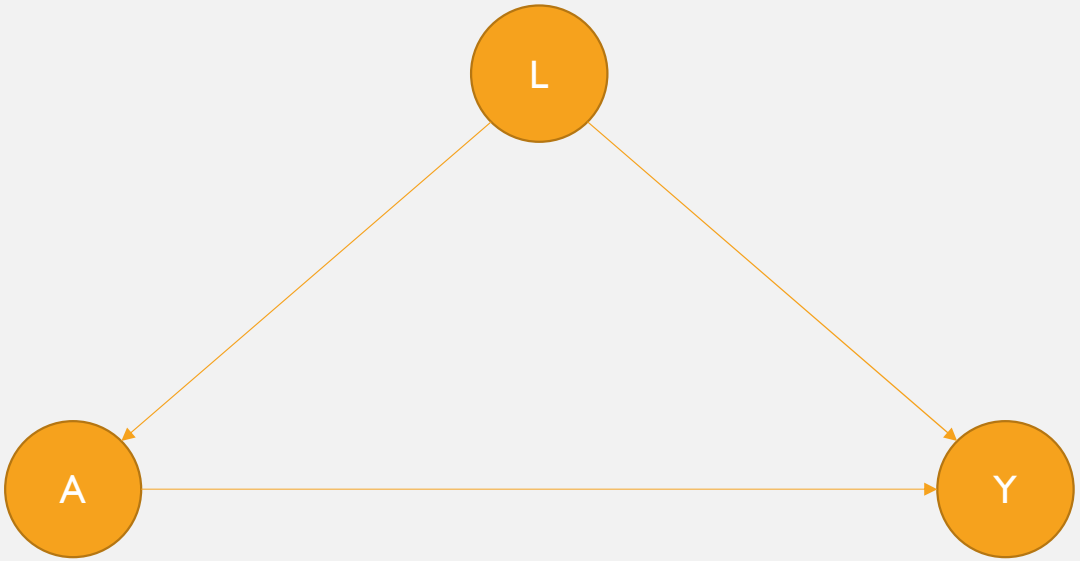
- Then the interior expectation can be re-written as $\mathrm{E}[Y(1) - Y(0)|L] = E[Y|L, A = 1] - E[Y|L, A = 0]$
- So, we can just estimate within each level of severity and then average them together!

A SLIGHTLY MORE COMPLEX SETTING

- What about if the covariates are continuous? What can we do then?
- We can estimate $\mathrm{E}\big[\mathrm{E}[Y(1)|L]\big]$ by first estimation of the conditional distribution of Y given L and A. Then we can combine over the distribution of L (i.e., compute the means separately and then take the difference) → **Outcome regression**
- Alternatively, we could estimate the propensity score $\Pr(A = 1|L)$ → **inverse probability weighting**
- Weight observations by the inverse propensity score:

$$E\left[\frac{A_i}{\Pr(A_i = 1|L)}Y_i\right] = E\left[\frac{1}{\Pr(A_i = 1|L)}E[A_iY_i|L_i]\right]$$
$$= E\left[\frac{\Pr(A_i = 1|L)}{\Pr(A_i = 1|L)}E[Y_i|A_i = 1, L_i]\right] = E\big[E[Y_i|L, A_i = 1]\big]$$

LANGUAGE 2: CAUSAL GRAPHS



LANGUAGE 3: STRUCTURAL EQUATIONS

- *Structural equations*

$$Y = f_Y(A, L, E_Y)$$
$$A = f_Y(A, L, E_Y)$$
$$X = f_X(E_X)$$

- 1921 Wright (genetics);
- 1943 Haavelmo (econometrics);
- 1975 Duncan (social sciences);
- 2000 Pearl (computer science).

OTHER STATISTICAL METHODS

- Matching and randomization inference
  - **Advantages**: transparent; easy implementation; can incorporate prior knowledge; ensures well overlapping covariates.
  - **Disadvantages**: Less efficient (though can be improved).
- Inverse probability weighting
  - **Advantages**: extends matching; generalizable to more complex problems
  - **Disadvantages:** can be unstable if the estimated probabilities are close to zero; not robust to model misspecification.
- Outcome regression
  - **Advantages**: can easily incorporate machine learning methods.
  - **Disadvantages**: not robust to model misspecification.
- Doubly robust estimator
  - **Advantages**: doubly robust; modelling bias is reduced
  - **Disadvantages**: can be unstable if the estimated probabilities are close to zero

DESIGN TRUMPS ANALYSIS

- **Design:**
  - Choose a strategy and collect relevant data
- **Analysis:**
  - Apply an appropriate statistical method to analyze the data
- Success of a research study: ≈99% depends on the design and how data are being collected
- Summarized by the following equation:

Causal estimator – True Causal effect = Design bias + Modeling Bias + Statistical Noise

CORROBORATION OF EVIDENCE

- About 20 years ago, when asked in a meeting what can be done in observational studies to clarify the step from association to causation, Sir Ronald Fisher replied: “Make your theories elaborate.”
- When constructing a causal hypothesis, one should envisage as many different consequences of its truth as possible, and plan observational studies to discover whether each of these consequences is found to hold...
- Falsifiability of scientific theory (K. Popper, 1959).
- “Inference to the best explanation” approach (abduction)
- Evolution of scientific paradigms—the social and collaborative nature of scientific progress (T. Kuhn, 1962).

RESOURCES

- Qingyuan Zhao's lecture notes (Statistical Audience)
- Brady Neal's lecture notes (Machine Learning Audience)
- The Causal Inference Book by Jamie Robins and Miguel Hernan
- An introduction to Causal Inference by Judea Pearl
- The Book of Why by Judea Pearl