Causal Inference I:

Introduction

Vincent Dorie

Columbia University

January 10, 2019

Causal Inference

Does X cause Y?

Causal Inference

Does Z cause Y?

Diet Science

- We should avoid salt [1], until we shouldn't [2], or it doesn't matter much [3, 4]

Diet Science

- We should avoid salt [1], until we shouldn't [2], or it doesn't matter much [3, 4]
- Eat "good" HDL cholesterol for your heart [6], but it won't help [6, 7], and it might hurt [8]

Diet Science

- We should avoid salt [1], until we shouldn't [2], or it doesn't matter much [3, 4]
- Eat "good" HDL cholesterol for your heart [6], but it won't help [6, 7], and it might hurt [8]
- Coffee can cause pancreatic cancer [9, 10], or maybe it doesn't [11, 12], or maybe it's good [13]

Diet Science

- We should avoid salt [1], until we shouldn't [2], or it doesn't matter much [3, 4]
- Eat "good" HDL cholesterol for your heart [6], but it won't help [6, 7], and it might hurt [8]
- Coffee can cause pancreatic cancer [9, 10], or maybe it doesn't [11, 12], or maybe it's good [13]
- Diet fads: Atkins, Dukan, lemon detox, ketogenic, zone, paleo, baby food, cabbage soup, etc

Philosophy

Generally have four philosophical requirements to say that Z “causes” Y

1. neo-Humean: cause and effect always (often) co-occur; cause preceeds effect
2. Weber/counterfactual: for units that are otherwise comparable, cause then effect and no cause, no effect
3. Manipulation: if cause is occassioned/manipulated, effect follows
4. Mechanism: can a pathway be described

See [Oxford Handbook of Political Science]

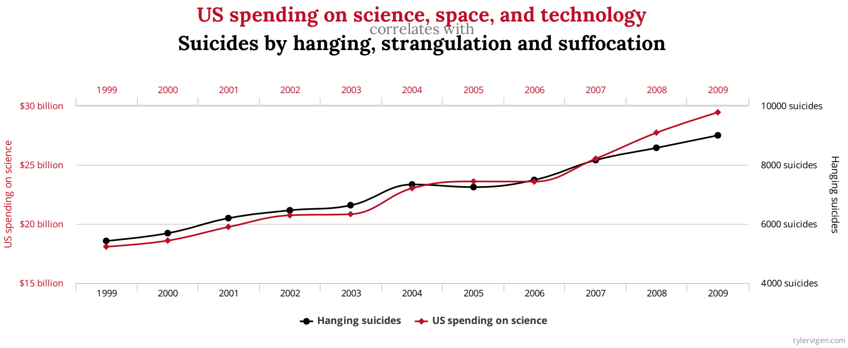
How Causal Inference is Different

In causal inference, the "truth" is never known

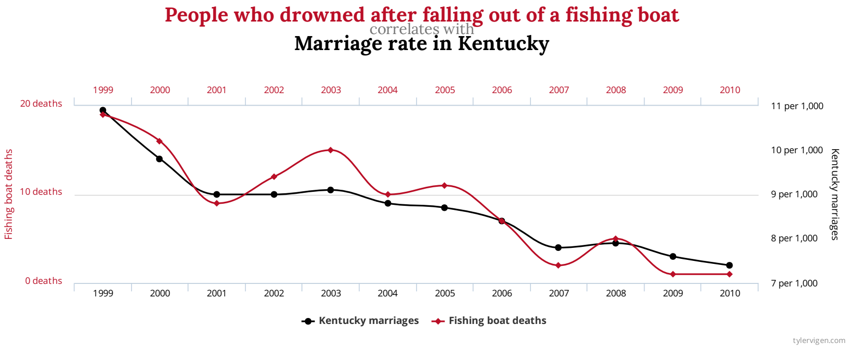
- Can't hold out data and use it to evaluate accuracy
- Have to follow an observational study with a randomized, controlled one

Causal inference is largely about making assumptions so that the true effect can be estimated and arguments that those assumptions are valid

Spurious Correlations



Spurious Correlations



Confounding

Child birth order is associated with an increase risk of Down's Syndrome

Confounding

Child birth order is associated with an increase risk of Down's Syndrome

Maternal age is the "cause" of both

Confounding

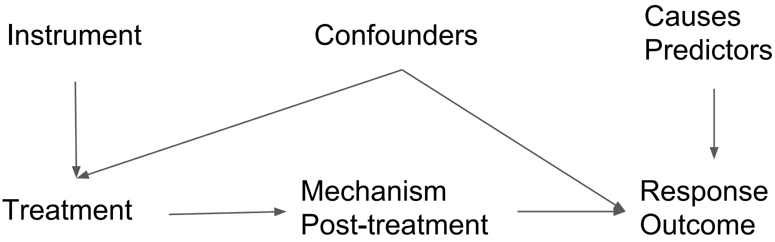
Bottle feeding of infants is associated with an increase in diarrhea (more chances for infection) over breast feeding, but in raw data the effect is reversed

Confounding

Bottle feeding of infants is associated with an increase in diarrhea (more chances for infection) over breast feeding, but in raw data the effect is reversed

The correct relationship appears again if we control for socioeconomic factors (education, wealth, etc)

Variable Types



Potential Outcomes

The *fundamental problem* of causal inference is that for any unit of observation, we can only see its outcome under the treatment or control condition and not both

- For every individual there is a $Y(0)$ and a $Y(1)$ and we observe only $Y(Z)$
- The unobserved value is its *counterfactual*
- Both together are *potential outcomes*

Need to be able to imagine a) what the potential outcomes are and b) that everyone could have received a different treatment

Researcher Sees

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treatment** z_i	**Observed outcome** y_i
Audrey	1	40	0	140
Anna	1	40	0	140
Bob	0	50	0	150
Bill	0	50	0	150
Caitlin	1	60	1	155
Cara	1	60	1	155
Dave	0	70	1	160
Doug	0	70	1	160

Example thanks to Prof Jennifer Hill, NYU

God of Statistics Sees

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**	135	140
Anna	1	40	0	**140**	135	140
Bob	0	50	0	**150**	140	150
Bill	0	50	0	**150**	140	150
Caitlin	1	60	1	160	**155**	155
Cara	1	60	1	160	**155**	155
Dave	0	70	1	170	**160**	160
Doug	0	70	1	170	**160**	160

1. What is the sample average treatment effect (SATE) just with the bolded outcomes?
2. What is the SATE once we can see all potential outcomes?
3. What is the SATE for men? What is the SATE for women? (CATE)

God of Statistics Sees

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**	135	140
Anna	1	40	0	**140**	135	140
Bob	0	50	0	**150**	140	150
Bill	0	50	0	**150**	140	150
Caitlin	1	60	1	160	**155**	155
Cara	1	60	1	160	**155**	155
Dave	0	70	1	170	**160**	160
Doug	0	70	1	170	**160**	160

1. What is the sample average treatment effect (SATE) just with the bolded outcomes? 12.5
2. What is the SATE once we can see all potential outcomes? -15
3. What is the SATE for men? What is the SATE for women? (CATE) -10, -5

Researcher View Again

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treatment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**	?	140
Anna	1	40	0	**140**	?	140
Bob	0	50	0	**150**	?	150
Bill	0	50	0	**150**	?	150
Caitlin	1	60	1	?	**155**	155
Cara	1	60	1	?	**155**	155
Dave	0	70	1	?	**160**	160
Doug	0	70	1	?	**160**	160

What if Instead?

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**		140
Anna	1	40	0		**135**	135
Bob	0	50	0	**150**		150
Bill	0	50	0		**140**	140
Caitlin	1	60	1	**160**		160
Cara	1	60	1		**155**	155
Dave	0	70	1	**170**		170
Doug	0	70	1		**160**	160

What if Instead?

Unit i	**Female** x_{1i}	**Age** x_{2i}	**Treat-ment** z_i	**if** $z_i = 0$, $y_i(0)$	**if** $z_i = 1$, $y_i(1)$	**Observed outcome** y_i
Audrey	1	40	0	**140**		140
Anna	1	40	0		**135**	135
Bob	0	50	0	**150**		150
Bill	0	50	0		**140**	140
Caitlin	1	60	1	**160**		160
Cara	1	60	1		**155**	155
Dave	0	70	1	**170**		170
Doug	0	70	1		**160**	160

The outcomes are *balanced* across covariates - this is what randomization tries to achieve

Notation

- $Y(1)$, $Y(0)$ - potential outcomes
- $Y_i(1)$, $Y_i(0)$ - potential outcomes for individual i
- Z - treatment
- $Y = Y(Z) = Y(1) \cdot Z + Y(0) \cdot (1 - Z)$ - observed response
- X - (observed) confounders

Estimands

Generally interested in an average treatment effect

- $E[Y(1) - Y(0)]$ - average treatment effect (ATE)
- $E[Y(1) - Y(0) \mid Z = 1]$ - average treatment effect on the treated (ATT or ATOT)
- $E[Y(1) - Y(0) \mid X = x]$ - conditional average treatment effect (CATE)

We can estimate for within the sample (SATE), or population (PATE, reweighting)

Assumptions

Required for causal estimands to be identifiable

Assumptions

Required for causal estimands to be identifiable

- Stable-Unit-Treatment Value Assumption: mechanism of treatment immaterial and assigning treatments to other units doesnt interfere

Assumptions

Required for causal estimands to be identifiable

- Stable-Unit-Treatment Value Assumption: mechanism of treatment immaterial and assigning treatments to other units doesnt interfere
- Ignorability: treatment assignment is independent of potential outcomes $\{Y(0), Y(1)\} \perp Z$

Assumptions

Required for causal estimands to be identifiable

- Stable-Unit-Treatment Value Assumption: mechanism of treatment immaterial and assigning treatments to other units doesnt interfere
- Ignorability: treatment assignment is independent of potential outcomes $\{Y(0), Y(1)\} \perp Z$
- Strong ignorability: treatment assignment is independent of potential outcomes, given confounders $\{Y(0), Y(1)\} \perp Z \mid X$, and also $0 < P(Z = 1 \mid X) < 1$ for all X in population of interest

Assumptions

Required for causal estimands to be identifiable

- Stable-Unit-Treatment Value Assumption: mechanism of treatment immaterial and assigning treatments to other units doesnt interfere
- Ignorability: treatment assignment is independent of potential outcomes $\{Y(0), Y(1)\} \perp Z$
- Strong ignorability: treatment assignment is independent of potential outcomes, given confounders $\{Y(0), Y(1)\} \perp Z \mid X$, and also $0 < P(Z = 1 \mid X) < 1$ for all X in population of interest

Ignorability: “selection on observables,” “all confounders measured,” “exchangeability,” the “conditional independence assumption,” and “no hidden bias”

Last: “common support,” or “overlap”, or “positivity”

Potential Outcomes and Causal Inference

The Rubin Causal Model and potential outcomes can be shown to be equivalent to almost every method of for causal inference.

Experimental Design

Causal inferences cannot be made in every type of experiment

- Require some (usually unknown) randomization mechanism
- Includes observational and possibly natural experiments
- Generally not quasi experiments unless potential outcomes can be concieved and assumptions met
- Also requires observations not under the treatment regimen

Validity

Useful but not mathematical concepts required to extrapolate from a sample

- *internal* - alternative explanation instead of treatment for the outcomes
- *construct* - generalizing from your program or measures to the concept of your program or measures
- *external* - generalizing from your study context to other populations

General Formula for CI

1. Define population of interest, treatment, outcome; analyze variables, keep confounders, throw out intermediate outcomes
2. Assume foundational assumptions hold for inferential unit
3. Produce a treatment effect estimate
4. Test sensitivity of estimate to violation of assumptions

Requires a lot of arguing about what is reasonable to assume!