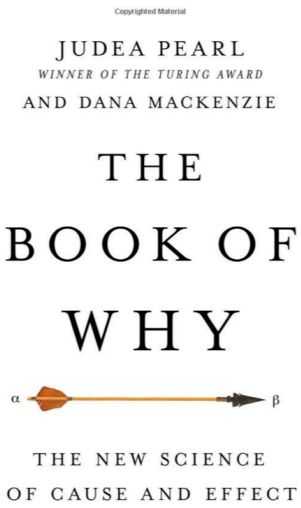
Causal Inference in Medicine and Public Health: An Introduction

Zhenke Wu (with materials from Liz Stuart and Constantine Frangakis)
Assistant Professor
Department of Biostatistics, University of Michigan
2019 Big Data Summer Institute

June 20, 2019

Recommended Summer Reading (Non-Technical)

- *The Book of Why: The New Science of Cause and Effect* (Pearl and Mackenzie, 2018)



- link to article "How a Pioneer of Machine Learning Became One of Its Sharpest Critics - The Atlantic"

Overview

- Methods for estimating causal effects: how to answer the question of "What is the effect of A on B?"
- Randomized designs
- Alternative designs when randomization is infeasible: matching methods, propensity scores, regression discontinuity, and instrumental variables. When are they most feasible?
- Broad spectrum of applications: health sciences, genetic studies, mental health, econometrics, public policy, education, social sciences … (That's why causal inference is very exciting)
- But…we need to define causal questions first.

Question Set 1 (causal questions)

- For a woman older >50 yrs, should she be getting regular screening for breast cancer?
- Do citizens of Los Angeles die because of air polution?
- What is the effect of heavy adolescent marijuana use on adult outcomes (e.g., earnings at age 40)?
- How much mortality and other burden was due to tobacco industry's misconduct?
- Does the Head Start program improve educational and health outcomes for children?
- Does a "healthy marriage" intervention improve relationship quality?
- How does the type of school affect a child's achievements later in life?
- …

Question Set 2 (we will not address such questions)

- Are parents more conservative than their children because they are older?
- Is there an effect of gender in this regression?

How do Question Set 1 and 2 differ?

How do Question Set 1 and 2 differ?

(Hint: "effect of cause" or "cause of effect")

A causal question is a problem with a manipulable *intervention*.

Causal Inference

- Important (and hot) topic right now
- Comparative effectiveness
- Debates regarding study design: "efficacy" versus "effectiveness"; observational versus randomized experiments
- Analytic challenges on modern study designs.
    - Community-level interventions (highway billboards, vaccination, new program…); matched-pair cluster-randomized trials (Wu et al., 2014, *Biometrics*; It's me)
    - Facebook A/B testing which ordering of ads/friends' posts makes you click
    - …

Today's Objectives

- Be able to formalize causal effect discussions
- Understand key elements of causal inference
- Resolve the Lord's paradox

What do we mean by a causal effect?

- What is the effect of some "treatment" T on an outcome Y?
  - Effect of a cause rather than cause of an effect
  - T must be a particular "intervention": something we can imagine giving or withholding
  - e.g. smoking a pack a day on lung cancer, Good Behavior Game on children's behavior and academic achievement

Key Elements in Rubin's Causal Model (Rubin, 1974, Journal of Educational Psychology)

- Units, at a particular place and time
- Treatments/interventions to compare (e.g., $T = 0$ for standard, $T = 1$ for new treatment)
- Potential outcomes, e.g., $Y_i(1), Y_i(0)$ are the outcomes that would be observed on the same subjects if assigned new, or alternatively, if assigned standard treatment
- Causal Effect (definition): comparisons of potential outcomes for the *same* subject or *same* groups of subjects.

Help us be very clear about the effects we are estimating. It helps create a data table with observed and unobserved potential outcomes.

Units

- The entity to which we apply or withhold the treatment
- e.g., individuals, schools, communities
- At a particular point in time
  - Me today and me tomorrow are two different units
- Example: adolescents, elderly people

Treatment

- The "intervention" that we could apply or withhold
  - Not "being male" or "being black"
  - Think of specific intervention that could happen
  - Example: Body mass index (BMI); heavy drug use during adolescence; trained nurses in a clinic to help manage the care for elderly people
- Defined in reference to some control condition of interest (!)
  - Defining control could be more difficult than the treatment
  - No treatment? Existing standard treatment?
  - Example: no or light drug use?

Potential Outcomes

- The potential outcomes that could be observed for each unit
  - $Y(T = 1) = Y(1)$: the outcome that could be observed if a unit gets the treatment
  - $Y(T = 0) = Y(0)$: the outcome that could be observed if a unit gets the control
- For example, your headache pain in two hours if you take an aspirin; your headache pain in two hours if not taking the aspirin
- Example: earnings if are heavy drug user ($Y_i(1)$); earnings if not ($Y_i(0)$)
- Causal effects are comparisons of these potential outcomes
- **No causal inference when exsitence of both $Y_i(1)$ and $Y_i(0)$ makes no sense.**

Setting

We assume the data we have is of the following form:

- Some "treatment", T, measured at a particular point in time
- Covariate(s) X observed on all individuals, measured (or applicable to) time before T
- Outcome(s) Y also observed on all individuals
- Ideally have X measured before T measured before Y

Note: Assume treatment administered at individual-level, but would work the same way for school or group-level treatments (consider the "group" as the "unit")

True "Data"

- e.g., effect of heavy adolescent drug use (T) on earnings at age 40 (Y)

Units	$Y_i(1)$	$Y_i(0)$
1	\$15,000	\$18,000
2	\$9,000	\$10,000
3	\$10,000	\$8,000
⋮	⋮	⋮
n	\$20,000	\$24,000

- Causal Effect for unit (individual) i: $Y_i(1) - Y_i(0)$
- Average causal effect: Average of $Y_i(1) - Y_i(0)$ across individuals

Observed Data

Units	$Y_i(1)$	$Y_i(0)$
1	\$15,000	?
2	?	\$10,000
3	?	\$8,000
⋮	⋮	⋮
n	\$20,000	?

- The fundamental problem of causal inference: $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$. Only observe $Y_i(1)$ or $Y_i(0)$ for each i.
- Causal inference as missing data problem
- So how can we estimate causal effects?

Two Types of Causal Effects

- Can't estimate individual-level causal effects
- So instead we aim to estimate average causal effects
  - e.g., effect of heavy drug use on males
  - Need to compare potential outcomes for males
- "ATE": average treatment effect
  - Average effect for everyone in population:
    $ATE = \frac{1}{N} \sum_{i=1}^{N} Y_i(1) - Y_i(0)$
  - e.g., effect of drug use on everyone, if forced everyone to use drugs
- "ATT": average treatment effect on the treated
  - Average effect for those in the treatment group:
    $ATT = \frac{1}{N_t} \sum_{i=1}^{N_t} (Y_i(1) - Y_i(0) | T_i = 1)$
  - e.g., effect of drug use on people who actually use drugs

Suppose we randomize subjects to $T = 1$ (new intervention) versus 0 (control)

- Estimate causal effect:
  - $ATE = 1/N \sum_{i=1}^{N} Y_i(1) - Y_i(0)$
  - That is: $\left(\sum_{i:T_i=1} Y_i(1) - \sum_{i:T_i=1} Y_i(0) + \sum_{i:T_i=0} Y_i(1) - \sum_{i:T_i=0} Y_i(0)\right) /N$
- **Problem**: cannot observe both $Y_i(1)$ and $Y_i(0)$
- Your goal is to estimate
  - $\sum_{i:T_i=1} Y_i(0)$ and
  - $\sum_{i:T_i=0} Y_i(1)$
- What does randomization ensure?

Statistical Concepts for Learning about Causal Effects

- Replication
- The Stable Unit Treatment Value Assumption (SUTVA)
- The assignment mechanism

Replication

- Need to have multiple units, some getting treatment and some getting control
- The number of potential outcomes grows

Stable Unit Treatment Value Assumption (SUTVA)

- No interference between units: treatment assignment of one unit does not affect potential outcomes of another unit.
  - Agricultural experiments (guard rows)
  - Drug use of one individual does not affect earnings of other individuals
- Only one version of each treatment
  - Lumping all "heavy" drug use together; estimating effect of any "heavy drug use"

Possible SUTVA Violation

- How could these be violated in headache example? Other examples?
  - Study showing benefits for infants of chickenpox vaccine (which they were too young to get):
  - http://www.ncbi.nlm.nih.gov/pubmed/22123875
  - (In this case wouldn't want to use infants as a comparison group!)
  - Recent media discussion of contagion (e.g., of obesity; Christakis and Fowler): http://www.nytimes.com/2011/08/09/health/09network.html?_r=1&ref=health
- Easier to deal with in design than analysis
  - e.g., guard rows
  - e.g., cluster randomized experiments (randomize schools rather than students or classrooms)
- Will talk briefly at the end of the term about some recent advances for modeling interference

Assignment Mechanism

- Process that determines which treatment unit receives
- Randomized experiments: Known (nice!) assignment mechanism
- Observational studies: have to posit an assignment mechanism (determine/model why some individuals become heavy drug users)
    - Propensity score models the assignment mechanism

Assignment Mechanism

- **High-level**: Assignment mechanism is the rule (possibly probabilistic) by which subjects get their actual treatments $\{T_i\}$. The assigned treatments unmask the potential outcomes $Y_i(T_i)$ (denoted by Y_i^{obs}), but mask the rest of potential outcomes, denoted by Y_i^{miss}
- Central to causal inference (look for it when reading literature; many not mentioning this)
- An extreme example: a doctor who always gives her patients the best treatment (no randomness given patient information and the doctor)
    - but over a population of doctors, the assignments can be summarized probabilistically.

Lord's Paradox

From Lord (1967, Page 304):
"A large university is interested in investigating the effects on the students of the diet provided in the university dining halls and any sex differences in these effects. Various types of data are gathered. In particular, the weight of each student at the time of his arrival in September and his weight the following June are recorded."

Distribution of weights for males and females the same in September and in June

Discussion based on Holland and Rubin (1983)

Two Contradictory Statisticians

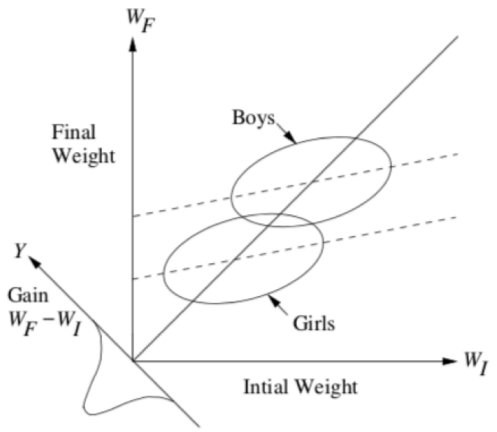


Figure 1: Lord's method of displaying no change in average gain ($W_F - W_I$) co-habitating with an increase in adjusted weight.

Two Contradictory Statisticians

- Statistician 1: No evidence of differential effect
  - Uses difference in mean weight gains
  - Neither group gains nor loses weight
  - Thus no effect for men or women
- Statistician 2: Diet has larger effect on men
  - For a man and woman of same September weight, man will weigh more in June, on average
  - Uses regression adjustment to compare average June weight for men and women with same September weight

Who is right?

Consider the framework:

- Units:
- Covariates:
- Potential outcomes:
- Treatment:
- Control:

Well, it depends

Consider the framework:

- Units: students
- Covariates: Sex, September weight
- Potential outcomes: June weight under treatment and control
- Treatment: University diet
- Control: ???

Lord's observed data

	Covariates (X)	June weight		
Students	Sex, Sept. weight	Y(0)	Y(1)	Impact
1	X_1	?	$Y_1(1)$	?
2	X_2	?	$Y_2(1)$	?
3	X_3	?	$Y_3(1)$	?
⋮	⋮	⋮		
N	X_N	?	$Y_N(1)$	?

Two control conditions

- Statistician 1:
  - June weight under control = September weight
- Statistician 2:
  - June weight under control a linear function of September weight
  - Models for male and female weights parallel
  - $E(Y(0)) = a + b * \mathrm{Sex} + \mathrm{c} * \mathrm{Weight}_{\mathrm{Sept}}$
- Either could be right, depending on assumptions made about the control condition

Recommended Reading

Pearl (2016) Lord's Paradox Revisited - (Oh Lord! Kumbaya!)

Common Objectives in Causal Inference Training/Reasearch

- Understand causal problems as potential interventions and how they are different from association questions
- Understand the framework to discuss/mathematicize causal inference (potential outcomes, assignment mechanism); graphical approaches (Pearl) not discussed here but hugely important for their intuitive appeals and robustness to probability specifications ("The Book of Why")
- Understand designs (ways to efficiently collect useful data) and methods for analyzing these data to answer causal questions
- Understand complications in causal studies, including missing data, noncompliance and many hidden biases.
- **Use these as powerful tools to critically review research with causal claims. And improve science!**

Main Points Once Again

- Causal inference is *counterfactual*
- Causal effect is a function of $(Y(1), Y(0))$
- Causal inference requires estimation of unobserved reponses - it makes sense when the estimation does
- Causal inference requires assignment mechanism
- Assignment mechanism known in randomized studies; must be assumed or modeled in observational studies (propensity scores)
- Causal inference makes *assumptions*, e.g., non-interference, etc.

- Thank you!
- zhenkewu.com
- zhenkewu@umich.edu