STA 790 (Fall 2022) — Bayesian Causal Inference

Chapter 1: Overview of Potential Outcome Framework

Fan Li

Department of Statistical Science
Duke University

Questions on Causation

- Relevant questions about causation:
  - the philosophical meaningfulness of the notion of causation
  - deducing the causes of a given effect
  - understanding the details of causal mechanism

- In this class we focus on measuring the effects of causes – a place where statistics, which is concerned with measurement, has most contributions to make.

Association vs. Causation

- The research questions that motivate most studies in statistics-based sciences are causal in nature.
- Standard statistical analysis is to infer *associations* among variables, based on which may do some *prediction*
- Causal analysis is one step further: it is about *counterfactual prediction*, predict what would have happened to the same units/subjects had they were exposed to a different (counterfactual) condition
- In most cases, *association does not imply causation*

Causal Inference

- *How to make the leap from association to causation*?
- Key: *causal assumptions* - structural and/or modeling
- Causal inference is about
  1. build a framework and define causal effects under general scenarios
  2. specify assumptions under which one can decare/identify causation from association
  3. assess the sensitivity to the causal assumptions and find ways to mitigate

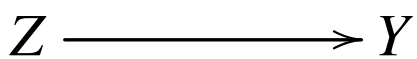
Notations

- Treatment (e.g. intervention, exposure) Z: for illustration we will mostly focus on binary treatments
- Outcome (e.g. disease status) Y
- Observed covariates or confounders X
- Unobserved covariates or confounders U
- Examples of question of interest
  - Causal effect of exposure on disease
  - Comparative effectiveness research: whether one drug or medical procedure is better than the other
  - Program evaluation in economics and policy

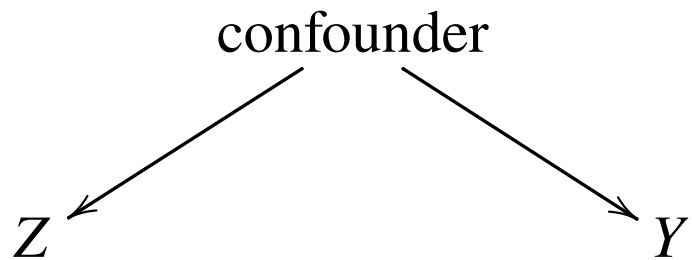
Confounding

- Confounding (or common cause) is the main complication/hurdle between association and causation
- Two Directed Acyclic Graphs

Causal relationship:



Confounding:



Examples of Confounding

- Ice cream consumption and number of people drown.
  Confounder: temperature
- A Covid-19 example:

COVID Vaccine Hesitancy and Risk of a Traffic Crash

DECEMBER 02, 2021 | 1:30PM - 3:00PM

Speaker: Donald Redelmeier, Professor of Medicine at the University of Toronto; Canada Research Chair in Medical Decision Sciences; Director of Clinical Epidemiology at Sunnybrook Health Sciences Centre; Senior Scientist at the Institute for Clinical Evaluative Studies in Ontario; Staff physician in the Division of General Internal Medicine at Sunnybrook Hospital

Abstract:

COVID vaccine hesitancy is a reflection of judgment, reasoning, and other psychological influences that may also contribute to traffic safety. We tested whether COVID vaccine hesitancy was associated with an increased risk of a serious traffic crash.

A total of 11,270,763 adults were identified, of whom 16% had not received a COVID vaccine and 84% had received a COVID vaccine. Those who had not received the vaccine accounted for a disproportionate number of crashes, equivalent to a significant increased traffic risk. The association between a lack of COVID vaccine and increased traffic risks extended to diverse patient subgroups, persisted after adjusting for measured baseline differences, applied across a spectrum of crash severity, and was similar to the relative risk associated with a diagnosis of sleep apnea.

We suggest that COVID vaccine hesitancy is associated with a significant increased risk of a serious traffic crash. An awareness of this counter-intuitive finding might contribute to more public support for the COVID vaccine.

A Classic Example—Smoking and Lung Cancer

Doll and Hill (1950 BMJ)

Figure: Sir Austin Bradford Hill (1897–1991)

- Smoking-cancer association
- Case-control study of lung cancer
- Risk ratio $\approx$ odds ratio, is roughly 9 even after adjusting for observed covariates:

$$\mathrm{RR}_{ZY}^{\mathrm{obs}} = \frac{\Pr(Y = 1 \mid Z = 1)}{\Pr(Y = 1 \mid Z = 0)} \approx 9$$

- Does smoking cause lung cancer?
- Box (2013) stopped smoking after seeing Doll and Hill (1950)

A Classic Example—Smoking and Lung Cancer

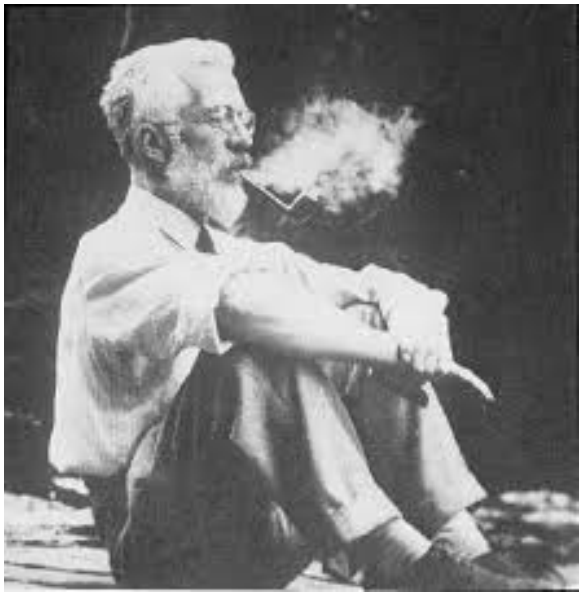


Figure: Sir Ronald Aylmer Fisher (1890–1962)

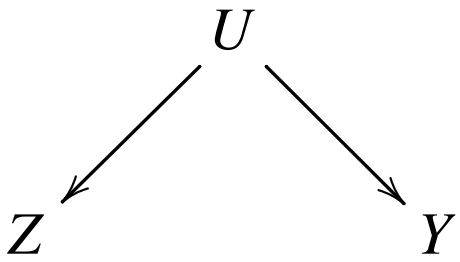
- Association does not imply causation
- "Common cause" (Reichenbach 1956, Fisher 1957 BMJ)
- Fisher (1957 BMJ):

  cigarette-smoking and lung cancer, though not mutually causative, are both influenced by a common cause, in this case the individual genotype.

Cornfield Inequality for Smoking and Lung Cancer

Cornfield et al. (1959 JNCI)

Common cause hypothesis



- Smoking Z
- Lung cancer Y
- Genetic factor U

- Assume Fisher is right
- The smoking-gene association must satisfy:

$$\mathrm{RR}_{ZU} \geq \mathrm{RR}_{ZY} \approx 9$$

- Such a genetic confounder is too strong to be realistic.
- Association must be due to causal.
- We will revisit this example later.

Frameworks for Causal Inference

- The purpose is to construct a model or a framework that is complex enough to allow us to formalize basic intuitions concerning cause and effect
- Two commonly used frameworks
  - The *potential outcome framework*, also known as the counterfactual framework, or the Neyman-Rubin Causal Model (Neyman, 1923; Rubin, 1974; Imbens and Rubin, 2015; Hernan and Robins, 2020)
  - The *causal diagram* framework (Pearl, 2009)
  - Mathematically the two frameworks are connected (Richardson and Robins, 2013), but each has different established goals, tools and applicable areas
  - This class focuses on the potential outcome framework, and will occasionally draw directed acyclic graphs (DAGs) for simple illustration

Potential Outcome Framework: Basic Setup

- Unit: The person, place, or thing upon which a treatment will operate, at a particular time
- *Target Population*: a well-defined **population** of units whose outcomes are going to be compared
- Data: a random sample of N units from a target population
- Treatment: an intervention, the effects of which (on some outcomes of the units) the investigator wishes to assess relative to no intervention (i.e., the control)
- For simplicity, consider a treatment with two levels: $z = 0, 1$
- For each unit i, we observe the (binary) treatment status $Z_i (= 0, 1)$, a vector of p covariates $X_i = (X_{i1}, \ldots X_{ip})$, and an outcome Y_i^{obs} (or simply denoted as Y_i later)

Potential Outcome Framework: Basic Setup

- Potential outcomes: the values of a unit's outcome (hypothetically) under treatment or control
- Each unit i has two potential outcomes: $Y_i(0), Y_i(1)$. Note: Such notation implicitly relies on SUTVA (more later)
- Causal effect: comparison between the potential outcomes under treatment and under control for *the same unit or a common set of units*. For example:
- Individual causal effect (ITE): $Y_i(1) - Y_i(0)$
- Average treatment effect (ATE): $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$, ITE averaged over a target population

The Fundamental Problem of Causal Inference

Holland, 1986, JASA

- The fundamental problem of causal inference: We can observe at most one of the potential outcomes for each unit, the other(s) are missing/counterfactual
- Causal inference under the potential outcome framework is essentially a missing data problem
- To identify causal effects from observed data, under any mathematical framework, one must make assumptions (structural or/and stochastic)

SUTVA

Assumption 1: Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1980):

- SUTVA includes two assumptions: (1) no interference, (2) no different versions of a treatment
- Seems trivial, actually very strong assumptions
- Under SUTVA, each unit has only two potential outcomes $Y_i(1), Y_i(0)$. Connect to observed outcome Y: $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$
- SUTVA connects the intervention we see (Z), with the causal intervention of interest (z)

Causal Estimands

- Causal estimands are functions of the potential outcomes
  - Conditional average treatment effect (CATE): conditional on a covariate value

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X = x]$$

  - Average treatment effect (ATE): important to clarify what population the expectation is taken on

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}_x[\tau(x)].$$

  - Average treatment effect for the treated units (ATT):

$$\tau = \mathbb{E}[Y_i(1) - Y_i(0)|Z_i = 1].$$

- Obviously these estimands are not identifiable without further assumptions
- So, what assumptions do we need?

Perfect Doctor

	Potential Outcomes			Observed Data		
	$Y(0)$	$Y(1)$	Z	$Y(0)$	$Y(1)$	Y^{obs}
	13	14	1	?	14	14
	6	0	0	6	?	6
	4	1	0	4	?	4
	5	2	0	5	?	5
	6	3	0	6	?	6
	6	1	0	6	?	6
	8	10	1	?	10	10
	8	9	1	?	9	9
True averages	7	5	Observed averages		5.4	11

Perfect Doctor: Comments

- The simple difference-in-means estimator does not return a valid estimate of the true causal effect. Why?
- Key: the assignment mechanism (Rubin, 1978) – the probabilistic rule that decides which unit gets assigned to which treatment

$$p(Z_i = 1 | X_i, Y_i(0), Y_i(1))$$

- Is the assignment mechanism in the perfect doctor example random? (in fact, what does random assignment mean?)
- The assignment depends on both $Y_i(0)$ and $Y_i(1)$ for each unit (in reality we will unlikely to have any such perfect doctor)
- The key identifying assumptions in causal inference are on the assignment mechanism

Ignorable Assignment

- A majority of causal studies assume versions of ignorable assignment, which consists of two sub-assumptions: (1) unconfoundedness; (2) overlap
- Note: there is no census in the literature of the terminology of unconfoundedness vs. (strong) ignorability: often used exchangeably
- The terminology corresponds to that of missing data mechanism in the missing data literature (Rubin, 1976).
- The meaning of “ignorability" is most relevant in Bayesian inference of causal effect: means that the assignment mechanism drops out from the data likelihood in estimating the causal effects

Positivity (or overlap)

Assumption 1: Positivity (or overlap):
$0 < \Pr(Z_i = 1 | X_i, Y_i(0), Y_i(1)) < 1$ for all i.

- Positivity requires, in large samples, for all possible values of the covariates there are both treated and control units.
- Testable from observed data

Unconfoundedness

Assumption 2: Unconfoundedness

$$\Pr(Z_i = 1|X_i, Y_i(0), Y_i(1)) = \Pr(Z_i = 1|X_i)$$

Often also written as $\{Y_i(0), Y_i(1)\} \perp W_i|X_i$

- Assumes that within subpopulations defined by values of observed covariates, the treatment assignment is random
- Rules out unmeasured confounders
- Randomized experiments satisfy unconfoundedness
- Untestable in most observational studies, but sometimes can be indirectly tested, and sensitivity can be checked
- $e_i(x) \equiv \Pr(Z_i = 1|X_i = x)$ is called the propensity score (Rosenbaum and Rubin, 1983)

Assumptions for Estimating ATT

- When the estimand is ATT, unconfoundedness and overlap can be weakened (Heckman, Ichimura, Todd, 1997)
  - Unconfoundedness for untreated: $Y_i(0) \perp Z_i \mid \mathbf{X}_i$,
  - Weak overlap: $\Pr(Z_i = 1|\mathbf{X}_i = \mathbf{x}) < 1$ for any $\mathbf{x}$.
- ATT also has population and sample version (more later).

Ignorable Assignment Mechanisms

- What assignment mechanisms are unconfounded, and more generally ignorable?
  - Randomized experiments? Yes, it is known and controlled by investigators, and *ignorable*
  - The perfect doctor example? No, because it depends on both $Y_i(1)$ and $Y_i(0)$
  - The smoking and lung cancer example? We do not know.
  - Most observational studies (e.g. smoking and lung cancer data): the assignment mechanism is unknown and uncontrolled.
- To make causal inference with *observational data*, we have to make (often strong) assumptions about the assignment mechanism

Identify causal effects under ignorability

- Under ignorability, we have

$$\Pr(Y(z)|X) = \Pr(Y|X, Z = z)$$

- The observed distribution of Y in treatment arm $Z = z$ equals the distribution of the potential outcome $Y(z)$.

- Thus we have the following two identification formula for ATE:

$$\begin{aligned} \tau^{\text{ATE}} &= E\{\mu_1(X) - \mu_0(X)\} \\ &= E\left\{\frac{ZY}{e(X)} - \frac{(1-Z)Y}{1-e(X)}\right\} \end{aligned}$$

  where $\mu_z(X) = E(Y|Z = z, X) = E(Y(z)|X)$ is the outcome model under treatment $z(z = 0, 1)$.

- This suggests two estimation strategies of ATE: (1) outcome regression; (2) inverse probability weighting

Method 1: Outcome Regression

- Specify two outcome regression model, one for each potential outcome

$$\mathbb{E}[Y(1)|X = x] = \mu_1(x), \quad \mathbb{E}[Y(0)|X = x] = \mu_0(x)$$

- Let $\hat{\mu}_z(X_i)$ denote the fitted potential outcome $Y_i(z)$ based on the regression models. Essentially imputing missing potential outcomes, can estimate any causal estimand
  - For ATE, the outcome-regression estimator is

$$\hat{\tau}^{\text{ATE}} = \sum_{i=1}^{N} Z_i(Y_i - \hat{\mu}_0(X_i)) + (1 - Z_i)(\hat{\mu}_1(X_i) - Y_i) \Big/ N$$

  - For ATT, the outcome-regression estimator is

$$\hat{\tau}^{\text{ATT}} = \sum_{i=1}^{N} Z_i \{Y_i - \hat{\mu}_0(X_i)\} \Big/ N_1$$

  - For CATE $\tau(x)$, the outcome-regression estimator is

$$\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$$

Outcome Regression: Linear models

- A common choice of outcome model is the linear regressions:

$$\mu_z(x) = \alpha_z + \beta_z X_i + \epsilon_{z,i}$$

- Then the outcome-regression estimator of ATE has the exact form as the ANCOVA estimator in randomized experiments

$$\widehat{\tau}^{\text{reg}} = \left\{\bar{Y}_1 - \widehat{\beta}_1(\bar{X}_1 - \bar{X})\right\} - \left\{\bar{Y}_0 - \widehat{\beta}_0(\bar{X}_0 - \bar{X})\right\}$$

  where $\widehat{\beta}_z$ is the OLS estimate of the coefficient of X in the regression $\mu_z(x)$

- Consistency
  - In randomized experiments, $\widehat{\tau}^{\text{reg}}$ is consistent for ATE even if the linear model is misspecified (Lin, 2013). Why?
  - But not the case in observational studies

Method 2: Inverse Probability Weighting

- Inverse probability weighting (IPW) does not involve an outcome model, it only involves the propensity score $e(X_i) = \Pr(Z_i = 1|X_i)$
- Estimator 1: Horvitz-Thompson (unnormalized)

$$\hat{\tau}^{HT} = \frac{\sum_{i=1}^{N} Z_i Y_i / \hat{e}(X_i)}{\sum_{i=1}^{N} Z_i} - \frac{\sum_{i=1}^{N} (1 - Z_i) Y_i / \{1 - \hat{e}(X_i)\}}{\sum_{i=1}^{N} (1 - Z_i)}$$

- Estimator 2: Hajék (normalize the weights so that the sum in each group is 1)

$$\hat{\tau}^{Hajek} = \frac{\sum_{i=1}^{N} Z_i Y_i / \hat{e}(X_i)}{\sum_{i=1}^{N} Z_i / \hat{e}(X_i)} - \frac{\sum_{i=1}^{N} (1 - Z_i) Y_i / \{1 - \hat{e}(X_i)\}}{\sum_{i=1}^{N} (1 - Z_i) / \{1 - \hat{e}(X_i)\}},$$

- Hajék is usually more stable and efficient than HT
- In both estimators, extreme (close to 0 or 1) propensity scores (i.e. violation of the overlap assumption) lead to large variance

Method 3: Combine Outcome regression and IPW – Doubly Robust Estimation

- The IPW estimator is consistent if the prop score model $e(x)$ is correct; the outcome regression estimator is consistent if the outcome model $\mu_z(x)$ is correct
- How about combine the two? Easy to verify: with true $\mu_z(X)$ and $e(X)$

$$\begin{aligned}
\tau &= \mathbb{E}\left\{\frac{ZY}{e(X)} - \frac{Z-e(X)}{e(X)}\mu_1(X)\right\} \\
&\quad - \mathbb{E}\left\{\frac{(1-Z)Y}{1-e(X)} + \frac{Z-e(X)}{1-e(X)}\mu_0(X)\right\} \\
&= \mathbb{E}\left[\mu_1(X_i) + \frac{Z_i\{Y_i-\mu_1(X_i)\}}{e(X_i)}\right] \\
&\quad - \mathbb{E}\left[\mu_0(X_i) + \frac{(1-Z_i)\{Y_i-\mu_0(X_i)\}}{1-e(X_i)}\right] \\
&= \mu_1 - \mu_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]
\end{aligned}$$

Doubly Robust Estimator

- In the previous formula, replace the true PS $e(x)$ and outcome $\mu_z(x)$ by the estimated ones from postulated models $\hat{e}(x)$ and $\hat{\mu}_z(x)$, we obtain two augmented estimators:

$$\begin{aligned}\hat{\tau}_{\text{dr}} &= \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{Z_i - \hat{e}(X_i)}{\hat{e}(X_i)}\hat{\mu}_1(X_i)\right\} - \frac{1}{N}\sum_{i=1}^{N}\left\{\frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)} + \frac{Z_i - \hat{e}(X_i)}{1-\hat{e}(X_i)}\hat{\mu}_0(X_i)\right\} \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_1(X_i) + \frac{Z_i\{Y_i - \hat{\mu}_1(X_i)\}}{\hat{e}(X_i)}\right] - \frac{1}{N}\sum_{i=1}^{N}\left[\hat{\mu}_0(X_i) + \frac{(1-Z_i)\{Y_i - \hat{\mu}_0(X_i)\}}{1-\hat{e}(X_i)}\right].\end{aligned}$$

- The two estimators are mathematically equivalent, but different statistical implications: the first estimator augments an IPW estimator by outcome regression (OR); the second augments an OR estimator by IPW

- The first estimator is usually refereed to as the doubly-robust (DR) estimator (Scharfstein et al. 1999; Lunceford and Davidian, 2004; Bang and Robins, 2005)

Doubly Robust Estimator

Double Robustness: $\hat{\tau}_{\text{dr}}$ is a consistent estimator of the ATE if either the propensity score model or the potential outcome model is, but not necessary both are, correctly specified

- DR is a large sample property
- Offers protection against model mis-specification: give you two chances to get it right (and wrong)!
- If $e(X)$ and $\mu_z(X)$ are modeled correctly, $\hat{\tau}_{\text{dr}}$ will have smaller variance than the IPW estimator (in large samples)
- If the outcome model $\mu_z(X)$ is correct, $\hat{\tau}_{\text{dr}}$ has larger variance (in large samples) than the outcome regression estimator
- ... but gives protection in the event it is not
- Semiparametric efficient (Robins et al. 1994)

Method 4: Matching

- Regression estimators impute the missing potential outcomes using the estimated regression function
- Matching estimators also impute the missing potential outcomes, using the outcomes of nearest (in terms of a certain distance measure) neighbours of the opposite treatment group
- Matching is similar to nonparametric kernel regression, essentially an imputation method
- They have often (but not exclusively) been applied in settings where
  - target estimand is ATT
  - a large reservoir of potential controls, allowing matching each treated unit to one or more distinct controls (matching without replacement)
- More general settings: both treated and control units are (potentially) matched and matching is done with replacement

Example: Nearest-Neighbor (NN) Matching

- let $\mathcal{M}_i$ be the set of the indices of the M closest matches of unit i in terms of the distance measure based on the norm $\|\cdot\|$

$$\sum_{j|Z_j \neq Z_i} 1\{\|X_j - X_i\| \leq \|X_l - X_i\|\} = M$$

- Let

$$\hat{Y}_i(0) = \begin{cases} \sum_{j \in \mathcal{M}_i} Y_j / M, & Z_i = 1, \\ Y_i, & Z_i = 0, \end{cases}$$

and

$$\hat{Y}_i(1) = \begin{cases} Y_i, & Z_i = 1, \\ \sum_{j \in \mathcal{M}_i} Y_j / M, & Z_i = 0. \end{cases}$$

- The treatment effect within a pair is then estimated as the difference in outcomes, and then average over pairs

$$\hat{\tau}^{\text{ATE}} = \sum_i \left(\hat{Y}_i(1) - \hat{Y}_i(0)\right) \Big/ N,$$

$$\hat{\tau}^{\text{ATT}} = \sum_i \left(Y_i - \hat{Y}_i(0)\right) Z_i \Big/ N_1.$$

Matching: Tuning and Connection to Others

- Matching involves lots of tuning decisions of the users
  - distance metric: Mahalanobis distance, propensity score, variance distributional discrepancies
  - high-dimensional covariates: covariate selection, order of importance of covariates
  - fixed or varying no. matches
  - for fixed M, number of matches
  - with or without replacement
  - caliper or nearest neighbor
  - how to deal with ties
- Many many matching methods, with some theory and general guidelines available
- But in general, choice of specific matching method is more of a personal habit
- With proper mathematical formulations, matching estimators can be viewed as nonparametric versions of IPW, regression or DR estimators based on nearest-neighbor regressions (Lin, Ding, Han, 2021)

Why Randomization is Special?

- In randomized experiments, assignment mechanism is known and controlled by investigators
- Randomization:
  - balances observed covariates: $Z \perp\!\!\!\perp X$
  - balances unobserved covariates: $Z \perp\!\!\!\perp U$
  - balance potential outcomes, i.e. guarantee ignorability (Rubin 1978)

$$Z \perp\!\!\!\perp (Y(1), Y(0))$$

- Observational studies do not have this luxury: trt and control units are often severely imbalanced
- In (frequentist) causal inference: a key step is to ensure covariate balance between groups, mimicking a randomized experiment

Propensity score

Definition (Rosenbaum and Rubin, 1983). The propensity score is defined as the conditional probability of receiving a treatment given pre-treatment covariates X:

$$e(X) = \Pr(Z = 1|X) = \mathbb{E}(Z|X), \tag{1}$$

where $X = (X_1, \ldots, X_p)$ is the collection of p covariates.

- Propensity score is a probability, analogous to a summary statistic (of the assignment mechanism)

Balancing property of propensity score

Property 1. The propensity score $e(X)$ balances the distribution of all X between the treatment groups:

$$Z \perp X \mid e(X).$$

Equivalently, $\Pr(Z_i = 1 \mid X_i, e(X_i)) = \Pr(Z_i = 1 \mid e(X_i))$.

- A balancing score $b(x)$ is a function of the covariates such that:

$$Z \perp X \mid b(X).$$

- Propensity score is a balancing score
- Rosenbaum and Rubin (1983) show that $e(X)$ is the coarsest balancing score: all balancing score is a function of $e(X)$

Remarks on the balancing property

1. If a subclass of units or a matched treatment-control pair is homogenous in $e(X)$, then the treatment and control units have the same distribution of X
2. If a subclass of units or a matched treatment-control pair is homogenous in both $e(X)$ and certain X, the other components of X within those refined class is also balanced – practical implication: estimating causal estimand in subpopulation, e.g. male or female group
3. The balancing property is a statement on the distribution of X, NOT on assignment mechanism or potential outcomes

Propensity score: Unconfoundedness

Property 2. If Z is unconfounded given X, then Z is unconfounded given $e(X)$, i.e.,

$$\{Y_i(1), Y_i(0)\} \perp Z_i \mid X_i \Longrightarrow \{Y_i(1), Y_i(0)\} \perp Z_i \mid e(X_i)$$

- Given a vector of covariates that ensure unconfoundedness, adjustment for differences in propensity scores removes all biases associated with differences in the covariates
- $e(X)$ can be viewed as a summary score of the observed covariates
- Causal inference can be drawn through stratification, matching, weighting, etc. using the scalar $e(X)$ instead of the high dimensional covariates.

Propensity score: remarks

- The propensity score balances the **observed** covariates, but does not generally balance **unobserved** covariates
- In most observational studies, the propensity score $e(X)$ is unknown and thus needs to be estimated
- There is a bias-variance tradeoff between modeling $e(X)$ and directly modeling the outcome $\Pr(Y(z) \mid X)$

Propensity score analysis of causal effects

Propensity score analysis typically involves two stages:

Stage 1 Estimate the propensity score $\hat{e}(X)$, e.g. by a logistic regression ($e(\boldsymbol{X}_i; \hat{\boldsymbol{\beta}}) = 1/(1 + \exp(-\boldsymbol{X}_i^T \hat{\boldsymbol{\beta}}))$) or machine learning methods.

Stage 2 Given the estimated propensity score, estimate the causal effects through one of these methods:

- Weighting: IPW or other weights, PS is essential
- Matching: PS as a distance metric
- Subclassification: a coarsened version of weighting
- Regression: PS as a covariate
- Mixed procedure of the above

References

Cole, S. R., Frangakis, C. E. (2009). The consistency statement in causal inference: a definition or an assumption?. Epidemiology, 20(1), 3-5.

Cole, S. R., Stuart, E. A. (2010). Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. American journal of epidemiology, 172(1), 107-115.

Cornfield, J., Haenszel, W., et al. (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. Journal of the National Cancer institute, 22(1), 173-203.

Ding, P, Li, F.(2018). Causal inference: a missing data perspective. *Statistical Science*. 33(2), 214-237.

Doll, R., Hill, A. B. (1954). The mortality of doctors in relation to their smoking habits. British medical journal, 1(4877), 1451.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1935). The Design of Experiments (1st ed.), Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1957). Dangers of cigarette-smoking. British Medical Journal, 2(5039), 297.

Hernan MA, Robins JM. Causal Inference. (2020). Boca Raton: Chapman & Hall/CRC.

Holland, P. (1986). Statistics and causal inference (with discussion). J. Amer. Statist. Assoc. 81, 945-970

References

Imbens, G. W. and Rubin, D. B. (2015). Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge Univ. Press, New York.

Li, F, Mealli, F. (2014). A conversation with Donald B. Rubin. *Statistical Science*. 29(3), 439-457.

Lin Z, Ding P, Han F. (2021) Estimation based on nearest neighbor matching: from density ratio to average treatment effect. arXiv:2112.13506

Neyman J. (1990). On the application of probability theory to agricultural experiments: Essay on Principles, Section 9. Masters Thesis. Portions translated into English by D. Dabrowska and T. Speed (1990). Statistical Science. 1923(5): 465-472

Pearl, J. (2009). Causality. Cambridge University Press.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. J. Educ. Psychol. 66 688-701.

Rubin, D. B. (1976). Inference and missing data. Biometrika 63 581-592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. Annals of Statistics. 6 34-58.

Rubin, D. B. (1980). Randomization analysis of experimental data: The Fisher randomization test comment. Journal of the American Statistical Association, 75(371), 591-593.