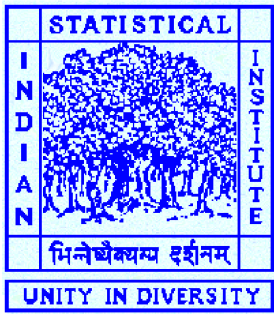
Unraveling Causality : Foundations and Advances in Causal Inference

D.Basu Presentation

Srijan Chattopadhyay

Indian Statistical Institute, Kolkata

August 22, 2024



I would like to dedicate this to the loving memory of Prof. Debabrata Basu in his birth centenary, whose profound contributions to statistics continue to inspire and illuminate our journey in this field. His legacy of brilliance and humility remains an enduring beacon for us all.

Contents

What is Causal Inference?

Introduction

- George Barnard

In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements “X is true” and “X is known to be true”.

Introduction

- George Barnard

In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "X is true" and "X is known to be true".

☞ "Correlation doesn't imply Causation".

Introduction

- George Barnard

In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "X is true" and "X is known to be true".

- "Correlation doesn't imply Causation".
- Most of the Superstitions basically arose from some bad statistics (in many cases, thinking association as cause, doing small sample inferences).

Introduction

- George Barnard

In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "X is true" and "X is known to be true".

- "Correlation doesn't imply Causation".
- Most of the Superstitions basically arose from some bad statistics (in many cases, thinking association as cause, doing small sample inferences).
- Causal inference tries to infer causal relationships from experimental or observational data.

Introduction

- George Barnard

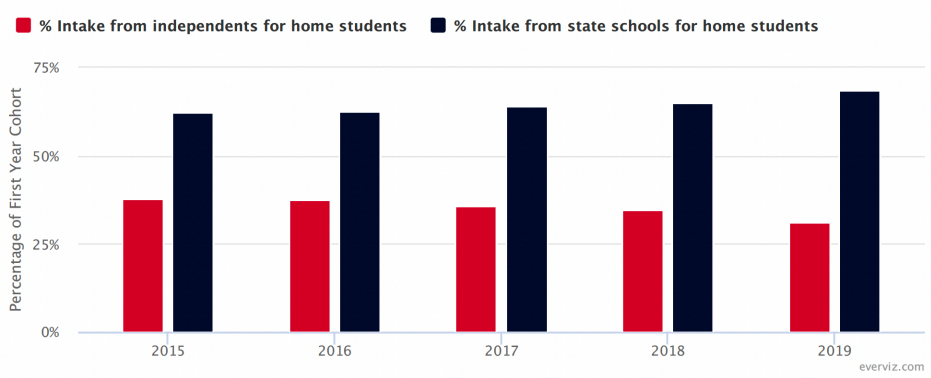
In statistical inference, as distinct from mathematical inference, there is a world of difference between the two statements "X is true" and "X is known to be true".

- "Correlation doesn't imply Causation".
- Most of the Superstitions basically arose from some bad statistics (in many cases, thinking association as cause, doing small sample inferences).
- Causal inference tries to infer causal relationships from experimental or observational data.
- Connects statistical theory with the real world:
  - Classical statistics: models $\rightarrow$ inference.
  - Machine learning: data $\rightarrow$ prediction.
  - Causal inference: models and inference $\leftrightarrow$ reality

Motivating Example 1

By the mid-1940s,
it had been observed that lung cancer cases had tripled over the previous three decades. A series of observational studies since 1950 reported overwhelmingly strong association between smoking and lung cancer. But 'nature is tricky' so can we be sure the association is causal? Some prominent statisticians including R A Fisher objected to the idea that this implies that smoking causes lung cancer, but no compelling competing hypothesis could be found. This led to one of the biggest public health intervention to reduce tobacco consumption.

Motivating Example 2



- Cambridge welcomes 68.7% of 2019 from maintained schools but falls far below a national average of 93% of state educated students.
- Some interesting quotes:"Considering 93% of pupils in England are taught in state schools, a figure of 68.7% means that state school students are still vastly underrepresented in the University.... **Cambridge's acceptance of state school applicants continues to be amongst the lowest in the UK**."

Motivating Example 2

So?

Does this mean Cambridge's admission is biased against state schools?

Motivating Example 2

So?

Does this mean Cambridge's admission is biased against state schools?

Not necessarily!

Motivating Example 2

So?

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.

Motivating Example 2

So?

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 or computer algorithms (Kusner and Loftus (2020)).

Languages of Causal Inference

$$\text{Causal inference} \approx \text{Causal language/model} + \text{Statistical inference}$$

There are three typical languages of Causal Inference.

1. Using Potential Outcomes/Counterfactuals
2. Using Structural Equation Modelling (SEM)
3. Using Graphs.

However, for today, we will mainly focus on Counterfactuals/Potential Outcomes.

1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see ...?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do ...? How?*
(What would Y be if I do X?
How can I make Y happen?)

EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done ...? Why?*
(Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

Randomized Experiments

What is Randomized Experiment?

- Randomization involves randomly allocating the experimental units across the treatment groups.
- For example, randomly allocating placebo or treatment to patients (doesn't matter whether the patient agrees or not!)
- Practically infeasible due to ethical issues.
- But simple to start with, so let's focus on randomized experiment for now!

Notations

For i th unit :

- Covariates : X_i
- Treatment : $A_i \in \mathcal{A} = \{0, 1\}$. $A_i = 1$ means treated.
- $[n] = \{1, \cdots, n\}, A_{[n]} = (A_1, \cdots, A_n)^T, X_{[n]} = (X_1, \cdots, X_n)^T$
- The assignment mechanism of the treatment is the conditional distribution $P(A_{[n]} = a_{[n]} | X_{[n]} = x_{[n]}) = \pi(a_{[n]} | x_{[n]})$. Examples include : Bernoulli Trial, Sample Without Replacement, Bernoulli Trials with Covariates etc.

An important treatment assignment mechanism

Pairwise Experiment

- Suppose n is even. The units are divided into $n/2$ pairs based on the covariates.
- Within each pair, one unit is randomly assigned to treatment.
- Let $B_i = B_i(x_{[n]})$ be the pair that unit i is assigned to.

$$\pi(a_{[n]}|x_{[n]}) = \frac{1}{2^{n/2}} 1\left(\sum_{i=1}^{n} a_i.1(B_i = j) = 1 \; \forall j = 1, \cdots, n/2\right)$$

"Implicit" Causal Inference

To estimate the causal effect of A on Y, one may

- Get the differences of the main effect estimates by fitting some anova/regression.

"Implicit" Causal Inference

To estimate the causal effect of A on Y, one may

- Get the differences of the main effect estimates by fitting some anova/regression.
- Compare the conditional expectations $E[Y|A=0]$ with $E[Y|A=1]$.

"Implicit" Causal Inference

To estimate the causal effect of A on Y, one may

- Get the differences of the main effect estimates by fitting some anova/regression.
- Compare the conditional expectations $E[Y|A=0]$ with $E[Y|A=1]$.
- Compare the conditional distributions $P(Y \leq y|A=0)$ with $P(Y \leq y|A=1)$.

"Implicit" Causal Inference

To estimate the causal effect of A on Y, one may

- Get the differences of the main effect estimates by fitting some anova/regression.
- Compare the conditional expectations $E[Y|A=0]$ with $E[Y|A=1]$.
- Compare the conditional distributions $P(Y \leq y|A=0)$ with $P(Y \leq y|A=1)$.
- Further condition on X and compare $E[Y|A=0; X=x]$ with $E[Y|A=1; X=x]$ or the conditional distributions.

"Implicit" Causal Inference

To estimate the causal effect of A on Y, one may

- Get the differences of the main effect estimates by fitting some anova/regression.
- Compare the conditional expectations $E[Y|A=0]$ with $E[Y|A=1]$.
- Compare the conditional distributions $P(Y \leq y|A=0)$ with $P(Y \leq y|A=1)$.
- Further condition on X and compare $E[Y|A=0; X=x]$ with $E[Y|A=1; X=x]$ or the conditional distributions.

This approach allows us to apply familiar statistical methodologies, but it has several limitations :

- Causal inference is only implicit and informal, as it seems that any difference can only be reasonably attributed to the different treatment assignments.
- Difficult to extend to non-iid treatment assignments.

Potential Outcomes

The potential outcome model avoids the above problems and provides a flexible basis for causal inference. It is first introduced by Neyman in his 1923 Master's thesis (Splawa-Neyman et al. (1990)) to study randomised experiments and later brought to observational studies by Rubin (Rubin (1974)).

This approach posits a potential outcome (or counterfactual), $Y_i(a_{[n]})$, for unit i under treatment assignment $a_{[n]}$. The potential outcomes (or counterfactuals) are linked to the observed outcome (or factuals) via the following assumptions.

SUTVA (Stable Unit Treatment Value Assumption)

Assumption (Consistency)

$Y_i = Y_i(A_{[n]})$ *for all* $i \in [n]$.

SUTVA (Stable Unit Treatment Value Assumption)

Assumption (Consistency)

$Y_i = Y_i(A_{[n]})$ *for all* $i \in [n]$.

Assumption (No Interference)

$Y_i(a_{[n]}) = Y_i(a_i)$ *for all* $i \in [n]$ *and* $a_{[n]} \in \mathcal{A}^n$

SUTVA (Stable Unit Treatment Value Assumption)

Assumption (Consistency)

$Y_i = Y_i(A_{[n]})$ *for all* $i \in [n]$.

Assumption (No Interference)

$Y_i(a_{[n]}) = Y_i(a_i)$ *for all* $i \in [n]$ *and* $a_{[n]} \in \mathcal{A}^n$

Because A_i is binary, we are only dealing with two potential outcomes, $Y_i(0)$ and $Y_i(1)$, for each unit i.

Fundamental Problem of Causal Inference

The difference $Y_i(1) - Y_i(0)$ is called the individual treatment effect for unit i, which can never be observed. This is often referred to as the "fundamental problem of causal inference" (Holland (1986)).

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	?	-3.7	1	-3.7
2	2.3	?	0	2.3
3	?	7.4	1	7.4
4	0.8	?	0	0.8
⋮	⋮	⋮	⋮	⋮

However, it should be possible to estimate the treatment effect at the population level if the treatment is randomised.

Average Treatment Effect and Estimation

Definition (ATE)

Sample Average Treatment Effect (SATE) $= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
Population Average Treatment Effect (PATE) $= E(Y_i(1) - Y_i(0))$

Average Treatment Effect and Estimation

Definition (ATE)

Sample Average Treatment Effect (SATE) $= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
Population Average Treatment Effect (PATE) $= E(Y_i(1) - Y_i(0))$

Note: PATE implicitly assumes that the n units are sampled from a super population, so $Y_i(0)$ and $Y_i(1)$ follow an unknown bivariate probability distribution.
Intuitively, it should be possible to estimate the treatment effect at the population level if the treatment is randomised. This can be formalised by the following assumption:

Average Treatment Effect and Estimation

Definition (ATE)

Sample Average Treatment Effect (SATE) $= \frac{1}{n}\sum_{i=1}^{n}(Y_i(1) - Y_i(0))$
Population Average Treatment Effect (PATE) $= E(Y_i(1) - Y_i(0))$

Note: PATE implicitly assumes that the n units are sampled from a super population, so $Y_i(0)$ and $Y_i(1)$ follow an unknown bivariate probability distribution.
Intuitively, it should be possible to estimate the treatment effect at the population level if the treatment is randomised. This can be formalised by the following assumption:

Assumption (Randomization)

$A_{[n]} \perp Y_{[n]}(a_{[n]})|X_{[n]}$ *for* $a_{[n]} \in \mathcal{A}^n$

Causal Identification

Theorem

Consider a Bernoulli trial with covariates, where $\{X_i, A_i, Y_i(a), a \in \mathcal{A}\}$ *are iid. Suppose the previous assumptions hold and* $P(A = a|X = x) > 0,\ \forall a \in \mathcal{A}, x$. *Then for all* $a \in \mathcal{A}$ *and* x,

$$(Y(a)|X = x) \stackrel{d}{=} (Y|A = a, X = x) \tag{1}$$

Proof.

$P(Y(a) \leq y|X = x) = P(Y(a) \leq y|A = a, X = x)$ [By Assumption 3]
$= P(Y \leq y|X = x)$[By Assumption 1] □

Why $P(A = a|X = x) > 0$?

This means under each covariate level both treatments and controls exist (for the sake of estimability of ATE under each covariate level)

Causal Identification

Corollary

Under the same set of assumptions as Theorem 1,

$$ATE = E(Y(1) - Y(0)) = E(E(Y|A = 1, X = x) - E(Y|A = 0, X = x))) \quad (2)$$

If $P(A = 1|X)$ *does not depend on* X*, then*

$$ATE = E[Y|A = 1] - E[Y|A = 0] \quad (3)$$

Proof.

First part is trivial. For Equation (3), we prove it in the case of discrete X. Since $A \perp X, P(X = x) = P(X = x|A = 1) = P(X = x|A = 0)$. By Theorem 1,
$E(Y(1)) = \sum_x E(Y|A = 1, X = x)P(X = x)$
$= \sum_x E(Y|A = 1, X = x)P(X = x|A = 1) = E(Y|A = 1)$

Why Causal Identification?

Results like Equation (1), (2), (3) are called causal identification, because they equate a **counterfactual quantity** on the left hand side with a **factual (so estimable) quantity** on the right hand side under a restrictive environment.

Causal Effect Estimator

The most intuitive estimator of ATE is $\hat{\beta} = \bar{Y_1} - \bar{Y_0}$, where $\bar{Y_1} = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i}$, $\bar{Y_0} = \frac{\sum_{i=1}^{n} Y_i(1-A_i)}{\sum_{i=1}^{n}(1-A_i)}$

Theorem

Under the previous assumptions and if A_i are sampled without replacement,
$E(\hat{\beta}|Y(0), Y(1)) = SATE$
$V(\hat{\beta}|Y(0), Y(1)) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{n}$, *where* $S_a^2 = \frac{\sum_{i=1}^{n}(Y_i(a) - \bar{Y(a)})^2}{n-1}$,
$S_{01}^2 = \frac{\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - SATE)^2}{n-1}$

Causal Effect Estimator

The most intuitive estimator of ATE is $\hat{\beta} = \bar{Y_1} - \bar{Y_0}$, where $\bar{Y_1} = \frac{\sum_{i=1}^{n} Y_i A_i}{\sum_{i=1}^{n} A_i}$, $\bar{Y_0} = \frac{\sum_{i=1}^{n} Y_i(1-A_i)}{\sum_{i=1}^{n}(1-A_i)}$

Theorem

Under the previous assumptions and if A_i are sampled without replacement,
$E(\hat{\beta}|Y(0), Y(1)) = SATE$
$V(\hat{\beta}|Y(0), Y(1)) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{n}$, *where* $S_a^2 = \frac{\sum_{i=1}^{n}(Y_i(a) - \bar{Y(a)})^2}{n-1}$,
$S_{01}^2 = \frac{\sum_{i=1}^{n}(Y_i(1) - Y_i(0) - SATE)^2}{n-1}$

Estimation of V

Note that, S_{01} can't be estimated due to fundamental problem of causal inference. So, it is common to estimate V by $\frac{\hat{S_0}^2}{n_0} + \frac{\hat{S_1}^2}{n_1}$, where $\hat{S_1}^2 = \frac{\sum_{i=1}^{n_1} A_i(Y_i - \bar{Y_1})^2}{n_1 - 1}$. Thus we get a conservative estimate of the variance.

Testing no Causal Effect

Fisher (Edwards (2005)) appears to be the first to grasp fully the importance of randomization for credible causal inference (Imbens and Rubin (2015)). Fisher considered testing the sharp null hypothesis (or exact null hypothesis)

$$H_0 : Y_i(0) = Y_i(1),\ \forall i \in [n]$$

Using H_0, we can impute all the missing values in the previous table.

i	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
1	-3.7	-3.7	1	-3.7
2	2.3	2.3	0	2.3
3	7.4	7.4	1	7.4
4	0.8	0.8	0	0.8

Then under the distribution π, simulate various $A_{[n]}$ and get all possible values of $\hat{\beta}$. Fisher proposed to test H_0 based on how extreme the observed $\hat{\beta}$ is compared to other potential values of $\hat{\beta}$.

An Example!

Under sampling without replacement, the following 6 scenarios are equally likely :

1. $A_{[4]} = (1, 1, 0, 0), \hat{\beta} = -4.8$
2. $A_{[4]} = (1, 0, 1, 0), \hat{\beta} = 0.3$
3. $A_{[4]} = (1, 0, 0, 1), \hat{\beta} = -6.3$
4. $A_{[4]} = (0, 1, 1, 0), \hat{\beta} = 6.3$
5. $A_{[4]} = (0, 1, 0, 1), \hat{\beta} = -0.3$
6. $A_{[4]} = (0, 0, 1, 1), \hat{\beta} = 4.8$

The observed realization is the second.

A More General Set up!

$H_0 : Y_i(1) - Y_i(0) = \beta; \forall i \in [n]$. This is still a very strong hypothesis: it says the individual treatment effect is always a fixed value β. Using the consistency assumption and H_0, we can impute :

$$Y_i(a) = \begin{cases} Y_i & \text{if } a = A_i, \\ Y_i + \beta & \text{if } a > A_i, \\ Y_i - \beta & \text{if } a < A_i \end{cases}$$

What Next?

- Then we choose a test statistic $T(A_{[n]}, X_{[n]}, Y_{[n]})$. An example is the difference-in-means estimator $\hat{\beta}$.
- Then, just test using the empirical distribution obtained by the randomzation.
- It turns out that it is indeed a level α test.
- Can take $T = \frac{|\hat{\beta}|}{\sqrt{var(\hat{\beta})}}$, or $T = \sum 1(A_i > A_j)1(Y_i > Y_j)$ etc.

No Unmeasured Confounders : Randomization Inference

What Next?

- In practical, we almost never get data which was collected through some randomized experiment due to potential ethical issues.

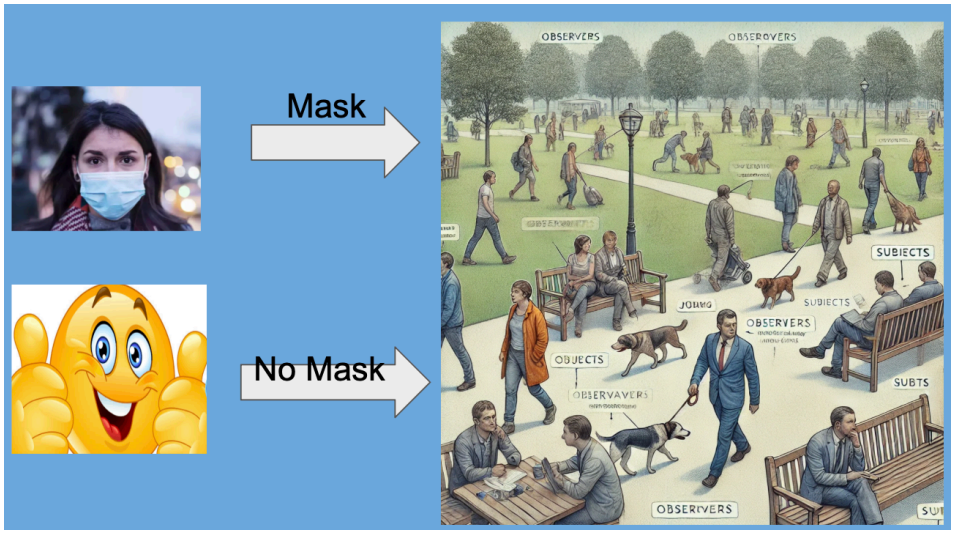
What Next?

- In practical, we almost never get data which was collected through some randomized experiment due to potential ethical issues.
- So, it is important to know how can we actually make "good causal inference" in the real world?

What Next?

- In practical, we almost never get data which was collected through some randomized experiment due to potential ethical issues.
- So, it is important to know how can we actually make "good causal inference" in the real world?
- Most data in the real world comes from some observational study!

What is Observational Study?



- An observational study draws inferences from a sample to a population where the independent variable is not under the control of the researcher because of ethical concerns or logistical constraints.
- This is in contrast with experiments, such as randomized controlled trials, where each subject is randomly assigned to a treated group or a control group.

Importance of the Design

Causal inference $\approx$ Causal language/model + Statistical inference.
Consider the following two situations.

1. **Situation 1:** Suppose we let half of the patients to receive the treatment **at random**. Significantly more treated participants have a better outcome. Therefore, the treatment must be beneficial.
2. **Situation 2:** Suppose the observed patients are **pair matched**, so that the patients in the same pair have similar demographics and medical history. In significantly more pairs, the treated patient has a better outcome. Therefore, the treatment must be beneficial.

Apart from statistical error and causality, a third possible explanation in Situation 2 is that the treated patients and the control patients are systematically different in some other way (eg, different lifestyles). So causal inference in observational studies is always abductive (inference to the best explanation).

Causal estimator - True causal effect = Design bias + Modelling bias + Statistical noise

No Unmeasured Confounders

- We will assume all relevant confounders are measured.
- Let $A_i \in \{0, 1\}$; be a binary treatment for individual i, Y_i be its the outcome of interest with two counterfactuals $Y_i(0)$ and $Y_i(1)$, and X_i be a p-dimensional vector of covariates.
- We assume $(X_i, A_i, Y_i(0), Y_i(1)), i = 1, \cdots, n$, are i.i.d.
- In other words, we eliminate one of the possible explanations to observed associations by assumption. This is convenient for studying statistical methodologies but obviously optimistic for practical applications.

Matching Algorithms

- Matching is a popular observational study design. Matching is essentially a preprocessing algorithm that aims to reconstruct a **pairwise randomised experiment** or a **stratified randomised experiment** from observational data.
- To simplify the exposition, we will assume $A_i = 1$ for $1 \leq i \leq n_1$ and $A_i = 0$ for $n_1 + 1 \leq i \leq n$.
- An essential element is a measure of distance $d(.,.)$ between two values of the covariates X.
- Following are 2 examples of distances.

Mahalanobis Distance

$$d_{MA}(X_i, X_j) = (X_i - X_j)^T \hat{\Sigma}^{-1} (X_i - X_j)$$

where

$$\hat{\Sigma} = \frac{1}{n}[\sum_{i=1}^{n_1}(X_i - \bar{X_1})(X_i - \bar{X_1})^T + \sum_{i=n_1+1}^{n}(X_i - \bar{X_0})(X_i - \bar{X_0})^T]$$

where $\bar{X_1}$ and $\bar{X_0}$ are the treatment and control sample means.

Propensity Score based Distance

Propensity score is defined as $\pi(x) = P(A = 1|X = x)$. It has been particularly famous in the recent years in Causal Modelling (Rosenbaum (2023)). The propensity score can be estimated from the observational data, commonly by fitting a logistic regression of A_i on X_i.

$$d_{PS}(X_i, X_j) = \left[\log\left(\frac{\hat{\pi}(X_i)}{1-\hat{\pi}(X_i)}\right) - \log\left(\frac{\hat{\pi}(X_j)}{1-\hat{\pi}(X_j)}\right)\right]^2$$

Sometimes, Mahalanobis distance with a propensity score caliper is used.

$$d(X_i, X_j) = \begin{cases} d_{MA}(X_i, X_j), & \text{if } d_{PS}(X_i, X_j) < \tau^2 \\ \infty, & \text{Otherwise} \end{cases}$$

where τ is treated as a tuning parameter.

Nearest-Neighbour Matching

- Given the distance measure d, this naive method matches a treated observation $1 \leq i \leq n_1$ with its nearest control observation.
- The problem with this method is that one control individual could be matched to several treated individuals, which never happens in a pairwise randomised experiment.
- We can fix this problem by a greedy algorithm that sequentially matches a treated i to its nearest control neighbor that has yet been selected. A drawback is that the result will then depend on the order of the input.

Optimal Matching

An improvement is the optimal matching that solves the following optimisation problem:

$$\text{minimize} \sum_{i=1}^{n_1} d(X_i, \sum_{j=n_1+1}^{n} M_{ij} X_j)$$

$$\text{subject to } M_{ij} \in \{0,1\} \; \forall 1 \leq i \leq n_1, \forall n_1 + 1 \leq j \leq n$$

$$\sum_{j=1}^{n_0} M_{ij} = 1, \; \forall 1 \leq i \leq n_1$$

$$\sum_{i=1}^{n_1} M_{ij} \leq 1, \; \forall n_1 + 1 \leq j \leq n$$

Optimal Matching

- M_{ij} is an indicator for the treated observation i being matched to the control observation j.
- The last two constraints mean that every treated is matched to exactly one control and every control is matched to at most one treated.
- Although combinatorial optimisation is generally NP-complete, the optimal matching problem can be recasted as a network flow problem and solved efficiently in polynomial time. (Rosenbaum (2020))

Checking Covariate Balance

- We can assess whether the matching is satisfactory by checking covariate balance.

Checking Covariate Balance

- We can assess whether the matching is satisfactory by checking covariate balance.
- A common measure of covariate imbalance is the standardised covariate differences (Rosenbaum and Rubin (1985))

$$B_k(M) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{ik} - \sum_{j=n_1+1}^{n} M_{ij}X_{ik})}{\sqrt{(s_{1k}^2 + s_{0k}^2)/2}}, k = 1, \cdots, p$$

Checking Covariate Balance

- We can assess whether the matching is satisfactory by checking covariate balance.
- A common measure of covariate imbalance is the standardised covariate differences (Rosenbaum and Rubin (1985))

$$B_k(M) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{ik} - \sum_{j=n_1+1}^{n} M_{ij}X_{ik})}{\sqrt{(s_{1k}^2 + s_{0k}^2)/2}}, k = 1, \cdots, p$$

- A rule of thumb is that the k-th covariate X_k is considered approximately balanced if $|B_k| < 0.1$, but obviously we would like the entire vector B to be as close to 0 as possible.

Checking Covariate Balance

- We can assess whether the matching is satisfactory by checking covariate balance.
- A common measure of covariate imbalance is the standardised covariate differences (Rosenbaum and Rubin (1985))

$$B_k(M) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{ik} - \sum_{j=n_1+1}^{n} M_{ij}X_{ik})}{\sqrt{(s_{1k}^2 + s_{0k}^2)/2}}, k = 1, \cdots, p$$

- A rule of thumb is that the k-th covariate X_k is considered approximately balanced if $|B_k| < 0.1$, but obviously we would like the entire vector B to be as close to 0 as possible.
- If the covariate balance is unsatisfactory, a common practice is to rerun the matching algorithm with a different distance measure or remove treated units that have extreme propensity scores. This is often called the **propensity score tautology**(Imai et al. (2008)).

Checking Covariate Balance

- We can assess whether the matching is satisfactory by checking covariate balance.
- A common measure of covariate imbalance is the standardised covariate differences (Rosenbaum and Rubin (1985))

$$B_k(M) = \frac{\frac{1}{n_1}\sum_{i=1}^{n_1}(X_{ik} - \sum_{j=n_1+1}^{n} M_{ij}X_{ik})}{\sqrt{(s_{1k}^2 + s_{0k}^2)/2}}, k = 1, \cdots, p$$

- A rule of thumb is that the k-th covariate X_k is considered approximately balanced if $|B_k| < 0.1$, but obviously we would like the entire vector B to be as close to 0 as possible.
- If the covariate balance is unsatisfactory, a common practice is to rerun the matching algorithm with a different distance measure or remove treated units that have extreme propensity scores. This is often called the **propensity score tautology**(Imai et al. (2008)).
- In modern optimal matching algorithms, it is possible to include $|B_k(M)| \leq \eta$; for all k as a constraint in the combinatorial optimisation problem.

Estimation

For simplicity, we assume treated observation i is matched to control observation $i + n_1$, $i = 1, \cdots, n_1$. Let $D_i = (A_i - A_{i+n_1})(Y_i - Y_{i+n_1})$ be the treated-minus-control difference in pair i. Can consider

$$\bar{D} = \frac{1}{n_1} \sum_{i=1}^{n_1} D_i$$

Randomization Inference

Assumption

We assume matching reconstructs a pairwise randomised experiment, so

$$P(A_{[2n_1]} = a | C_{[2n_1]}, A_{[2n_1]} \in M) = 2^{-n_1}.1(a \in M)$$

Randomization Inference

Assumption

We assume matching reconstructs a pairwise randomised experiment, so

$$P(A_{[2n_1]} = a | C_{[2n_1]}, A_{[2n_1]} \in M) = 2^{-n_1}.1(a \in M)$$

Consider the sharp null hypothesis $H_0 : Y_i(1) - Y_i(0) = \beta, \ \forall i$, where β is given. Under H_0 and by the consistency assumption, the counterfactual values of $D_{[n_1]}$ can be imputed as

$$D_i(a_{[2n_1]}) = (a_i - a_{i+n_1})(Y_i(a_i) - Y_{i+n_1}(a_{i+n_1})) = \begin{cases} D_i, & a_i = 1, a_{i+n_1} = 0, \\ 2\beta - D_i, & a_i = 0, a_{i+n_1} = 1 \end{cases}$$

Randomization Inference

Consider any test statistic $T = T(D_{[n_1]})$. Next we construct a randomisation test based on the randomisation distribution of $T(D_{[n_1]}(A_{[2n_1]}))$. Let $F(t)$ denote its cumulative distribution function given $C_{[2n_1]}$ and $A_{[2n_1]} \in M$ under H_0,

$$\begin{aligned} F(t) &= P(T \leq t | C_{[2n_1]}, A_{[2n_1]} \in M) \\ &= \sum_{a_{[2n_1]} \in M} \left(\frac{1}{2}\right)^{n_1} .1(T \leq t) \end{aligned}$$

We then compute the p-value for this randomisation test as $P_2 = F(T)$ and reject the hypothesis H_0 if P_2 is less than a significance threshold $0 < \alpha < 1$.

Randomization Inference

Consider any test statistic $T = T(D_{[n_1]})$. Next we construct a randomisation test based on the randomisation distribution of $T(D_{[n_1]}(A_{[2n_1]}))$. Let $F(t)$ denote its cumulative distribution function given $C_{[2n_1]}$ and $A_{[2n_1]} \in M$ under H_0,

$$\begin{aligned} F(t) &= P(T \leq t | C_{[2n_1]}, A_{[2n_1]} \in M) \\ &= \sum_{a_{[2n_1]} \in M} \left(\frac{1}{2}\right)^{n_1} .1(T \leq t) \end{aligned}$$

We then compute the p-value for this randomisation test as $P_2 = F(T)$ and reject the hypothesis H_0 if P_2 is less than a significance threshold $0 < \alpha < 1$.

Theorem

$P(P_2 \leq \alpha) \leq \alpha$ under H_0 for all $0 \leq \alpha \leq 1$.

No Unmeasured Confounders : Various Estimators

A Motivating Example! I

Suppose that we are interested in giving teenagers interesting math and stat topics to read to discourage them from smoking. A random subset of 5% of teenagers in Baranagar, and a random subset of 20% of teenagers in Kharagpur are eligible for the study.

	Baranagar		Kharagpur	
	Non-S.	Smoker	Non-S.	Smoker
Treat.	152	5	581	350
Control	2362	122	2278	1979

	Baranagar + Kharagpur	
	Non-Smoker	Smoker
Treatment	733	355
Control	4640	2101

- Within each city, we have a randomized controlled study, and in fact readily see that the treatment helps.

A Motivating Example! II

- $\hat{\tau}_B = \frac{5}{5+152} - \frac{122}{122+2362} = -1.7\%$
- $\hat{\tau}_K = \frac{350}{350+581} - \frac{1979}{1979+2278} = -8.9\%$
- $\hat{\tau}_{B+K} = \frac{355}{355+733} - \frac{2101}{2101+4640} = 0.01\%$
- However, looking at aggregate data is misleading, and it looks like the treatment hurts.
- Once we aggregate the data, this is no longer an RCT because Kharagpur people are both more likely to get treated, and more likely to smoke whether or not they get treated.
- In order to get a consistent estimate of the ATE, we need to estimate treatment effects in each city separately.
- $\hat{\tau} = \frac{2641}{2641+5188}\hat{\tau}_B + \frac{5188}{2641+5188}\hat{\tau}_K = -6.5\%$

Aggregating difference-in-means estimators

$$\hat{\tau}_{AGG} = \sum_{x \in \chi} \frac{n_x}{n} \hat{\tau}(x)$$

where

$$\hat{\tau}(x) = \frac{1}{n_{x1}} \sum_{X_i = x, A_i = 1} Y_i - \frac{1}{n_{x0}} \sum_{X_i = x, A_i = 0} Y_i$$

Theorem

$\sqrt{n}(\hat{\tau}_{AGG} - \tau) \xrightarrow{D} N(0, V_{AGG})$, *where* $V_{AGG} = V(\tau(X_i)) + E\left(\frac{\sigma^2(X_i)}{\pi(X_i)(1-\pi(X_i))}\right)$, *and* $\sigma^2(x) = V(Y(1)|X = x) = V(Y(0)|X = x)$ *[Remarkably, the asymptotic variance* V_{AGG} *does not depend on* $|\chi| = p$*]*

But....

What to do in case of continuous X? because we won't be able to get enough samples for each possible value of $x \in \chi$ to be able to get a reasonable estimate of $\tau(x)$.

Propensity stratification

1. Sort the observations according to their propensity scores.
2. Split the sample into J evenly size strata using the sorted propensity score and, in each stratum $j = 1, 2, \cdots, J$, compute the simple difference in-means treatment effect estimator for the stratum

$$\hat{\tau}_j = \frac{\sum_{j=[(j-1)n/J]+1}^{[jn/J]} A_i Y_i}{\sum_{j=[(j-1)n/J]+1}^{[jn/J]} A_i} - \frac{\sum_{j=[(j-1)n/J]+1}^{[jn/J]} (1 - A_i) Y_i}{\sum_{j=[(j-1)n/J]+1}^{[jn/J]} (1 - A_i)}$$

3. Estimate the average treatment by

$$\hat{\tau}_{STRATA} = \frac{1}{J} \sum_{i=1}^{J} \hat{\tau}_j$$

Outcome Regression

$$\hat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i))$$

where $\hat{\mu}_i$ is estimated via OLS or some nonparametric regression between Y and $X_1, \cdots, X_p$ differently for $A = 0$ and $A = 1$.

Note

Whether or not the true effect function $\mu_w(x)$ is linear, OLS always reduces the asymptotic variance of difference in means estimator (Little (2019)).A worst case for OLS is when $\beta_{(1)} = \beta_{(0)} = 0$, i.e., when OLS asymptotically just does nothing, then $\hat{\tau}_{OR}$ reduces to Difference in means estimator.

Inverse-propensity weighting (IPW)

Another, algorithmically simpler way of exploiting unconfoundedness is via inverse-propensity weighting.

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{A_i Y_i}{\hat{\pi}(X_i)} - \frac{(1-A_i)Y_i}{1-\hat{\pi}(X_i)}\right)$$

where $\hat{\pi}(x)$ is estimated via non-parametric regression or some other way.

Note

We see that the performance of the oracle IPW estimator (i.e. IPW with the true $\pi(x)$ value) is somewhat disappointing when compared to $\hat{\tau}_{AGG}$ in discrete case. Despite having access to the true propensity score $\pi(x)$, it always under-performs $\hat{\tau}_{AGG}$. But it can be proved that it is consistent for τ [proved by using the oracle IPW].

Augmented IPW I

- Given that the average treatment effect can be estimated in two different ways, i.e., by first non-parametrically estimating $\pi(x)$ or by first estimating $\hat{\mu}_{(0)}(x)$ and $\hat{\mu}_{(1)}(x)$, it is natural to ask whether it is possible to combine both strategies.
- This turns out to be a very good idea, and yields the augmented IPW (AIPW), also called Doubly Robust estimator.
-

$$\hat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}_{(1)}(X_i) - \hat{\mu}_{(0)}(X_i) + A_i\frac{Y_i - \hat{\mu}_{(1)}(x)}{\hat{\pi}(X_i)} - (1 - A_i)\frac{Y_i - \hat{\mu}_{(0)}(x)}{1 - \hat{\pi}(X_i)})$$

Augmented IPW II

Note

AIPW is consistent if either the $\hat{\mu}$ are consistent or $\hat{\pi}$ is consistent. Furthermore, one might also argue that, in a modern statistical setting, one should expect practitioners to appropriate non-parametric estimators for both $\mu_{(w)}(x)$ and $\pi(x)$ such that both are consistent; in which case both $\hat{\tau}_{OR}$ and $\hat{\tau}_{IPW}$ would already be consistent on their own, and so the above double robustness statement (namely consistency of $\hat{\tau}_{AIPW}$) doesn't buy us much. However, if both the models are consistent, then we get asymptotic normality of $\hat{\tau}_{AIPW}$.

Augmented IPW III

Note

Sometimes, a **Cross fitting estimator** is used to avoid idiosyncrasies of the specific model adjustment we chose to use. Cross-fitting first splits the data (at random) into two halves I_1 and I_2, and then uses an estimator which uses the model trained in I_2 to predict that of I_1 and vice versa and take an weighted average of these two. Under some mild model accuracy conditions, it corrects bias due to overfitting and works better than the usual one and gives asymptotic normality as well.

Note

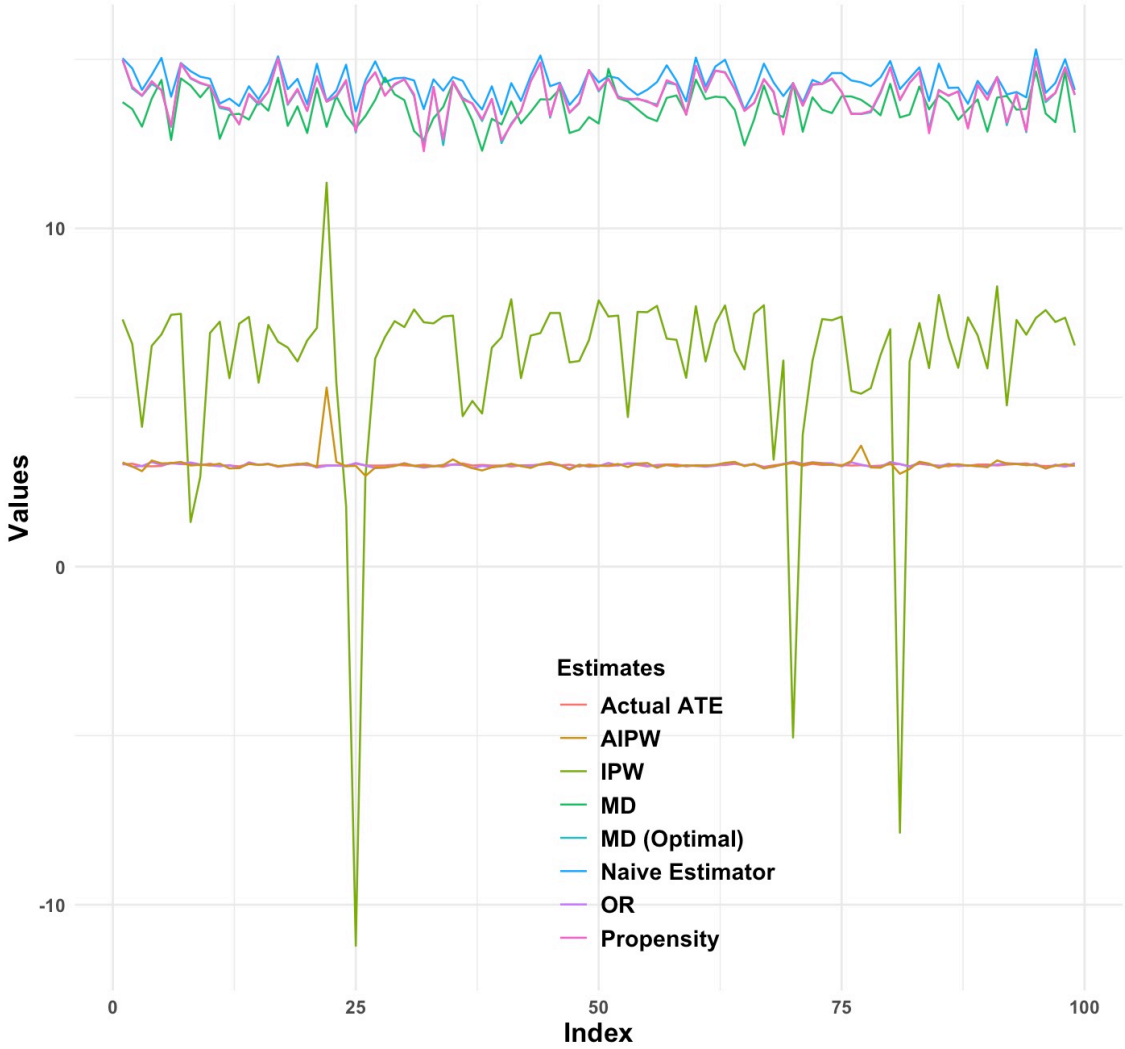
Sometimes jack-knife type or k-fold estimates are also used.

Simulation Studies

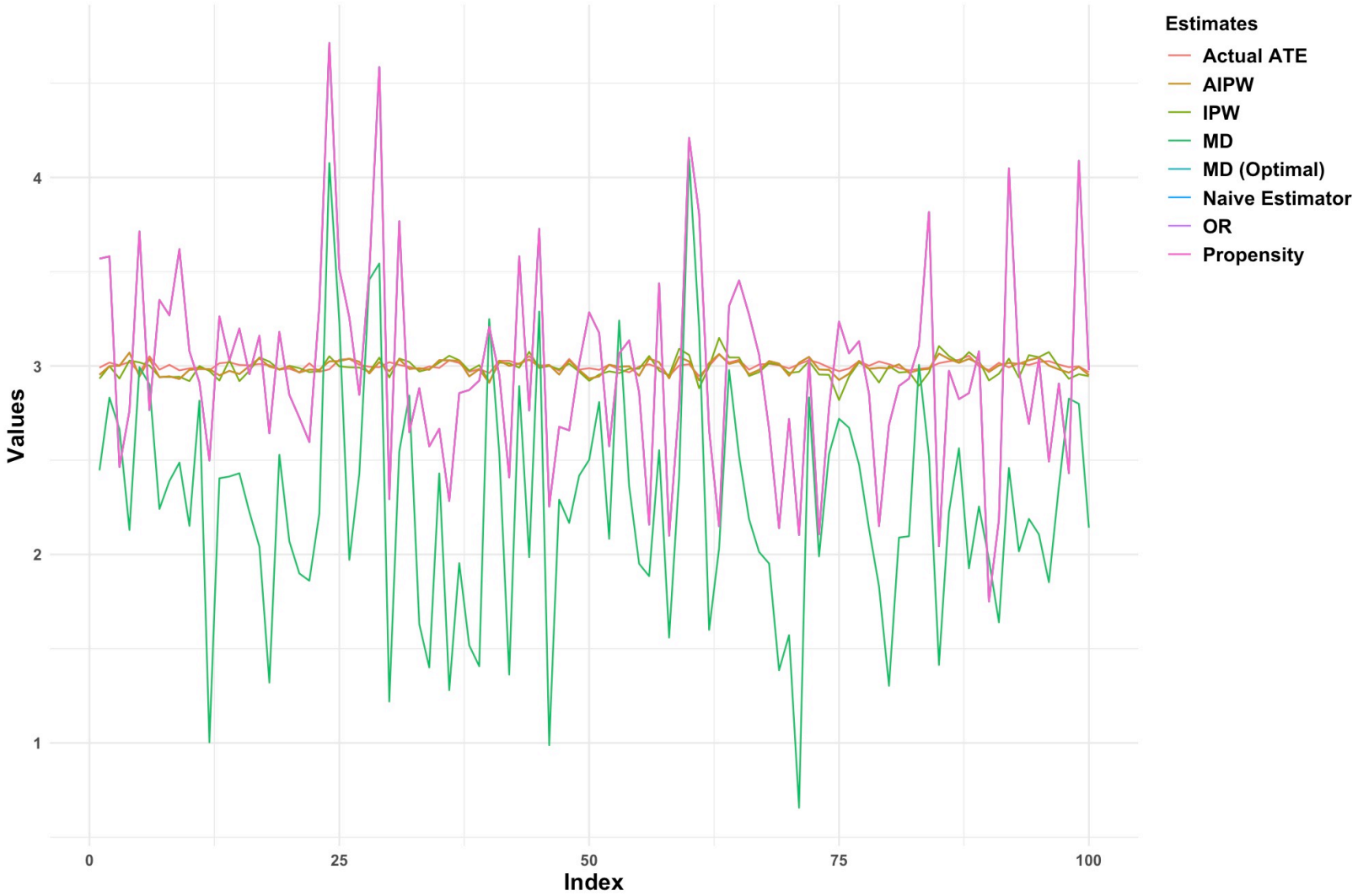
Let's work with simulated examples now!

- We generate 1000 samples $(Y, A, X_1, X_2, X_3, X_4, X_5)$.
- $X_1 \sim N(3,1), X_2 \sim N(5,4), X_3 \sim t_5, X_4 \sim Ber(0.3), X_5 \sim Pois(2)$. We scale all the variables.
- 1. Situation 1: $A_i \sim Ber(p_i)$, where $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, where $\eta_i = X_{1,i} + 5X_{2,i} + 6X_{3,i} + 2X_{4,i} + 3X_{5,i} + N(0,1)$
  2. Situation 2: Randomly assign $n/2$ people to $A = 1$.
  3. Situation 3: $A_i \overset{iid}{\sim} Ber(0.5)$
  4. Situation 4: $A_i \sim Ber(p_i)$, where $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, where $\eta_i = X_{1,i}^2 + 5X_{2,i} + 6X_{3,i} + 2X_{4,i}^2 + 3X_{5,i} + N(0,1)$
- $Y_i = (3 + 9X_{1,i} + 3X_{2,i} + 8X_{3,i} + 4X_{4,i} + 5X_{5,i} + N(0,1/4))1(A_i = 1) + (X_{1,i} + 2X_{2,i} + X_{3,i} + 2X_{4,i} + 4X_{5,i} + N(0,1/4))1(A_i = 0)$
- Repeat the whole process 100 times, i.e. generate 100 such multivariate samples.

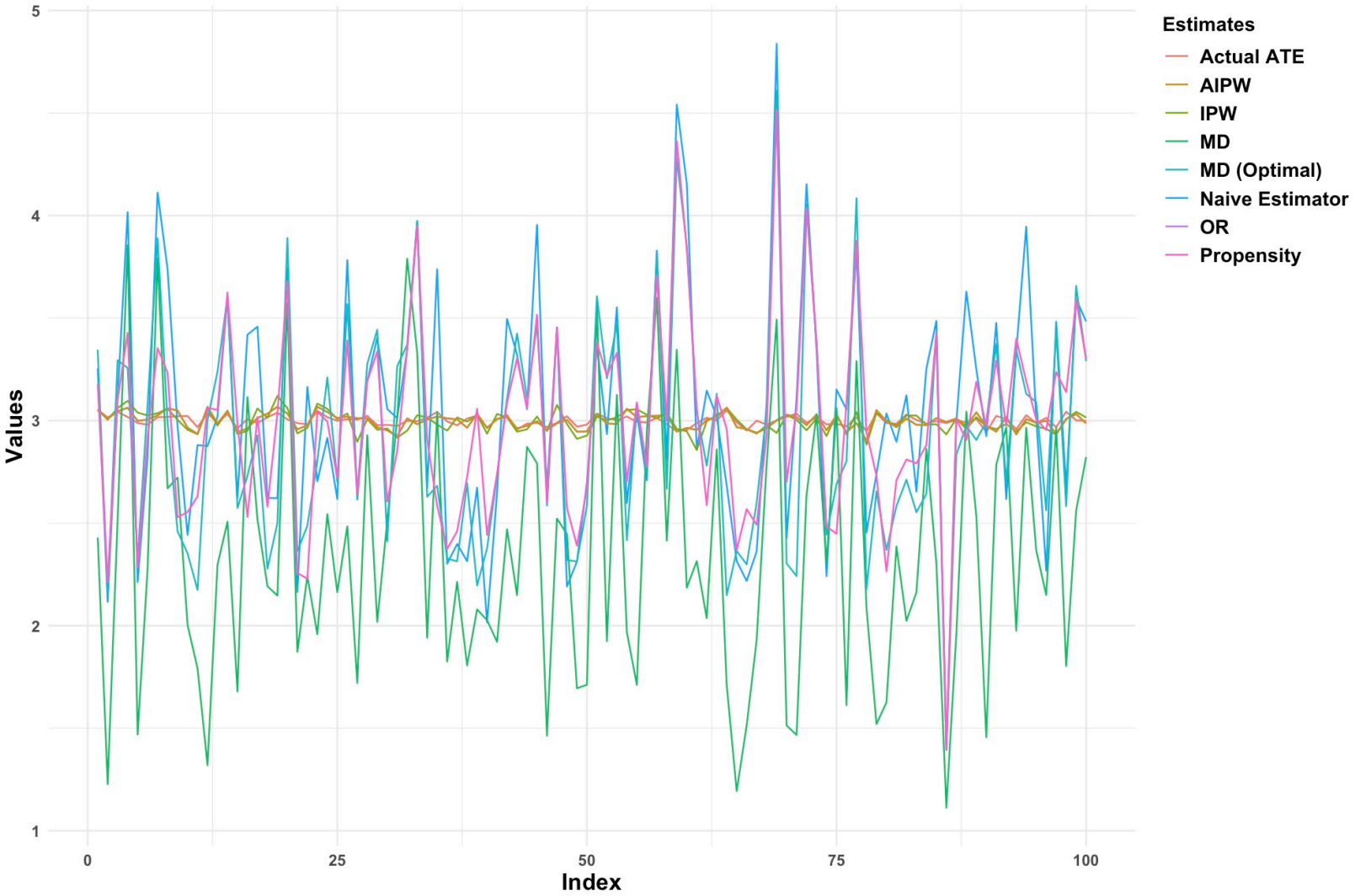
Comparing the estimates (Bernoulli With linearly mixed Covariates)



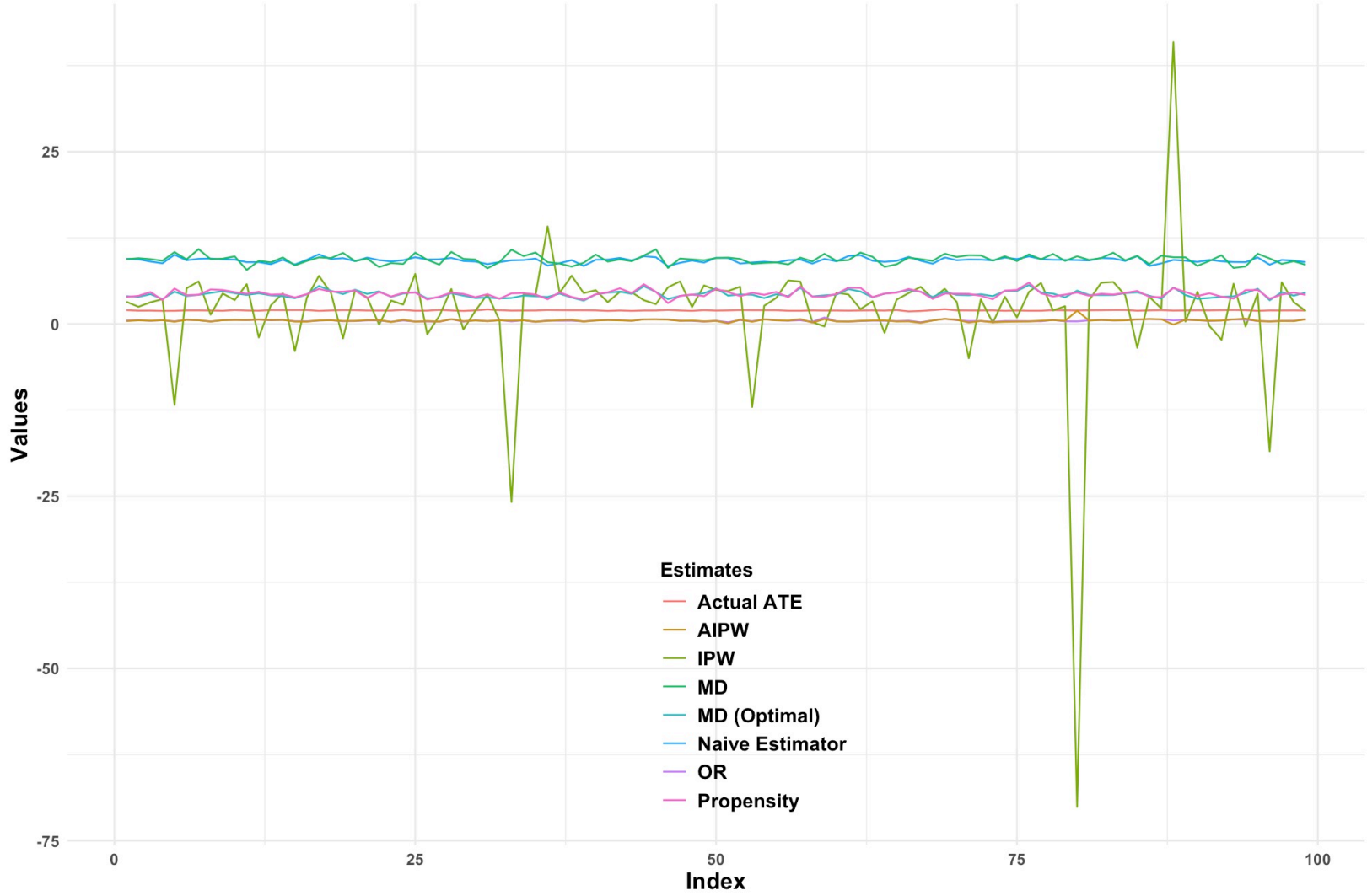
Comparing the estimates (Assigning to n/2 people)



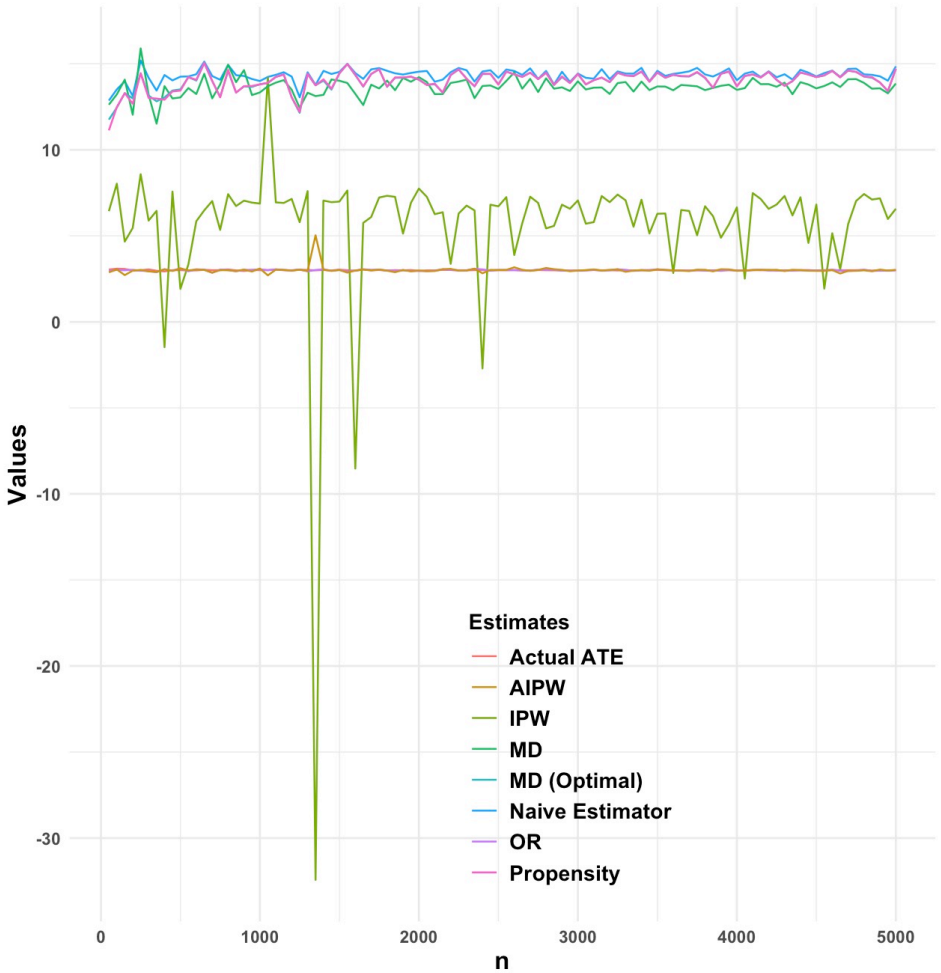
Comparing the estimates (Bernoulli (0.5))



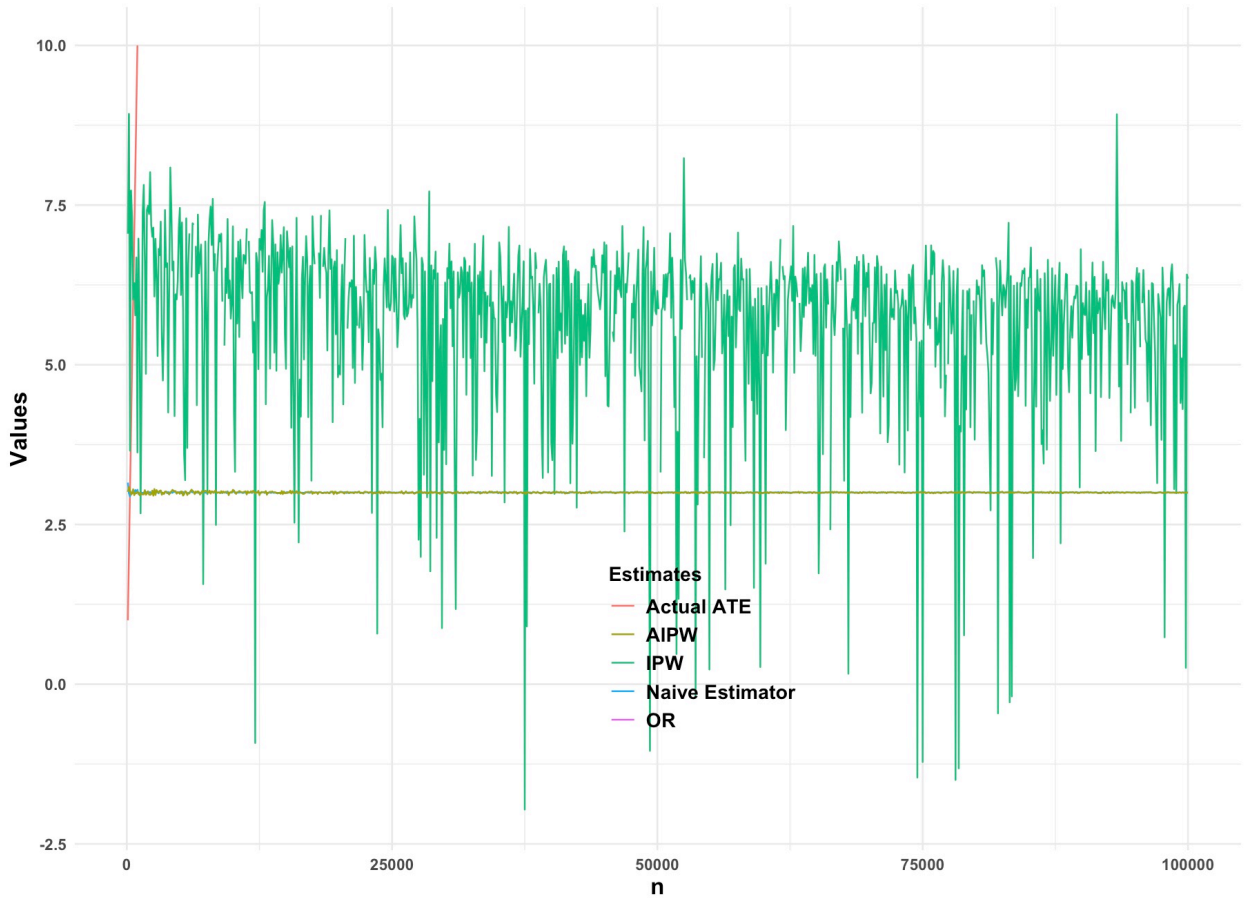
Comparing the estimates (Bernoulli With non-linearly mixed Covariates)



Comparing the estimates (Bernoulli With linearly mixed Covariates, but for various n)



Comparing the estimates (Bernoulli With linearly mixed Covariates, but for various n)



Unmeasured confounders: A Brief Discussion!

Roadmap I

- No unmeasured confounders is a rather optimistic assumption in practice.
- So the question is, does a limited violation of this assumption render our statistical analysis useless?
- For that we should do some sensitivity analysis or re-modeling to validate the outcomes.
- We will brifely discuss some of the important ones here.

Roadmap I

- No unmeasured confounders is a rather optimistic assumption in practice.
- So the question is, does a limited violation of this assumption render our statistical analysis useless?
- For that we should do some sensitivity analysis or re-modeling to validate the outcomes.
- We will brifely discuss some of the important ones here.
- Rosenbaum (1987) proposed a sensitivity analysis which is a randomisation inference for matched observational studies. It specifies a family of distributions for the full data of all the relevant factuals and counterfactuals, where η is a sensitivity parameter, then test the null hypothesis $\eta = 0$.

Roadmap I

- No unmeasured confounders is a rather optimistic assumption in practice.
- So the question is, does a limited violation of this assumption render our statistical analysis useless?
- For that we should do some sensitivity analysis or re-modeling to validate the outcomes.
- We will brifely discuss some of the important ones here.
- Rosenbaum (1987) proposed a sensitivity analysis which is a randomisation inference for matched observational studies. It specifies a family of distributions for the full data of all the relevant factuals and counterfactuals, where η is a sensitivity parameter, then test the null hypothesis $\eta = 0$.
- Robins (1999) estimates the design bias using estimated propensity score. We can then estimate the ATE by $\hat{\beta} - \hat{bias}$.

Roadmap II

- "Instrumental Variable Regression" was invented to estimate the price elasticities of (the causal effects of price on) demand and supply (Stock and Trebbi (2003)). This is a challenging problem because price is determined simultaneously by demand and supply. To estimate the causal effect of price on supply, we cannot use observational data that correspond to different demand and supply curves. Instead, we need to use "exogenous" events that change the demand but not the supply. For example, we can use the COVID-19 outbreak as an instrumental variable for the demand of masks.

A different Causal Inference : Measuring Directional Dependence

- Blöbaum et al. (2019) addresses the problem of inferring the causal direction between two variables by comparing the least-squares errors of the predictions in both possible directions. They have considered only a unconfounded situation.
- Their main idea was to simply compare the MSE of regressing Y on X and the MSE of regressing X on Y. It is somewhat similar to the Structural Equation Modeling approach (where a linear relationship between the variables are assumed).
- It is very natural to assume that $\mathbb{E}[(E - \mathbb{E}(E|C))^2] \leq \mathbb{E}[(C - \mathbb{E}(C|E))^2]$, or, equivalently, $\mathbb{E}(Var(E|C)) \leq \mathbb{E}(Var(C|E))$ where C is the Actual Cause and E is the Effect.
- Define $\phi(c) = \mathbb{E}(E|c)$ and Noise variable $N = E - \phi(C)$ and to study the limit of an almost deterministic relation in a mathematically precise way, they considered a family of effect variables E_α by $E_\alpha = \phi(C) + \alpha N$, where $\alpha \in \mathbb{R}^+$ is a parameter controlling the noise level.
- They showed under some assumptions, $\lim_{\alpha \to 0} \frac{\mathbb{E}(Var(C|E'_\alpha)}{\mathbb{E}(Var(E'_\alpha|C)} \geq 1$, where E'_α is the rescaled and centered version of E_α. Based on this result, they propose the following algorithm.

Algorithm 1

Algorithm 1 The proposed causal inference algorithm.

function RECI(X, Y) ▷ X and Y are the observed data.
  $(X, Y) \leftarrow \text{RescaleData}(X, Y)$
  $f \leftarrow \text{FitModel}(X, Y)$ ▷ Fit regression model $f : X \rightarrow Y$
  $g \leftarrow \text{FitModel}(Y, X)$ ▷ Fit regression model $g : Y \rightarrow X$
  $\text{MSE}_{Y|X} \leftarrow \text{MeanSquaredError}(f, X, Y)$
  $\text{MSE}_{X|Y} \leftarrow \text{MeanSquaredError}(g, Y, X)$
  if $\text{MSE}_{Y|X} < \text{MSE}_{X|Y}$ **then**
    return X causes Y
  else if $\text{MSE}_{X|Y} < \text{MSE}_{Y|X}$ **then**
    return Y causes X
  else
    return No decision
  end if
end function

Algorithm 2

Algorithm 2 Causal inference algorithm

function RECI(X, Y, t) ▷ X and Y are the observed data and $t \in [0, 1]$ is the confidence threshold for rejecting a decision.

$(X, Y) \leftarrow \text{RescaleData}(X, Y)$

$f \leftarrow \text{FitModel}(X, Y)$ ▷ Fit regression model $f : X \to Y$

$g \leftarrow \text{FitModel}(Y, X)$ ▷ Fit regression model $g : Y \to X$

$\text{MSE}_{Y|X} \leftarrow \text{MeanSquaredError}(f, X, Y)$

$\text{MSE}_{X|Y} \leftarrow \text{MeanSquaredError}(g, Y, X)$

$\xi \leftarrow 1 - \frac{\min(\text{MSE}_{X|Y}, \text{MSE}_{Y|X})}{\max(\text{MSE}_{X|Y}, \text{MSE}_{Y|X})}$

if $\xi \geq t$ **then**

  if $\text{MSE}_{Y|X} < \text{MSE}_{X|Y}$ **then**

    return X causes Y

  else

    return Y causes X

  end if

else

  return No decision

end if

end function

A Modified Approach

- Pascual-Marqui et al. (2024) uses Chatterjee's Correlation (Chatterjee (2021)) and uses the same type of algorithm as of Blöbaum et al. (2019) to infer the directional causal relation.
- Consider $(x_{(1)}, y_{(1)}), \cdots, (x_{(N)}, y_{(N)})$ and r_i be the rank of $y_{(i)}$. Then, $Ch[y = f(x)] = 1 - \frac{3\sum_{k=1}^{N-1} |r_{k+1} - r_k|}{N^2 - 1}$ is also interpreted as a measure of how well "y is predicted by x".
- Then they define $\Delta(x \to y) = Ch[y = f(x)] - Ch[x = g(y)]$. We say X causes Y if $\Delta(x \to y) > 0$ and Y causes X if $\Delta(x \to y) < 0$.

More on Causal Inference I

- Yao et al. (2021) talks about relaxing the crucial assumptions like SUTVA, Unconfoundedness, and Positivity etc.
- Unconfoundedness assumption is not testable in practice. Kallus et al. (2018) suggests to combine the experimental data and observational data together which makes strictly weaker assumptions than existing approaches. D'Amour et al. (2021) argues that the positivity assumption is a strong assumption and is more difficult to be satisfied in the high-dimensional datasets.
- Dawid (2000) criticizes the counterfactual approach due to unverifiable assumptions and proposes a bayesian decision analytic approach for causal inference. The principal difficulty with the counterfactual approach is that the desired inference depends on the joint probability structure of the complementary variables (Y_0, Y_1). However, at a time only one is observable. Dawid (2000) claims that by using bayesian decision analytic approach, this problem somewhat becomes negligible.

More on Causal Inference II

- Guo et al. (1906) applies Graph Convolutional Networks into a causal inference model is an approach to handle the network data as independence assumption is not applicable to network data.
- Causal inference has been discovered for time series data (Runge et al. (2023)), spatial data (Akbari et al. (2023)) as well.

References I

Akbari, K., Winter, S., and Tomko, M. (2023). Spatial causality: A systematic review on spatial causal inference. *Geographical Analysis*, 55(1):56–89.

Blöbaum, P., Janzing, D., Washio, T., Shimizu, S., and Schölkopf, B. (2019). Analysis of cause-effect inference by comparing regression errors. *PeerJ Computer Science*, 5:e169.

Chatterjee, S. (2021). A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022.

D'Amour, A., Ding, P., Feller, A., Lei, L., and Sekhon, J. (2021). Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654.

Dawid, A. P. (2000). Causal inference without counterfactuals. *Journal of the American statistical Association*, 95(450):407–424.

Edwards, A. W. (2005). Ra fischer, statistical methods for research workers, (1925). In *Landmark writings in western mathematics 1640-1940*, pages 856–870. Elsevier.

Guo, R., Li, J., and Liu, H. (1906). Learning individual treat-ment effects from networked observational data (2019). *Preprint arXiv*.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960.

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 171(2):481–502.

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.

References II

Kallus, N., Puli, A. M., and Shalit, U. (2018). Removing hidden confounding by experimental grounding. *Advances in neural information processing systems*, 31.

Kusner, M. J. and Loftus, J. R. (2020). The long road to fairer algorithms. *Nature*, 578(7793):34–36.

Little, R. J. (2019). Comment: "models as approximations i: Consequences illustrated with linear regression" by a. buja, r. berk, l. brown, e. george, e. pitkin, l. zhan and k. zhang.

Pascual-Marqui, R. D., Kochi, K., and Kinoshita, T. (2024). Distance-based chatterjee correlation: a new generalized robust measure of directed association for multivariate real and complex-valued data. *arXiv preprint arXiv:2406.16458*.

Robins, J. M. (1999). Association, causation, and marginal structural models. *Synthese*, 121(1/2):151–179.

Rosenbaum, P. R. (1987). Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26.

Rosenbaum, P. R. (2020). Modern algorithms for matching in observational studies. *Annual Review of Statistics and Its Application*, 7(1):143–176.

Rosenbaum, P. R. (2023). Propensity score. In *Handbook of matching and weighting adjustments for causal inference*, pages 21–38. Chapman and Hall/CRC.

Rosenbaum, P. R. and Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39(1):33–38.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688.

Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Earth & Environment*, 4(7):487–505.

References III

Splawa-Neyman, J., Dabrowska, D. M., and Speed, T. P. (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472.

Stock, J. H. and Trebbi, F. (2003). Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194.

Yao, L., Chu, Z., Li, S., Li, Y., Gao, J., and Zhang, A. (2021). A survey on causal inference. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(5):1–46.

Lecture Notes on Causal Inference, Qingyuan Zhao, May 30, 2022.

STATS 361: Causal Inference, Stefan Wager, Stanford University, Spring 2022

AI image and text generator, Google

Thank You!!