Identification Under Strong Ignorability

Hyunseung Kang

Stat 992: Topics in Causal Inference
Apr. 10, 2024

Review: Causal Identification Under Complete Randomized Experiment

Under an ideal, complete randomized experiment, the following assumptions are satisfied:

- (A1,SUTVA): $Y = AY(1) + (1 - A)Y(0)$
- (A2, Randomization of A): $A \perp X, Y(1), Y(0)$
- (A3, Positivity/Overlap): $0 < P(A = 1) < 1$

These assumptions were motivated from missing data literature:

	$Y(1)$	$Y(0)$	Y	A	X_{Age}
John	NA	0.9	0.9	0	38
Sally	0.8	NA	0.8	1	30
Kate	NA	0.6	0.6	0	23
Jason	0.6	NA	0.6	1	26

To identify the column mean of $Y(1)$, we can take the observed $Y(1)$. This was valid as long as the missingness was completely at random, i.e. $\mathbb{E}[Y(1)] = \mathbb{E}[Y|A = 1]$ if $A \perp Y(1)$.

Stratified/Block Randomized Experiments

- In most randomized experiments, treatment is not randomized completely at random.
- Often, treatment is randomized within a pre-defined block of individuals based on their covariates X in order to improve precision.
- This type of randomized experiment is broadly known as **stratified/blocked experiments.**
  - Subdividing above table by $X_{\text{under30}} = I(X_{\text{Age}} < 30)$ and randomizing treatment within each block.
  - Twin experiments where treatment is randomized within each twin.
- Treatment probabilities can be different across blocks. But, within each block, the treatment is assigned randomly.

Assumptions Behind Stratified Randomized Experiments

We can formalize the assumptions under a stratified randomized experiment as follows:

- (A1, SUTVA): $Y = AY(1) + (1 - A)Y(0)$
- (A2c, Conditional randomization of A): $A \perp Y(1), Y(0)|X$
- (A3c', Positivity/Overlap): $0 < P(A = 1|X = x) < 1$ for all x

For example, if $X = X_{\text{under30}}$:

- (A2c) states that conditional on different age group, treatment A is randomly assigned to individuals
- (A3c) states that conditional on different age group, each person has a non-zero probability of receiving treatment or control. Note that the treatment probability may be different across age groups, i.e. $P(A = 1|X_{\text{under30}} = 1)$ may not be equal to $P(A = 1|X_{\text{under30}} = 0)$.

Assumptions (A2c) and (A3c) are known as **strong ignorability** (Rosenbaum and Rubin (1983))

Connection to Complete Randomized Experiments

Stratified randomized experiment:

$$(\mathrm{A2c}) : A \perp Y(1), Y(0)|X, \quad (\mathrm{A3c}) 0 < P(A = 1|X = x) < 1 \forall x$$

Complete randomized experiment:

$$(\mathrm{A2}) : A \perp Y(1), Y(0), X, \quad (\mathrm{A3c}) 0 < P(A = 1) < 1$$

Intuitively, if the treatment was randomized completely at random to everyone, the treatment is also randomized to a subgroup of individuals defined by their covariates.

- Formally, we can show (A2) implies (A2c); see lecture notes.
- We can also show (A3) and (A2) implies (A3c). Without (A2), it's not always the case that (A3) implies (A3c); see lecture notes

Connection to Missing Data

Assumptions (A2c) and (A3c) have connections to the **missing at random (MAR)** assumption in the missing data literature.

Consider the data table partitioned by age:

	$Y(1)$	$Y(0)$	Y	A	X_{under30}
John	NA	0.9	0.9	0	0
Sally	0.8	NA	0.8	1	0
Kate	NA	0.6	0.6	0	1
Jason	0.6	NA	0.6	1	1

- (A2c): within the rows of the sub-table where Xs are identical (i.e. conditional on X), the missingness indicator A is completely independent of the columns $Y(1), Y(0)$.
- (A3c) states that within the rows of the sub-table, some values of $Y(1)$ (or $Y(0)$) are observed and this holds for every sub-table.

Causal Identification Under Strong Ignorability (i.e. Assumptions (A1),(A2c), and (A3c))

Identification of the ATE

Under a stratified randomized experiment (i.e. where strong ignorability holds), identification of the ATE among a subgroup defined by X, i.e. $\tau(x) = \mathbb{E}[Y(1) - Y(0)|X = x]$ is immediate.

- Intuitively, identification is achieved by considering the sub-table of people with $X = x$.
- Then, similar to a complete randomized experiment, we can identify $\mathbb{E}[Y(1)|X = x]$ by taking the average of the observed $Y(1)$ within the sub-table.
- $\tau(x)$ is known as **the conditional average treatment effect (CATE)**.

We can take the average of $\tau(X)$ over the distribution of X to identify the ATE:

$$\mathbb{E}[Y(1) - Y(0)] = \mathbb{E}[\mathbb{E}[Y|A = 1, X]] - \mathbb{E}[\mathbb{E}[Y|A = 0, X]]$$

Formal Proof

$$\begin{aligned}
\mathbb{E}[Y|A=1,X=x] &= \mathbb{E}[AY(1)+(1-A)Y(0)|A=1,X=x] && \text{(A1)}\\
&= \mathbb{E}[Y(1)|A=1,X=x] \\
&= \mathbb{E}[Y(1)|X=x] && \text{(A2c)}
\end{aligned}$$

Assumption (A3c) ensures that the conditioning event $\mathbb{E}[Y|A=1,X=x]$ is well-defined.

By the law of total expectation, we can also identify the unconditional mean $\mathbb{E}[Y(1)]$ as follows

$$\begin{aligned}
\mathbb{E}[Y(1)] &= \mathbb{E}[\mathbb{E}[Y(1)|X]] && \text{Law of total expectation}\\
&= \mathbb{E}[\mathbb{E}[Y|A=1,X]] && \text{Argument from above}
\end{aligned}$$

Using a similar argument, we get $\mathbb{E}[Y(0)] = \mathbb{E}[\mathbb{E}[Y|A=0,X]]$.

Identification of the Average Treatment Effect Among the Treated (ATT)

Another popular causal estimand is the average treatment effect among the treated (ATT)

$$ATT = \mathbb{E}[Y(1) - Y(0) \mid A = 1]$$

	$Y(1)$	$Y(0)$	Y	A	X_{Age}
John	NA	0.9	0.9	0	38
Sally	0.8	NA	0.8	1	30
Kate	NA	0.6	0.6	0	23
Jason	0.6	NA	0.6	1	26

The ATT represents the average difference of $Y(1) - Y(0)$ among Sally and Jason, both of whom were treated.

Note that the ATT is different than the ATE, which is the average of $Y(1) - Y(0)$ for both treated and untreated individuals.

A Minor Change in Assumptions Under ATT

A unique feature of the ATT is that you can estimate this causal effect by a weaker version of strong ignorability, i.e.

$$(\mathrm{A2c.0}): \quad A \perp Y(0) \mid X$$

where A does not have to be independent of $Y(1)$ given X, i.e.

$$(\mathrm{A2c}): \quad A \perp Y(1), Y(0) \mid X$$

Missing data perspective: we only need the missingness indicator to be independent of the column $Y(0)$, not necessarily with the column $Y(1)$.

From my experience, the practical difference between (A2c) and (A2c.0) where investigators discuss whether plausibility of assumptions in observational studies, is minor.

Formal Proof

The term $\mathbb{E}[Y(1) \mid A = 1]$ can be identified with just (A1):

$$
\begin{aligned}
&\mathbb{E}[Y(1) \mid A = 1] \\
=&\mathbb{E}[\mathbb{E}[Y(1) \mid A = 1, X] \mid A = 1] && \text{Law of total expectation} \\
=&\mathbb{E}[\mathbb{E}[Y \mid A = 1, X] \mid A = 1] && \text{(A1)}
\end{aligned}
$$

The term $\mathbb{E}[Y(0) \mid A = 1]$ can be identified with (A1),(A2c.0) and (A3).

$$
\begin{aligned}
&\mathbb{E}[Y(0) \mid A = 1] \\
=&\mathbb{E}[\mathbb{E}[Y(0) \mid A = 1, X] \mid A = 1] && \text{Law of total expectation} \\
=&\mathbb{E}[\mathbb{E}[Y(0) \mid A = 0, X] \mid A = 1] && \text{(A2c.0) and (A3)} \\
=&\mathbb{E}[\mathbb{E}[Y \mid A = 0, X] \mid A = 1] && \text{(A1)}
\end{aligned}
$$

Hence, under (A1), (A2c.0), and (A3), we can identify the ATT via

$$\mathbb{E}[Y(1)-Y(0) \mid A = 1] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, X] \mid A = 1]-\mathbb{E}[\mathbb{E}[Y \mid A = 0, X] \mid A = 1]$$

Identification of Other Measures of Causal Effects: Causal Relative Risk (CRR) and Causal Odds Ratio (COR)

Under a binary outcome, some popular causal estimands are the causal relative risk (CRR) or causal odds ratio (COR):

$$\text{CRR} = \frac{\mathbb{E}[Y(1)]}{\mathbb{E}[Y(0)]} = \frac{\mathbb{P}(Y(1)=1)}{\mathbb{P}(Y(0)=1)}$$

$$\text{COR} = \frac{\frac{\mathbb{P}(Y(1)=1)}{1-\mathbb{P}(Y(1)=1)}}{\frac{\mathbb{P}(Y(0)=1)}{1-\mathbb{P}(Y(0)=1)}}$$

- There are some issues with defining causal odds ratios (or more generally odds ratios). I recommend using CRRs instead of CORs unless the scientific question is expressed in odds ratios.
- The original ATE $E[Y_i(1) - Y_i(0)]$, or a linear contrast of the outcomes, is still well-defined for binary outcomes.

Formal Proof

Identification of the CRR or the COR often proceeds by identifying $\mathbb{E}[Y(a)]$ for any a.

Formally, we have

$$\begin{aligned}
\mathbb{E}[Y(a)] &= \mathbb{E}[\mathbb{E}[Y(a) \mid X]] && \text{Law of total expectation} \\
&= \mathbb{E}[\mathbb{E}[Y(a) \mid A = a, X]] && \text{(A2c) and (A3c)} \\
&= \mathbb{E}[\mathbb{E}[Y \mid A = a, X]] && \text{(A1)}
\end{aligned}$$

Note that we need (A3c) to ensure that the conditioning event $\{A = a, X\}$ is well-defined.

Then, under (A1), (A2c), and (A3c), CRR and COR are identified as

$$\mathrm{CRR} = \frac{\mathbb{E}[\mathbb{E}[Y \mid A = 1, X]]}{\mathbb{E}[\mathbb{E}[Y \mid A = 0, X]]}$$

$$\mathrm{COR} = \frac{\frac{\mathbb{E}[\mathbb{E}[Y|A=1,X]]}{1-\mathbb{E}[\mathbb{E}[Y|A=1,X]]}}{\frac{\mathbb{E}[\mathbb{E}[Y|A=0,X]]}{1-\mathbb{E}[\mathbb{E}[Y|A=0,X]]}}$$

Identification of Single, Static, Optimal Treatment Regime/Policy (OTR)

In personalized medicine, the goal is to develop an optimal treatment assignment policy where the patient receives the treatment that maximizes the patient's outcome.

Formally, consider a policy function $\pi : \mathcal{X} \to \{0, 1\}$ which assigns either treatment (i.e 1) or control (i.e 0) based on the individual's characteristic $X \in \mathcal{X}$.

The goal is to find the best π, denoted as π_{OTR}, that maximizes the expected counterfactual outcome:

$$\pi_{\text{OTR}} = \underset{\pi \in \Pi}{\text{argmax}}\ \mathbb{E}[Y(\pi(X))]$$

Π represents all policy functions of the form $f : \mathcal{X} \to \{0, 1\}$

Value Function

$$\pi_{\text{OTR}} = \underset{\pi \in \Pi}{\text{argmax}}\ \mathbb{E}[Y(\pi(X))]$$

The term $Y(\pi(X))$ is the counterfactual outcome if treatment is assigned based on π and can be written as

$$Y(\pi(X)) = Y(1)I(\pi(X) = 1) + Y(0)I(\pi(X) = 0)$$

The term $\mathbb{E}[Y(\pi(X))]$ takes an average of the counterfactual outcome under policy π and is called the **value** of π.

- For example, the value of a policy that always assigns treatment, i.e. $\pi(X) = 1$, is $\mathbb{E}[Y(\pi(X))] = \mathbb{E}[Y(1)]$
- The value of a policy that assigns control, i.e. $\pi(X) = 0$, is $\mathbb{E}[Y(\pi(X))] = \mathbb{E}[Y(0)]$

Causal Identification of the Value Function

Given any policy π, we can identify its value under assumptions (A1), (A2c), and (A3c)

$$
\begin{aligned}
&\mathbb{E}[Y(\pi(X))] \\
=&\mathbb{E}[Y(1)I(\pi(X)=1)+Y(0)I(\pi(X)=0)] && \text{Definition} \\
=&\mathbb{E}[\mathbb{E}[Y(1)I(\pi(X)=1)+Y(0)I(\pi(X)=0)\mid X]] && \text{Law of total exp.} \\
=&\mathbb{E}[I(\pi(X)=1)\mathbb{E}[Y(1)\mid X]+I(\pi(X)=0)\mathbb{E}[Y(0)\mid X]] \\
=&\mathbb{E}[I(\pi(X)=1)\mathbb{E}[Y\mid A=1,X]+I(\pi(X)=0)\mathbb{E}[Y\mid A=0,X]] && \text{(A1), (A2c), (A3c)}
\end{aligned}
$$

The last equality follows from the identification of the ATE.

Note that the identification result holds for any policy π.

Causal Identification of Optimal Policy

Once we identified the value function, we don't need any more assumptions to identify the optimal policy.

Let $\mu_a(x) = \mathbb{E}[Y \mid A = a, X = x]$. Then,

$$
\begin{aligned}
\pi_{\mathrm{OTR}} &= \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}[Y(\pi(X))] \\
&= \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}[I(\pi(X) = 1)\mu_1(X) + I(\pi(X) = 0)\mu_0(X)] \\
&= \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}[\pi(X)\mu_1(X) + (1 - \pi(X))\mu_0(X)] \\
&= \underset{\pi}{\operatorname{argmax}}\ \mathbb{E}[\pi(X)(\mu_1(X) - \mu_0(X))] \\
&= I(\mu_1(X) - \mu_0(X) \geq 0)
\end{aligned}
$$

The optimal treatment policy π_{OTR} for a person with characteristic X is to check whether the expected outcome among people with X is larger under treatment (i.e. $\mu_1(X)$) or under control (i.e. $\mu_0(X)$).

- If $\mu_1(X) \geq \mu_0(X)$, the person should be treated.
- If $\mu_1(X) < \mu_0(X)$, the person should get control.

Some Details About Proof

Let $\Delta(x) = \mu_1(x) - \mu_0(x)$. Then, the second to the last equality becomes

$$\begin{aligned}
&\mathbb{E}[\pi(X)(\mu_1(X) - \mu_0(X))] \\
=&\mathbb{E}[\pi(X)\Delta(x)\{I(\Delta(x) \geq 0) + I(\Delta(X) < 0)\}] \\
=&\underbrace{\mathbb{E}[\pi(X)\Delta(x)I(\Delta(x) \geq 0)]}_{\text{non-negative}} + \underbrace{\mathbb{E}[\pi(X)\Delta(x)I(\Delta(X) < 0)]}_{\text{non-positive}}
\end{aligned}$$

To find π that maximize the above expression, we need

- $\pi(X) = 0$ whenever $\Delta(X) < 0$ to maximize the non-positive term
- $\pi(X) = 1$ whenever $\Delta(X) > 0$ to maximize the non-negative term

Combining these two observations, we arrive at $\pi_{\text{OTR}}(X) = I(\Delta(X) \geq 0)$.

Observational Studies and Strong Ignorability

When studying observational studies for causal effects, several works assume that we have measured pre-treatment covariates X where the treatment A can be considered "as-if" random conditional on them, akin to a stratified randomized experiment.

Another way to interpret these assumptions in the context of observational studies are

- We measured all the **confounders** in the observational study (i.e. X) and these variables satisfy (A2c) and (A3c) above.
- There are **no unmeasured confounders** U that can influence the propensity for someone to be treated (or receive control). A bit more formally, we do not have the case where

$$A \perp Y(1), Y(0)|X, U \quad \text{but} \quad A \not\perp Y(1), Y(0)|X$$

- Self-selection into treatment (or control) does not depend on anything except X.
- If (A2c) and (A3c) hold in an observational study, we must adjust/control for X in order to identify the ATE.

Observational Study and Randomized Experiments

See Cochran (1965),Rubin (2007), and a very recent, nice article by Small (2024) for further discussion about studying observational studies from the lens of a randomized experiment.

- **In a randomized experiment, the propensity score $e(X)$ is known by the investigator**. In contrast, in an observational study, $e(X)$ is not known since individual's selection into treatment cannot be controlled by the investigator.
- There is a push in observational studies to blind the outcome, akin to a randomized experiment where the investigator is blind to the outcome by design. Specifically, investigators should focus on X and treatment assignment A, especially achieving balance in the form of $X \perp A \mid e(X)$, before seeing the outcome.

Central Role of the Propensity Score $\mathbb{P}(A = 1|X)$

We highlight the two most important properties of the propensity score.

Consider any function $b(X)$ of the covariates. This function b is called a balancing score if conditional on $b(X)$, the treatment is independent of X, i.e.

$$A \perp X|b(X)$$

A couple of remarks:

- A trivial function b that satisfies this condition is the identity function $b(X) = X$.
- Theorem 1 of Rosenbaum and Rubin (1983) showed that the propensity score $e(X)$ is a balancing score; see their Theorem 1.

Propensity Score is the Coarsest Balancing Score

Theorem 2 of Rosenbaum and Rubin (1983): $b(X)$ is a balancing score if and only if $b(X)$ is finer than the propensity score $e(X)$, i.e. if there exists a function g where $e(X) = g(b(X))$.

- The propensity score contains the "smallest'' amount of information to achieve $A \perp X|b(X)$; **the propensity score is the coarsest balancing score**.
- To intuitively check this, consider setting $b(X) = X$. This is not only a balancing score, but also provides much more information (i.e. finer information) than the propensity score $P(A = 1|X = x)$, which is a number between 0 and 1.
- In the above case, $e(X) = e(b(X))$ where $g = e$.

Propensity Score Is Sufficient for Strong Ignorability

Theorem 3 of Rosenbaum and Rubin (1983): Let $e(X) = \mathbb{P}(A = 1|X)$. If conditions (A1), (A2c), and (A3c) hold, then we have

$$A \perp Y(1), Y(0)|e(X) \text{ and } 0 < \mathbb{P}(A = 1|e(X)) < 1$$

Some implications:

- If (A1),(A2c), and (A3c) hold for X, then these assumptions also hold for a scalar summary of X, i.e.$e(X)$.
- We can identify the ATE via

  $$\mathbb{E}[Y(1)-Y(0)] = \mathbb{E}[\mathbb{E}[Y \mid A = 1, e(X)]]-\mathbb{E}[\mathbb{E}[Y \mid A = 1, e(X)]]$$

  The proof of this follows directly from the proof of the identification of the ATE where we replace X with $e(X)$.
- In completely randomized trial where (A2) and (A3) held, we had $A \perp X$ and covariates were balanced. Under (A2c) and (A3c), we now have $A \perp X \mid e(X)$ or covariates are balanced conditional on the propensity score $e(X)$.

References

Cochran, William G. 1965. "The Planning of Observational Studies of Human Populations." *Journal of the Royal Statistical Society. Series A (General)* 128 (2): 234–66.

Greenland, Sander, James M Robins, and Judea Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 14 (1): 29–46.

Hernán, Miguel A, David Clayton, and Niels Keiding. 2011. "The Simpson's Paradox Unraveled." *International Journal of Epidemiology* 40 (3): 780–85.

Rosenbaum, Paul, and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55.

Rubin, Donald B. 2007. "The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels with the Design of Randomized Trials." *Statistics in Medicine* 26 (1): 20–36.

Small, Dylan S. 2024. "Protocols for Observational Studies: Methods and Open Problems." *arXiv Preprint arXiv:2403.19807*.