Rubin's causal model

Adam Gleave

25th September 2017

University of California Berkeley

Potential outcome model

- Notation introduced by Neyman.
- Suppose there are M possible treatments; e.g. different drugs.
- Let $Y(j)$ denote the potential outcome for treatment j. This is what we would measure if treatment j were administered.
- Once a treatment k has been chosen, $Y(k)$ is observed and $Y(j)$ for $j \neq k$ are counterfactuals.
- For simplicity, we will take $M = 2$, and call $j = 1$ treatment and $j = 0$ control.

The causal estimand

- N units; e.g. individual people.
- Each unit i has:
  - a set of covariates X_i, that are independent of the treatment.
  - potential outcomes $Y_i(1)$ and $Y_i(0)$ for the treatment and controls.
- Does this definition make sense?

SUTVA

- Need to make Stable Unit Treatment Value Assumption (SUTVA) for the causal estimand to be well-defined.
  - No interference between units: $Y_i(1)$ and $Y_i(0)$ are unaffected by what treatment *other* units received.
  - No hidden treatments: $Y_i(j)$ will be observed no matter how treatment j is administered.
- Further assumption: covariates and potential outcomes are not affected by the study.
- How realistic are these assumptions?

Individual causal effects

- We can define individual causal effects in terms of $Y_i(1)$ and $Y_i(0)$.
- Typically we use the difference, $\delta_i = Y_i(1) - Y_i(0)$.
- Sometimes the ratio $\frac{Y_i(1)}{Y_i(0)}$ is used instead, especially when calculating risk ratios.
- Can only ever observe one of $Y_i(1)$ and $Y_i(0)$. So no way to compute Δ_i. This is the fundamental problem of causal inference.

Assignment mechanism

- Let $W = (W_1, \ldots, W_N)$ be a random variable, where W_i gives the treatment assignment for unit i.
- An assignment mechanism is a probability distribution on W, conditional on the covariates X, and potential outcomes $Y(0)$ and $Y(1)$:

$$\mathbb{P}(W \mid X, Y(1), Y(0)).$$

- We can define the observed and missing outcomes in terms of W:

$$Y_{\text{obs},i} = W_i Y_i(1) + (1 - W_i) Y_i(0),$$

$$Y_{\text{mis},i} = (1 - W_i) Y_i(1) + W_i Y_i(0).$$

Randomized experiment

- Randomized experiments are a special type of assignment mechanism, satisfying two conditions.
- First, they are ignorable:

$$\mathbb{P}(W \mid X, Y(1), Y(0)) = \mathbb{P}(W \mid X, Y_{\text{obs}}).$$

- Second, they assign non-zero probability to both treatment and control:

$$0 < \mathbb{P}(W_i = 1 \mid X, Y_{\text{obs}}) < 1.$$

- A stronger condition is that the assignment mechanism is unconfounded:

$$\mathbb{P}(W \mid X, Y(0), Y(1)) = \mathbb{P}(W \mid X).$$

Summary causal effects

- Let S be a subset of the units in the causal estimand. A summary causal effect is a comparison between the ordered sets $\{Y_i(1) \mid i \in S\}$ and $\{Y_i(1) \mid i \in S\}$.
- The Average Treatment Effect (ATE) is:

$$\text{ATE} = \mathbb{E}[\Delta] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0],$$

  where $\Delta = Y^1 - Y^0$.
- Average Treatment Control (ATC) and Average Treatment Treatment (ATT) correspond to the ATE restricted to the control and treatment group respectively:

$$\text{ATC} = \mathbb{E}[\Delta \mid W = 0],$$

$$\text{ATT} = \mathbb{E}[\Delta \mid W = 1].$$

Estimating the ATE

- Decomposing the ATE:

$$\begin{aligned}\text{ATE} = \pi\mathbb{E}[Y^1 \mid W = 1] + (1-\pi)\mathbb{E}[Y^1 \mid W = 0] \\ - \pi\mathbb{E}[Y^0 \mid W = 1] - (1-\pi)\mathbb{E}[Y^0 \mid W = 0],\end{aligned}$$

  where $\pi = \mathbb{E}[W]$, the proportion given treatment.
- Observational data gives us consistent and unbiased estimators for π, $\mathbb{E}[Y^1 \mid W = 1]$ and $\mathbb{E}[Y^0 \mid W = 0]$. But $\mathbb{E}[Y^1 \mid W = 0]$ and $\mathbb{E}[Y^0 \mid W = 1]$ are unknown.
- The naive estimator is the difference between the sample means of the treatment and control group. This converges to the contrast $\mathbb{E}[Y^1 \mid W = 1] - \mathbb{E}[Y^0 \mid W = 0]$.
- When W is unconfounded, then $\mathbb{E}[Y^i \mid W = j] = \mathbb{E}[Y^i]$, so the naive estimator is an unbiased and consistent estimator for the ATE. Moreover, ATE = ATC = ATT.

Bias in the naive estimator

Using the decomposition of the ATE, the naive estimator converges to:

$$\mathbb{E}[Y^1 \mid W = 1] - \mathbb{E}[Y^0 \mid W = 0] = \mathbb{E}[\Delta] + \text{Baseline} + \text{DTE},$$

where the baseline bias is:

$$\text{Baseline} = \left(\mathbb{E}[Y^0 \mid W = 1] - \mathbb{E}[Y^0 \mid W = 0]\right),$$

and the differential treatment effect bias is:

$$\text{DTE} = (1 - \pi)\left(\mathbb{E}[\Delta \mid W = 1] - \mathbb{E}[\Delta \mid W = 0]\right).$$

Connections to Pearl's causal model

- Pearl represents an ideal experiment with the $\mathrm{do}(.)$ operator.
- Causal quantities are given by under-intervention distributions:

$$\mathbb{P}(Y \mid \mathrm{do}(W = 1)) \quad \text{and} \quad \mathbb{P}(Y \mid \mathrm{do}(W = 0)),$$

  and not the pre-intervention distributions:

$$\mathbb{P}(Y \mid W = 1) \quad \text{and} \quad \mathbb{P}(Y \mid W = 0).$$

- $Y \mid \mathrm{do}(W = j)$ is analogous to Y^j in potential outcome notation.
- For example, the ATE is $\mathbb{E}[Y^1] - \mathbb{E}[Y^0]$ in potential outcome notation, and $\mathbb{E}[Y \mid \mathrm{do}(W = 1)] - \mathbb{E}[Y \mid \mathrm{do}(W = 0)]$ in Pearl's notation.