MAS 420: Potential Outcomes Review

Adam Rohde
Department of Statistics
University of California, Los Angeles

October 17, 2022

Recommended resources

We'll review potential outcomes and how we can use them to study causal effects.

Useful resources

1. Causal Inference: What If - Hernán and Robins (2020) [link]
2. Causal Inference: the Mixtape - Cunningham (2021) [link]
3. Mostly Harmless Econometrics - Angrist and Pischke (2009) [link]

What are potential outcomes

What do we mean by the causal effect of treatment D on outcome Y for unit i?
How would Y_i have looked if D_i had been 1,
relative to how Y_i have looked if D_i had been 0.

We use potential outcomes to represent these possible versions of Y_i.

- Y_i when D_i took the value 1 is written Y_{1i}.
- Y_i when D_i took the value 0 is written Y_{0i}.
- $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$

Individual treatment effects and fundamental problem of causal inference

For unit i, the treatment effect could be written as $\tau_i = Y_{1i} - Y_{0i}$.

This means we can write $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i = Y_{0i} + \tau_i D_i$.

For unit i we only get to observe either Y_{1i} or Y_{0i}, since we cannot observe unit i when they had been treated and when they were not treated at the same time. [And measuring at different times means we're really observing two separate outcomes ($Y_{i,t}$ and $Y_{i,t'}$).]

The goal of causal inference is to find ways to "fill in" the missing potential outcomes using what we observe.

Built-in assumptions

Consistency: The statement that $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$ is really an assumption that Y_{di} is the Y_i that we would have seen if D_i was d.

No-interference: Do unit j's attributes (e.g., outcome or treatment) affect unit i's outcome? We could write $Y_{D_i=d,Y_{\mathbf{j}=y},i}$ or $Y_i(D_i = d, Y_{\mathbf{j}=y})$ to represent a potential outcome for having COVID or not, where D is getting vaccinated. Whether or not unit i has COVID is affected by unit i's vaccination status as well as by whether or not unit j has COVID. We often assume that this sort of thing is not happening: $Y_i(D_i = d, Y_{\mathbf{j}=y}) = Y_i(D_i = d)$.

One version of treatment: We also assume that what we mean by $D = d$ is the same thing in practice for all units. If we are interested in the effect of aspirin on headaches, we don't want $D = d$ to mean "take some aspirin". This could be 1 pill or 20 pills, which are substantively different. We want $D = d$ meaning "take 500mg of aspirin".

Measures of causal effects

There are many different ways to measure a causal effect. Due to the fundamental problem of causal inference, we are often interested in some aggregation of individual causal effects.

- ATE $= \mathbb{E}[\tau_i] = \mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$
- ATT $= \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 1]$, ATC $= \mathbb{E}[Y_{1i} - Y_{0i} | D_i = 0]$
- CATE $= \mathbb{E}[Y_{1i} - Y_{0i} | X_i = x]$
- Causal mean ratio: $\mathbb{E}[Y_{1i}]/\mathbb{E}[Y_{0i}]$
- Binary outcomes:
  - Causal risk difference: $p(Y_{1i} = 1) - p(Y_{0i} = 1)$
  - Causal risk ratio: $p(Y_{1i} = 1)/p(Y_{0i} = 1)$
  - Causal odds ratio: $\frac{p(Y_{1i}=1)/p(Y_{1i}=0)}{p(Y_{0i}=1)/p(Y_{0i}=0)}$

Different goals might require different measures. If you want to understand the total number of cases of a disease under different treatments you might want a risk difference but if you want to understand how much treatment increases disease risk, then you might use the risk ratio. There are many other measures of causal effects.

Identification and ignorability

How might we actually "fill in" the missing potential outcomes?
We do this by "identifying" some causal effect measure (say ATE $= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}]$) with an expression of quantities that can be estimated from data.

$$\begin{aligned}
\mathbb{E}[Y_{1i}] &= \sum_y y \times p(Y_{1i} = y) \\
&= \sum_y y \times p(Y_{1i} = y | D_i = 1) \quad \text{if } Y_{di} \perp\!\!\!\perp D_i\text{, which we call "ignorability"} \\
&= \sum_y y \times p(Y_i = y | D_i = 1) \quad \text{by consistency: } Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) D_i \\
&= \mathbb{E}[Y_i | D = 1]
\end{aligned}$$

So we "filled in" the average treated potential outcomes by using the units that we observed to have been treated **and an assumption ($Y_{di} \perp\!\!\!\perp D_i$)**.
We can also do something similar for $\mathbb{E}[Y_{0i}]$.

Ignorability and experiments

But where did the $Y_{di} \perp\!\!\!\perp D_i$ come from? Perhaps which units treated or not is random. That is, maybe we ran a randomized experiment or are studying a natural experiment.

How does random assignment of D give us $Y_{di} \perp\!\!\!\perp D_i$? If a unit's D value is assigned at random, then no other features of that unit or its environment will be systematically associated with D_i. (Though, in a small sample, chance associations between D and other variables are possible.)

Since D in a randomized experiment is no longer systematically associated with any other features of the units or environment, we say that it is "ignorable," and we can write things like $\mathbb{E}[Y_{1i}] = \mathbb{E}[Y_{1i}|D_i = 1] = \mathbb{E}[Y_i|D = 1]$.

Ignorability and experiments

Example (Test Prep and SAT Scores)

Suppose we have a set of linear relationships that govern SAT score (Y_i), whether or not someone went to a test prep course (D_i), and parental education (Z_i; say that parental education increases prob. that you do test prep and your score). Say Z_i is unobserved.

$$
\begin{aligned}
Z_i &= \eta_i && \text{where } \eta_i, \xi_i, \epsilon_i \text{ are independent noise,} \\
D_i &= \Phi(\gamma_0 + \gamma_1 Z_i + \xi_i) && \text{and } \Phi() \text{ is the CDF of the normal dist.,} \\
Y_i &= \alpha_0 + \alpha_1 Z_i + \alpha_2 D_i + \epsilon_i && \text{and } \gamma_0, \gamma_1, \alpha_0, \alpha_1, \alpha_2 \in \mathbb{R}. \\
\implies Y_{di} &= \alpha_0 + \alpha_1 Z_i + \alpha_2 d + \epsilon_i
\end{aligned}
$$

We want to understand the effect that test prep (D_i) has on SAT score (Y_i), α_2.
D_i is not independent of Y_{di} because Z_i appears in both the equation for D_i and for Y_{di}.
If we were to randomize the assignment of D, this would mean that $D_i \sim \text{Bernoulli}(p)$ and so Z_i would no longer be a cause of D_i; D_i and Y_{di} would be independent.

Difference in means

From a couple slide ago, we saw that, when we have ignorability ($Y_{di} \perp\!\!\!\perp D_i$), we could identify the ATE as $\mathbb{E}[Y_{1i} - Y_{0i}] = \mathbb{E}[Y_i|D = 1] - \mathbb{E}[Y_i|D = 0]$.
We can estimate this with $\frac{1}{N_1}\sum_{i=1:D_i=1}^{N_1} Y_i - \frac{1}{N_0}\sum_{i=1:D_i=0}^{N_0} Y_i$.

When we don't have ignorability, we can write (see Cunningham (2021))

$$\begin{aligned}
&\overbrace{\mathbb{E}[Y_i|D = 1] - \mathbb{E}[Y_i|D = 0]}^{\text{Difference in Means}} \\
&= \underbrace{\mathbb{E}[Y_{1i} - Y_{0i}]}_{\text{ATE}} + \underbrace{\mathbb{E}[Y_{0i}|D = 1] - \mathbb{E}[Y_{0i}|D = 0]}_{\text{“Selection” Bias}} + \underbrace{(1 - p(D = 1))(\text{ATT} - \text{ATC})}_{\text{Heterogeneous Treatment Effect Bias}}
\end{aligned}$$

where
$\text{ATT} = \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1]$ and
$\text{ATC} = \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0]$.
Do you see why ignorability makes DIM = ATE?

Conditional ignorability

What might we do when randomization is not possible?
We might consider a "no **unobserved** confounders" argument.
That is, we might assume conditional ignorability holds: $Y_{di} \perp\!\!\!\perp D_i | Z_i$ which means that, within strata of Z, it is like treatment is randomly assigned.

Example (Test Prep and SAT Scores)

Recall our test prep example. We want to study the effect that test prep (D_i) has on SAT score (Y_i), α_2. Z_i is parental education, which now say we observe.

$$D_i = \Phi(\gamma_0 + \gamma_1 Z_i + \xi_i)$$
$$Y_{di} = \alpha_0 + \alpha_1 Z_i + \alpha_2 d + \epsilon_i$$

D_i is not independent of Y_{di} because Z_i appears in both the equation for D_i and for Y_{di}. But if we look within strata of Z_i (i.e., compare people whose parents have the same education), we see that D_i is independent of Y_{di}, since no other causes of Y are related to D. (Note the last statement is an assumption of no unobserved confounders)

Identification and conditional ignorability

With conditional ignorability, we can also identify the distribution over the potential outcomes as

$$
\begin{aligned}
p(Y_{di}) &= \sum_z p(Y_{di} = y, Z_i = z) \\
&= \sum_z p(Y_{di} = y, Z_i = z) \frac{p(Z_i = z)}{p(Z_i = z)} \\
&= \sum_z p(Y_{di} = y | Z_i = z) p(Z_i = z) \\
&= \sum_z p(Y_{di} = y | D_i = d, Z_i = z) p(Z_i = z) \text{ by } Y_{di} \perp\!\!\!\perp D_i | Z_i \\
(*) &= \sum_z p(Y_i = y | D_i = d, Z_i = z) p(Z_i = z) \text{ by consistency}
\end{aligned}
$$

Identification and conditional ignorability

We can then identify causal effects, like the ATE.

$$\begin{aligned}
\mathbb{E}[Y_{di}] &= \sum_y y \times p(Y_{di} = y) \\
&= \sum_y y \times \left[\sum_z p(Y_i = y | D_i = d, Z_i = z) p(Z_i = z) \right] \text{ by } (*) \\
&= \sum_z \left[\sum_y y \times p(Y_i = y | D_i = d, Z_i = z) \right] p(Z_i = z) \\
&= \sum_z \mathbb{E}[Y_i | D_i = d, Z_i = z] p(Z_i = z)
\end{aligned}$$

Stratification or outcome modelling

So we saw that, with conditional ignorability, we could identify $\mathbb{E}[Y_{1i}] = \sum_z \mathbb{E}[Y_i | D_i = 1, Z_i = z] p(Z_i = z)$.

We can estimate this as

$$\frac{1}{N} \sum_{i=1}^{N} \hat{\mathbb{E}}[Y_i | D = 1, Z_i]$$

where $\hat{\mathbb{E}}[Y_i | D = 1, Z_i]$ is either the Z-strata-specific mean or a model we've fit for the outcome Y using Z as predictors, where both use our observed data for treated units.

We are predicting the value for Y for each observation under the assumption that $D = 1$ and using the observed value for Z. We could do something similar for $\mathbb{E}[Y_{0i}]$.

Inverse probability weighting (IPW)

We could also write $\mathbb{E}[Y_{di}]$ as

$$\begin{aligned}
\mathbb{E}[Y_{di}] &= \sum_y y \times p(Y_d = y) = \sum_y \sum_z y \times p(Y = y | D = d, Z = z) p(Z = z) \text{ by } (*) \\
&= \sum_y \sum_z y \times \frac{p(Y = y, D = d | Z = z)}{p(D = d | Z = z)} p(Z = z) \\
&= \sum_y \sum_z y \times \frac{p(Y = y, D = d, Z = z)}{p(D = d | Z = z)} \\
&= \sum_y \sum_z \sum_d y \times \mathbb{1}_{D=d} \times \frac{p(Y = y, D = d, Z = z)}{p(D = d | Z = z)} = \mathbb{E}\left[\frac{Y \times \mathbb{1}_{D=d}}{p(D = d | Z = z)}\right]
\end{aligned}$$

We can then model the probability of treatment $\hat{p}(D = d | Z = z)$. This is often called a "propensity score." This can be estimated with $\frac{1}{n} \sum_{i=1}^n \frac{Y_i \times \mathbb{1}_{D_i = d}}{\hat{p}(D_i = d | Z_i = z)}$.
These are simple estimators; they have short comings.

Any remaining time

questions / break