

Economics 696F, Causal Inference and Program Evaluation

Lecture Note 2: Potential Outcomes and Randomized Experiments

Reading for next week:

- Rubin 1977
- Imbens 2004 (listed as 2003 on earlier versions of syllabus)
- Abadie and Imbens (focus on basic idea - we won't go through all the details)

1 Potential Outcomes Model

Recall: for units $i = 1, \dots, n$,

$$\begin{aligned}T_i &= 0, 1 \text{ treatment} \\Y_i(0) &= \text{potential outcome under control} \\Y_i(1) &= \text{potential outcome under treatment} \\Y_i &= (1 - T_i)Y_i(0) + T_iY_i(1)\end{aligned}$$

Formal framework:

Probability space (Ω, \mathcal{B}, P) , where Ω is sample space, \mathcal{B} is a σ -algebra of subsets of Ω , and P is a probability measure. A random variable/vector (say X) is a measurable function $X(\cdot)$ from Ω to \mathbb{R}^k . Usually it's fine to take $\Omega = [0, 1]$ and \mathcal{B} the usual Borel σ -algebra, and P the uniform distribution.

Probability model for a single individual: we regard $(T_i, Y_i(0), Y_i(1))$ as a random vector defined on the probability space (Ω, \mathcal{B}, P) . Very loosely, we can think of Ω as the (infinite) population of individuals, and ω being the specific individual.

Probability model for random sample of size n : modify the probability space to be the n -fold product of (Ω, \mathcal{B}, P) . Each individual's variables $(T_i, Y_i(0), Y_i(1))$ are then functions of the coordinate projection of Ω^n .

The upshot is that $(T_i, Y_i(0), Y_i(1))$ are IID, with some joint distribution, which we will also denote P in a slight abuse of notation.

Since Y_i is a function of $T_i, Y_i(0)$, and $Y_i(1)$, then (T_i, Y_i) also has some joint distribution.

1.1 What is a Treatment?

Holland: a treatment is a potential manipulation that we can imagine.

“NO CAUSATION WITHOUT MANIPULATION.”

Discussion point: is “gender” a valid treatment?

1.2 SUTVA

Stable Unit Treatment Value Assumption (SUTVA, Rubin 1974): The value of the outcome for unit i when exposed to treatment t will be the same regardless of the treatments that other units receive.

SUTVA implies that the representation by potential outcomes is adequate.

We can formalize this by starting with a more general model, in which a person’s potential outcomes could in principle depend on everyone’s treatment:

$$Y_i(t_1, t_2, \dots, t_n).$$

Then SUTVA implies that $Y_i(t_1, \dots, t_n)$ depends only on t_i .

Note that SUTVA is quite restrictive. It rules out peer effects, “general equilibrium” effects, etc. We may want to think about relaxing SUTVA in some applications.

Another implicit assumption (sometimes considered part of SUTVA): the mechanism used to assign the treatments does not matter.

This says that the assigning the treatments in a different way, does not constitute a different treatment (again, the representation by potential outcomes is adequate).

Hawthorne effect: people might change their behavior, if they know they are being observed.

1.3 Estimands

- Average Treatment Effect:

$$ATE := E[Y_i(1) - Y_i(0)].$$

- Average Effect on the Treated:

In some cases, we are interested in some policy or treatment which we will make available to individuals, but we will not force them to take the treatment. So

we would like to know how effective the treatment will be, for those individuals who elect to take the treatment. We could define this as:

$$TT := E[Y_i(1) - Y_i(0)|T_i = 1].$$

In general, TT could be different from ATE . For example, if individuals who are likely to benefit from the treatment are the ones who end up taking it, then we could have $TT > ATE$.

We should be a little careful in using this definition. If information about the efficacy of the treatment becomes widely known, individuals may change their take-up behavior over time. So TT is arguably somewhat more sensitive to problems of “external generalizability.”

- Subgroup Average Effect:

Suppose we also observe some individual characteristic (“covariate” or “attribute”) X_i , another random variable defined on the probability space. We assume that X is not affected by the treatment.

$$ATE(x) := E[Y_i(1) - Y_i(0)|X_i = x].$$

This is the average effect for individuals with covariate value x . It could be particularly useful to a social planner who wants to make treatment assignments on the basis of individual characteristics.

2 Randomized Experiments

Naive Estimator:

$$\begin{aligned} \hat{\tau}_n &= E[\widehat{Y_i|T_i = 1}] - E[\widehat{Y_i|T_i = 0}] \\ &= \frac{\sum_{i=1}^n Y_i 1(T_i = 1)}{\sum_{i=1}^n 1(T_i = 1)} - \frac{\sum_{i=1}^n Y_i 1(T_i = 0)}{\sum_{i=1}^n 1(T_i = 0)}. \end{aligned}$$

In general, this estimator does not consistently estimate ATE:

$$\begin{aligned} \hat{\tau}_n &= E[\widehat{Y_i|T_i = 1}] - E[\widehat{Y_i|T_i = 0}] \\ &\xrightarrow{p} E[Y_i|T_i = 1] - E[Y_i|T_i = 0] \\ &= E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] \\ &\neq E[Y_i(1)] - E[Y_i(0)]. \end{aligned}$$

If T_i is “randomly assigned,” then it should be independent of $Y_i(0), Y_i(1)$. Let us assume that:

$$T_i \perp (Y_i(0), Y_i(1)).$$

Here, “ \perp ” means “independent of.”

Then

$$\begin{aligned} E[Y_i(1)|T_i = 1] &= E[Y_i(1)] \\ E[Y_i(0)|T_i = 0] &= E[Y_i(0)], \end{aligned}$$

so

$$\hat{\tau}_n \xrightarrow{P} E[Y_i(1)] - E[Y_i(0)] = ATE.$$

So randomized experiments permit consistent estimation of the ATE, without strong distributional or functional form assumptions. We say the ATE is *nonparametrically identified*.

An alternative definition of identification (roughly equivalent): ATE (or other object of interest) is *identified* if we can recover it from the distribution of observables — this is, from the distribution of (T_i, Y_i) .

2.1 Identification vs. Estimation

Having shown that *ATE* is identified, we can turn to estimation. A natural choice is, of course, the estimator $\hat{\tau}_n$. Under standard regularity conditions, we can construct asymptotically valid standard errors and confidence intervals, and test hypotheses about *ATE*. (More on testing later in the course.)

However, we could also consider alternative estimators. We might want to make some further parametric assumptions about the distribution of $Y_i(1)$ and the distribution of $Y_i(0)$, and take advantage of these assumptions in estimating the treatment effect.

For example, suppose we think that the distributions of $Y_i(1)$ and $Y_i(0)$ are both lognormal:

$$\begin{aligned} Y_i(0) &\sim LN(\mu_0, \sigma_0^2); \\ Y_i(1) &\sim LN(\mu_1, \sigma_1^2). \end{aligned}$$

(Note: we are only specifying the *marginal* distributions of $Y_i(0)$ and $Y_i(1)$, not their joint distribution. We could try to specify the joint distribution, but is there information in the data about the correlation between $Y_i(0)$ and $Y_i(1)$??)

So (under randomization of treatment),

$$\begin{aligned} E[Y_i|T_i = 0] &= E[Y_i(0)] = \exp\left(\mu_0 + \frac{1}{2}\sigma_0^2\right); \\ E[Y_i|T_i = 1] &= E[Y_i(1)] = \exp\left(\mu_1 + \frac{1}{2}\sigma_1^2\right). \end{aligned}$$

We can estimate $\mu_0, \sigma_0, \mu_1, \sigma_1$ by MLE, and then plug in to get the MLE estimate of ATE .

Discussion point: If we are going to make parametric assumptions (like lognormality), then why not use them when verifying identification? In other words, why not simply look at *parametric* identification rather than nonparametric identification?

3 Examples

3.1 Example 1: Lalonde (1986)

National Supported Work Demonstration (NSW): randomized evaluation of a job training program.

Difference in means between treated and controls: females \$851, males \$886.

Also used regression methods to control for observed background characteristics (this can improve precision of estimates) - similar results

“Nonexperimental estimates”: don’t use experimental controls. Instead, construct a control group from various national surveys (PSID, CPS).

Use regression methods, selection correction methods to try to control for differences and obtain estimates of treatment effects.

Nonexperimental methods did poorly - in many cases not close to the experimental results.

3.2 Example 2: Vitamin C

(see Rosenbaum, Ch.1)

Cameron and Pauling (1976): gave vitamin C to 100 patients believed to be terminally ill from cancer.

Comparison group constructed by matched sampling: for each treated patient, select 10 patients from historical records with same type of cancer and other characteristics (age, gender)

Patients receiving vitamin C lived about 4 times longer than controls, highly significant.

Later, careful randomized experiment conducted at the Mayo Clinic. Patients randomly assigned to receive vitamin C or a placebo.

No evidence that vitamin C prolonged survival.

Discuss: what might be explanation? Can we express this in terms of potential outcomes?

4 Population vs. Superpopulation Treatment Effects

We have been assuming an infinite population, and defining effects like ATE in terms of the distribution of potential outcomes in the infinite population.

Some work in causal inference focuses instead on the effect *within the finite population* $i = 1, \dots, n$.

$$ATE_n := \frac{1}{n} \sum_{i=1}^n Y_i(1) - Y_i(0).$$

(similar expressions for TT , etc.)

This is the treatment effect for the n individuals we actually have. To make a distinction, some authors would call ATE_n the “population” treatment effect and would call ATE the “superpopulation” treatment effect.

For clarity, we will call ATE_n the “finite-population” average treatment effect.

Discuss: when (if ever) is ATE_n the more interesting estimand?

Asymptotic arguments involving finite-population effects becomes a bit tricky. What does it mean to consistently estimate ATE_n ? We have a moving target. One way to operationalize might be to look for:

$$(\hat{\tau}_n - ATE_n) \xrightarrow{p} 0.$$

Asymptotic distribution theory is tricky:

$$\begin{aligned} \sqrt{n}(\hat{\tau}_n - ATE_n) &= \sqrt{n}(\hat{\tau}_n - ATE + ATE - ATE_n) \\ &= \sqrt{n}(\hat{\tau}_n - ATE) + \sqrt{n}(ATE - ATE_n) \end{aligned}$$

The second term is also converging in distribution to something!

Or we could instead look for estimators that are unbiased (for fixed sample size n).

5 Sharp Null Hypothesis of No Treatment Effect

Continue with the finite-population perspective for now.

In the finite-population view, the $Y_i(0)$ and $Y_i(1)$ are fixed quantities. Only the vector $\mathbf{T} = (T_1, \dots, T_n)$ is random, leading to randomness in the outcome vector $\mathbf{Y} = (Y_1, \dots, Y_n)$.

Randomization Mechanism some possibilities for the distribution of \mathbf{T} :

1. T_i are independent Bernoulli, $Pr(T_i = 1) = 1/2$.
Thus, there are 2^n possible configurations for \mathbf{T} , each equally likely.
Small probability of all treated, or all untreated.
2. Equal number of treated and untreated.
Suppose n even. There are $\binom{n}{n/2}$ ways to assign $n/2$ units to treatment and $n/2$ units to control, each with probability $1/\binom{n}{n/2}$.
Note that under this mechanism, T_i is not independent of T_j , $j \neq i$.
3. Could subdivide sample into strata based on observed characteristics, and apply (2) within each strata.

Some version of mechanism (2) is the most commonly used. Sometimes people ignore the distinction between (1) and (2), even though they are formally distinct. For large n , it might be possible to show that the two mechanisms are “close.”

Now, suppose we are interested in the following null hypothesis: the treatment has no effect for any individual:

$$H_0 : Y_i(1) = Y_i(0) \quad \forall i.$$

Under this *sharp null hypothesis*, there is no missing data: for each i ,

$$Y_i(0) = Y_i(1) = Y_i.$$

Test statistic: some real-valued function $s(\mathbf{T}, \mathbf{Y})$. Example: a t-statistic.

Want to calculate

$$Pr [s(\mathbf{T}, \mathbf{Y}) \geq c | H_0].$$

Here c is the critical value of the test.

We know distribution of \mathbf{T} , and under H_0 we know **all** the $Y_i(0), Y_i(1)$.

Let Ω denote the space of possible configurations for the vector \mathbf{T} . Use \star to denote element-by-element multiplication of vectors, and let ι_n denote an n -vector of ones. Then

$$Pr [s(\mathbf{T}, \mathbf{Y}) \geq c | H_0] = \sum_{t \in \Omega} 1(s(t, \mathbf{Y}(t)) \geq c) Pr(\mathbf{T} = t),$$

where

$$\mathbf{Y}(t) := (\iota_n - t) \star \mathbf{Y}(\mathbf{0}) + t \star \mathbf{Y}(\mathbf{1}).$$

Hence, we can get exact p-values by simulation:

1. Draw $\tilde{\mathbf{T}}$ using our knowledge of the randomization mechanism.
2. Form $\tilde{\mathbf{Y}}$ based on draw for $\tilde{\mathbf{T}}$.
3. Form $s(\tilde{\mathbf{T}}, \tilde{\mathbf{Y}})$.
4. Repeat

In some cases, exact p-values can be obtained analytically. For example, if outcome is binary, exact distribution of t-stat is hypergeometric.