Lecture Notes
MS&E: Causal Inferences (Autumn 2020)
@ Stanford University

Tianyu Du

November 5, 2020

1 Potential Outcome Framework

1.1 Rubin Causal Model / Neyman-Rubin Potential Causal Framework

Define assignments and potential outcomes

- Let $i = 1, 2, \dots, N$ be indices of N units (subjects).
- For simplicity, assume binary intervention Z_i with value $z_i \in \{0, 1\}$ or $\{\text{control, treatment}\}$.
- Let vector of random variables $\mathbf{Z} = (Z_1, Z_2, \dots, Z_N)$ be the population assignment. A realization of $\mathbf{Z}$, $\mathbf{z} = (z_1, z_2, \dots, z_N)$, denotes actual treatment assignments to the population.
- $Y_i(\mathbf{z})$ denotes the potential outcome of unit i when the entire population receives treatment $\mathbf{z}$.
- A $\mathbf{z}$ defines an universe / realization of $\mathbf{Z}$, therefore, there are 2^N potential outcomes for each unit i.

Assumption 1.1. The Stable Unit Treatment Value Assumption (SUTVA).

1. No interference between units: outcome of unit i is not affected by treatment of player j for all $j \neq i$.

$$\mathbf{z}_i = \mathbf{z}'_i \implies Y_i(\mathbf{z}) = Y_i(\mathbf{z}') \tag{1}$$

   With the first assumption holds, we can write the potential outcome of unit i as a function of z_i only: $Y_i(z_i)$.

2. No hidden version of treatments: $Y_i(z_i)$ is a well-defined function.

$$z_i = z'_i \implies Y_i(z_i) = Y(z'_i) \tag{2}$$

Given SUTVA, there are only 2 potential outcomes for each unit i.

The Science Denote the vector of potential outcomes $\mathbf{Y}(0) := (Y_1(0), Y_2(0), \dots, Y_N(0))$ and $\mathbf{Y}(1) := (Y_1(1), Y_2(1), \dots, Y_N(1))$. The science (aka. schedule of potential outcomes) is defined as

$$\underline{\mathbf{Y}} := (\mathbf{Y}(0), \mathbf{Y}(1)) \tag{3}$$

$\underline{\mathbf{Y}}$ tells outcomes in all factual and counterfactual scenarios.

Definition 1.1. An **assignment mechanism** is a distribution of $\mathbf{Z}$. A generic mechanism is often denoted as $\mathbf{Z} \sim \eta$. An assignment mechanism is a **randomized experiment** if

1. *Probabilistic*: $0 < P(z_i | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) < 1$.
2. *Known assignment mechanism*: the assignment can be expressed explicitly.
3. *Individualistic*: $P(z_i | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(z_i | X_i, Y_i(0), Y_i(1))$.
4. *Unconfoundness* $P(\mathbf{z} | \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)) = P(\mathbf{z} | \mathbf{X})$, where $\mathbf{X}$ is known and observed covariates.

Remark 1.1. The unconfoundness assumption cannot be tested empirically since we never have full knowledge on the science.

1.2 Neymanian Inference

Average Treatment Effect and Difference in Mean Estimator

- Define the **average treatment effect** (ATE) to be $\tau = \overline{\mathbf{Y}(1)} - \overline{\mathbf{Y}(0)}$.
- Let $\mathbf{z}^{obs}$ and $\mathbf{y}^{abs}$ denote the observed treatment assignment and outcome.
- We may construct the **difference in mean** estimator for ATE: $\hat{\tau}^{DIM} = \overline{\mathbf{y}^{obs}(1)} - \overline{\mathbf{y}^{obs}(0)}$.
- Given the $\underline{\mathbf{Y}}$, each realization of $\mathbf{z}$ leads to one value of $\hat{\tau}$. The $\mathbf{z} \sim \eta$ induces a distribution on $\hat{\tau} \sim P_\eta$.

Definition 1.2. Given an assignment mechanism η, the **bias** of an estimator is

$$Bias_\eta(\hat{\tau}, \tau, \underline{\mathbf{Y}}) = \mathbb{E}_{\mathbf{z} \sim \eta}(\hat{\tau}(\mathbf{z}, \underline{\mathbf{Y}})) - \tau(\underline{\mathbf{Y}}) \tag{4}$$

An estimator is **unbiased** for τ under design η if for all $\underline{\mathbf{Y}}$, $Bias_\eta(\hat{\tau}, \tau, \underline{\mathbf{Y}}) = 0$.

We can rewrite the observed outcome as

$$y_i = z_i y_i(1) + (1 - z_i) y_i(0) \tag{5}$$

1.3 Causal Estimands

Clarification

- Estimand: the quantity of interest to be estimated.

- Estimator: a procedure to approximate estimand from data.

Example 1.1. Examples of causal estimands include

- Individual treatment effect: $\tau_i := y_i(1) - y_i(0)$.
- Average treatment effect: $\tau^{ATE} := \overline{\mathbf{y}_i(1)} - \overline{\mathbf{y}_i(0)}$.
- Conditional average treatment effect: let X_i be the controlled co-variate, $\tau_x := \frac{1}{N_x} \sum_{i=1}^{N} \mathbb{1}\{X_i = x\}\tau_i$.
- Lift: $L := \frac{\overline{\mathbf{y}_i(1)} - \overline{\mathbf{y}_i(0)}}{\overline{\mathbf{y}_i(0)}}$.

Super-population Estimands So far, we have fixed and finite population without a model. τ is specific to the population of size N. We may assume the N units are actually i.i.d. samples from a super-population:

$$(Y_i(0), Y_i(1)) \overset{i.i.d.}{\sim} P \tag{6}$$

with

$$\mathbb{E}_P(Y_i(0)) = \mu_0 \tag{7}$$
$$\mathbb{E}_P(Y_i(1)) = \mu_1 \tag{8}$$

One super-population estimand is $\theta = \mathbb{E}(\tau) = \mu_1 - \mu_0$, we may construct models to estimate parameters μ_0 and μ_1.

1.4 No Causation without Manipulation

2 Randomized Experiments: Neyman v.s. Fisher Inferences

2.1 The randomization based framework

Throughout this section, we assume SUTVA. However, the randomization removes the need for most assumptions beyond SUTVA.

Reasoned-Basis for Inference

- Recall: observed $y_i = z_i y_i(1) + (1 - z_i) y_i(0)$.
- Classical statistics:
  1. Assume $y_i | (z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2),\ y_i | (z_i = 0) \sim \mathcal{N}(\mu_0, \sigma_0^2)$.
  2. Compute MLE $\hat{\mu}_1^{MLE}$ and $\hat{\mu}_0^{MLE}$.
  3. Estimate $\hat{\tau}^{MLE} = \hat{\mu}_1^{MLE} - \hat{\mu}_0^{MLE}$.
- Randomization-based inference:

1. Consider $\underline{\mathbf{Y}}$as fixed, but unobserved.
2. $\mathbf{z}$is a random variable.
3. $\mathbf{y}(\mathbf{z})$ is the observed realization from a function of $\mathbf{z}$. (recall: $\mathbf{z}$defines $\mathbf{y}$through η.)
4. No other assumptions.
5. All randomness came from $\mathbf{z}$, the distribution of $\mathbf{z} \sim P_\eta(\mathbf{z})$ will play a crucial role.

Assignment Mechanism

Example 2.1 (Bernoulli Assignment)**.** Let $p \in (0,1)$, $P(z_i = 1) = p$, $P(\mathbf{z}) = \prod_{i=1}^N p^{z_i}(1-p)^{1-z_i}$.

Example 2.2 (Completely Randomized Design (CRD))**.** CRD(N_1, N) randomly draws N_1 units from N units and assigns them treatment.

$$P(z_i = 1) = \frac{N_1}{N} \tag{9}$$

$$P(\mathbf{z}) = \begin{cases} \binom{N}{N_1}^{-1} & \text{if } \sum z_i = N_1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Proposition 2.1. Let $\tau = \tau^{ATE}$ and $\hat{\tau} = \hat{\tau}^{DIM}$, suppose $\eta = CRD(N_1, N)$ with $0 < N_1 < N$, then the difference-in-mean estimator is unbiased. That is, for all $\underline{\mathbf{Y}}$, $Bias_\eta(\hat{\tau}, \tau; \underline{\mathbf{Y}}) = 0$.

Proof.

$$\hat{\tau}^{DIM} \equiv \frac{1}{\sum z_i}\sum_{i=1}^N z_i y_i - \frac{1}{\sum(1-z_i)}\sum_{i=1}^N (1-z_i)y_i \tag{11}$$

$$= \frac{1}{N_1}\sum_{i=1}^N z_i y_i - \frac{1}{N-N_1}\sum_{i=1}^N (1-z_i)y_i \text{ by CRD} \tag{12}$$

$$\implies \mathbb{E}_\eta[\hat{\tau}^{DIM}] = \frac{1}{N_1}\sum_{i=1}^N \mathbb{E}_\eta[z_i y_i] - \frac{1}{N-N_1}\sum_{i=1}^N \mathbb{E}_\eta[(1-z_i)y_i] \tag{13}$$

$$= \frac{1}{N_1}\sum_{i=1}^N \mathbb{E}_\eta[z_i y_i(1)] - \frac{1}{N-N_1}\sum_{i=1}^N \mathbb{E}_\eta[(1-z_i)y_i(0)] \tag{14}$$

$$= \frac{1}{N_1}\sum_{i=1}^N y_i(1)\mathbb{E}_\eta[z_i] - \frac{1}{N-N_1}\sum_{i=1}^N y_i(0)\mathbb{E}_\eta[(1-z_i)] \tag{15}$$

$$= \frac{1}{N_1}\sum_{i=1}^N y_i(1)\frac{N_1}{N} - \frac{1}{N-N_1}\sum_{i=1}^N y_i(0)\frac{N-N_1}{N} \tag{16}$$

$$= \frac{1}{N}\sum_{i=1}^N y_i(1) - y_i(0) \tag{17}$$

■

Definition 2.1. The variance of estimator $\hat{\tau}$ is defined as

$$Var_\eta(\hat{\tau}) := \mathbb{E}_\eta[(\hat{\tau} - \mathbb{E}_\eta[\hat{\tau}])^2] \tag{18}$$

Proposition 2.2. Let $\tau = \tau^{ATE}$ and $\hat{\tau} = \hat{\tau}^{DIM}$, suppose $\eta = CRD(N_1, N)$ with $0 < N_1 < N$, define $N_0 = N - N_1$. Then,

$$Var_\eta(\hat{\tau}) = \frac{V_1}{N_1} + \frac{V_0}{N_0} - \frac{V_{1,0}}{N} \tag{19}$$

where V_a is the variance of potential outcome a and $V_{1,0}$ is the variance of treatment effect.

$$V_a = \frac{1}{N-1}\sum_i (y_i(a) - \bar{y}(a))^2 \;\; a \in \{0, 1\} \tag{20}$$

$$V_{1,0} = \frac{1}{N-1}\sum_i (\tau_i - \tau)^2 \tag{21}$$

2.2 Inferences: Neymanian

- Suppose we have a normal approximation is large samples:

  $$\frac{\hat{\tau} - \tau}{\sigma(\hat{\tau})} \sim \mathcal{N}(0, 1) \tag{22}$$

  Then,

  $$CI_{1-\alpha} = [\hat{\tau} - q_{1-\alpha/2}\sigma(\hat{\tau}), \hat{\tau} + q_{1-\alpha/2}\sigma(\hat{\tau})] \tag{23}$$

  The confidence interval satisfies

  $$P_\eta(\tau \in CI_{1-\alpha}) \approx 1 - \alpha \tag{24}$$

- However, v and $\sigma \equiv \sqrt{v}$ depends on unknown quantities (true variances of potential outcomes and treatment effects). We need to estimate $\hat{v}$ using data.

- In general, we wish construct a conservative estimation $\hat{v}$ such that $\mathbb{E}[\hat{v}] \geq v$, which leads to a larger confidence interval satisfying $P_\eta(\tau \in \hat{CI}_{1-\alpha}) \geq 1 - \alpha$.

- Specifically, we can use conventional sample estimations for V_1 and V_0 while ignoring the variance of treatment effects. Doing so leads to a conservative estimation of variance.

Proposition 2.3. The Neyman estimator of variance is

$$\hat{v} = \frac{\hat{V}_1}{N_1} + \frac{\hat{V}_0}{N_0} \tag{25}$$

where

$$\hat{v}_1 = \frac{1}{N_1 - 1} \sum_{i=1}^{N} z_i \left(y_i^{obs} - \overline{y}^{obs} \right)^2 \tag{26}$$

$$\hat{v}_0 = \frac{1}{N_0 - 1} \sum_{i=1}^{N} (1 - z_i) \left(y_i^{obs} - \overline{y}^{obs} \right)^2 \tag{27}$$

under $CRD(N_1, N_0)$,

$$\mathbb{E}_\eta[\hat{v}] \geq Var_\eta(\hat{\tau}) \tag{28}$$

2.3 Hypothesis Testing: As a Stochastic Proof by Contradiction

- $H_0 : \overline{Y}(1) = \overline{Y}(0)$ (i.e., $\tau^{ATE} = 0$).
- Define $T^{obs} = \frac{\hat{\tau} - 0}{\sqrt{\hat{v}}} = \frac{\hat{\tau}}{\sqrt{\hat{v}}}$.
- Define the p-value as $1 - \Phi(T^{obs})$ (one-sided) or $2(1 - \Phi(T^{obs}))$ (two-sided).
- Then,

$$P_\eta(p \leq \alpha | H_0) \leq \alpha \tag{29}$$

- That is, we firstly suppose H_0 to be true, in this case, T^{obs} should follow $\mathcal{N}(0, 1)$ (the null distribution). If we observe some T^{obs} that is unlikely under the null distribution, that is, T^{obs} contradicts the null distribution, we reject H_0.

2.4 Horvitz-Thompson Estimator for τ^{ATE}

Definition 2.2. Let η be any design that is a randomized experiment, let $\Pi_i = P_\eta(Z_i = 1)$ and $0 < \Pi_i < 1$. Define

$$\hat{\tau}^{HT} = \frac{1}{N} \sum_{i=1}^{N} \frac{z_i}{\Pi_i} y_i + \frac{1}{N} \sum_{i=1}^{N} \frac{1 - z_i}{1 - \Pi_i} y_i \tag{30}$$

Note that $\hat{\tau}^{HT}$ is a special case of inverse propensity-score weighting (IPW) estimators.

Proposition 2.4. Let η be any design that is a randomized experiment, then the HT estimator $\hat{\tau}^{HT}$ is unbiased for τ^{ATE}.

Proof.

$$\begin{aligned}
\mathbb{E}_\eta \hat{\tau}^{HT} &= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[\frac{z_i}{\Pi_i}y_i] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[\frac{1-z_i}{1-\Pi_i}y_i] && (31)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[\frac{z_i}{\Pi_i}y_i(1)] + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[\frac{1-z_i}{1-\Pi_i}y_i(0)] && (32)\\
&= \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[z_i]\frac{1}{\Pi_i}y_i(1) + \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\eta[1-z_i]\frac{1}{1-\Pi_i}y_i(0) && (33)\\
&= \frac{1}{N}\sum_{i=1}^{N}y_i(1) + \frac{1}{N}\sum_{i=1}^{N}y_i(0) && (34)\\
&= \tau^{ATE} && (35)
\end{aligned}$$

■

2.5 Fisher Randomization Test

- Neymanian makes very few assumptions,
- but requires asymptotic arguments: $T^{obs} \sim \mathcal{N}(0,1)$,
- for CRD we may derive the asymptotic distribution easily, but this is much harder for other designs.
- FRT assumes nothing beyond SUTVA.

Nayman's H_0 v.s. Fisher's H_0 H_0^{Fisher} is stronger than H_0^{Neyman}:

$$\begin{aligned}
H_0^{Neyman} &: \tau^{ATE} = 0 && (36)\\
H_0^{Fisher} &: Y_i(0) = Y_i(1)\ \forall i \in [N] && (37)
\end{aligned}$$

FRT Workflow

- Set $\underline{\mathbf{Y}}$ fixed but unknown.
- Observe $\mathbf{z}^{obs} \sim \eta$ and $\mathbf{y}^{obs} = y(\mathbf{z}^{obs})$.
- Compute $T(\mathbf{z}^{obs}, \mathbf{y}^{obs})$ such as $\frac{1}{N_1}\sum z_i y_i + \frac{1}{N_0}\sum(1-z_i)y_i$.
- Suppose the null hypothesis H_0 is true, such as $Y_i(0) = Y_i(1)$.
- Deduce $\underline{\mathbf{Y}}^*$ based on H_0 and $\mathbf{y}^{obs}$.
- Given the deduced $\underline{\mathbf{Y}}^*$ how likely is it that we observe T^{obs}?
- Iterate over all possible $\mathbf{z}'$, compute $\mathbf{y}' = \underline{\mathbf{Y}}^*(\mathbf{z}')$ and $T(\mathbf{z}', \mathbf{y}')$.

- The distribution of computed $T(\mathbf{z}', \mathbf{y}')$ is called the null distribution.

- The p-value is $P_\eta(T(\mathbf{z}, \underline{\mathbf{Y}}^*(\mathbf{z})) \geq T^{obs}|H_0)$ and measures how likely T^{obs} occurs under H_0.

```
Algorithm 1: Fisher's Randomization Test
Inputs: $\mathbf{z}^{obs}, \mathbf{y}^{obs}, T(\cdot), \eta$;
Returns: estimated $p$-value.;
$T^{obs} \leftarrow T(\mathbf{z}^{obs}, \mathbf{y}^{obs})$;
Deduce $\underline{\mathbf{Y}}^*$ from $H_0$;
for $k = 1, 2, \ldots, K$ do
    Sample $\mathbf{z}^{(k)} \sim \eta$;
    $\mathbf{y}^{(k)} \leftarrow \underline{\mathbf{Y}}^*(\mathbf{z}^{(k)}$;
    $T^{(k)} = T(\mathbf{z}^{(k)}, \mathbf{y}^{(k)})$;
Compute Monte-Carlo approximation of $p$-value:
    $\widehat{pval} = \frac{1}{K} \sum \mathbb{1}\{T^{(k)} \geq T^{obs}\}$
```

Theorem 2.1. Under H_0, as $K \to \infty$,

$$P_\eta(\widehat{pval} \leq \alpha|H_0) \leq \alpha \tag{38}$$

Given confidence level α, we reject H_0 if and only if $\hat{p} \leq \alpha$. The theorem says the chance of falsely rejecting H_0 (type I error) is less than α. This theorem suggests a rejection criterion based on the output of FRT is justifiable.

2.6 Power and Choice of Test Statistics

A good test much control type I error and have high power to detect certain volition of the H_0. The power of a test is H_1-specific. Given an alternative hypothesis

$$Power(H_1) = P_\eta(pval \leq \alpha|H_1) \tag{39}$$

If H_1 is true, we wish to reject H_0 as often as possible by choosing a larger α, which leads to increase chance of type I error.