The Potential Outcomes Framework

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- We discussed basic concepts in statistical analysis.
- The estimand is the theoretical quantity we want to estimate.
- The estimator is a mapping from data to a number (the estimate).
- We hope that the estimator is well behaved.
- Desirable properties include unbiasedness, consistency, efficiency, and asymptotic normality.
- We also want to quantify the uncertainty around our estimate.
- This process is known as statistical inference.

Review

- Typically, we first derive the variance of the estimator.
- Then, we use its sample analogue to estimate the variance.
- It is acceptable if the variance estimate is conservative.
- If the estimator converges to a normal distribution with the root-N rate, we can construct confidence intervals using normal critical values.

Why do we care about causality?

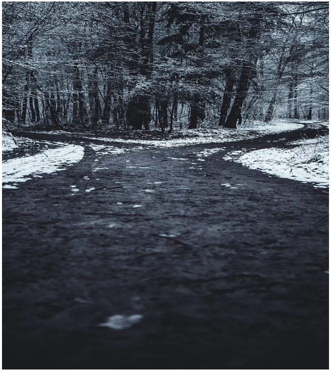
- This lecture introduces basic concepts in causal inference.
- Most theories we are interested in take the form of causal relationships.
- What would happen to Y if D changes?
  - Does economic growth cause democratization?
  - Do political ads change viewers' political preference?
  - Can trade reduce the probability of war?
- We call Y the outcome and D the treatment.
- The better we understand causal relationships, the better we can design policy interventions.

How do we define causality?

- There has been a long history of defining causality.
- Aristotle (four causes), Hume (does it exist?), Mill (the method of agreement/difference)...
- We follow the current practice and define causality using counterfactual.
- Ideally, we travel back to the past with a time machine and alter the value of D.
- We then observe what would happen to Y.

How do we define causality?

- "Two roads diverged in a wood, and I—
  I took the one less traveled by,
  And that has made all the difference."
  — Robert Frost, *The Road Not Taken*



- This simple idea is captured by the Neyman-Rubin model.

The Neyman-Rubin model

- We possess a sample of N units.
- Denote the outcome of interest for unit i as Y_i and the treatment as $D_i \in \{0, 1\}$.
- We may also have some pre-treatment covariates $\mathbf{X}_i$.
- Then, we have

$$Y_i = \begin{cases} Y_i(0), \ D_i = 0 \\ Y_i(1), \ D_i = 1. \end{cases}$$

- $Y_i(d)$ is called the "potential outcome."
- $\tau_i = Y_i(1) - Y_i(0)$ is the individualistic treatment effect.

The Neyman-Rubin model

- We call the average of τ_i, $\tau = \frac{1}{N}\sum_{i=1}^{N}\tau_i$, the average treatment effect (ATE).
- When N is sufficiently large, we can also write the ATE as $E[\tau_i]$.
- Obviously,

$$\tau = E[\tau_i] = E[Y_i(1)] - E[Y_i(0)].$$

- Note that $Y_i(d)$ could be a complex function of both observable and unobservable factors:

$$Y_i(d) = f_d(\mathbf{X}_i, U_i),$$

  where U_i represents unobservable factors.
- τ is a quantity that marginalizes over all these factors.

The Neyman-Rubin model

- The model includes several implicit assumptions:
  1. Consistency,
  2. Manipulable treatment,
  3. Stable Unit Treatment Value Assumption (SUTVA).

The Neyman-Rubin model

- Consistency is a philosophical concern on the interpretation of potential outcomes.
- Manipulable treatment restricts the scope of problems we could study.
- It forces us to focus on the "effects of causes" rather than "causes of effects."
- SUTVA can be relaxed in many cases.

The Neyman-Rubin model

- The idea of potential outcomes was first established by Neyman when analyzing agricultural experiments (Neyman 1923).
- It was formalized by Rubin in his 1974 paper ("the science").
- One motivation of the model is Heisenberg's uncertainty principle.
- It was independently developed in other disciplines (Roy model, Pearl's DAG, etc.).

The fundamental problem of causal inference

- We observe either $Y_i(0)$ or $Y_i(1)$ in practice, but never both.
- "The fundamental problem of causal inference" (Holland 1986)
- The unobserved potential outcome is called the counterfactual.
- Causal inference aims to impute the counterfactual based on assumptions.
- "Causal inference is a missing data problem." —Donald Rubin
- Qualitative studies can be similarly understood (Coppock and Kaur 2022)

The fundamental problem of causal inference

Unit	$Y_i(1)$	$Y_i(0)$	D_i
1	3	2	
2	5	3	
3	4	5	

- The ATE equals to $(1 + 2 - 1)/3 = 2/3$.

The fundamental problem of causal inference

Unit	$Y_i(1)$	$Y_i(0)$	D_i
1	3	NA	1
2	NA	3	0
3	4	NA	1

The scientific solution

- How do we test Newton's second law of motion?
- "When a constant force acts on a massive body, it causes it to accelerate."
- We need two assumptions: temporal stability and unit homogeneity.
- Neither is credible in social science.

The statistical solution

- The statistical solution relies on a large sample.
- We divide the sample into the treatment group and the control group.
- Idea: John Mill's method of difference.
- If the two groups are the same in all the aspects, their difference in the average outcome could be attributed to the treatment.
- Yet this does not work in practice.

Treatment assignment

- Suppose there are 20 binary covariates that affect the outcome variable.
- To apply the method of difference, we need a sample of $2^{20} \approx 1$ million units.
- Instead, we rely on randomization of treatment assignment.
- Suppose there exists a probability $0 < p_i < 1$ for each unit i, such that
$$P(D_i = 1) = p_i.$$
- This is an individualistic and probabilistic assignment mechanism (Imbens and Rubin 2015).

Treatment assignment

- In theory, we can assign the treatment vector $\mathbf{d} = (d_1, d_2, \ldots, d_N)$ altogether.

Assignment	Probability
$(1, 0, 1)$	0.4
$(0, 0, 1)$	0.6

- There are at most 2^N possibilities.
- We can assign each possibility a probability such that these probabilities sum up to one.
- These probabilities are known as an assignment mechanism.
- But we usually assume that treatment assignment is decided by one's own attributes (individualistic) and the probability is strictly between 0 and 1 (probabilistic).
- Individualistic assignment does not mean that the probabilities are independent across units.

Treatment assignment

- In this case, p_i may still be a function of all the variables in sample:
$$p_i = g(\mathbf{X}_i, U_i, Y_i(1), Y_i(0)).$$
- This assignment mechanism is unconfounded if $p_i = p(\mathbf{X}_i)$.
- If the assignment mechanism is individualistic, probabilistic, and unconfounded, we have a **classical randomized experiment**.
- From now on, we further assume that p_i does not depend on $\mathbf{X}_i$.
- There are two common assignment mechanisms in practice.
- Bernoulli trial: $p_i = p$ for any i.
- Complete randomization:

$$P(\mathbf{d}) = \begin{cases} \frac{1}{\binom{N}{N_1}}, & \text{if } \sum_{i=1}^{N} d_i = N_1 \\ 0, & \text{otherwise.} \end{cases}$$

Treatment assignment

- Let's define $N_1 = \sum_{i=1}^{N} D_i$ and $N_0 = \sum_{i=1}^{N}(1 - D_i)$.
- Obviously, $N = N_1 + N_0$.
- They are random variables under Bernoulli trial.
- If $p = 0.5$, N_1 can be either 60 or 40 in one assignment.
- Under complete randomization, N_1 and N_0 are pre-fixed numbers.
- Complete randomization gives you the group size you want.
- It is like a lottery.

Treatment assignment

- But complete randomization is not possible in certain contexts.
- E.g., decide whether a patient is treated or not upon their arrival.
- It is easier to analyze the Bernoulli trial as probabilities are independent to each other.
- When $N \to \infty$, the difference between the two mechanisms disappears.
- Therefore, we use the Bernoulli trial as the benchmark.

Bernoulli trial vs. complete randomization

```
## Under Bernoulli trial, we have 1009 treated units, and
##  991 untreated units.
```

```
## Under complete randomization, we have 1000
##  treated units, and  1000 untreated units.
```

The statistical solution (continued)

- Treatment assignment is randomized in a classical randomized experiment, hence

$$D_i \perp \{Y_i(0), Y_i(1)\}_{i=1}^{N},$$
$$1 - \varepsilon < P(D_i = 1) < \varepsilon,$$

  and

$$E[Y_i|D_i = 1] = E[Y_i(1)|D_i = 1] = E[Y_i(1)],$$
$$E[Y_i|D_i = 0] = E[Y_i(0)|D_i = 0] = E[Y_i(0)].$$

- Hence,

$$E[Y_i|D_i = 1] - E[Y_i|D_i = 0] = E[Y_i(1)] - E[Y_i(0)] = E[\tau_i].$$

- Remember that $E[\tau_i] = \tau$ is the ATE.

The statistical solution (continued)

- The power of randomization was first recognized by Ronald Fisher.
- Randomization creates an exogenous variation so that causal identification becomes possible.
- Due to randomization, all the other factors that affect Y_i are balanced in expectation: $E[\mathbf{X}_i|D_i = 1] = E[\mathbf{X}_i|D_i = 0]$.
- This is no guarantee that $\frac{1}{N_1}\sum_{i:D_i=1} X_i = \frac{1}{N_0}\sum_{i:D_i=0} X_i$.
- As N grows, the probability for $\sum_{i:D_i=1} X_i$ to be significantly different from $\sum_{i:D_i=0} X_i$ will get smaller.
- People typically test the null hypothesis that $\frac{1}{N_1}\sum_{i:D_i=1} X_i = \frac{1}{N_0}\sum_{i:D_i=0} X_i$ for each X_i.
- Rejection of the null implies the failure of randomization.

The statistical solution (continued)

- If D_i is not randomly assigned, there will be $\mathbf{X}_i$ affecting both Y_i and D_i.
- The causal relationship between Y_i and D_i will be confounded by $\mathbf{X}_i$, hence we call them confounders.
- Causal inference studies how to utilize existing randomization from either experiments or hypothetical experiments to identify causal relationships.
- It is about inference rather than creating causality from nowhere.

Estimand vs. estimator

- No individualistic treatment effect is identifiable under statistical solutions.
- We focus on the average effect over a fixed population.
- These average effects are our estimands.
- It could be the ATE, the ATT ($\tau_{ATT} = E[Y_i(1) - Y_i(0)|D_i = 1]$), or the CATE ($\tau(\mathbf{x}) = E[Y_i(1) - Y_i(0)|\mathbf{X}_i = \mathbf{x}]$).
- We sometimes differentiate these estimands in the sample and them in the population.
- E.g., the SATE ($\frac{1}{N}\sum_{i=1}^{N}[Y_i(1) - Y_i(0)]$) vs. the PATE ($E[Y_i(1) - Y_i(0)]$).

Estimand vs. estimator

- Our estimands are functionals of the joint distribution of $\{Y_i(1), Y_i(0)\}$, $F(y_1, y_0)$.
- Such a distribution is unknown to the researcher.
- For example, the population average treatment effect (PATE) equals to

$$\tau_{PATE} = E[Y_i(1) - Y_i(0)] = \int (y_1 - y_0) dF(y_1, y_0)$$

- In the sample, we only have access to the observed outcome: $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$.
- Denote the joint distribution of $\{Y_i, D_i, \mathbf{X}_i\}, i \in \{1, 2, \dots, N\}$ as $G(y, d, \mathbf{x})$.
- Our estimator $\hat{\tau}$ is a functional of $G(y, d, \mathbf{x})$.
- Causal identification means that there exists a $\hat{\tau}$ such that $\hat{\tau}(G) = \tau(F)$ when N is infinite.

References I

Coppock, Alexander, and Dipin Kaur. 2022. "Qualitative Imputation of Missing Potential Outcomes." *American Journal of Political Science*.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81 (396): 945–60.

Imbens, Guido W, and Donald B Rubin. 2015. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Neyman, Jerzy S. 1923. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9.(tlanslated and Edited by Dm Dabrowska and Tp Speed, Statistical Science (1990), 5, 465-480)." *Annals of Agricultural Sciences* 10: 1–51.