# Statistics 5210 Sample Survey and Causal Inference
# Chapter 1: Introduction
# Part 2: Causal Inference

Zhengyuan Zhu [1]

Iowa State University

January 22, 2025

[1] Based on notes by Emily Berg and Jae-Kwang Kim

# Causal Inference

- We are interested in investigating the effect of treatment over control on the outcome $Y$ of interest.

| Subject | Treatment ($T$) | Outcome ($Y$) |
|---|---|---|
| Clinical trial | New drug | Health outcome |
| Labor economics | Job training | Employment status |
| Politics | Canvassing | Vote turnout |

- In this course, we assume that $T$ is binary: $T = 1$ for treatment and $T = 0$ for control.
- Two potential outcomes for $Y$: $Y(0)$ for $T = 0$ and $Y(1)$ for $T = 1$
- Terminology for $T$: action, manipulation, treatment, intervention

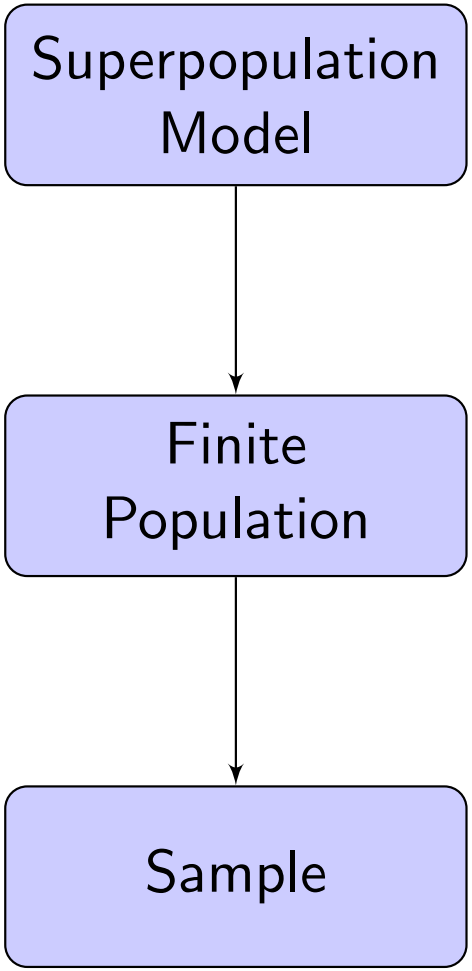
# Presumption

1. Although a unit was (at a particular point in time) exposed to a particular action, the same unit could have been exposed to an alternative action (at the same point in time).
2. Interpreting causation as a deterministic relation means that if A causes B, then A must always be followed by B. In this sense, war does not cause deaths, nor does smoking cause cancer or emphysema. As a result, many turn to a notion of probabilistic causation. Informally, A ("The person is a smoker") probabilistically causes B ("The person has now or will have cancer at some time in the future"), if the information that A occurred increases the likelihood of Bs occurrence. That is, $P(B \mid A) \geq P(B)$.

# Randomized Experiment vs Observational Study

- Randomized experiment (e.g. clinical trial): the event for $T_i = 1$ is completely determined by a pure random mechanism.
- Observational study: Each unit $i$ is assigned to $T_i = 0$ or $T_i = 1$ by other factors (such as physician's discretion or participants' choice).

# Superpopulation framework




Potential outcome random variables $(Y(0), Y(1))$

Obtain $\mathcal{F} = \{(Y_i(1), Y_i(0)); i = 1, \ldots, N\}$

Observe $T_i$ and $Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$ for $i = 1, \ldots, N$

# Defining Causal Effects

- For each unit $i$, we observe $(X_i, T_i, Y_i)$ where $Y_i = Y_i(T_i)$.
- We observe only one potential outcome for each unit. So, it is a missing data problem.
- In some literature, the unobserved potential outcome is called the *counterfactual outcome*. If $T_i = 1$, then we observe $Y_i(1)$ only and $Y_i(0)$ is the counterfactual outcome for unit $i$.
- Intuitively: if we could observe the counterfactual outcomes, then the difference $Y_i(1) - Y_i(0)$ is attributable to the treatments.
- Causal effect for unit $i$: $\tau_i = Y_i(1) - Y_i(0)$.

# Fundamental Problem of Causal Inference (Holland, 1986):

- Fundamental problem: It is impossible to estimate $\tau_i$ from the data, as we can only observe one potential outcome per participant.
- Two solutions to the fundamental problem of causal inference
  1. Scientific solution: We could use scientific theory to measure both potential outcomes. (eg: testing new diet on genetically identical twins).
  2. Statistical solution: We randomly assign treatment to individuals.

# Average Treatment Effect

- Unit-level causal effects are difficult to estimate.
- (Finite-) Population Average Treatment effect

$$\bar{\tau}_N = \frac{1}{N}\sum_{i=1}^{N}\tau_i = \frac{1}{N}\sum_{i=1}^{N}\{Y_i(1) - Y_i(0)\}$$

- (Superpopulation) Average Treatment Effect:

$$\mathrm{ATE} = E\left(Y(1) - Y(0)\right) := \tau$$

# Other causal parameters

- Average treament effect for the treated:

$$\text{ATT} = E\left(Y(1) - Y(0) \mid T = 1\right)$$

- Conditional average treatment effect (CATE):

$$\tau(\mathbf{x}) = E\left(Y(1) - Y(0) \mid \mathbf{X} = \mathbf{x}\right)$$

Applications to precision medicine and micro-targeting.

# Formal causal problem

- Data: IID observed data

$$(X_i, T_i, Y_i), \quad i = 1, \ldots, N$$

  where $Y_i = Y_i(1)$ if $T_i = 1$ and $Y_i = Y_i(0)$ if $T_i = 0$.
- The causal parameter is a function of potential outcomes. However, we do not observe the potential outcomes directly.
- Goal: We wish to estimate the causal parameters (such as ATE) from the observed data.
- Problem: Under what assumptions can we do this?

# Stable Unit Treatment Value Assumption (SUTVA)

- **Assumption 2.1 (no interference)** Potential outcomes for an individual are not affected by treatments received or potential outcomes of other individuals.
- **Assumption 2.2 (consistency)** There are no other versions of the treatment. The outcome $Y_i$ observed for the individual $i$, who received treatment $A_i$, is the same as his potential outcome for that treatment regardless of the conditions under which he received that treatment.
- Rubin (1980) called the Assumptions 2.1 and 2.2 above together the Stable Unit Treatment Value Assumption (SUTVA).
- Mathematical expression of SUTVA:

$$Y_i = Y_i(1)T_i + Y_i(0)(1 - T_i), \quad i = 1, \ldots, N \tag{1}$$

# Identification formula

- Under (1), we have

$$E(Y \mid T = t) = E(Y(t) \mid T = t) \tag{2}$$

  for $t = 0, 1$.
- The LHS is the conditional expectation in terms of the observation, while the RHS is the conditional expectation in terms of the potential outcome.

# Justification

# Randomized studies

- Randomization ensures treatment assignment is independent of all other factors, including individual characteristics
- Thus, we have

$$\{Y(1), Y(0)\} \perp T \tag{3}$$

- **Main Result**: Under (1) and (3), we have

$$E(Y \mid T = t) = E(Y(t)) \tag{4}$$

for $t = 0, 1$.

# Implication of (4)

- We can use

$$\hat{\tau} = \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i - \frac{1}{N_0}\sum_{i=1}^{N}(1 - T_i) Y_i \tag{5}$$

  to estimate $\tau = E\{Y(1)\} - E\{Y(0)\}$, where $N_1 = \sum_{i=1}^{n} T_i$ and $N_0 = \sum_{i=1}^{N}(1 - T_i)$.
- That is, $\hat{\tau}$ is unbiased for $\tau = E\{Y(1) - Y(0)\}$.
- The estimator in (5) is called the difference-in-means estimator (DIME).

# Justification

# Remark

- By algebra, we can obtain $\hat{\tau}$ in (5) as the joint minimizer of the following quantity:

$$Q(\alpha, \tau) = \sum_{i=1}^{N} (Y_i - \alpha - \tau T_i)^2$$

- That is, we can compute DIME as the slope in the ordinary regression of $Y_i$ on $T_i$.

# Justification

# Statistical interpretation

- Assume that the superpopulation model is

$$Y_i(t) = \mu_t + e_i(t)$$

for $t = 0, 1$, where $e_i(t) \sim (0, \sigma_t^2)$. We allow $e_i(1)$ and $e_i(0)$ to be correlated.

- Instead of observing $Y_i(1)$ and $Y_i(0)$, we observe

$$Y_i = T_i Y_i(1) + (1 - T_i) Y_i(0)$$

- In this case, we can express

$$Y_i = \mu_0 + \tau T_i + e_i$$

where $e_i = T_i e_i(1) + (1 - T_i) e_i(0)$.

- Under (3), we can obtain $e_i \sim (0, \sigma^2)$ for some $\sigma^2$.

# Justification

# Justification

# A toy example ($N = 4$)

- Let's look at an artificial data (of size $N = 4$).

| ID | $Y_i(1)$ | $Y_i(0)$ | $\tau_i = Y_i(1) - Y_i(0)$ |
|---|---|---|---|
| 1 | 4 | 1 | 3 |
| 2 | 6 | 3 | 3 |
| 3 | 7 | 5 | 2 |
| 4 | 8 | 6 | 2 |

- Population average treatment effect

$$\bar{\tau}_N = \frac{1}{4}(3 + 3 + 2 + 2) = 2.5$$

- The value of $Y_i(t)$ is observed only for $T_i = t$. We never observed $Y_i(1)$ and $Y_i(0)$ jointly.

- Suppose that we assign $N_1 = 2$ units to treatment group ($T_i = 1$) and assign $N_0 = 2$ units to control group.
- 6 possible group assignment

| case | $T = 1$ group ID | $T = 0$ group ID | DIME |
|---|---|---|---|
| 1 | 1, 2 | 3,4 | -0.5 |
| 2 | 1, 3 | 2,4 | 1.0 |
| 3 | 1, 4 | 2,3 | 2.0 |
| 4 | 2, 3 | 1,4 | 3.0 |
| 5 | 2, 4 | 1,3 | 4.0 |
| 6 | 3, 4 | 1,2 | 5.5 |

- For example, for case 3, we obtain

$$\hat{\tau} = \frac{1}{2}\{Y_1(1) + Y_4(1)\} - \frac{1}{2}\{Y_2(0) + Y_3(0)\} = 6.0 - 4.0 = 2.0$$

# Completely randomized experiment

- Completely randomized experiment (CRE): Assign the same selection probability to all possible assignments

| case | Trtment group ID | DIME ($\hat{\tau}$) | selection probability |
|---|---|---|---|
| 1 | 1, 2 | -0.5 | 1/6 |
| 2 | 1, 3 | 1.0 | 1/6 |
| 3 | 1, 4 | 2.0 | 1/6 |
| 4 | 2, 3 | 3.0 | 1/6 |
| 5 | 2, 4 | 4.0 | 1/6 |
| 6 | 3, 4 | 5.5 | 1/6 |

- In this case, the DIME has a discrete probability distribution.

- Probability mass function of $\hat{\tau}$:

$$P(\hat{\tau} = y) = \begin{cases} 1/6 & \text{if } y \in \{-0.5, 1.0, 2.0, 3.0, 4.0, 5.5\} \\ 0 & \text{otherwise.} \end{cases}$$

- Unbiased

- Variance

# Remark

- No model assumption about $y_i$ in the example: agnostic approach
- Design-based approach: the reference distribution is the sampling distribution generated by the repeated application of the given selection (or assignment) mechanism.
- Why randomization approach ?
  1. It creates comparable treatment and control groups on average.
  2. It serves as a "reasoned basis" for statistical inference.

# Comparison

| Area | Survey Sampling | Causal Inference |
|---|---|---|
| Target Population | Finite population | Potential outcome model |
| Parameter of interest | Descriptive parameter (ex: Total, mean, etc) | Causal parameter (ex: ATE, ATT, etc) |
| Gold standard (model-free) | Probability sampling | Randomized experiment |
| Partial Failure of randomization | Nonresponse | Non-compliance |
| No-randomization (model-based) | Non-probability sample | Observational study |

# Fisher randomization test (FRT)

- Interested in testing

$$H_0 : Y_i(1) = Y_i(0), \quad \forall i = 1, \ldots, n \tag{6}$$

- The above null hypothesis is called the *sharp null hypothesis* (or *strong null hypothesis)*.
- Idea for FRT: Under $H_0$, we can construct the sampling distribution of any test statistic

$$Q = Q(T, Y)$$

  where $T = (T_1, \ldots, T_n)$ and $Y = (Y_1, \ldots, Y_n)$.
- Let $T^{(1)}, \ldots, T^{(M)}$ be all possible vectors of $T$ under CRE. The sampling distribution of $Q$ is known due to the design of the CRE.

- For example, if $n_1 = n_0 = 2$, then the sampling distribution (or randomization distribution) of $Q = Q(T, Y)$ is summarized as follows.

| case $(k)$ | $T$ | Y | $Q$ | Probability |
|---|---|---|---|---|
| 1 | $T^{(1)} = (1, 1, 0, 0)$ | $Y^{(1)}$ | $Q(T^{(1)}, Y^{(1)})$ | 1/6 |
| 2 | $T^{(2)} = (1, 0, 1, 0)$ | $Y^{(2)}$ | $Q(T^{(2)}, Y^{(2)})$ | 1/6 |
| 3 | $T^{(3)} = (1, 0, 0, 1)$ | $Y^{(3)}$ | $Q(T^{(3)}, Y^{(3)})$ | 1/6 |
| 4 | $T^{(4)} = (0, 1, 1, 0)$ | $Y^{(4)}$ | $Q(T^{(4)}, Y^{(4)})$ | 1/6 |
| 5 | $T^{(5)} = (0, 1, 0, 1)$ | $Y^{(5)}$ | $Q(T^{(5)}, Y^{(5)})$ | 1/6 |
| 6 | $T^{(6)} = (0, 0, 1, 1)$ | $Y^{(6)}$ | $Q(T^{(6)}, Y^{(6)})$ | 1/6 |

- For example, $Y^{(3)} = (Y_1(1), Y_2(0), Y_3(0), Y_4(1))$ for case 3. Under the strong null hypothesis, we have $Y^{(1)} = \ldots = Y^{(6)}$ and we can compute $Q^{(k)} = Q(T^{(k)}, Y^{(k)})$ from the realized sample.

- Since the sampling distribution of $Q = Q(T, Y)$ can be constructed under $H_0$, we can compute the $p$-value given by

$$p_{\mathrm{FRT}} = \frac{1}{M} \sum_{k=1}^{M} \mathbb{I}\{Q(T^{(k)}, Y) \geq Q(T, Y)\}$$

- The above p-value measures the extremeness of the value of the realized test statistic with respect to its randomization distribution.
- This is the basic idea of Fisher's exact test. It is finite-sample exact in the sense that, under $H_0$ in (6),

$$P\left(p_{\mathrm{FRT}} \leq u\right) \leq u, \quad \forall u \in (0, 1).$$

- Fisher's exact test is not computationally feasible if $n$ is large.

# REFERENCES

Holland, P. (1986), 'Statistics and causal inference', *Journal of the American Statistical Association* **81**, 945–960.

Rubin, D. B. (1980), 'Randomization analysis of experimental data: The fisher randomization test comment', *Journal of the American Statistical Association* **75**, 591–593.