Modelling Causal Inference via Randomized Experiments

Gauranga Kr. Baishya, Chennai Mathematical Institute (CMI)

May 14, 2025

Analytical Smoking Example

Suppose we want to study the impact of smoking marijuana on life longevity. Consider the potential outcome model:

$$Y(d) = \eta_0 + \eta_1 d \tag{1}$$

where η_0 is baseline health and $\eta_1 = 0$ (no treatment effect). The treatment selection is given by:

$$D = 1(\nu > 0), \tag{2}$$

where ν is an unobserved factor affecting smoking choice. Assume:

$$E[\eta_0 \nu] < 0. \tag{3}$$

In other words, people who are more likely to smoke (higher ν) tend to have worse baseline health (η_0). Suppose that smoking marijuana has no causal effect on life longevity, then:

$$Y = Y(0) = Y(1) = \eta_0,$$

where η_0 represents the baseline life longevity.

An Analytical Smoking example

The expected outcome for smokers ($D = 1$) is given by:

$$E[Y \mid D = 1] = E[\eta_0 \mid \nu > 0] < E[\eta_0]. \tag{4}$$

since η_0 and ν are negatively correlated. &
The expected outcome for non-smokers ($D = 0$):

$$E[Y \mid D = 0] = E[\eta_0 \mid \nu \leq 0] > E[\eta_0]. \tag{5}$$

since η_0 and ν are negatively correlated.
Therefore (4) and (5),

$$\implies E[Y \mid D = 1] < E[Y \mid D = 0]. \tag{6}$$

Smokers appear to have worse health outcomes, but this is due to selection bias, not a causal effect of smoking.

Selection Bias

It is useful to emphasize that the main reason for having selection bias is that

$$E[Y(d) \mid D = 1] \neq E[Y(d)]$$

whenever D is not independent of $Y(d)$. If D and $Y(d)$ were independent, then

$$E[Y(d) \mid D = 1] = E[Y(d)]$$

would hold, since in this case, D is uninformative about the potential outcome and drops out from the conditional expectation. To sum up, the problem with observational studies like our contrived example is that the "treatment" variable D is determined by individual behaviors which may be linked to potential outcomes. This linkage generates selection bias - the disagreement between APE and ATE. There are many ways of addressing selection bias, one of which is through Randomised Control Trials (RCT)

RCT

Assumption (Random Assignment/Exogeneity) The treatment status is randomly assigned. Namely, D is statistically independent of each potential outcome $Y(d)$ for $d \in \{0, 1\}$, which is denoted as:

$$D \perp\!\!\!\perp Y(d)$$

and satisfies:

$$0 < P(D = 1) < 1.$$

Theorem : (Randomization Removes Selection Bias) Under the above Assumption, the average outcome in treatment group d recovers the average potential outcome under the treatment status d:

$$E[Y \mid D = d] = E[Y(d) \mid D = d] = E[Y(d)],$$

for each $d \in \{0, 1\}$. Hence, the average predictive effect (π) and the average treatment effect (δ) coincide:

$$\pi := E[Y \mid D = 1] - E[Y \mid D = 0] = E[Y(1)] - E[Y(0)] =: \delta.$$

Setup: Randomized Controlled Trial (RCT)

We observe an independent sample:

$$\{(Y_i, D_i)\}_{i=1}^{n}$$

where:

- Y_i is the observed outcome for unit i.
- D_i is the treatment indicator:

$$D_i = \begin{cases} 1, & \text{if unit } i \text{ receives treatment} \\ 0, & \text{if unit } i \text{ is in control group} \end{cases}$$

- Each (Y_i, D_i) is drawn i.i.d. from the same distribution as (Y, D).

Estimating the Two Group Means

We estimate the expected outcomes in each group:

$$\theta_d = E[Y \mid D = d], \quad d \in \{0, 1\}.$$

The sample estimates are:

$$\hat{\theta}_d = \frac{\sum_{i=1}^{n} Y_i \cdot 1(D_i = d)}{\sum_{i=1}^{n} 1(D_i = d)}$$

- $\hat{\theta}_0$ estimates the mean outcome for **control** group.
- $\hat{\theta}_1$ estimates the mean outcome for **treatment** group.

Treatment Effect and Regression Interpretation

The **Average Treatment Effect (ATE)** is:

$$\delta = \theta_1 - \theta_0.$$

The sample estimate:

$$\hat{\delta} = \hat{\theta}_1 - \hat{\theta}_0.$$

Claim: Under mild regularity conditions:

$$\sqrt{n}\begin{bmatrix}\hat{\theta}_0 - \theta_0 \\ \hat{\theta}_1 - \theta_1\end{bmatrix} \overset{a}{\sim} N(0, V),$$

where V is the variance-covariance matrix:

$$V = \begin{bmatrix}\frac{\text{Var}(Y|D=0)}{P(D=0)} & 0 \\ 0 & \frac{\text{Var}(Y|D=1)}{P(D=1)}\end{bmatrix}.$$

The variance of the estimated treatment effect:

$$\text{Var}(\hat{\delta}) = V_{11} + V_{22}.$$

Proof: Statistical Inference of 2 Means

We have an independent sample from an RCT,

$$\{(Y_i, D_i)\}_{i=1}^n,$$

where Y_i is the outcome and D_i is the treatment indicator ($d \in \{0,1\}$).
We want to estimate the group means:

$$\theta_d = E[Y \mid D = d], \quad d \in \{0,1\}.$$

A natural estimator for each θ_d is the sample average:

$$\hat{\theta}_d = \frac{\mathbb{E}_n[Y \cdot 1\{D = d\}]}{\mathbb{E}_n[1\{D = d\}]} = \frac{\frac{1}{n}\sum_{i=1}^n Y_i \, 1\{D_i = d\}}{\frac{1}{n}\sum_{i=1}^n 1\{D_i = d\}}.$$

Proof: Asymptotic Normality of the Group Means

Let the true mean be $\theta_d = E(Y \mid D = d)$. Then,

$$\hat{\theta}_d - \theta_d = \frac{\frac{1}{n}\sum_{i=1}^{n} Y_i \, 1\{D_i = d\}}{\frac{1}{n}\sum_{i=1}^{n} 1\{D_i = d\}} - \frac{\frac{1}{n}\sum_{i=1}^{n} \theta_d \, 1\{D_i = d\}}{\frac{1}{n}\sum_{i=1}^{n} 1\{D_i = d\}}.$$

Simplifying, we have:

$$\hat{\theta}_d - \theta_d = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \theta_d)\, 1\{D_i = d\}}{\frac{1}{n}\sum_{i=1}^{n} 1\{D_i = d\}} = \frac{\bar{Z}}{P_n(d)},$$

where

$$Z_i = (Y_i - \theta_d)\, 1\{D_i = d\} \quad \text{and} \quad P_n(d) = \frac{1}{n}\sum_{i=1}^{n} 1\{D_i = d\}.$$

Proof: CLT and Computing the Variance

Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$. Then by the Central Limit Theorem (CLT),

$$\sqrt{n}\,(\bar{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

Applying this to $Z_i = (Y_i - \theta_d)\,1\{D_i = d\}$:

$$\mathrm{Var}(Z_i) = E(Z_i^2) = E\left[(Y_i - \theta_d)^2 1\{D_i = d\}\right] = P(D = d) \cdot \mathrm{Var}(Y \mid D =$$

Thus,

$$\sqrt{n}\,\bar{Z} \xrightarrow{d} N\Big(0,\ P(D = d) \cdot \mathrm{Var}(Y \mid D = d)\Big).$$

Hence, using our earlier result,

$$\sqrt{n}(\hat{\theta}_d - \theta_d) = \frac{\sqrt{n}\,\bar{Z}}{P_n(d)} \xrightarrow{d} N\left(0, \frac{\mathrm{Var}(Y \mid D = d)}{P(D = d)}\right).$$

Joint Distribution of Both Groups

Moreover,

$$\sqrt{n}\begin{bmatrix}\hat{\theta}_0 - \theta_0 \\ \hat{\theta}_1 - \theta_1\end{bmatrix} \overset{d}{\sim} N\Big(0, V\Big),$$

where

$$V = \begin{bmatrix}\frac{\text{Var}(Y|D=0)}{P(D=0)} & 0 \\ 0 & \frac{\text{Var}(Y|D=1)}{P(D=1)}\end{bmatrix}.$$

Since the groups $D = 0$ and $D = 1$ are mutually exclusive, the off-diagonal elements are zero.

Estimating the Difference between Group Means

The Average Treatment Effect (ATE) is defined as:

$$\lambda = \theta_1 - \theta_0.$$

The estimator is:

$$\hat{\lambda} = \hat{\theta}_1 - \hat{\theta}_0.$$

From the previous result,

$$\sqrt{n}(\hat{\lambda} - \lambda) \xrightarrow{d} N\left(0, \frac{\text{Var}(Y \mid D = 0)}{P(D = 0)} + \frac{\text{Var}(Y \mid D = 1)}{P(D = 1)}\right).$$

The *Relative Effectiveness* is defined as:

$$f(\theta) = \frac{\theta_1 - \theta_0}{\theta_0} = \frac{\delta}{\theta_0}.$$

The corresponding estimator is:

$$f(\hat{\theta}) = \frac{\hat{\theta}_1 - \hat{\theta}_0}{\hat{\theta}_0}.$$

Proof Using the Delta Method

By a Taylor expansion,

$$f(\hat{\theta}) = f(\theta) + \nabla f(\theta)^T(\hat{\theta} - \theta) + R_n,$$

where R_n is a remainder term such that $\sqrt{n}\, R_n \to 0$. Multiplying by $\sqrt{n}$ gives:

$$\sqrt{n}\left(f(\hat{\theta}) - f(\theta)\right) = \nabla f(\theta)^T \sqrt{n}(\hat{\theta} - \theta) + \sqrt{n}\, R_n.$$

Since $\sqrt{n}\, R_n \to 0$, it follows that:

$$\sqrt{n}\left(f(\hat{\theta}) - f(\theta)\right) \xrightarrow{d} \nabla f(\theta)^T \sqrt{n}(\hat{\theta} - \theta).$$

Given that

$$\sqrt{n}(\hat{\theta} - \theta) \overset{d}{\sim} N(0, V),$$

by the properties of linear transformations of normal random variables,

$$\sqrt{n}\left(f(\hat{\theta}) - f(\theta)\right) \xrightarrow{d} N\left(0,\ G^T V G\right),$$

where $G = \nabla f(\theta)$.

Proof: Confidence Intervals for the Treatment Effect

The 95% confidence interval for δ is:

$$\hat{\delta} \pm z_{\alpha/2}\sqrt{\widehat{\mathrm{Var}}(\hat{\delta})}.$$

where:

- $z_{\alpha/2}$ is the critical value from the normal distribution.
- $\widehat{\mathrm{Var}}(\hat{\delta})$ is estimated using:

$$\widehat{\mathrm{Var}}(\hat{\delta}) = \frac{\widehat{\mathrm{Var}}(Y \mid D = 0)}{n_0} + \frac{\widehat{\mathrm{Var}}(Y \mid D = 1)}{n_1}.$$

Pre-treatment Covariates and Heterogeneity

Pre-treatment covariates (denoted as W): If we use these extra details, like age, we can get more precise estimates of how effective the treatment really is. Consider the **Conditional Average Treatment Effect (CATE)**:

$$\delta(W) = E[Y(1) \mid W] - E[Y(0) \mid W],$$

which compares the average potential outcomes conditional on a set of covariates W. Similarly **Conditional Average Predictive Effects (CAPE)**,

$$\pi(W) = E[Y \mid D = 1, W] - E[Y \mid D = 0, W],$$

However, these CAPE values will generally not agree with the CATE unless the treatment is assigned **randomly and independently** of the covariates W.

Pre-treatment Covariates and Heterogeneity

If we drop observations with missing data, we might unknowingly introduce post-treatment bias because missingness itself can be influenced by the treatment. For example, treated individuals are more likely to respond to follow-ups than the control group.

Theorem (Randomization with Covariates): Under the above assumption, the expected value of Y conditional on treatment status $D = d$ and covariates W coincides with the expected value of the potential outcome $Y(d)$ conditional on covariates W:

$$E[Y \mid D = d, W] = E[Y(d) \mid D = d, W] = E[Y(d) \mid W],$$

for each d. Hence, the conditional predictive effects and average treatment effects agree:

$$\pi(W) = \delta(W).$$

Limitations of RCT: SUTVA fails to hold

SUTVA: potential outcomes for a given observation respond only to its own treatment status; potential outcomes are invariant to random assignment of others

If a large number of people in a community get vaccinated, the disease spreads less overall because fewer people can catch and transmit it – herd immunity. SUTVA wouldn't hold.

Want to study the earning effect of getting a college degree. If only a small group gets a college degree, their wages may increase without affecting the job market. But if many people get degrees, the job market adjusts, reducing the wage advantage.

Limitations of RCT: Ethical, Practical, and Generalizability Concerns

Ethical: Many RCTs are infeasible because implementing them would be unethical. The key ethical principles are "Respect for persons," "Beneficence," and "Justice". E.g., A RCT where individuals are assigned to a smoking treatment group. The trial would violate the principle of "beneficence" – causing physical harm to study participants.

Practical: RCTs may also face practical issues. They can be prohibitively expensive when the treatment is costly, data collection costs are high, or the sample size required for adequate power is high.

Generalizability: Even if an RCT is well-executed and provides a reliable treatment effect, its findings may not generalize broadly. Differences in local conditions, implementation capacity, or the scale of the intervention can affect outcomes when applied in a different setting.

However..

RCTs have a profound influence on business, economics and science more generally. For example, RCTs are routinely used to study the efficacy of drugs and efficacy of various programs in labor and development economics, among other subfields of economics. The FDA moved to RCTs as the gold standard of proving that treatments work in 1970s-80s.

The expansion of the use of RCT in economics is associated with the work of Richard Thaler, the recipient of the 2017 Alfred Nobel Memorial Prize in Economics; Abhijit Banerjee, Esther Duflo, and Michael Kremer, the recipients of the 2019 Alfred Nobel Memorial Prize in Economics; John List, among many others.

References

- Chapter 2, Applied Causal Inference Powered by ML and AI, Victor Chernozhukov et al.
- Chapter 3: Randomized Controlled Trials, Statistical Tools for Causal Inference Sylvain Chabé-Ferret.