The Potential Outcome Framework

Alessandro Caggia

June 2025

Abstract

Plan for Research

1. What is the **Causal relationship** of interest?
2. What is the **Ideal Experiment** to capture it?
3. What is the **identification strategy**? that is, how to use what you have to actually approximate 2? eg. you have an exogenous Source of Variation (DiD, RD etc)
4. What is your **mode of statistical inference**? what is the population of interest? what is the sample? What assumptions for SE?

Causality

- Definition **in terms of counterfactual**: difference between the direct effect of that cause and the state that would have been observed had that specific intervention not taken place (the counterfactual).
- **Correlation does not imply causation**: we can define correlation based on the joint distribution of two variables but for causation we need further assumptions (think at OVB).

To define causality, three possible structures:

- A. Potential Outcomes model (Rubin 1974)
- B. Structral model
- C. Granger Sims causality: views causality as a prediciton property

1. The Potential Outcome Framework

Causality is tied to an action applied to some units (eg treatment is a job training program).[1]

- We primarily consider settings with **2 actions**:
  - $D_i = 1$ if exposed to the treatment,
  - $D_i = 0$ if exposed to no treatment (control).
- For each individual i, there are **two Potential Outcomes**:
  - Y_{1i}: potential outcome if individual i is treated ($D_i = 1$).[2]
  - Y_{0i}: potential outcome if individual i is not treated ($D_i = 0$).[3]
- We define the **Individual Causal (or Treatment) Effect** of D as: $Y_{1i} - Y_{0i}$.
- Switching equation:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i.$$

SUTVA

- **The potential outcomes for any unit do not vary with the treatments assigned to other units (ie no spillovers).**
- **Strong behavioral assumption** that rules out social interactions, peer effects, network effects and many types of general equilibrium (GE) effects and other spillovers.
- Possible workaround: when spillovers are an important part of the analysis, we can choose a more aggregate unit, e.g. one school, or school class instead of one student.

[1] Note: careful in distinguishing actions form attributes (e.g. gender).
[2] This has not to be a single value, as you will see in the following example it could be two values or a distribution of values)
[3] Remark: 1 vs 0 has nothing to do with time!

Example

- Effect of taking an aspirin on having a headache.
- Actions: $D_i = 1$ if aspirin, $D_i = 0$ if no aspirin
- Unit: individual i
- Outcome:
  - $Y_i = Y_{1i}$: headache ($Y_{1i} = 1$) or no headache ($Y_{1i} = 0$). Only observed if took an aspirin.
  - $Y_i = Y_{0i}$: headache ($Y_{0i} = 1$) or no headache ($Y_{0i} = 0$). Only observed if did not take an aspirin.
- Individual Causal Effect: $Y_{1i} - Y_{0i} \in \{-1, 0, 1\}$

		Y_{1i}	Y_{0i}	$Y_{1i} - Y_{0i}$
1	Headache gone only if you take aspirin	0	1	-1
2	Headache not gone even if you take aspirin	1	1	0
3	Headache gone even without aspirin	0	0	0
4	Headache gone only if no aspirin taken	1	0	1

1.1. ATE, ATT, ATNT

In practice, we have multiple units (i = 1...N) and **we want to study aggregate casual effects**. Aggregate the individual causal effects to get:

- **ATE(AverageTreatmentEffect)**:

$$\text{ATE} = \mathbb{E}(Y_{1i} - Y_{0i}) = \mathbb{E}(Y_{1i}) - \mathbb{E}(Y_{0i})$$

  for every individual one takes the difference in potential outcomes.
- **ATT(AverageTreatmentEffectontheTreated)**:

$$\text{ATT} = \mathbb{E}(Y_{1i} - Y_{0i} \mid D_i = 1) = \mathbb{E}(Y_{1i} \mid D_i = 1) - \mathbb{E}(Y_{0i} \mid D_i = 1)$$

  takes the difference in potential outcomes only for treated individuals! (more infra: we will have to reconstruct just the counterfactual outcome when untreated y_0 from the control group)
- **ATNT(AverageTreatmentEffectontheNon-Treated)**:

$$\text{ATNT} = \mathbb{E}(Y_{1i} - Y_{0i} \mid D_i = 0) = \mathbb{E}(Y_{1i} \mid D_i = 0) - \mathbb{E}(Y_{0i} \mid D_i = 0)$$

  This represents the average difference, for non-treated individuals, between their counterfactual outcome under treatment (unobserved) and their observed outcome under no treatment.
- If we could observe both Y_{1i} and Y_{0i} for all units:

$$\text{ATE} = \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}), \quad \text{ATT} = \frac{1}{N_{D=1}}\sum_{i:D_i=1}(Y_{1i} - Y_{0i})$$

- We can also focus on more general functionals of potential outcomes (e.g., medians, percentiles, quantiles).

1.2. Fundamental problem of causal inference

Above we made a crucial, and unrealistic, underlying assumption: **for the same individual, we assumed to have both the outcomes when treated and when not treated**. **BUT**, in practice, **we only observe one of Y_{1i} or Y_{0i} for each unit**, this is **equivalent to missing data problem**. Hence, we cannot directly estimate the

individual Treatment Effect because we only see one potential outcome for each individual, and we cannot use the formulas above for the ATE and ATT. For example, for a treated individual we only observe Y_{1i}, whereas Y_{0i} is the unobserved Counterfactual (what would have been the outcome if not treated).

1.3. Solving the fundamental problem of causal inference

- We have a population of N units: i=1,...,N.
- **The problem:** we never observe both Y_{1i} and Y_{0i} for the same individual only one of them is observed depending on treatment status D_i.
- A common first approach is to compare group means:

$$\begin{aligned}\mathbb{E}(Y_i \mid D_i=1)-\mathbb{E}(Y_i \mid D_i=0) &= \mathbb{E}(Y_{1i} \mid D_i=1)-\mathbb{E}(Y_{0i} \mid D_i=0)\\ &= \underbrace{\mathbb{E}(Y_{1i} \mid D_i=1)-\mathbb{E}(Y_{0i} \mid D_i=1)}_{\text{ATT}}\\ &\quad + \underbrace{\mathbb{E}(Y_{0i} \mid D_i=1)-\mathbb{E}(Y_{0i} \mid D_i=0)}_{\text{Selection Bias}}\end{aligned}$$

  - To get to the first step we used the switching equation: if $D_i = 1 \rightarrow Y_i = Y_{1i}$
  - To get to the second step you added and subtracted $\mathbb{E}(Y_{0i} \mid D_i = 1)$
- The root of the problem is **Selection Bias**: the observed control group may not be a valid counterfactual for the treated group.
  - **The fact is that you don't have the outcome the treated would have had hadn't been treated** $\mathbb{E}(Y_{0i} \mid D_i = 1)$. **You use as counterfactual the outcome of the control**, but if this does not correctly mimic the characteristics of the treated group you have selection bias.
  - For example, treated individuals may have self-selected into treatment due to the fact they had higher benefits from treatment (ATT > ATE).

1.4. Relationshp between the ATT and the ATE

What is the relation between ATT and ATE?

$$\begin{aligned}ATE =& \mathbb{E}(Y_{1i}-Y_{0i})\\ =& \mathbb{E}[\mathbb{E}(Y_{1i}-Y_{0i}|D_i)] \quad \text{(by LIE)}\\ =& \mathbb{P}(D_i=1)\cdot\mathbb{E}(Y_{1i}-Y_{0i}|D_i=1)+\mathbb{P}(D_i=0)\cdot\mathbb{E}(Y_{1i}-Y_{0i}|D_i=0)\\ =& \gamma ATT+(1-\gamma)ATNT\end{aligned}$$

Comment: the ATE is a weighted average of the ATT and the ATNT with the weights (γ) being determined by the fraction of the population being treated.

Useful to express ATT in terms of ATE and treatment effect heterogeneity ($ATT - ATNT$):

$$ATT = ATE + (1-\gamma)(ATT - ATNT)$$

Comment: The formula above recomposes the ATT! So the TE on the Treated is already there, while the TE on the fraction of the non-Treated is corrected by adding the gap they have with respect to the non-Treated.

Hence now we can recognize that **our initial simple difference in means is equal to**:

$$\begin{aligned}\mathbb{E}(Y_i \mid D_i=1)-\mathbb{E}(Y_i \mid D_i=0) &= ATT + \text{SelectionBias}\\ &= ATE + \underbrace{(1-\gamma)(ATT-ATNT)}_{\text{Treatment Effect heterogeneity}} + \text{SelB}\end{aligned}$$

Comment: Even with no selection bias, recovery of ATE requires equal treatment effects across treated and non-treated: $ATT = ATNT$. So we need both the absence of *a priori* and *a posteriori* differences! This is solved by randomization (for the same reasons why selection bias is solved by randomization), see below!

1.5. Random Assignment

To solve this problem, we need to impose the following crucial assumption:

- **Random Assignment**: treatment is independent of potential outcomes.

$$(Y_{1i}, Y_{0i}) \perp\!\!\!\perp D_i$$

- Here the idea is that **the control group has now the exact same composition as the treatment group**. So the baseline outcome y0 of treated individuals will be the same as the baseline outcome y0 of treated individuals:

$$\mathbb{E}(Y_{0i} \mid D_i = 1) = \mathbb{E}(Y_{0i} \mid D_i = 0)$$

- An (extra) logic step further: think again at the formulas above for the ATE and the ATT.
  - Now for the ATE we will need to assign to each individual in the treatment group the counterfactual outcome y0 extracted from a (similar) individual in he control group. And to each individual in the control group we will assign the counterfactual outcome y1 extracted from a similar individial on the treatment group.
  - for the ATT we will just consider the outcome y1 of treated individuals and subtract their counterfactual outocme extracted from the outocome y0 of (similar) individuals in the control group.
  - for the ATNT we will just consider the outcome y0 of control individuals and subtract it to the counterfactual outocme y1 extracted from the outcome y1 of (similar) individuals in the treatment group.
- **Random Assignment solves the selection problem** by guaranteeing:

$$D_i \perp (Y_{0i}, Y_{1i}) \quad \Rightarrow \quad \mathbb{E}(Y_{0i} \mid D_i = 1) = \mathbb{E}(Y_{0i} \mid D_i = 0)$$

So, the naive difference in means equals the true treatment effect:

$$\begin{aligned}\mathbb{E}(Y_i \mid D_i=1)-\mathbb{E}(Y_i \mid D_i=0) &= \mathbb{E}(Y_{1i} \mid D_i=1)-\mathbb{E}(Y_{0i} \mid D_i=1)\\ &= \mathbb{E}(Y_{1i}-Y_{0i} \mid D_i=1) = ATT\end{aligned}$$

Moreover, given independence:

$$ATT = \mathbb{E}(Y_{1i}-Y_{0i} \mid D_i=1) = \mathbb{E}(Y_{1i}-Y_{0i}) = ATE$$

Example

- $D_i = 1$ if individual has a bank account. Y_i: total savings.
- If we compare individuals with and without a bank account:

$$\mathbb{E}(Y_i \mid D_i=1)-\mathbb{E}(Y_i \mid D_i=0).$$

- We will not get the ATT (the causal effect of interest) if

$$\mathbb{E}(Y_{0i} \mid D_i=1)-\mathbb{E}(Y_{0i} \mid D_i=0) > 0.$$

- **Selection bias**: those with a bank account ($D_i = 1$) have a priori higher savings than those who do not, and would have had more savings even without a bank account (Y_{0i}) than those without one ($D_i = 0$).
- As $\mathbb{E}(Y_{0i} \mid D_i = 1)$ is unobservable we need assumptions or methods (e.g. matching, IV, RCT) to ensure:

$$\mathbb{E}(Y_{0i} \mid D_i=1) = \mathbb{E}(Y_{0i} \mid D_i=0).$$

More on Random assignment

- The key requirement is that Treated and Control units must be comparable. Hence, understanding the **assignment mechanism**i.e., the rule that determines treatment statusis fundamental.
- In most observational settings, the assignment mechanism is unknown or uncontrolled. An important exception is the **Randomized Control Trial (RCT)**, which ensures a well-behaved assignment via randomization. RCTs satisfy the following properties:
  - **Unconfounded Assignment**: the treatment is independent of potential outcomes.

- **Probabilistic Assignment**: every unit has a positive probability of receiving either treatment status (no deterministic assignment).
- **Individualistic Assignment**: the treatment probability of one unit does not depend on the characteristics or outcomes of other units.

2. Statistical Inference

Now that you have run your experiment, is time to run statistical inference!

2.1. Neyman Inference

- We use statistical inference to calculate standard errors, confidence intervals, and conduct hypothesis testing.
- Standard approach: **Neyman's repeated sampling framework**, where:
  - Sampling is assumed to be drawm from an infinite **superpopulation**.
  - **Uncertainty** (variance) arises from this sampling process (no uncertainty if we had the full population). The concept is that **the true parameter is fixed**, and the **uncertainty we face is driven by the fact that we are analysing a subsample from a population**.
- To derive standard errors, we rely on the **CLT** and construct confidence intervals using the normal distribution (for n large enough the sampling distribution of each sample mean (mean of T vs mean of C) converges to a normal distribution!)

2.2. Randomization Inference

- Idea: Different perspective, uncertainty comes from the fact that we only observe one assignment realization, while many others were possible.
- Randomization inference becomes possible under the **sharp null hypothesis**:

$$H_0 : Y_{1i} = Y_{0i} \quad \text{for all } i$$

That is, treatment has **no effect on any unit**. Hence, under this hypothesis, each unit's outcome is invariant to the treatment status:

- **Simulate** counterfactual scenarios by:
  1. **Plot their distribution of the observed outcomes** $Y_1, \dots, Y_N$ in the particular treatment realization you had.
  2. **Build the counterfactual**
     - 2a. **Re-randomize the treatment assignments** D_i according to the known experimental design (e.g., randomization with fixed number treated)
     - 2b. **Recompute the test statistic** (e.g., difference in means) under each simulated assignment;
  3. **Building the randomization distribution** of the statistic under the null;
  4. **Comparing** the observed test statistic (in 1.) to this null distribution (in 3.) to compute the p-value.

Unit i	Observed D_i	Observed Y_i	D_i^*	Y_i^*	D_i^{**}	Y_i^{**}
1	1	4	0	4	0	4
2	1	6	1	6	0	6
3	0	2	1	2	1	2
4	0	5	0	5	1	5
5	1	3	1	3	1	3
6	0	1	0	1	0	1
Means (Observed D_i)			Treated: 1,2,5	$\bar{Y}_T = 4.33$	Control: 3,4,6 **ATE = +1.67**	$\bar{Y}_C = 2.67$
Means (D_i^*)			Treated: 2,3,5	$\bar{Y}_T = 3.67$	Control: 1,4,6 **ATE = +0.33**	$\bar{Y}_C = 3.33$
Means (D_i^{**})			Treated: 3,4,5	$\bar{Y}_T = 3.33$	Control: 1,2,6 **ATE = −0.33**	$\bar{Y}_C = 3.67$

Remark: what changes from one column to the outher is not the outcomes but the composition of the treated and control groups.

- **Key Insight:** Under the sharp null H_0, outcomes are invariant to treatment, so we can treat them as known constants and generate the entire distribution of the test statistic using only variation in D_i. So the following is possible because we assume we have the same counterfactual outcomes and we are just redrawing treatment.
- This makes randomization inference a **nonparametric** (not assuming any parametric distribution for the outcomes (e.g., normality, linearity)) method for hypothesis testing, justified by our sharp null hypothesis.
- **In small samples**, randomization inference is often preferred for valid exact p-values, or at least should be reported alongside parametric results to show robustness.

3. Regression Analysis of Experiments

- We can estimate treatment effects either:
  - Non-parametrically: by comparing conditional means (up to now);
  - Parametrically: using a regression framework.
- Define the observed outcome as:

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$$

- Assume a constant treatment effect: $\rho = Y_{1i} - Y_{0i}$ and define:

$$\alpha = \mathbb{E}(Y_{0i}), \quad \eta_i = Y_{0i} - \mathbb{E}(Y_{0i})$$

take the expected value and define η_i in a way to center the regression residuals around 0 leading to the regression model:

$$Y_i = \alpha + \rho D_i + \eta_i$$

- From this regression, the expected difference in outcomes between treated and control is:

$$\mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0) = \rho + \mathbb{E}(\eta_i \mid D_i = 1) - \mathbb{E}(\eta_i \mid D_i = 0)$$

- This shows: **Selection Bias = correlation between D_i and η_i**. Another (microeconometric) way to frame the problem of endogeneity! Imagine: OVB creates Selection Bias
- only if the assumption above holds we can interpret ρ as the causal effect of D on Y. ρ = ATT = ATEy
- In correctly implemented randomized experiments, this condition holds **by design**, so OLS recovers the causal effect.

3.1. Do you need controls in your RCT regression?

- **In theory: Randomization ensures independence between treatment assignment and any covariate**. So:
  - The treatment should be independent of any control variable, thus it is not necessary to include controls in order to get an unbiased estimate for the treatment effect (think of condition for omitted variable bias: treatment is now independent form any other covariate!).;
- **In practice:** You must **check whether randomization actually balanced the covariates**. To verify:
  1. **Conductbalancetests**: compare pre-treatment covariate means in treatment vs. control groups. *Remark: expect differences to be statistically significant at 5% level in 5% of cases (1 out of 20 variables) due to false positives.*
  2. **RunajointF-test** to check if covariates jointly predict treatment.
  3. **Checkwhetherincludingcontrolschangesthecoefficienton** D. It shouldn't under successful randomization.
  4. **RobustnessRule:** You should **always include controls for unbalanced variables** to absorb bias and improve credibility.
- **Why include controls if there is no imbalance?**
  - If covariates help predict Y, they reduce residual variance:

$$\text{If } \sigma^2_{Y|D,X} < \sigma^2_{Y|D}, \text{ then precision increases.}$$

- But in general after the first few covariates, gains in precision diminish sd those covariates are weakly related to Y.

Remark: **Never include controls that might be affected by the treatment**. They would capture part of the effect of interest and including them would bias the results. Eg effect of being in the back of the class on grades. You are in the back of the class, you feel less pressure by prof, you use more the pc, you get lower grade. If you control for pc you stop the transition channel (will anlayse individuals at the same level of pc use!).

3.2. Interacting Covariates with the Treatment Dummy

- In addition to including covariates linearly, we can interact them with the treatment dummy D_i.
- This is useful when we expect the association between covariates and the outcome to vary by treatment status (i.e., heterogeneous covariate effects).
- These interactions allow the covariate slopes to differ across treatment and control, potentially improving the precision of the treatment effect estimate as well.
- A useful specification is:

$$Y_i = \alpha + \rho D_i + X_i\beta + D_i(X_i - \overline{X})\gamma + \eta_i$$

where:

- X_i is a vector of covariates,
- $(X_i - \overline{X})$ is mean-centered (demeaned) to ensure ρ is interpretable as the average treatment effect (ATE)[1],
- $D_i(X_i - \overline{X})$ allows covariate effects to vary with treatment.

[1]**Remark:** If we use the raw (non-demeaned) interaction D_iX_i, the coefficient $\hat{\rho}$ reflects the treatment effect only for the special case when $X_i = 0$. In this case, to recover the ATE we would need to adjust $\hat{\rho}$ by adding the average effect of the interaction term: ATE $= \hat{\rho} + \hat{\gamma} \cdot \overline{X}$. By demeaning X_i, the average of the interaction term $D_i(X_i - \overline{X})$ is zero by construction, so $\hat{\rho}$ directly estimates the ATE without further correction.

3.3. Randomization Techniques

Randomization is a critical component of experimental design. It ensures that treatment assignment is independent of potential outcomes and covariates.

3.3.1. How do we actually randomize?

- Randomization can be conducted **privately** (e.g., via random number generators) **or publicly** (e.g., via lotteries). Trade-offs, e.g., lotteries, are better to create trust, but they let people know they are treated.
- Several techniques are available to improve covariate balance:

1. **Pure Randomization (Single Draw)**: Assign units randomly. By chance, imbalances may occur (e.g., 1 in 10 covariates may differ at the 10% significance level).
2. **Stratification (Blocking)**: Divide units into strata based on key covariates (e.g., baseline outcomes, gender, location), then randomize within each stratum. This is to increase the likelihood covariates are balanced (10 females-90 males, randomizing globally could have 5-45. Randomize within stratum: sample 5 on 10 and then 45 on 90). Issue: if you create too many subcategories you don't have enough T and C for each subgroup.
3. **Pair-wise Matching**: Form pairs based on similarity in covariates; assign one unit in each pair to treatment and the other to control.
4. **Re-randomization methods**: Repeat randomization many times and select the one with the best covariate balance:
   - Redraw if any covariate shows imbalance (e.g., $p < 0.05$).
   - Choose the draw minimizing (between) the worst t-statistic from regressors (within) on treatment.

   Lost internal validity: your sample is no longer a random draw from the population ⇒ No longer Neyman inference, need adjustments.

3.3.2. Which randomization method is better?

- Athey and Imbens (2016) recommend **stratified randomization** as superior, especially when covariates are predictive of outcomes (i.e., we really need balance around those covariates).
- Stratification should be done *ex-ante* to improve balance, rather than relying on regression adjustment *ex-post*.
- If:
  - Covariates are weak predictors of outcomes.
  - Sample size is large (> 300)

  Then balancing on them does not matter much (quite obvious)

3.3.3. Randomization recommendations (Bruhn & McKenzie, 2009):

- The method of randomization affects which covariates to control for later: there is no true variability in sample draws; the sample is synthetic and we need to fix se. To do this, we add controls to resort to within-group comparisons.
  - *Stratification:* include strata dummies in regressions (not for bias, actually we statifies just to give balance).
  - *Pair-wise matching:* include matching variables in reg.
  - *Re-randomization:* include all variables for which balance was checked to decide whether to re-randomize.

3.3.4. Clustered Randomized Experiments

- When spillovers are possible (e.g., peer effects), it is better to randomize at a more aggregate level, ie the **cluster level**.
- Instead of assigning treatment randomly to units within a cluster, we assign treatment to entire clusters. All units within a cluster receive the same treatment.
- Then, econometric analysis can be done:
  - At the cluster level (preferred for inference).
  - At the individual level, using cluster-robust standard errors.

3.4. Power Computation

How large a sample do we need in order to detect a meaningful effect with high probability? That is, how much will I be able to learn from this experiment? Power computation! Suppose we aim to estimate the effect of a randomly assigned treatment D_i on an outcome Y_i using:

$$Y_i = \alpha + \delta D_i + \eta_i,$$

Power computation

- **Power:** If the true effect is δ, what is the probability of getting an estimate big enough to reject $\delta = 0$ (reject when false)?
- To achieve a power of $1 - \beta$ at a confidence level α, we need the true effect to be such as (see image below):

$$\delta / SE(\hat{\delta}) > (t_\beta + t_{\alpha/2})$$

- Then, the minimum detectable effect (MDE) is:

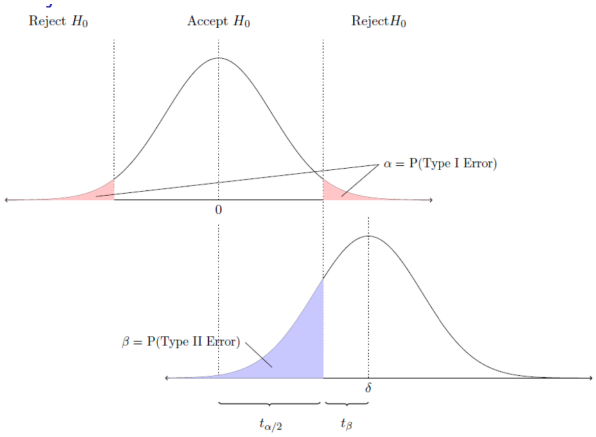
$$\delta_{MDE} = (t_\beta + t_{\alpha/2}) \cdot SE(\hat{\delta}) = (t_\beta + t_{\alpha/2}) \cdot \sqrt{\frac{\sigma^2}{P(1-P)N}}.$$

This tells us that we will be able to detect a smaller effect size if:

- Power increases with sample size N,

- Power is maximized when $P = 0.5$ (equal-sized treatment and control),
- Power increases when error variance σ^2 is smaller. High error variance means unexplained noise in the data.

Remark: why $\delta/SE(\hat{\delta}) > (t_\beta + t_{\alpha/2})$



- Upper panel assumes H_0 is true (ATE $= 0$).
  - To control the Type I error at a rate of α, we reject H_0 only if the absolute value of the observed t-statistic is equal to or larger than the critical value $t_{\alpha/2}$.
- Bottom panel assumes H_0 is false (ATE $= \delta$).
  - The power to detect an effect of size δ is the fraction of the area under the distribution that falls to the right of $t_{\alpha/2}$.
  - In that region we correctly reject the null hypothesis.
  - We limit the Type II error to some $\beta > \alpha$ (and then power is $1 - \beta$), such β will be characterized by some t_β.
  - the δ plotted in the bottom panel is exaclty the MDE! The MDE is the smallest value for which we can correctly reject H_0 with probability $1 - \beta$ at a significance level α.

3.4.1. Interpretation of Minimum Detectable Effect (MDE)

- If we aim at power $1 - \beta = 0.8$ with confidence level $\alpha = 0.05$
- Then, the minimum detectable effect is:

$$\delta_{MDE} = 2.8 \times \frac{\sigma}{\sqrt{P(1-P)N}}$$

- If $P = 0.5$, and $N = 1000$, we can detect an effect of 0.18σ, i.e., 0.18 standard deviations of the outcome.
- Is 0.18 standard deviations a meaningful effect size? It depends on the ratio between the mean and standard deviation of the outcome!
- Example 1: If the ratio is 1 (e.g., mean and sd of employment rate are both 0.5), the effect is 18% of the mean (or 9 percentage points), which seems large.
- Example 2: If the ratio is 2 (e.g., mean = 100, sd = 50), then $0.18\sigma = 9$, which is only 9% of the mean. This may not seem like a large effect.

3.4.2. Required Sample Size:

we want to kwnot what is the required smapel size to detect a pre-specified effect δ_0, we rearrange the MDE formula:

$$N > (t_\beta + t_{\alpha/2})^2 \cdot \frac{\sigma^2}{\delta_0^2 P(1-P)}.$$

Take $(t_\beta + t_{\alpha/2})^2 = 2.8^2$, $P = 0.5$, $1 - \beta = 0.8$, if tge effect is of 2 dollars with variance $\sigma^2 = 100$:

$$N > (2.8)^2 \cdot \frac{100}{4 \cdot 0.5 \cdot 0.5} = 784.$$

Interpretation: We need at least 784 observations to detect a 2 dollar effect with 80% power at the 5% significance level.

4. Internal Validity and Its Threats

- **Internal Validity**: ability of a study to estimate causal effects within the study sample.
- **External Validity**: ability to generalize results beyond the study sample.
- To maximize both, randomization should ideally be performed in two stages:
  1. Take a random sample from the population of interest (maximize external validity).
  2. Randomly allocate part of the sample to the treatment group and part to the control group (guarantees internal validity).

4.1. Main Threats to Internal Validity

An experiment eliminates the most important threats to internal validity in non-experimental settings (e.g. omitted variable bias). Yet, randomization is not a panacea.

- **Attrition:** Some individuals drop out of the sample after assignment.
- **Partial Compliance:** Some treated individuals do not receive treatment, or some controls receive it.
- **Externalities:** Spillovers across units (e.g., treated individuals affect control individuals).
- **Experimental Effects:** Subjects may change behavior due to being observed (Hawthorne/Henry effects).
- **Data Mining:** Testing many hypotheses without proper correction can lead to spurious findings.

These are also threats to non-experimental settings, except maybe experimental effects.

4.1.1. Attrition and Its Implications

A. Intuition

- If **random (unrelated to treatment)**: only reduces statistical power.
- If **correlated with treatment or outcomes** (e.g., low-benefit individuals drop out), introduces **attrition bias**.

B. See it in the model

- R_{1i}, R_{0i}: potential response indicators if the same individual is treated or controlled.
- switching eq: $R_i = R_{0i} + D_i(R_{1i} - R_{0i})$
- Y_i is only observed if $R_i = 1$. Meaning that, clearly, I can observe outcomes only of those who respond.
- Suppose $D_i \perp (Y_{0i}, Y_{1i}, R_{0i}, R_{1i})$. This means treatment is randomly assigned and independent of both potential outcomes and potential response behavior.
- However, the above could actually be violated. Outcomes and response can still be correlated: the issue arises *ex post*, after treatment is assigned. That is, the treatment itself may not affect average response rates, but eg:
  1. Individuals assigned to the control group may choose not to respond; and
  2. The probability of responding may depend on potential outcomes (e.g., people with lower Y_{0i} are less likely to respond if assigned to control).

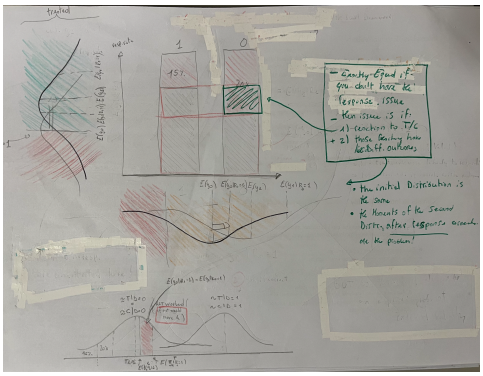
  Even though treatment is randomly assigned, the **composition of respondents across groups may differ**, leading to biased comparisons unless response is also independent of outcomes.

- What you observe: $\Delta = \mathbb{E}[Y_i \mid D_i = 1, R_i = 1] - \mathbb{E}[Y_i \mid D_i = 0, R_i = 1]$. Then by the switching equation:

$$\Delta = \underbrace{\mathbb{E}[Y_{1i} \mid R_{1i} = 1]}_{\text{Treated who responds}} - \mathbb{E}[Y_{0i} \mid R_{0i} = 1]$$
$$= \underbrace{\mathbb{E}[Y_{1i} \mid R_{1i} = 1] - \mathbb{E}[Y_{0i} \mid R_{1i} = 1]}_{\text{ATT on respondents when treated}} + \underbrace{(\mathbb{E}[Y_{0i} \mid R_{1i} = 1] - \mathbb{E}[Y_{0i} \mid R_{0i} = 1])}_{\text{Differential response across treat arms bias}}$$

  where second equation adds and subtract that term to give economic meaning. Note we have assumed the experiemtn has been successfull, so we have no selection bias. **Two problems:**

  1. $\underbrace{\mathbb{E}[Y_{1i} - Y_{0i}]}_{\text{Overall Outcome}} \neq \underbrace{\mathbb{E}[Y_{1i} - Y_{0i} \mid R_{1i} = 1]}_{\text{outcome just for respondents}}$ threatens External validity: not generalizable to the whole population, as these are only those who responded.
  2. Treatment induces different units to respond: $\mathbb{E}[Y_{0i} \mid R_{1i} = 1] \neq \mathbb{E}[Y_{0i} \mid R_{0i} = 1]$, Internal validity: I may be estimating the wrong treatment effect if the baseline outcome of treated respondent is different from the baseline outcome of control respondents. So in the Control you have eg less respondents and those who are responding less are individuals with lower outcomes (would create downwards bias).



C. Reasons for attrition

- Question not clear for T = 0, they don0t understand, what if those who don't understand have different outocmes? (non-natives?)
- t = 1 has more up to date contact (updated the email)
- unhappy to be in the control group!

*note always need systematically different outcomes (or is it just an issue of power)!

D. Suggestions

1. **Present the attrition rate** in control vs treatment.
2. **Compare baseline characteristics** of attritors vs non-attritors.
3. **Run a regression**:

$$\text{Attritor}_i = \alpha + \beta D_i + \gamma X_i + \delta(D_i \cdot X_i) + \varepsilon_i$$

   and test significance:

   - Coefficient on treatment evidence of differential attrition between treatment arms.
   - Interaction Tests whether the relationship between covariates and attrition differs by treatment status. That is: do certain subgroups drop out more in treatment than in control?

4. **If attrition is related to treatment**, apply:

   - Parametric methods: model attrition using covariates (control for).
   - Non-parametric methods: **bounding approaches**. However note: Bounds are going to be uninformative if attrition is very large.

E. Bounding Methods: Manski Bounds (Binary Outcomes: eg employed 1, unemployed 0)

- facile facile: la riposta è 1 (emplyed) o 0 (unemployed). ti hanno risposto solo l'80% ed hai una media di 0.4. LB: assumi il restante 20% abbia tutta 0. UB: oppure tutto 1.
- focus on the effects on the treated for the moment
- Your expected valye is, reasonably, a weighted average of the outcome of the reposndnets and the outcome of non reponsnents. $\mathbb{E}[R \mid D = 1]$ is the share of respondents among the treated:

$$\mathbb{E}[Y \mid D = 1] = \mathbb{E}[R \mid D = 1]\,\mathbb{E}[Y \mid D = 1, R = 1] \tag{1}$$

- $\mathbb{E}[Y \mid D = 1, R = 0]$ is **unobserved**, but bounded: $0 \leq \mathbb{E}[Y \mid D = 1, R = 0] \leq 1$
- Thus, bound $\mathbb{E}[Y \mid D = 1]$. So trivial: see the expression [1] above, the dummy takes 0,1. So worst case scenario all the unobserved share is multiplied by an expectaiton of y equal to 0 (LB), best case sceanrio all the unobserved share is multiplied by an expectation of y equal to 1.
- Provides ATE bounds. ATE no longer point identified but set identified.

Bounding Methods: Lee Bounds (for Continuous or Unbounded Outcomes)

- Suppose treatment increases the probability of response (suppose response rate is 10% higher in the treated group). Then **the treated sample may include people who would have not responded if assigned to control**. So the treated sample includes people for which you do not have a valid counterfactual in the control group.
- Lee bounds(non-binary): trim the treatment group to only those who would have responded in the control group.
- Two assumptions:
  - Extra 10% are those with **highest** 5% and **lowest** 5% outcomes: drop top (bottom) 5% from treated. ATT:
- We take them out from the treated mean to set-identify $\mathbb{E}[Y_1 \mid R_0 = 1]$

  $\mathbb{E}[Y_1 \mid R_0 = 1] \in [\mathbb{E}[Y \mid D = 1, R = 1, Y \leq p_{95}], \mathbb{E}[Y \mid D = 1, R = 1, Y \geq$
- Look at the CI above, the more you trim (the more the difference in response rate or the larger the % you trim) the larger the CI (the first CI element is the lower bound, because it is the expected value with the top 5% out, the second element is the upper bound, because you are removing the worst 5%). So the idea is that the lower bound gets lower and the upper bound gets larger.
  Remark: another option could have been to predict the probability of response on covariates, etc.

4.1.2. Partial compliance

Definition

- When only a fraction of the individuals who are offered the treatment take it up.
- Or, when some members of the control group manage to get the treatment.
- Default case in "encouragement" designs, individuals randomly encouraged to take treatment. Randomization affects the probability of getting treatment, not the treatment itself.

Solutions

- Focus on the *Intention-To-Treat* (ITT) effect:
  - Compare all treated *assigned* to treatment to all controls *assigned* to no-treatment.
  - Effect of being offered treatment.
- Try to estimate the ATE instead of the ITT?
  - We will see that we can use the dummy for being offered treatment as an Instrumental Variable (IV) for treatment, but retrieves LATE.
- We can use bound analysis (partial identification) to provide bounds for the ATE (assumptions of TE on Compliers and NT: e.g., you estimate proportions from actual take-up rates and then you impose assumptions on the outcomes e.g., Manski bounds).

How does Partial Compliance affect power?

- Assume that only a fraction c of the treatment group receives the treatment, and a fraction d of the control group receives it.
- Then:

$$\delta_{MDE} = 2.8 \cdot \frac{1}{c-d} \cdot \frac{\sigma}{\sqrt{P(1-P)N}}$$

- With $P = 0.5$, and $N = 1000$, with $c - d = 1$, we can detect an effect of 0.18σ.
- If $c - d = 0.5$, we can detect an effect of only 0.36σ.
- To ensure the same MDE, the sample size under 50% partial compliance would be almost 4 times larger than the sample under **full compliance (no linear relationship!)**. Why? recall you have to put N out of the sq root.

4.1.3. Externalities / spillovers

Individuals in the control group might be affected by the treatment. Examples: vaccines, imitation effects.

SUTVA is violated: the potential outcome of each individual depends on the vector of allocations to treatment and comparison groups.

- One solution: randomize at a level that allows capturing externalities (e.g., school level instead of individual level).
- Smart design used leveraging the violation of the SUTVA: compare adoption of fertilizer among the friends of treated farmers to that of the friends of control farmers.
- Actually could be interesting for research: to study peer effects, randomly assign individuals to different peer groups (e.g., Moving to Opportunity experiment in the US).

4.1.4. Experimental Effects (intrinsic to experiments)

The evaluation itself may lead individuals to change their behavior: experimenter effects. Example: farmers working harder in the experimental plot with randomly allocated fertilizer than in the rest of their farm.

- If the behavioral change is in the treatment group, we call it Hawthorne Effect (e.g., observed workers exhibit higher productivity).
- If in the control group: John Henry Effect.(Legendary American Steel driver, when he heard his output was being compared with that of a steam drill, he worked harder to outperform the machine; he died in the process).
- funny to try PLACEBO TESTS!
- how to fight these effects? justify saying they are not very strong, OR INDUCE them and bound their strength (eg through placebo)!

4.1.5. Data Mining: P-hacking when Studying Heterogeneity of Impact

- Underlying Issue: As for balancing: if you estimate many treatment effects (and interactions) you may have significance by chance (false positives)... and then you report just the significant!
- If we estimate the treatment effects for many different subgroups we need to **correct p-values for multiple testing.**
- Extra Good Practices:
  - write **pre-analsyis plan**
  - Stratify before randomizing: be sure there are no imbalances across characteristics in T and C group. Pre-randomization stratification ensures balance in key subgroups, thus Increasing power to detect true heterogeneity.
  - Avoid complex modeling to recover unbiased ATE. MAKE ANALYSSI SIMPLER AND MORE CREDIBLE Recall: One argument in favor of randomization is that everyone will get the same result and the researcher is bounded by the ex-ante design (no need to choose method, variables, specification).
  - study heterogeneity with machine learning causal forest. Non parametric (no linear form, normality etc): 1) flexible (no functional form assumed for treatemtn efeftc), 2) algoritmic discovery (he does the job, no manipulation), data drievn learning

4.1.6. External Validity

- Concerns about external validity are more relevant when there is significant **treatment effect heterogeneity** (cross-country analysis).
- **Generalization / Extrapolation** of results may not be reliable if:
  - **Issue of scalability**
  - It is difficult to **extrapolate results to other populations** (e.g., will poor farmers in Uganda respond like those in South America?).
- One approach to improve external validity is to **accumulate evidence across different settings and locations**.
- Another approach is to **account for differences in the distribution of characteristics** across settings that may drive treatment effect heterogeneity. Effect: So you see which characteristics are driving the treatment effect heterogeneity, then you can try to successfully predict the ATE on a training sample of which you know the values for those characteristics + the ATE (works)
- **Behavioral theory** can guide whether and how results should be extrapolated to new contexts.
- In general, randomized evaluations cannot capture **general equilibrium effects**. These effects may be observed if the unit of observation is large enough (e.g., country, state, or city).

5. Natural Experiments

5.1. Idea

- The exogenous source of variation comes from a randomly assigned variable that can be used as an instrument for treatment.
- Classic examples: Vietnam draft lottery to study military service effects, Mariel boatlift for immigration impacts, changes in divorce laws, etc.
- These settings mimic random assignment, enabling causal inference without explicit experimental control.
- **Threats to internal validity** increase:
  - policy endogeneity (to an area (unobervabels)-> affects treatemtn and outcome),

  - omitted variable bias (Suppose we estimate the effect of a health insurance reform but do not control for baseline regional health infrastructure),
  - outcome trends (DId with wrong pretrends),
  - simultaneity (A government might increase subsidies for schools in response to declining test scores),
  - sample selection (f only a non-random subset of the population is observed or affected by the treatment, estimates will not be representative or causal.).
- Similar challenges as RCTs regarding **external validity**, with one key difference: natural experiments typically do not require informed consent.
- Rosenzweig and Wolpin (2000) emphasize *natural natural experiments* those that rely solely on natural exogenous events, not policy changes or institutional assignments.

5.2. Debate

- Natural experiments and quasi-experimental methods are central to the so-called **credibility revolution** in microeconomics.
- Critics argue that **overenthusiasm** for these methods can lead to oversimplified claims e.g., reliance on single-equation models with robust standard errors without deeper understanding.