DegreeLive

Bachelor of Commerce (B2001) | Level 3 • Part B Major | Monash University

Applied econometrics

Course Code: ETC3410 • Module 1 Notes

Module 1: Causal Inference, Potential Outcomes & Randomised Experiments

1.1 From Association to Causation in Applied Microeconometrics

Empirical microeconometrics examines the economic behaviour of individual agents—workers, households, students, patients, and firms—using micro-level datasets such as the Household, Income and Labour Dynamics in Australia (HILDA) Survey, Australian Bureau of Statistics (ABS) Longitudinal Labour Force microdata, and Australian Taxation Office (ATO) administrative tax records. A central distinction separates descriptive prediction from causal inference. Descriptive regression estimates the conditional expectation $E[Y_i \mid X_i = x]$, which summarises how an outcome Y_i co-varies with observable covariates X_i in a population. Causal inference asks a counterfactual question: how would Y_i change if an intervention or policy externally manipulated a treatment variable D_i from 0 to 1, holding all other determinants of Y_i fixed?

In Australian economic policy evaluation—such as measuring whether completing a Victorian TAFE Certificate III qualification raises subsequent annual earnings, or whether an energy-efficiency subsidy reduces household electricity consumption—observational correlations frequently mislead policymakers. Individuals who select into training programmes or apply for subsidies systematically differ from non-participants in unobserved motivation, prior wealth, local labour market exposure, and discount rates. Formalising causal claims requires a mathematical framework that defines counterfactual states independently of the observed assignment mechanism.

1.2 The Rubin Causal Model and Potential Outcomes

The modern econometric paradigm for causal inference is the **Rubin Causal Model (RCM)**, originally introduced for randomised agricultural trials by Jerzy Neyman (1923) and generalised to observational studies by Donald Rubin (1974). Consider a population of N units indexed by $i = 1, \dots, N$, and a binary treatment indicator $D_i \in \{0, 1\}$, where $D_i = 1$ denotes receiving the treatment (for example, participating in a Commonwealth wage-subsidy programme) and $D_i = 0$ denotes the control state.

Each unit i possesses two **potential outcomes** defined prior to treatment assignment:

- $Y_i(1)$**:** The outcome unit i would realise if assigned to treatment ($D_i = 1$).
- $Y_i(0)$**:** The outcome unit i would realise if assigned to control ($D_i = 0$).

The **Individual Treatment Effect (ITE)** for unit i is the difference between its two potential states:

$$\tau_i = Y_i(1) - Y_i(0)$$

Where:

- τ_i is the causal effect of treatment on unit i.
- $Y_i(1)$ is unit i's potential outcome under treatment.
- $Y_i(0)$ is unit i's potential outcome under control.

The Switching Equation and the Fundamental Problem of Causal Inference

For any given unit i, the econometrician observes only the realised treatment status D_i and a single realised outcome Y_i, linked to the potential outcomes via the **switching equation**:

Degreelive

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = Y_i(0) + D_i [Y_i(1) - Y_i(0)]$$

Where:

- Y_i is the observed outcome for unit i.
- $D_i \in \{0, 1\}$ is the binary treatment indicator.
- $Y_i(1) - Y_i(0)$ is the latent individual causal gain that activates solely when $D_i = 1$.

Because unit i cannot simultaneously receive and not receive the treatment at the same point in time, either $Y_i(1)$ or $Y_i(0)$ is permanently missing. Holland (1986) termed this the **Fundamental Problem of Causal Inference**: individual treatment effects τ_i are inherently unobservable, turning causal inference into a missing-data problem at the population level.

Stable Unit Treatment Value Assumption (SUTVA)

Writing potential outcomes solely as $Y_i(1)$ and $Y_i(0)$ implicitly imposes the **Stable Unit Treatment Value Assumption (SUTVA)**, which comprises two strict requirements:

1. **No Interference Between Units:** The potential outcome of unit i depends exclusively on unit i's own treatment assignment D_i, and not on the treatment vector $\mathbf{D}_{-i}$ assigned to other units. In Australian labour market policies, SUTVA fails if subsidising job-seekers in Western Sydney displaces unsubsidised workers competing for the same retail vacancies (general equilibrium or spillover effects).
2. **No Hidden Variations of Treatment:** There is only one version of $D_i = 1$ and one version of $D_i = 0$. If a university mentoring scheme varies drastically in intensity across campuses—for example, Clayton offering 12 weekly tutorials while Caulfield offers 2 online webinars—then $Y_i(1)$ is ill-defined without disaggregating treatment dosage.

1.3 Population Causal Estimands: ATE, ATT, ATU, and CATE

Since individual causal effects τ_i cannot be identified without strong parametric restrictions, applied econometricians target population or subpopulation averages of τ_i. Four primary estimands govern policy analysis:

$$\text{ATE} = E[Y_i(1) - Y_i(0)] = E[\tau_i]$$

Where:

- ATE is the **Average Treatment Effect** across the entire target population.
- $E[\cdot]$ denotes the expectation over the joint distribution of $(Y_i(1), Y_i(0))$.

$$\text{ATT} = E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

Where:

- ATT is the **Average Treatment Effect on the Treated**, measuring the mean causal gain specifically among units that actually received the programme.
- $D_i = 1$ conditions the expectation on the treated subpopulation.

DegreeLive

$$\text{ATU} = E[Y_i(1) - Y_i(0) \mid D_i = 0]$$

Where:

- ATU is the **Average Treatment Effect on the Untreated**, representing the hypothetical gain if the control group were enrolled.

$$\text{CATE}(x) = E[Y_i(1) - Y_i(0) \mid X_i = x]$$

Where:

- $\text{CATE}(x)$ is the **Conditional Average Treatment Effect** for the subgroup sharing baseline covariate profile $X_i = x$.

Let $\pi = P(D_i = 1)$ denote the population share treated. By the Law of Iterated Expectations, ATE is a weighted average of ATT and ATU:

$$\text{ATE} = \pi \cdot \text{ATT} + (1 - \pi) \cdot \text{ATU}$$

Where:

- $\pi = P(D_i = 1)$ is the marginal treatment probability in the population.
- $1 - \pi = P(D_i = 0)$ is the marginal control probability.

Estimand	Formal Definition	Policy Question Answered	Australian Policy Example
ATE	$E[Y_i(1) - Y_i(0)]$	What is the expected impact of mandating the policy universally?	Effect of compulsory Superannuation Guarantee rate rise on total household saving
ATT	$E[Y_i(1) - Y_i(0) \mid D_i = 1]$	Did existing programme participants benefit relative to their counterfactual?	Earnings gain for unemployed Australians who actually enrolled in Workforce Australia training
ATU	$E[Y_i(1) - Y_i(0) \mid D_i = 0]$	What would happen if a voluntary scheme expanded to non-participants?	Expected electricity savings if non-adopter households received rooftop solar rebates
CATE(x)	$E[Y_i(1) - Y_i(0) \mid X_i = x]$	Which demographic or regional subgroups benefit most?	Effect of childcare subsidy reform conditional on metropolitan vs. regional postcode

1.4 Selection Bias Decomposition

In observational data, an analyst might naively compare the average observed outcome of treated units against the average observed outcome of untreated units: $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$. Applying the switching equation and adding and subtracting the unobservable counterfactual $E[Y_i(0) \mid D_i = 1]$ decomposes the naive difference-in-means into two distinct components:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \underbrace{E[Y_i(1) - Y_i(0) \mid D_i = 1]}_{\text{ATT}} + \underbrace{\{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\}}_{\text{Selection Bias}}$$

DegreeLive

Where:

- $E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0]$ is the observable naive difference in sample means.
- $E[Y_i(1) - Y_i(0) \mid D_i = 1]$ is the true ATT.
- $E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$ is the **Selection Bias**, capturing pre-existing baseline differences in the untreated potential outcome $Y_i(0)$ between those who enter treatment and those who do not.

Moreover, if treatment effects are heterogeneous ($\text{ATT} \neq \text{ATU}$), comparing the naive estimator to the population ATE introduces an additional **heterogeneous treatment effect bias** term $(1 - \pi)(\text{ATT} - \text{ATU})$:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = \text{ATE} + \text{Selection Bias} + (1 - \pi)(\text{ATT} - \text{ATU})$$

Where:

- $(1 - \pi)(\text{ATT} - \text{ATU})$ reflects sorting on gains (Roy-model selection), where individuals with larger expected causal benefits τ_i disproportionately select into treatment.

Figure 1.1: Selection Bias vs. Causal Effect Decomposition



Worked Numerical Example 1.1: Selection Bias in an Australian Wage-Subsidy Programme

Suppose the Department of Employment and Workplace Relations evaluates a voluntary vocational upskilling voucher ($D_i = 1$) across $N = 1,000$ unemployed Australians, where 400 workers enrol ($\pi = 0.40$) and 600 do not ($1 - \pi = 0.60$). Let potential annual post-programme earnings (in AUD) take the following population subgroup means:

- Among participants ($D_i = 1$): mean counterfactual earnings without training are $E[Y_i(0) \mid D_i = 1] = \text{A\$46,000}$, and mean realised earnings with training are $E[Y_i(1) \mid D_i = 1] = \text{A\$52,000}$.
- Among non-participants ($D_i = 0$): mean realised earnings without training are $E[Y_i(0) \mid D_i = 0] = \text{A\$40,000}$, while hypothetical counterfactual earnings if trained would be $E[Y_i(1) \mid D_i = 0] = \text{A\$43,500}$.

Step 1 — Compute the Naive Difference-in-Means:

$$\Delta_{\text{naive}} = E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = 52,000 - 40,000 = \text{A\$12,000}$$

Where:

- Δ_{naive} is the raw observable gap in annual earnings between participants and non-participants.

Step 2 — Compute ATT, ATU, and ATE:

$$\text{ATT} = 52,000 - 46,000 = \text{A\$6,000}, \qquad \text{ATU} = 43,500 - 40,000 = \text{A\$3,500}$$

$$\text{ATE} = 0.40(6,000) + 0.60(3,500) = 2,400 + 2,100 = \text{A\$4,500}$$

Where:

- $\text{ATT} = \text{A\$6,000}$ is the true causal earnings gain for those who took up the voucher.
- $\text{ATU} = \text{A\$3,500}$ is the counterfactual gain for non-enrollees.
- $\text{ATE} = \text{A\$4,500}$ is the population-wide average causal effect.

Step 3 — Quantify Selection Bias and Sorting-on-Gains Bias:

$$\text{Selection Bias} = E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] = 46,000 - 40,000 = \text{A\$6,000}$$

$$\text{Naive} = \text{ATT} + \text{Selection Bias} = 6,000 + 6,000 = \text{A\$12,000}$$

Where:

- The naive comparison of A\$12,000 overstates the true ATT (A\$6,000) by exactly A\$6,000 (100% upward bias) because higher-earning-capacity workers voluntarily enrolled, and overstates the population ATE (A\$4,500) by A\$7,500 because $(1-\pi)(\text{ATT} - \text{ATU}) = 0.60(6,000 - 3,500) = \text{A\$1,500}$.

1.5 Randomised Controlled Trials (RCTs): Identification and Estimation

Randomised Controlled Trials resolve the selection bias problem by design. When treatment D_i is assigned via a known randomisation mechanism—such as a computer-generated coin flip with fixed probability $p \in (0,1)$—treatment status is statistically independent of both potential outcomes and all baseline characteristics:

$$(Y_i(1), Y_i(0), X_i) \perp D_i$$

Where:

- $\perp$ denotes statistical independence between the vector of potential outcomes and covariates on the left and random assignment D_i on the right.

Under random assignment, conditional expectations equal unconditional expectations: $E[Y_i(0) \mid D_i = 1] = E[Y_i(0) \mid D_i = 0] = E[Y_i(0)]$, and similarly $E[Y_i(1) \mid D_i = 1] = E[Y_i(1)]$. Selection bias vanishes identically to zero, and $\text{ATT} = \text{ATU} = \text{ATE}$:

$$E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1)] - E[Y_i(0)] = \text{ATE}$$

Where:

- The simple sample difference-in-means $\hat{\tau}_{\text{DIM}} = \bar{Y}_1 - \bar{Y}_0$ is an unbiased estimator of the population ATE.

DegreeLive

Neyman-Rubin Variance and Regression Equivalence

Consider a completely randomised experiment with N_1 treated units and $N_0 = N - N_1$ control units. Under the Neyman finite-population framework (where potential outcomes are fixed and randomness arises solely from the assignment vector), the sampling variance of $\hat{\tau}_{\text{DIM}} = \bar{Y}_1 - \bar{Y}_0$ is:

$$\text{Var}(\hat{\tau}_{\text{DIM}}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}$$

Where:

- $S_1^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(1) - \bar{Y}(1))^2$ is the finite-population variance of treated potential outcomes.
- $S_0^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(0) - \bar{Y}(0))^2$ is the finite-population variance of control potential outcomes.
- $S_\tau^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\tau_i - \bar{\tau})^2$ is the variance of individual treatment effects across units.

Because S_τ^2 cannot be identified without observing both $Y_i(1)$ and $Y_i(0)$ for the same unit, the standard Neyman conservative variance estimator omits $-S_\tau^2/N$ and substitutes sample variances s_1^2 and s_0^2:

$$\widehat{\text{Var}}_{\text{Neyman}}(\hat{\tau}_{\text{DIM}}) = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}$$

Where:

- s_1^2 and s_0^2 are the sample variances within the treated and control arms, respectively.
- This conservative estimator is algebraically identical to the Eicker-Huber-White (HC2) heteroskedasticity-robust OLS variance of $\hat{\beta}_1$ in the bivariate regression $Y_i = \beta_0 + \beta_1 D_i + u_i$.

Regression Adjustment and Lin (2013) Interacted Estimator

Although the unadjusted difference-in-means $\hat{\tau}_{\text{DIM}}$ is unbiased, pre-treatment covariates X_i (such as prior earnings, age, or baseline test scores) frequently explain substantial variance in Y_i. Including demeaned covariates $\widetilde{X}_i = X_i - \bar{X}$ improves precision without introducing bias when estimated via Lin's (2013) fully interacted specification:

$$Y_i = \alpha + \tau D_i + \gamma' \widetilde{X}_i + \delta' (D_i \cdot \widetilde{X}_i) + u_i$$

Where:

- $\widetilde{X}_i = X_i - \bar{X}$ is the vector of baseline covariates centred at their full-sample means.
- $\hat{\tau}$ is the OLS estimate of the ATE, which remains consistent even if the linear model is misspecified.
- γ captures the baseline relationship between covariates and $Y_i(0)$, while δ allows covariates to shift treatment responsiveness.

Freedman (2008) showed that adding covariates without interactions ($Y_i = \alpha + \tau D_i + \gamma' X_i + u_i$) can slightly degrade asymptotic precision relative to unadjusted difference-in-means when treatment ratios differ from 0.5 and treatment effects are heterogeneous. Lin (2013) proved that including the interaction $D_i \cdot \widetilde{X}_i$ guarantees asymptotic efficiency at least as high as both the unadjusted estimator and the additive covariate regression.

DegreeLive

Randomisation Inference (Fisher's Exact Test)

When sample sizes are small or outcome distributions are highly skewed—common in Australian firm-level innovation grants or regional pilot trials—asymptotic normal approximations can misstate Type I error rates. **Fisher's Randomisation Inference (RI)** tests the sharp null hypothesis of zero treatment effect for every unit:

$$H_0^{\text{sharp}} : Y_i(1) = Y_i(0) \quad \text{for all } i = 1, \dots, N$$

Where:

- Under H_0^{sharp}, every unit's counterfactual outcome is identical to its observed outcome, so the full schedule of potential outcomes is known.

The exact randomisation p-value is computed by permuting the treatment vector $\mathbf{D}^{(m)}$ across all M admissible randomisations under the experimental design (e.g., $\binom{N}{N_1}$ combinations or $10,000$ Monte Carlo draws) and calculating the fraction whose test statistic $|T(\mathbf{D}^{(m)}, \mathbf{Y})|$ equals or exceeds the observed statistic $|T_{\text{obs}}|$:

$$p_{\text{RI}} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}(|T(\mathbf{D}^{(m)}, \mathbf{Y})| \geq |T_{\text{obs}}|)$$

Where:

- M is the total number of permissible treatment permutations.
- $\mathbf{1}(\cdot)$ is the indicator function equal to 1 if the condition holds and 0 otherwise.

1.6 Experimental Design Extensions: Stratification and Clustering

Real-world field experiments conducted by Australian agencies (such as the Behavioural Economics Team of the Australian Government, BETA) rarely rely on simple unstratified coin flips. Two design architectures dominate applied practice.

Stratified (Blocked) Randomisation

In stratified randomisation, units are partitioned into S mutually exclusive strata (e.g., states/territories or baseline income quartiles) prior to randomisation, and treatment is randomly assigned within each stratum s with proportion p_s. Stratified randomisation prevents accidental covariate imbalances across key subgroups and reduces sampling variance. The stratified ATE estimator aggregates stratum-specific differences weighted by stratum population shares $w_s = N_s/N$:

$$\hat{\tau}_{\text{strat}} = \sum_{s=1}^{S} w_s(\bar{Y}_{1s} - \bar{Y}_{0s}), \qquad \widehat{\text{Var}}(\hat{\tau}_{\text{strat}}) = \sum_{s=1}^{S} w_s^2 \left(\frac{s_{1s}^2}{N_{1s}} + \frac{s_{0s}^2}{N_{0s}}\right)$$

Where:

- $w_s = N_s/N$ is the share of the experimental sample belonging to stratum s.
- $\bar{Y}_{1s}$ and $\bar{Y}_{0s}$ are the treated and control sample means within stratum s.
- N_{1s} and N_{0s} are the treated and control sample counts inside stratum s.

DegreeLive

When treatment probabilities p_s vary across strata, regression estimation must include stratum fixed effects λ_s (and interact D_i with demeaned stratum indicators if targeting the ATE rather than a precision-weighted average).

Cluster-Randomised Trials (CRTs) and Design Effect

When interventions operate at the school, clinic, or local government area (LGA) level—for instance, trialling a new financial literacy curriculum across 60 Victorian secondary schools—randomising individual students inside the same classroom risks severe SUTVA violations via peer spillovers. Instead, entire clusters $c = 1, \dots, C$ of average size m are randomised. Positive correlation among individuals within the same cluster reduces effective sample size, quantified by the **Intracluster Correlation Coefficient (ICC)** ρ and the **Design Effect (DEFF)**:

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_w^2}, \qquad \text{DEFF} = 1 + (m-1)\rho$$

Where:

- σ_b^2 is the between-cluster variance component of the outcome.
- σ_w^2 is the within-cluster variance component.
- m is the number of individual observations per cluster.
- DEFF is the factor by which variance increases relative to an individually randomised trial of the same total headcount $N = C \cdot m$.

Consequently, standard errors in a CRT must be clustered at the level of randomisation (the school or LGA). Failing to cluster standard errors dramatically understates sampling variability and inflates false positive rates.

Worked Numerical Example 1.2: Precision, Regression Adjustment, and Cluster Design Effect

BETA evaluates a digital superannuation nudging trial across $C = 40$ workplaces, each employing $m = 25$ workers ($N = 1,000$ total workers). Twenty workplaces ($N_1 = 500$ workers) are randomly assigned to receive personalised contribution prompts ($D_c = 1$), and twenty workplaces ($N_0 = 500$ workers) receive standard annual statements ($D_c = 0$). Observed annual voluntary contributions yield $\bar{Y}_1 = \text{A\$2,450}$ ($s_1 = \text{A\$1,100}$) and $\bar{Y}_0 = \text{A\$2,100}$ ($s_0 = \text{A\$1,000}$). The workplace intracluster correlation is $\rho = 0.08$.

Step 1 — Point Estimate of ATE:

$$\hat{\tau}_{\text{DIM}} = \bar{Y}_1 - \bar{Y}_0 = 2,450 - 2,100 = \text{A\$350}$$

Where:

- $\hat{\tau}_{\text{DIM}} = \text{A\$350}$ is the estimated average increase in annual voluntary superannuation contributions.

Step 2 — Naive (Unclustered) Neyman Standard Error:

$$\text{SE}_{\text{naive}} = \sqrt{\frac{1,100^2}{500} + \frac{1,000^2}{500}} = \sqrt{2,420 + 2,000} = \sqrt{4,420} \approx \$66.48$$

Where:

- SE_{naive} ignores within-workplace correlation, producing an artificially high t-statistic of $350/66.48 = 5.26$.

Step 3 — Design Effect and Cluster-Adjusted Standard Error:

DegreeLive

$$\text{DEFF} = 1 + (m-1)\rho = 1 + (25-1)(0.08) = 1 + 1.92 = 2.92$$

$$\text{SE}_{\text{cluster}} = \text{SE}_{\text{naive}} \times \sqrt{\text{DEFF}} = 66.48 \times \sqrt{2.92} \approx 66.48 \times 1.7088 = \$113.60$$

Where:

- The effective sample size drops from $1,000$ to $N_{\text{eff}} = 1,000/2.92 \approx 342.5$ independent observations.
- The true cluster-adjusted t-statistic is $t = 350/113.60 = 3.08$ ($p < 0.01$ with $df = C - 2 = 38$).

1.7 Covariate Balance Diagnostics and Internal Validity Threats

Randomisation guarantees balance in expectation across infinite replications, but finite samples can still exhibit chance imbalances, or implementation failures (such as caseworkers overriding random assignment). Before estimating treatment effects, applied econometricians systematically verify baseline covariate balance.

Standardised Mean Differences and Omnibus Tests

For each pre-treatment covariate X_{ki}, simple t-tests are sensitive to sample size: in massive administrative trials ($N = 100,000$), trivial differences appear statistically significant, whereas in small trials ($N = 80$), substantive imbalances go undetected. Consequently, best practice reports the **Normalised (Standardised) Difference** (Imbens & Rubin, 2015):

$$\Delta_{\text{std},k} = \frac{\bar{X}_{k,1} - \bar{X}_{k,0}}{\sqrt{\frac{s^2_{k,1} + s^2_{k,0}}{2}}}$$

Where:

- $\bar{X}_{k,1}$ and $\bar{X}_{k,0}$ are the treated and control sample means of baseline covariate k.
- $s^2_{k,1}$ and $s^2_{k,0}$ are the treated and control sample variances.
- A threshold of $|\Delta_{\text{std},k}| < 0.10$ indicates good balance, while $|\Delta_{\text{std},k}| > 0.25$ signals serious imbalance requiring investigation or regression adjustment.

To avoid multiple-testing errors across K covariates, analysts complement pairwise standardised differences with a single **omnibus joint F-test** from regressing the treatment indicator D_i on all baseline covariates $X_{1i}, \ldots, X_{Ki}$ (with heteroskedasticity- or cluster-robust errors), testing $H_0 : \gamma_1 = \cdots = \gamma_K = 0$.

Threats to Internal Validity in Field Experiments

Even with flawless baseline randomisation, four post-randomisation threats can compromise causal interpretation:

Threat	Mechanism	Econometric Consequence	Diagnostic / Mitigation
Imperfect Compliance	Control units access treatment (always-takers) or treated units refuse (never-takers)	$\bar{Y}_1 - \bar{Y}_0$ estimates Intention-to-Treat (ITT), not Treatment-on-the-Treated (TOT)	Use random assignment Z_i as an instrumental variable for actual take-up D_i (Module 3)
Differential Attrition	Follow-up survey response rates differ systematically between treated and control arms	Reintroduces selection bias into observed post-treatment outcomes	Test attrition rate parity; compute Manski or Lee (2009) trimming bounds

DegreeLive

Threat	Mechanism	Econometric Consequence	Diagnostic / Mitigation
Spillovers (SUTVA failure)	Treated workers share information or compete against control workers in the same market	Biases difference-in-means toward zero (information) or exaggerates gap (displacement)	Two-stage cluster saturation designs (Crepon et al., 2013)
Hawthorne / John Henry Effects	Subjects alter behaviour simply from knowing they are observed or competing	Conflates experimental scrutiny with the intrinsic policy mechanism	Natural field experiments with unobtrusive administrative outcome tracking (ATO/Medicare)

Lee (2009) Attrition Bounds

Suppose follow-up response probability is higher in the treatment arm ($p_1 = 0.88$) than in the control arm ($p_0 = 0.80$), creating differential attrition of 8 percentage points. Lee (2009) derives sharp non-parametric bounds on the treatment effect for the "always-observed" subpopulation by assuming monotonicity (treatment affects attrition in only one direction). The excess treated observations, fraction $q = (p_1 - p_0)/p_1$, are trimmed first from the bottom tail of the treated outcome distribution (yielding the upper bound $\hat{\tau}_U$) and then from the top tail (yielding the lower bound $\hat{\tau}_L$).

1.8 Heterogeneous Treatment Effects and Multiple Hypothesis Testing

Average treatment effects frequently mask critical distributional variation. A job-search coaching intervention might raise earnings substantially for long-term unemployed youth while having zero impact on mature workers with recent industry experience. Estimating heterogeneous treatment effects (HTEs) moves beyond ATE to $\text{CATE}(x) = E[Y_i(1) - Y_i(0) \mid X_i = x]$.

Linear Interaction Models for CATE

When pre-specified subgroups $G_i \in \{0, 1\}$ (such as regional vs. metropolitan residency) are of primary policy interest, heterogeneity is estimated via an interaction regression:

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2 G_i + \beta_3 (D_i \times G_i) + u_i$$

Where:

- β_1 is the CATE for the reference subgroup ($G_i = 0$).
- $\beta_1 + \beta_3$ is the CATE for the interacting subgroup ($G_i = 1$).
- β_3 directly measures the treatment effect heterogeneity across subgroups, tested via $H_0 : \beta_3 = 0$.

Data-Driven Discovery: Causal Trees and Honest Estimation

When researchers search across dozens of baseline covariates for unexpected treatment heterogeneity, standard in-sample searching overfits noise. Athey and Imbens (2016) introduced **Causal Trees** with **honest estimation**, which splits the RCT dataset into two disjoint subsamples:

1. **Splitting Sample (I_{split}):** Used exclusively to partition the covariate space into leaves $\ell(x)$ that maximise heterogeneity in estimated treatment effects while penalising within-leaf variance.
2. **Estimation Sample (I_{est}):** Used exclusively to compute the leaf-specific treatment effect $\widehat{\text{CATE}}(x) = \bar{Y}_{1,\ell(x)} - \bar{Y}_{0,\ell(x)}$ and valid standard errors without post-selection bias.

Extending this idea to ensembles produces **Causal Forests** (Wager & Athey, 2018), which estimate smooth individualised CATE functions $\hat{\tau}(X_i)$ with valid asymptotic confidence intervals.

DegreeLive

Multiple Hypothesis Testing Corrections

Testing treatment effects across M outcomes or M subgroups inflates the probability of falsely rejecting at least one true null hypothesis (the Family-Wise Error Rate, $\text{FWER} = 1 - (1-\alpha)^M$; for $M = 10$ independent tests at $\alpha = 0.05$, $\text{FWER} \approx 40.1\%$). Econometricians control error rates using two main procedures:

- **Bonferroni / Holm-Bonferroni (FWER Control):** Bonferroni rejects null j only if $p_j \leq \alpha/M$. Holm's step-down procedure orders p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(M)}$ and compares $p_{(k)}$ against $\alpha/(M-k+1)$, strictly dominating Bonferroni in statistical power while preserving exact $\text{FWER} \leq \alpha$.
- **Benjamini-Hochberg (False Discovery Rate, FDR Control):** Rather than guarding against a single false positive, BH controls the expected proportion of false rejections among all rejected hypotheses at level q. Ordering p-values $p_{(1)} \leq \cdots \leq p_{(M)}$, find the largest index k^* such that $p_{(k^*)} \leq \frac{k^*}{M} q$, and reject all hypotheses $1, \ldots, k^*$.

Worked Numerical Example 1.3: Benjamini-Hochberg FDR vs. Holm-Bonferroni Adjustment

An Australian RCT tests a household energy-tariff rebate across $M = 5$ independent demographic subgroups at target error threshold $\alpha = q = 0.05$, obtaining raw unadjusted p-values: $p_A = 0.008$, $p_B = 0.019$, $p_C = 0.032$, $p_D = 0.044$, and $p_E = 0.210$.

Rank (k)	Subgroup	Raw $p_{(k)}$	Holm Threshold $\frac{0.05}{5-k+1}$	BH Threshold $\frac{k}{5}(0.05)$	Holm Decision	BH Decision
1	Subgroup A	0.008	0.0100	0.0100	Reject H_0	Reject H_0
2	Subgroup B	0.019	0.0125	0.0200	Fail to reject (stops)	Reject H_0
3	Subgroup C	0.032	0.0167	0.0300	Fail to reject	Fail to reject
4	Subgroup D	0.044	0.0250	0.0400	Fail to reject	Fail to reject
5	Subgroup E	0.210	0.0500	0.0500	Fail to reject	Fail to reject

Interpretation: Under naive unadjusted testing ($p < 0.05$), four subgroups (A, B, C, D) appear statistically significant. Under strict FWER control (Holm-Bonferroni), only Subgroup A survives ($0.008 \leq 0.0100$), as Subgroup B fails $0.019 > 0.0125$. Under False Discovery Rate control (Benjamini-Hochberg at $q = 0.05$), the largest k^* satisfying $p_{(k)} \leq 0.01k$ is $k^* = 2$ ($0.019 \leq 0.0200$), so both Subgroups A and B are declared discoveries while Subgroups C and D are screened out as likely false positives.

1.9 Statistical Power, Minimum Detectable Effect, and Sample Size Planning

Before launching a field experiment, applied econometricians conduct prospective power calculations to ensure the sample size N is large enough to detect policy-relevant effects while avoiding wasteful over-recruitment. Under a two-sided test of $H_0 : \tau = 0$ at significance level α (typically 0.05) and target statistical power $1-\kappa$ (typically 0.80), the **Minimum Detectable Effect (MDE)** is defined as the smallest true ATE detectable with probability $1-\kappa$:

$$\text{MDE} = (t_{1-\alpha/2} + t_{1-\kappa}) \sqrt{\frac{\sigma^2}{p(1-p)N}}$$

Where:

- $t_{1-\alpha/2}$ is the critical value for the Type I error threshold (approximately 1.96 for $\alpha = 0.05$ in large samples).
- $t_{1-\kappa}$ is the quantile corresponding to target power (approximately 0.84 for 80% power, or 1.28 for 90% power).

- σ^2 is the residual outcome variance (which shrinks to $\sigma^2(1 - R^2)$ when baseline covariates explain fraction R^2 of outcome variation).
- $p = N_1/N$ is the share of units assigned to treatment (maximising power at balanced allocation $p = 0.5$, where $p(1 - p) = 0.25$).
- N is the total sample size across both arms.

For a Cluster-Randomised Trial with C clusters of equal size m and intracluster correlation ρ, the MDE incorporates the Design Effect $\text{DEFF} = 1 + (m - 1)\rho$:

$$\text{MDE}_{\text{CRT}} = (t_{1-\alpha/2} + t_{1-\kappa}) \sqrt{\frac{\sigma^2[1 + (m - 1)\rho]}{p(1 - p)Cm}}$$

Where:

- C is the total number of clusters (e.g., schools, clinics, or LGAs).
- m is the number of surveyed individuals per cluster.
- Because ρ multiplies $(m - 1)$, adding more individuals m within fixed clusters C yields rapidly diminishing power gains compared to recruiting additional clusters C.

Worked Numerical Example 1.4: Powering an Australian Financial Counselling Trial

Suppose ASIC and Financial Counselling Australia plan an individual-level RCT to test whether proactive SMS budgeting reminders reduce annual payday-loan borrowing among low-income households. Prior ABS data indicate baseline annual payday borrowing has standard deviation $\sigma = \text{A\$900}$, and baseline credit-score covariates predict $R^2 = 0.20$ of outcome variance. Treatment allocation is balanced ($p = 0.50$), with $\alpha = 0.05$ ($z_{0.975} = 1.96$) and 80% power ($z_{0.80} = 0.84$).

Step 1 — Effective Residual Variance with Covariate Adjustment:

$$\sigma^2_{\text{adj}} = \sigma^2(1 - R^2) = 900^2(1 - 0.20) = 810,000 \times 0.80 = 648,000$$

Where:

- σ^2_{adj} is the conditional residual variance after absorbing baseline covariates.

Step 2 — Solve for Required Sample Size N to Detect $\text{MDE} = \text{A\$120}$:

$$N = \frac{(z_{1-\alpha/2} + z_{1-\kappa})^2 \sigma^2_{\text{adj}}}{p(1 - p) \cdot \text{MDE}^2} = \frac{(1.96 + 0.84)^2 \times 648,000}{0.25 \times 120^2} = \frac{7.84 \times 648,000}{3,600} = \frac{5,080,320}{3,600} = 1,411.2$$

Where:

- Rounding up to an even integer yields $N = 1,412$ households (706 treated, 706 control). Without covariate adjustment ($R^2 = 0$), the trial would require $1,411.2/0.80 = 1,764$ households—a 25% larger recruitment budget.

Degree Live

1.10 Varying Propensity Scores: Inverse Probability Weighting (IPW)

In multi-site Australian experiments or stratified trials where assignment probabilities $e(X_i) = P(D_i = 1 \mid X_i)$ vary deliberately across strata (for example, oversampling regional remote communities into treatment at $p = 0.70$ while urban centres receive $p = 0.30$), pooling raw treated and control means estimates a biased mixture. Under unconfoundedness $(Y_i(1), Y_i(0)) \perp D_i \mid X_i$ and strict overlap $0 < e(X_i) < 1$, the **Horvitz-Thompson Inverse Probability Weighting (IPW)** estimator recovers the true population ATE:

$$\hat{\tau}_{\text{IPW}} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{D_i Y_i}{e(X_i)} - \frac{(1 - D_i) Y_i}{1 - e(X_i)} \right)$$

Where:

- $e(X_i) = P(D_i = 1 \mid X_i)$ is the known (or estimated) propensity score for unit i.
- Each treated unit receives weight $1/e(X_i)$ and each control unit receives weight $1/(1 - e(X_i))$, re-balancing the treated and control arms to represent the common target population.

Because extreme weights near $e(X_i) \approx 0$ or $e(X_i) \approx 1$ inflate finite-sample variance, applied researchers typically normalise the weights so they sum to 1 within each arm (the **Hajek IPW estimator**) or combine IPW with outcome regression to form the **Augmented Inverse Probability Weighting (AIPW)** doubly robust estimator:

$$\hat{\tau}_{\text{AIPW}} = \frac{1}{N} \sum_{i=1}^{N} \left[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i) + \frac{D_i (Y_i - \hat{\mu}_1(X_i))}{e(X_i)} - \frac{(1 - D_i)(Y_i - \hat{\mu}_0(X_i))}{1 - e(X_i)} \right]$$

Where:

- $\hat{\mu}_1(X_i)$ and $\hat{\mu}_0(X_i)$ are regression-predicted conditional outcomes under treatment and control.
- **Double Robustness:** $\hat{\tau}_{\text{AIPW}}$ remains consistent if *either* the outcome model $\hat{\mu}_d(X_i)$ *or* the propensity score model $e(X_i)$ is correctly specified.

Worked Numerical Example 1.5: Exact Randomisation Inference (Permutation p-Value)

Consider a pilot trial of an intensive mentoring grant across $N = 6$ regional Victorian small businesses, where $N_1 = 3$ firms are randomly allocated grants ($D_i = 1$) and $N_0 = 3$ serve as controls ($D_i = 0$). Observed annual revenue growth (in A$1,000s) is $\{18, 22, 14\}$ for the three treated firms and $\{8, 10, 12\}$ for the three control firms.

Step 1 — Compute the Observed Test Statistic:

$$T_{\text{obs}} = \bar{Y}_1 - \bar{Y}_0 = \frac{18 + 22 + 14}{3} - \frac{8 + 10 + 12}{3} = 18.0 - 10.0 = 8.0$$

Where:

- $T_{\text{obs}} = 8.0$ (A$8,000) is the observed difference-in-means.

Step 2 — Enumerate the Reference Distribution Under $H_0^{\text{sharp}} : Y_i(1) = Y_i(0)$:

There are $\binom{6}{3} = 20$ equally likely ways to select 3 firms into treatment from the fixed outcome set $\{22, 18, 14, 12, 10, 8\}$ (whose sum is 84, mean 14). Let S_1 be the sum of the 3 firms assigned to treatment under a permutation, so $\bar{Y}_1 - \bar{Y}_0 = \frac{S_1}{3} - \frac{84 - S_1}{3} = \frac{2S_1 - 84}{3}$. To achieve $|T| \geq 8.0$, we require $|2S_1 - 84| \geq 24$, meaning

either $S_1 \geq 54$ or $S_1 \leq 30$.

- Combinations with $S_1 \geq 54$: only $\{22, 18, 14\}$ gives $S_1 = 54$ ($T = +8.0$). Every other triple sums to at most $22 + 18 + 12 = 52$ ($T = +6.67$).
- Symmetric combinations with $S_1 \leq 30$: only $\{12, 10, 8\}$ gives $S_1 = 30$ ($T = -8.0$).

Step 3 — Compute Exact Two-Sided Randomisation p-Value:

$$p_{\text{RI}} = \frac{2}{20} = 0.10$$

Where:

- Even though the treated group earned the three highest revenue gains in the entire sample, the minimum achievable two-sided exact p-value with $\binom{6}{3} = 20$ allocations is $2/20 = 0.10$, illustrating why small pilot trials cannot achieve $p < 0.05$ under exact finite-sample randomisation inference regardless of effect size.

1.11 External Validity and Scaling RCT Evidence to Policy

While randomisation secures internal validity within the experimental sample, policymakers at the Australian Treasury or Productivity Commission ultimately require **external validity**—whether an ATE estimated in a trial generalises to a broader target population or national rollout. Four structural divergences commonly separate experimental results from population-scale outcomes:

- **Site Selection and Volunteer Bias:** Organisations volunteering to host trials (e.g., high-capacity metropolitan schools or progressive employers) often differ systematically from non-participating entities. Reweighting experimental estimates using **transportability weights** $w(X_i) = \frac{P(S_i=0|X_i)}{P(S_i=1|X_i)}$, where $S_i = 1$ denotes trial participation and $S_i = 0$ denotes the target national dataset (such as ABS Census), bridges observable covariate shifts.
- **General Equilibrium and Market Feedback Effects:** Small-scale job-training trials hold market wages and vacancy queues fixed, whereas nationwide scaling shifts aggregate labour supply, alters equilibrium wages, and triggers displacement of untrained job-seekers.
- **Implementation Fidelity ("Voltage Drop"):** Pilot programmes supervised by academic researchers achieve tighter delivery quality than routine administration across hundreds of regional service centres.
- **Time Horizon and Fade-Out:** Short-term 6-month follow-ups may capture temporary salience boosts that fade over 3 years, or conversely miss long-run human capital compounding captured only via linked administrative tax and welfare records.

DegreeLive