DegreeLive

Bachelor of Commerce (B2001) | Level 2 • Part B Major | Monash University

Statistical thinking

Course Code: ETC2420 • Module 4 Notes

Module 4: Experimental Design, Observational Studies, Multiple Testing & Statistical Communication

1. Causal Inference, Potential Outcomes, and Randomised Experiments

A central task of statistical thinking is distinguishing **association** (variables moving together in observed data) from **causation** (how an outcome would change under an active intervention). Predictive models exploit correlations regardless of origin, whereas policy, medical, and commercial decisions—such as changing an Australian superannuation default fee, launching a digital marketing campaign, or adopting a clinical protocol—depend on causal effects.

1.1 The Potential Outcomes Framework and Selection Bias

Under the **Neyman-Rubin potential outcomes framework**, every unit i (customer, patient, firm, or region) in a study with binary treatment indicator $D_i \in \{0, 1\}$ possesses two latent potential outcomes:

- $Y_i(1)$: the outcome unit i would experience if exposed to treatment ($D_i = 1$).
- $Y_i(0)$: the outcome unit i would experience if assigned to control ($D_i = 0$).

The unit-level causal effect is $\tau_i = Y_i(1) - Y_i(0)$, and the population **Average Treatment Effect (ATE)** is $\tau_{\text{ATE}} = \text{E}[Y_i(1) - Y_i(0)]$. The **Fundamental Problem of Causal Inference** is that for each unit i, we observe only the realised outcome $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$; the counterfactual outcome is never observed simultaneously.

When an analyst naively compares average observed outcomes between treated and untreated units in non-randomised data, the difference decomposes into the true causal effect on the treated plus **selection bias**:

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

Where: ATT is the Average Treatment Effect on the Treated; $\text{E}[Y_i(0) \mid D_i = 1]$ is the unobservable baseline outcome treated units would have experienced had they not been treated; and selection bias arises whenever treated and control groups differ systematically at baseline (for example, healthier individuals electing a voluntary wellness program or higher-spending customers opting into a loyalty tier).

1.2 Randomisation, Blocking, Cluster Trials, and Factorial Designs

Randomised Controlled Trials (RCTs) and online **A/B tests** solve selection bias by assigning D_i via an exogenous random mechanism (a fair coin flip or pseudo-random hash) independent of all baseline characteristics and potential outcomes: $(Y_i(1), Y_i(0)) \perp D_i$. Independence guarantees $\text{E}[Y_i(0) \mid D_i = 1] = \text{E}[Y_i(0) \mid D_i = 0] = \text{E}[Y_i(0)]$ and $\text{E}[Y_i(1) \mid D_i = 1] = \text{E}[Y_i(1)]$, eliminating selection bias in expectation and making the simple difference in sample means $\bar{Y}_1 - \bar{Y}_0$ an unbiased estimator of the ATE.

Well-designed experiments often go beyond simple complete randomisation:

- **Blocking (Stratified Randomisation):** When an observable baseline covariate (such as Australian state, store size, or customer tenure) strongly predicts the outcome, units are grouped into homogeneous *blocks* and randomised separately within each block. Blocking eliminates between-block variation from the treatment comparison, substantially reducing standard errors and increasing statistical power without bias.
- **Cluster Randomised Trials:** When treatment must be delivered at a group level (such as schools, clinics, or Australian postcodes) to prevent contamination, entire clusters of average size m are randomised. Because observations within the same cluster correlate positively with intraclass correlation ρ_{ICC}, the variance of cluster means inflates by the **Design Effect** $\text{DEFF} = 1 + (m - 1)\rho_{\text{ICC}}$. For instance, randomising 20 schools of $m = 30$ students each ($N = 600$ total) with $\rho_{\text{ICC}} = 0.05$ yields $\text{DEFF} = 1 + 29(0.05) = 2.45$, meaning the effective sample size is only $600/2.45 \approx 245$ independent individuals.
- **Factorial Designs:** Instead of testing one factor at a time, a 2×2 factorial experiment randomises units across all four combinations of two interventions (Factor A and Factor B), estimating both *main effects* with full sample efficiency and testing whether an *interaction effect* exists.

Worked Example 1.1: Analysing a 2×2 Factorial Pricing and Reminder Experiment

A Sydney subscription publisher randomises $4{,}000$ lapsed subscribers equally ($1{,}000$ per cell) across two factors: Factor A (Price Discount: $A = -1$ None vs. $A = +1$ 20% Off) and Factor B (Reminder Channel: $B = -1$ Email Only vs. $B = +1$ Email + SMS). Observed renewal rates are:

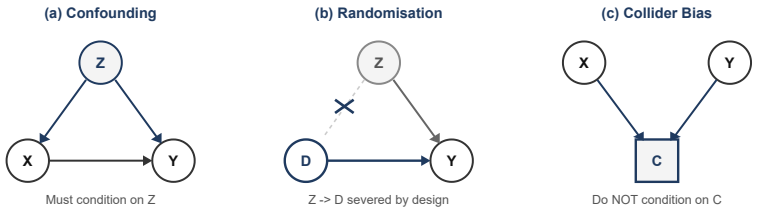
Cell Combination	Factor A (Discount)	Factor B (SMS)	Renewal Rate $\bar{Y}$
Control	-1 (None)	-1 (No SMS)	8.0%
SMS Only	-1 (None)	$+1$ (SMS)	10.0%
Discount Only	$+1$ (20% Off)	-1 (No SMS)	14.0%
Discount + SMS	$+1$ (20% Off)	$+1$ (SMS)	20.0%

DegreeLive

Step 1: Compute the **Main Effect of A** (average effect of the discount across both SMS levels): $\frac{1}{2}[(14.0\% - 8.0\%) + (20.0\% - 10.0\%)] = \frac{1}{2}[6.0\% + 10.0\%] = +8.0\%$.

Step 2: Compute the **Main Effect of B** (average effect of SMS across both discount levels): $\frac{1}{2}[(10.0\% - 8.0\%) + (20.0\% - 14.0\%)] = \frac{1}{2}[2.0\% + 6.0\%] = +4.0\%$.

Step 3: Compute the **Interaction Effect ($A \times B$)** (half the difference between the discount effect with SMS and without SMS): $\frac{1}{2}[(20.0\% - 10.0\%) - (14.0\% - 8.0\%)] = \frac{1}{2}[10.0\% - 6.0\%] = +2.0\%$ (an extra $+4.0\%$ synergy when combined, demonstrating a positive quantitative interaction that separate one-factor-at-a-time experiments would miss completely).



2. Sample Size Determination, Statistical Power, and A/B Testing Practice

Running an underpowered experiment wastes resources and produces erratic, unreplicable estimates; running an unnecessarily huge experiment wastes time and exposes users to suboptimal variants. Before collecting data, analysts determine required sample sizes through formal **power analysis**.

2.1 Type I Error, Type II Error, Power, and Sample Size Formulas

In a two-sided hypothesis test of $H_0 : \mu_1 - \mu_0 = 0$ versus $H_1 : \mu_1 - \mu_0 = \delta \neq 0$ at significance level α:

- **Type I error (α):** Rejecting H_0 when H_0 is actually true (false positive).
- **Type II error (β):** Failing to reject H_0 when a genuine effect of size δ exists (false negative).
- **Statistical Power ($1 - \beta$):** The probability of correctly rejecting H_0 when the true effect equals the **Minimum Detectable Effect (MDE)** δ, typically targeted at 80% or 90%.

For a balanced two-arm experiment assigning n subjects to Treatment and n subjects to Control (total $2n$) with common outcome variance σ^2, setting the rejection threshold $z_{1-\alpha/2}\sqrt{2\sigma^2/n}$ equal to $\delta - z_{1-\beta}\sqrt{2\sigma^2/n}$ yields the per-group sample size formula:

$$n = \frac{2\,(z_{1-\alpha/2} + z_{1-\beta})^2\,\sigma^2}{\delta^2}$$

Where: n is the required sample size per arm; $z_{1-\alpha/2}$ is the standard Normal critical value (e.g., 1.96 for $\alpha = 0.05$); $z_{1-\beta}$ is the power quantile (e.g., 0.8416 for 80% power, 1.2816 for 90% power); σ^2 is the within-group outcome variance; and $\delta = |\mu_1 - \mu_0|$ is the Minimum Detectable Effect. Notice the inverse-square law: halving the target MDE δ quadruples the required sample size n.

When the outcome is a binary conversion indicator with baseline rate p_0 and target rate $p_1 = p_0 + \delta$, substituting the Bernoulli variances yields the approximate per-arm sample size for comparing two proportions:

$$n = \frac{\left(z_{1-\alpha/2}\sqrt{2\bar{p}(1-\bar{p})} + z_{1-\beta}\sqrt{p_0(1-p_0) + p_1(1-p_1)}\right)^2}{(p_1 - p_0)^2}, \qquad \bar{p} = \frac{p_0 + p_1}{2}$$

Where: p_0 is the baseline conversion probability; p_1 is the alternative conversion probability; and $\bar{p}$ is the pooled proportion under H_0.

Worked Example 2.1: Sizing an A/B Test for an Australian Fintech App

A Sydney fintech startup plans an A/B test on its mobile onboarding screen. Historical analytics show a baseline onboarding completion rate of $p_0 = 0.12$. Product management specifies an MDE of $\delta = 0.03$ (detecting an uplift to $p_1 = 0.15$) with two-sided $\alpha = 0.05$ ($z_{0.975} = 1.96$) and 80% power ($z_{0.80} = 0.8416$).

Step 1: Compute the pooled proportion $\bar{p}$ and variance terms:

$$\bar{p} = \frac{0.12 + 0.15}{2} = 0.135, \qquad 2\bar{p}(1-\bar{p}) = 2(0.135)(0.865) = 0.23355$$

Degree Live

$$p_0(1-p_0)+p_1(1-p_1) = 0.12(0.88)+0.15(0.85) = 0.1056+0.1275 = 0.2331$$

Where: $2\bar{p}(1-\bar{p})$ governs the null standard error; and $p_0(1-p_0)+p_1(1-p_1)$ governs the alternative standard error.

Step 2: Substitute into the per-arm sample size formula:

$$n = \frac{(1.96\ \cancel{0.23355} + 0.8416\ \cancel{0.2331})^2}{(0.03)^2} = \frac{(1.96(0.48327)+0.8416(0.48280))^2}{0.0009} = \frac{(0.94721+0.40633)^2}{0.0009} \approx \frac{1.8321}{0.0009} \approx 2{,}036$$

Where: $n \approx 2{,}036$ users per arm, requiring $4{,}072$ total users across Control and Treatment. Using the simpler Rule-of-Thumb approximation $n \approx 16\bar{p}(1-\bar{p})/\delta^2 = 16(0.1168)/0.0009 \approx 2{,}076$ gives virtually the identical planning figure.

2.2 Variance Reduction via CUPED and Practical A/B Testing Pitfalls

When pre-experiment historical outcomes X_i (such as prior 30-day customer spend before randomisation) correlate strongly with the experimental outcome Y_i, modern digital experimentation teams accelerate tests using **CUPED (Controlled-experiment Using Pre-Experiment Data)**. Because randomisation ensures $\mathrm{E}[X_i \mid D_i = 1] = \mathrm{E}[X_i \mid D_i = 0]$, subtracting a linear covariate adjustment leaves the ATE estimator unbiased while slashing variance:

$$\hat{\tau}_{\mathrm{CUPED}} = (\bar{Y}_1 - \bar{Y}_0) - \hat{\theta}(\bar{X}_1 - \bar{X}_0), \qquad \hat{\theta} = \frac{\mathrm{Cov}(Y,X)}{\mathrm{Var}(X)}, \qquad \mathrm{Var}(\hat{\tau}_{\mathrm{CUPED}}) = (1-\rho_{X,Y}^2)\,\mathrm{Var}(\bar{Y}_1 - \bar{Y}_0)$$

Where: $\hat{\theta}$ is the pooled OLS slope of post-period Y on pre-period X; and $\rho_{X,Y}$ is the correlation between pre- and post-experiment outcomes. If $\rho_{X,Y} = 0.70$, CUPED reduces variance by $1 - 0.70^2 = 51\%$, cutting the required experiment duration or sample size in half without changing α or power.

Even with ample sample size, operational pitfalls can invalidate A/B tests:

- **Optional Stopping ("Peeking"):** Checking p-values repeatedly as data accrue and stopping immediately upon observing $p < 0.05$ severely inflates Type I error (peeking 10 times at $\alpha = 0.05$ raises cumulative false positives near 19%). Legitimate interim monitoring uses **group sequential alpha-spending** boundaries (e.g., O'Brien-Fleming, which requires very stringent thresholds like $|z| \geq 3.47$ at the first of three looks and $|z| \geq 2.01$ at the final look so cumulative Type I error stays exactly 0.05).
- **SUTVA Violations and Network Interference:** If treated users interact with control users in a shared Australian marketplace (e.g., ride-sharing drivers or accommodation listings), cluster randomisation by suburb or market is required instead of user-level randomisation.
- **Novelty Effects and Sample Ratio Mismatch (SRM):** Short-term curiosity spikes decay over time, while an unexpected deviation from the intended $50/50$ traffic split signals telemetry or redirect bugs that invalidate causal claims.

3. Observational Studies, Confounding, and Simpson's Paradox

When randomisation is unethical, illegal, or infeasible—such as studying the health effects of air pollution, the wage returns to university degrees, or corporate governance reforms—analysts rely on **observational studies** (cross-sectional surveys, longitudinal cohorts, or natural experiments). In observational data, treatment assignment D_i is entangled with covariates Z_i.

3.1 Confounding, Back-Door Paths, and Simpson's Paradox

In Pearl's Directed Acyclic Graph (DAG) formalism, a variable Z is a **confounder** if it lies on an unblocked *back-door path* $X \leftarrow Z \rightarrow Y$ (Figure 1(a)). Failure to condition on Z mixes the causal effect of X with spurious correlation induced by Z. In extreme cases, pooling data across heterogeneous strata produces **Simpson's Paradox**: an association observed within every subgroup reverses direction completely when the subgroups are aggregated.

Worked Example 3.1: Simpson's Paradox in a Regional Employment Program

Suppose an Australian state agency evaluates a voluntary digital-skills job placement program ($D = 1$ Participant vs. $D = 0$ Non-Participant) across 800 jobseekers in two labour markets: Metropolitan Melbourne (high baseline placement rate) and Regional Victoria (tighter local labour market). Overall pooled results show:

- **Program Participants ($D = 1$):** 220 placed out of 400 total (55.0% placement rate).
- **Non-Participants ($D = 0$):** 248 placed out of 400 total (62.0% placement rate).

Naively comparing the pooled rates suggests the program *reduces* job placement by 7.0 percentage points! However, disaggregating by region reveals why:

DegreeLive

Region (Confounder Z)	Participants ($D = 1$)	Non-Participants ($D = 0$)	Within-Stratum Difference
Metropolitan Melbourne	$75/100 = 75.0\%$	$210/300 = 70.0\%$	$+5.0\%$ (Program higher)
Regional Victoria	$145/300 = 48.3\%$	$38/100 = 38.0\%$	$+10.3\%$ (Program higher)
Pooled Total (Unadjusted)	$220/400 = 55.0\%$	$248/400 = 62.0\%$	-7.0% **(Reversal!)**

Why does the reversal occur mathematically? Because 75% of program participants lived in Regional Victoria ($300/400$), where baseline placement rates are lower for everyone, whereas 75% of non-participants lived in Metropolitan Melbourne ($300/400$), the unadjusted comparison mixes a positive within-region program effect ($+5.0\%$ and $+10.3\%$) with severe regional imbalance. Standardising both groups to the overall population distribution (400 Metro and 400 Regional, weights 0.50 each) yields the adjusted ATE estimate $0.50(5.0\%) + 0.50(10.33\%) = +7.67$ percentage points.

3.2 Propensity Scores, IPTW, Collider Bias, and Natural Experiments

When all confounders Z are measured (the *conditional ignorability* assumption, $(Y(1), Y(0)) \perp D \mid Z$), analysts adjust for Z using multivariable regression, matching, or **Inverse Probability of Treatment Weighting (IPTW)** based on the **propensity score** $e(Z) = P(D = 1 \mid Z)$. Treated units receive weight $w_i = 1/e(Z_i)$ and control units receive weight $w_i = 1/(1 - e(Z_i))$, creating a balanced pseudo-population where Z is independent of D. In Worked Example 3.1, Metro participants have $e(\text{Metro}) = 100/400 = 0.25$ (IPTW weight 4), Metro controls have $1 - e = 0.75$ (weight $4/3$), Regional participants have $e(\text{Reg}) = 300/400 = 0.75$ (weight $4/3$), and Regional controls have $1 - e = 0.25$ (weight 4). Weighting each group reproduces the exact standardised placement rates: $0.5(75\%) + 0.5(48.33\%) = 61.67\%$ for participants versus $0.5(70\%) + 0.5(38\%) = 54.00\%$ for controls (ATE $= +7.67\%$).

Conversely, analysts must never condition on a **collider** C (a common effect $X \to C \leftarrow Y$ as in Figure 1(c)) or a post-treatment mediator unless specifically performing mediation analysis; conditioning on a collider opens a spurious non-causal path between X and Y. When unobserved confounders are suspected, researchers turn to **natural experiments**—such as **Difference-in-Differences (DiD)**, which compares the pre-to-post outcome change of a policy-treated Australian state against an untreated control state under the *parallel trends* assumption, or **Regression Discontinuity Designs (RDD)**, which exploit sharp eligibility cutoffs (such as an ATAR threshold or age-based pension cutoff) to compare units just above and just below the threshold.

Worked Example 3.2: Difference-in-Differences Evaluation of a State Energy Rebate

Suppose Victoria introduces an energy-efficiency rebate for small retail businesses in Year 2, while neighbouring New South Wales (NSW) does not. Analysts measure average quarterly electricity bills (in AUD) across 500 Victorian and 500 NSW cafes in Year 1 (Pre-policy) and Year 2 (Post-policy):

State Group	Year 1 (Pre-Policy)	Year 2 (Post-Policy)	Within-State Change (Δ)
Victoria (Treated)	\$4,200	\$3,950	$-\$250$
New South Wales (Control)	\$3,800	\$3,900	$+\$100$

Step 1: Note that comparing Year 2 alone ($\$3{,}950 - \$3{,}900 = +\$50$) falsely suggests Victorian bills are higher, because Victoria had higher baseline electricity costs in Year 1 ($\$4{,}200$ vs. $\$3{,}800$).

Step 2: Compute the Difference-in-Differences estimator:

$$\hat{\tau}_{\text{DiD}} = (\bar{Y}_{\text{VIC, Post}} - \bar{Y}_{\text{VIC, Pre}}) - (\bar{Y}_{\text{NSW, Post}} - \bar{Y}_{\text{NSW, Pre}}) = (-\$250) - (+\$100) = -\$350$$

Where: subtracting NSW's $+\$100$ common macroeconomic tariff rise from Victoria's $-\$250$ change estimates a causal quarterly reduction of $\$350$ per business—valid provided Victoria and NSW would have followed parallel cost trajectories absent the rebate.

4. Multiple Hypothesis Testing, FWER, and False Discovery Rate (FDR)

Modern empirical analysts routinely evaluate dozens or thousands of hypotheses simultaneously: screening 20 regional sub-segments in an A/B test, testing 100 ASX factor anomalies, or scanning $10{,}000$ genomic markers. Applying a conventional $\alpha = 0.05$ threshold independently to every test guarantees an epidemic of false discoveries.

4.1 Error Inflation and Family-Wise Error Rate (FWER)

Suppose all m null hypotheses $H_{0,1}, \ldots, H_{0,m}$ are genuinely true (the global null) and test statistics are independent. The probability of committing at least one Type I error across the entire family—the **Family-Wise Error Rate (FWER)**—escalates rapidly with m:

$$\text{FWER} = P(\text{at least one false rejection}) = 1 - (1 - \alpha)^m$$

Where: α is the per-test significance level; and m is the total number of tests performed. At $\alpha = 0.05$: for $m = 10$ tests, FWER $= 1 - 0.95^{10} = 40.1\%$; for $m = 20$ tests, FWER $= 64.2\%$; and for $m = 100$ tests, FWER $= 99.4\%$.

4.2 Controlling FWER: Bonferroni and Holm-Bonferroni

- **Bonferroni Correction:** Reject H_{0_j} only if $p_j \leq \alpha/m$ (equivalently, adjusted p-value $p_j^{\text{Bonf}} = \min(mp_j, 1) \leq \alpha$). By Boole's inequality, FWER $\leq \alpha$ under any dependence structure, though it sacrifices considerable power when m is large.
- **Holm-Bonferroni Step-Down Procedure:** Uniformly more powerful than Bonferroni while maintaining identical FWER $\leq \alpha$. Sort the p-values $p_{(1)} \leq \cdots \leq p_{(m)}$. Starting from $k = 1$, compare $p_{(k)}$ against $\frac{\alpha}{m-k+1}$. Reject while $p_{(k)} \leq \frac{\alpha}{m-k+1}$, and **stop** at the first index k^* where $p_{(k^*)} > \frac{\alpha}{m-k^*+1}$. Its step-down adjusted p-values are $p_{(k)}^{\text{Holm}} = \max_{j \leq k} \min((m-j+1)p_{(j)},\ 1)$.

4.3 Controlling the False Discovery Rate: Benjamini-Hochberg (BH)

In exploratory screening across many tests, guaranteeing zero false positives anywhere (FWER) is often overly restrictive. Benjamini and Hochberg (1995) proposed controlling the **False Discovery Rate (FDR)**—the expected proportion of false discoveries among all rejected hypotheses, FDR $= \mathrm{E}[V / \max(R, 1)]$, where V is the number of false rejections and R is total rejections. To control FDR $\leq q$, the **Benjamini-Hochberg step-up procedure** sorts $p_{(1)} \leq \cdots \leq p_{(m)}$, finds the **largest** rank k^* satisfying:

$$p_{(k)} \leq \frac{k}{m} q$$

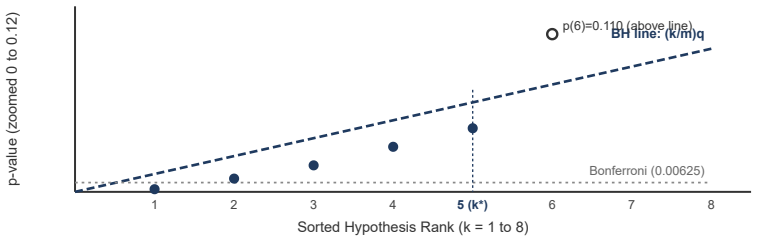
Where: $k \in \{1, \ldots, m\}$ is the sorted rank; m is the total number of hypotheses tested; and q is the target FDR level. All hypotheses $H_{0,(1)}, \ldots, H_{0,(k^*)}$ are rejected. Equivalently, BH-adjusted p-values (q-**values**) are computed backward from $k = m$ down to 1 as $q_{(k)} = \min_{j \geq k} \min(\frac{m}{j} p_{(j)},\ 1)$.

Worked Example 4.1: Comparing Bonferroni, Holm, and BH on Eight Economic Indicators

An analyst tests whether $m = 8$ distinct retail sub-sectors experienced abnormal sales growth after an RBA rate decision, observing sorted p-values: 0.002, 0.009, 0.018, 0.031, 0.044, 0.110, 0.340, 0.720. Compare Bonferroni ($\alpha = 0.05$), Holm-Bonferroni ($\alpha = 0.05$), and Benjamini-Hochberg ($q = 0.10$) along with exact BH q-values:

Rank k	$p_{(k)}$	Bonferroni ($\frac{0.05}{8}$)	Holm ($\frac{0.05}{9-k}$)	BH Cutoff ($\frac{k}{8} \times 0.10$)	BH q-value
1	0.002	0.00625 (Reject)	0.00625 (Reject)	0.0125 (≤)	0.0160 (Reject)
2	0.009	0.00625 (Retain)	0.00714 (Retain — Stop)	0.0250 (≤)	0.0360 (Reject)
3	0.018	0.00625 (Retain)	0.00833 (Retain)	0.0375 (≤)	0.0480 (Reject)
4	0.031	0.00625 (Retain)	0.01000 (Retain)	0.0500 (≤)	0.0620 (Reject)
5	0.044	0.00625 (Retain)	0.01250 (Retain)	0.0625 (≤)	0.0704 (Reject)
6	0.110	0.00625 (Retain)	0.01667 (Retain)	0.0750 (>)	0.1467 (Retain)
7	0.340	0.00625 (Retain)	0.02500 (Retain)	0.0875 (>)	0.3886 (Retain)
8	0.720	0.00625 (Retain)	0.05000 (Retain)	0.1000 (>)	0.7200 (Retain)

Result summary: Bonferroni and Holm-Bonferroni control FWER at 0.05 and reject only rank 1 ($p_{(1)} = 0.002$). Benjamini-Hochberg at $q = 0.10$ finds largest rank $k^* = 5$ with $p_{(5)} = 0.044 \leq 0.0625$ (and $q_{(5)} = \frac{8}{5}(0.044) = 0.0704 \leq 0.10$), rejecting ranks 1 through 5 while bounding the expected false-discovery share among those 5 flagged sectors below 10%.



5. Questionable Research Practices, Reproducibility, and Statistical Communication

Even mathematically valid inference procedures collapse when analysts exploit undisclosed analytical flexibility or miscommunicate uncertainty to decision-makers.

DegreeLive

5.1 p-Hacking, HARKing, Winner's Curse, and Post-Selection Inference

***p*-Hacking** (data dredging) encompasses post-hoc analytical choices motivated by achieving $p < 0.05$: trying multiple covariate sets, selectively dropping "outliers" after inspecting significance, switching between mean and median tests, splitting or merging categories, or peeking at accruing data. Even without conscious dishonesty, Gelman and Loken's **garden of forking paths** warns that choosing which test to run *after* observing the data makes the nominal *p*-value invalid because it ignores all alternative tests that would have been chosen under other data realisations.

Closely related distortions include **HARKing** (Hypothesising After the Results are Known—presenting exploratory post-hoc discoveries as if they were pre-specified confirmatory tests) and **publication bias** (the file-drawer problem, where journals and corporate teams publish statistically significant findings while burying null results). When an analyst screens hundreds of noisy predictors and reports only the highest-ranked significant estimate, that estimate suffers from the **winner's curse** (systematic upward bias in magnitude because only positive random noise spikes clear the selection threshold). One clean remedy is **sample splitting**: using 50% of the data for exploratory screening and hypothesis generation, and reserving the untouched 50% validation holdout for unbiased confirmatory estimation and *p*-value calculation.

5.2 Interpreting p-Values and Reproducible Workflows

The American Statistical Association's formal statement on *p*-values emphasises six core principles for sound practice:

1. A *p*-value can indicate how incompatible data are with a specified statistical model, not prove that a theory is true or false.
2. A *p*-value does **not** measure the probability that the null hypothesis is true, $P(H_0 \mid \text{data})$, nor the probability that data were produced by chance alone.
3. Scientific and business decisions should never rest solely on whether a *p*-value crosses an arbitrary bright-line threshold such as 0.05.
4. Proper inference requires full transparency: reporting all hypotheses tested, exclusions made, and analyses conducted.
5. **Statistical significance is not practical significance:** With $n = 500{,}000$ observations, a trivial 0.05% click-rate difference can yield $p < 0.001$ while being economically worthless; conversely, a wide 95% confidence interval spanning a large potential harm and benefit warrants caution even if $p = 0.08$.
6. By itself, a *p*-value does not measure the size of an effect or the importance of a result; always report point estimates alongside confidence or credible intervals.

5.3 Communicating Statistical Uncertainty to Decision-Makers

When presenting findings to Australian boards, regulators, or clinical panels, technical accuracy must be paired with clarity without exaggerating certainty. Consider how framing alters perception: stating that a preventative program "cuts adverse events by 50%" (relative risk reduction) sounds dramatic even when the underlying risk drops merely from 2 per $1{,}000$ to 1 per $1{,}000$—an **Absolute Risk Reduction (ARR)** of 0.001 and a **Number Needed to Treat (NNT)** of $1/0.001 = 1{,}000$ participants per prevented event. Expressing probabilities as **natural frequencies** alongside absolute risk changes and interval bounds prevents framing distortion.

Reporting Dimension	Flawed Practice	Best-Practice Standard
Design & Pre-Registration	Post-hoc hypothesis selection (HARKing)	Pre-register primary outcomes, sample size, and stopping rules
Effect Magnitude	Reporting only "$p < 0.05$" or relative %	Report point estimate, 95% CI/credible interval, and absolute effect
Multiple Comparisons	Cherry-picking significant subgroups	Report total m tests and apply Holm-Bonferroni or BH FDR control
Missing Data & Outliers	Silently deleting inconvenient records	Document exclusion rules pre-analysis and show sensitivity checks
Computational Reproducibility	Manual spreadsheet edits without logs	Scripted pipeline under version control with fixed random seed

Worked Example 5.1: Rewriting Technical Results for Executive Decision-Making

Consider a flawed analyst summary from an Australian bank's credit-retention pilot: *"Proactive outreach significantly reduced default ($p = 0.032 < 0.05$), cutting default risk by 25%."* Although statistically accurate in narrow terms, this summary omits baseline rates, interval uncertainty, sample scope, and cost-benefit balance.

A decision-grade statistical communication reframes the identical empirical evidence transparently: *"In a randomised trial of $4{,}000$ at-risk mortgage accounts over 6 months, 16 of $2{,}000$ outreach accounts defaulted (0.80%) compared with 24 of $2{,}000$ control accounts (1.20%). This represents an absolute reduction of 0.40 percentage points (95% CI: -0.04 to $+0.84$ points; $p = 0.032$ one-sided / 0.064 two-sided, NNT ≈ 250 calls per prevented default). Because calling 250 accounts costs $\$3{,}750$ in staff time against an average default loss of $\$22{,}000$, the net expected return is positive, though uncertainty spans a near-zero benefit; we recommend scaling to the highest-risk decile first while continuing monitoring."* This formulation equips executives with natural frequencies, absolute impacts, precision bounds, and economic decision criteria.

6. Advanced Experimental Design, Variance Reduction (CUPED) & Sequential Testing

Modern digital and financial enterprises in Australia conduct thousands of controlled experiments annually to optimise credit underwriting interfaces, mobile banking navigation, and consumer marketing campaigns. However, standard randomised controlled trials (A/B tests) face significant operational constraints: large sample size requirements when detecting subtle conversion or revenue increments, lengthy testing

windows exposing experiments to seasonality and macroeconomic shocks, and managerial pressures to peek at accumulating results before nominal sample sizes are reached. Advanced statistical thinking addresses these friction points through pre-experimental covariate adjustment (CUPED) and sequential testing methodologies that protect nominal Type I error rates.

Continuous Monitoring, Peeking Bias, and Alpha Spending

In standard fixed-sample testing, hypotheses are evaluated exactly once after collecting a predetermined sample size N. In commercial practice, decision-makers frequently inspect live dashboards daily. If an analyst declares significance as soon as the observed p-value dips below $\alpha = 0.05$, the true false positive rate inflates rapidly across repeated inspections due to random walk properties of test statistics:

$$alpha_{textinflated} = 1 - (1 - alpha)^{K} quad textunderindependence, orapproximately 30% textto 40% textacrosscontinuousmonitoring$$

Where:

$\alpha_{textinflated}$ is the empirical Type I error across repeated interim analyses;

α is the nominal significance level per individual test (e.g., 0.05);

K is the number of interim evaluations.

To control false positives under continuous monitoring, practitioners deploy **Alpha Spending Functions** (e.g., Lan-DeMets with O'Brien-Fleming or Pocock boundaries) or the **Sequential Probability Ratio Test (SPRT)**. An alpha spending function $\alpha(t)$ allocates portions of the total error budget α as information fraction $t = n/N$ progresses from 0 to 1:

$$alpha_{textOBF}(t) = 2 - 2 Phi left(fracZ_{alpha/2}sqrtt right)$$

Where:

$\alpha_{textOBF}(t)$ is the cumulative Type I error spent by information fraction t;

$\Phi($ $cdot)$ is the standard normal cumulative distribution function;

$Z_{alpha/2}$ is the standard normal critical value (1.96 for two-sided 5% level);

t is the fraction of total planned sample size collected to date ($0 < t$ $le 1$).

Under O'Brien-Fleming boundaries, critical values are extremely conservative in early stages (e.g., requiring $Z > 4.0$ at $t = 0.25$), preserving the vast majority of the alpha budget for the terminal evaluation while permitting early stopping only for overwhelming treatment effects.

Controlled-Experiment Using Pre-Experiment Data (CUPED)

Controlled experiments evaluating continuous commercial metrics (such as customer transaction spend, loan origination balance, or website revenue) frequently suffer from high baseline variance: a handful of high-net-worth customers or frequent purchasers generate immense dispersion that inflates the standard error of the mean difference. **CUPED (Controlled-Experiment Using Pre-Experiment Data)** eliminates pre-existing individual variation by utilising pre-experiment covariate data X measured prior to randomisation.

Let Y denote the target metric measured during the experiment window, and let X denote the identical metric (or a highly correlated metric) measured for the same user during the pre-experiment period. The variance-reduced CUPED metric $\widetilde{Y}$ is defined as:

DegreeLive

$$tildeY_i = Y_i - theta(X_i - mu_X)$$

Where:

$\widetilde{Y}_i$ is the adjusted target metric for individual user i;

Y_i is the observed metric during the experimental intervention;

X_i is the user's pre-experiment covariate value;

μ_X is the known or sample mean of the pre-experiment covariate across the combined population;

θ is the adjustment coefficient that minimises the variance of $\widetilde{Y}$.

The optimal coefficient θ^* that minimises $\mathrm{Var}(\widetilde{Y})$ is obtained from ordinary least squares regression of Y on X:

$$theta^* = fractextCov(Y,X)textVar(X)$$

Substituting θ^* into the variance expression reveals the variance reduction factor:

$$textVar(tildeY) = textVar(Y)left(1 - rho_{XY}^2right)$$

Where ρ_{XY} is the Pearson correlation coefficient between the pre-experiment covariate X and the post-randomisation outcome Y. If the pre-experiment metric correlates with post-experiment spend at $\rho_{XY} = 0.70$, CUPED eliminates $0.70^2 = 49\%$ of the outcome variance, allowing the experimenter to achieve identical statistical power with approximately half the original sample size.

Worked Numerical Example: CUPED Variance Reduction on an Australian Fintech Checkout Experiment

An Australian buy-now-pay-later (BNPL) provider tests a redesigned checkout interface intended to increase average monthly gross transaction volume (GTV) per active customer. In the unadjusted historical data, the sample variance of monthly user spend is $s_Y^2 = 40{,}000$ (standard deviation $s_Y = 200$ AUD). The product team wishes to detect a true minimum detectable effect (MDE) of $\delta = 8.00$ AUD per customer with two-sided significance $\alpha = 0.05$ ($Z_{alpha/2} = 1.96$) and statistical power $1 - beta = 0.80$ ($Z_{beta} = 0.8416$).

Prior to randomisation, the team extracts each user's spend during the preceding 30 days (X). Analysis of the pre-experiment data yields $s_X^2 = 36{,}000$ and covariance $\mathrm{Cov}(Y,X) = 28{,}800$.

Step 1: Compute Correlation and Optimal CUPED Coefficient
The sample correlation between pre-experiment spend and experiment-period spend is:

$$rho_{XY} = fractextCov(Y,X)s_Y cdots_X = frac28{,}800200times189.737 = frac28{,}80037{,}947.33 = 0.7589$$

DegreeLive

The optimal linear adjustment coefficient θ^* is:

$$theta^* = fractextCov(Y,X)textVar(X) = frac28,80036,000 = 0.800$$

Step 2: Calculate Variance Reduction and Residual Variance
The variance of the CUPED-adjusted metric $\widetilde{Y}$ is:

$$textVar(tildeY) = textVar(Y) left(1 - rho_{XY}^2 right) = 40,000 times left(1 - 0.7589^2 right) = 40,000 times(1 - 0.5759) = 16,964$$

The adjusted standard deviation drops from 200.00 AUD to:

$$s_{tildeY} = sqrt16,964 = 130.25 textAUD$$

CUPED reduces the unexplained error variance by 57.59%, effectively removing customer-level baseline spending habits that are unrelated to the experimental checkout design.

Step 3: Sample Size Comparison Between Standard and CUPED Designs
For an equal-allocation two-sample test, the required sample size per treatment arm under the standard unadjusted design is:

$$n_{textstandard} = frac2left(Z_{alpha/2} + Z_{beta}right)^2 cdots_Y^2 delta^2 = frac2(1.96 + 0.8416)^2 times40,0008.00^2 = frac2(7.8490)times40,00064 = 9,811 textusersperarm$$

Under the CUPED-adjusted design utilising pre-experiment covariate data:

$$n_{textCUPED} = frac2left(Z_{alpha/2} + Z_{beta}right)^2 cdots_{tildeY}^2 delta^2 = frac2(7.8490)times16,96464 = 4,161 textusersperarm$$

DegreeLive

CUPED contracts the required sample size from 19,622 total active users to 8,322 total active users—a 57.6% reduction in testing traffic. In practice, this enables the fintech platform to conclude the experiment in 12 days rather than 28 days, substantially reducing exposure to macroeconomic shocks and accelerating product deployment.

7. Quasi-Experimental Evaluation: Difference-in-Differences & Regression Discontinuity

In many economic, public policy, and corporate finance contexts, randomised experiments are operationally infeasible, legally prohibited, or ethically unacceptable. For example, evaluating the impact of an Australian state payroll tax reform, an APRA macroprudential lending restriction, or a university tuition subsidy cannot be achieved by randomly assigning individuals to treatment and control conditions. Modern applied econometrics addresses these scenarios using **quasi-experimental identification strategies** that exploit exogenous institutional rules, geographical boundaries, or policy timing.

Difference-in-Differences (DiD) and Parallel Trends

The **Difference-in-Differences (DiD)** design identifies causal effects by comparing the temporal change in outcomes for a treated group before and after policy enactment against the corresponding temporal change observed for an untreated control group:

$$
\begin{aligned}
hattau_{textDiD} = \\
left(\\
barY_{texttreat,post} - \\
barY_{texttreat,pre} \\
right) - \\
left(\\
barY_{textcontrol,post} - \\
barY_{textcontrol,pre} \\
right)
\end{aligned}
$$

Where:

$\hat{tau}_{textDiD}$ is the estimated average treatment effect on the treated (ATT);

$\bar{Y}_{texttreat,post}$ and $\bar{Y}_{texttreat,pre}$ are sample outcome averages for the treated unit across post- and pre-reform periods;

$\bar{Y}_{textcontrol,post}$ and $\bar{Y}_{textcontrol,pre}$ are sample outcome averages for the control unit across identical periods.

The fundamental identifying assumption of DiD is the **Parallel Trends Assumption**: in the counterfactual absence of the treatment, the average outcome of the treated group would have evolved in parallel with the average outcome of the control group. While parallel counterfactual trends cannot be directly verified post-treatment, researchers inspect pre-treatment event-study leads:

$$
\begin{aligned}
Y_{it} = \\
alpha_i + \\
lambda_t + \\
sum_{k=-K,kne-1}^{M} \\
beta_k \\
cdotD_{it}^{k} + \\
varepsilon_{it}
\end{aligned}
$$

Where β_k measures treatment-control differences k periods relative to the policy event at $k = -1$. If pre-treatment coefficients β_{-K}, $dots$, $beta_{-2}$ are statistically indistinguishable from zero, empirical support for parallel trends is established.

Recent Advances: Heterogeneous Treatment Effects in Staggered DiD

Econometric research by Callaway & Sant'Anna (2021) and Goodman-Bacon (2021) demonstrates that classical Two-Way Fixed Effects (TWFE) linear regression with staggered policy adoption assigns negative implicit weights to earlier-treated units used as controls for later-treated units. When treatment effects evolve dynamically over time, TWFE estimates can be severely biased—even reversing signs. Modern applied statistical practice requires group-time average treatment effect estimators $\mathrm{ATT}(g,t)$ that compare newly treated cohorts strictly against never-treated or not-yet-treated units.

DegreeLive

Regression Discontinuity Designs (RDD)

Where policy interventions are allocated based on whether a continuous rating variable X exceeds a strict administrative threshold c, **Regression Discontinuity Design (RDD)** provides near-experimental causal identification. Units just below the threshold ($c -$ $varepsilon$) serve as valid counterfactual controls for units just above the threshold ($c +$ $varepsilon$), because idiosyncratic noise prevents agents from perfectly sorting across the threshold.

In a **Sharp RDD**, treatment receipt changes deterministically from 0 to 1 at c:

$$tau_{textSRD} = lim_{xdownarrowc} E[Y_i midX_i = x] - lim_{xuparrowc} E[Y_i midX_i = x]$$

In a **Fuzzy RDD**, crossing the cutoff creates a discontinuous jump in the probability of treatment receipt ($0 <$ $DeltaP < 1$), requiring local instrumental variable estimation (Wald ratio):

$$tau_{textFRD} = fraclim_{xdownarrowc} E[Y_i midX_i = x] - lim_{xuparrowc} E[Y_i midX_i = x]lim_{xdownarrowc} E[D_i midX_i = x] - lim_{xuparrowc} E[D_i midX_i = x]$$

Worked Numerical Example: Difference-in-Differences Evaluation of an Australian Regional Youth Employment Subsidy

The Victorian State Government implements a targeted wage subsidy program in regional Geelong on 1 July 2023, offering employers a $4,000 annual subsidy for hiring unemployed youth aged 18–24. The regional city of Ballarat, possessing similar macroeconomic, demographic, and industrial structures, does not receive the subsidy and serves as the primary control economy.

Economists collect longitudinal quarterly survey data on the regional youth employment rate (percentage of youth cohort employed) across 4 quarters prior to reform (FY2022–23) and 4 quarters following reform (FY2023–24). The recorded empirical averages are summarized below:

Region	Pre-Reform Average Employment Rate (%)	Post-Reform Average Employment Rate (%)	Observed Regional Difference (ΔY)
Geelong (Treated Region)	58.40%	64.10%	+5.70%
Ballarat (Control Region)	56.20%	58.80%	+2.60%

Step 1: Calculate the 2x2 Difference-in-Differences Point Estimate
The secular macroeconomic time trend observed in the control region is:

$$Delta barY_{textcontrol} = barY_{B,textpost} - barY_{B,textpre} = 58.80\% - 56.20\% = +2.60\%$$

In the absence of the wage subsidy, assuming parallel trends, Geelong's youth employment rate would have risen by the identical 2.60 percentage points to counterfactual level:

$$Y_{G,textcounterfactual} = 58.40\% + 2.60\% = 61.00\%$$

Degree Live

The causal Difference-in-Differences treatment effect (ATT) is the difference between Geelong's actual outcome and its counterfactual:

$$hattau_{textDiD} = (64.10\% - 58.40\%) - (58.80\% - 56.20\%) = +5.70\% - 2.60\% = +3.10\%$$

The wage subsidy caused a 3.10 percentage point increase in Geelong's youth employment rate over and above broader state macroeconomic growth.

Step 2: Regression Specification and Standard Error Inference
In regression form, the model is estimated using individual-level survey data ($N = 4{,}000$ youth respondents, 1,000 per cell):

$$Y_{it} = beta_0 + beta_1 cdot textTreated_i + beta_2 cdot textPost_t + delta_{textDiD} cdot(textTreated_i times textPost_t) + varepsilon_{it}$$

The estimated regression coefficients are:

- $\hat{beta}_0 = 56.20\%$ (Ballarat pre-reform baseline).
- $\hat{beta}_1 = 58.40\% - 56.20\% = +2.20\%$ (pre-existing structural level difference between Geelong and Ballarat).
- $\hat{beta}_2 = 58.80\% - 56.20\% = +2.60\%$ (statewide economic trend affecting both regions).
- $\hat{delta}_{textDiD} = +3.10\%$ (interaction term matching our 2x2 calculation).

Cluster-robust standard errors clustered at the regional municipal level yield $SE(hatdelta_{textDiD}) = 0.85\%$. Computing the test statistic and 95% confidence interval:

$$t = frachatdelta_{textDiD}textSE = frac0.03100.0085 = 3.647, qquadp = 0.00027$$

95

Because the 95% confidence interval excludes zero and $p < 0.001$, the policy impact is statistically significant at conventional academic standards.

Step 3: Economic Cost-Effectiveness Evaluation
Geelong possesses a youth labour force of approximately 16,000 individuals. A 3.10 percentage point increase in the employment rate implies:

DegreeLive

$$Delta \\ textEmployedYouth = 16{,}000 \\ times 0.0310 = 496 \\ textadditionalemployedyouth$$

If the state paid wage subsidies across 1,200 eligible hiring vouchers at $4,000 each, total fiscal expenditure was $1{,}200 \\ times 4{,}000 = 4{,}800{,}000$ AUD. The fiscal cost per net new job created is:

$$textFiscalCostperNetJob = \\ frac{4{,}800{,}000}{textAUD}{496} = 9{,}677.42 \\ textAUD$$

Comparing this fiscal cost against long-term societal benefits (reduced youth welfare dependency, income tax receipts, and reduced youth contact with the justice system) confirms that the program delivered highly cost-effective public welfare enhancement.

DegreeLive