Quantitative Research Methods

Randomised Controlled Trials

Advanced module A1 — optional self-study

Prof. Dr. Christoph Weisser

HSBI

Summer Semester 2026

Where this module fits

Course chapter	What this module builds on it
Ch. 0 — pre-course refresher	The two-sample *t*-test *is* the basic RCT analysis
Ch. 3 — linear regression	Regression adjustment for precision; robust standard errors
Ch. 4 — classification	Binary outcomes (conversion) as experiment metrics
Ch. 5 — resampling	Simulated sampling distributions; power by simulation
Ch. 13 — multiple testing	Peeking and metric panels as multiplicity problems

A new question — not a better model

Chapters 2–10 estimated $\hat{f}(x) \approx E[Y \mid X = x]$ — **correlational** by design, and exactly right for prediction. This module asks the **interventional** question: what happens to Y if *we* set the treatment D? No amount of model flexibility answers it; different *data* — randomised data — do.

One of four optional advanced modules; self-study, with a companion lab notebook.

Contents

Optional and advanced material is collected in the **appendix** at the end of the deck.

Notation in this module

Symbol	Meaning
$Y_i(1),\ Y_i(0)$	potential outcomes of unit i with / without treatment
D_i	treatment indicator (1 = treated); Z_i: *assignment* when they differ
$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$	the observed outcome
$\tau_i = Y_i(1) - Y_i(0)$	individual treatment effect (ITE)
$\tau = E[\tau_i]$	average treatment effect (ATE)
ATT, ATU	average effect on the treated / the untreated
$\pi = \Pr(D_i = 1)$	share treated
$n_1,\ n_0;\ \ \bar{Y}_1,\ \bar{Y}_0$	arm sizes and arm means; $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$
$S_1^2,\ S_0^2$	sample variances within the arms
σ^2	outcome variance used in power planning
δ	effect size to detect; MDE: minimum detectable effect
$\alpha,\ 1 - \beta$	two-sided test size and power; z_q: standard normal quantile
ρ	intraclass correlation within a cluster; DE: design effect
X_i	pre-treatment covariates

Sources and references

Primary textbook (course reference)

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning, with Applications in Python.* Springer.

Foundations for this module

Neyman (1923); Rubin (1974) — potential outcomes. Imbens, G. & Rubin, D. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences.* Cambridge University Press. Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy Online Controlled Experiments.* Cambridge University Press.

Learning objectives

By the end of this module you will be able to:

- **Explain** potential outcomes, the fundamental problem of causal inference, and SUTVA.
- **Derive** the selection-bias decomposition of a naive group comparison.
- **Analyse** a randomised experiment with the difference in means, the two-sample t-test, and regression adjustment with robust standard errors.
- **Design** an experiment: power, sample size, minimum detectable effect, stratified and cluster randomisation.
- **Recognise** the threats — attrition, non-compliance, peeking — and **apply** the intention-to-treat principle.
- **Frame** industrial A/B testing as the RCT it is.

Roadmap of this module

From correlation to causation

1. Why prediction is not causation: a confounded training programme.
2. Potential outcomes, estimands, SUTVA — and the selection-bias decomposition.
3. Randomisation: why it works; the difference-in-means estimator.
4. Analysis in practice: regression adjustment, robust standard errors.
5. Design: power, minimum detectable effect, blocking, clusters.
6. Threats and practice: attrition, non-compliance, balance, peeking, A/B tests.

Where this module is used in industry

Sector	Concrete application	Which decision it drives
Tech, e-commerce	A/B tests on features, ranking, pricing UX	Ship or roll back the change
Pharma	Phase-III randomised, double-blind trials	Whether the drug is approved
Development, policy	Cash-transfer and nudge experiments	Which programme is scaled up
Marketing	Incrementality / uplift tests with holdout groups	Where the ad budget goes
Banking, insurance	Randomised pricing, CRM and collection pilots	Which policy is rolled out
Marketplaces, mobility	Switchback and geo experiments	Pricing and matching algorithms

In industry — The common thread

Whenever the decision is *causal* — ship it, approve it, scale it — the gold standard is the same design: create the counterfactual by **randomisation** instead of assuming it away. Tech simply industrialised what agriculture and medicine invented: Fisher's randomised plots became feature flags.

Industry case in depth: a phase-III drug trial

In industry — The most regulated RCT there is

A candidate drug reaches phase III: two arms, **randomised 1:1** and **double-blind**, one **pre-registered primary endpoint** (say, 12-month survival), and a sample size from a power calculation — large enough to detect the smallest clinically relevant effect with 80–90% power. The protocol and the statistical analysis plan are frozen *before* the first patient is enrolled.

The rules the regulator enforces

Analysis is by **intention to treat**: every randomised patient counts in the arm they were assigned to, whether or not they took the drug — anything else re-opens the door to self-selection. Interim looks are allowed only through pre-specified group-sequential boundaries (a Chapter 13 idea). Secondary endpoints need a multiplicity strategy before they can support a claim.

Two honest caveats

ITT estimates the effect of *offering* the drug, not of taking it; and the trial population (consenting, screened patients) is not the treated population — internal validity is bought first, external validity argued second.

Outline

A training programme that "hurts": prediction is not causation

A firm offers an optional training programme. We simulate 1000 workers (`rng(2024)`; wages in \$1000s) where the *true* built-in effect of training is $+4.1$ on average — and then compare the two groups the way a dashboard would.

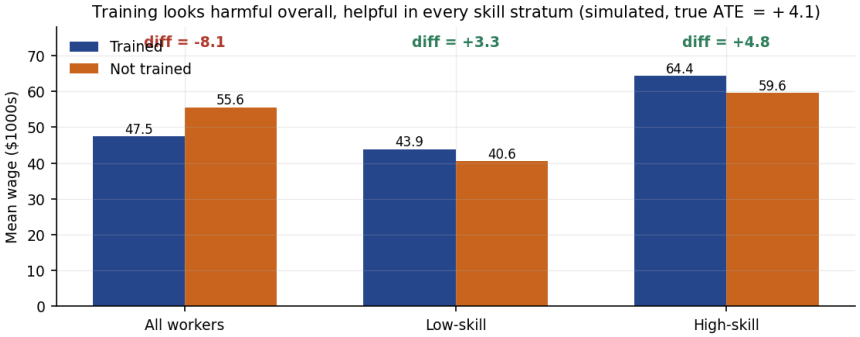
The naive comparison

- Mean wage of the 491 trained workers: 47.5; of the 509 untrained: 55.6.
- Naive difference: $47.5 - 55.6 = -8.1$ — training "costs" \$8100?
- The truth built into the simulation: training *raises* wages by $+4.1$.

Takeaway

The comparison is computed correctly; the *causal reading* is wrong. Who takes training is not random: in this population, 80% of low-skill workers enrol but only 20% of high-skill workers. The groups differ *before* treatment does anything.

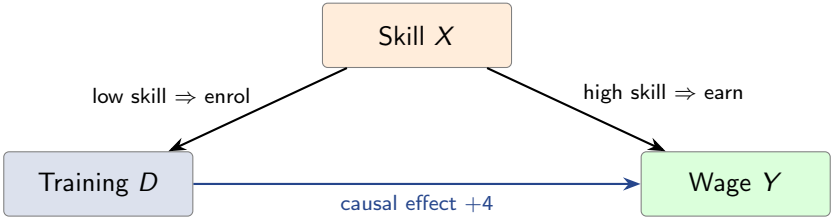
The sign flips inside every stratum (Simpson's paradox)



Takeaway

Overall: -8.1. Within low-skill workers: $+3.3$ (404 trained vs 107 not); within high-skill: $+4.8$ (87 vs 402). The aggregate flips sign because 82% of the trained are low-skill against only 21% of the untrained.

Confounding, drawn (native diagram)



Reading the diagram

The naive comparison measures *both* arrows into Y: the causal path $D \to Y$ **plus** the backdoor path $D \leftarrow X \to Y$. Comparing within skill strata closes the backdoor — but only for confounders you *observe*; motivation, health or ability would confound invisibly.

In industry — Marketing: attribution

"Customers who saw the ad bought 30% more" — but the ad was targeted at likely buyers. Until a randomised holdout (ghost ads) says otherwise, that lift is selection, not persuasion.

Outline

The Neyman–Rubin framework: two outcomes per unit, one observed

Potential outcomes and estimands

Each unit i carries *two* potential outcomes: $Y_i(1)$ if treated, $Y_i(0)$ if not; the individual effect is $\tau_i = Y_i(1) - Y_i(0)$. We observe only

$$\boxed{Y_i \;=\; D_i\, Y_i(1) + (1 - D_i)\, Y_i(0).}$$

Estimands: $\text{ATE} = E[\tau_i]$, $\text{ATT} = E[\tau_i \mid D_i = 1]$, $\text{ATU} = E[\tau_i \mid D_i = 0]$, with $\text{ATE} = \pi\,\text{ATT} + (1 - \pi)\,\text{ATU}$.

How to read this

- $Y_i(1), Y_i(0)$ exist *before* treatment is decided — treatment only selects which one you get to see.
- ATE: effect of treating *everyone*; ATT: effect on those who were treated — a mandatory rollout needs the ATE, a voluntary programme the ATT.

The God's-eye view: a potential-outcomes table

The fundamental problem: each unit reveals only one potential outcome

Unit i	$Y_i(1)$	$Y_i(0)$	τ_i	D_i	Y_i **observed**
1	*57*	57	*+0*	0	57
2	71	*70*	*+1*	1	71
3	*59*	52	*+7*	0	52
4	*56*	56	*+0*	0	56
5	*60*	59	*+1*	0	59
6	*60*	59	*+1*	0	59
7	85	*77*	*+8*	1	85
8	76	*73*	*+3*	1	76
9	79	*77*	*+2*	1	79
10	80	*79*	*+1*	1	80

Grey italic cells are counterfactuals: never observed. God's-eye ATE $= 2.40$; naive difference $= 78.2 - 56.6 = +21.6$ because treatment went to the five highest $Y_i(0)$.

Takeaway

In this simulated table the true ATE is $+2.40$ (and ATT $= 3.00$, ATU $= 1.80$). An analyst sees only the coloured cells; because treatment went to the five highest $Y_i(0)$, the naive difference is $78.2 - 56.6 = +21.6$ — nine times the truth.

SUTVA — the assumption the whole notation rests on

Stable Unit Treatment Value Assumption

Writing $Y_i(1), Y_i(0)$ at all assumes: (i) **no interference** — unit i's outcome does not depend on *other* units' treatments; (ii) **one version of treatment** — "treated" means the same thing for every unit.

Where it honestly fails

- Vaccination trials: my risk falls when *you* are vaccinated (herd immunity).
- Marketplace tests: a discount for treated buyers depletes inventory or driver supply for the controls — the control group is contaminated.
- Social features: "invite your friends" spills across arms by design.

SUTVA is a claim about the *world*, not a modelling choice — and randomisation does **not** buy it. When it fails, redefine the unit (clusters, regions, time slices) so that it holds.

In industry — Two-sided marketplaces

Why the naive comparison lies: the selection-bias decomposition

Derivation (two lines)

$$
\begin{aligned}
E[Y \mid D=1] - E[Y \mid D=0] &= E[Y(1) \mid D=1] - E[Y(0) \mid D=0] \\
&= \underbrace{E[Y(1) - Y(0) \mid D=1]}_{\text{ATT}} + \underbrace{E[Y(0) \mid D=1] - E[Y(0) \mid D=0]}_{\text{selection bias}}
\end{aligned}
$$

$$\boxed{\text{naive difference} \;=\; \text{ATT} \;+\; \text{selection bias.}}$$

How to read this

First line: each group reveals its own potential outcome. Second line: add and subtract $E[Y(0) \mid D=1]$ — the *counterfactual* of the treated. Selection bias asks: how would the treated have differed **anyway**, untreated? (Replacing ATT by ATE adds a third term, $(1-\pi)(\text{ATT} - \text{ATU})$ — Exercise A1.2.)

The training programme decomposed

Exercise A1.1 — Potential outcomes by hand [Concept]

Exercise

Six workers have the following (God's-eye) potential outcomes; D_i marks who actually took the training:

i	1	2	3	4	5	6
$Y_i(1)$	5	9	6	11	6	11
$Y_i(0)$	4	7	3	9	5	8
D_i	0	1	0	1	0	1

1. Compute every individual effect τ_i and the ATE.
2. Which outcomes does the analyst actually observe? Is any τ_i observable?
3. Compute the naive difference in observed means.
4. Explain the gap using the decomposition: compute ATT and the selection-bias term.

Solution — Exercise A1.1

Solution

Part 1 — effects. $\tau = (1, 2, 3, 2, 1, 3)$, so ATE $= \frac{1+2+3+2+1+3}{6} = 2.0$.

Part 2 — observed data. Only $Y_2(1) = 9$, $Y_4(1) = 11$, $Y_6(1) = 11$ and $Y_1(0) = 4$, $Y_3(0) = 3$, $Y_5(0) = 5$. No unit shows both potential outcomes, so no τ_i is observable — the fundamental problem.

Part 3 — naive difference. $\bar{Y}_1 - \bar{Y}_0 = \frac{9+11+11}{3} - \frac{4+3+5}{3} = 10.33 - 4.00 = 6.33$ — more than three times the ATE.

Part 4 — decomposition. ATT $= \frac{2+2+3}{3} = 2.33$; selection bias $= E[Y(0) \mid D{=}1] - E[Y(0) \mid D{=}0] = \frac{7+9+8}{3} - 4 = 8 - 4 = 4.00$. Check: $2.33 + 4.00 = 6.33$ ✓. Treatment went to workers who would have done better anyway.

Common mistake

Calling naive − ATE = 4.33 "the" selection bias. The exact bias term is $E[Y(0) \mid D{=}1] - E[Y(0) \mid D{=}0] = 4.00$; the remaining 0.33 is the ATT–ATE gap (treatment effects differ across groups).

Exercise A1.2 — The three-term decomposition [Math]

Exercise

Let $\pi = \Pr(D = 1)$. Starting from naive $= \text{ATT} +$ selection bias and the identity $\text{ATE} = \pi\,\text{ATT} + (1 - \pi)\,\text{ATU}$:

1. Show that naive $= \text{ATE} +$ selection bias $+ (1 - \pi)\,(\text{ATT} - \text{ATU})$.
2. Interpret the third term in one sentence.
3. Verify the formula on the table of Exercise A1.1 (there $\pi = \frac{1}{2}$).

Solution — Exercise A1.2

Solution

Part 1 — derivation. From the identity,

$$\text{ATT} - \text{ATE} = \text{ATT} - \pi\,\text{ATT} - (1-\pi)\,\text{ATU} = (1-\pi)\,(\text{ATT} - \text{ATU}).$$

Substituting $\text{ATT} = \text{ATE} + (1-\pi)(\text{ATT} - \text{ATU})$ into the two-term decomposition gives

$$\text{naive} = \text{ATE} + \underbrace{E[Y(0) \mid D{=}1] - E[Y(0) \mid D{=}0]}_{\text{selection bias}} + \underbrace{(1-\pi)(\text{ATT} - \text{ATU})}_{\text{differential effect}}.$$

Part 2 — interpretation. Even with no baseline selection, the naive difference misses the ATE when those who take treatment *benefit differently* from those who do not.

Part 3 — check. $\text{ATU} = \frac{1+3+1}{3} = 1.67$, so $2.00 + 4.00 + \frac{1}{2}(2.33 - 1.67) = 2.00 + 4.00 + 0.33 = 6.33$ ✓ — exactly the naive difference of A1.1. (In the training programme: $-8.07 = 4.07 - 12.04 - 0.10$.)

Common mistake

Believing "no selection bias" makes the naive comparison equal the ATE. It equals the *ATT*; the differential-effect term must also vanish before it equals the ATE.

Outline

Randomisation kills both bias terms — in expectation

What a coin flip buys

Random assignment forces $D \perp (Y(1), Y(0))$, hence $E[Y(0) \mid D{=}1] = E[Y(0) \mid D{=}0]$ (selection bias $= 0$) and ATT $=$ ATU $=$ ATE. The naive comparison becomes the estimand:

$$\boxed{E[Y \mid D = 1] - E[Y \mid D = 0] \;=\; \tau.}$$

How to read this

The coin cannot peek at $Y_i(0)$ or $Y_i(1)$ — treated and control are random samples of the *same* population, so **all** confounders, observed and unobserved, balance in expectation. That last clause is the entire magic: no list of control variables is ever needed, or ever complete.

Takeaway

"In expectation" is doing real work: any *single* randomisation can be unlucky. The standard error on the next slide quantifies exactly that design noise — bias is gone, variance remains.

The difference in means, its standard error — and Ch. 0

Estimator, standard error, test

$$\hat{\tau} = \bar{Y}_1 - \bar{Y}_0, \qquad \widehat{\mathrm{SE}} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}}, \qquad \hat{\tau} \pm z_{0.975}\,\widehat{\mathrm{SE}}, \qquad t = \frac{\hat{\tau}}{\widehat{\mathrm{SE}}}.$$

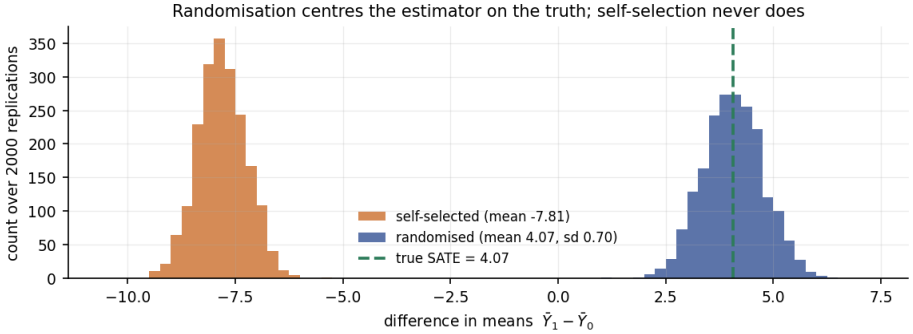
This is the pre-course two-sample t-test

Identical to the Welch statistic from the Chapter 0 refresher: the humble t-test *is* a complete RCT analysis. Neyman (1923) showed this SE is (weakly) **conservative** for the randomisation variance — the honest direction to err (derivation in the appendix).

Training population — randomised for once

Assign 500/500 at random: one draw gives $\widehat{\mathrm{SE}} \approx 0.70$. Across 2000 re-randomisations, the estimator's actual standard deviation is 0.70 — the formula tracks the true design variability.

Seeing it: the estimator's sampling distribution under both designs



Takeaway

Same population, 2000 replications. Randomised (blue): mean 4.07 — exactly the truth 4.07 — sd 0.70, and the 95% CI covers the truth in 95.1% of draws. Self-selected (orange): mean -7.81, a structural bias of -11.9 that **no sample size reduces** — the orange histogram is narrow *and* wrong.

Exercise A1.3 — Analysing a small RCT [Math]

Exercise

A randomised experiment ends with $n_1 = n_0 = 100$, treated mean $\bar{Y}_1 = 54.0$ ($S_1 = 10$), control mean $\bar{Y}_0 = 50.0$ ($S_0 = 10$).

1. Compute $\hat{\tau}$ and its standard error.
2. Compute the t-statistic and the 95% confidence interval (use $t_{0.975,\,198} = 1.972$).
3. Decide at $\alpha = 0.05$ and state the conclusion in one honest sentence.
4. What happens to the SE if both arms are doubled to 200?

Solution — Exercise A1.3

Solution

Part 1. $\hat{\tau} = 54.0 - 50.0 = 4.0$; $\widehat{\mathrm{SE}} = \sqrt{\frac{100}{100} + \frac{100}{100}} = \sqrt{2} = 1.414$.

Part 2. $t = 4.0/1.414 = 2.83$; CI: $4.0 \pm 1.972 \times 1.414 = [1.21,\ 6.79]$; $p \approx 0.005$.

Part 3. Reject $H_0 : \tau = 0$. "Treatment raised the outcome by an estimated 4.0 points (95% CI 1.2 to 6.8)" — because assignment was randomised, this difference *is* the causal effect estimate, not a correlation.

Part 4. $\widehat{\mathrm{SE}} = \sqrt{\frac{100}{200} + \frac{100}{200}} = 1.0$: doubling both arms shrinks the SE by $1/\sqrt{2}$ — precision is bought at quadratic cost, which is why power planning (Section 5) comes *before* launch.

Common mistake

Computing the SE as if the 200 observations were one sample ($S/\sqrt{200} = 0.71$). The SE of a *difference* adds the two arms' variances — it is $\sqrt{2}$ times larger.

Extended Exercise A1.1 — Randomisation vs selection in simulation [Python]

Extended exercise (15 min)

Build the training population from the lecture with `np.random.default_rng(2024)`: `high` $\sim$ Bernoulli(0.5), enrolment probability 0.8/0.2 for low/high skill, $Y(0) = 40 + 20\,\texttt{high} + \mathcal{N}(0, 5^2)$, $\tau = 4 + \mathcal{N}(0, 2^2)$, $Y(1) = Y(0) + \tau$, $n = 1000$.

1. Simulate 2000 replications of (a) **self-selected** uptake and (b) **randomised** 500/500 assignment; record the difference in means.
2. Report mean and sd of both estimator distributions, and the bias of (a).
3. Under randomisation, compute the coverage of the 95% CI $\hat{\tau} \pm 1.96\,\widehat{\mathrm{SE}}$.
4. In two sentences: what does randomisation fix, and what does it not fix?

Run this live

The worked solution sits in `advanced_01_rcts_lab.ipynb` under *Lecture exercises — worked Python solutions*: the cell headed *Extended Exercise A1.1 — Randomisation vs selection in simulation [Python]*.

Solution — Extended Exercise A1.1 (1/2)

Solution — code: population and the two designs

```
import numpy as np
rng = np.random.default_rng(2024) # module seed
n = 1000
high = rng.random(n) < 0.5 # skill stratum
p_sel = np.where(high, 0.2, 0.8) # self-selection propensities
_ = rng.random(n) < p_sel # the observational draw
y0 = 40 + 20*high + rng.normal(0, 5, n) # potential outcome untreated
tau = 4 + rng.normal(0, 2, n) # heterogeneous ITEs, mean 4
y1 = y0 + tau # potential outcome treated

est_r, est_s, cover = [], [], 0
for _ in range(2000): # 2000 replications
    Dr = np.zeros(n, bool); Dr[rng.permutation(n)[:500]] = True # randomised
    yr = np.where(Dr, y1, y0)
    d = yr[Dr].mean() - yr[~Dr].mean(); est_r.append(d)
    se = np.sqrt(yr[Dr].var(ddof=1)/500 + yr[~Dr].var(ddof=1)/500)
    cover += abs(d - tau.mean()) <= 1.96*se # CI covers truth?
    Ds = rng.random(n) < p_sel # self-selected uptake
    ys = np.where(Ds, y1, y0)
```

Solution — Extended Exercise A1.1 (2/2)

Solution — numbers and interpretation

Parts 2–3 — printout (true SATE = 4.07):

```
randomised  mean 4.07  sd 0.70  coverage 0.951
self-selected  mean -7.81  sd 0.56  bias -11.88
```

- **Randomised**: centred on the truth; the Neyman CI covers it in 95.1% of draws — the advertised guarantee, delivered.
- **Self-selected**: bias −11.9, stable across replications — collecting the same biased data 2000 times over changes nothing.

Part 4. Randomisation removes *bias* (both decomposition terms are zero in expectation); it does not remove *variance* — a single unlucky draw can still be off, which is what the SE and CI quantify. It also does not buy SUTVA or fix attrition and non-compliance (Section 6).

Common mistake

Expecting the self-selected estimator to be *noisier*. It is actually slightly *more* stable here (sd 0.56 vs 0.70) — precision is no defence against bias: a tight interval around the wrong number is the most dangerous

Outline

Should we adjust an RCT with regression? (bridge to Ch. 3)

Regression adjustment

Fit the Chapter 3 model on experimental data:

$$Y_i = \alpha + \tau D_i + X_i^\top \beta + \varepsilon_i.$$

Randomisation makes $X \perp D$, so $\hat{\tau}$ still estimates the ATE — the covariates soak up outcome variance and **shrink the standard error**.

The rules that keep it honest

- Covariates must be **pre-treatment** (measured before assignment) and **pre-specified** (chosen before results are seen).
- Adjustment here buys *precision*, not identification — unlike in observational data, the estimate is valid even with *no* covariates.
- With very unequal arms or strong effect heterogeneity, add $D \times (X - \bar{X})$ interactions (Lin 2013) — the safe default at scale.

A semi-synthetic experiment on Wage ($n = 3000$)

Real covariates and wages from `Wage.csv`; we *randomise* a fake programme and build in a true effect of $+5$:

```
import numpy as np, pandas as pd
import statsmodels.formula.api as smf
wage = pd.read_csv("Wage.csv") # course data, n = 3000
rng = np.random.default_rng(2024)
wage["D"] = (rng.random(len(wage)) < 0.5).astype(int) # coin flip: 1475 vs 1525
wage["y"] = wage["wage"] + 5.0 * wage["D"] # truth by construction: tau = 5

m0 = smf.ols("y ~D", data=wage).fit(cov_type="HC2") # robust (HC2) SEs
print(m0.params["D"], m0.bse["D"]) # 4.122 1.523
```

Why `cov_type="HC2"`

Treatment may change the *variance* of Y, not just its mean; classical OLS SEs assume equal error variance across arms. Heteroskedasticity-consistent (HC) errors drop that assumption — HC2 exactly reproduces the Neyman variance in a two-arm experiment, so it is the natural default for RCTs.

What adjustment buys: the numbers

Unadjusted vs adjusted — true effect +5; HC2 SEs

Model	$\hat{\tau}$	SE	95% CI
`y ~ D`	4.12	1.52	[1.14, 7.11]
`y ~ D + age + education`	3.73	1.31	[1.15, 6.31]

SE ratio 0.86 $\Rightarrow$ variance down 25% $\Rightarrow$ worth 34% more sample, for free. The covariates explain $R^2 \approx 0.26$ of the outcome — that is the fuel of the precision gain.

Takeaway

Adjustment buys **precision, not truth**: both estimates are unbiased and both CIs cover the built-in +5; the adjusted one is simply less noisy. The estimates differ (4.12 vs 3.73) only through sampling noise — with one finite randomisation, that is expected, not alarming.

In industry — CUPED: the same idea at scale

Exercise A1.4 — Regression adjustment on the Wage experiment [Python]

Exercise

Using `Wage.csv` and the semi-synthetic experiment from the lecture (seed 2024, coin-flip D, $y = \texttt{wage} + 5D$):

1. Fit the unadjusted model `y ~ D` with HC2 standard errors and report $\hat{\tau}$ and its SE.
2. Add `age` and `education`; report the new $\hat{\tau}$ and SE.
3. Compute the implied variance reduction and the equivalent sample-size gain.
4. A colleague proposes also adjusting for a satisfaction survey answered *after* the programme. Why must you refuse?

Run this live

The worked solution sits in `advanced_01_rcts_lab.ipynb` under *Lecture exercises — worked Python solutions*: the cell headed *Exercise A1.4 — Regression adjustment on the Wage experiment [Python]*.

Solution — Exercise A1.4 (1/2)

Solution — code

```
import numpy as np, pandas as pd
import statsmodels.formula.api as smf
wage = pd.read_csv("Wage.csv") # n = 3000
rng = np.random.default_rng(2024) # lecture seed
wage["D"] = (rng.random(len(wage)) < 0.5).astype(int)
wage["y"] = wage["wage"] + 5.0 * wage["D"] # true effect +5

m0 = smf.ols("y ~D", data=wage).fit(cov_type="HC2")
m1 = smf.ols("y ~D + age + education", data=wage).fit(cov_type="HC2")
print(round(m0.params["D"], 3), round(m0.bse["D"], 3)) # 4.122 1.523
print(round(m1.params["D"], 3), round(m1.bse["D"], 3)) # 3.732 1.315
```

Solution — Exercise A1.4 (2/2)

Solution — reading the output

Parts 1–2. Unadjusted: $\hat{\tau} = 4.122$ (SE 1.523). Adjusted: $\hat{\tau} = 3.732$ (SE 1.315). Both within one SE of the built-in truth $+5$.

Part 3. Variance ratio $(1.315/1.523)^2 = 0.746$: adjustment cut the variance by 25.4%. Equivalent sample gain: $1/0.746 = 1.34$ — the adjusted analysis of 3000 people is as precise as an unadjusted one of about 4000.

Part 4. The survey is **post-treatment**: the programme may change satisfaction, so conditioning on it splits the sample by a *consequence* of D — a "bad control" that re-introduces exactly the selection bias randomisation eliminated. Only pre-treatment covariates are admissible.

Common mistake

Reading $\hat{\tau}$ moving from 4.12 to 3.73 as the adjusted model "finding a smaller effect". Both estimate the same τ; the wobble is sampling noise. What changed systematically is the *SE* — that is the whole point of adjusting.

Outline

Power and the minimum detectable effect

Sample size for a two-arm trial (equal arms, two-sided α)

$$n_{\text{per arm}} = \frac{2\,\sigma^2\,(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2} \quad \Longleftrightarrow \quad \text{MDE} = (z_{1-\alpha/2} + z_{1-\beta})\,\sigma\,\sqrt{\tfrac{2}{n}}.$$

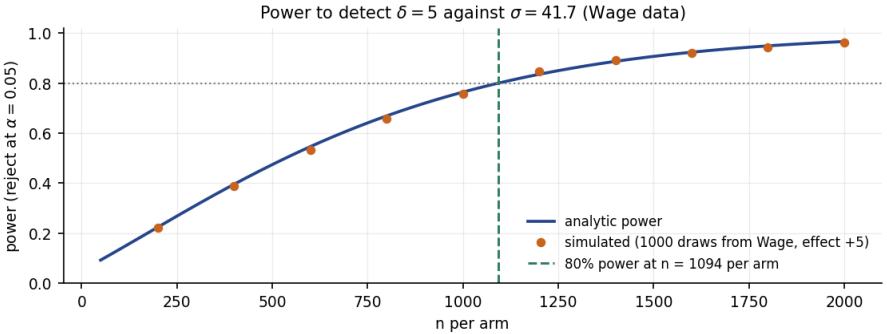
How to read this

- σ: outcome sd; δ: smallest effect worth detecting; α: test size; $1-\beta$: power.
- For $\alpha = 0.05$, power 80%: $z_{0.975} = 1.96$, $z_{0.80} = 0.84$, so $(z_{1-\alpha/2} + z_{1-\beta})^2 = 7.85$ — the handy rule $n \approx 15.7\,\sigma^2/\delta^2$ per arm.
- The MDE is the same equation read backwards: given my n, what can I honestly hope to see?

The wage experiment — planned properly

$\sigma = 41.7$ (sd of `wage`), $\delta = 5$: $n = 2 \times 7.85 \times 41.7^2/25 = 1094$ per arm. Conversely, at $n = 1000$ per arm the MDE is $2.80 \times 41.7 \times \sqrt{2/1000} = 5.23$.

The power curve: analytic formula vs simulation (bridge to Ch. 5)



Takeaway

Orange dots: 1000 simulated experiments per n, resampling real `Wage` data with a $+5$ effect — at $n = 500$ the simulation says 0.477, the formula 0.474. Both agree: 80% power needs 1094 per arm; at 500 per arm you toss a slightly bent coin.

Design refinements: blocking and cluster randomisation

Stratified (blocked) randomisation

Randomise *within* strata of a prognostic variable (skill, country, device). Balance on the stratum is exact by construction, not just in expectation; analyse with stratum fixed effects. Precision gain $\approx$ regression adjustment, guaranteed by design.

Cluster randomisation and the design effect

When treatment operates on groups (schools, cities, stores) or interference within groups breaks SUTVA, randomise *clusters*. Outcomes within a cluster are correlated (ρ), so n raw observations are worth fewer:

$$\boxed{\mathrm{DE} = 1 + (m-1)\,\rho} \qquad n_{\text{effective}} = n/\mathrm{DE}.$$

Worked example

200 schools $\times$ 50 pupils, $\rho = 0.05$: $\mathrm{DE} = 1 + 49 \times 0.05 = 3.45$ — the 10 000 pupils carry the information of $10\,000/3.45 \approx 2899$ independent ones.

Exercise A1.5 — Sample size for an A/B test [Math]

Exercise

A sign-up funnel converts at $p_0 = 10\%$. You must detect an absolute lift of 2 percentage points (to $p_1 = 12\%$) at $\alpha = 0.05$ (two-sided) with 80% power. For proportions, the per-arm sample size is

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2 \left[p_0(1-p_0) + p_1(1-p_1)\right]}{(p_1 - p_0)^2}.$$

1. Compute n per arm (use $(z_{1-\alpha/2} + z_{1-\beta})^2 = 7.85$).
2. The product owner halves the MDE to 1 point ($p_1 = 11\%$). Recompute n.
3. State the approximate scaling law of n in the MDE δ.

Solution — Exercise A1.5

Solution

Part 1. Variance term: $0.10 \times 0.90 + 0.12 \times 0.88 = 0.1956$.

$$n = \frac{7.85 \times 0.1956}{0.02^2} = \frac{1.535}{0.0004} \approx 3839 \text{ per arm } (\approx 7700 \text{ users in total}).$$

Part 2. Variance term: $0.10 \times 0.90 + 0.11 \times 0.89 = 0.1879$;

$$n = \frac{7.85 \times 0.1879}{0.01^2} \approx 14\,749 \text{ per arm} \quad (3.84 \times \text{Part 1}).$$

Part 3. $n \propto 1/\delta^2$: halving the detectable effect roughly **quadruples** the sample (here $3.84\times$, slightly less than 4 because the variance term also shrinks as p_1 moves toward p_0). Sensitivity is the most expensive word in experimentation.

Common mistake

Quoting n for a *relative* lift as if it were absolute. A "2% lift" on a 10% baseline is 0.2 points, not 2 — a $100\times$ larger required sample. Always spell out absolute percentage points.

Outline

Attrition: the quiet return of selection bias

Randomisation happens at assignment — but people *leave* afterwards, and they do not leave at random.

When dropout hurts — and how to see it

- Harmless: attrition unrelated to treatment and outcome (pure noise, less n).
- Harmful: **differential** attrition — e.g. the control group's weakest participants quit, so the surviving controls look artificially strong.
- Diagnostics: compare attrition *rates* by arm; compare *pre-treatment covariates* of the stayers by arm; if suspicious, report worst-case (Manski/Lee) bounds rather than a point estimate.

Takeaway

You randomised the *assigned*; you analyse the *observed*. The moment those two populations differ by arm, the selection-bias term of Section 2 is back — with the same algebra and the same danger.

Non-compliance and the intention-to-treat principle

People assigned to treatment ($Z_i = 1$) do not always take it ($D_i = 0$) — and compliance is self-selected.

Intention-to-treat (ITT)

Compare by *assignment*, ignoring uptake:

$$\boxed{\widehat{\mathrm{ITT}} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}}$$

Z is randomised even when D is not — ITT inherits the full causal guarantee.

How to read this

ITT estimates the effect of *offering* the treatment — exactly the quantity a policy or product decision needs, since rollout can only ever offer. It is diluted by non-compliance: with compliance share c and no effect on never-takers, $\mathrm{ITT} = c \times$ effect on compliers.

Worked example

Per-protocol analysis vs the compliers' effect (CACE)

Per-protocol: the tempting, biased alternative

Comparing those who *took* treatment with everyone else discards randomisation: compliers are self-selected (keener, healthier, more engaged), so the selection-bias decomposition applies in full.

The instrumental-variables repair

Under randomised Z, exclusion (Z affects Y only through D) and one-sided non-compliance:

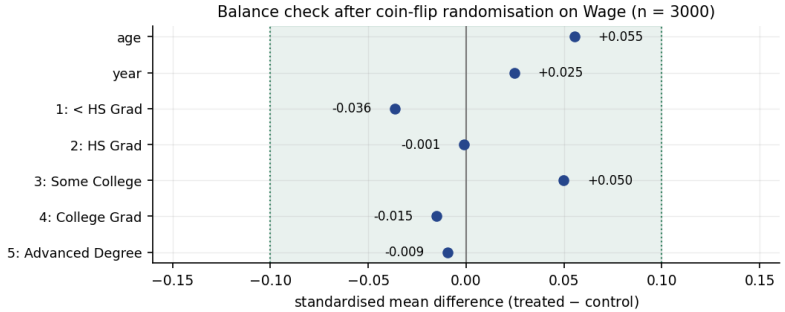
$$\widehat{\mathrm{CACE}} = \frac{\widehat{\mathrm{ITT}}_Y}{\widehat{\mathrm{ITT}}_D} = \frac{\bar{Y}_{Z=1} - \bar{Y}_{Z=0}}{\text{compliance share}}$$

(the Wald estimator; derivation in the appendix).

Worked example

$\widehat{\mathrm{CACE}} = 6.0/0.75 = 8.0$: the offer moved outcomes by 6.0 on average *because* the 75%

Balance checks: auditing the randomiser

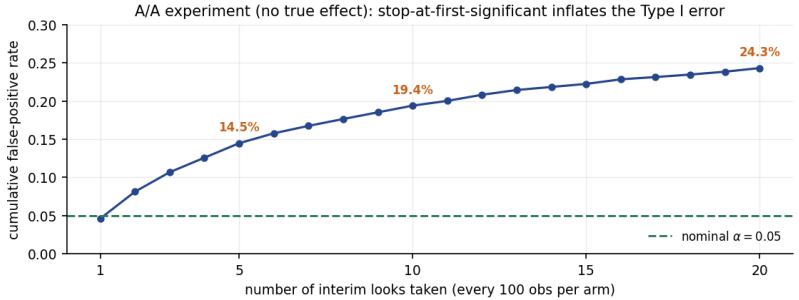


Takeaway

Standardised mean differences for the Wage experiment: every covariate sits within ± 0.06 — comfortably inside the $|\mathrm{SMD}| < 0.1$ rule of thumb (largest: age at $+0.055$). This is what a healthy randomisation looks like.

Balance-test theatre

Peeking: multiple testing in the time dimension (bridge to Ch. 13)



Takeaway

An A/A test (no true effect), checked after every 100 users per arm and stopped at the first star: the realised false-positive rate is 4.6% after one look but 14.5% by 5 looks, 19.4% by 10, 24.3% by 20. Each look is another test on overlapping data — correlated, so below the independent-tests bound $1 - 0.95^{20} = 64\%$, yet five times nominal α.

The A/B playbook: this module, spoken in tech

This module	A/B practice
Unit and randomisation	User, assigned by hashing a stable id into variants
Treatment D	The variant behind a feature flag
Primary outcome	One pre-registered north-star metric per experiment
Power / MDE	Traffic forecast $\rightarrow$ MDE $\rightarrow$ run length, fixed in advance
ITT	Analyse every *bucketed* user, exposed or not
SUTVA repair	Switchback / geo / cluster designs for marketplaces
Peeking control	Fixed horizon, or sequential (always-valid) methods
Secondary metrics, segments	FDR control and "exploratory" labels — Chapter 13

In industry — A/B testing is the RCT of tech

Large platforms run tens of thousands of experiments a year; the experiment config *is* the pre-registration, guardrail metrics are the safety endpoints, and the ramp-up plan is the dose escalation. Every hard problem in this section — interference, peeking, multiplicity — is a live engineering concern, not a textbook footnote.

Exercise A1.6 — ITT, per-protocol and compliance [Concept]

Exercise

An app randomises 2000 users 1:1: the treatment arm is *offered* a coaching feature; the control arm is not. In the offered arm, 75% activate the feature; no control user can. The offered arm's mean engagement ends 6.0 points above the control arm's.

1. State the ITT estimate. What causal quantity does it measure?
2. An analyst proposes comparing the 750 activators with all 1250 non-activating users (both arms pooled). What exactly is wrong?
3. Estimate the effect on the users who actually activate (state your assumptions).
4. Which number belongs in the rollout decision, and why?

Solution — Exercise A1.6

Solution

Part 1. $\widehat{\mathrm{ITT}} = 6.0$ points: the causal effect of *offering* the feature, guaranteed unbiased because the offer was randomised.

Part 2. Activators are self-selected — the engaged users opt in. The comparison equals ATT + selection bias (Section 2), and the bias term is almost surely positive here: activators would have been more engaged *anyway*. It also pools controls with unmotivated treated users, muddying both sides.

Part 3. With one-sided non-compliance and exclusion (Z moves engagement only via activation): $\widehat{\mathrm{CACE}} = 6.0/0.75 = 8.0$ points among compliers.

Part 4. The **ITT** (6.0): shipping the feature can only ever *offer* it, so the offer effect is the rollout effect. The CACE (8.0) answers a different, product-design question — how valuable the feature is to those who engage.

Common mistake

Reporting the CACE as "the effect of the feature" without the assumptions (randomised offer, exclusion, one-sidedness) — and without saying it applies to compliers only, not to the average user.

Extended Exercise A1.2 — Designing the free-shipping experiment [Integrative]

Extended exercise (15 min)

An online shop converts $p_0 = 5.0\%$ of visitors. Marketing wants to test a free-shipping threshold and cares about an absolute lift of 0.5 percentage points. Traffic: 20 000 visitors per week, split evenly between arms.

1. Compute the required sample per arm ($\alpha = 0.05$, power 80%, $(z_{1-\alpha/2} + z_{1-\beta})^2 = 7.85$) and the run time in weeks.
2. The product manager suggests checking daily and stopping at the first significant result. Quantify the danger and give two acceptable alternatives.
3. Should you randomise *sessions* or *users*? Justify via SUTVA and independence.
4. Twelve secondary metrics will be tracked. State a defensible reporting rule (bridge to Chapter 13).

Solution — Extended Exercise A1.2 (1/2)

Solution — power arithmetic and run time

Part 1. Variance term: $0.05 \times 0.95 + 0.055 \times 0.945 = 0.0475 + 0.0520 = 0.0995$.

$$n = \frac{7.85 \times 0.0995}{0.005^2} = \frac{0.781}{0.000025} \approx 31\,231 \text{ per arm.}$$

At 10 000 visitors per arm per week: $31\,231/10\,000 \approx 3.1$ weeks $\Rightarrow$ **run 4 full weeks** — rounding *up* to whole weekly cycles keeps weekday/weekend mix balanced.

MDE honesty check. If management demanded an answer after one week (10 000 per arm), the MDE would be $2.80\sqrt{2 \times 0.05 \times 0.95/10\,000} \approx 0.86$ points — the test could *not* see the 0.5-point effect anyone cares about. Say so before launch, not after.

Solution — Extended Exercise A1.2 (2/2)

Solution — peeking; units; metric discipline

Part 2. Daily checks over 4 weeks = up to 28 looks; the lecture's A/A simulation already reaches a 24.3% false-positive rate by 20 looks — roughly one ineffective change in four would ship. Alternatives: (i) fixed horizon, one look at 4 weeks; (ii) pre-specified group-sequential boundaries or always-valid sequential inference, which permit monitoring at controlled α.

Part 3. Randomise **users**. Sessions of the same visitor are not independent, and a returning visitor would bounce between variants — violating the "one version of treatment" half of SUTVA and contaminating both arms. (Analysing sessions while randomising users then requires cluster-robust SEs — the design effect of Section 5.)

Part 4. Pre-register conversion as the *primary* metric, judged at $\alpha = 0.05$. Report the twelve secondaries with Benjamini–Hochberg FDR control and label anything found only in segments as *exploratory* — Chapter 13's rules, applied verbatim.

Common mistake

Treating the 3.1-week answer as "we can stop Wednesday of week 4". The horizon was chosen to protect α and power — stopping on schedule is part of the test's validity, not bureaucracy.

Outline

Python lab — run it live in the notebook

Companion notebook: `advanced_01_rcts_lab.ipynb`

The code in this module is meant to be **run, not read**. Open the companion Jupyter notebook and work through it cell by cell:

- §1, *A confounded training programme* — simulate the population; naive vs stratified comparisons;
- §2, *Potential outcomes and selection bias* — the God's-eye table and the decomposition, term by term;
- §3, *Randomisation in action* — sampling distributions and CI coverage;
- §4, *Regression adjustment on Wage* — the semi-synthetic experiment with HC2 errors;
- §5, *Power* — the analytic curve against simulation;
- §6, *Peeking* — the A/A false-positive simulation.

Data loads via the `ISLP` package or the bundled CSV files; the notebook ends with worked solutions to the [Python] exercises and open exercises of its own.

Outline

Module A1 in one slide

What we accomplished

Added the causal question — and the one design that answers it cleanly — to a course that, until now, deliberately only predicted.

- Potential outcomes $Y(1)$, $Y(0)$; the fundamental problem; SUTVA stated honestly.
- Naive comparison = ATT + selection bias (+ differential effect): derived and computed.
- Randomisation zeroes both bias terms; the difference in means *is* the Ch. 0 t-test.
- Regression adjustment (pre-specified, pre-treatment, HC errors) buys precision, not truth.
- Design: $n = 2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2/\delta^2$; blocking; clusters via the design effect.
- Threats: attrition, non-compliance (ITT!), balance, peeking — Ch. 13 in the time dimension.

Key formulas at a glance

Observed outcome $Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$; $\tau_i = Y_i(1) - Y_i(0)$.

Decomposition $\text{naive} = \text{ATT} + \big(E[Y(0) \mid D{=}1] - E[Y(0) \mid D{=}0]\big) = \text{ATE} + \text{sel. bias} + (1 - \pi)(\text{ATT} - \text{ATU})$.

Randomisation $D \perp (Y(1), Y(0)) \;\Rightarrow\; E[Y \mid D{=}1] - E[Y \mid D{=}0] = \tau$.

Estimator $\hat{\tau} = \bar{Y}_1 - \bar{Y}_0$, $\widehat{\text{SE}} = \sqrt{S_1^2/n_1 + S_0^2/n_0}$ (conservative; = Welch t).

Sample size $n_{\text{per arm}} = 2\sigma^2 (z_{1-\alpha/2} + z_{1-\beta})^2 / \delta^2$; $\text{MDE} = (z_{1-\alpha/2} + z_{1-\beta})\, \sigma \sqrt{2/n}$.

Design effect $\text{DE} = 1 + (m - 1)\rho$; $n_{\text{eff}} = n / \text{DE}$.

Non-compliance $\text{ITT} = \bar{Y}_{Z=1} - \bar{Y}_{Z=0}$; $\text{CACE} = \text{ITT}_Y / \text{ITT}_D$.

Vocabulary checklist

Term	Meaning
Potential outcomes $Y(1), Y(0)$	Outcomes under treatment / control, both pre-existing
Fundamental problem	Only one potential outcome ever observed per unit
ATE / ATT / ATU	Average effect: overall / on treated / on untreated
SUTVA	No interference; one version of treatment
Selection bias	$E[Y(0) \mid D{=}1] - E[Y(0) \mid D{=}0]$
Randomisation	Assignment independent of potential outcomes
ITT / per-protocol	Analysis by assignment (valid) / by uptake (biased)
CACE (LATE)	Effect on compliers; Wald ratio $\mathrm{ITT}_Y / \mathrm{ITT}_D$
MDE	Smallest effect a design can detect at target power
Design effect	Information loss from cluster randomisation
Balance check (SMD)	Standardised covariate difference; audit of the randomiser
Peeking	Repeated interim testing; inflates false positives

Decision rules of thumb

- If you can randomise, **randomise** — no model rescues selection you could have designed away.
- Power *before* launch: no MDE, no experiment. $n \propto 1/\delta^2$.
- Pre-register the primary metric, the analysis model, and the horizon.
- Analyse as randomised: **ITT** first; CACE as the labelled follow-up.
- Adjust only for pre-treatment, pre-specified covariates; use HC (robust) standard errors.
- Fix the horizon or use sequential methods — never stop at the first star.
- Groups treated together or interfering? Randomise clusters and pay the design effect honestly.

Common pitfalls

Don't

1. Compare volunteers with non-volunteers and call the difference an effect.
2. Drop non-compliers or dropouts "for a cleaner comparison".
3. Peek daily and stop at the first significant result.
4. Add or remove covariates until the treatment effect is significant.
5. Ignore clustering because "we have lots of rows".
6. Read a non-significant underpowered test as "no effect".
7. Test twenty segments and report the one that won (Chapter 13!).

Self-check questions

Quick self-assessment

1. Why can no dataset, however large, reveal a single τ_i?
2. A naive comparison shows $+6$; the ATT is $+2$. What is the selection-bias term (assume no differential-effect term)?
3. Your RCT's control arm lost 30% of participants, the treatment arm 5%. What is threatened, exactly?
4. Why is the ITT the right number for a rollout decision even at 60% compliance?
5. Your A/B platform shows a live p-value updating hourly. What must the analysis plan say about it?

Answers — compressed

1: both potential outcomes never coexist. 2: $+4$. 3: differential attrition re-introduces selection bias among the observed. 4: rollout can only *offer*; ITT is the offer's effect. 5: it is for monitoring *health*, not for stopping — the decision waits for the pre-set horizon or a

Appendix — what is in here, and why

Nothing in the appendix is needed to follow the module: each slide is a formal derivation, a longer worked exercise, or a side topic. Read it when you want the full story, or when a later module sends you back here.

Contents of the appendix

- **Neyman's variance** — why the difference-in-means SE is conservative; moved out because the main deck only needs the direction of the inequality.
- **Fisher's randomisation test** — exact p-values from re-randomising; a bridge to Chapter 13's permutation tests, beyond the module's core.
- **The Wald/CACE derivation** — the two-line algebra behind $\mathrm{ITT}_Y/\mathrm{ITT}_D$; the main deck states only the result.
- **The design effect, derived** — where $1 + (m-1)\rho$ comes from.
- **Further reading** — the canonical books and papers per topic.

Neyman's variance: why the SE is conservative

Over the randomisation distribution (finite population of n units, n_1 treated),

$$\mathrm{Var}(\hat{\tau}) = \frac{S^2_{(1)}}{n_1} + \frac{S^2_{(0)}}{n_0} - \frac{S^2_\tau}{n}, \qquad S^2_\tau = \frac{1}{n-1}\sum_{i=1}^{n}\left(\tau_i - \bar{\tau}\right)^2,$$

where $S^2_{(1)}, S^2_{(0)}$ are the population variances of $Y(1)$ and $Y(0)$.

How to read this

- The plug-in estimator $S_1^2/n_1 + S_0^2/n_0$ estimates only the first two terms; the third, $S^2_\tau/n \geq 0$, is **unestimable** — it involves the never-observed joint distribution of $(Y(1), Y(0))$.
- Omitting a non-negative term means the estimated variance is too *large*: CIs are honest or wider than needed.
- Equality holds when τ_i is constant — then $S^2_\tau = 0$ and the usual SE is exact.

Takeaway

Fisher's randomisation test (bridge to Ch. 13 resampling)

The sharp null and the exact test

Under H_0: $Y_i(1) = Y_i(0)$ for *every* i, the observed outcomes are fixed regardless of assignment — so the null distribution of $\hat{\tau}$ can be *generated*: re-randomise D many times and recompute the statistic.

How to read this

- Empirical p-value: $p = \dfrac{1 + \#\{|\hat{\tau}^{*b}| \geq |\hat{\tau}|\}}{1 + B}$ over B re-randomisations — the same construction as Chapter 13's permutation p-values; the "+1" keeps it strictly positive.
- Exact for any n, any outcome distribution — no normality, no large-sample appeal; the randomisation itself is the probability model.
- Note the null is *sharp* (no effect for anyone), stronger than Neyman's average null $\tau = 0$.

Takeaway

The Wald/CACE estimator, derived

Let Z be the randomised *offer* and $D(z)$ the treatment actually taken. With one-sided non-compliance ($D(0) = 0$: no access without the offer) and exclusion (Z affects Y only through D):

$$\begin{aligned}\mathrm{ITT}_Y &= E[Y \mid Z{=}1] - E[Y \mid Z{=}0] \\ &= E\big[D(1)\,(Y(1) - Y(0))\big] = \Pr\big(D(1) = 1\big)\; E\big[Y(1) - Y(0) \mid D(1) = 1\big].\end{aligned}$$

How to read this

First equality: randomisation of Z. Second: under $Z = 1$ only the takers ($D(1) = 1$) have their outcome moved; never-takers contribute zero by exclusion. Dividing by the compliance share $\Pr(D(1) = 1) = \mathrm{ITT}_D$ gives

$$\boxed{\mathrm{CACE} = \frac{\mathrm{ITT}_Y}{\mathrm{ITT}_D}}$$

— the effect *on compliers*, not on everyone. With two-sided non-compliance the same ratio identifies the LATE under the extra assumption of monotonicity (no defiers).

The design effect, derived

Cluster of size m, outcomes equicorrelated with intraclass correlation ρ (variance σ^2). The variance of the cluster mean is

$$\mathrm{Var}(\bar{Y}_{\text{cluster}}) = \frac{1}{m^2}\Big[m\,\sigma^2 + m(m-1)\,\rho\,\sigma^2\Big] = \frac{\sigma^2}{m}\,\big[1 + (m-1)\rho\big].$$

How to read this

- The bracket is the **design effect**: the factor by which the variance exceeds that of m *independent* observations.
- $\rho = 0$: DE $= 1$, clustering costless. $\rho = 1$: DE $= m$ — each cluster is effectively *one* observation, however many members it has.
- Sample-size practice: compute n as usual, then multiply by DE; or divide the achieved n by DE to get the effective sample.

Takeaway

Even tiny within-cluster correlation bites at scale: at $m = 50$, $\rho = 0.05$ triples the required sample (DE $= 3.45$). Ignoring it produces standard errors that are confidently, quietly wrong.

Further reading

Where to go next

- **Foundations** — Imbens & Rubin (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press: the potential-outcomes canon.
- **Field experiments** — Gerber & Green (2012), *Field Experiments: Design, Analysis, and Interpretation*, W. W. Norton; Duflo, Glennerster & Kremer (2007), "Using Randomization in Development Economics Research", *Handbook of Development Economics*.
- **Online experiments** — Kohavi, Tang & Xu (2020), *Trustworthy Online Controlled Experiments*, Cambridge University Press: peeking, CUPED, guardrails, at industrial scale.
- **Regression adjustment** — Lin (2013), "Agnostic notes on regression adjustments to experimental data", *Annals of Applied Statistics*.
- **Sequential testing** — Johari et al. (2017), "Peeking at A/B tests", *KDD*: always-valid inference.