

Type 2 Error Statistics:

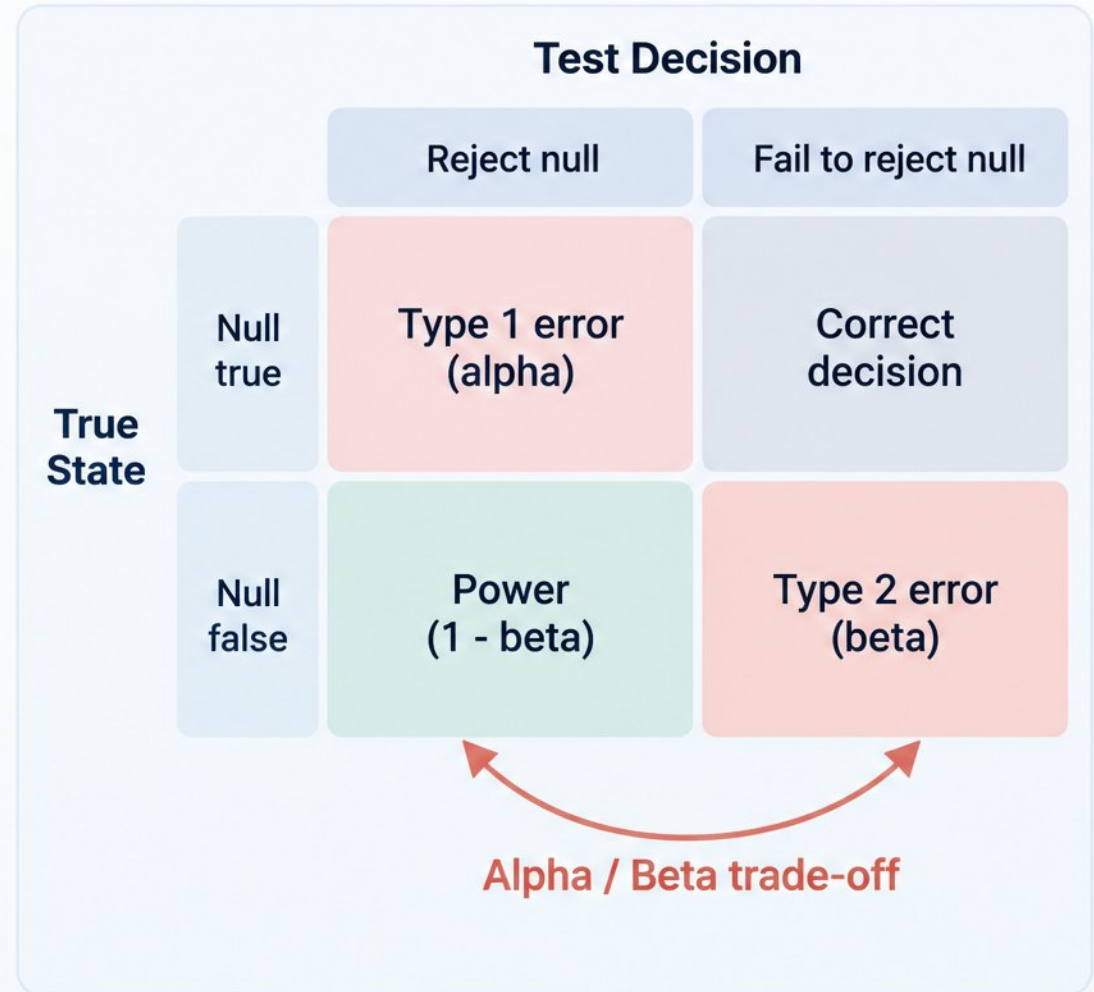
Introduction and Why Missed Effects Matter

- Type 2 error (beta): failing to reject a false null; a false negative.
- Four outcomes: correct non-rejection, Type 1 error, Type 2 error, correct rejection (power).
- Beta is not fixed like alpha; it depends on the specific true alternative.
- Practical cost: abandoned treatments, false 'no relationship' claims, shelved improvements, missed safety signals.
- Worked example: effective drug yields $p = 0.08$ vs $\alpha = 0.05$, concluding 'ineffective'.
- A p-value is not the probability of a Type 2 error.



The Decision Framework: Alpha, Beta, and Power

- Power = 1 - beta: correctly rejecting a false null
- 80% power means beta = 0.20, a four-to-one trade-off
- Alpha and beta are not complements
- Lowering alpha from 0.05 to 0.01 raises beta
- Four levers: sample size, effect size, alpha, variability
- Power is always stated for a specific alternative



Levers for Greater Power



Sample size



Effect size

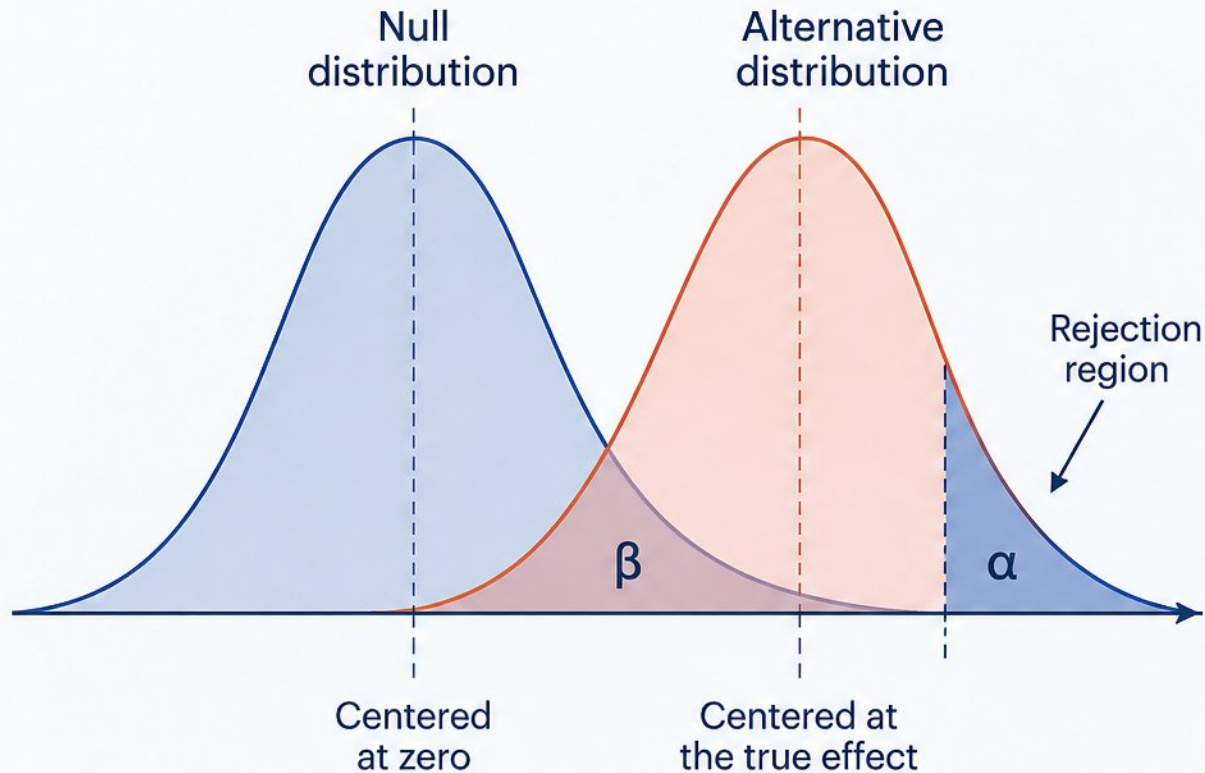


Alpha



Variability

Visualizing Beta: Null and Alternative Sampling Distributions



- Two overlapping curves: null centered at zero, alternative at the true effect
- Rejection region fixed by alpha; beta is the alternative's area inside non-rejection
- Larger samples narrow both curves, shrinking overlap and lowering beta
- Bigger true effects shift the alternative right, cutting overlap and raising power
- Power falls sharply as the true effect approaches the null value

Factors That Move Type 2 Error Risk

- Sample size: the most controllable lever; more observations shrink standard error and beta
- Smaller true effects are harder to detect; power collapses near the null value
- High outcome variability inflates standard error and silently raises false negatives
- Raising alpha adds power but sacrifices false-positive control
- One-tailed tests gain power only if opposite-direction effects count as no effect

Sample size



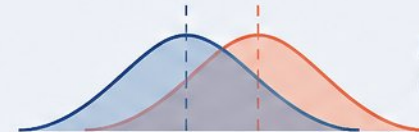
More separation
(lower beta)

Effect size



More separation
(lower beta)

Outcome variability



More overlap
(higher beta)

Alpha level

α



More separation
(higher power)

One-tailed test



More separation
(higher power)

A Priori Power Analysis and Sample Size Planning



Fix any three: effect size, alpha, power, sample size—the fourth is determined



Anchor assumed effect on SESOI or policy relevance, never on convenient N



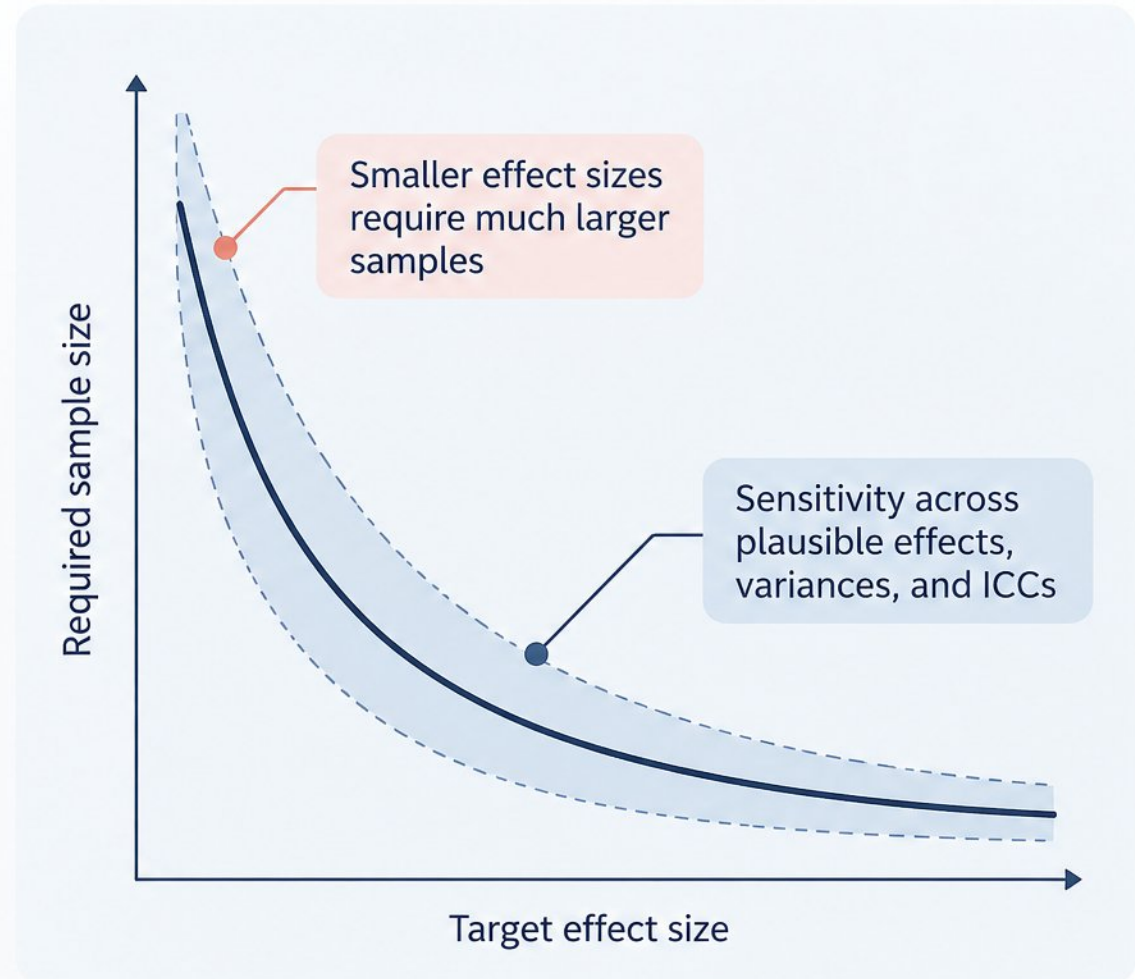
Halving the target effect roughly quadruples the required sample size



Inflate for attrition and clustering: 20% dropout means recruiting 1.25x



Report a sensitivity curve across plausible effects, variances, and ICCs



Calculating and Estimating Beta in Practice



Closed-form power formulas cover z-tests, t-tests, proportions, and ANOVA



Complex designs — multilevel, repeated measures, clustered — usually need simulation



Tools: G*Power, R (pwr, Superpower), Python, online calculators



Minimum detectable effect (MDE): smallest effect your affordable sample can detect



Pilot effects are inflated; use the lower confidence bound



Uncertainty is real — report an MDE curve, not one power number

Methods and Tools for Power Analysis



Analytical Formulas



G*Power



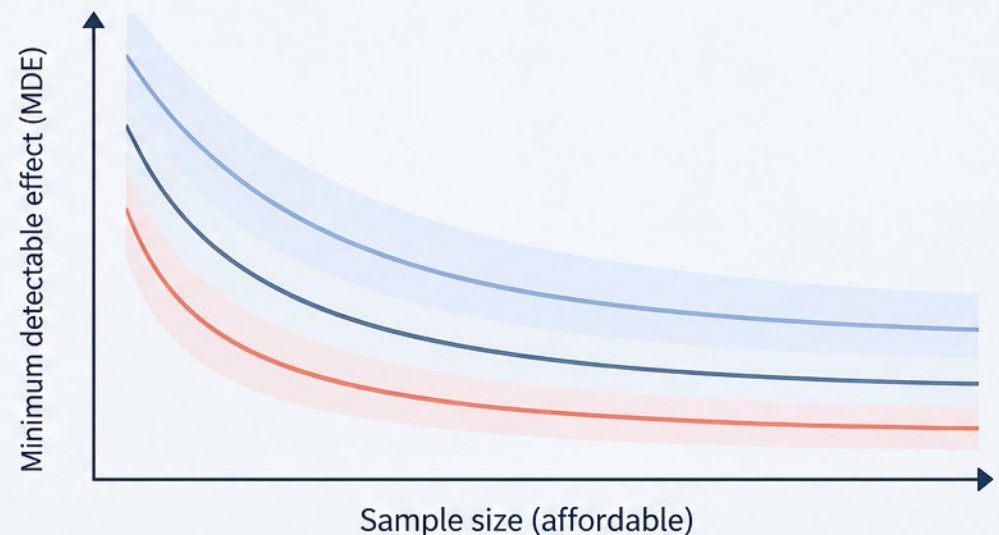
R (pwr, Superpower), Python



Simulation

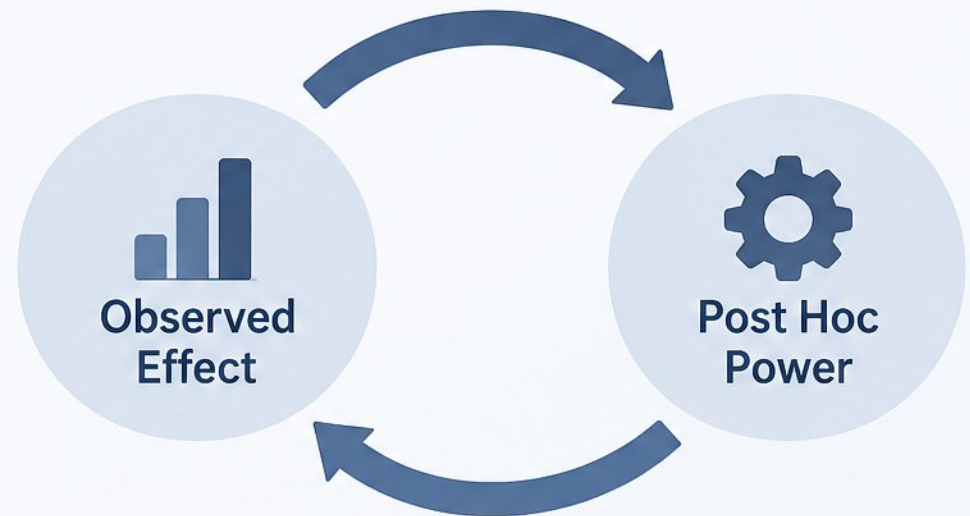
Minimum Detectable Effect (MDE) Curve

Smaller effects become detectable with larger sample sizes.



Post Hoc Power and Observed Power: Use With Caution

- Post hoc power is just a function of the observed p -value
- It adds no new information to the analysis
- Always below 50% when p exceeds alpha
- Assumes the observed effect equals the true effect
- A high value is not an upper bound on true effects
- Prefer confidence intervals, equivalence tests, and sensitivity analysis
- Power belongs at the design stage, not after data collection



**Always below 50%
when p exceeds alpha**

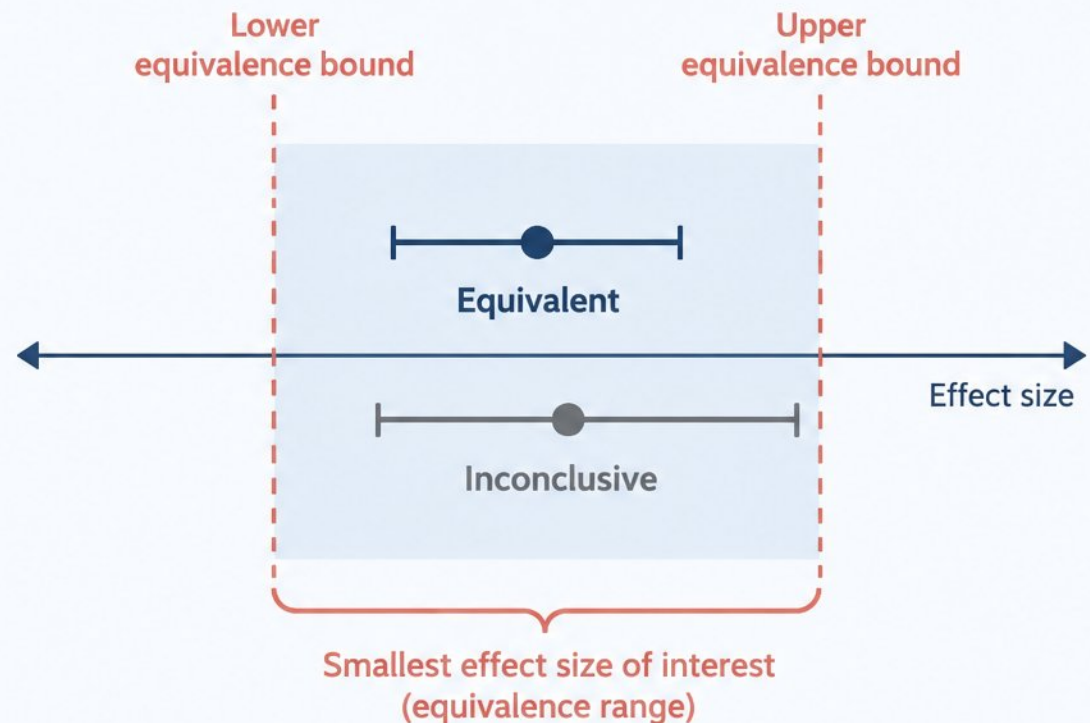


Interpreting a Non-Significant Result Without Overclaiming

- 1 Absence of evidence is not evidence of absence.
- 2 Null results have many causes: no effect, small effect, unlucky sample, noise, confounding.
- 3 Equivalence testing (TOST) sets bounds from the smallest effect size of interest.
- 4 A 90% CI inside the equivalence bounds supports practical equivalence.
- 5 Report the equivalence range explicitly; distinguish "equivalent" from "undetermined."

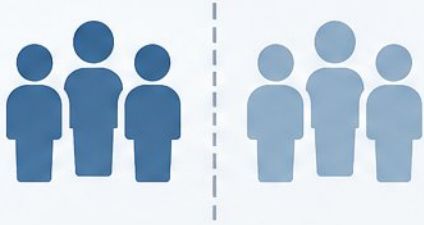
Equivalence Testing (TOST)

The smallest effect size of interest defines equivalence bounds. A 90% CI inside the bounds supports practical equivalence.



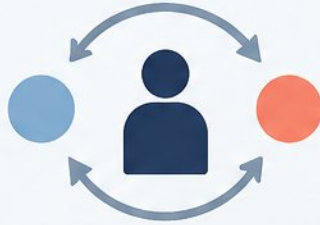
Study Design Choices That Reduce Missed Effects

Between-Subjects Design



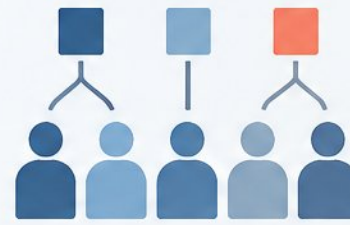
- ↑ Higher variance
- ↑ Higher standard error
- ↓ Lower power (higher beta)

Within-Subjects Design



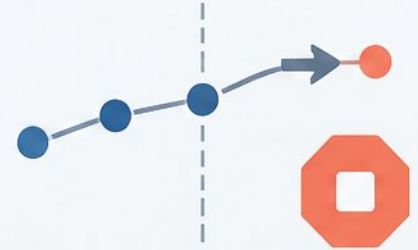
- ↓ Lower variance
- ↓ Lower standard error
- ↑ Higher power (lower beta)

Stratified Randomization, ANCOVA, and Repeated Measures



- ↓ Lower variance
- ↓ Lower standard error
- ↑ Higher power (lower beta)

Sequential Designs with Alpha-Spending



Allows earlier stopping safely

- Paired and within-subject designs cut between-person variance, boosting power
- Stratified randomization, ANCOVA, and repeated measures add precision without inflating alpha
- Cut measurement error and tighten manipulation to enlarge effect size
- Sequential designs with alpha-spending allow earlier stopping safely
- Clustering, unequal allocation, and multiple outcomes erode effective sample size



Type 2 Errors in A/B Tests and Product Experiments



Underpowered tests are the top cause of 'no result' — and real wins die quietly



Set MDE from the smallest lift worth shipping, then compute sample



Never reverse-engineer MDE from the traffic you happen to have



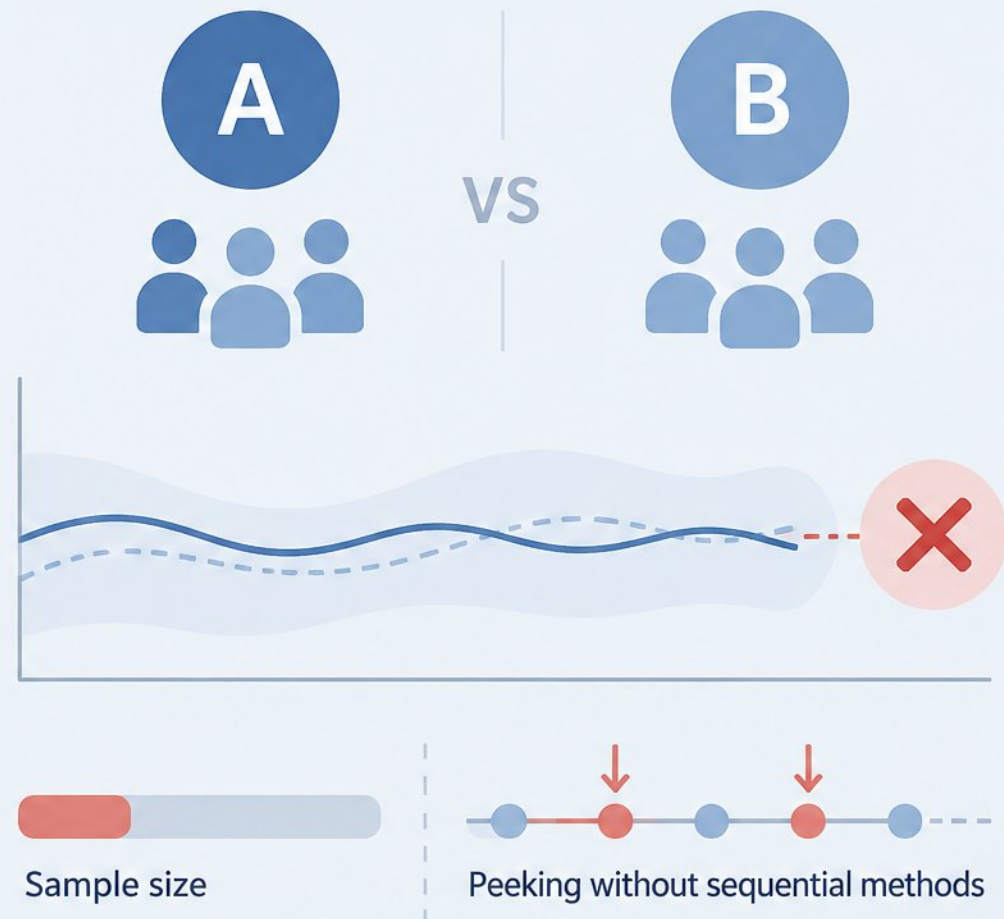
Peeking without sequential methods inflates false positives and skews power



Guardrails, novelty, and delayed effects can hide true business impact



A non-significant result from an underpowered test is silence, not evidence



Missed Effects in Clinical, Social, and Policy Research



Psychology studies average power near 0.35–0.50, fueling the replication crisis



Clinical trials: missed benefits or harms carry ethical and regulatory consequences



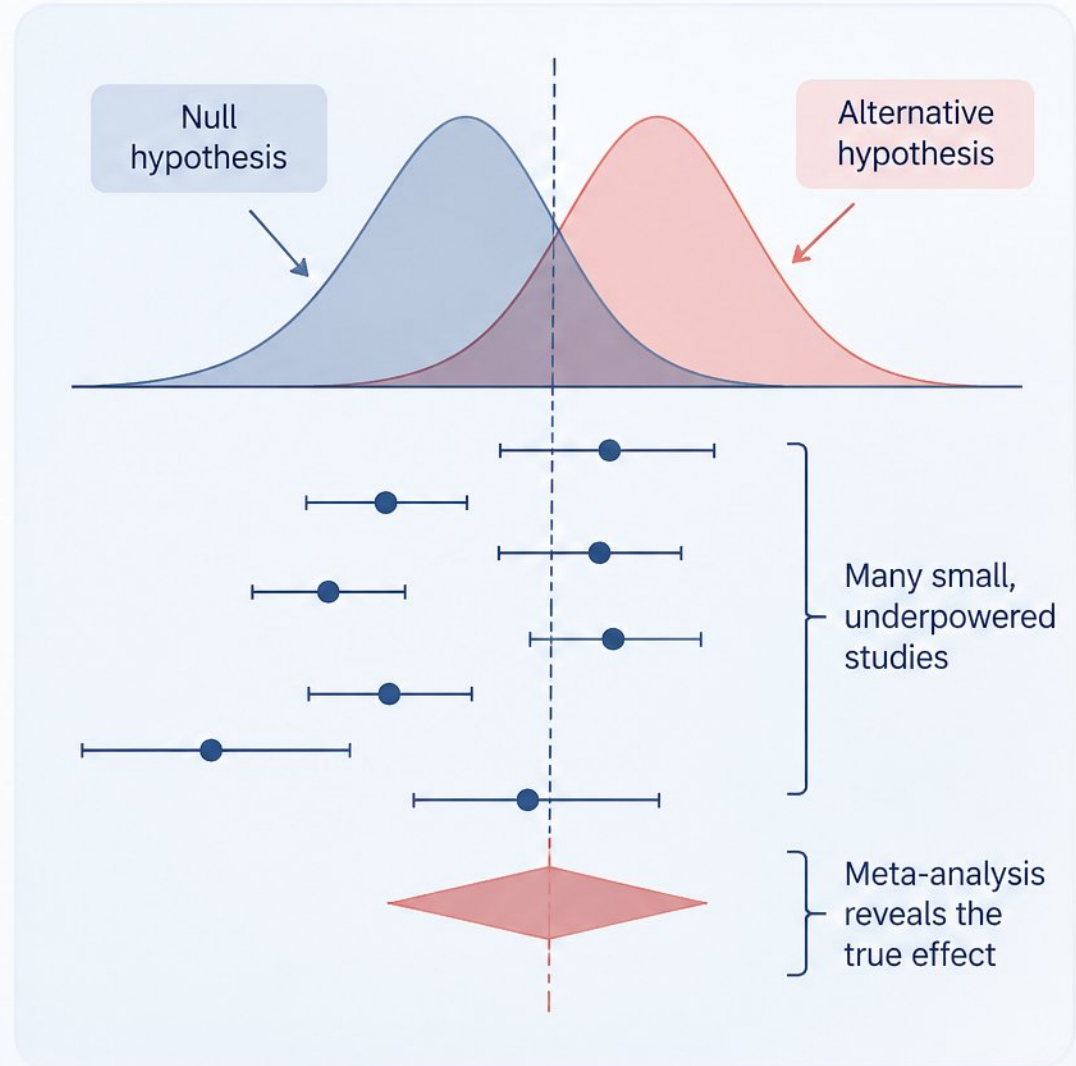
Replication samples ignore effect-size uncertainty, so replication rates fall short



Meta-analysis and cumulative synthesis rescue underpowered individual studies



Inflated published effects spawn a cycle of still-underpowered research



STRONGER EVIDENCE
BRIGHTER DECISIONS

Teaching and Communicating Type 2 Error and Power



Common misconceptions:
beta equals alpha; non-significance
proves the null



Interactive simulations beat
formulas for building power
intuition



Diagnostic questions with
targeted feedback confront
misconceptions directly



Use analogies carefully; check
which error matters more in context



For non-technical audiences, say:
“the chance we would miss a real
effect of this size”

Explore with an Interactive Simulation

Mean



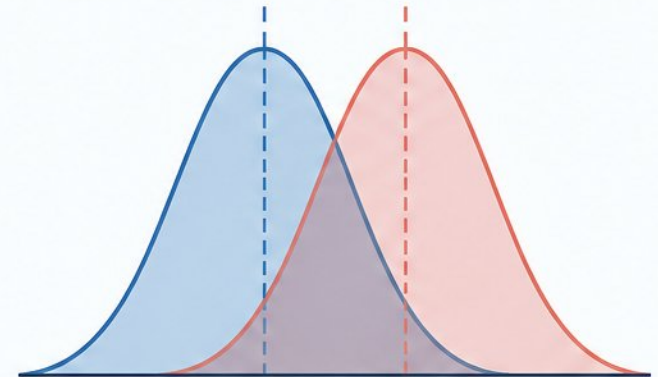
Standard deviation



Sample size



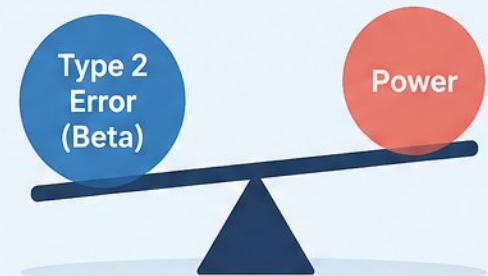
Alpha



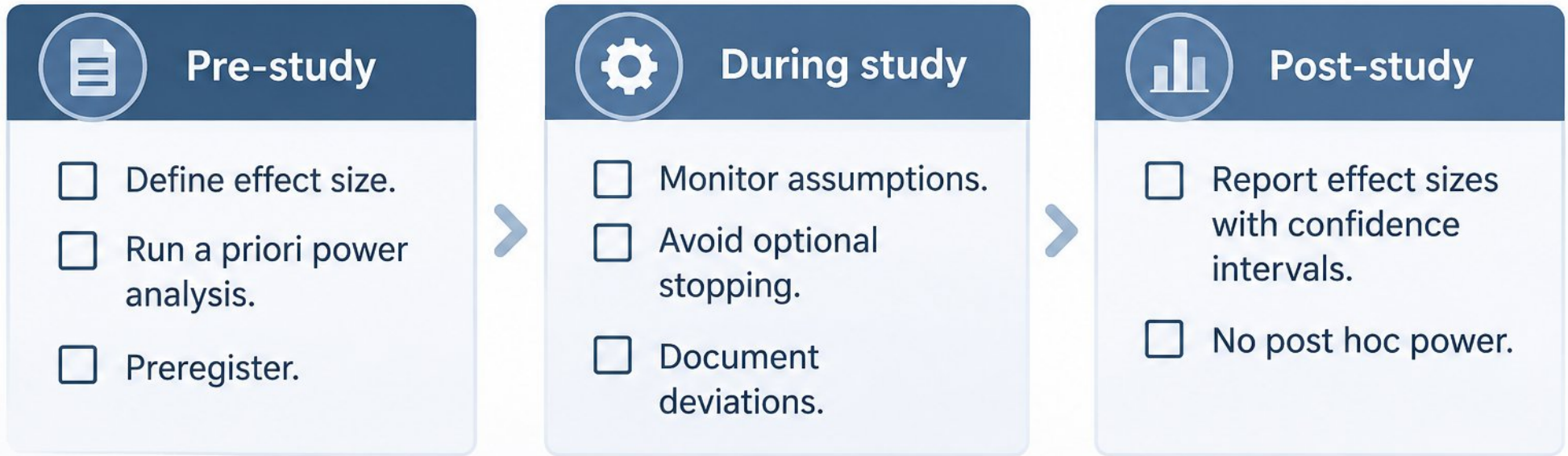
Check Your Understanding

Match each misconception
to the best response.

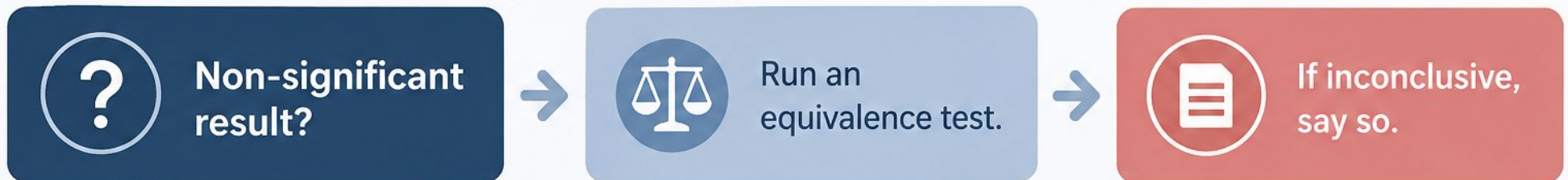
The Power / Beta Trade-off



Reporting and Decision Guide for False Negatives



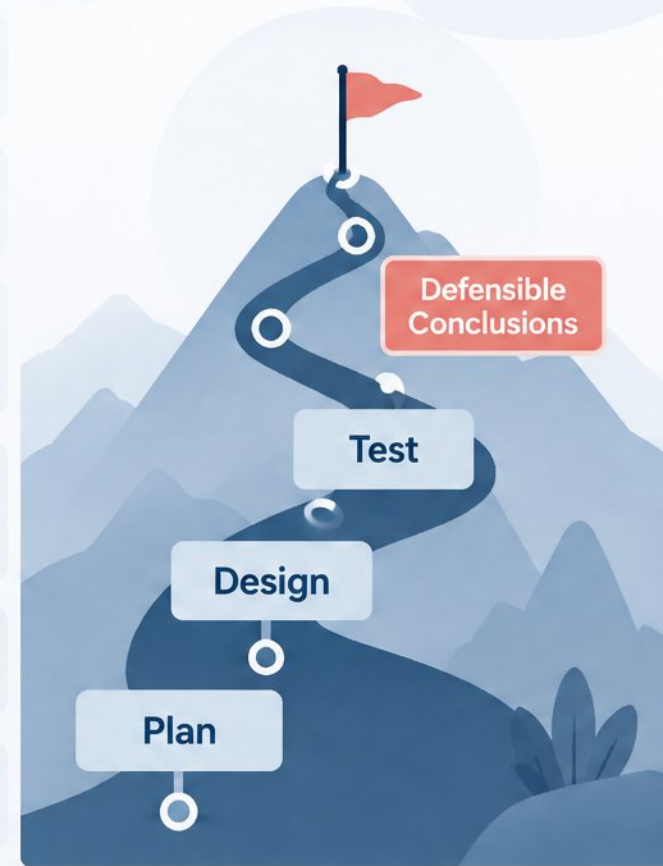
 **CONSORT 2025:** prespecify outcomes, separate prespecified from post hoc analyses.



Key Takeaways and Further Learning

Better
Questions
Stronger
Evidence

- 1 Type 2 error is a design problem, not an analysis problem
- 2 Its main controllable driver: sample size relative to effect size
- 3 Read non-significant results through power, MDE, and confidence intervals
- 4 Equivalence testing turns an ambiguous null into a defensible claim
- 5 Next steps: simulation tools, an equivalence test, a team MDE policy
- 6 Keep learning: G*Power, pwr, TOSTER, Lakens, CONSORT 2025



Further Learning Resources



G*Power



pwr



TOSTER



Lakens



CONSORT 2025

PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/1108a8c7/public