Power Analysis for Two-Group Experiments

Sample Size, Power, and MDES — Concepts and Practice

Justin S. Eloriaga

September 15, 2025

ECON 521

Roadmap

Why Power Analysis?

Set-up and Assumptions

Power, Sample Size, MDES

Design Choices that Matter

Validation and Reporting

Figures from the Notebook

Further Reading

Why Power Analysis?

What power analysis answers

- **Sample size**: How many units are needed to detect a target effect?
- **Power**: With a given n, what is the probability of detecting a true effect?
- **MDES**: With a given n, what is the smallest effect we can reliably detect?
- **Design tradeoffs**: Tails (one- vs two-sided), allocation ratio, clustering, covariates, multiple tests, attrition, noncompliance.

Key definitions (two-group difference in means)

- **Type I error** α: false positive rate (often 0.05).
- **Type II error** β: false negative rate.
- **Power** $= 1 - \beta$: probability of rejecting H_0 when the effect is real.
- **Effect size** (Cohen's d): $d = \dfrac{\mu_1 - \mu_0}{\sigma}$ (standardized difference).
- **MDES**: the minimum d detectable at chosen α, power, and design.

Set-up and Assumptions

Canonical set-up (independent samples t-test)

- Two groups: treatment (n_1) and control (n_0); independent sampling.
- Outcome approximately normal or n large (CLT).
- Often assume *equal variances*; if not, use Welch's t or plan for variance heterogeneity.
- **Two-sided** vs **one-sided** alternative must be justified *a priori*.

Effect sizes: standardized and natural units

- **Standardized** (d) is portable across measures; helpful for planning and comparisons.
- **Natural units** (e.g., percentage points, dollars) matter for interpretation and policy relevance.
- Always translate between the two: choose a substantively meaningful target effect, then map to d via an anticipated σ.

Power, Sample Size, MDES

Determinants of power (intuition)

- **Larger effects** $\Rightarrow$ higher power.
- **Larger samples** $\Rightarrow$ smaller SEs $\Rightarrow$ higher power.
- **One-sided** tests (when justified) $\Rightarrow$ higher power than two-sided.
- **Less noise** (variance reduction via design or covariates) $\Rightarrow$ higher power.
- **Lower** α (stricter) $\Rightarrow$ lower power, all else equal.

Back-of-the-envelope sample size (equal n per group)

For small-to-moderate effects and two-sided α:

$$n_{\text{per group}} \approx \frac{2\,(z_{1-\alpha/2} + z_{1-\beta})^2}{d^2}$$

- d = target standardized effect; z_p = normal quantile.
- **Implication**: halving d *quadruples* sample size.

MDES: minimum detectable effect

- With fixed n, α, and power, MDES solves the same equation for d.
- **Diminishing returns**: doubling n reduces MDES, but not linearly.
- **Communicate both**: report MDES in d and in the outcome's natural units.

Visuals to include from the notebook

- **Power curves**: power vs. per-group n for several d (e.g., $0.10, 0.20, 0.50$).
- **MDES curve**: MDES vs. per-group n at fixed α and target power.
- **Sampling distributions schematic**: overlap of null vs alternative to illustrate power.

Design Choices that Matter

Two-sided vs one-sided tests

- **Two-sided** is default when effects could plausibly be in either direction, or for conservative inference.
- **One-sided** gains power but must be justified before seeing data; direction must be theoretically constrained.
- Switching post hoc invalidates error control.

Allocation ratio and costs

- Equal allocation ($n_1 = n_0$) maximizes power for a fixed total N when costs are equal.
- **Unequal allocation** (e.g., scarce treatment or differential costs) reduces power for fixed N; plan larger N.
- If control outcomes are cheaper/easier to collect, a modestly larger control group can be cost-effective.

Variance assumptions and measurement

- If variances differ ($\sigma_1^2 \neq \sigma_0^2$), use robust methods (Welch) and plan with conservative variance.
- Reduce variance via better measurement, careful protocols, and pre-specification of outcome definitions.
- Consider transformations or rank-based tests if distributions are heavy-tailed.

Clustering and the design effect

- In cluster settings (classrooms, clinics), outcomes within clusters are correlated.
- **Design effect:** $DE = 1 + (m - 1)\rho$ where m = average cluster size, ρ = ICC.
- Effective sample size: $n_{\text{eff}} = n/DE$. **Plan** to *inflate* n by DE or design at the cluster level.

Covariate adjustment and pre-post (ANCOVA)

- Adjusting for predictive baseline covariates reduces residual variance.
- Rough rule: variance scales by $(1 - R^2)$ where R^2 is from regressing the outcome on covariates.
- Pre-post designs: using baseline outcome as a covariate (ANCOVA) typically outperforms change scores.

Multiple testing and families of outcomes

- Testing many hypotheses inflates the chance of at least one false positive.
- Control familywise error (Bonferroni/Holm) or false discovery rate (Benjamini–Hochberg).
- **Planning implication**: per-test power falls under correction $\Rightarrow$ larger n or fewer primary outcomes.

Attrition and noncompliance

- **Attrition**: anticipate loss-to-follow-up; inflate planned n accordingly.
- **Noncompliance**: ITT effects are diluted relative to treatment-on-the-treated; plan for reduced detectable effects.
- Mitigation: strong tracking, incentives, clear protocols, and pre-registered handling of missing data.

Validation and Reporting

Why validate with simulation (conceptually)

- **Cross-check** analytic power under your exact design features (tails, allocation, clustering, covariates).
- **Stress-test** against plausible deviations: higher variance, lower ICC, heavier tails, unequal n.
- **Document** that simulated rejection rates align with target power near the chosen n.

What to report (checklist)

- Hypotheses, test type (two-sample t or variant), tails, α, target power.
- Target effect in **natural units** and in **standardized** d; source/justification for σ.
- Allocation ratio, anticipated attrition, clustering (ICC, cluster sizes), covariates assumed and expected R^2.
- Any multiplicity adjustments and identification of **primary** vs **secondary** outcomes.
- Figures: power curves, MDES vs n, design diagram; brief note on simulation validation.

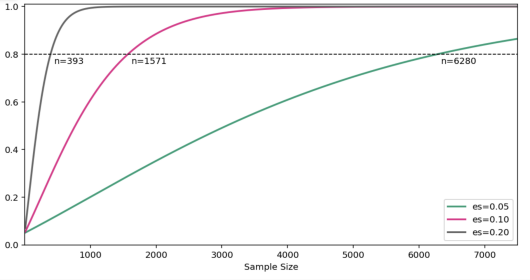
Common pitfalls & remedies

- **Over-optimistic σ or R^2**: use conservative estimates or sensitivity ranges.
- **Post hoc one-sided switch**: pre-specify tails to preserve validity.
- **Ignoring clustering**: even small ICCs can have large effects when clusters are big.
- **Outcome creep**: too many outcomes $\Rightarrow$ low per-test power; prioritize.
- **No attrition buffer**: build realistic margins and tracking plans.

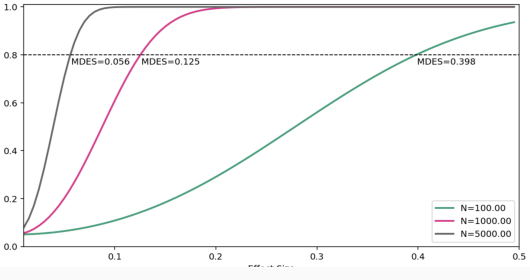
Figures from the Notebook

Figures

Power curves



MDES vs *n*



Further Reading

Further reading (accessible)

- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.).
- Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit.
- Lakens, D. (2022). Sample size justification. *Collabra: Psychology*.
- Gelman, A., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*.
- List, J. et al. (2022). *The Voltage Effect* (for the "meaningful effect" perspective).

The end

Questions?