Program Evaluation and Randomized Experiments

Chris Conlon

Tuesday 14th October, 2025

Applied Econometrics

Overview

This set of lectures will cover (roughly) the following papers:

Theory:

- Angrist and Imbens (1994)
- Heckman Vytlacil (2005/2007)
- Abadie and Imbens (2006)

And draw heavily upon notes by

- Guido Imbens
- Richard Blundell and Costas Meghir

The Evaluation Problem

- The issue we are concerned about is identifying the effect of a policy or an investment or some individual action on one or more outcomes of interest
- This has become the workhorse approach of the applied microeconomics fields (Public, Labor, etc.)
- Examples may include:
  - The effect of taxes on labor supply
  - The effect of education on wages
  - The effect of incarceration on recidivism
  - The effect of competition between schools on schooling quality
  - The effect of price cap regulation on consumer welfare
  - The effect of indirect taxes on demand
  - The effects of environmental regulation on incomes
  - The effects of labor market regulation and minimum wages on wages and employment

Potential Outcomes

- Consider a binary treatment $D_i \in \{0, 1\}$.
  - Some people use $T_i \in \{0, 1\}$ instead.
- We observe the outcome Y_i. But there are two potential outcomes
  - $Y_i(1)$ the outcome for i if they are treated.
  - $Y_i(0)$ the outcome for i if they are not treated (control).
- We are generally interested in $\beta_i \equiv Y_i(1) - Y_i(0)$ which we call the treatment effect.
  - Individuals have heterogeneous treatment effects.
  - In an ideal world we could fully characterize $f(\beta_i)$

Some Challenges

Stable unit treatment value assumption (SUTVA)

- We assume a *ceteris paribus* version of treatment effects
- We need β_i to be a policy invariant (structural) parameter.
- Your β_i doesn't respond to whether or not another individual is treated.
- Two common limitations:
  - Peer effects: Whether you respond to job training program depends on whether your spouse is also treated.
  - Equilibrium effects: if we sent everyone to college, returns to college would be quite different.

Some Challenges

Fundamental Problem of Causal Inference

- We don't observe the counterfactual $Y_i(D_i)$.
- For a single individual we either observe $Y_i(1)$ or $Y_i(0)$ but never both!
  - ex: We don't see what your wage would have been if you didn't attend college.
  - ex: We might know your cholesterol before you took Lipitor, but we don't know what it would be today if you didn't take Lipitor.

$$Y_i = D_i \cdot Y_i(1) + (1 - D_i) \cdot Y_i(0)$$

i	$Y_i(1)$	$Y_i(0)$	D_i	D_i
1	1	?	1	1
2	0	?	1	0
3	?	0	0	0
		$\vdots$		
n	?	1	0	1

Structural vs. Reduced Form

- Usually we are interested in one or two parameters of the distribution of β_i (such as the average treatment effect or average treatment on the treated).
- Most program evaluation approaches seek to identify one effect or the other effect. This leads to these as being described as reduced form or quasi-experimental.
- The structural approach attempts to recover the entire joint $f(\beta_i, \varepsilon_i)$ distribution but generally requires more assumptions, but then we can calculate whatever we need.

Treatment Effects Parameters

Most approaches to estimating treatment effects will recover some moments of $f(\beta_i)$ instead of the entire distribution

Average Treatment Effect (ATE) corresponds to $\mathbb{E}[\beta_i]$.

Average Treatment on Treated (ATT) corresponds to $\mathbb{E}[\beta_i \mid D_i = 1]$.

Average Treatment on Control/Untreated (ATUT) corresponds to $\mathbb{E}[\beta_i \mid D_i = 0]$.

We also have that if the probability of treatment Pr$(D_i = 1) = \pi$

$$ATE = \pi \cdot ATT + (1 - \pi) \cdot ATUT$$

The Selection Problem

- Let's start with the easy cases: run OLS and see what happens.

$$Y_i = \alpha + \beta_i \cdot D_i + \varepsilon_i$$

- OLS compares mean of treatment group with mean of control group (possibly controlling for other X)

$$\begin{aligned}\beta^{OLS} &= \mathbb{E}(Y_i \mid D_i = 1) - \mathbb{E}(Y_i \mid D_i = 0) \\ &= \underbrace{\mathbb{E}[\beta_i \mid D_i = 1]}_{\text{ATT}} + \left(\underbrace{\mathbb{E}[\varepsilon_i \mid D_i = 1] - \mathbb{E}[\varepsilon_i \mid D_i = 0]}_{\text{selection bias}}\right)\end{aligned}$$

- Even in absence of heterogeneity $\beta_i = \beta$ we can still have selection bias.
- $Y_i^0 = \alpha + \varepsilon_i$ may vary within the population (this is quite common).

Why worry about selection?

Unless we have random assignment...

$$Y_i = \alpha + \beta_i D_i + u_i$$

- People often choose D_i with β_i in mind.
- The problem: $D_i \perp u_i$ and/or $D_i \perp \beta_i$ are likely violated.
- We can get positive or negative selection bias:
  - e.g. Who goes to college? those likely to benefit more than most!
  - e.g. who gets risky surgeries/drugs? people who are very sick.

The Magic of Randomization

$$\begin{aligned}\mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 0\right] &= \\ \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 1\right] &+ \underbrace{\mathbb{E}\left[Y_i(0) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 0\right]}_{\text{selection bias}}\end{aligned}$$

- If treatment is assigned randomly, then $Y_i(0)$ and T_i should be independent. Consequently,

  $$\mathbb{E}\left[Y_i(0) \mid D_i = 1\right] = \mathbb{E}\left[Y_i(0) \mid D_i = 0\right],$$

  recalling that if $Y_i(0)$ and D_i are independent, $\mathbb{E}\left[Y_i(0) \mid D_i\right] = \mathbb{E}\left[Y_i(0)\right]$.
- Thus, randomization of treatment eliminates selection bias.

The Magic of Randomization II

- We've just shown that randomization gives us the average effect of treatment on the treated (ATT) without selection bias.

$$\mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 0\right] = \mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 1\right]$$

- Furthermore, randomization of treatment also implies that the ATT equals the ATE. If D_i is independent of $Y_i(0)$ and $Y_i(1)$, then

$$\begin{aligned}\mathbb{E}\left[Y_i(1) \mid D_i = 1\right] - \mathbb{E}\left[Y_i(0) \mid D_i = 1\right] &= \mathbb{E}\left[Y_i(1)\right] - \mathbb{E}\left[Y_i(0)\right] \\ &= \mathbb{E}\left[Y_i(1) - Y_i(0)\right].\end{aligned}$$

What Does Randomization Do?

1. Randomization of treatment eliminates selection bias.
2. Randomization of treatment ensures that the ATE=ATT.

Randomization and Endogeneity I

- Selection bias has to do with the fact that *baseline* outcomes for the treated and untreated groups may differ. Example: schoolchildren who get the treatment of having small class sizes (private schools) are also children who have access to private tutors and well-educated parents.
- This is a version of an *endogeneity* problem, and it's a fundamental problem for causal inference. Randomizing treatment solves the problem, for it means that baseline outcomes should no longer be correlated with the treatment.

Randomization and Endogeneity II

- Simplifying the situation by assuming $\beta = Y_i(1) - Y_i(1)$ for all i, we can put this back in the regression equation framework,

$$Y_i = \beta_0 + \beta_1 \, D_i + \varepsilon_i$$

where $\beta_0 = \mathbb{E}\left[Y_i(0)\right]$ and $\varepsilon_i = Y_i(0) - \beta_0$.

- If heterogeneity in baseline outcomes $Y_i(0)$ is correlated with treatment status D_i, then the error term ε_i is correlated with the regressor ε_i, violating the strict exogeneity assumption, and leading to biased estimates of β.
- In theory, randomization of regressors is a way of ensuring that the strict exogeneity assumption holds: we randomly control D_i so that it won't be related to whatever is in ε_i.

Randomization and Endogeneity III

- In practice, randomization doesn't *guarantee* there is no endogeneity problem:
  - Random control of treatment may be imperfect: attrition and manipulation.
  - Treatment status may directly influence ε in some cases: placebo effects and behavioral responses to treatment.
  - Blind and double-blind studies aim to mitigate these concerns.
- This requires more notation to discuss within the potential outcome framework – we need to distinguish between whether subject *believes* they are being treated or not, as well as whether they are *actually* being treated or not – but it is easier to talk about in the regression framework.

Randomization and Endogeneity IV

- What does it mean for the disturbance to be influenced by a regressor? Can't we just consider this an effect of the regressor?

- It depends: often there will be an external validity issue, in the sense that the effect won't extrapolate outside of the study.

- Examples:
  - With placebo effect, we could understand this as treatment influencing the disturbance term, but maybe we're happy to enjoy the placebo effect. On the other hand, what if we're considering whether to add a nutrient to a pill or not, rather than deciding whether or not to give the subjects a pill or not? Perhaps we won't get the placebo effect twice.
  - If treatment influences error term because of manipulation on the part of the researcher implementing the study, then the relationship between treatment and outcomes we see typically won't extrapolate outside of the study. **Very important to avoid this!**

Randomization and Heterogeneity

- The fact that ATE$\neq$ ATT is a separate issue having to do with *heterogeneity* of treatment effects. This issue is also “solved” by randomization in a way, but is this an issue we want to solve?

- Example: the people who take anti-depressants benefit more than the people who don’t. The ATE for a given drug in the whole population might be low, but that doesn’t mean the drug is ineffective. If the ATE within the group of people diagnosed with depression is high, and the people who end up taking the drug fall within that group, the ATT in practice might correspond closely to the sub-population ATE.

- Bottom line: differences between ATE and ATT don’t reflect problems of causal inference, but they reflect the importance of **context** and understanding the population of interest. There’s a reason clinical trials for new cancer drugs typically focus on people that have cancer.

Experiment with Regression Controls

- Consider the model

$$Y_i = \beta_0 + \beta_1 D_i + \beta_2' \mathbf{X}_i + \varepsilon_i$$

  where D_i is assigned randomly.

- Does controlling for **X** matter in the experimental context? Note that even if **X** is omitted from the regression model, there is **no problem of omitted variables bias** because D and **X** are uncorrelated.

- However, controlling for covariates may improve precision of the estimates, especially in small sample sizes where **X** may not be balanced across the treatment and control groups.

Randomization with Small Samples

- In a finite sample, there's always a chance that we end up with subjects that look very different across the treatment and control groups.
  - For observable characteristics, it is customary to check that the two groups have similar means and medians. This is often Table 1 in experimental papers.
- What would you do if you randomly assigned subjects to the two groups, and, before running the experiment, you notice that characteristics are not balanced? Would it be bad to re-randomize?
- Related ideas:
  - Student (1938), the t-test guy, argued against randomization in agricultural trials. Simple random sampling vs. systematic sampling.
  - Stratified sampling, clustered sampling, sampling theory.
  - Chassang et al (2012) explore the idea of **selective trials**.

Experiments in the Social Science

- Randomized Controlled Trials have long been the gold standard for research in the natural sciences.
- In social sciences, many important questions don't lend themselves well to randomization.
  - Macroeconomic policy, other large-scale policy issues, especially when there are spillovers across markets/countries. E.g., what is global impact of EU's decision to implement carbon pricing?
  - Mergers and antitrust, other situations where regulatory questions highly context-specific.
- However, experiments are becoming increasingly popular in some fields
  - Lab experiments: behavioral economics
  - Field experiments: development, labor, education

Example: Banerjee et al (2007)

- Background: getting kids into schools in India seemingly had unimpressive impacts on educational attainment. School quality (educational inputs) is also important.
- Experimental treatment: remedial education. Third and fourth grade students identified as at risk for falling behind are assigned an extra teacher for two hours/day.
    - Group A (50% of schools): third grade classrooms treated in 2001-2, fourth grade classrooms treated in 2002-3
    - Group B (50% of schools): fourth grade classrooms treated in 2001-2, third grade classrooms treated in 2002-3.
- Note that this experiment design allows for the estimation of several treatment effects.

Example: Banerjee et al (2007)

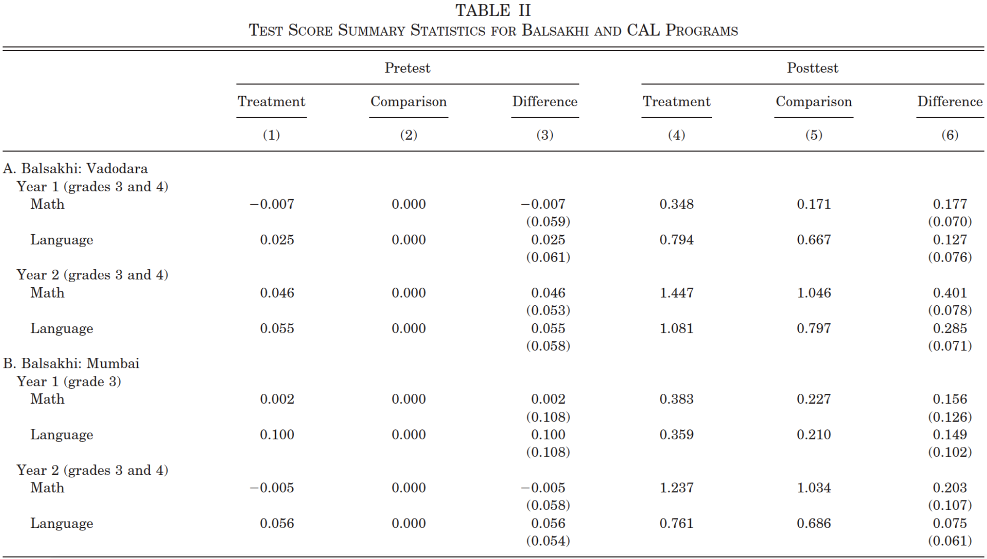


TABLE II
Test Score Summary Statistics for Balsakhi and CAL Programs

	Pretest			Posttest		
	Treatment	Comparison	Difference	Treatment	Comparison	Difference
	(1)	(2)	(3)	(4)	(5)	(6)
A. Balsakhi: Vadodara						
Year 1 (grades 3 and 4)						
Math	−0.007	0.000	−0.007	0.348	0.171	0.177
			(0.059)			(0.070)
Language	0.025	0.000	0.025	0.794	0.667	0.127
			(0.061)			(0.076)
Year 2 (grades 3 and 4)						
Math	0.046	0.000	0.046	1.447	1.046	0.401
			(0.053)			(0.078)
Language	0.055	0.000	0.055	1.081	0.797	0.285
			(0.058)			(0.071)
B. Balsakhi: Mumbai						
Year 1 (grade 3)						
Math	0.002	0.000	0.002	0.383	0.227	0.156
			(0.108)			(0.126)
Language	0.100	0.000	0.100	0.359	0.210	0.149
			(0.108)			(0.102)
Year 2 (grades 3 and 4)						
Math	−0.005	0.000	−0.005	1.237	1.034	0.203
			(0.058)			(0.107)
Language	0.056	0.000	0.056	0.761	0.686	0.075
			(0.054)			(0.061)

But don’t get too excited

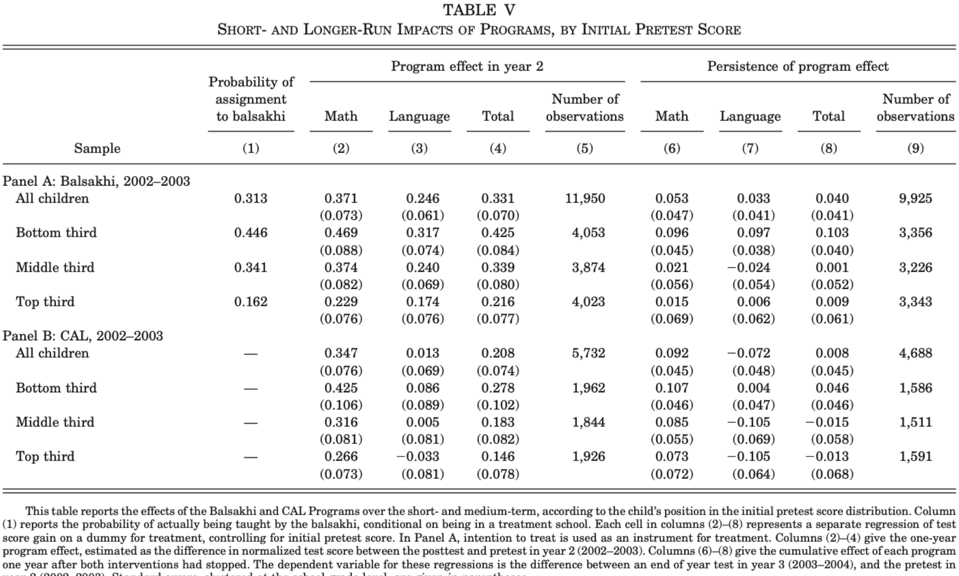


TABLE V

SHORT- AND LONGER-RUN IMPACTS OF PROGRAMS, BY INITIAL PRETEST SCORE

		Program effect in year 2				Persistence of program effect			
	Probability of assignment to balsakhi	Math	Language	Total	Number of observations	Math	Language	Total	Number of observations
Sample	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Balsakhi, 2002–2003									
All children	0.313	0.371	0.246	0.331	11,950	0.053	0.033	0.040	9,925
		(0.073)	(0.061)	(0.070)		(0.047)	(0.041)	(0.041)	
Bottom third	0.446	0.469	0.317	0.425	4,053	0.096	0.097	0.103	3,356
		(0.088)	(0.074)	(0.084)		(0.045)	(0.038)	(0.040)	
Middle third	0.341	0.374	0.240	0.339	3,874	0.021	−0.024	0.001	3,226
		(0.082)	(0.069)	(0.080)		(0.056)	(0.054)	(0.052)	
Top third	0.162	0.229	0.174	0.216	4,023	0.015	0.006	0.009	3,343
		(0.076)	(0.076)	(0.077)		(0.069)	(0.062)	(0.061)	
Panel B: CAL, 2002–2003									
All children	—	0.347	0.013	0.208	5,732	0.092	−0.072	0.008	4,688
		(0.076)	(0.069)	(0.074)		(0.045)	(0.048)	(0.045)	
Bottom third	—	0.425	0.086	0.278	1,962	0.107	0.004	0.046	1,586
		(0.106)	(0.089)	(0.102)		(0.046)	(0.047)	(0.046)	
Middle third	—	0.316	0.005	0.183	1,844	0.085	−0.105	−0.015	1,511
		(0.081)	(0.081)	(0.082)		(0.055)	(0.069)	(0.058)	
Top third	—	0.266	−0.033	0.146	1,926	0.073	−0.105	−0.013	1,591
		(0.073)	(0.081)	(0.078)		(0.072)	(0.064)	(0.068)	

This table reports the effects of the Balsakhi and CAL Programs over the short- and medium-term, according to the child’s position in the initial pretest score distribution. Column (1) reports the probability of actually being taught by the balsakhi, conditional on being in a treatment school. Each cell in columns (2)–(8) represents a separate regression of test score gain on a dummy for treatment, controlling for initial pretest score. In Panel A, intention to treat is used as an instrument for treatment. Columns (2)–(4) give the one-year program effect, estimated as the difference in normalized test score between the posttest and pretest in year 2 (2002–2003). Columns (6)–(8) give the cumulative effect of each program one year after both interventions had stopped. The dependent variable for these regressions is the difference between an end of year test in year 3 (2003–2004), and the pretest in year 2 (2002–2003). Standard errors, clustered at the school-grade level, are given in parentheses.

Designing Randomized Experiments

Nightmare Scenario

- You spend big money running an RCT in the field.
  - You recruit your sample
  - You pre-register your statistical analysis
  - You run your experiment and collect your data
  - Congratulations, you increased the probability of your outcome by 15% from 1% to 1.15%
  - However, is your sample large enough for a statistically significant effect? (Often no!)

Goal

- Design an A/B test (randomized experiment) to measure **lift** from an online ad campaign.
- Typical outcome: user-level **conversion** (binary) or continuous metric (revenue per user, spend).
- Decide sample size per arm n for target Type I error α and power $1 - \beta$.

Define the estimand: lift

- Baseline conversion rate (control): p_0.
- Treatment conversion rate: p_1.
- Absolute lift: $\Delta = p_1 - p_0$.
- Relative lift (percent): Lift $= \dfrac{p_1 - p_0}{p_0}$, so $p_1 = p_0(1 + \text{Lift})$.

Difference in Binary Outcomes (Wald Test)

Let $\widehat{p}_1$ and $\widehat{p}_0$ be the sample proportions of succesfful sales in treatment and control.

Under random assignment and large n_1, n_0:

$$\widehat{p}_j \sim \mathcal{N}\left(p_j,\ \frac{p_j(1-p_j)}{n_j}\right), \quad j \in \{0, 1\}$$

For large n, the difference-in-means is approximately normal:

$$\widehat{p}_1 - \widehat{p}_0 \sim \mathcal{N}\left(p_1 - p_0,\ \frac{p_1(1-p_1)}{n_1} + \frac{p_0(1-p_0)}{n_0}\right).$$

The Wald test statistic divides by the estimated standard error under $H_0 : p_1 = p_0$:

$$Z = \frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/n_1 + \widehat{p}_0(1-\widehat{p}_0)/n_0}} \quad \xrightarrow{d} \quad \mathcal{N}(0,1)$$

This large-sample approximation underpins both the power and sample size formulas.

Large-sample approx.: two-sample difference-in-proportions

Test statistic (Wald) for two independent samples (equal n per arm):

$$\hat{p}_1 - \hat{p}_0 \sim \mathcal{N}\left(p_1 - p_0,\ \frac{p_1(1-p_1)}{n} + \frac{p_0(1-p_0)}{n}\right).$$

Solving for n to achieve power $1-\beta$ at two-sided significance α gives:

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

Here z_q is the q-quantile of the standard normal (e.g. $z_{0.975} \approx 1.96$, $z_{0.8} \approx 0.842$).

Continuous outcome (two-sample t approximation)

If the outcome is continuous with (assumed) common variance σ^2:

$$\hat{\mu}_1 - \hat{\mu}_0 \sim \mathcal{N}\left(\mu_1 - \mu_0,\ 2\frac{\sigma^2}{n}\right)$$

so

$$n = \frac{2\sigma^2\left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{(\mu_1 - \mu_0)^2}.$$

If you think in terms of standardized effect size $d = \dfrac{\mu_1 - \mu_0}{\sigma}$:

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

Worked example (binary conversion)

Suppose:

- Baseline conversion $p_0 = 0.02$ (2%).
- Desired **relative lift** = 10% $\Rightarrow p_1 = 0.02 \times 1.10 = 0.022$.
- Two-sided $\alpha = 0.05$ and power $1 - \beta = 0.80$.

Compute per-arm sample size from the formula:

$$n = \frac{(1.96 + 0.8416)^2 \, (0.022(1 - 0.022) + 0.02(1 - 0.02))}{(0.022 - 0.02)^2}$$

Numerically this gives

$$n \approx 80{,}681 \quad \text{(per arm)}.$$

Interpretation: To detect a 10% relative increase from a very small base rate (2%) requires a huge sample because the absolute change is very small (0.2 percentage points).

Alternate example: bigger lift

If relative lift = 50% (so $p_1 = 0.02 \times 1.5 = 0.03$), same α and power:

$$n \approx 3{,}823 \quad \text{(per arm).}$$

- So power is very sensitive to the *absolute* difference $p_1 - p_0$.
- Of course if you knew p_1 you wouldn't need to do all this work in the first place!

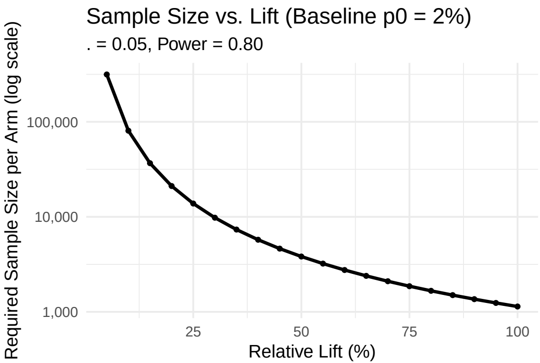
R code: compute sample size (binary)

```r
# two-sided test (approx normal)
p0 <- 0.02
lift <- 0.10
p1 <- p0*(1+lift)
alpha <- 0.05
power <- 0.80
z_alpha <- qnorm(1-alpha/2)
z_beta  <- qnorm(power)

n_per_arm <- ((z_alpha + z_beta)^2 * (p1*(1-p1) + p0*(1-p0))) / ((p1 - p0)^2)

ceiling(n_per_arm)
# Alternatively: use power.prop.test for approximation (one-sided/two-sided differences)
power.prop.test(p1 = p0, p2 = p1, power = power, sig.level = alpha,
                alternative = ”two.sided”)
```

Calculating the Sample Size/Lift Curve



```
p0 <- 0.02
alpha <- 0.05
power <- 0.8
z_alpha <- qnorm(1-alpha/2)
z_beta <- qnorm(power)

lift_grid <- seq(0.05, 1.0, by=0.05)
df <- data.frame(
  lift = lift_grid,
  p1 = p0 * (1 + lift_grid)
)
df$n <- ((z_alpha + z_beta)^2 * (df$p1*(1-df$p1) + p0*(1-p0)
        )) / (df$p1 - p0)^2

ggplot(df, aes(x = lift*100, y = n)) +
  geom_line(linewidth=1.2) +
  geom_point() +
  scale_y_log10(labels = scales::comma) +
  labs(
    x = ”Relative Lift (%)”,
    y = ”Required Sample Size per Arm (log scale)”,
    title = ”Sample Size vs. Lift (Baseline p0 = 2%)”,
    subtitle = ”  = 0.05, Power = 0.80”
  ) +
  theme_minimal(base_size = 14)
```

R code: continuous outcome

```
# Two-sample t approximation
sigma <- 10.0       # estimated SD of outcome
delta <- 1.0        # desired absolute lift (mu1 - mu0)
alpha <- 0.05
power <- 0.80
z_alpha <- qnorm(1-alpha/2)
z_beta  <- qnorm(power)

n_per_arm <- (2 * sigma^2 * (z_alpha + z_beta)^2) / (delta^2)
ceiling(n_per_arm)

# or use built-in
power.t.test(delta = delta, sd = sigma, power = power,
             sig.level = alpha, type = ”two.sample”, alternative = ”two.sided”)
```

Minimum Detectable Effect (MDE)

Often we have fixed n and want the smallest detectable absolute difference Δ at given α, β:

$$\Delta = (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\frac{p_1(1-p_1) + p_0(1-p_0)}{n}}.$$

For planning you can use the conservative plug-in $p_1 \approx p_0$, giving

$$\Delta_{\text{approx}} \approx (z_{1-\alpha/2} + z_{1-\beta})\sqrt{\frac{2p_0(1-p_0)}{n}}.$$

Convert to relative lift as Lift $= \Delta/p_0$.

Practical considerations

- **Equal vs unequal allocation:** If treatment fraction $\pi \neq 0.5$, replace $1/n + 1/n$ by $1/(n\pi) + 1/(n(1-\pi))$ or solve for each arm separately.
- **Clustering / interference:** If randomization is by cluster (e.g., by geography), inflate n by the design effect: $DE = 1 + (m-1)\rho$, where m is cluster size, ρ ICC.
- **Multiple testing / sequential peeking:** Correct α (Bonferroni, group sequential, alpha-spending).
- **Non-compliance / attrition:** Adjust expected treated proportion and effective sample size.
- **Heterogeneous treatment effects:** If variance differs by arm, use $p_1(1-p_1)$ and $p_0(1-p_0)$ explicitly.
- **Pre-launch A/B to estimate p_0 and σ:** Accurate estimates of baseline rate and SD drastically change required n.

Quick checklist before running

1. Estimate baseline p_0 from recent data (not stale).
2. Pick practically meaningful relative lift.
3. Choose α (often 0.05) and desired power (often 0.8 or 0.9).
4. Account for clustering, multiple comparisons, and unequal exposure.
5. If sample sizes are enormous, consider: targeting a segment, increasing effect size via stronger treatment, or switching to a continuous metric with lower variance.

Summary

- For low baseline rates, even modest relative lifts require large sample sizes because absolute differences are tiny.
- Use the formula for difference-in-proportions for binary outcomes; for continuous outcomes the two-sample t formula is natural.
- R functions ‘power.prop.test()‘ and ‘power.t.test()‘ provide convenient built-in calculations.
- Adjust for clustering, attrition, etc. before launching the experiment.