Experimental methods II

ADEC781001: Empirical Behavioral Economics

Lawrence De Geest (lrdegeest.github.io)

BOSTON COLLEGE
Woods College of Advancing Studies

Further reading

- A lot of this lecture (especially the power analysis) is derived from two great sources
  - Moffatt, Peter G. Experimetrics: Econometrics for experimental economics. Macmillan International Higher Education, 2015.
  - List, John A., Sally Sadoff, and Mathis Wagner. "So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design." Experimental Economics 14, no. 4 (2011): 439.

Power Analysis

Do you have the power?

- You run a study and estimate an ATE
  - focus here is on difference in means with t-test (e.g. via regression)
  - also applies to Wilcoxon test
- But did your study have the power to yield a reliable estimate?
- The power of a statistical test is P(detect true result | true result exists)
  - e.g. P(ATE significant | significant ATE exists)
- Recall there are two types of errors
  - Type 1 or false positive (reject H_0 when it is true)
    - also known as test size or significance $\alpha \in [0, 1]$
    - generally considered "costlier" (so you want to minimize probability of making it)
  - Type 2 or false negative (reject to reject H_0 when it is false)
    - also known as β
    - implies P(reject H_0 | H_0 is false) is $1 - \beta$
    - $\pi = 1 - \beta$ is power
  - Note: for fixed n you can't reduce probability of one error without increasing the other
- Convention $\alpha = 0.05$, $\pi = 0.8$ (so $\beta = 0.2$)
  - implies 4:1 tradeoff between Type 2 and Type 1 error
  - objective is to find minimum n that satisfies π

One sample
Set-up

- Continuous outcome Y (e.g. share of a pie offered in an ultimatum game)
- Population mean is μ
- Hypotheses:
  - $H_0 : \mu = \mu_0$, $H_A : \mu = \mu_1, \mu_1 > \mu_0$
- test statistic: t-test, $t = \frac{\bar{y}-\mu_0}{SE} \sim t_df$, $SE = \frac{s}{\sqrt{n}}$, $df = n - 1$
  - Given α, rejection rule is $t > t_{df,\alpha}$
  - Assume you will draw sufficiently large n so Central Limit Theorem binds
  - Then rejection rule is $t > z_\alpha$ (i.e. you compare to standard normal distribution)

One sample
Power

- What is probability test statistic t greater than z_α if $\mu = \mu_1$?

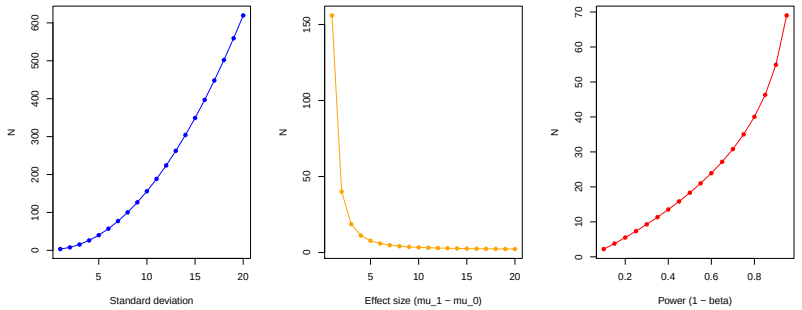
$$\begin{aligned} P(t > z_\alpha \mid \mu = \mu_1) &= P\left(\frac{\bar{y} - \mu_0}{SE} > z_\alpha \,\middle|\, \mu = \mu_1\right) \\ &= P\left(\bar{y} > \mu_0 + \frac{z_\alpha}{SE} \,\middle|\, \mu = \mu_1\right) \\ &= P\left(\frac{\bar{y} - \mu_1}{SE} > \frac{\mu_1 - \mu_0 - z_\alpha SE}{SE} \,\middle|\, \mu = \mu_1\right) \\ &= \Phi\left(\frac{\mu_1 - \mu_0 - z_\alpha SE}{SE}\right) \end{aligned}$$

- To get a power of $1 - \beta$: $\left(\frac{\mu 1 - \mu_0 - z_\alpha SE}{SE}\right) = z_\beta$
- Solve for n: $n = \frac{s^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}$

One sample
Power

- Plug in our values
  - $z_\alpha = \Phi(1 - 0.05) = \texttt{qnorm}(1 - 0.05) = 1.645$
  - $z_\beta = \Phi(0.80) = \texttt{qnorm}(0.80) = 0.841$
  - $n = \frac{6.17s^2}{(\mu_1 - \mu_0)^2}$
- Let $\mu_1 = 12, \mu_0 = 10, s = 5$
  - $n = \frac{6.17 \times 25}{4} = 38.6 \rightarrow 39$ (need integers for subjects!)
- In R: `power.t.test(power = .80, delta = 2, sd=5, type = "one.sample", alternative = "one.sided")`

One sample
Comparative statics



assumes $\alpha = 0.05$

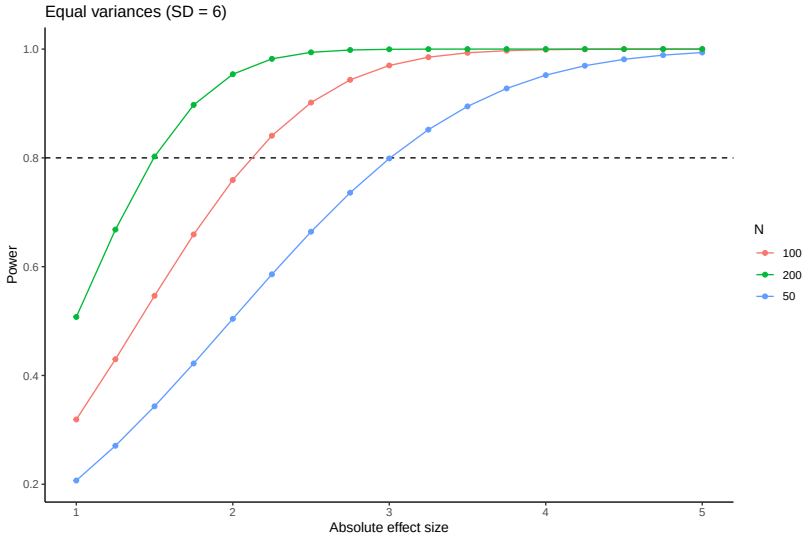
Two samples

- Calculating ATE in an experiment usually implies control vs treatment group (2 samples)
- μ_T is mean of treatment, μ_C is mean of control
  - effect size: $d = \mu_T - \mu_C$
  - estimate d from previous studies/priors/pilot studies
- $H_0 : d = 0$
  - test statistic: $t = \frac{\bar{y_T} - \bar{y_C}}{s_p \sqrt{\frac{1}{n_T} + \frac{1}{n_C}}}$
    - s_p is pooled variance, $s_p = \sqrt{\frac{(n_T - 1)s_T^2 + (n_C - 1)s_C^2}{n_T + n_C - 2}}$

Two samples
Equal sample sizes

- Let $n = n_T = n_C$
- Then $t = \frac{\bar{y_T} - \bar{y_C}}{s_p \sqrt{2/n}}$
- Power of the test: $P(t > z_\alpha \mid d) = \Phi\left(\frac{d - z_\alpha s_p \sqrt{2/n}}{s_p \sqrt{2/n}}\right)$
- For test power $1 - \beta$: $z_\beta = \frac{d - z_\alpha s_p \sqrt{2/n}}{s_p \sqrt{2/n}}$
- Solve for n: $n = \frac{2s_p^2 (z_\alpha + z_\beta)^2}{d^2}$

Two samples
Comparative statics



Two samples
Equal sample sizes, unequal variances

- Suppose as before $d = 2$, and $s_T = 7.84$, $s_C = 4$
  - estimate these from previous studies/priors/pilot studies
  - s_p is about 6 (average of s_T and s_C)
    - Recall s_p is a weighted average where the weights are the sample degrees of freedoms
  - $n = \frac{12.35 \times 36}{4} = 112$
  - R: `MESS::power_t_test(n=NULL, sd=4, power=.8, ratio=1, sd.ratio=7.84/4, delta=2, alternative = "one.sided")`
    - Returns $n = 121$
    - Pretty close to hand calculation

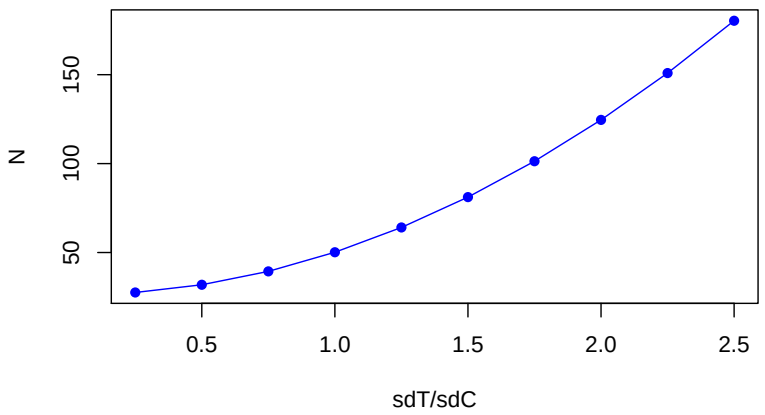
Two samples

Equal sample sizes, unequal variances

- In reality you will have a budget constraint
- Rule of thumb: choose sample sizes so $\frac{n_T}{n_C} \propto \sqrt{\frac{c_C}{c_T}}$
  - c_C is cost per control subject
  - c_T is cost per treatment subject
- Example: experiment varies incentives (high-incentive treatment, low-incentive control)
  - suppose $c_T = 4c_C$
  - then we should expect about twice as many subjects in low-incentive control
- R: `MESS::power_t_test(n=NULL, sd=7.84, power=.8, ratio=2, sd.ratio=7.84/4, delta=2, alternative = "one.sided")`
  - see R script (`be_bc_power.R`) for explanation

Two samples

Comparative statics



Power analysis with groups

- So far we have assumed independence between subjects
  - each subject is their own group
- But this won't hold in a strategic setting where payoffs and thus actions are dependent
  - i.e. when subject are clustered
- Let's suppose subject i is put into group j
  - outcomes between groups are independent (i.e. each group is an independent observation)
  - but outcomes within groups are dependent
- Let u_j be the group-specific error-term
  - model: $Y_{ijT} = \alpha + \beta_T + \mu_j + \varepsilon_{ij}$, $T = \{0, 1\}$
- Basic idea: dependence "inflates" the variation
- So you need a "variance inflation factor" that increases n

Power analysis with groups

Variance inflation factor

- Assume equal sample sizes and variances
- Then List (2011) shows $n = \left(\frac{2s_p^2(z_\alpha + z_\beta)^2}{d^2}\right)(1 + (c-1)\rho)$
  - $1 + (c-1)\rho$ is the variance inflation
  - c is group size
  - ρ is "coefficient of intracluster correlation": $\rho = \frac{var(u_j)}{var(u_j) + var(\varepsilon_{ij})}$
- Suppose no differences between groups
  - then $var(u_j) = 0 \implies \rho = 0 \implies$ no change in n
- Suppose differences between groups - but all individuals within groups behave identically
  - then $var(\varepsilon_{ij}) = 0 \implies \rho = 1 \implies$ multiply n by c
- In reality we expect some intergroup differences and intersubject differences
  - i.e. $var(u_j) \neq var(\varepsilon_{ij}) \neq 0$

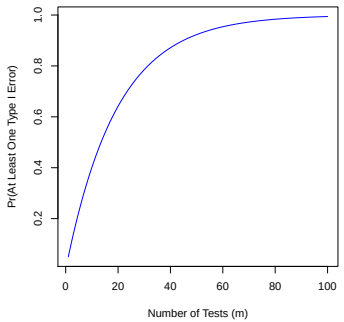
Power analysis with groups

Example

- Same as before: $n = \frac{2s_p^2(z_\alpha + z_\beta)^2}{d^2} = 112$
- But now group size is $c = 4$
- Suppose $\rho = 0.05$
- Inflation factor is 1.15
- New sample size is $112 \times 1.15 = 129$
  - need number divisible by c: round down (128) or round up (132)
- Where do you get estimates of $var(\varepsilon_{ij})$ and $var(u_j)$?
  - hard to find in other papers (not always reported)
  - best-case: pilot studies

Multiple hypothesis testing

- Suppose you have three treatments $T \in \{0, 1, 2\}$
- When you estimate the ATEs you are now testing two hypotheses (assuming $T = 0$ is the reference)
- Testing multiple hypothesis at once leads to "α inflation"
  - P(make Type 1 error) = α
  - P(not make Type 1 error) = $1 - \alpha$
  - P(not make Type 1 error in m tests) = $(1 - \alpha)^m$
  - P(make at least **one** Type 1 error in m tests) = $1 - (1 - \alpha)^m$



Multiple hypothesis testing

Correcting α

- You have to cut your α if you are testing multiple hypotheses
  - in terms of power this means you are going to need more data
- Many ways to adjust α, still an open discussion
  - For a detailed discussion see List et al. (2016)[1]
- Some methods
  - Bonferroni adjustment
  - False Discovery Rate (FDR)
  - Though these don't adjust for dependence in hypotheses
  - See List et. al (2016)
    - adjustments much more complicated for dependent hypotheses

[1] List, John A., Azeem M. Shaikh, and Yang Xu. "Multiple hypothesis testing in experimental economics." Experimental Economics (2016): 1-21.

Multiple hypothesis testing

Bonferroni and FDR adjustments

- Most conservative approach: Bonferonni correction
  - Reject H_0 if $p < \frac{\alpha}{m}$ where m is the number of hypotheses
  - assumes hypotheses are independent
  - problem: as m grows it leads to high Type 2 error (false negative) rate
    - i.e. power goes down
- False Discovery Rate (FDR)
  - basically a Bonferonni adjustment on *ordered* p-values
  - first order the p-values smallest to largest
  - then check if k^{th} ordered p-value greater than $\frac{\alpha}{m}$

Takeaways

- Power analysis generates suggested sample sizes under best-case scenarios
- Lots of tradeoffs to make
  - N, α, β, etc.
  - requires estimates of variance and treatment sizes, often difficult to obtain without pilot studies
- Overall: helpful to get you thinking about your design and analysis
  - Good to do many power calculations for different scenarios
- Other practical benefits
  - minimize cost of data collection
  - many grants require power calculations
- P.S. Other software to calculate power
  - Stata: `power`
  - G*Power: http://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower.html

Randomization

Approaches to randomization

- Randomization is the key to identification
  - as in identifying (and estimating) the ATE, a causal relationship
- Simplest case is a completely randomized design
  - draw a random sample from the subject pool
  - randomly assign subjects to control/treatment
- Pros: ensures no correlation between treatment assignment and subject characteristic
- Cons: sample sizes are random to each treatment so possibility for high variance
  - high variance $\implies$ harder to identify ATE

Factorial designs

- Assign pre-determined sample size to each treatment
  - e.g. each treatment has n subjects
- Example: 2x2 dictator game
  - low stakes (L) or high stakes (H)
  - communication (C) or no communication (NC)
  - 2x2 = 4 treatments (L-C, H-C, L-NC, H-NC)
  - assign n subjects to each treatment
  - "full factorial design" because all treatment combinations are covered
  - allows you to estimate ATEs as well as interactions
    - e.g. average effect of L and interaction L-C
- In general: factorial design requires 2^m trials for m treatments
- Make sure to randomize assignment to treatments
  - e.g. don't assign treatments (or roles within treatments) by order in which subjects arrive (since early arrival may be correlated with behavior in the game)

Block design

- $Y = \alpha_i + \beta T + \mathbf{X}\gamma + \varepsilon$
  - T is treatment
  - $\mathbf{X}$ are observable subject characteristics (e.g. gender)
- If goal is to remove role of $\mathbf{X}$ on treatment then randomize *within* (not between) "blocks"
- "Blocking factor" is source of variation not of primary interest
  - the variable on which "blocking" is applied is the blocking factor
  - e.g. block on gender to control for variation due to gender (and not treatment)

Within-subject design

- Extension of block design
  - assign same subject to multiple treatments
  - experimenter blocks on a single subject
- $Y = \alpha_i + \beta T + \mathbf{X}\gamma + \varepsilon$
  - α_i is subject-specific effect
  - easier to estimate (and thus improve precision of ATE estimate) using within design
- Let β_{ws} be the ATE of the within design and β_{bs} the between design
  - $V(\beta_{ws}) = V(\beta_{bs}) - \frac{2}{N} V(\alpha_i)$
    - $V(\cdot)$ is the variance
  - If subjects are identical $V(\alpha_i) = 0$
    - no difference in within or between design
  - If subjects are not identical $V(\alpha_i) \neq 0$
    - benefit of within design increases with $V(\alpha_i)$