Section 6: Power Calculations

Jon Cohen
October 29, 2021
Outline

Power Calculations

Parametric Power Calculations

Simulation Power Calculations

Potpourri of Power Calculation Issues

Concluding Thoughts
1. How big of a sample size do you “need”?

2. Conditional on sample size, how “should” you allocate across arms?
1. How big of a sample size do you “need”?

2. Conditional on sample size, how “should” you allocate across arms?

**General intuition:** Make *ex ante* assumptions about how your experiment *will* look to understand properties of eventual analysis.
Components of a Power Calculation

- **Specify data generating process**
  - Randomly assign $n$ observations into treatment and control group
  - Variance of outcomes $\sigma^2$

- Specify estimand of interest
  - ATE: $E[Y|D=1] - E[Y|D=0]$

- Specify estimator and its properties
  - Difference in means $\mu_1 - \mu_0$ with sample sizes $N_1, N_2$
  - False positives (size/Type I error): $\alpha$ fraction of the time
  - False negatives (power/Type II error): $1 - \beta$ fraction of the time
  - Minimum detectable effect size $\delta$
Components of a Power Calculation

- **Specify data generating process**
  - Randomly assign $n$ observations into treatment and control group
  - Variance of outcomes $\sigma^2$

- **Specify estimand of interest**
  - ATE: $E[Y|D = 1] - E[Y|D = 0]$
Components of a Power Calculation

• Specify data generating process
  ■ Randomly assign $n$ observations into treatment and control group
  ■ Variance of outcomes $\sigma^2$

• Specify estimand of interest
  ■ ATE: $E[Y|D = 1] - E[Y|D = 0]$

• Specify estimator and its properties
  ■ Difference in means $\mu_1 - \mu_0$ with sample sizes $N_1, N_2$
  ■ False positives (size/Type I error) $\alpha$ fraction of the time and false negatives (power/Type II error) $1 - \beta$ fraction of the time
  ■ Minimum detectable effect size $\delta$
You should walk away from this recitation knowing...

1. How to analytically solve for a simple power calc
2. The idea behind simulating an arbitrarily complex power calc
3. Why you shouldn’t commit the cardinal sin of calculating “post hoc power”
Useful References

  - “So You Want To Run An Experiment, Now What? Some Simple Rules of Thumb For Optimal Experimental Design”

  - “Using Randomization in Development Economics Research: A Toolkit”
Outline

Power Calculations

Parametric Power Calculations

Simulation Power Calculations

Potpourri of Power Calculation Issues

Concluding Thoughts
Parametric Power Calc Verbal Intuition

1. Draw outcome distributions under the null and a specific alternative hypothesis.
2. Assume $\sigma$ and $n$ to get distribution of the (random variable) estimator.
3. Calculate rejection regions of relevant curves.
Visual Intuition: Rejection Threshold and Region if Null is True
Visual Intuition: Rejection Threshold if Small Alternative is True

Suppose true effect were 1 SE (Standard Error):
Visual Intuition: Rejection Region if Small Alternative is True

Power would only be approximately 0.17

Probability density

0

0

Null

Under 1 SE effect
Visual Intuition: Rejection Threshold if Large Alternative is True

Suppose true effect were 3 SE’s (Standard Errors):
Visual Intuition: Rejection Region if Large Alternative is True

Power would be approximately 0.85
Visual Intuition: MDE Controls Size and Power Appropriately

How the power calculation formula works

- Null distn.
- Effect distn.
- $t_{\alpha/2}$ size
- $t_{1-\kappa}$ power
(Same visual intuition with more notation)
Parametric Power Calculation Math for MDE $\delta$

1. $\hat{\delta} \sim N(\delta, \sigma_{\hat{\delta}})$ by CLT, getting $\sigma_{\hat{\delta}}$ with reasonable assumptions on outcome variance

2. For confidence level $\alpha$, true parameter $\delta$, and power $1 - \beta$:

$$P\left(\frac{\hat{\delta}}{\sigma_{\hat{\delta}}} > t_{\alpha/2} | \delta\right) = 1 - \beta$$  (probability of correctly rejecting null)

$$P\left(\frac{\hat{\delta} - \delta}{\sigma_{\hat{\delta}}} > t_{\alpha/2} - \frac{\delta}{\sigma_{\hat{\delta}}} | \delta\right) = 1 - \beta$$  (recenter by subtraction)

$$\Phi\left(\frac{\delta}{\sigma_{\hat{\delta}}} - t_{\alpha/2}\right) = 1 - \beta$$  (by normality of $\delta$ and symmetry of $\Phi(\cdot)$)

$$\frac{\delta}{\sigma_{\hat{\delta}}} - t_{\alpha/2} = t_{1-\beta}$$  (since $t_k \equiv$ threshold under which $k\%$ of $\Phi(\cdot)$ lies)

$$\delta_{MDE} = (t_{1-\beta} + t_{\alpha/2})\sigma_{\hat{\delta}}$$  Calculated by Stata command `sampsi`
Sanity Check with OLS, Two Groups, and No Covariates

- \( Y_i = \alpha + \delta D_i + \epsilon_i \)
- \( D_i \in \{0, 1\} \text{ with } P(D_i = 1) = p \)
- \( \epsilon_i \text{ i.i.d. with } Var(\epsilon) = \sigma^2 \)

What is the formula for \( \hat{\sigma}_\delta \) given the above setup?
Sanity Check with OLS, Two Groups, and No Covariates

- $Y_i = \alpha + \delta D_i + \epsilon_i$
- $D_i \in \{0, 1\}$ with $P(D_i = 1) = p$
- $\epsilon_i$ i.i.d. with $Var(\epsilon) = \sigma^2$

What is the formula for $\hat{\sigma}_\delta$ given the above setup?

$$\hat{\sigma}_\delta = \sqrt{\frac{1}{p(1-p)} \frac{\sigma^2}{N}}$$
More General Setup

1. $Y_{iD} = \alpha_i + X_i \beta + (\bar{\delta} + \delta_i) D_i + \epsilon_i$
2. $\sigma^2_1 - \sigma^2_0 = \text{Var}(\delta_i | X)$
3. $\sigma_{\hat{\delta}} = \sqrt{\frac{\sigma^2_1}{N_1} + \frac{\sigma^2_0}{N_0}}$
More General Setup

- $Y_{iD} = \alpha_i + X_i \beta + (\tilde{\delta} + \delta_i) D_i + \epsilon_i$
- $\sigma_1^2 - \sigma_0^2 = Var(\tilde{\delta}_i | X)$
- $\sigma_\delta = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_0^2}{N_0}}$
- In theory, want to allocate a given overall $N$ in proportion to outcome variance
  - Analogous results for arm cost differences given an overall budget
- In practice, researchers rarely deviate from equal arm size
Extension #1: Imperfect Compliance

Why does this affect the MDE?
Extension #1: Imperfect Compliance

Why does this affect the MDE?

1. Reduced-form (ITT): $MDE_{\text{perfect comp.}} = MDE_{\text{partial comp.}} \times \text{complier share}$

2. Not as straightforward for instrumental variables (LATE)
   - See Austin Frakt’s blog for a derivation
Extension #2: Group-level Randomization

Why does this affect the MDE?

Explicitly correct for intra-cluster correlation between observations... Scale $\sigma^2$ by $\sqrt{1 + (n_{\text{group size}} - 1) \rho}$, where $\rho$ is the intra-cluster correlation (i.e. % of overall variance explained by within-group variance)

Stata command: loneway or sampclus

...or collapse outcomes to the unit of randomization and apply previous results
Extension #2: Group-level Randomization

Why does this affect the MDE?

1. Explicitly correct for intra-cluster correlation between observations...
   - Scale $\sigma_\delta$ by $\sqrt{1 + (n_{\text{groupsize}} - 1)\rho}$, where $\rho$ is the intra-cluster correlation (i.e. % of overall variance explained by within-group variance)
   - Stata command: `loneway` or `sampclus`

2. ...or collapse outcomes to the unit of randomization and apply previous results
Extension #3: Controlling for Covariates

- Pros?

- Cons?

- Alternatives?
Extension #3: Controlling for Covariates

- **Pros?**
  - Can soak up residual variance in outcomes

- **Cons?**
  - Can undo randomization that was the point in the first place
  - Do not want to control for mediating factors

- **Alternatives?**
  - Stratify randomization on covariates
Why does this affect the MDE?
Extension #4: Between vs. Within-Subjects Designs

**Why does this affect the MDE?**

- Within-subject can be thought of as stratifying treatment at the subject-level

\[
\text{Var}(\hat{\delta}) = \frac{\sigma_1^2}{N_W} + \frac{\sigma_0^2}{N_W} - \frac{2\sigma_1\sigma_0\rho}{N_W}
\]

where \( \rho \) is within-subject correlation in outcomes

- Very related to McKenzie (2012) JDE
  “Beyond baseline and follow-up: The case for more T in experiments”
Extension #5: Continuous Treatment

- Suppose I think the effect is linear. Does it matter what values of treatment I randomize?
- What if I think the effect is quadratic?
- See Section 6 of List, Sadoff, and Wagner
Extension #6: Spillovers

- What if the stable unit treatment value assumption (SUTVA) is violated? (i.e. your treatment affects my outcome)
  - Classic example is the *Miguel and Kremer (2004)* de-worming paper
Extension #6: Spillovers

- What if the stable unit treatment value assumption (SUTVA) is violated? (i.e. your treatment affects my outcome)
  - Classic example is the *Miguel and Kremer (2004)* de-worming paper
- Identification: Carefully specify estimand for MDE. Need both individual and “market”-level randomization.
- Inference: Hard. Best to simulate.
• What if the stable unit treatment value assumption (SUTVA) is violated? (i.e. your treatment affects my outcome)
  ■ Classic example is the Miguel and Kremer (2004) de-worming paper
• Identification: Carefully specify estimand for MDE. Need both individual and “market”-level randomization.
• Inference: Hard. Best to simulate.
• See Aronow, Eckles, Samii, and Zonszein (2020) for modern methods
Extensions Takeaways

- The variance term is more complicated in more complicated designs
- But simulations are good to avoid annoying derivations
Outline

Power Calculations

Parametric Power Calculations

Simulation Power Calculations

Potpourri of Power Calculation Issues

Concluding Thoughts
1. Use an underlying model to generate (arbitrarily complex!) data
2. Run (arbitrarily complex!) estimation on simulated data from (1)
3. Given confidence level $\alpha$, record whether the result from (2) is significant
4. Repeat (1)-(3) many times
5. Power is fraction of rejections
Power Calc Simulation Implementation

1. Code it up yourself
2. DeclareDesign
   - Available in R with additional Stata packages
   - Its blog nicely emphasizes steps in pre-specifying model, parameters of interest, and empirical strategy to gauge power and bias
   - (I personally haven’t found the command that intuitive)
Outline

Power Calculations

Parametric Power Calculations

Simulation Power Calculations

Potpourri of Power Calculation Issues

Concluding Thoughts
Potpourri #1: Power Calculations are Ex Ante!

- It’s tempting to plug the observed effect size and standard deviation into the power formula to see how much an estimate should move your priors

Source: Daniel Lakens' blog (see also Gelman 2018)

Simulated from DGP with 50% Power

Simulated from DGP with 90% Power
Potpourri #1: Power Calculations are Ex Ante!

- It’s tempting to plug the observed effect size and standard deviation into the power formula to see how much an estimate should move your priors

- **DO NOT DO THIS!** “POST-HOC POWER” IS SIMPLY A MONOTONIC TRANSFORMATION OF THE P-VALUE

- Source: Daniel Lakens’ blog (see also Gelman 2018)
Potpourri #2: Underpowered Experiments

- Why is an underpowered (e.g. low $\beta = 0.06$) experiment bad?

Source: Andrew Gelman's blog (based on Gelman and Carlin 2014)
Potpourri #2: Underpowered Experiments

- Why is an underpowered (e.g. low $\beta = 0.06$) experiment bad?
- “Type S” error: Conditional on significant result, probability it’s wrong-signed
- “Type M” error: Conditional on significant result, expected overstatement

Source: Andrew Gelman’s blog (based on Gelman and Carlin 2014)
Potpourri #3: Factorial Designs

- Two binary treatments $D_1$ and $D_2$
- Interested in effect of treatment 1 relative to control
- Fully saturated “long” specification: $Y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} T_{2i} + \epsilon_i$
- Commonly used “short” specification: $Y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \epsilon_i$
- Why might the “short” specification have different power/size properties?
• Muralidharan, Romero, and Wuthrich (2020) WP derives the properties
• World Bank blog has accessible write-up on these problems
  - Pre-testing and running short regression isn’t uncommon!
  (e.g. the Amy’s 2018 SNAP paper!)

Note: Simulations are based on sample size N, normal iid errors, and 10,000 repetitions. The size for figures 1c and 1a is \( \alpha = 0.05 \).
Outline

Power Calculations

Parametric Power Calculations

Simulation Power Calculations

Potpourri of Power Calculation Issues

Concluding Thoughts
Art of the Power Calculation

1. Standard deviation of outcome $\hat{\sigma}_y$
   - Pilot study/previous studies
   - Survey data

2. MDE $\delta^{MDE}$
   - What would be “interesting” or cost-effective
   - Compare to interventions with similar goals
   - Use information from theory/calibrated models

3. Sample size $N$
   - What would be feasible given implementation partner and budget constraints
Potential Connections to Other Papers

- Power calculations emphasize sampling-based uncertainty
  - How could you incorporate design-based uncertainty a la Abadie et al. (2020) ECMA?
- Power calculations emphasize statistical significance
  - Is it more reasonable to focus only on $\sigma_\delta$ a la Abadie (2020) AERI?