J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB

Essentials of Sample Size and Power

Course Overview

1. Why Evaluate
2. Theory of Change & Measurement
3. Why & When to Randomize
4. How to Randomize
5. Sample Size & Power
6. Ethical Considerations for Randomized Evaluations
7. Threats & Analysis
8. Randomized Evaluation from Start to Finish
9. Applying & Using Evidence
10. The Generalizability Framework

Power tracks

Essentials of Sample Size and Power: The lecture will cover the intuition behind power calculations and go over some basic principles for determining a study size that minimizes the probability of false negatives. It is aimed at policymakers and practitioners who wish to understand the essentials of power and how various components can be tweaked when designing a study.

→ **This is where you are now!**

Mechanics of Power Calculations: The lecture is designed for participants who are looking to discuss statistical power in more depth and may be planning on conducting power calculations in the near future. The lecture provides the statistical framework for power, introduces its components, and provides practical guidance for power and sample size calculations. This lecture might be right for you if you:

- Have taken at least one class on probability theory, statistics, or econometrics
- Have at least some experience working with data
- Have at least some experience reading academic literature

Learning objectives for Essentials of Power and Sample Size

Takeaways:

- A basic understanding of what statistical power is and why it is important
- An understanding of the difference between "absence of evidence" and "evidence of absence" and the pitfalls of an underpowered study
- A basic understanding of the relationship between statistical power, sample size, and effect size

Who has heard of **sample size** or **power**?

What are some challenges you have experienced or questions you have about sample size or power?

Session outline

- Statistical power
  - False positives and negatives
  - What is power and why does it matter?
  - Importance of avoiding an underpowered study – absence of evidence versus evidence of absence
- The relationship between **sample size**, **power**, and **effect size**
  - Introduce sample size and effect size
  - Other key components of power
  - Rules of thumb for power and sample size

Key concept for this session: sampling & variation

Why care about the sample size?

- The larger the sample,
  - The less likely it is that the result of a positive impact is a "false positive"
  - The less likely it is that the result of no impact is a "false negative"
- Question: How can a result be "false?"
- Answer: Statistical uncertainty

Why care about the sample size?

- The larger the sample,
  - The less likely it is that the result of a positive impact is a "false positive"
  - The less likely it is that the result of no impact is a "false negative"
- Question: How can a result be "false?"
- Answer: Statistical uncertainty

Credit: Montagehealth.org

Evaluation results versus underlying truth

What we really want to know but cannot observe

Reality/underlying truth

What we actually measure/learn

Evaluation results

	No impact	**Impact**
No impact detected		
Impact detected		

Evaluation results versus underlying truth

Important note!
This is a thought experiment! We can't measure "underlying truth" so we are using our evaluation results to approximate it

		Reality/underlying truth	
		No impact	**Impact**
Evaluation results	**No impact** detected		
	Impact detected		

Results versus underlying truth

Let's use a recognizable example:

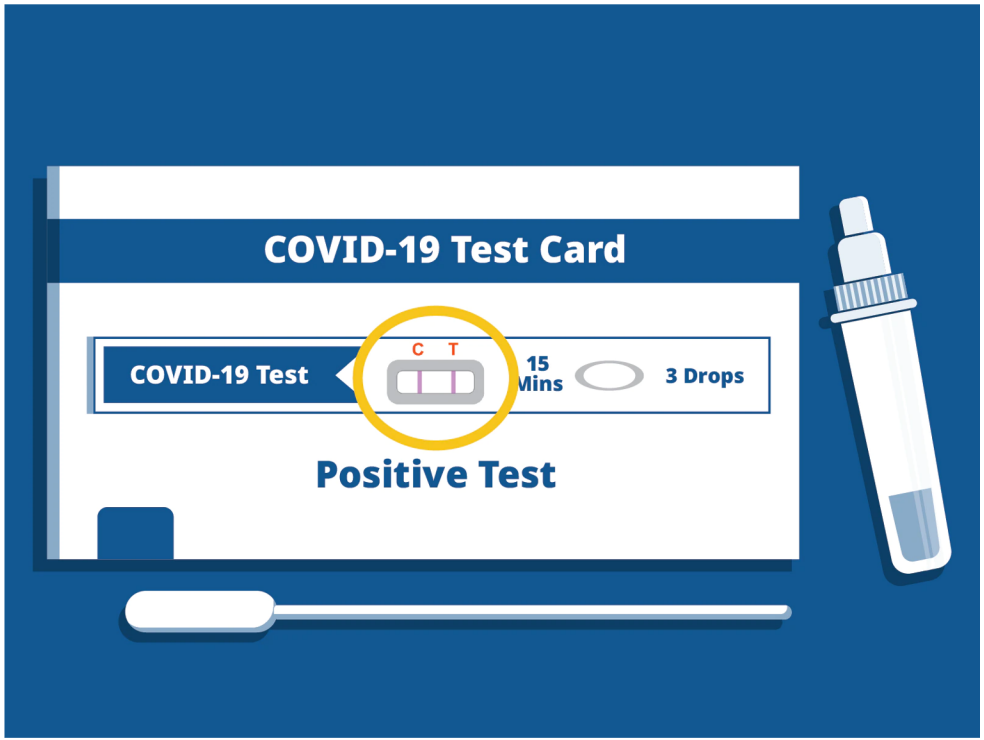


Image: mass.gov

Results versus underlying truth

Reality/underlying truth:
Are you infected with Covid-19?

Rapid test results: Do you test positive or negative?	**Not sick**	**Sick**
Negative test results	**Accurate/true negative**	**False negative:** you conclude you are NOT sick when you are
Positive test results	**False positive:** you conclude that you are sick when you are not!	**Accurate/true positive**

How does this relate to RCTs and impact evaluation?

- The underlying logic is similar:
  - We can't observe the underlying reality/underlying truth
  - We **take a sample** and use that sample to try to learn something about the underlying truth
  - And we want to minimize false positives and false negatives to the extent possible!

Evaluation results versus underlying truth

Reality/underlying truth

Evaluation results	**No impact**	**Impact**
No impact detected	GREAT!	**False negative:** you conclude there is NO impact when there is Mismatch!
Impact detected	**False positive:** you conclude there is impact when there is not Mismatch!	GREAT!

What are some risks if you find a **false positive** (finding an impact when there isn't actually an impact)?

What are some risks if you find a **false negative** (finding no impact when there actually is an impact)?

What is statistical power?

The **statistical power** of an evaluation is the probability of detecting an impact when there actually is one

In other words, statistical power is the likelihood of **avoiding a false negative** (concluding there is no impact when there actually is one)

By convention, we aim for 80% power

- This means that we expect that 80% of the time we will be able to detect an effect if there is one
- 20% of the time we will falsely conclude there is no impact of our program

Quick aside: Statistical significance

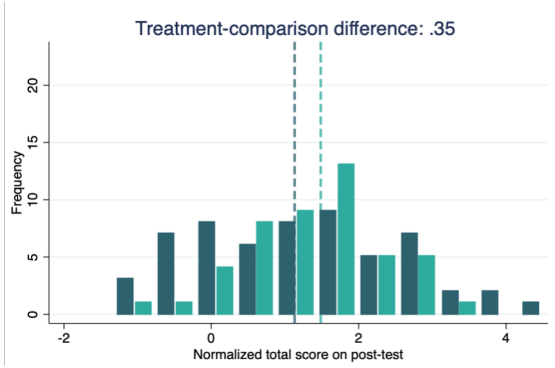
What do we mean when we say "detect an effect?"

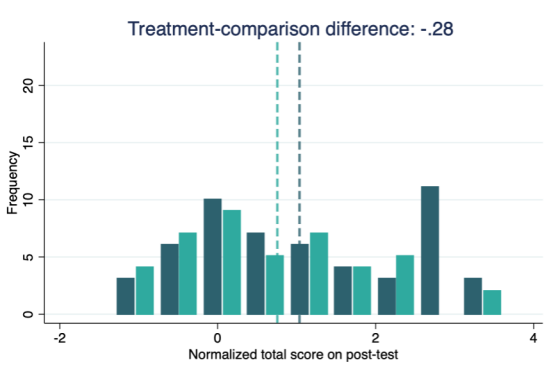
→ find a **statistically significant** impact

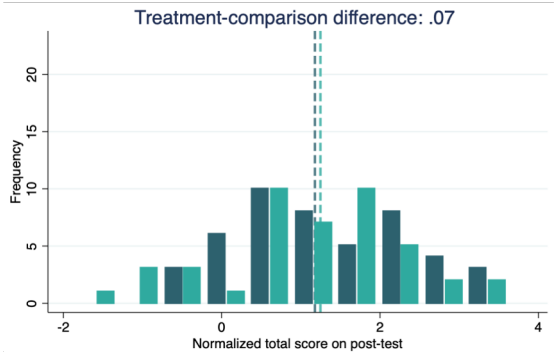
A **statistically significant result** is unlikely to have been produced by chance

Statistical significance is about **avoiding a false positive** (concluding your program had an impact when it did not)

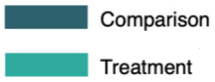
Different random samples from the same population lead to different treatment effect size estimates







Challenge: Is the difference between groups due to chance variation or an effect of the program?



Samples of size **200** drawn from the same underlying data.

Source: Banerjee, Abhijit; Cole, Shawn; Duflo, Esther; Linden, Leigh, 2017, "Balsakhi", Harvard Dataverse,

Research publication: Abhijit Banerjee, Shawn Cole, Esther Duflo, Leigh Linden; "Remedying Education: Evidence from Two Randomized Experiments in India", The Quarterly Journal of Economics 122(3), 1235–1264.

Statistical significance, continued

We want to be sure that the measured program effect is due to the program itself and not due to natural variation or random chance

If it is reasonably unlikely (5% or less) that we would observe this outcome due to random chance (natural variation in participants), then we conclude that the result was **statistically significant**

However, if it is likely that we could observe this outcome due to random chance, then we do **not** have a statistically significant result

Quick recap: statistical significance & statistical power

Statistical significance

- "Detecting an effect"
- Avoiding false positives a.k.a. falsely concluding there was an impact when there is none
- By convention, we set statistical significance to 95% or higher
  - This means that 5% (or less) of the time we will falsely conclude our program had an impact when the differences observed were actually due to random chance

Statistical power

- Probability of finding an effect if there actually is one!
- Avoiding false negatives a.k.a. falsely concluding there is no impact
- By convention, we aim for 80% power
  - This means that we expect that 20% of the time we will falsely conclude there is no impact of our program

Absence of evidence or evidence of absence?

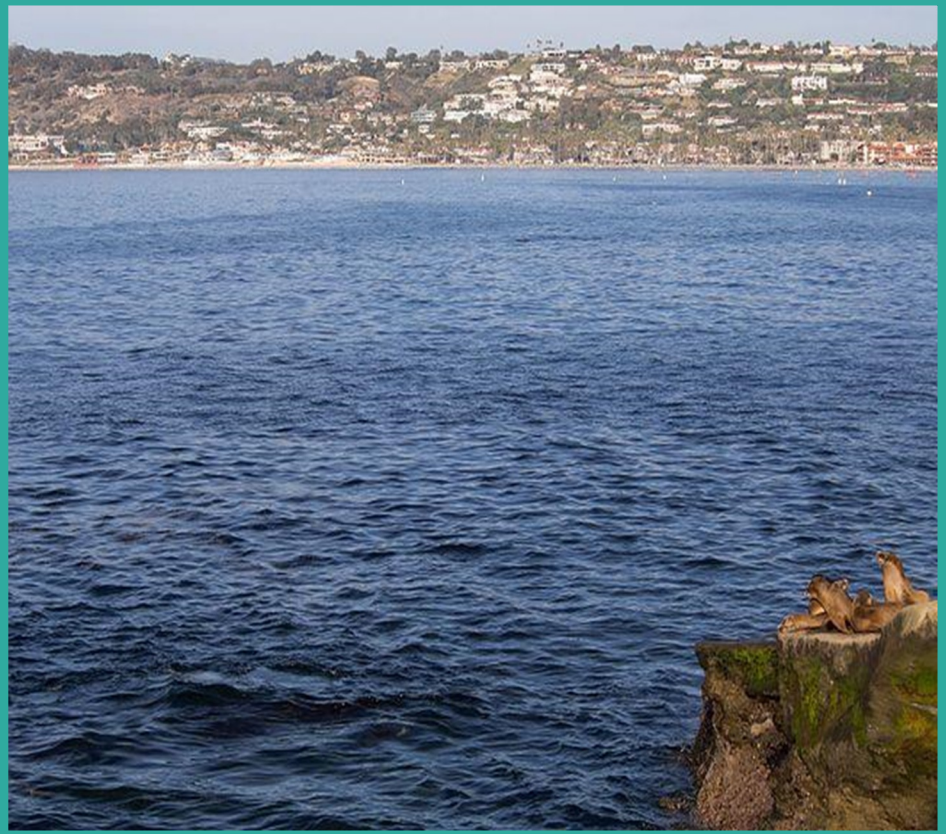
If we do **not** have a statistically significant result, there are two interpretations:

1. There is no effect of our program (true negative!)
2. There is an effect, but we don't have enough statistical power to observe it (false negative!)

Without adequate power, an evaluation may not teach us much

Failure to find a statistically significant effect can be misinterpreted as the failure of the program, rather than the failure of the evaluation

Is there evidence of sharks?

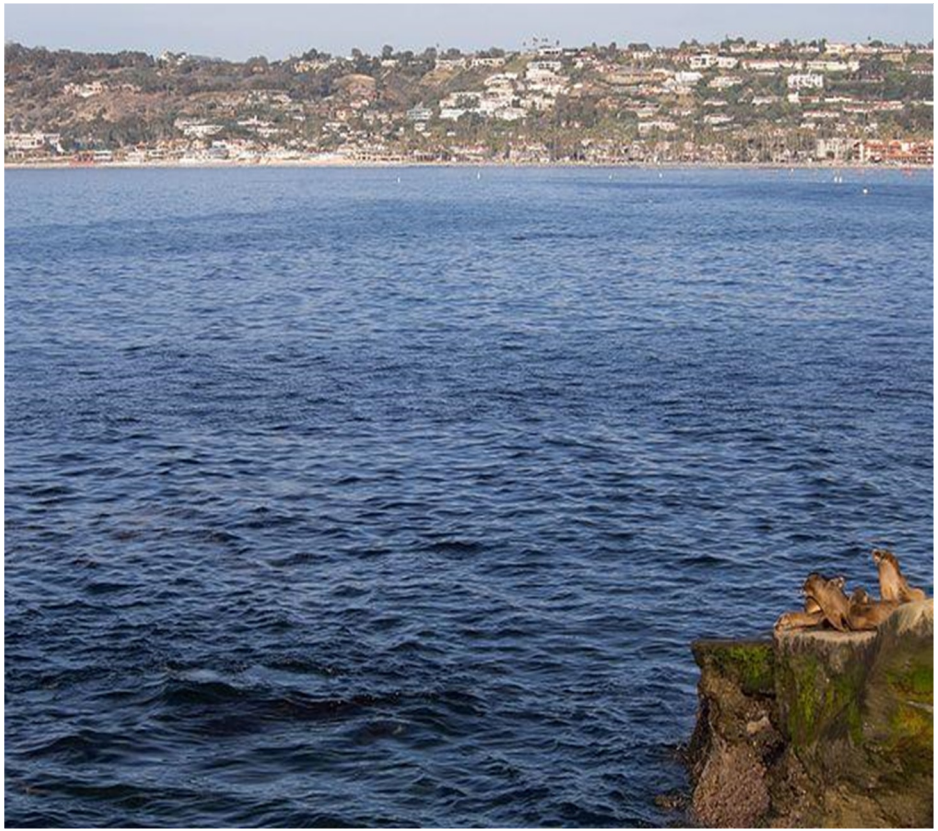


In other words:

If I tell you there are sharks in this water, is there evidence to support my claim?

Source: Wikimedia Commons

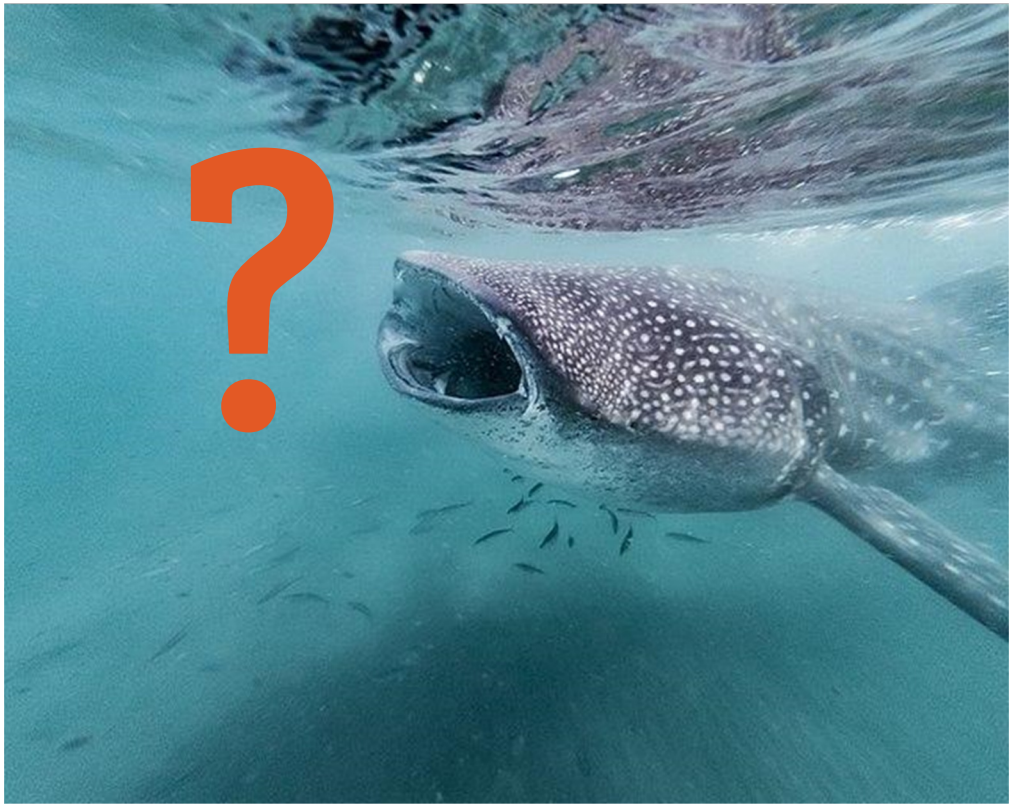
Absence of evidence



This photo represents an "absence of evidence" of sharks

Source: Wikimedia Commons

Absence of evidence



We cannot conclude that there are no sharks under the surface of the water.

Source: Wikimedia Commons

Is there evidence of sharks?

Source: Wikimedia Commons

In other words:

If I tell you there are no sharks in this pool, is there evidence to support my claim?

Evidence of absence

Source: Wikimedia Commons

In this case, we can conclude with certainty that there are no sharks in the pool

We have evidence of no sharks

Absence of evidence or evidence of absence?

Absence of evidence

Evidence of absence

Source: Wikimedia Commons

Absence of evidence or evidence of absence?

Reality/underlying truth

Evaluation results	**No impact**	**Impact**
No impact detected		
Impact detected		

Source (dog): @kingmajesty_and_princessrose, Instagram
Source (others): Wikimedia Commons

Absence of evidence or evidence of absence?

Reality/underlying truth

Evaluation results

Session outline

- Statistical power
  - False positives and negatives
  - What is power and why does it matter?
  - Importance of avoiding an underpowered study – absence of evidence versus evidence of absence
- The relationship between **sample size**, **power**, and **effect size**.
  - Introduce sample size and effect size
  - Other key components of power
  - Rules of thumb for power and sample size

What determines a well-powered study?

Key inputs in determining statistical power

- Effect size (accounting for participation or take up rate)
- Sample size (accounting for attrition)
- Variation in the outcome
- Proportion of the sample in the program group
- Unit of randomization

Example: An evaluation to measure the impact of a financial literacy program on savings

Effect size & minimum detectable effect (MDE)

- What: The impact your program has on the outcome(s) of interest
  - The MDE is the minimum effect size that can be detected with given statistical power (e.g., 80%)
- Example: The financial literacy program might lead to a $50 increase (on average) in savings for the treated group
  - Consider take-up: If you offer the seminar to 300 people and only 150 show up, all 300 would still be in your treatment group!
- How does it affect power? The larger the effect size, the smaller the sample needed to detect an effect

How to choose a MDE

- Choosing a reasonable MDE is the hardest part
- Think about what constitutes meaningful impact in your context!
  - A study should be powered for the smallest effect size that is still decision-relevant
    - Evidence of (meaningful) impact
    - Evidence of no (meaningful) impact
  - Decisions could be based on a number of things: cost vs. benefit, alternative program options, academic literature, etc.
- Reminder:
  - What is *not* decision-relevant?
    - No evidence of impact → underpowered study
    - Hyper-precision

What is the smallest effect size that is meaningful to you?

Why is that a meaningful effect size?

Examples:

- Covers program costs
- Comparable to the effect other programs have achieved

Sample size

- What: The number of individuals or units in your evaluation
  - Remember to account for attrition (individuals dropping out of the evaluation/program)!
- Example: A financial literacy program may serve 300 people from one community
  - Consider attrition: 300 people may enroll, but only 100 persist through the program. Attrition can be either someone dropping out of your program or missing from data collection

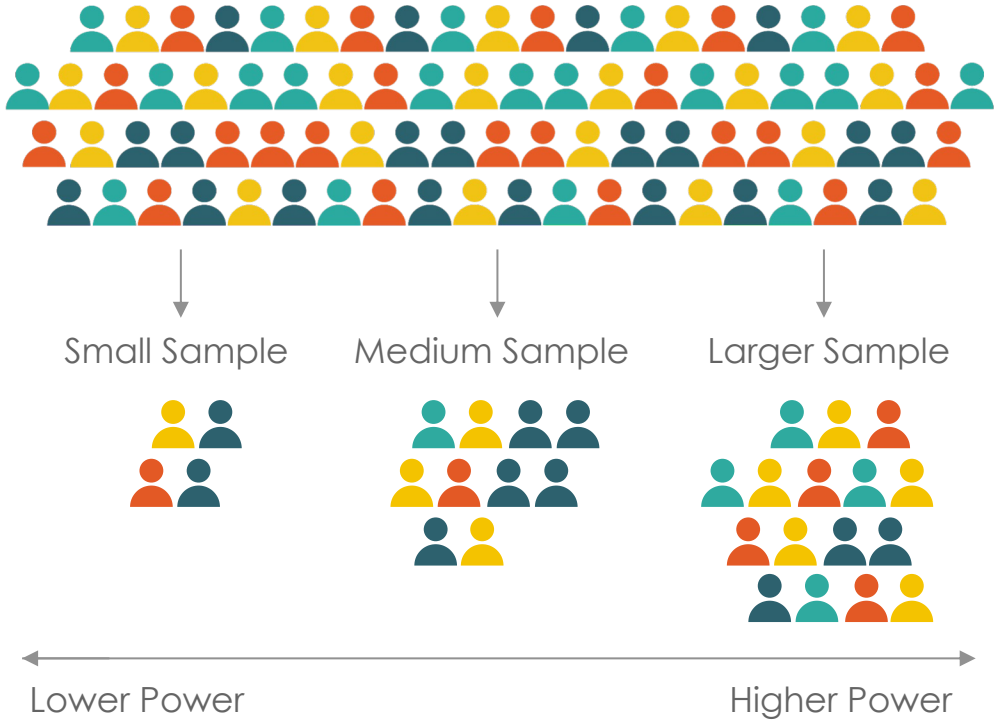
How does sample size affect power?

- For a given effect size: larger sample will yield more statistical power

- For a given power level: larger sample means you can detect a smaller effect

This is the input to power that you have the most control over and will thus often be your main lever to maximize power!

A larger sample increases the power of the evaluation



If the expected effect size is small, the evaluation will need a larger sample

To illustrate this:
Try to detect differences in these two images

Differences between the two images symbolizes effect size

The size of the images symbolizes the sample size

Houses by sripfoto from NounProject.com

If the expected effect size is small, the evaluation will need a larger sample

Differences between the two images symbolizes effect size

The size of the images symbolizes the sample size

When the images get bigger (sample size increases), it is possible to detect small differences (smaller effect size)

With a bigger sample, you can detect smaller differences

Houses by sripfoto from NounProject.com

If the expected effect size is small, the evaluation will need a larger sample

When the images differ a lot (effect size is larger), you can still see those differences even when the images are relatively small (sample size is smaller)

Castle by Eucalyp from NounProject.com;
House by sripfoto from NounProject.com

How to determine sample size

Think about:

- How many individuals does the program serve?
- Over how long?
- What is the demand for the program?
- How big a sample does the study require? (Consider the burden of research participation)
- How will sample be allocated to the program group vs. the comparison group? (Covered soon)
- Will the sample be clustered? (Covered soon)

Key inputs in determining statistical power

We've discussed:

- Effect size (including take-up rate)
- Sample size (including attrition)

Other important factors:

- Variation in the outcome
- Proportion of the sample in the treatment group
- Unit of randomization

Variation in the outcome

- What: How similar are individuals in your sample to each other (on the outcome of interest)?
- How does it affect power? If the underlying population has high variation in outcomes, the evaluation needs a larger sample
- How to determine? Program data (for example, if measuring test scores, previous test scores are a good source of information)
- What to do about it: Control variables; Program targeting

Why does variation matter?

This relates to the earlier discussion on statistical significance and natural variation!

Before intervention

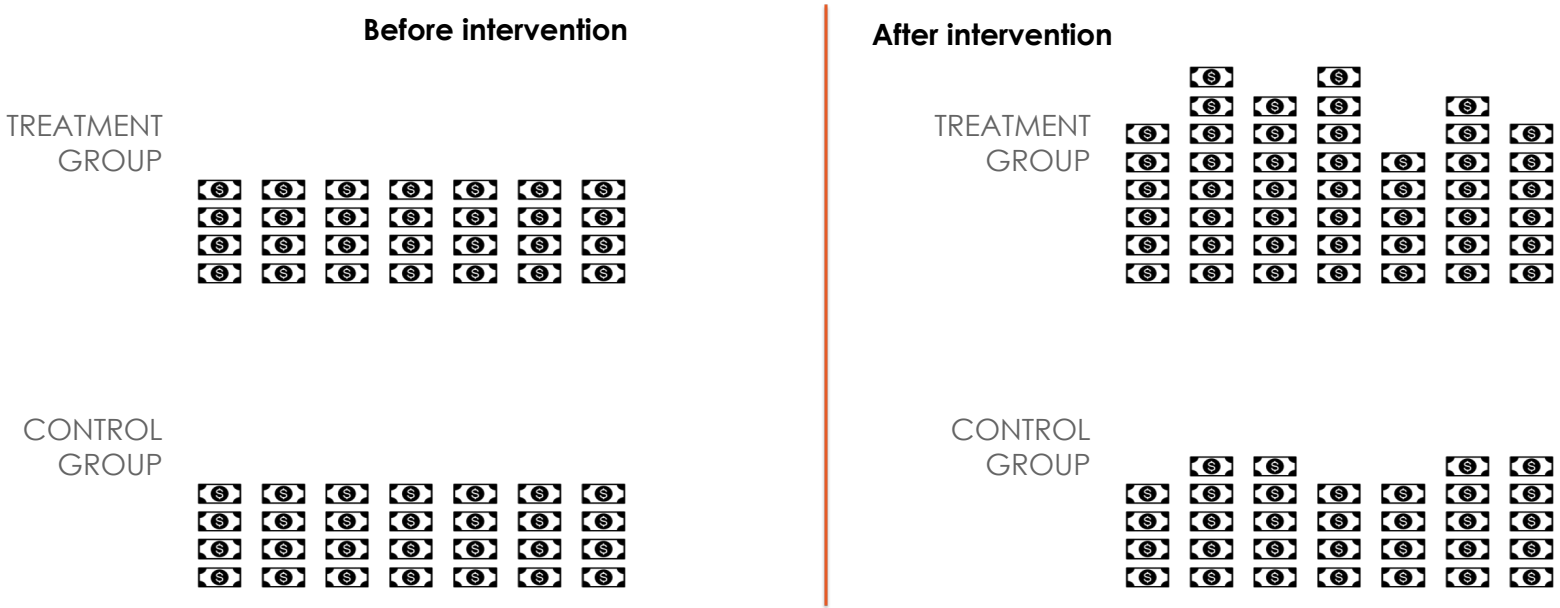
TREATMENT GROUP

CONTROL GROUP

Before the intervention, the treatment and control groups have exactly the same savings levels and **low variation** across participants

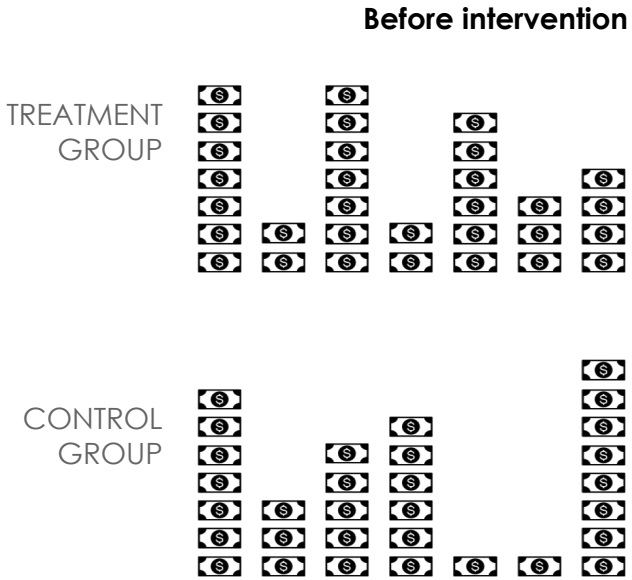
Dollar bills by Hasanudin, Noun Project

Why does variation matter?



Dollar bills by Hasanudin, Noun Project

Why does variation matter?

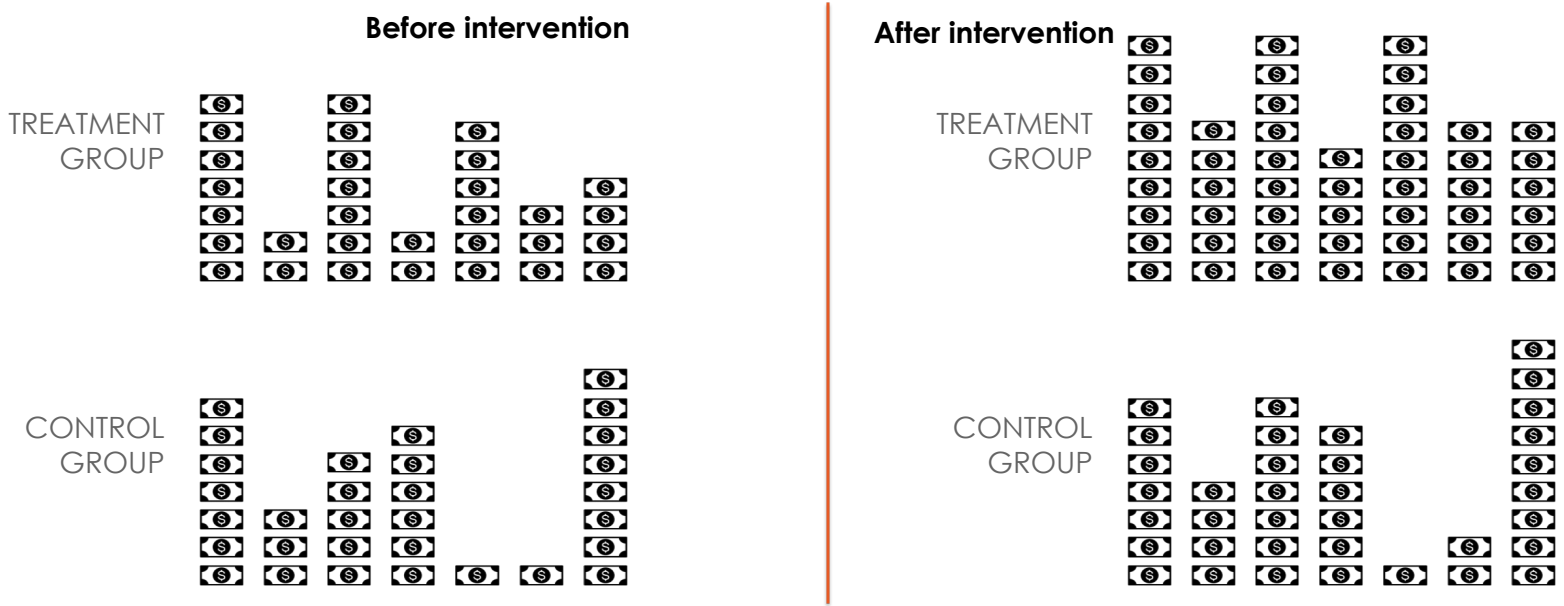


Before the intervention, the treatment and control groups have exactly the same average savings levels and **high variation** across participants

Example: study of an intervention aiming to increase the amount of money that people save. Variation in initial money in savings accounts.

Dollar bills by Hasanudin, Noun Project

Why does variation matter?



Example: study of an intervention aiming to increase the amount of money that people save. Variation in initial money in savings accounts.

Dollar bills by Hasanudin, Noun Project

Proportion of the sample in the treatment group

- What: How the sample size is split up between the program group and the comparison group
  - For example: 65% treatment and 35% comparison
- How does this affect power: An **equal allocation** to treatment and comparison groups will maximize power, all else equal
  - Small deviations from 50/50 have minimal impact on power
  - Increasing the sample size increases power, even if proportion changes
- How to allocate: It depends!
  - Program supply and demand
  - Cost of program vs. cost of data collection

Power is maximized when the sample is equally split between treatment and control groups

- For a fixed sample size, power maximized with even split between treatment and control.

- However, **power increases when the sample size increases** even if that comes through adding individuals just to one group (or unequally)
  - Even you are unable to add more people to the treatment group due to resource constraints, you **could add more individuals to the control group**

Unit of randomization

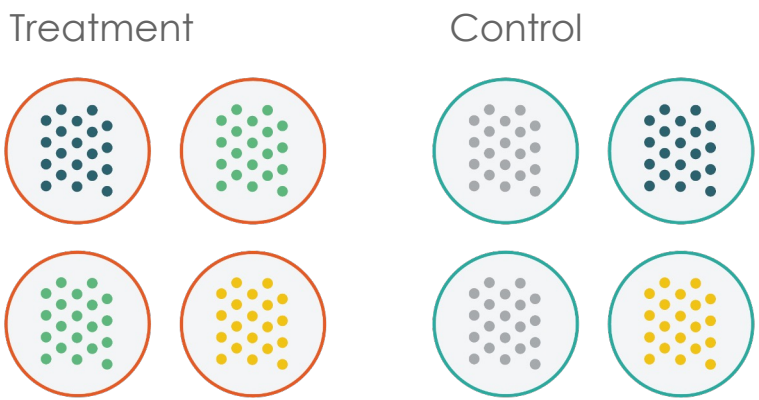
Discussed in "How to Randomize"

- What: the level at which treatment assignment is randomized. A "clustered design" randomizes each group or "cluster" of units to treatment or comparison
  - Example: You might randomize the financial literacy intervention at the community center or neighborhood level rather than the individual level
  - Even when you randomize at the community level, you are still measuring data at the individual level
- How does this affect power: All else equal, randomizing at the individual level maximizes power
- How to choose? It depends! Often a feasibility consideration

Unit of randomization, continued

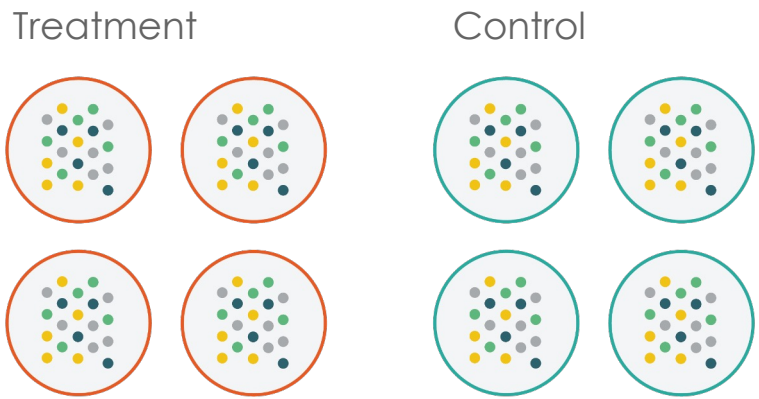
- You can increase power by adding more clusters and/or by adding more units to existing clusters
  - Example: adding more community centers (adding more clusters) and adding more individuals from within a community center to the program (adding more units to existing clusters)
- Typically, the addition of more clusters will increase power more than the addition of more units to existing clusters
- The optimal number of clusters and the size of each cluster will depend on how similar individuals within a cluster are to each other, the "intra-cluster correlation coefficient" (ICC)

When randomizing at the cluster level, the more similar the outcomes of individuals within clusters are, the larger the sample needs to be



More similarity within clusters, larger sample needed

When randomizing at the cluster level, the more similar the outcomes of individuals within clusters are, the larger the sample needs to be



More variation within clusters, smaller sample needed

Calculating **statistical power**

- Many elements go into determining the power of an evaluation
  - A key element that we can often control is the **sample size**
- There are multiple levers we can pull to increase power
  - Some are a lot more powerful than others
  - Not everything is in our control
- Two ways to approach power calculations:
  - **If sample size is flexible:** What sample size has a "good chance" of picking up an effect (if there is one)?
  - **If sample size is fixed:** What effect size would you be able to "pick up," and is this effect reasonable?

Power calculation formula (one variation)

N = sample size

Standard values for the designated significance and power levels

σ^2 = variance

$$N = (t_{1-\kappa} + t_{\frac{\alpha}{2}})^2 \frac{1}{P(1-P)} \times \frac{\sigma^2}{MDE^2}$$

1-κ = statistical power
α = significance level

By convention, we aim for 80% power and set α to 5%.

P = proportion of the sample in the program group (a.k.a. treatment group)

MDE (minimum detectable effect) is the smallest effect size that can be detected given the other inputs. (This might factor in participation or "take up rate")

Don't worry about this, this slide is mostly for your reference later on!

More detail: https://www.povertyactionlab.org/resource/power-calculations

Rules of thumb for statistical power

All else equal:

- ↑ sample size = ↑ power
  - ↑ attrition = ↓ power
- ↓ expected effect = ↑ sample size needed
  - ↓ take-up = ↑ sample size needed
- ↑ variation in outcomes = ↑ sample size needed
- Power is maximized when sample evenly split between T and C
- When individuals within clusters are similar ⇒ ↑ clusters needed

Recap: Essentials of Power and Sample Size

- Power is the probability of detecting an impact of a given size if there is one
- Power is about setting your study up for success in a world where there is statistical uncertainty about what underlying effects really are
- Goal: balance the risk of false positives and false negatives so that study results can be informative, policy-relevant, and an efficient use of resources

Additional resources

Written resources

- J-PAL Research Resource: Power calculations
- J-PAL Research Resource: Quick guide to power calculations
- J-PAL Blog: Six rules of thumb for understanding statistical power
- EGAP: 10 Things to Know About Statistical Power

Tools and code

- EGAP's Power Calculator is a useful tool to test how the power changes with different parameters for a simple study design with individual or cluster randomization
- Optimal design (instructions) can be used for complex study designs
- power command in STATA and the pwrcalc package in R can be useful
- J-PAL has sample code for conducting power calculations using in-built commands and simulations in both Stata and R on GitHub

Reuse and citation

To reference this lecture, please cite as:

J-PAL. "Lecture: Essentials of Sample Size and Power." Abdul Latif Jameel Poverty Action Lab. 2023. Cambridge, MA