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. Randomized Evaluation from Start to Finish
7. Threats & Analysis
8. Ethical Considerations
9. Generalizing & Applying Evidence

Learning Objectives

- 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

Motivation and definitions

I. Sampling variation, false positives, and false negatives
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

Session outline

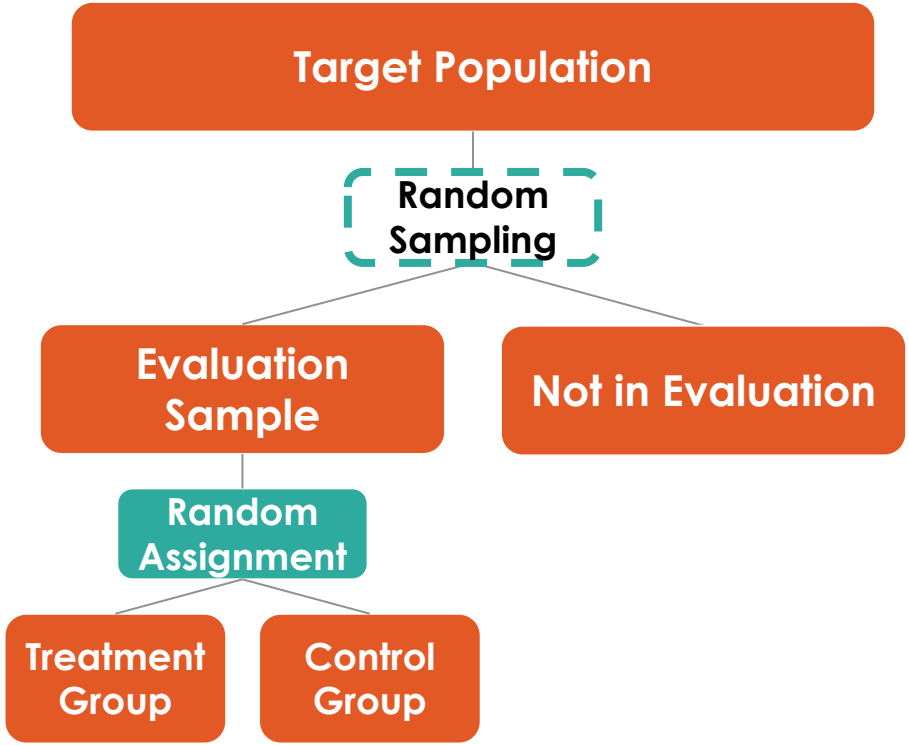
Motivation and definitions

I. **Sampling variation, false positives, and false negatives**
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

Key concepts: sampling & variation



Key concepts: sampling & variation

Why care about the sample size?

55Dental Kids Toothbrush Set of Soft Giraffe Toothbrush for Kids 3-9. Easy-Grip, Bristle Cover, Self-...

Kid

4

newrichbee 8 Packs Kids Toothbrushes, Extra Soft Lovely Little Deer Toothbrush for Kids 2-...

Kid · 8 Count (Pack of 1)

5,051

How does this relate to RCTs and impact evaluation?

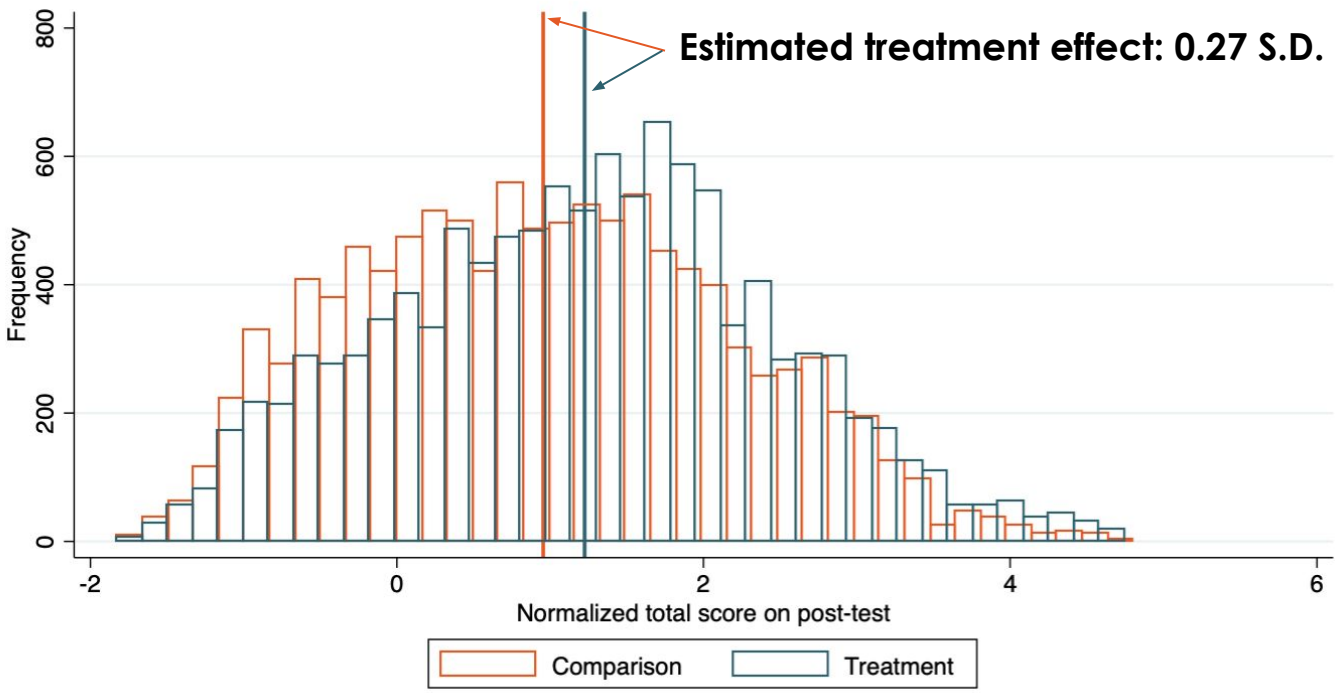
The underlying logic is similar:

- 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!

Example: Balsakhi tutoring program

Study: Balsakhi remedial tutoring program in India

Sample size: More than 23,000 students



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.

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

Samples of size 200 drawn from the same underlying data

Sample 1 Estimate	Sample 2 Estimate	Sample 3 Estimate	Overall Treatment Effect
0.35	0.07	-0.28	0.27

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

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.

Larger samples lead to less random variation in treatment effects

Samples of size 2,000 drawn from the same underlying data

Sample 1 Estimate	Sample 2 Estimate	Sample 3 Estimate	Overall Treatment Effect
0.24	0.23	0.26	0.27

When the sample size is larger, any observed difference is more likely to be caused by the program than sampling variation.

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.

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**
Conclude no impact		
Conclude impact		

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

Evaluation results	Reality/underlying truth: No impact	Reality/underlying truth: Impact
Conclude no impact		
Conclude impact		

Evaluation results versus underlying truth

Reality/underlying truth

Evaluation results	**No impact**	**Impact**
Conclude no impact	**GREAT!**	**False negative:** you conclude there is no impact when there is (type II error) **Mismatch!**
Conclude impact	**False positive:** you conclude there is impact when there is not (type I error) **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)?

Session outline

Motivation and definitions

I. Sampling variation, false positives, and false negatives
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

How to "detect an effect"

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

We also want to balance the risks of false positives and false negatives

We do this by conducting a thought experiment (hypothesis testing)

Statistical significance

Ask: How likely would it be to observe this outcome due to random chance alone (natural variation in participants)?

- If it is reasonably unlikely (5% probability or less), then we conclude that the result was **statistically significant**

  → This is what we mean when we say "detect an effect"

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

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

Statistical power

Ask: If there were a true underlying effect, how likely would we be able to detect it in this experiment?

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 recap: statistical significance & statistical power

Statistical significance

- Avoiding a false positive (falsely concluding there was an impact when there is none)
  - "Rejecting the null hypothesis"
  - "Detecting an effect"
- 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 a false negative (type II error) 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

Session outline

Motivation and definitions

I. Sampling variation, false positives, and false negatives
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

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?

Source: Wikimedia Commons

In other words:

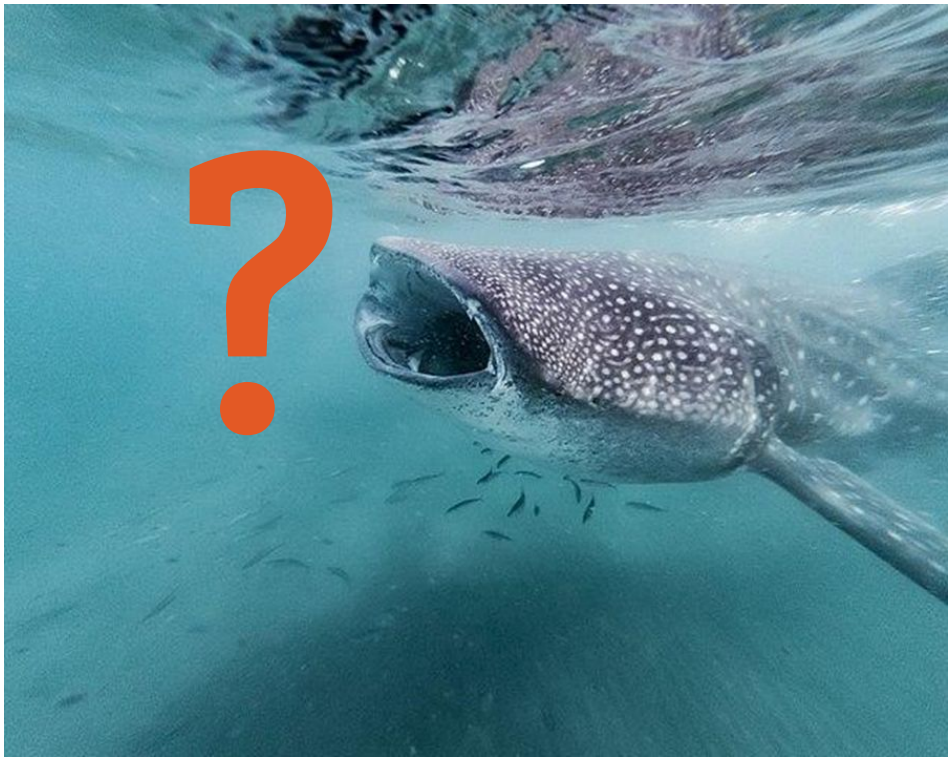
If I tell you there are sharks in this water, is there evidence to support my claim?

Absence of evidence

Source: Wikimedia Commons

This photo represents an "absence of evidence" of sharks

Absence of evidence



Source: Wikimedia Commons

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

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

Session outline

Motivation and definitions

I. Sampling variation, false positives, and false negatives
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

Key inputs in determining statistical power

- Effect size: the minimum detectable effect (MDE) for a given power level
  - Accounting for the take up rate
- Sample size
  - Accounting for attrition
- Variation in the outcome
- Proportion of the sample in the program group
- Unit of randomization (i.e. clustering)

Power calculations: two approaches

- **If sample size is flexible:** Calculate sample size that ensures 80% power for a given minimum true effect size. Is this sample size reasonable?

- **If sample size is fixed:** Calculate minimum true effect size required to achieve 80% power for a given sample size. Is this effect size reasonable?

If the expected effect size is large, the evaluation will need a smaller 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

Source: NounProject.com

If the expected effect size is large, the evaluation will need a smaller 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)

Source: NounProject.com

If the expected effect size is large, the evaluation will need a smaller 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)

Source: NounProject.com

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

With more subtle differences (effect size is smaller), it is harder to see when the images are relatively small

Source: NounProject.com

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

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

Source: NounProject.com

How to choose an MDE

Discussed in Power Handout

- 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**
  - 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

How does the unit of randomization affect power?

Discussed in "How to Randomize"

- In practice, we often randomize at units larger than the individual, while still measuring outcomes at the individual-level
  - Schools, classrooms, households, villages
- Challenge: Units within clusters are not independent of one another
  - Students from same school likely to have similar family income, test scores, etc.
  - People within households likely to have similar levels of education, political preferences, etc.
- Impact of clustering depends on how "similar" units within a given cluster are (intra-cluster correlation coefficient, ranging from 0 to 1)

Example: Clustering and power

- Research question: Who will win the next local election in your town?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of who will win
- Who do you poll?:
  a) All four people in 50 households
  b) One person in 200 households
  c) Somewhere in between

Example: Clustering and power

- Research question: Who will win the next local election in your town?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of who will win
- Who do you poll?:
  a) All four people in 50 households
  b) One person in 200 households
  c) Somewhere in between

High intra-cluster correlation:
Units within clusters are very similar to each other → adding more units within a cluster adds little information about the underlying distribution

Example: Clustering and power

- Research question: Do people prefer strawberry or raspberry flavor?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of what people prefer
- Who do you poll?:
  a) All four people in 50 households
  b) One person in 200 households
  c) Somewhere in between

Example: Clustering and power

- Research question: Do people prefer strawberry or raspberry flavor?
  - Population consists of 10,000 inhabitants: 2,500 households with 4 people in each
- You have resources to poll 200 people and want to get the best possible estimate of what people prefer
- Who do you poll?:
  - **a) All four people in 50 households**
  - b) One person in 200 households
  - **c) Somewhere in between**

Low intra-cluster correlation:

Units within clusters are not very similar to each other → adding more units within a cluster or adding new clusters both add information about the underlying distribution

Clustering and power

- 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 more similar individuals within clusters are, the larger the sample needs to be
  - 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)

Session outline

Motivation and definitions

I. Sampling variation, false positives, and false negatives
II. Statistical power and statistical significance
III. Importance of avoiding an underpowered study

Power calculations and designing well-powered studies

IV. Components of power calculations
V. Rules of thumb for power and sample size

Rules of thumb for statistical power

Component	Relationship to Power	Relationship to MDE
1 True effect size increases	**Increases** power	N/A
2 Sample size increases	**Increases** power	**Decreases** the MDE
3 Take up increases	**Increases** power	**Decreases** the MDE
4 Outcome variance decreases	**Increases** power	**Decreases** the MDE
5 Equal treatment allocation	**Increases** power	**Decreases** the MDE
6 ICC increases	**Decreases** power	**Increases** the MDE
7 Attrition increases	**Decreases** power	**Increases** the MDE

Design levers to improve power

- Increase the **sample size**
  - Increase the number of units or clusters
  - Reduce attrition
  - Conduct individual-level random assignment when possible
- Increase the **effect size**
  - Increase the intensity of the treatment
  - Increase take-up/compliance
- Simplify the **design**
  - Reduce the number of treatment arms or the number of hypotheses you test
  - The study needs to be powered for the smallest MDE among the intended comparisons
- Reduce **variation** in the outcome
  - Stratify the randomization to ensure baseline balance on important observables
  - Control for covariates (especially baseline measures of the outcome) to reduce residual variance

Practical tips for conducting power calculations

- Perform power calculations early in the design phase (before the program is implemented)
- Don't panic about the number of assumptions required
  - Power calculations should be considered *a rough guide* in the decision of whether to carry out the study and provide an *estimate* of how large the sample should be.
- Conduct sensitivity analyses to test how power changes with changes to any critical assumptions
  - Create "best case" scenarios and "worst case" scenarios and evaluate those
  - If the best case scenario MDE is unrealistically high or requires an unrealistically large sample size, consider how to tweak the design to increase power
  - If sufficient power cannot be achieved, an RCT might not be the best way forward

Power calculation formula (solving for sample size)

σ^2 = variance

$$N = (1.96+0.84)^2 \cdot \frac{1}{P(1-P)} \cdot \frac{\sigma^2}{MDE^2}$$

N = sample size

Standard values for the designated significance and power levels

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, this slide is mostly for your reference later on!

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

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

- 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

Appendix

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

Compliance/take-up

- What: Actual participation rates in the program in treatment and comparison groups
- How does it affect power? Low participation rates decrease effect sizes because everyone randomized is included in analysis but only those who participate will benefit from the program
- Example: A financial literacy program causes a $50 average increase in savings for the treated group
  - If 100% participate: the average increase is $50
  - If 50% participate: the average increase is $25
- How to address: Knowledge of sample & program; intake & enrollment process; incentives

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?

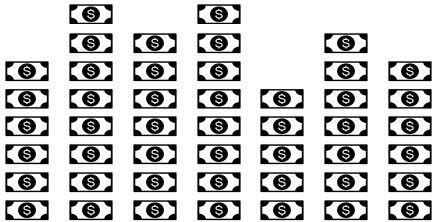
Before intervention

TREATMENT GROUP

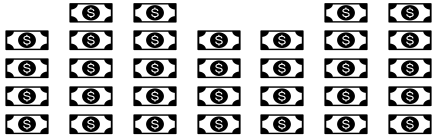
CONTROL GROUP

After intervention

TREATMENT GROUP

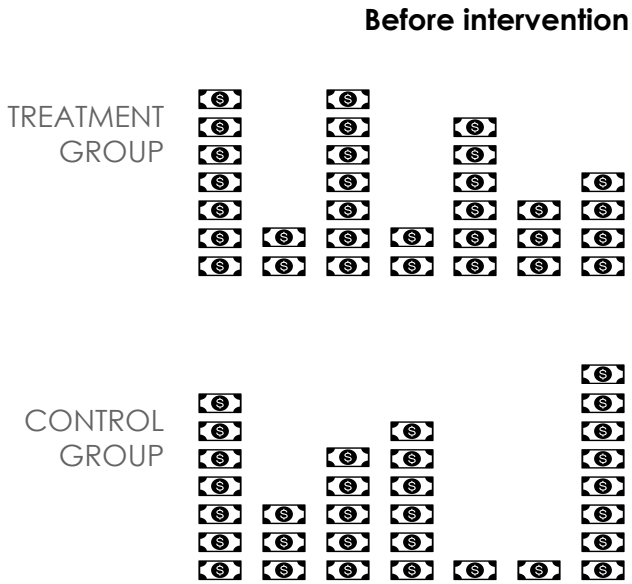


CONTROL GROUP



Dollar bills by Hasanudin, Noun Project

Why does variation matter?

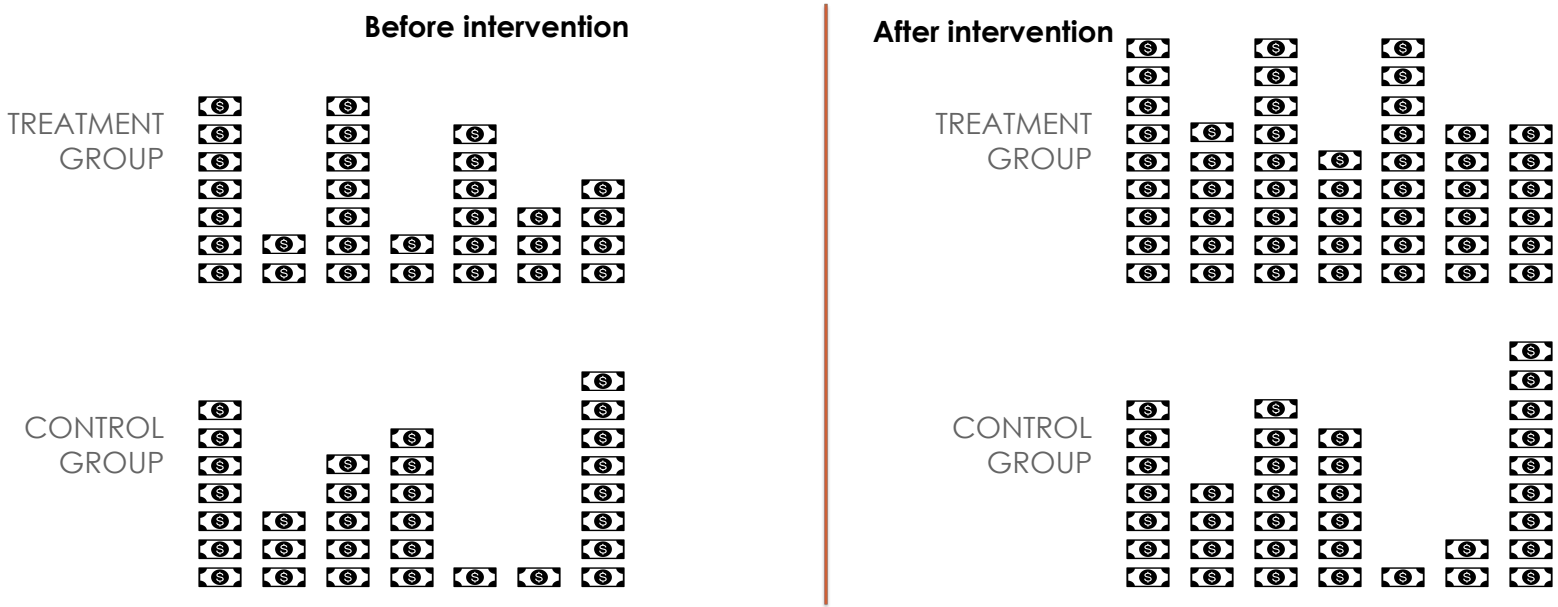


Before the intervention, the treatment and control groups have exactly the same average savings levels and **high variation** across participants

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

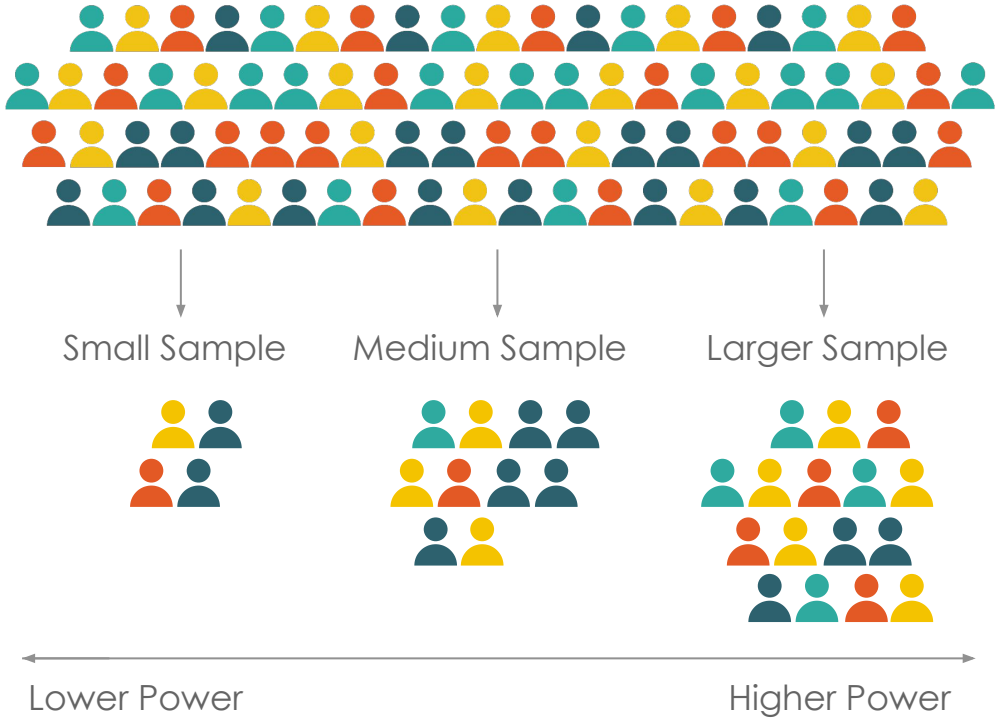
- For a fixed sample size, power maximized with even split between treatment and control.

- 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**

Sample size

- What: The number of individuals or units in your evaluation
  - Remember to account for **attrition**! Attrition can be either someone dropping out of your program or missing from data collection. Consider the sample size at the end of the evaluation.
- 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

A larger sample increases the power of the evaluation



How to determine sample size

- How many individuals could you enroll?
  - Budget
  - Timeline
  - Total number of units available
- What is the demand for the program?
- How will sample be allocated to the program group vs. the comparison group? (Covered soon)
- Will the sample be clustered? (Covered soon)

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

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

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. 2024. Cambridge, MA