SIX RULES OF THUMB FOR DETERMINING SAMPLE SIZE AND STATISTICAL POWER

THE ABDUL LATIF JAMEEL POVERTY ACTION LAB (J-PAL)

J-PAL
ABDUL LATIF JAMEEL POVERTY ACTION LAB
NORTH AMERICA

SIX RULES OF THUMB FOR DETERMINING SAMPLE SIZE AND STATISTICAL POWER

SUMMARY:

The ability of an evaluation to detect a meaningful impact of a program is determined by the evaluation's sample size and statistical power. This is a tool for policymakers and practitioners that describes some of the factors that affect statistical power and sample size. Further information on the dangers of running an evaluation with inadequate power can be found in a companion resource available here.

WHAT IS STATISTICAL POWER, AND WHAT IS THE PURPOSE OF STATISTICAL POWER ANALYSIS?

The statistical power, or power, of an evaluation reflects the likelihood of detecting any meaningful changes in an outcome of interest brought about by a successful program. In the process of designing a randomized evaluation, researchers conduct power analyses to inform decisions such as:

- Whether to conduct the evaluation
- At which unit to randomize (e.g., individual, household, or group)
- How many units to randomize
- How many units or individuals to survey
- How many times to survey each unit or individual over the course of the evaluation
- How many different program alternatives to test
- How much baseline information to collect
- Which outcomes to measure
- How to measure the outcomes of interest

It is important to understand how the factors above are interrelated and affect the overall power and sample size needed for a randomized evaluation. The rules of thumb outline the key relationships between the determinants of statistical power and sample size, and demonstrate how to design a high-powered randomized evaluation.

COVER PHOTO: ANTON GVOZDIKOV | SHUTTERSTOCK.COM

SIX RULES OF THUMB FOR DETERMINING SAMPLE SIZE AND STATISTICAL POWER

RULE OF THUMB #1: A LARGER SAMPLE INCREASES THE STATISTICAL POWER OF THE EVALUATION

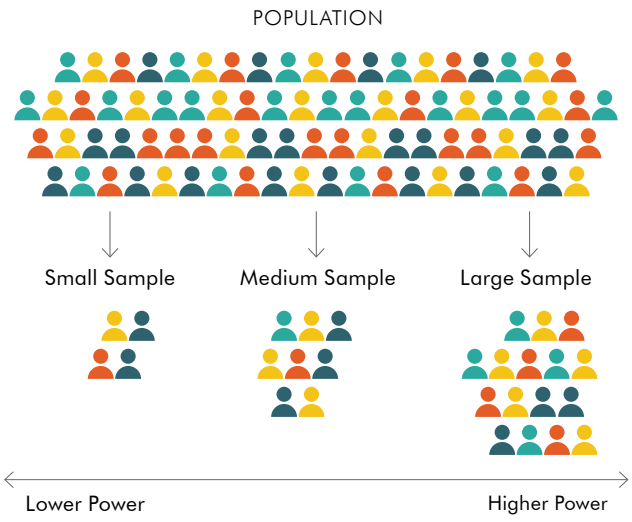
Researchers run evaluations on samples that are selected from a larger population. When designing an evaluation, the research team must determine the number of participants to include in the sample.

In the extreme scenario, a researcher would be able to include the whole population of interest in the study sample. In this case, the sample is the population and is therefore the best representation of the population. However, in most cases, the study sample is a subset of the broader population.

Larger samples are more likely to be representative of the original population (see Figure 1.1) and are more likely to capture impacts that would occur in the population. Additionally, larger samples increase the precision of impact estimates and the statistical power of the evaluation.

When designing an evaluation, it is important to take into account expected levels of attrition, since attrition reduces sample size and power. If you anticipate that you will not be able to collect outcome data on some study participants, increase your initial sample size to ensure that you will have sufficient power to detect the impact of the program at the conclusion of the intervention.

FIGURE 1.1

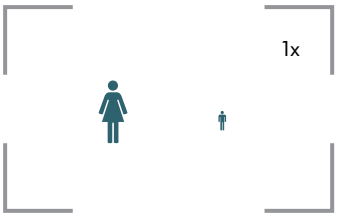


RULE OF THUMB #2: IF THE EFFECT SIZE OF A PROGRAM IS SMALL, THE EVALUATION NEEDS A LARGER SAMPLE TO ACHIEVE A GIVEN LEVEL OF POWER

The effect size of an intervention is the magnitude of the impact of the intervention on a particular outcome of interest. When designing an evaluation, the research team wants to ensure that they are able to identify the effect of the program with precision. When an evaluation has sufficient power, impact estimates are precise. Both the effect size and sample size affect precision.

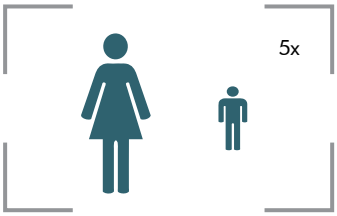
We can think about precision by thinking about the confidence with which images can be identified based on their size. Imagine that you are looking at the images in Figure 2.1 through a camera lens. Which one can you more precisely identify as a female? Is the image on the right the same as the image on the left?

FIGURE 2.1



Zooming in, we see that the image on the right is different from the image on the left. This difference is easier to identify when we increased our zoom. Large images can be precisely identified without much zoom, while smaller images require more zoom.

FIGURE 2.2



The size of the images represents the effect size, and the level of zoom represents the sample size of the evaluation. For a given level of power, large effects can be precisely detected with a smaller sample size, while smaller effects can only be precisely detected with larger sample sizes.

Think of a larger sample as allowing you to zoom in on a smaller effect size, or image. A larger image requires less zoom, or a smaller sample. A smaller image requires more zoom, or a larger sample.

RULE OF THUMB #3: AN EVALUATION OF A PROGRAM WITH LOW TAKE-UP NEEDS A LARGER SAMPLE

Randomized evaluations are designed to detect the average effect of a program over the entire sample that is assigned to the treatment group. Therefore, lower take-up decreases the magnitude of the average effect of the program. Since a larger sample is required to detect a smaller effect (see rule of thumb #2), it is important to plan ahead if low take-up is anticipated and run the evaluation with a larger sample.

To illustrate the relationship between take-up, effect size, and sample size, consider this simplified example: four households are randomly selected to receive encouragement to enroll in a program, and four do not receive encouragement to enroll. Once a household enrolls and participates, the program is expected to increase savings by $100 for each household that participates.[1]

If, as in Figure 3.1, 100 percent of the treatment group enrolls in the program, the average effect of the program, or the effect size, is $100.

If, as in Figure 3.2, only 50 percent of the treatment group enrolls in the program, the effect size is $50.

FIGURE 3.1

EFFECT SIZE*

$$\left(\frac{100 + 100 + 100 + 100}{4}\right) - \left(\frac{0}{4}\right) = \$100$$

Treatment Control ✓ Enroll in program ($100 in savings)

* i.e., average difference between treatment and control

FIGURE 3.2

EFFECT SIZE*

$$\left(\frac{100 + 0 + 100 + 0}{4}\right) - \left(\frac{0}{4}\right) = \$50$$

Treatment Control ✓ Enroll in program ($100 in savings)

* i.e., average difference between treatment and control

[1] We use eight households purely for illustrative purposes; an actual evaluation would need many more households to have sufficient statistical power.

RULE OF THUMB #4: IF THE UNDERLYING POPULATION HAS HIGH VARIATION IN OUTCOMES, THE EVALUATION NEEDS A LARGER SAMPLE

Say a nutrition and exercise program is implemented in schools to decrease the rate of childhood obesity. First, consider a scenario in which there is no variation in the incidence of obesity as measured by Body Mass Index (BMI); each student has the same BMI. Absent the program and absent a randomized evaluation, if you observe the average BMI for the entire group over a given period of time, you would expect to see little change in BMI for the entire group.

If you conducted a randomized evaluation, and introduced the program to the randomly-selected treatment group, you might see that the BMI in the treatment group drops (Figure 4.1). In this case, you can be confident that this effect can be attributed to the program.

Alternatively, consider a scenario in which there is high variation in the incidence of obesity as measured by Body Mass Index (BMI); some students have a high BMI and others have a low BMI. Absent the program and randomized evaluation, over a given period of time, the BMI for the sample might change due to naturally occurring variation within the population.

When the program is administered to the randomly-selected treatment group, you might observe that the BMI in the treatment group drops (Figure 4.2), but since BMI varies within the population, it is more challenging to attribute this change in BMI to the program rather than to the natural variation in BMI within the sample.

If the evaluation is conducted on this high-variance sample, we still do not know whether the nutrition and exercise program caused the average BMI in the treatment group to fall, or whether the change in average BMI of the treatment group is due to naturally occurring variation that was present before the program was introduced.

In a population with high variation in key outcome measures (e.g., BMI), it is challenging to disentangle the effect of the program from the effect of random variation in these outcome measures.

Especially when running an evaluation on a population with high variance, selecting a larger sample increases the likelihood that you will be able to distinguish the impact of the program from the impact of naturally occurring variation in key outcome measures. Larger samples in the presence of high variance make it easier to identify the causal impact of a program (Figure 4.3).

FIGURE 4.1

Low variation in BMI

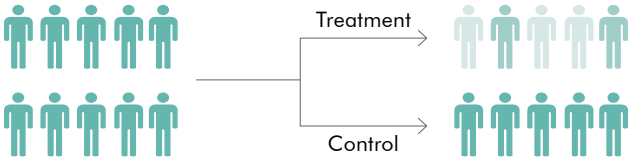
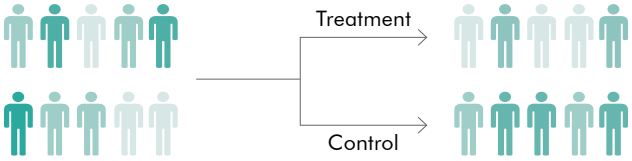


FIGURE 4.2

High variation in BMI



Key for Figure 4.1 and 4.2:

FIGURE 4.3

RULE OF THUMB #5: FOR A GIVEN SAMPLE SIZE, POWER IS MAXIMIZED WHEN THE SAMPLE IS EQUALLY SPLIT BETWEEN THE TREATMENT AND CONTROL GROUP

To achieve maximum power for a given sample size, the sample should be evenly divided between the treatment group and control group. If you have the opportunity to add study participants, regardless of whether you add them to the treatment or control group, power will increase because the overall sample size is increasing. However, the most efficient way to increase power by expanding the sample size is to add participants to achieve or maintain balance between the treatment and control groups.

Why might you split a sample unevenly between the treatment and control groups?

Taking resource constraints, intervention costs, data collection costs, and multiple treatment arms into account, research teams may decide on an uneven ratio of treatment to control participants. Adding treatment participants to a study is likely more expensive than adding control participants.

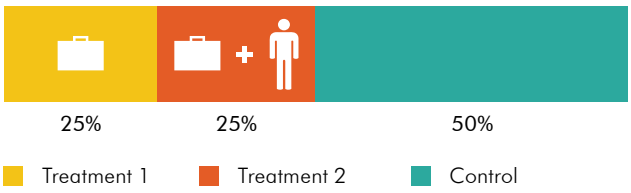
Evaluations with multiple treatment arms (i.e., different versions or combinations of treatments) help researchers to disentangle mechanisms, determine which aspect of a treatment bundle drives impact, and identify whether the components of the treatment bundle are complements or substitutes. The main research questions inform decisions made regarding the proportion of individuals assigned to each arm. Unequal proportions may be an optimal decision to ensure that the evaluation is sufficiently powered to answer the questions of interest.

For example, say you want to evaluate the impact of an employment program with two, potentially complementary components. A research team might design an evaluation with two treatment arms and one control arm. Of the individuals who apply to the program, 25 percent could be offered a job, 25 percent could be offered a job and a career coach, and the remaining 50 percent of participants could be randomized into the control group.[2]

A research team might design their study in this way so that they can examine the pooled impact of both treatments compared to the control condition. With a large sample, this allocation strategy also equips researchers to compare the impact of the job-only treatment to the job and career coach treatment. However, since the sample is cut in half to compare the job-only group to the job and career coach group, the evaluation has less power to detect relative impact of the two treatments than to detect the pooled impact of both treatments compared to the control condition. Additionally, if the sample is not large enough, the evaluation may not have sufficient power to compare the job-only treatment to the career coach and job treatment.

FIGURE 5.1

ALLOCATION OF SAMPLE TO STUDY ARMS



[2] Alternatively, a research team might decide to assign equal proportions of the sample to each of the three groups (i.e., 33 percent treatment one, 33 percent treatment two, 33 percent control).

RULE OF THUMB #6: FOR A GIVEN SAMPLE SIZE, RANDOMIZING AT THE CLUSTER LEVEL AS OPPOSED TO THE INDIVIDUAL LEVEL REDUCES THE POWER OF THE EVALUATION. THE MORE SIMILAR THE OUTCOMES OF INDIVIDUALS WITHIN CLUSTERS ARE, THE LARGER THE SAMPLE NEEDS TO BE.

When designing an evaluation, the research team must choose the unit of randomization. For example, individuals can be randomly assigned to the treatment group or control group. Alternatively, randomization can be done by "clusters." By this method, groups of individuals are treated as units, whether they are households, classrooms, schools, or neighborhoods, and each cluster is randomly assigned to the treatment group or the control group.

For a given sample size, randomizing clusters as opposed to individuals decreases the power of the study. The reason for this relates to how similar the outcomes of individuals within a cluster are to each other. Consider a classroom-level evaluation with eight classrooms, each of which has twenty students. In this case, each classroom is a cluster.

If students within a classroom are similar – in observable (e.g., GPA) or unobservable (e.g., level of motivation) ways – their outcomes (e.g., test scores and attendance) will likely be similar as well. In the extreme scenario, everyone within a classroom is identical, and their outcomes are identical. Dots of the same color represent identical individuals in Figure 6.1. In this case, the power of the evaluation would reflect the power of an individual-level randomized evaluation that only sampled eight students.

On the other extreme, when each classroom contains students with differing characteristics, as depicted in Figure 6.2, even though randomization was implemented at the cluster, or classroom, level, the evaluation would be powered as if randomization occurred at the individual level.

Usually, the number of clusters is a bigger determinant of power than the number of people per cluster. Therefore, if you are looking to increase your sample size, and individuals within a cluster are similar to each other on the outcome of interest, the most efficient way to increase the power of the evaluation is to increase the number of clusters rather than increasing the number of people per cluster.

FIGURE 6.1

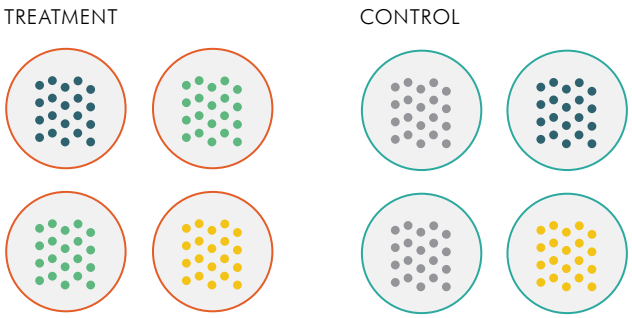
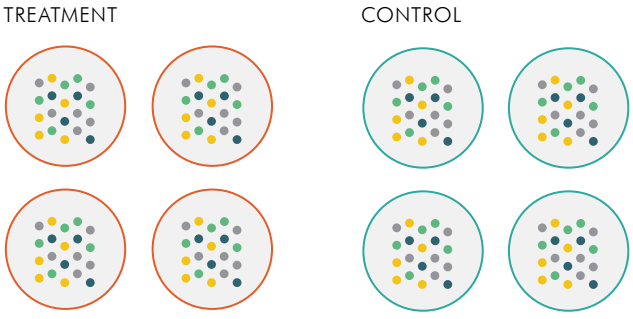


FIGURE 6.2



Key for Figure 6.1 and 6.2:

Cluster Treatment Control

GLOSSARY

Attrition: When individuals drop out of the control or treatment group over the course of the evaluation.

Effect size: The magnitude of the impact of an intervention on a particular outcome of interest.

Statistical power: The likelihood that an evaluation will be able to detect a treatment effect of a certain size.

FURTHER READING

Glennerster, Rachel, and Kudzai Takavarasha. 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton: Princeton University Press.

J-PAL. 2017. "Sampling and Sample Size." Accessed May 26. https://www.povertyactionlab.org/sites/default/files/documents/L5_Sampling%20and%20Sample_Glennerster_2016-06-15.pdf.