

7.7.1: Power and Sample Size

Large N and Small Effects

There is an intriguing relationship between N (sample-size) and power. As N increases, so does power to detect an effect. Additionally, as N increases, a design is capable of detecting smaller and smaller effects with greater and greater power. Let's think about what this means.

Imagine a drug company told you that they ran an experiment with 1 billion people to test whether their drug causes a significant change in headache pain. Let's say they found a significant effect (with power = 100%), but the effect was very small, it turns out the drug reduces headache pain by less than 1%, let's say 0.01%. For our imaginary study we will also assume that this effect is very real, and not caused by chance.

Clearly the design had enough power to detect the effect, and the effect was there, so the design did detect the effect. However, the issue is that there is little practical value to this effect. Nobody is going to buy a drug to reduce their headache pain by 0.01%, even if it was "scientifically proven" to work. This example brings up two issues. First, increasing N to very large levels will allow designs to detect almost any effect (even very tiny ones) with very high power. Second, sometimes effects are meaningless when they are very small, especially in applied research such as drug studies.

These two issues can lead to interesting suggestions. For example, someone might claim that large N studies aren't very useful, because they can always detect really tiny effects that are practically meaningless. On the other hand, large N studies will also detect larger effects too, and they will give a better estimate of the "true" effect in the population (because we know that larger samples do a better job of estimating population parameters).

Additionally, although really small effects are often not interesting in the context of applied research, they can be very important in theoretical research. For example, one theory might predict that manipulating X should have no effect, but another theory might predict that X does have an effect, even if it is a small one. So, detecting a small effect can have theoretical implication that can help rule out false theories. Generally speaking, researchers asking both theoretical and applied questions should think about and establish guidelines for "meaningful" effect-sizes so that they can run designs of appropriate size to detect effects of "meaningful size".

Small N and Large Effects

Note

All other things being equal, would you trust the results from a study with small N or large N ?

This isn't a trick question, but sometimes people tie themselves into a knot trying to answer it. We already know that large sample-sizes provide better estimates of the distributions the samples come from. As a result, we can safely conclude that we should trust the data from large N studies more than small N studies.

At the same time, you might try to convince yourself otherwise. For example, you know that large N studies can detect very small effects that are meaningless in real life. You also know that small N studies are only capable of reliably detecting very large effects. So, you might reason that a small N study is better than a large N study because if a small N study detects an effect, that effect must be big and meaningful; whereas, a large N study could easily detect an effect that is tiny and meaningless.

This line of thinking needs some improvement. First, just because a large N study can detect small effects, doesn't mean that it *only* detects small effects. If the effect is large, a large N study will easily detect it. Large N studies have the power to detect a much wider range of effects, from small to large. Second, just because a small N study detected an effect, does not mean that the effect is real, or that the effect is large. For example, small N studies have more variability, so the estimate of the effect size will have more error. Also, there is 5% (or alpha rate) chance that the effect was spurious. Interestingly, there is a pernicious relationship between effect-size and type I error rate.

Type I errors: Convincing with Small Samples?

So what is this pernicious relationship between Type I errors and effect-size? Mainly, this relationship is pernicious for small N studies. Imagine a situation in which the null hypothesis is false, that there really is no mean differences between the groups. This

is true, for example, on math performance by gender for children before puberty; girls and boys do equally well on math before the age of 13.

We know that under the null, researchers will find differences between groups that are similar about 5% of the time or less ($p < .05$), that is the definition. So, if a researcher measured math scores on 10-year olds in 100 experiments, they would expect to find a significant difference for 5 of their experiments by random chance. If the sample size is small, we are more likely to accidentally find a gender difference because only a few girls that have more experience with math could shift the mean of the girls' math scores higher. This is the pernicious aspect. When you make a type I error for small N , your data will make you think there is no way it could be a type I error because the effect is just so big! When N is very large, like 1000 (say, 500 boys and 500 girls), it is very unlikely that you'd get a wonky sample when there is no differences between the groups. It would be very unlikely to get enough girls with more experience in math to actually shift a mean based on a 500 scores. So if you find a difference between the two groups when you have a large sample, it means that there probably is a real difference between the groups.

Probably.

Summary

Power: The chance of detecting a real difference, if there is one, between the sample mean and the population mean.

The easiest way to increase power is to increase the size of the sample (N).

Contributors and Attributions

- [Matthew J. C. Crump](#) ([Brooklyn College of CUNY](#))
- [Dr. MO](#) ([Taft College](#))

This page titled [7.7.1: Power and Sample Size](#) is shared under a [CC BY-SA 4.0](#) license and was authored, remixed, and/or curated by [Michelle Oja](#).

- **12.4: Some considerations** by [Matthew J. C. Crump](#) is licensed [CC BY-SA 4.0](#). Original source: <https://www.crumplab.com/statistics/>.