

How to Select and Evaluate a Sample Size

- Sample size affects the power of a statistical test, or its capacity to detect a true difference
- The required sample size for a study can be calculated *a priori* using a power analysis
- Decisions about sample size and power should be reported, with justification

by ALEXANDER THORPE, GARSTON LIANG, & QUENTIN F. GRONAU

1 Knowledge is Power

There's an old saying that "knowledge is power". In statistics, this is literally true. A test's *power* is its capacity to detect a true difference. There are several ways to increase the power of a statistical test—you can accept a lower standard of evidence, i.e., a higher p-value, or a lower effect size. However, the only way to increase power without compromising on these other factors is by increasing the *sample size*, which means taking more observations or recruiting more participants. The greater the sample, the more powerful the test. Knowledge really is power!

So why not just collect data from as many participants as you can? For starters, it's unethical to ask for more than you need from a participant group, especially when they are from a clinical population and participating could put them at risk. Also, data collection takes time and resources that may not be abundant. Ideally, we should collect data from as many participants as we need to ensure a reliable result, but no more than is necessary. We can identify this ideal sample size by undertaking a *power analysis*. Note that the information below only relates to *frequentist* statistics. In *Bayesian* statistics, power analyses simply tell us how much evidence a sample provides.

2 How to Select a Sample Size

Before we discuss power analysis, we need to define some relevant statistical terms—Type I and Type II errors, and effect size. *Type I errors* are cases where a test returns a false positive result, or says "yes" when the answer is really "no". *Type II errors* are the opposite—cases where a test says "no" when the answer is really "yes". The probability of committing a Type I error is commonly denoted by the Greek symbol α , while the probability of committing a Type II error is denoted by β . Table 1 shows the four possible outcomes of a statistical test. The top-left and bottom-right corners are where we'd like to be. Typically, α is set at .05, and if a test's p-value falls below this value, we are confident of avoiding a Type I error (for more on p-values, see another primer in

Table 1: Possible outcomes of a statistical test and their probabilities.

	Genuinely Positive	Genuinely Negative
Reported Positive	True Positive $1 - \alpha$	Type I Error α
Reported Negative	Type II Error β	True Negative $1 - \beta$

this series, “What a P-Value Can and *Cannot* Tell You”). The value we choose for β determines the power of our test. For example, if it is set at .1, the power of our test is 90%, or $1 - \beta$. This means a 10% chance of committing a Type II error.

An *effect size* is the standard size of the difference between the groups being compared. For example, if a *t*-test produces an effect size of Cohen’s $d=0.5$, the mean of group A is 0.5 standard deviations higher or lower than the mean of group B. The bigger the effect size, the stronger the effect.

2.1 Power Analysis

As a power analysis estimates the required sample size, it should naturally be carried out before data collection begins. There are freely available tools that allow researchers to undertake power analysis, such as G*Power, for a plethora of statistical tests and experimental designs. For the purpose of this primer, let’s imagine a study comparing two groups with an independent-samples *t*-test. The parameters required for the analysis depend on the kind of test, but there are four key parameters that are all closely related—significance (α), power (β), effect size, and sample size. Change one, and the outcome of the analysis will change. We already mentioned we can increase power by increasing α or lowering our expected effect size, but as discussed above, this may not be desirable. The only parameter that you can raise without lowering another is sample size.

We have discussed α and β , but choosing an effect size requires some difference choices. We could choose a value based on conventions. For example, a Cohen’s d is considered *large* if it is above 0.8, *medium* above 0.5, and *small* above 0.2. We could also base the decision on previous literature, as different research areas may expect to see different effect sizes. We could also calculate an effect size based on what would represent a medically significant outcome. However, this requires us to know something about the variance in the population of interest, as it is part of the effect size calculation. Again, this information could be available in the literature.

In our imagined study, we have decided to set α at .05, in line with conventions. We also expect to see a *medium* effect size, or a Cohen’s d of 0.5. Finally, we set β at .05, for a power of 95%. Using these parameters, our required sample size is 210 participants, or 105 in each group. However, if we lower our power to 80%, or $\beta = .2$, our required sample size falls to 128, or 64 participants per group. It must be noted that, in the latter case, we are accepting a 20% chance of committing a Type II error, or missing a true difference. G*Power allows us to plot required sample size against power; Figure 1 shows how power increases as sample size increases. The two required sample sizes from the above analyses are marked with dotted lines.

The above is an example of estimating required sample size by setting the other three key parameters, but we can also estimate any of the other three parameters the same way. For example, we may have only managed to collect data from 96 participants. If we set our α at .05 and our effect size at 0.5, we can estimate the power of our *t*-test—68% or a β of .32. That’s a 32% chance

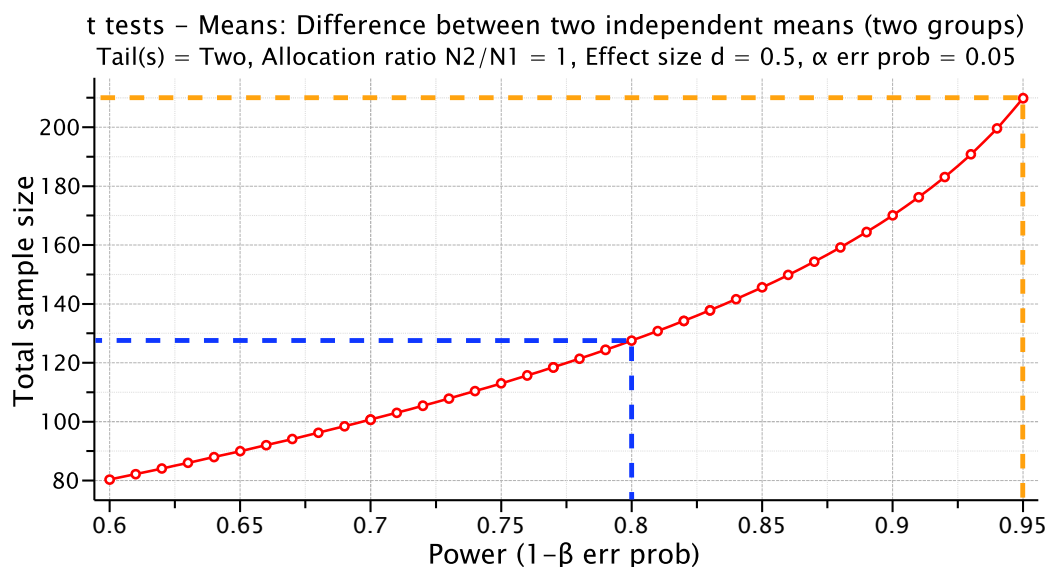


Figure 1: Required sample sizes as a function of power for an example test. Required samples are marked for 80% (blue) and 95% (orange) power.

of landing in the bottom-left of Table 1!

3 Evaluating Sample Sizes

When estimating required sample size, power analysis is a prospective calculation. Running the test after we've already run the study would not tell anything we don't already know. It is therefore not useful to assess previous studies' sample sizes using retrospective power analysis. Instead, it is important to read the Methods and Results sections of previous research critically—do the authors report an *a priori* power analysis? Do they justify their sample size based on statistical or pragmatic necessity, for example patient availability? If the study is longitudinal or uses a test-retest methodology, did they account for participants dropping out? Finally, do the authors justify expected effect sizes, significance, and power? In medical research, there are many legitimate reasons for a small sample size, or even for a study to ultimately be underpowered, but a well-designed (and well-reported) study should have good answers to all the above questions.

4 Further Reading

Below are some open-source resources that go into more detail about choosing sample sizes and power analyses. We have also included a link to download G*Power from the Heinrich Heine Universität Düsseldorf.

- G*Power download: <https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower>
- Pourhoseingholi, M. A., Vahedi, M., & Rahimzadeh, M. (2013). Sample size calculation in medical studies. *Gastroenterology and hepatology from bed to bench*, 6(1), 14–17.

- Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, 31(1), 010502. <https://doi.org/10.11613/BM.2021.010502>
- Uttley, J. (2019). Power Analysis, Sample Size, and Assessment of Statistical Assumptions—Improving the Evidential Value of Lighting Research. *LEUKOS*, 15(2–3), 143–162. <https://doi.org/10.1080/15502724.2018.1533851>