Sample size: More than just power

E-mail distributed on 17-04-2023

Dear all,

I often receive questions about setting sample size for studies, to which I have given similar replies over the years, emphasizing mainly that sample size should not be determined only with power in mind. There is **more to sample size than power**. In fact, I believe there are 6 criteria for deciding what should be a good sample size for your study, and not all of them are statistical:

1) **Power:** As per tradition in frequentist statistics, you want a sample size that can detect the expected effect at a certain significance level and a desired power level. This is the most formal way of setting sample size, with G*Power the most popular tool for calculation in social science. However, this calculation often faces practical problems (e.g., choosing effect size, complex models, etc.), and it may not necessarily satisfy some of the subsequent criteria.

2) **Reliability:** You want a sample size that gives reliable/precise mean estimates. Larger samples will reduce the influence of measurement error on mean/parameter estimates. Here, some literature provides rules-of-thumb for linear regression that say *N* should be equal to at least 50 + (8 × *P*), where *P* is the number of predictors in the model (Kutner et al., 2005), regardless of power considerations.

3) **Distributional convergence:** You want a sample size so that mean/parameter estimates converge to the required distribution for inferential tests (e.g., normal, chi-square). Here, the traditional number cited is 20 or 30. At these numbers, mean estimates often start to converge to a normal distribution regardless of the distribution of the raw data (if they are independent-and-identically distributed data[1]). The convergence will be perfect for infinite sample size, a mathematical truth known as the "central limit theorem" in statistics.

4) **Generalizability:** You want a sample size that enables you to *generalize* to the population of interest. Practically speaking, this criterion is nearly impossible to satisfy for most studies in social science. If the target population is "humanity" (as implied by many studies), thousands would not be enough to generalize to the population, so this criterion could negate almost every other justification on this list. This is why in some fields—such as vaccine-research—clinical trials are required to be run across vast demographical strata, to show that the vaccine is safe for use among all people, regardless of any power considerations. On the other hand, if the target population is very small (e.g., children with a rare disorder), even a small sample

[1] If the distribution has a defined mean and variance

could be representative and afford generalization of results. Generalizability does not always equate to "large" sample size.

5) **Feasability:** Your sample size may be restricted by what is practically available or feasible, given restrictions on time, resources, and exclusion criteria. The rarer your target population, the less realistic it is too obtain a large sample. Likewise, some designs and methods are expensive (e.g., longitudinal tracking, brain imaging). Unless you have the luxury to crowd-source data using a simple survey, there will be practical limits to your sample size.

6) **Anticipated problems:** Finally, you want your sample size to compensate for expected problems, both at the design-level and the analysis-level. This includes for example drop-out of participants (e.g., in longitudinal studies), unequal samples (e.g., one group may be more difficult to sample than another), non-normal response data, or unequal variances. While many researchers account for drop-out, they routinely neglect analysis problems, even when they are widely known in the field (e.g., variance increases in older age groups). Recall that statistical tests with variance corrections usually shrink the degrees-of-freedom of the reference distribution (e.g., Welch *t*-test). This means you may end up with a (much) lower effective power than your nominally calculated sample size suggested!

When setting a sample size for your study, all of these criteria could be invoked to justify it. For example, for a large effect in a within-subjects design, G*Power may return a low sample size estimate, such as 15 or 20. While this would be statistically correct, it would be prudent to expand that number to 30 or 50, for reliability, distributional convergence, and generalizability. Possibly more are needed if drop-out is anticipated.

In my opinion, it sometimes makes more sense to set sample size at a **large common-sense number** (e.g., 100), than to make a painstaking effort to justify the number mathematically (if feasibility permits it).

Finally, some of the above criteria apply to setting the trial sample size as well as the participant sample size. In multiple-trial studies, having more trials permits more precise, distributionally normal, and generalizable within-subject estimates, while affording more power than equivalent single-trial (i.e., between-subjects) studies do.

Best,
Ben

--
Ben Meuleman, Ph.D.
Statistician
Swiss Center for Affective Sciences
University of Geneva | Campus Biotech
Chemin des Mines 9 | CH-1202 Genève
ben.meuleman@unige.ch | +41 (0)22 379 09 79