Pilot studies for power calculation

E-mail distributed on 21-04-2025

Dear all,

A question that regularly comes up is how to set the sample size for an adequately powered pilot study. In my opinion this may be a misguided question, in that finding your effect of interest should not, in fact, be the primary purpose of a pilot study. Foremost, I think a pilot should function as a feasibility study on the proposed methodology, and clarify practical questions such as whether a session has the right duration, if the instructions, materials and questionnaires are understood by participants, and whether the tasks function from a purely technical point of view (as a proof-of-concept). This assessment may not even involve data analysis but simply observation and informal discussions with the pilot participants.

Even if resources were abundant, running large pilot studies would be discouraged from the preceding perspective, since if some part of the experiment turns out to be flawed or unfeasible, all of the collected data may be unusable. Similarly, when resources are scarce (e.g., hard-to-sample populations, time limits, funding limits), it is best to waste as few valuable participants as possible.

For finding the target effect, a pilot will be underpowered almost by definition, so no panic is warranted if it should not be significant (but if it is, great). Moreover the descriptive effect size can be highly informative to guide the power analysis of the final experiment. Thus, while I doubt it makes sense to set a proper sample size for a pilot study, its results can itself inform the sample size for the final experiment. In doing so, one should not just use the effect size's point estimate but also its confidence interval, and determine the qualitative range in which the effect size is likely to be observed (e.g., small, medium, large). This range will allow to calculate optimal sample sizes under pessimistic and optimistic conditions.

1. Non-standard power parameters

Using the pilot study to estimate the effect size may not be necessary in many cases, when its value is more or less established theoretically and/or in previous studies. Even in this case, it is advised to sample effect sizes from several studies (or their confidence intervals) and likewise consider a *range* of possible effects when setting your own sample size. Unfortunately, this approach will typically not work for other types of statistical parameters that nonetheless influence power and sample size, such as:

- The non-sphericity parameter epsilon in repeated measures ANOVA

- Random effects parameters in multilevel regression
- Bayesian priors and their hyperparameters.

For example, in G*Power (Faul et al., 2009), the window for repeated measures ANOVA contains extra fields for the correlation among repeated measures and the non-sphericity correction epsilon, which measures the departure from spherical repeated measures correlation (Fig. 1). Both of these can dramatically impact the required sample size, with higher correlation requiring fewer participants, and lower epsilon (i.e., stronger departure from sphericity) requiring more participants. The default values suggested by G*Power are respectively too low (repeated measures tend to be moderately to strongly) and too high, meaning realistic estimates would be preferred.[1] However, neither of these values is commonly reported in papers. While one could set epsilon to its theoretical lower bound (i.e., 1/(k-1), where *k* is the number of levels in the within-subject factor), this is very conservative and doing so in the example of Fig. 1 would raise the required sample size from 29 to 93!

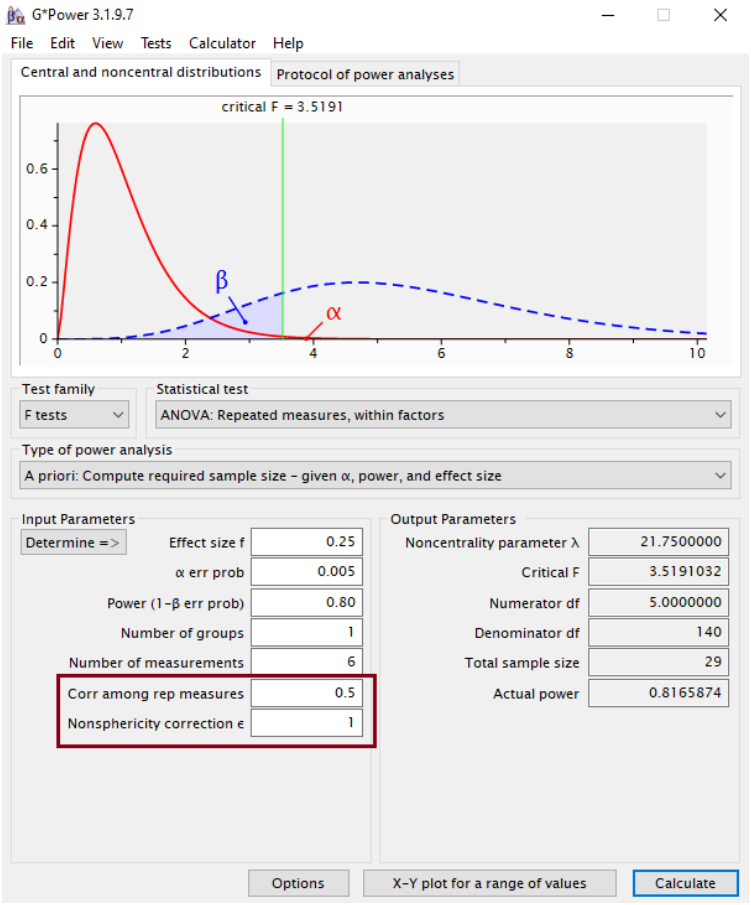

*Figure 1. Sample size for a repeated measures ANOVA in G*Power, with one 6-level within-subjects factor, no between-subjects factors, a medium effect size, a significance level of 0.005, and a power level of 0.80.*

[1] Only applies to within-subject designs with factors that have more than 2 levels

In this example, a pilot study would be perfect for providing useful reference values, which can then be plugged into the sample size calculation. As with the effect size, one can again consider a *range* of possible values for these parameters to calculate the sample size under pessimistic or optimistic conditions. For the repeated measures correlation, this could be based directly on its confidence interval, while for epsilon, one could choose between the Greenhouse-Geisser estimate or the less conservative Huyn-Feldt estimate.[2]

2. Assumption violations

A second important area where pilot studies can inform future analyses is in assumption violations. In a previous post, I have recommended that sample size calculations should take into account "anticipated problems", such as missing data and assumption violations. For example, having non-normal residuals or heteroscedastic residuals in traditional regression may necessitate either the use of non-parametric tests (e.g., Spearman rank correlation) or corrections to the parametric degrees of freedom (e.g., Welch *t*-test). In both cases, power will be lost. Because these violations are typically difficult to foresee, sample size calculations mostly ignore them. Here again a pilot study can give an initial sense of whether non-normal and/or heteroscedastic residuals should be expected, and whether an alternative analysis model should be planned for.

One caution attached to this application is that diagnosing assumption violations should **not be done with inferential tests** (e.g., Shapiro-Wilk test, *F*-ratio test). Such tests are problematic in a variety of scenarios and typically have low power in small samples (see Part 1 of my workshop on non-parametric data analysis). As noted, the pilot sample size is expected to be underpowered anyway, therefore diagnostic tests are expected to have almost no practical utility. Instead, diagnostics should be done primarily visually, which is generally recommended for all data analyses.

References

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

--

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

[2] A theoretical confidence interval for epsilon does not exist, although an empirical CI could be constructed with bootstrapping.