A Researcher's Guide to Power Analysis

Though conducting a power analysis is an essential part of any research plan, the process is often a confusing one for researchers. The purpose of this document is to provide a brief tutorial to assist in understanding, performing, and reporting a power analysis. Resource links to supporting literature and power software are also provided.

Introduction

Understanding the definitions of effect size, p-values, and power, as well as their relationship to one another, is a crucial step in understanding the concept of power analysis.

Effect sizes, P-values, & Power

In research, there are two measures used to report study results - effect sizes & p-values.

What is an effect? An effect is just the study result, i.e. what was found. If you were looking for the difference in scores between two groups, the effect is the difference found.

An effect size (ES) measures the strength of the result and is solely magnitude based – it does not depend on sample size. So the effect size is pure – it is what actually was found in the study for the sample studied, regardless of the number of subjects. But is what was found generalizable to a population?

This is where p-values come into play. A p-value gives you the likelihood that what you found is not due to chance. P-values very much depend on sample size.

Let's look at a very simple example. Suppose we want to look at the difference in height between basketball players and the general population. We have two basketball players handy whom we measure and who are found to be 7 feet tall and we randomly select two people from the population, who happen to be 5 feet tall. So we would see a huge difference, i.e. a very large effect in this poorly designed study. But because our sample in so small, we would suspect the result found may not be representative of what we would expect in the population, i.e. a chance finding.

Looking at a more realistic example, suppose we wish to know if there is a difference in mean scores on some measure between the intervention & control group in a pilot study with a sample size of 8?

$N_{int} = 4$, $Mean_{int} = 90$, $SD_{int} = 5$
$N_{ctl} = 4$, $Mean_{ctl} = 85$, $SD_{ctl} = 5$
$ES = (Mean_{int} – Mean_{ctl})/\text{pooled SD} = (90-85)/5 = 1.0$ (which is considered to be a large effect)

If we now proceed and use an inferential test such as a Mann-Whitney U or t-test to test for significant difference in mean score between groups, we find we have a non-significant p-value, even though there is a large effect. This is likely because we have a small sample. A power analysis would have shown us that at least 14 subjects are needed in each group to prove this effect with inferential statistics (i.e. using p-values). Small studies (< 100) may have medium or large effects but not yield statistically significant p-values. Large studies (> 2000) may have small and often inconsequential effects but be statistically significant. And mid-size studies (> 100 and < 2000) usually have agreement in that medium to large effects generally also yield a p-value < .05.

It is therefore important in ALL studies to report both effect sizes and p-values and to do a power analysis. Sufficient power to find statistical significance (i.e. via p-value) for a given effect size minimizes chance findings & is critical to funding research, conducting statistical analysis, and publishing results. The one exception is pilot studies, which often rely on effect sizes.

The Language of Power Analysis

What is a power analysis? A power analysis is just a process by where one of several statistical parameters can be calculated given others. Usually, a power analysis calculates needed sample size given some expected effect size, alpha, and power.

There are four parameters involved in a power analysis. The research must 'know' 3 and solve for the 4th. They are as follows:

1. Alpha:
    - Probability of finding significance where there is none
    - False positive
    - Probability of a Type I error
    - Usually set to .05
2. Power
    - Probability of finding true significance
    - True positive
    - 1 – beta, where beta is :
        - Probability of not finding significance when it is there
        - False negative
        - Probability of a Type II error
    - Usually set to .80
3. N:
    - The sample size (usually the parameter you are solving for)
    - May be known and fixed due to study constraints
4. Effect size:
    - Usually the 'expected effect' is ascertained from:
        - Pilot study results
        - Published findings from a similar study or studies
            - May need to be calculated from results if not reported
            - May need to be translated as design specific using rules of thumb
        - Field defined 'meaningful effect'
        - Educated guess (based on informal observations and knowledge of the field)

Since alpha is usually set to .05 and power to .80, the researcher primarily needs to be concerned with the sample size and the effect size. In most cases, the researchers is interested in solving for the sample size, so the majority of the work needed to do a power analysis relates to determining the expected effect to be used in the power analysis.

The most commonly asked question at this point is 'How can I know the expected effect when I haven't done the study yet?' In most cases, this expected effect is not known unless pilot data is available. Without pilot data, a literature search for a similar study or studies is the next best option. As published literature is not likely to be identical in study design and purpose as your study, this search may not be an easy one. The goal is to have a realistic expectation of what type of effect you are likely to find (small, medium, or large) and to be certain it is a valued effect in your field of study. If, for example, prior studies have shown no relevant effect, does it make sense to do your study as it will likely yield similar inconsequential results? If, though, another somewhat similar study found a moderate effect in some demographic subgroup, then a study that looks at other groups or the population as a whole might be worthwhile and this prior effect found could be used as a best estimate of expected effect in this larger study.

If there is no pilot data and there are no similar studies (i.e. you are venturing into new research), it is often wise to start with a pilot study. If this is not possible, then your expected effect must be based on either your knowledge of the field or what you view as meaningful (small, medium, or large). Keep in mind that with the exception of pilot studies, funders do not look too kindly towards researcher expectations that are not backed by pilot studies or published literature, so this approach is a last resort.

Types of Power Analysis

There are four types of power analysis, defined by which one of these four parameters you wish to solve for:

- A priori: compute N, given alpha, power, ES
- Post-hoc: compute power, given alpha, N, ES
- Criterion: compute alpha, given power, ES, N
- Sensitivity: compute ES, given alpha, power, N

The a priori power analysis is what is usually done when designing a study. This tells you what sample size is needed to detect some level of effect with inferential statistics (i.e. with p-values). Funding agencies of course want to avoid chance findings, so an a priori power analysis is needed in all study proposals, except pilot studies.

A post-hoc power analysis at the completion of a study is also wise, as your expected effect and actual effect may not align. This post-hoc power analysis tells you if you had sufficient subjects to detect with inferential statistics the actual effect you found.

A criterion power analysis is seldom used by researchers.

A sensitivity power analysis is used when the sample size is predetermined by study constraints. For example, if there are only 20 subjects available, determining how many you need is less relevant. Instead, one determines what level of effect you could find with the subjects you have. This is referred to as the minimal detectable effect (MDE).

It is important to note that study design impacts power calculations and the interpretation of effect sizes. We saw in a previous example that when looking at the difference in mean scores between an intervention and control group, we had an effect size of 1.0 which was considered to be quite large (see the first line in the table below). If instead we had been looking at the difference between three means, we might have an ANOVA design and an effect of just .4 using f, not to be confused with F) would be large. As a researcher may be reviewing published literature with varying designs when trying to determine the expected effect for their proposed study, it is essential to understand that what is small, medium, and large varies with the design.

	Effect Size Benchmarks		
Statistic	Small	Medium	Large
Means - Cohen's d	0.2	0.5	0.8
ANOVA - f	0.1	0.25	0.4
ANOVA - eta squared	0.01	0.06	0.14
Regression f-test	0.02	0.15	0.35
Correlation - r or point serial	0.1	0.3	0.5
Correlation - r squared	0.01	0.06	0.14
Association - 2 x 2 table -OR	1.5	3.5	9
Association - Chi-square - w or Phi	0.1	0.3	0.5

Conducting a Power Analysis

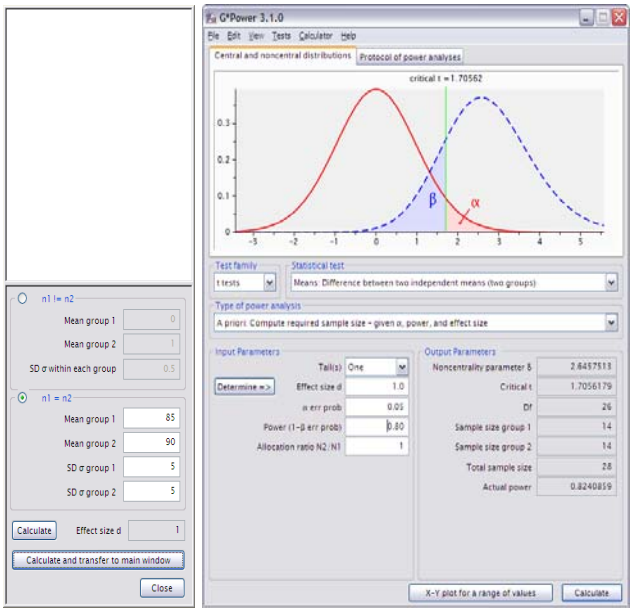
There many software products available for performing a power analysis, some of which are quite easy to use. The most popular by far is GPower and with good reason. This user-friendly product is a free download available for both the PC and Mac. It supports many designs (t-test, ANOVA, ANCOVA, repeated measures, correlations, regression, logistic, proportions, Chi-square, nonparametric equivalents). It includes an effect size calculator and an online tutorial manual is also available to train users in its use.

For multi-level designs, the Optimal Design product is available for the PC, also at no cost. A free very lengthy tutorial manual is available, but like GPower the product works with pull-down selections, much like most Windows applications. The product expects a two-group randomized control trial design with up to three levels.

Other software options that come at a price include SPSS Sample Power (an SPSS add-on), SAS Proc Power, Pint, and PASS.

The steps involved in conducting a power analysis are as follows:
1. Select the type of power analysis desired (a priori, post-hoc, criterion, sensitivity)
2. Select the expected study design that reflects your hypotheses of interest (e.g. t-test, ANOVA, etc.)
3. Select a power analysis tool that supports your design
4. Provide 3 of the 4 parameters (usually alpha=.05, power = .80, expected effect size, preferably supported by pilot data or the literature)
5. Solve for the remaining parameter, usually sample size (N)

e.g. Using the prior pilot data with an ES=1 presented earlier, determine the sample size needed to detect this level of expected effect using inferential statistics (i.e. p-values)

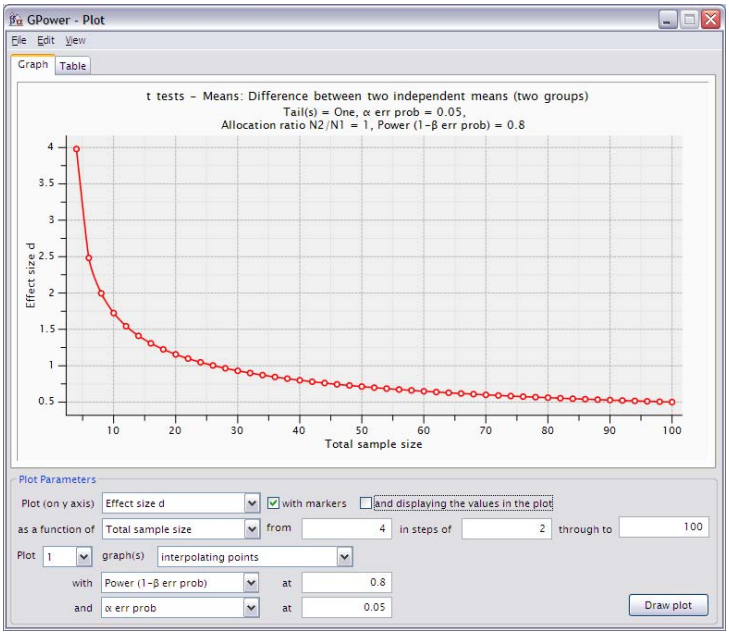


In GPower we would first indicate that we wished to perform an a priori power analysis, then choose a one-tail t-test between two independent samples with equal group sizes. Next we enter the expected effect size from pilot data of 1.0 (or press 'Determine' next to effect and enter the means and SDs, letting GPower calculate the effect size from the pilot data for us). Then enter an alpha of .05 and a power of .80 and press 'Calculate'. GPower displays on the same screen the needed N of 28 (14 per group) to detect this level of effect (ES=1.0) with p-values using a t-test.

Running power analysis

Suppose instead we wish to check the effect size as the study progresses to see if the expected effect we gleamed from the literature is realistic in our study, with its unique design and purpose. If our effect size is higher or lower thus far, we may wish to adjust recruitment accordingly yet retain sufficient power to show the effect inferentially. To do this, use a running power analysis, also available through GPower.

If our pilot data, for example, showed a very large effect size of 1.0 as in the prior example, but we found our two group means comparison study was yielding a lesser yet still large effect size of .8, we would need more than 28 subjects to find significance with inferential statistics. Our running power analysis below suggests we should continue recruitment to 40 subjects, if our study design allows a rolling recruitment.



A word of caution! A running power analysis is ethical if it allows a worthwhile effect to be proved inferentially, where it can be shared in the literature as an important and statistically significant finding. Using a running power analysis to increase sample size to prove inferentially a meaningless effect is highly questionable and can waste valuable time and financial resources.

The purpose of the running power analysis is to support a relevant effect with inferential statistics. What is a worthwhile effect is very specific to the field of study and requires consideration of risk, cost, and benefit … sometimes a small effect might be highly relevant, while more often a medium or large effect is needed.

Reporting a Power Analysis

Usually a researcher performs a power analysis for the main hypothesis of interest. In some more complex study plans, the researcher may perform a separate power analysis for each experiment or hypothesis, and select the larger sample size needed from among them as a basis of recruitment. Other considerations in performing and reporting power analysis for sample size estimation include attrition and control for possible mediators or moderators. Expected mediators/moderators can be considered when selecting the study design in the power analysis software. And sample size used for recruitment can be adjusted to allow for expected attrition.

Though every study differs, writing a power analysis summary in a grant proposal need not be complex. A simple paragraph or two that reflects your study plans and addresses possible contingencies for additional variables, attrition, etc. usually is satisfactory. Most funding agencies are interested in study feasibility with a well thought out study plan, which a power analysis is a component of.

Below is a sample power analysis paragraph for a simple design, which can be modified easily to reflect the study specifics:

Sample size estimation

A statistical power analysis was performed for sample size estimation, based on data from *pilot study/published study X* (N=…), comparing …. to ….. The effect size (ES) in this study was …., considered to be *extremely large/large/medium/small* using Cohen's (1988) criteria. With an alpha = .05 and power = 0.80, the projected sample size needed with this effect size (GPower 3.1 or other software) is approximately N = ….. for this simplest *between/within* group comparison. Thus, our proposed sample size of ..N+.. will be more than adequate for the main objective of this study and should also allow for expected attrition and our additional objectives of *controlling for possible mediating/moderating factors/subgroup analysis, etc.*

Summary

Be aware that though power analysis is extremely important in study design and post study analysis, it is indeed fuzzy science.

When using power analysis to calculate N, the expected ES may not align with the actual effect found as each study is unique in protocol, population studied, covariates & factors considered, etc. (i.e. the expected effect size is an educated guess).

When using power analysis to calculate the minimal detectable effect (MDE), the expected sample size may not align with the final N due to missing data or differing attrition rates (i.e. the expected N is an educated guess).

The study design used in the power analysis to calculate N (or MDE) may not align with that used in the actual study as the data may not meet the assumptions of the proposed method (i.e. the expected study design is an educated guess).

Given all these educated guesses, an a priori power analysis may not be accurate! Its purpose is solely to show the feasibility of the proposed study.

Resources

UCLA Power Analysis Seminar:
http://www.ats.ucla.edu/stat/seminars/intro_power/default.htm

GPower free download & tutorial manual (Mac or PC):
http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

Optimal Design for multilevel RCT (for PC):
http://sitemaker.umich.edu/group-based/optimal_design_software

Seminal reference for power analysis:
Cohen, J. (1969) Statistical Power Analysis for the Behavioral Sciences. NY: Academic Press