Research & Innovation

University of the West of England BRISTOL

Sample Size and Power in Clinical Trials

Version 1.0 May 2011

1. Power of a Test
2. Factors affecting Power
3. Required Sample Size

RELATED ISSUES

1. Effect Size
2. Test Statistics
3. Variation
4. Significance Level
5. One Sided and Two Sided Tests

North Bristol NHS Trust

Statistical Power

The power of a statistical test is the probability that the test will reject the null hypothesis when the null hypothesis is false i.e. it will not make a Type II error, or a false negative decision. As the power increases, the chances of making a Type II error decrease. The probability of a Type II error is referred to as the false negative (β). Therefore power is equal to 1-β, which is also known as the sensitivity.

Uses of Power Analysis

A power analysis can be used to calculate the minimum sample size required so that one can be reasonably likely to detect an effect of a given size. It can also be used to calculate the minimum effect size that is likely to be detected in a study using a given sample size. Additionally, the concept of power can be used in making comparisons between different statistical testing procedures: for example, between a parametric and a nonparametric test of the same hypothesis.

The two sample problem

Statistical tests use data from samples to assess, or make inferences about a population. In the concrete setting of a two-sample comparison, the goal is to assess whether the mean values of some attribute obtained for individuals in two sub-populations differ. For example, to test the null hypothesis that the mean heart rate of smokers and non-smokers do not differ, samples of smokers and non-smokers are drawn and their heart rate is recorded. The mean heart rate of one group is compared to that of the other group using a statistical test such as the two-sample t-test. The power of the test is the probability that the test will find a statistically significant difference between smokers and non-smokers, as a function of the size of the true difference between those two populations. Conceptually, power is the probability of finding a difference that does exist, as opposed to the likelihood of declaring a difference that does not exist (which is known as a Type I error or “false positive”).

Factors Influencing Power

There are many factors that affect the power of a study. In part power depends upon the test statistic chosen at analysis (that why we always aim to chose the "best" statistic for any given situation). The power of the test is also dependent on the amount of variation present; the greater the variability the lower the power. This can be partly offset by consideration of the best design and a careful consideration about the design of the study can lead to increased power. In general, large differences are easier to find than small differences. The relative size of the difference is known as the effect size (difference) and in general we have increasing power with increasing effect size. It also stands to reason that we are "more likely" to make a correct inference with a big "representative" sample. Accordingly sample size affects power. Finally the power of a test is dependent of the nominal significance level chosen as this is quite key in determining the decision rule. The power of a study can be greatly affected by the study design (which can alter the effect size). We will consider each of these aspects qualitatively.

1) An appropriate Test Statistic

There are many ways of analysing a given set of data. In any one situation we may be able to argue that one particular method or one particular technique or one particular approach is better than another. For instance, suppose we are going to compare two independent samples for differences in location. Possible methods of analysis might include the two-sample independent *t*-test or the Mann Whitney test or you might even invent your own test. If the underpinning assumptions for the two-sample independent *t*-test are satisfied (or are not grossly violated) then we would argue that the *t*-test is the best one to apply. Note that by "best" we really mean the "most powerful" test i.e. the one that is most likely to allow us to claim a statistically significant difference when in fact the null hypothesis is false.

Theory shows that **when** the assumptions underpinning the t-test are true **then** the *t*-test is the most powerful test to apply. Likewise when the assumptions underpinning the two-samples *t*-test are slightly violated then simulation studies show that the *t*-test still retains relatively high power compared with other general techniques (such as the Mann Whitney test). However when the assumptions underpinning the two samples *t*-test are grossly violated then the Mann Whitney test can have greater power.

Power depends upon the test statistic chosen. An appropriate test statistic permits a defendable way of analysis data in any particular situation and it impacts on the probability of obtaining a statistically significant result. For this reason we devote considerable effort in deciding on the best test statistic to use.

Power also depends on the approach. A modern branch of statistics is the computer intensive approach known as the bootstrap; for non-normal populations the bootstrapping approach can have greater power than corresponding parametric or non-parametric techniques. **Power depends upon the data analysis approach.**

1) Variation

Statistical inference is concerned with identifying a "message" in the data. The message (e.g. a difference in means) can be obscured by variation in the data. If the variation in data can be reduced then it is easier to obtain a clear message (i.e. we would ideally want a high "signal-to-noise" ratio). Studies should be designed to minimize error variance relative to the systematic variance and/or to partition out (remove or account for) some aspects of variation in the analysis phase. For instance in a "before and after" study we would have data on experimental units (e.g. participants) at the outset and then later on. At the outset the experimental units would, most likely, differ. These initial differences might be large and would add to the variation in the data and this "between experimental unit" variation could cloud the message in the data. To reduce the effect of the variation we would look at the *changes* between the before and after pairs. Doing this means that we are focussing directly on the phenomenon of interest (the change between before and after) and are removing some of the variation due to initial differences in the experimental units. [2]

In general if we are designing a study and if we have the choice of generating "paired" data or "unpaired" data then we would probably want to opt for the paired design as it gives [1] a direct focus on the phenomenon of interest and [2] at the analysis phase we can remove or account for some variation. These two benefits means that a Blocked or repeated measures design tends to be more powerful than the corresponding independent or between subjects design.

Alternatively suppose that a researcher is looking to see if differences in performance exist between those that ingest caffeine and those that do not. If such effects do exist then it more likely that the effects would be apparent in a study comparing an intake of 500 mg of caffeine against 0 mg of caffeine than a study comparing 25 mg of caffeine against 0 mg of caffeine.

In general, if the variation is relatively large then the power of the test would be relatively low (all other things remaining equal).

Careful thought about the design of the study can lead to a more powerful test.

2) Effect Size

If population mean values are widely separated relative to the variation then we would anticipate this being reflected in the data. The greater the true separation the easier it will be to detect the difference using a sample. Consequently we have increasing power with increasing differences between means (all other things being equal). For comparing two means the effect size (ES) is defined as "the difference between means relative to the standard deviation" (note standard deviation and not standard error of the means). See Section 3 for comparing two means.

Similarly consider a correlation study. If two variables are strongly correlated then we anticipate this strong correlation being reflected in a sample. The greater the true correlation the easier it will be to detect the correlation in the sample. Consequently we have increasing power with increasing strength of population correlation. In a correlation analysis the correlation coefficient is the measure of effect size.

More generally we have increasing power with increasing size of effect.

3) Sample Size

Generally speaking the precision of a statistic increases with increasing sample size. By way of example suppose we consider the sample mean $\bar{x}$ which is used to estimate a population mean μ. For different samples each of size *n* the sample means (say $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$, $\bar{x}_4$, ...) will not all have the same value. If we are dealing with random samples then the variance of the sample estimates is given by σ^2 / n where σ^2 is the variance of the population. In any practical setting the population variance σ^2 is a fixed positive number. The ratio σ^2 / n decreases as *n* increases (i.e. $n \to \infty$, $\sigma^2 / n \to 0$). Because the "error variance" of the statistic diminishes with increasing sample size we are more likely to make a correct inference. **"More likely" to make a correct inference indicates that the power of the test increases with increasing sample size (all other things being equal).**

Increasing sample size is typically associated with an increase in power.

4) Significance Level

Contemporary practice is to perform hypothesis tests using a nominal significance level of $\alpha = 0.05$. The significance level indicates the probability of committing a Type I error (i.e. the probability of incorrectly rejecting the null hypothesis when in fact it is true). In some situations the consequences of making a Type I error may be judged as not being "too serious" and in these cases we may choose to work with a bigger nominal significance level (e.g. $\alpha = 0.10$). The bigger the significance level the easier it is to reject the null hypothesis. If the null hypothesis is false and the criterion to reject the null hypothesis has been relaxed then the test will be more powerful than it would have been if the criterion had not been relaxed.

Sometimes the consequences of making a Type I error may be viewed as being "very serious". In these cases we may require compelling evidence before rejecting the null hypothesis and opt to use a more stringent (lower) significance level (e.g. $\alpha = 0.001$). Making it harder to reject the null hypothesis means that the power of the test will be reduced (all other things being equal).

In general we have decreasing power with decreasing nominal significance level.

5) One-Sided or Two-Sided Test

In some very rare instances theory might exist so that some possibilities concerning the direction of an effect can be eliminated. In these cases a researcher might carry out a one-sided test. In one-sided tests there is a focus in one direction only (e.g. a positive difference) and all other things being equal then one-sided tests will be more powerful than two sided tests.

Standards for Power

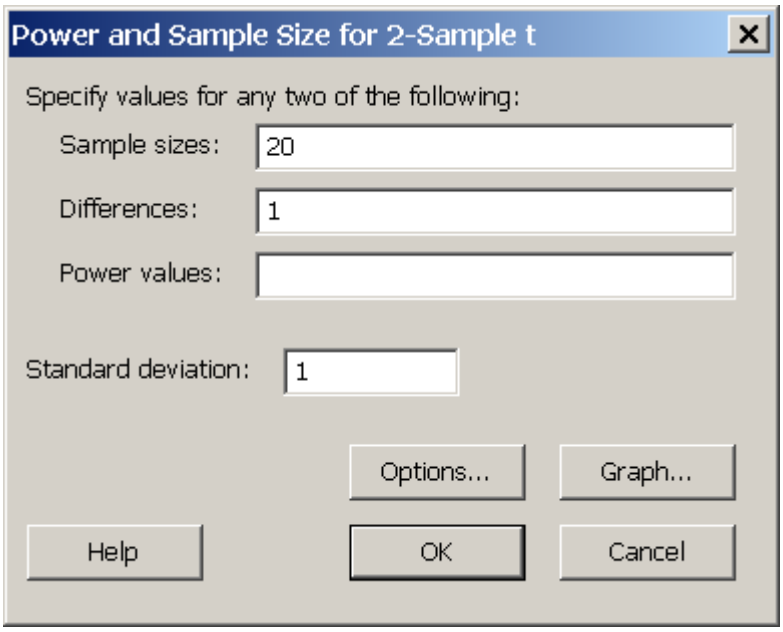
Although there are no formal standards for power, most researchers assess the power of their tests using 0.80 (80%) as a standard for adequacy. This convention implies a four-to-one trade off between β-risk and α-risk. (β is the probability of a Type II error, α is the probability of a Type I error, 0.2 = 1 - 0.8 and 0.05 are conventional values for β and α). However, there will be times when this 4-to-1 weighting is inappropriate. In medicine, for example, tests are often designed in such a way that no false negatives (Type II errors) will be produced. But this inevitably raises the risk of obtaining a false positive (a Type I error). The rationale is that it is better to tell a healthy person "we may have found something – let's test further," than to tell a diseased person "all is well".

Power analysis is appropriate when the concern is with the correct rejection, or not, of a null hypothesis. In many contexts, the issue is about determining if there is or is not a difference but rather with getting the more refined estimate of the population effect size. For example, if we were expecting a population correlation between height and weight of around 0.50, a sample size of 20 will give us approximately 80% power (α = 0.05, two-sided) to reject the null hypothesis of zero correlation. However, in doing this study we are probably more interested in knowing whether the correlation is 0.30 or 0.60 or 0.50. In this context we would need a much larger sample size in order to reduce the confidence interval of our estimate to a range that is acceptable for our purposes. Techniques similar to those employed in a traditional power analysis can be used to determine the sample size required for the width of a confidence interval to be less than a given value.

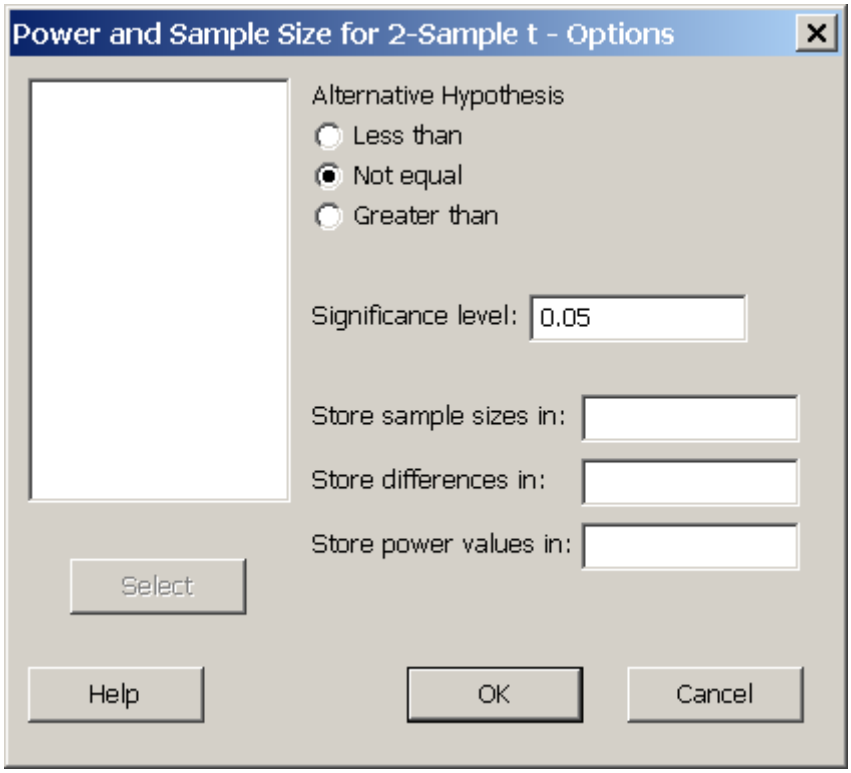
An Example Power Analysis

A clinician wants to select the better of two treatments for Tinea Capitis (a fungal infection of the scalp or ringworm of the scalp) and the primary outcome of interest is “time to complete disappearance of the fungus on the scalp” measured in days. We will assume that the two sample t-test will be used; that a mean difference of size 1 is anticipated and that the population standard deviation is estimated to be 1. We will assume that a researcher can take a sample of size 20 per group and then wonder what the power of the test would be.

The example information has been put into Minitab as shown



Under Options we have selected to perform our analysis at the standard 0.05 significance level.



Output from running the power routine using the above parameters is as follows:

```
2-Sample t Test

Testing mean 1 = mean 2 (versus not =)
Calculating power for mean 1 = mean 2 + difference
Alpha = 0.05  Assumed standard deviation = 1

            Sample
Difference    Size     Power
         1      20  0.868953
The sample size is for each group
```

From the above output we conclude that if the null hypothesis is false and the true mean difference is +1 and if the population standard deviation is also equal to 1 then if we obtain a sample of size *n* = 20 per group we have an 87% chance of correctly rejecting the null hypothesis (again assuming that the underpinning assumptions of the two-sample *t*-test would not be grossly violated).

Applying for Grants

Funding agencies, ethics boards and research review panels frequently request that a researcher perform a power analysis, for example to determine the minimum number of test subjects needed for an experiment to be informative. This is because and underpowered study is unlikely to allow one to choose between hypotheses at the desired significance level. Power remains the most convenient measure of how much a given experiment size can be expected to refine prior beliefs. A study with low power is unlikely to lead to a large change in prior beliefs.