

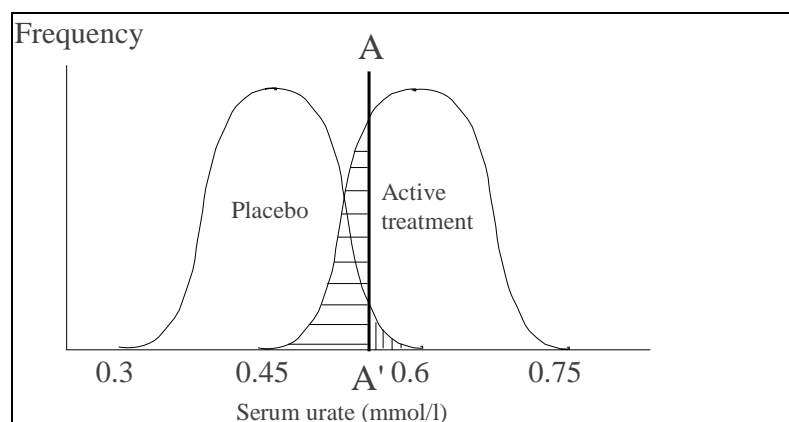
Power and sample size calculations

Sample size calculations are an essential component of planning a research trial and of the application for ethical approval. The Ethics Committee's view of power calculations stems from the principle that to be ethical a trial must have scientific validity. This view extends to planning a study in such a way as to allow unambiguous statistical interpretation, which in turn demands estimation of sample size. If a study is too small, resources may be wasted on an inconclusive study, and if it is too large, potential treatment benefits could be denied to patients randomized to placebo or the previous standard treatment. For this reason, the Committee is changing its emphasis on power calculations and now requires full disclosure.

1. The need for power calculations: Type I (α) and type II (β) errors

Consider a parallel group study in which a drug's effect on plasma urate is to be determined. Values of urate are measured in two groups given active treatment or placebo (denoted by t and p respectively) containing n_t and n_p subjects (n_t and n_p are not necessarily equal). Fig 1 (derived from *Colton T, Statistics in Medicine. Little, Brown & Co, Boston, 1974, p 121*) shows the sampling distribution curves for each treatment. In the hypothesis being studied the drug is assumed to increase plasma urate, and the possibility that urate is decreased is excluded. This allows application of what is called a "one-tail" test since the sampling distribution with the test drug is being tested against only the upper tail of the placebo distribution of urate. If the hypothesis was that the drug changes urate, without specifying the direction of change, then a "two-tail" test would be required. This alters the sample size requirements. Trials normally require to be powered under a 2-tailed test.

The measured mean urate on treatment is 0.6 mmol/L compared to untreated (placebo) urate of 0.45 mmol/L (see figure). First, it is assumed that the drug makes no difference (the Null Hypothesis, designated as H_0). Then, the probability that the experimental data is at least as inconsistent as H_0 if H_0 is true is derived. The second negative in the previous sentence is required because the original hypothesis is itself expressed in the negative; it is not strictly correct to state that the statistical test measures the probability that H_0 is true. The usual convention is that H_0 will be rejected if the probability of the data is at least this extreme falls below 1 in 20 (0.05), but more stringent tests (say 0.01, equivalent to a 1 in 100 probability) are sometimes used. The prior probability is called α .



This is equivalent to saying the H_0 can be rejected if the treatment sample mean lies outside the value of the plasma urate (0.55 mmol/L; line AA') which divides the area of the placebo (left-hand) distribution into a "tail" of 5% and a "body" of 95%. If this occurs there is justification to reject H_0 and therefore to

conclude that the drug has a significant effect. However, there is by definition a 1 in 20 chance that the rejection of H_0 will be incorrect because the observed treatment effect is actually due to chance.

A type I or α error occurs when H_0 is rejected when it is true (false positive). Alpha is the probability of a type I error. The probability of not making this error, given by $(1 - \alpha)$, increases as α decreases. For example, if the required value of p is < 0.01 then $\alpha = 0.01$ and $(1 - \alpha) = 0.99$; there is now a 1% probability of a type I error and a 99% probability of a correct result. There is an inverse relationship between α and sample size when β error (see below) is constant; a smaller α means a larger sample requirement.

Now consider the possibility that H_0 is indeed false, i.e. treatment has a real effect. A reciprocal source of error, known as a type II error, now exists.

A type II or β error occurs when H_0 is not rejected when it is false (false negative). The risk of a type II error is denoted by β , and its complement $(1 - \beta)$ is the probability of correctly rejecting H_0 when the treatment groups actually differ. $(1 - \beta)$ is known as the POWER of the study.

Consider the figure and suppose that H_0 is false (i.e. there is a real effect of treatment which is to increase the mean plasma urate to 0.6 mmol/L). Since a sample mean of less than 0.55 mmol/L fails to supply evidence to reject H_0 , the risk of not rejecting H_0 is represented by the area under the right hand distribution which is to the left of this value (the horizontal shaded area). If this area is 12.5 percent of the distribution centred on a mean of 0.6 mmol/L, $\beta = 0.125$.

The diagram demonstrates an important principle, namely that if α is decreased with the aim of decreasing the probability of a type I error (equivalent to moving the line AA' further towards the right-hand tail of the placebo distribution), the probability of a type II error (β) is necessarily increased. *Thus there is always a trade-off between the probability of a type I and type II error.* The investigator must consider what level of protection should be provided by the study to avoid two undesirable but mutually exclusive possible outcomes: on the one hand, that an ineffective intervention might be found by chance to be effective and wrongly admitted into clinical practice, and on the other that an efficacious one might be discarded because of a chance observation that it is ineffective. Both these outcomes have obvious ethical implications, which explains the interest of the Committee in sample size calculations.

2. Sample size estimations

For a 2-treatment crossover study (suitable for paired t test), the equation for n, the number of subjects for a two-tailed test is

$$n = \left[\frac{(z_\alpha - z_\beta) \cdot \sigma_d}{(\mu_t - \mu_p)} \right]^2 \quad (1)$$

where z_α and z_β are critical values (obtained from tables) corresponding to the specified α and β levels, σ_d is the SD of the difference between paired samples, and μ_t and μ_p are the mean values of the primary response variable obtained on each treatment. For $\alpha = \beta = 0.05$, $(z_\alpha - z_\beta) = 3.61$ and the equation simplifies to

$$n = 13 \times \frac{\sigma_d^2}{(\mu_t - \mu_p)^2} \quad (2)$$

For example, if you wish to measure the effect of an antihypertensive drug which decreases BP by 5 mmHg when the SD of blood pressure difference between paired values is 5 mmHg and using the above power values, about 13 subjects will be required.

	Beta				
Alpha	0.01	0.05	0.1	0.2	0.3
0.01	24	18	14.9	11.7	9.7
0.05	18	13	10.5	7.8	6.2

The table shows the value of the leading coefficient (13 in equation (2)) for various combinations of α and β for a two-tailed paired t test. As the table implies, it is generally more acceptable to contain sample

size by compromising on β rather than α . This can be rationalised by accepting that the trial is unlikely to be conceived without some indication that the drug has the postulated effect. This represents the rationale for the study. Hence one may feel intuitively that it is less important to protect against the possibility of a type II error because H_0 is more likely to be false than true. Nevertheless, the compromise inherent in this logic must be recognised.

The corresponding equation for an unpaired t test is

$$n = 2 \times \left[\frac{(z_\alpha - z_\beta) \cdot \sigma}{(\mu_t - \mu_p)} \right]^2 \quad (3)$$

where μ and z have the same meaning as before, but σ is now the pooled population SD obtained by application of a standard formula (refer to statistical texts). Equation (3) can be rewritten as

$$n = 26 \times \frac{\sigma^2}{(u_t - \mu_p)^2} \quad (4)$$

when $\alpha = \beta = 0.05$. The coefficient for other α/β combinations in a parallel study is obtained by doubling the value given in the table and substituting in equation 4. Note that σ in equations 2 and 4 has different meanings. If the above study for an antihypertensive drug is done as a parallel group study (the more common study design) and the expected fall in BP with treatment is 10 mmHg and the SD of blood pressure is 15 mmHg, about 58 subjects per group will be required under the above power assumptions. A crossover study design is usually more sparing of resources and patients, but each patient is studied twice.

The above equations illustrate an important point, namely that different statistical tests have different power when applied to the same data. Sample size calculations are made in respect of a specified statistical test (and hence of a specified trial design), and are not valid for other tests. For such tests, eg tests of proportions or of regression, statistical advice should be sought.

Researchers often state that power calculations are not possible because they do not know the results of the study in advance, i.e they do not know the value of μ or σ in the above equations. This is very rarely valid because the primary aim should be written in such a way as to specify $(\mu_t - \mu_p)$ in advance, and often some guide to σ is known from other studies or may even be estimated from personal experience. The investigator not only has to hypothesize that the treatment has an effect, he should also state the expected size of the effect. If this is not known, then the expected minimum clinically

meaningful effect can be used. The Committee will critically assess whether the statement that power calculations are not possible is in fact true, and may decline an application if the claim is unsupported. If it is true that no *a priori* data are available, the Committee will regard the project as a pilot study. The pilot study should be sufficiently large to allow reasonable estimates of μ and σ , and therefore to allow planning for the definitive study.

Finally, note the need to include some correction in studies which compare more than two groups. This arises because a type I error can arise by chance for every 20 comparisons made within a study when $\alpha = 0.05$. One solution is to apply the Bonferroni method and power the study for $\alpha_{\text{corrected}} = \alpha / n$, where n is the number of tests. Thus α_{overall} is 0.05.

Royal Perth Hospital Power and sample size calculations document is acknowledged as the basis of this document.