PROFESSIONAL DEVELOPMENT

Sample Size

By
Egyptian Group for Surgical Science and Research
Said Rateb, EGSSR Moderator
Nabil Dowidar, EGSSR Secretary General
Mohamed Farid
Ahmed Hussein
Ahmed Hazem

INTRODUCTION

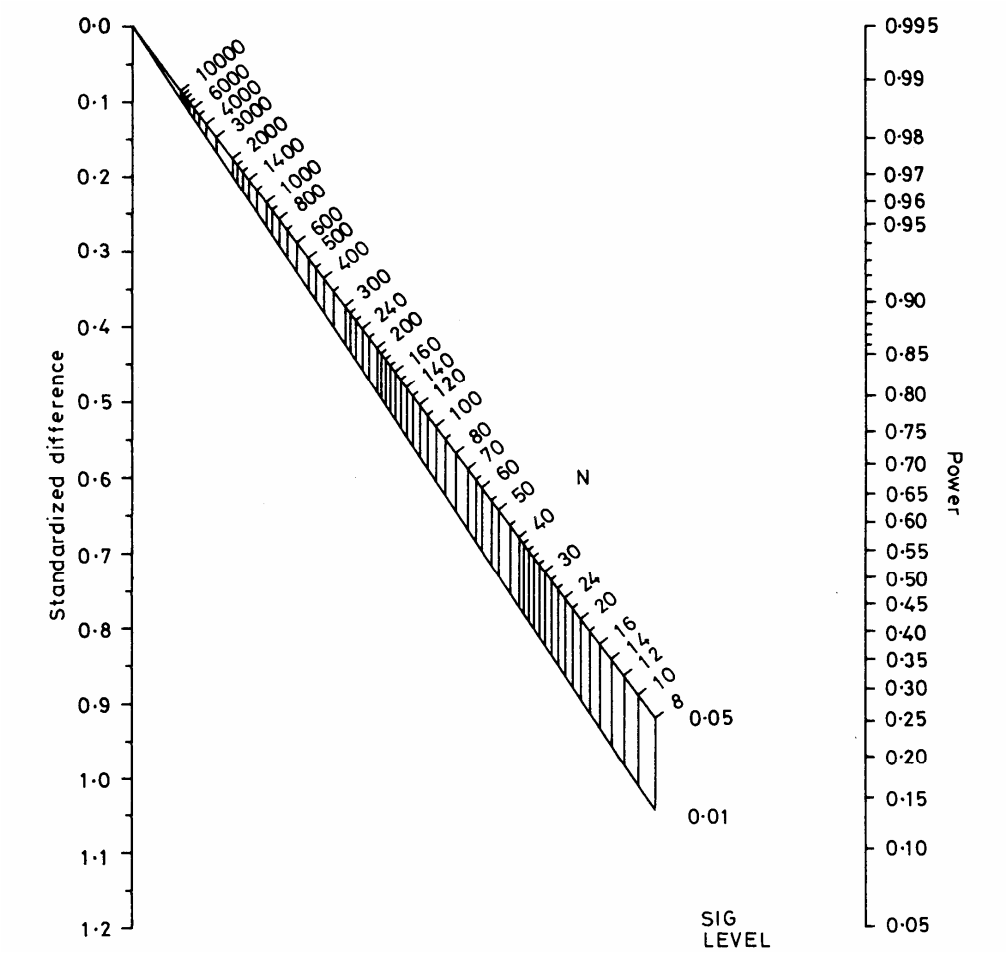
The power of a test is the probability that a study of a given size would detect as statistically significant a real difference of a given magnitude. The medical literature contains many trials that are far too small to have a good chance of detecting clinically worthwhile differences between the treatments being investigated. It is clear from many reviews of published trials that the majority have been carried out with no statistical calculation of the appropriate sample size. Unless the true treatment effect is large, small trials cannot yield a statistically significant result.

SAMPLE SIZE, HYPOTHESIS TESTS AND POWER

We can use the power of a hypothesis test to calculate the appropriate sample size for a clinical trial if we can specify the smallest true difference between the treatments that would be clinically valuable. It is this requirement that is somewhat artificial and difficult to define. In practice, however, it is usually possible to specify the degree of benefit that the new treatment would need to have over the old one for it to be a worthwhile treatment.

The main idea behind the sample size calculations is to have a high chance of detecting, as statistically significant, a worthwhile effect if it exists, and thus to be reasonably sure that no such benefit exists in reality if it is not found in the trial.

The necessary sample size is usually obtained from complicated formulae or tables, but it is much simpler to use a nomogram (see figure). All sample size calculations are based on the quantity known as the standardized difference. This is calculated in a different way for continuous or categorical data (outcome variables), but in principle it is based in each case on the ratio of the difference of interest to the standard deviations. In other words, we express the difference of interest as a multiple of the standard deviation. As we would expect, the smaller this ratio is, the larger the required size of the trial.



Nomogram for calculating sample size or power

(a) Continuous data – two independent groups

For studies of two independent groups of patients with a continuous outcome measure we need to specify the following quantities:

1. Standard deviation of the variable (in each group) (s).
2. Clinically relevant difference (δ).
3. The significance level (α - two-sided).
4. The power (1 - β).

And it is assumed that the variable has a Normal distribution in the population. The total sample size is *N*. It is common to require a power of between 80% and 90%.

The standardized difference is calculated simply as the ratio of the difference of interest to the standard deviation, that is δ/s. We can use Figure 2 to calculate the necessary sample size from the standardized difference for any desired power, choosing either 5% or 1% level of significance.

For example, suppose that we are planning a feeding trial in patients with cancer, to see if a daily supplement of fish oil will lead to an increased gain in weight after curative surgery, compared with a control group. We know from published data, for example, that patients with cancer gain on average about 6 Kg after successful surgery, with a standard deviation of 2 Kg. Suppose that the effect of the fish oil on weight gain will be considered important if it is at least 0.5 Kg more than the usual gain of 6 Kg.

We want a high probability of detecting such a difference, so we set the power to be 0.9 (90%) and choose a 1% significance level. The standardized difference is 0.5/2.0 = 0.25. We can now use the nomogram to calculate the necessary sample size. We 'draw' a straight line from the value 0.25 on the scale for the standardized difference to the value 0.90 on the scale for power, and read off the value for *N* on the line corresponding to $\alpha = 0.01$, which gives a total sample size of 900, i.e. 450 in each group.

It is easy to calculate the sample size for any combination of input values (s, δ, α, 1-β), and we can always change the sample size by altering the input values. However, it is preferable to decide in advance what the requirements are. While some modest relaxation of these is acceptable, in general if the calculated sample size exceeds what seems practical, then the study can be extended either in time, or by running the study at more centres. If it is not possible to get near to the required size of study, then the study may best be abandoned.

(b) Continuous data – paired or within person studies

The appropriate sample size for paired studies, or within person studies such as crossover trials, is obtained in a very similar way. The main difference is that the standard deviation we use is the standard deviation of the changes expected (sd). Unfortunately, an estimate of this standard deviation is often not available. If we do have a reasonable estimate of sd, we can calculate the standardized difference as $2\delta/sd$, and then use the nomogram as before. (Note the similarity to the formula for independent groups, apart from the multiplier of 2.)

(c) Categorical data

The nomogram in Figure 2 can also be used for studies which have a binary outcome variable. If the outcome variable has more than two categories, it is necessary to create a binary variable of interest. For example, if patients are to be assessed as 'improved', 'no change' or 'worse', then the sample size calculation could be based on whether or not the patient has improved.

The calculation of sample size for comparing proportions makes use of the normal approximation to the Binomial distribution, discussed earlier. It is based on the following information:

1. The expected proportion with the specified outcome in each group (p1 and p2).
2. The significance level (α - two-side)
3. The power (1 - β).

The usual way of thinking about specifying p1 and p2 is that previous knowledge should allow us to predict the proportion with the outcome in the control group (say p1), and so we need to specify the proportion with the outcome in the experimental group that would present an important improvement.

Given specified values of p1 and p2 we can calculate the standardized difference as

$$\frac{p_1 - p_2}{\sqrt{\overline{p}(1-\overline{p})}}$$

where $\overline{p} = (p1 + p2)/2$.

For example, suppose we are planning a trial to compare two analgesics in the control of pain in patients with advanced cancer. One group is to be given a new kind of analgesic, and the other group will receive a standard analgesic. On the basis of published evidence we expect that in the standard analgesic group 15% of patients will remain pain-free at 3 months. We would be interested in an improvement to 30% in the group given the new analgesic. The proportions to be compared are 0.30

and 0.15. Suppose that we want an 85% probability of detecting such a difference, if it really exists, as statistically significant at the 5% level. We can use the nomogram to work out the necessary sample size for the trial.

We have p1 = 0.30 and p2 = 0.15 so $\overline{p}$ = (0.30 + 0.15)/2 = 0.225. Using the above formula the standardized difference is given as

$$\frac{0.30 - 0.15}{\sqrt{0.225(1 - 0.225)}}$$

or 0.36. We connect the standardized difference of 0.36 to the power of 0.85 in the nomogram and read off the necessary sample size for the trial from the central axis corresponding to a significance level of 0.05, which gives N = 280. To meet the conditions specified for the trial we thus need to have 140 patients in each group.

(d) Unequal sample size

The nomogram can be used for trials in which the sample size in the two groups will be different. Sometimes it is felt desirable or necessary to use unequal (weighted) randomization. As long as the imbalance is not great, the loss in power is small.

To use the nomogram to plan a study with unequal groups, we must first calculate N as if we are using equal groups, and then calculate the modified sample size N'. If k = n1 / n2 is the ratio of the sample sizes in the two groups, then the required total sample size is

$$N' = N(1 + k)2 / 4k$$

and the two sample sizes are given by $N'/(1 + k)$ and $kN'/(1 + k)$. So, for example, if we wish to put twice as many subjects on the experimental treatment than on the control, we have k = 2, and so $N' = 9N/8$, a fairly small increase, but for k = 3 we have $N' = 16N/12$, which is an increase of a third over equal sample sizes.

(f) Getting enough patients

Often the sample size calculations reveal a required sample size that exceeds the recruiting capability of a single centre. Rather than carry out a trial that is low in power, it is often worth trying to get other centres to collaborate in a 'multicentre' trial.

Another problem is that the expected rate of accrual of patients to a trial can be much less than anticipated by the trial organizers. While this may be partly through over-optimism, it is often largely because of a failure to appropriate the effect of the trial's eligibility criteria. Restricting eligibility may lead to failure to achieve the planned sample size, and thus affect the usefulness of the trial as well as generalizability of the results. Another factor here is the proportion of eligible patients who refuse to participate. If these rates cannot be reliably estimated, then it is prudent to make an allowance for them when planning the sample size for the trial.

Many of the difficulties can be avoided by having a pilot study, which is also valuable for assessing the quality of the data collection forms, and for checking the logistics of the trial, such as the expected time to examine each patient, which affects the number that can be seen in a session. A pilot study may also provide more reliable estimates for use in sample size calculations.