Medicine and Mathematics

Statistics and ethics in medical research

III How large a sample?

DOUGLAS G ALTMAN

Whatever type of statistical design is used for a study, the problem of sample size must be faced. This aspect, which causes considerable difficulty for researchers, is perhaps the most common reason for consulting a statistician. There are also, however, many who give little thought to sample size, choosing the most convenient number (20, 50, 100, etc) or time period (one month, one year, etc) for their study. They, and those who approve such studies, should realise that there are important statistical and ethical implications in the choice of sample size for a study.

A study with an overlarge sample may be deemed unethical through the unnecessary involvement of extra subjects and the correspondingly increased costs. Such studies are probably rare. On the other hand, a study with a sample that is toc small will be unable to detect clinically important effects. Such a study may thus be scientifically useless, and hence unethical in its use of subjects and other resources. Studies that are too small are extremely common, to judge by surveys of published research.[1, 2] The ethical implications, however, have only rarely been recognised.[3, 4]

The approach to the calculation of sample size will depend on the complexity of the study design. I will discuss it here in the context of trying to ascertain whether a new treatment is better than an existing one, since it will help if the ideas are illustrated by one of the most common types of research.

Significant tests and power

Despite their widespread use in medical research significance tests are often imperfectly understood. In particular, few medical researchers know what the power of a test is. This is perhaps because most simple books and courses on medical statistics do not discuss it in any detail, even though it is a concept fundamental to understanding significance tests. Some of the general implications, however, are well appreciated, such as the awareness that the more subjects there are, the greater the likelihood of statistical significance.

Formally, the power of a significance test is a measure of how likely that test is to produce a statistically significant result for a population difference of any given magnitude. Practically, it indicates the ability to detect a true difference of clinical importance. The power may be calculated retrospectively to see how much chance a completed study had of detecting (as significant) a clinically relevant difference. More importantly, it may be used prospectively to calculate a suitable sample size. If the smallest difference of clinical relevance can be specified we can calculate the sample size necessary to have a high probability of obtaining a statistically significant result—that is, high power—if that is the true difference. For a continuous variable, such as weight or blood pressure, it is also necessary to have a measure of the usual amount of variability. A simple example will, I hope, illustrate the relation between the sample size and the power of a test.

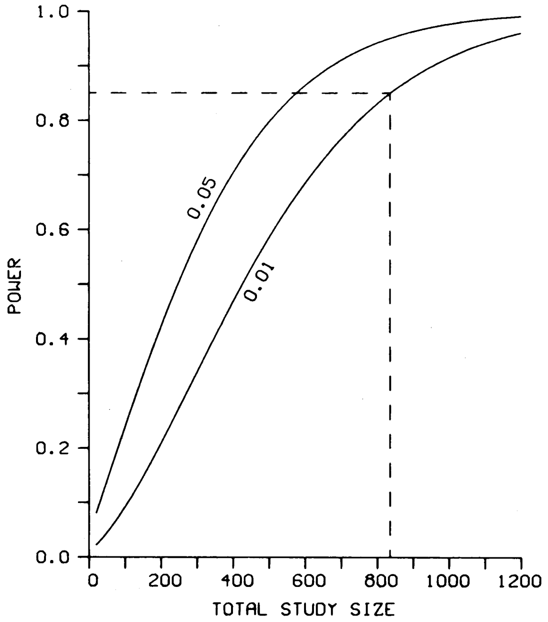

FIG 1—Relation between sample size and power to detect as significant ($p<0.05$ or $p<0.01$) a difference of 0·5 cm when standard deviation is 2 cm.

AN EXAMPLE

Suppose we wish to carry out a milk-feeding trial on 5-year-old children when a random half of the children are given extra milk every day for a year. We know that at this age children's height gain in 12 months has a mean of about 6 cm and a standard deviation of 2 cm. We consider that an extra increase in height in the milk group of 0·5 cm on average will be an important difference, and we want a high probability of detecting a true difference at least that large.

Figure 1 shows the power of the test for a true difference of

Division of Computing and Statistics, Clinical Research Centre, Harrow, Middx HA1 3UJ

DOUGLAS G ALTMAN, BSC, medical statistician (member of scientific staff)

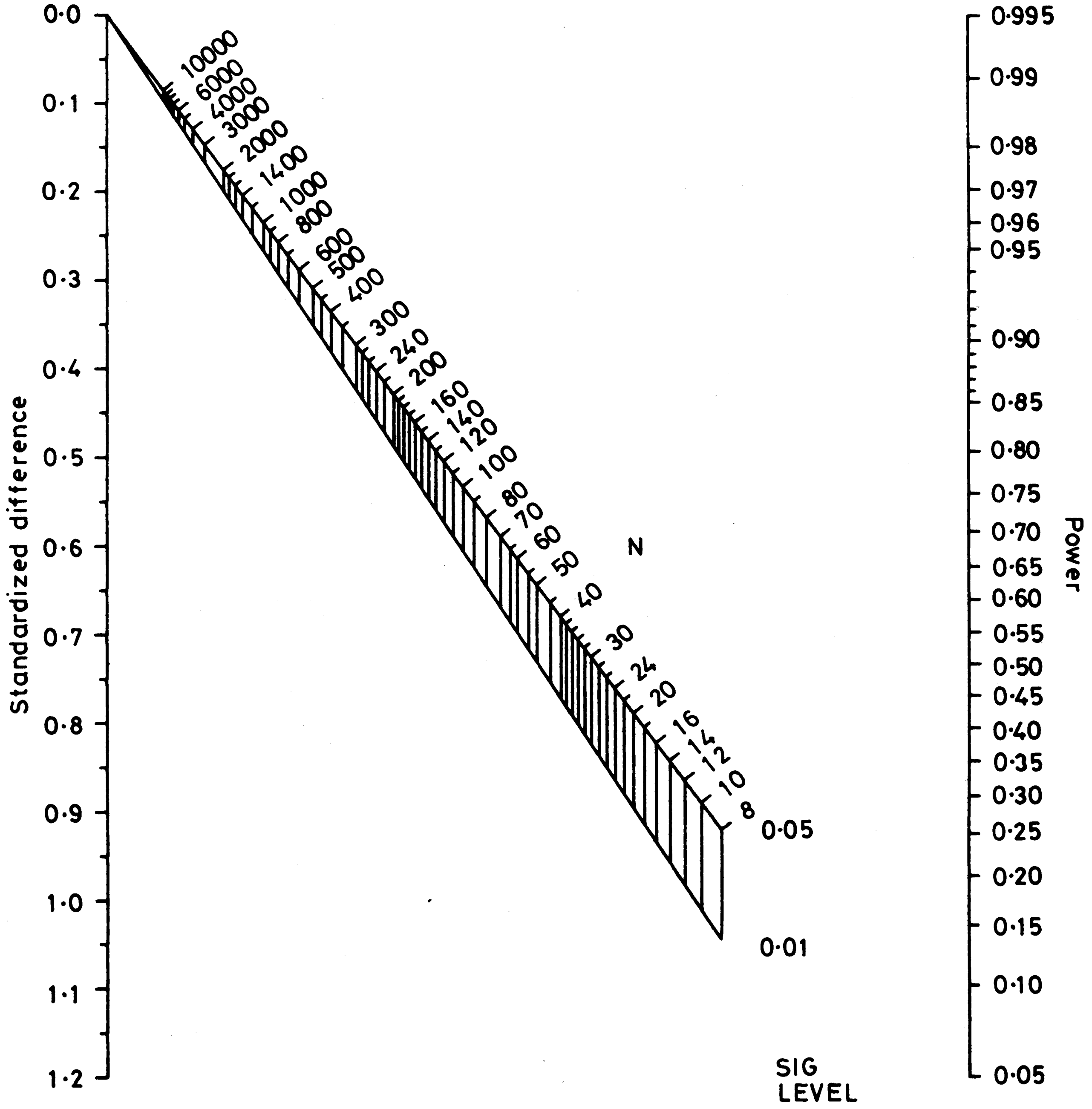

FIG 2—Nomogram for a two-sample comparison of a continuous variable, relating power, total study size, the standardised difference, and significance level.

0·5 cm. The increase in power with increasing sample size is clearly seen, as is the relation with the significance level. For any given sample size the probability of obtaining a result significant at either the 5% or 1% level, given a true difference in growth of 0·5 cm, can be read off. Power of 80-90% is recommended; fig 1 shows that to achieve an 85% chance of detecting the specified difference of 0·5 cm significant at the 1% level, we would need a total of about 840 children.

If we are told that we can have at most 500 children in all, what will the power be now? Figure 1 shows that the power drops from 85% to 60%. We are now more than twice as likely to miss a true difference of 0·5 cm at the 1% level, although the power is still about 80% for a test at the 5% level of significance. Alternatively, and not shown by fig 1, this size of study achieves the same power as the larger one for a difference of 0·65 cm instead of 0·5 cm. Whether or not this is thought sufficient will depend on how far one is prepared to alter one's criteria of acceptability for the sake of expediency. Although they are to some extent arbitrary, it is generally advisable to stick closely to the prestated criteria.

A NEW SIMPLE METHOD

The formula on which these calculations are based is not particularly simple. Graphs are preferable, but because so many variables are concerned, a large set of graphs like fig 1 would be necessary to calculate sample size for any problem. Greater flexibility, however, is achieved by the nomogram shown in fig 2. This makes use of the standardised difference, which is equal to the postulated true difference (usually the smallest medically relevant difference) divided by the estimated standard deviation. So in the previous example the standardised difference of interest was 0·5/2·0=0·25. The nomogram is appropriate for calculating power for a two-sample comparison of a continuous measurement with the same number of subjects in each group. The only restriction is the common requirement that the variable that is being measured is roughly Normally distributed.

The nomogram gives the relation between the standardised difference, the total study size, the power, and the level of significance. Given the significance level (5% or 1%),* by joining with a straight line the specific values for two of the variables the required value for the other variable can easily be read off the third scale. By using this nomogram, it is both simple and quick to assess the effect on the power of varying the sample size, the effect on the required sample size of changing the difference of importance, and so on. It is easy to confirm the earlier calculations for the milk-feeding trial.

An estimate of the standard deviation should usually be available, either from previous studies or from a pilot study. Note that the nomogram is not strictly appropriate for retrospective calculations. Although it will be reasonably close for samples larger than 100, for smaller samples it will tend to overestimate the power.

QUALITATIVE DATA

For many studies the outcome measure is not continuous but qualitative—for example, where one is looking for the presence or absence of some condition or comparing survival rates. Peto *et al*[5] have discussed calculating sample size for such studies, and they emphasise the problem of getting enough subjects when either the condition is rare or the expected improvement is not large. For example, about 1600 subjects would be needed to have a power of 90% of detecting (at $p < 0{\cdot}05$) a reduction in mortality from 15% to 10%. Although the sample size will in general need to be much larger for studies including qualitative outcome measures, the logic behind the calculations is exactly the same as with continuous data, except that a prior estimate of the standard deviation is not needed. Several authors have published graphs for general use.[6-8]

OTHER TYPES OF STUDY

Sequential designs are similarly amenable to the incorporation of considerations of power at the design stage. Indeed, it is probably much more common here than for ordinary randomised studies. For these, and for more complicated designs, it may be particularly helpful to enlist the aid of a statistician when thinking about sample size.

Conclusions

The idea behind using the concept of power to calculate sample size is to maximise, so far as practicable, the chances of finding a real and important effect if it is there, and to enable us to be reasonably sure that a negative finding is strong grounds for believing that there is no important difference. The effect of the approach outlined above is to make clinical importance and statistical significance coincide, thus avoiding a common problem of interpretation.

Before embarking on a study the appropriate sample size should be calculated. If not enough subjects are available then the study should not be carried out or some additional source of subjects should be found.[5] (It should also be borne in mind that expected accession rates tend to be over-optimistic.) The calculations affecting sample size and power should be reported when publishing results. A study[2] of 172 randomised controlled trials published in the *New England Journal of Medicine* and the *Lancet* from 1973 to 1976 found that none mentioned a prior estimate of the required sample size, and none specified a clinically relevant difference that might allow calculation of the power of their study. Obviously in most of these studies such calculations were not done.

It is surprising and worrying that in such an ethically sensitive area as clinical trials so little attention has been given to an aspect that can have major ethical consequences. If the sample size is too small there is an increased risk of a false-negative finding. A recent survey[1] of 71 supposedly negative trials found that two-thirds of them had at least a 10% risk of missing a true improvement of 50%. In only one of the 71 studies was power mentioned as having been considered before carrying out the study. It is surely ethically indefensible to carry out a study with only a small chance of detecting a treatment effect unless it is a massive one, and with a consequently high probability of failure to detect an important therapeutic effect.

*As in the example these are two-tailed significance levels.

This is the third in a series of eight articles.

No reprints will be available from the authors.

References

[1] Freiman JA, Chalmers TC, Smith H, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978;**299**:690-4.

[2] Ambroz A, Chalmers TC, Smith H, Schroeder B, Freiman JA, Shareck EP. Deficiencies of randomized control trials. *Clinical Research* 1978; **26**:280A.

[3] Newell DJ. Type II errors and ethics. *Br Med J* 1978;iv:1789.

[4] Anonymous. Controlled trials: planned deception? *Lancet* 1979;i:534-5.

[5] Peto R, Pike MC, Armitage P, *et al.* Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I Introduction and design. *Br J Cancer* 1976;**34**:585-612.

[6] Aleong J, Bartlett DE. Improved graphs for calculating sample sizes when comparing two independent binomial distributions. *Biometrics* 1979;**35**:875-81.

[7] Boag JW, Haybittle JL, Fowler JF, Emery EW. The number of patients required in a clinical trial. *Br J Radiol* 1971;**44**:122-5.

[8] Mould RF. Clinical trial design in cancer. *Clin Radiol* 1979;**30**:371-81.

A right-handed 46-year-old stonemason developed a right axillary vein thrombosis. No haematological, biochemical, or physical abnormalities were found to account for his thrombosis, and he has recovered well taking anticoagulants. Might his condition have been related to his occupation?

It might have been, especially if he had had a spell off work. Axillary vein thrombosis commonly results from unaccustomed use of the arm, including upward movements that compress the vein between clavicle and first rib.

What are the health hazards of taking small babies to public swimming pools?

Mother and baby bathing is a rewarding experience for both parent and child. It aids physical development of the baby and augments the psychological "bonding." Many public bathing pools have special mother (father) and baby bathing sessions, and those interested are advised to try to use this facility. There is the safety advantage of a poolside attendant being present. The best age to start for the baby is from 9 to 12 months, although some enthusiasts may start earlier. Much depends on the development of the baby and the confidence of the parent. The pool should be reasonably warm, between 80-85°F (26-30°C) (most public baths are 70-75°F (21-24°C)), and it is most important to let the baby gain confidence by holding him and only gradually allowing independence in the water. It is preferable to have only parents and babies in the pool, as excited older children shouting and splashing may be frightening. It is unwise to take a baby bathing until at least 1-1½ hours after his last meal. There is no more risk of contracting any infection than in any other social activity, and provided the parent is not over-enthusiastic the chance of an accident is negligible. Small babies take to bathing readily, and parents who have used the special sessions confirm that parent and baby bathing is well worth while.