How confident are you in your study?

Editorial

'How many participants do I need for my study?' During many years of providing statistical support to clinical researchers, this has been one of the most frequently asked questions. It is posed by medical students preparing dissertation projects through to seasoned professors preparing major grant proposals. The repetition of this question can be frustrating, but does at least indicate that the researchers are aware – if only vaguely – of the importance of calculating in advance the number of participants needed for their study.

But why is this so important? Put simply, failure to consider a study's optimum sample size may render the study unethical. For how can it be ethical to ask patients to undergo the rigours and possible risks of a clinical trial if there is no evidence that the sample size is sufficient to have an acceptable chance of achieving a clinically useful result? This is why many medical research ethical committees now routinely ask for a power (sample size) calculation. (If yours does not, it should!). The mathematics involved is not particularly onerous and there are inexpensive software packages available that can help (e.g. *nQuery*, *Power and Precision*). If a study sample is too small to detect a clinically important effect, future patients may be denied the benefits of an effective treatment until a more definitive (powerful) study is completed.

Studies can also be too large. Some years ago, a researcher provided his Research Ethics Committee with a power calculation showing that a sample of 160 patients was the optimum number for his study objective. However, he proceeded to recruit 240 patients. On terminating the study, the difference between the treatments was found to be statistically highly significant ($p<0.001$). Had he recruited 160 patients as originally planned, the difference would still have been significant ($p<0.01$). Therefore, 80 patients had been through the study needlessly and 40 of them had received an inferior treatment.

The sample size for a comparative study is determined in order to provide an acceptably high chance (power) of detecting a clinically significant difference between the study groups, if one exists. Conventionally, power levels of 80 to 90% are considered desirable. For example, a study to investigate the differences in verbal IQ between children with and without ADHD may choose a measure of verbal IQ that has an inherent standard deviation of 20 units and that a difference of 12 units would be considered clinically significant. A simple calculation indicates that a study of 45 ADHD and 45 control children would have 80% power (i.e. an 80% chance of detecting a real difference of 12 units or greater). Increasing the sample size to 60 participants per group would raise the power to 90%. However, it is common for groups as small as 15 or even 12 children to be used in studies of this type. Such samples provide about 33% power which means that if there is a real difference in verbal IQ of 12 or more units between the two groups, there is only about a 1 in 3 chance of detecting it.

For a survey, sample size is determined in advance but in a slightly different way: a simple survey to estimate, for example, the proportion of ADHD children with a verbal IQ level below the 16th centile at age 3 years is often carried out by testing consecutive samples of eligible children at one or more centres. The sample obtained is usually reasonably random and so can be regarded as an acceptable proxy for all children with ADHD. Nevertheless, the sample, no matter how carefully selected, will be subject to sampling error: the proportion of children in the sample with impaired verbal IQ will not be exactly equal to the proportion in the wider clinical population. The probable size of this discrepancy is best shown using a 95% confidence limit, e.g. 32% (24 to 40%). The best estimate would be, therefore, that 32% of the study group would be verbally impaired at age 3 years; we can be 95% certain that the real proportion would be somewhere in the range 24 to 40%; the precision of our estimate is ±8%. If a total of 20 children are studied, the confidence limits may be as wide as ±22%. Increasing the sample to 50 improves the precision to ±14%. If the desired precision was ±10% a sample of just under 100 children would be needed; for a precision of ±5% almost 500 children would need to be studied.

DMCN, like many other journals, receives more papers than it is able to publish. Peer review is used to select those of the highest quality for publication. For this reviewer, the inclusion of a power/sample size statement is clear evidence that the authors have considered carefully the likely quality of their study procedures and this invariably indicates a high quality paper. When a study reports a negative (null) finding, as a statistician my immediate reaction is that this could be a type II error (false-negative) result, in other words, the study was too small and insufficiently powerful to detect a clinically significant difference or effect. However, a clear statement of how large a difference the study was designed to detect usually provides the necessary reassurance that the chances of the negative finding being incorrect are acceptably small. Similarly, for a survey, an appropriately worded power statement indicates whether the precision achieved was sufficiently high for the reported findings to be clinically useable.

A power/sample size statement indicates to the journal how confident the researchers are about the ability of their study to reach a clinically useful end-point and this improves the degree of confidence the journal has in accepting the paper for publication.

DOI: 1017/S0012162203000434

Brian Faragher MSc PhD FSS
Senior Lecturer, University of Manchester Institute of Science & Technology.

We are grateful to Brian Faragher for this editorial and for the statistical advice he has given to DMCN over the years. Ed.