Chapter 14
Setting the Size

14.1   Sample Size

The most important question that a researcher should ask when planning a study is: 'How large a sample do I need?' If the sample size is too small, even a well-conducted study may fail to answer its research question or to detect important effects or associations or may estimate those effects or associations too imprecisely. Similarly, if the sample size is too large, the study will be more difficult and costly, and the size may even lead to a loss in accuracy. Hence, optimum sample size is an essential component of any research. Careful consideration of sample size and power analysis during the planning and design stages of clinical research are crucial [1].

Statistical power is the probability that an empirical test will detect a relationship when a relationship exists. In other words, statistical power explains the generalisability of the study results and its inferential power to explain population variability. Sample size is directly related to power; all else being equal, the bigger a sample, the higher the statistical power. If the statistical

power is low, this does not necessarily mean that an undetected relationship exists but does indicate that the research is unlikely to find such links if they exist [2].

With the study design and the makeup of the study sample determined, the sample size estimates can be obtained. Fundamental to estimating sample size are the concepts of statistical hypothesis testing, type-I error, type-II error and power. In planning clinical research, it is necessary to determine the number of subjects required to ensure that the study achieves sufficient statistical power to detect the hypothesised effect. If the reader is not familiar with the concept of statistical hypothesis testing, introductory biostatistics texts and many websites cover this topic. Briefly, in trials to demonstrate improved efficacy of a new treatment over placebo/standard treatments, the null hypothesis is that there is no difference between treatments, and the alternative hypothesis is that there is a treatment difference. The research hypothesis usually corresponds to the alternative hypothesis, which represents a minimal meaningful difference in clinical outcomes. Statistically, either we reject the null hypothesis in favour of the alternative hypothesis or we fail to reject the null hypothesis.

Typically, the sample size is computed to provide a fixed level of power under a specified alternative hypothesis. Power is an important consideration for several reasons. Low power can cause a true difference in clinical outcomes between study groups to go undetected. However, too much power may yield statistically significant results that are not meaningfully different to clinicians. The probability of a type-I error (>) of 0.05 (two sided) and powers of 0.80 and 0.90 has been widely used for sample size estimation in clinical trials. The sample size estimate will also allow estimation of the total cost of the proposed study [3].

A clinical trial that is conducted without attention to sample size or power information carries the risk of either failing to detect clinically meaningful differences (type-II error) because

not enough subjects have taken part or of taking an unnecessarily excessive number of samples for a study. Both cases fail to adhere to the ethical guidelines of the American Statistical Association, which recommend avoiding the use of an excessive or inadequate number of research subjects by making informed recommendations for study size [3].

14.2   What Information Is Needed to Calculate Power and Sample Size?

The components that most sample size programmes require for input include:

- Choose type-I error.
- Choose power.
- Choose clinical outcome variable and effect size (differences between means, proportions, survival times and regression parameters).
- Variation estimate.
- Allocation ratio [4].

14.3   Clinical Outcome Measures

Clearly describe the clinical outcomes that will be analysed by the statistician. The variable type and distribution of the primary outcome measurement must be defined before sample size and power calculations can proceed. The sample size estimates are mainly needed for the primary outcome. However, providing power estimates for secondary outcomes is often helpful to reviewers [4].

14.4   Effect Size

As an example, suppose a parallel group study is being designed to compare systolic blood pressure between two treatments, and the investigators want to be able to detect a mean difference of 10 mmHg between groups. This 10-mmHg difference is referred to as the effect size, detectable difference or minimal expected difference [4].

14.5   How Is the Effect Size Determined?

An effect size is chosen that is based on clinical knowledge of the primary endpoint. A sample size that B worked with in a published paper is no guarantee of success in a different setting. The selected effect size is unique to the study intervention and the specific type of participants in the study sample and, perhaps, constitutes an aspect of the outcome measurement that is unique to the clinic or laboratory [5].

   The investigator and statistician examine the literature, the investigator's own past research or a combination of the above to determine a study effect size. To investigate a difference in mean blood pressure between two treatments, the effect size options might be 2, 6, 10 or 20 mmHg. Which of these differences do you need to have the ability to detect? This is a clinical question, not a statistical question. Effect size is a measure of the magnitude of the treatment effect and represents a clinically or biologically important difference. Choosing a 20-mmHg effect size yields a smaller sample size than a 10-mmHg effect size because it is easier to statistically detect the larger difference. However, an effect size of 10 mmHg, or a smaller magnitude, may be a more realistic treatment effect and less likely to result in a flawed or wasted study [4].

14.6   Variation Estimates for Sample Size Calculations

In addition to effect size, we may need to estimate how much the outcome varies from person to person. For a continuous outcome, the hypothesised difference in systolic blood pressure, for example, is an effect size of 10 mmHg, and a study with a blood pressure SD of 22 mmHg will have lower power than a study where the SD is 14 mmHg. For a continuous outcome such as blood pressure, a measure of the variation is another part of the formula needed to compute the sample size. An estimate of variation can be derived from a literature search or from the investigator's preliminary data. Obtaining this information can be a challenge for both the clinical investigator and the statistician [4].

Consider sample size scenarios for detecting differences in blood pressure when comparing two treatments based on a *t* test. An SD of 14 mmHg is chosen to estimate the variation. Sample sizes are calculated for powers of 0.80 and 0.90 at the two-sided 0.05 significance level. Notice that the smaller effect sizes require a larger sample size and that the sample size increases as the power increases from 0.80 to 0.90. Determining a reasonable and affordable sample size estimate is a team effort. There are practical issues such as budgets or recruitment limitations that may come into play. Too large a sample size could preclude the ability to conduct the research. The research team will assess scenarios with varying detectable differences and power. Typically, a scenario can be worked out that is both clinically and statistically viable. The elements of sample size calculations presented here pertain to relatively simple designs. Cluster samples or family data need special statistical adjustments. For a longitudinal or repeated measures design, the correlation between the repeated measurements is incorporated into the sample size calculations [6, 7].

The power of a study tells us how confidently we can exclude an association between two parameters. For example, regarding the previous research question of the association between NCC and epilepsy, a negative result might lead one to conclude that there is no association between NCC and epilepsy. However, the study might not have been sufficiently powered to exclude any possible association, or the sample size might have been too small to reveal an association [1].

The sample sizes seen in the two meningitis studies mentioned earlier (?) are calculated numbers. Using estimates of prevalence of meningitis in their respective communities, along with variables such as the size of the expected effect (expected rate difference between treated and untreated groups) and level of significance, the investigators in both studies would have calculated their sample numbers ahead of enrolling patients. Sample sizes are calculated based on the magnitude of effect that the researcher would like to see in his treatment population (compared with placebo). It is important to note that variables such as prevalence, expected confidence level and expected treatment effect need to be predetermined to calculate sample size. As an example, Scarborough et al. [8] stated that, 'On the basis of a background mortality of 56 % and an ability to detect a 20 % or greater difference in mortality, the initial sample size of 660 patients was modified to 420 patients to detect a 30 % difference after publication of the results of a European trial that showed a relative risk of death of 0.59 for corticosteroid treatment'. Determining existing prevalence and effect size can be difficult in areas of research where such numbers are not readily available in the literature. Ensuring adequate sample size has an impact on the final results of a trial, particularly negative trials. An improperly powered negative trial could fail to detect an existing association simply because not enough patients were enrolled. In other words, the result of the sample analysis would have failed to reject the null hypothesis (that there is no difference between the new treatment and the alternative treatment),

when in fact it should have been rejected, which is referred to as type-II error. This statistical error arises because of inadequate power to explain population variability. Careful consideration of sample size and power analysis is one of the prerequisites of medical research [1].

References

1. Suresh K, Thomas SV, Suresh G. Design, data analysis and sampling techniques for clinical research. Ann Indian Acad Neurol. 2011;14(4):287–90. doi:10.4103/0972-2327.91951.
2. Tamela FD, Ketchen Jr JD. Organizational configuration and performance: the role of statistical power in extant research. Strateg Manage J. 1999;20:385–95.
3. American Statistical Association. Ethical guidelines for statistical practice: executive summary. Amstat News. April 1999:12Y15.
4. Adams-Huet B, Ahn C. Bridging clinical investigators and statisticians: writing the statistical methodology for a research proposal. J Investig Med. 2009;57(8):818. doi:10.231/JIM.0b013e3181c2996c.
5. Lenth RV. Some practical guidelines for effective sample size determination. Am Stat. 2001;55:187Y193.
6. Ahn C, Jung SH. Effect of dropouts on sample size estimates for test on trends across repeated measurements. J Biopharm Stat. 2005;15:33Y41.
7. Ahn C. Sample size requirement for clinical trials with repeated binary outcomes. Drug Inf J. 2008;42:107Y113.
8. Scarborough M, Gordon S, Whitty C, French N, Njale Y, Chitani A. Corticosteroids for bacterial meningitis in adults in sub-Saharan Africa. N Engl J Med. 2007;357:2441–50.