STATISTICAL BRIEFING: STATISTICAL POWER

Christopher R. Lamb

Statistical power is the probability that a statistical test will indicate a significant difference when there truly is one. The power of a statistical test is analogous to the sensitivity of a diagnostic test. Just as obtaining a negative result for a sensitive diagnostic test can be used to rule out a diagnosis,[1] a negative result for a powerful statistical test allows us to be sure that no difference exists. Knowing the statistical power of a study becomes particularly important when the result of a study is negative because there is a need to distinguish a true negative result (i.e., there really is no difference) from a false negative result (i.e., there is an underlying difference, but the study was not powerful enough to find it). Adequate power of a statistical test is usually considered to be 0.8. This means there is an 80% chance of detecting a difference if one exists.

It is normal practice to determine the power of a study before any data are collected in order to recognize when a lack of power could be an issue, and, if possible, to take steps to avoid that problem. One of the major determinants of statistical power is the sample size, hence calculations done in the planning stages of a study use assumptions about the desired power and the effect under study to estimate the minimum sample size.[2–4]

A critical factor in these calculations is the smallest difference of interest, which should represent a clinically relevant value. For continuous data, the smallest difference of interest may be expressed as a multiple of the standard deviation of observations, in which case it is called the standardized difference. For example, if planning a study to compare outcomes of conventional surgery and a novel method to treat intervertebral disc extrusion in dogs, the period of hospitalization might be considered an important outcome, and the smallest difference in hospitalization of interest might be 1 day. If, say, the mean (SD) hospitalization of previous admissions was 15 (4) days, the smallest difference of interest expressed as the standardized difference $= 1/4 = 0.25$.

Lehr's formula[5] is a quick method for estimating sample size for studies that will compare two groups using an unpaired t-test or χ^2 test: Sample size $= 16/(\text{Standardized difference})^2$

This formula assumes the probability of Type 1 and Type 2 errors to be 0.05 and 0.2, respectively. Using this method for the hypothetical study of disk extrusion treatments, sample size $= 16/(0.25)^2 = 256$ dogs. Hence, collecting hospitalization data on 256 dogs (128 having surgery and 128 having the novel treatment) would be necessary to detect a difference of 1 day in hospitalization with a sensitivity of 80%.

Including a sample size calculation in the methods section of a study report provides explicit evidence that the study was properly planned and that some thought was given to the size of effect that would be clinically important. Problems can arise if the calculation indicates the need for an unfeasibly large sample. Depending on study design and statistical methodology, it may be possible to modify data collection to reduce the sample size to a more achievable number by

- using a continuous measurement of the relevant variable rather than assigning its value to a category (e.g., small, medium, and large),
- using more precise measurement methods (or repeated measurements),
- using paired measurements, e.g. instead of comparing the average lesion size in a group of treated patients with that in a control group, measuring the change in lesion size in each patient after treatment allows each patient to serve as his or her own control and yields more statistical power,
- accepting unequal group sizes, e.g. use more control subjects if these are easy to collect,
- accepting a reduced sensitivity by increasing the smallest difference of interest. For example, in the hypothetical study of disk extrusion treatments, increasing the smallest difference of interest from 1 day to 2 days reduces the sample size from 256 to 64 dogs.

REFERENCES

1. Lamb CR. Statistical briefing: SpPInS and SnNOuts. Vet Radiol Ultrasound 2007;48:486–487.
2. Eng J. Sample size estimation: how many individuals should be studied? Radiology 2003;227:309–313.
3. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. BMJ 1995;311:1145–1148.
4. Scally AJ, Brealey S. Confidence intervals and sample size calculations for studies of film-reading performance. Clin Radiol 2003;58:238–246.
5. Lehr R. Sixteen S-squared over D-squared: a relation for crude sample size estimates. Stat Med 1992;11:1099–1102.

From Department of Veterinary Clinical Sciences, The Royal Veterinary College, Hawkshead Lane, North Mymms, Hertfordshire, AL9 7TA, UK

Address correspondence and reprint requests to Christopher R. Lamb, at the above address. E-mail: clamb@rvc.ac.uk

doi: 10.1111/j.1740-8261.2009.01546.x