Towards estimation and confidence intervals

Independent Television currently screens a dog food advertisement which claims that "eight out of 10 owners who expressed an opinion said that their dogs preferred it." Such a simple statement needs amplifying before its meaning can be assessed.

Let us assume, however, that it was a blinded crossover comparison, that the sample size of dog owners who participated was rather more than 10, and that the proportion of those approached who refused to participate was not so large as to make the validity of the sample doubtful. In a scientific journal we should then have expected to see the result assessed by a statistical significance test with a probability value telling us how likely it was that the observed result arose by chance.

Given a significant result with $P<0.05$ many of us would have taken matters no further; we should assume that an 80% expressed preference was both a real and important difference since 80% is clearly far removed from the 50% expected by chance. In doing so we would be assuming, however, that the sample surveyed was truly representative and that it gave an accurate assessment of the responses of the entire population of owners. In reality such assumptions would be most unlikely to be justified. Recognising that, we might reasonably go on to ask to what extent the sample of respondents could be used to estimate the likely true response of the population. Intuitively we should expect that the larger the sample then the more precise would be the assessment of the response; and ideally we should want to be given a range of values that is highly likely to include the true rate. This range is known as a confidence interval, and from 1 July authors of papers submitted to the *BMJ* will be expected to calculate confidence intervals whenever the data warrant this approach.

In an attempt to make that task easier Gardner and Altman (p 746) show how estimates of the likely size of differences between the responses of populations (be they cure rates, concentrations of serum constituents, or whatever) can be obtained by simple mathematics.

The standard deviation measures the spread of individual results round a mean value. It tells us nothing about the mean itself, but it can be used to calculate the standard error of the mean by taking the sample size into account and so giving a figure to express the variability or uncertainty of that mean. If the data are Normally distributed—and therefore suitable for conventional *t* testing—the calculated standard error and the *t* statistic itself can easily be used to derive a confidence interval that tells us the range within which the true population value is likely to lie.

How does this help the clinician? In a worked example Gardner and Altman consider a difference in blood pressure between groups of diabetics and non-diabetics of 6 mm Hg with a 95% confidence interval of 1.1 to 10.9 mm Hg. Given the large sample sizes (100 in each group), the difference was statistically significant ($P<0.02$). Not only is the mean difference of 6 mm Hg rather small, however, and so unlikely to have practical relevance—but we also say using the confidence interval that there is only a 2.5% chance that the true difference in the population at large is greater than 10.9 mm Hg. Again this is a figure which is unlikely to be of clinical importance; so the conclusion must be that we are unlikely to be missing a large and clinically important difference.

Suppose, however, that the mean difference in blood pressure had been four times as great (24 mm Hg) with the same numbers and spread of individual observations so as to give an unchanged standard deviation—17.7 mm Hg. We can then calculate the 95% confidence interval (see appendix II of the paper for the method) and find that it is 19.1 to 28.9 mm Hg. Since this is a difference of about 20-30 mm Hg it is likely to be of some importance. If we had wanted a greater degree of certainty we might have chosen a 99% confidence interval; the limits would then be 17.5 and 30.5 mm Hg.

Contrast this with the findings if our samples studied included only 20 diabetics and 20 non-diabetics, with the same spread of individual observations giving a standard deviation of 17.7 mm Hg. Then we can say that there is a 95% chance that the range +12.7 to +35.3 mm Hg includes the population difference. We are therefore left in some uncertainty whether the mean difference found—which is statistically significant—is an important one or not.

We are well used to the idea that simple probability values help us to decide how likely it was that an observed change in a set of values might have arisen by chance and thus by inference whether the difference might represent something other than random variation. We have always needed to go on to generalise to the population at large and so to decide what the true difference and its clinical importance might be. Confidence intervals help us to do so.

M J S LANGMAN

Professor of Therapeutics,
University Hospital,
Nottingham NG7 2UH