STATISTICS FROM THE INSIDE

3. Estimation

M J R Healy

If (as I suggested in note 2 of this series) significance testing often fails to get the essential information out of a body of data, what alternative statistical approaches are there which might be more effective? Some possibilities can be derived by considering the types of question that medical research workers are aiming to answer. Often these questions relate to a new treatment, and it might be supposed that the appropriate question would be 'Is the new treatment *effective*?'—is the average effect of the treatment (over and above that of a placebo or a standard treatment) different from zero? In fact, such a question is seldom adequate. Usually the question a clinician wants an answer to is more likely to be 'Is the new treatment *effective enough*?'—is the average effect of the treatment large enough to make it worth adopting, having regard to cost, side effects, etc? The question asks not merely whether the effect is different from zero but rather how large its size is, and the answer will be quantitative rather than 'yes' or 'no' or even 'probably'.

Suppose we are looking at a treatment intended to increase the forced expiratory volume in one minute (FEV₁) in asthmatic patients, and a well conducted trial gives us a sample average effect of +0.521 with a standard error of 0.10. The effect amounts to no less than 5.2 standard errors and is very highly significant by any reckoning; the probability of obtaining such a large result if the true average effect is equal to zero is vanishingly small, and in this sense the treatment is almost certainly having a non-zero average effect on FEV₁. But how large is this average effect? In statistical terms, what can we say about the mean of the population of patients from which our sample has been drawn (and which we reckon will contain the future candidates for the new treatment)? Surely, more than the mere statement that it is different from zero. What we would like is some kind of *estimate* of how large the population mean might be in the light of our data.

If one's arm were twisted hard enough, it would be tempting to say that the population mean was 0.52, equal to that of the sample. In statistical jargon, this single number is a *point estimate* of the quantity we are interested in. In a certain sense it is the best point estimate we can achieve, the one most likely to be close to the truth. But it is clearly far from satisfactory—there is no reason at all to believe that the population mean and the mean of the actual sample should be exactly equal. We ought to take other possibilities into account.

If then we have to consider other numbers as possible point estimates, what numbers might be acceptable? One possibility that can be quite confidently ruled out is zero. We have seen that our sample mean differs very significantly from zero—if the population mean really is equal to zero, a sample mean as large as that which we have actually observed would be extremely improbable. Zero is not an acceptable estimate of the population mean.

So 0.52 is a possible point estimate and zero is not. What about some intermediate values, 0.20 say, or 0.40? Taking the first of these as a null hypothesis, the deviation of the sample mean from the hypothetical value amounts to 0.32 which is 3.2 standard errors, corresponding still to a very small significance probability. This hypothetical value also renders the actual sample mean too improbable for it to be acceptable as a point estimate. But a population mean of 0.40 gives a deviation of 1.2 standard errors and a significance level of around 0.23, not a very small probability—if the population mean is equal to 0.40, a sample mean as large as 0.52 will occur quite frequently.

Thus any assumed value for the population mean attaches a significance probability to the observed sample mean. We might say that the sample is 'reasonably consistent' with the assumed value if this significance level is not too small. Let us agree to insist upon a particular significance level, α say, to define 'not reasonably consistent'. We can then calculate a range of values such that the sample mean is 'reasonably consistent' with any population mean within the range, in the sense that any such population mean gives the sample mean a significance level greater than α . This range of values is called a *confidence interval* for the population mean, specifically a $100(1-\alpha)\%$ confidence interval, and its extremities are called *confidence limits*. As a matter of detail, the confidence limits for the mean of a Normal population can easily be calculated as:

$$\text{sample mean} \pm (t \times \text{standard error})$$

where t is the appropriate value of Student's t . The 95% limits correspond to the 0.05 level of significance and the corresponding value of t is close to 2 except in very small samples.

It will be appreciated that the width of the confidence interval depends upon the degree of confidence demanded. If we are content with 90% confidence (a significance level of 0.10), a

relatively narrow range of values will result; if we demand 99% confidence, more extreme values of the population mean cannot be ruled out and the interval will be that much wider.

The confidence intervals for Normal means described above are symmetric about their central value. This is not an essential property; indeed, by using one tailed significance tests it is possible to construct one sided confidence intervals which extend to infinity in one direction. These are not often called for, but asymmetric intervals are appropriate when the underlying distribution is itself asymmetric or skew. Such intervals are produced when data are analysed after logarithmic transformation and the results are transformed back to the original scale for presentation purposes—odds ratios from 2×2 tables are often treated in this way. I intend to discuss data transformation later in this series.

It is clear that a confidence interval is vastly more informative than a mere significance test. Still with a standard error of 0.10, suppose that the sample mean effect is 0.22. This is 2.2 standard errors away from zero, giving a significance probability of around 0.03, beyond the conventional 5% level. Yet the 95% confidence interval is from 0.02 to 0.42, and the trial data are 'reasonably consistent' with a population average which is too small to be of any practical interest.

The added information is even more apparent when the sample mean does not attain the conventional significance level at $p=0.05$. With a standard error of 0.10, a sample mean of 0.18 gives a 95% confidence interval from -0.02 to +0.38. We cannot at this level of confidence assert an effect which is definitely positive; but we can rule out values of the population mean that are seriously negative and see that the true effect may in fact be quite substantial, ample information to justify further investigation. On

the other hand, a very large trial giving a sample mean of +0.03 with a standard error of 0.02 provides a 95% confidence interval from -0.01 to +0.07, showing not only that zero is an acceptable value for the population mean but (much more importantly) that any true effect in either direction is likely to be quite small. We have, with 95% confidence, established the 'quasi-null' hypothesis that the true mean effect is 'nearly' equal to zero.

There is one important defect in the confidence interval approach that has been illustrated in the previous paragraph. It is possible (and very desirable) to play down the malign influence of the conventional 5% level of significance by quoting the actual significance level of the data, rather than merely labelling them 'significant' or 'NS'. To present a confidence interval, though, the confidence level must be preset, and in practice the 95% level of confidence is almost invariably adopted. The magic 5% level of significance is thus reintroduced by the backdoor. This argues against the practice of presenting confidence limits in place of standard errors. The latter permit the reader to calculate her own confidence limits at any desired level of confidence using only the simplest of mental arithmetic.

Confidence limits can be calculated for means, differences between means, variances, regression and correlation coefficients, odds ratios, etc, etc. They are provided by the better statistical computer packages. For details of their calculation, see the BMJ publication *Statistics with Confidence* (1989) edited by M J Gardner and D G Altman. The important issue of the different types of question asked by clinical research workers is discussed at length in *Clinical Trials* by D Schwartz, R Flamant, and J Lellouch (Academic Press, 1980) and more briefly in my article 'Truth and consequences in medical research', *Lancet* 1978;ii:1300-1.