STATISTICS FROM THE INSIDE

6. Data structures (continued)

M J R Healy

Two level structures

Situations arise when experimental units (subunits) can be combined to form larger units (main units) which can be treated as a whole. For example, responses to several concentrations of an allergen could be measured at sites on the forearms of individual subjects. It might be required to investigate the effect on these responses of two alternative systemic drug treatments. The forearm sites (which receive the different concentrations of allergen) are then the subunits, while the subjects (who receive the drug treatments) are the main units. A design of this kind is usually referred to as a *split unit* layout. When the results are analysed there will be two different error variances, one relating to comparisons between subunits and the other to comparisons between main units. The first of these will normally be the smaller of the two, perhaps by a substantial margin.

The commonest form of split unit experiment in clinical work is the familiar 2×2 crossover trial in which two treatments A and B are administered successively to each subject, half the subjects receiving the sequence AB and the other half the sequence BA across the two treatment periods. The main units are the subjects, the subunits are the treatment occasions. The contrasts between treatments and between periods are subunit (within-subject) comparisons and as such are likely to be fairly precise. However, for these comparisons to be valid it is important that the two treatment sequences do not give different results. The contrast between sequences is a main unit (between-subject) comparison and so is likely to be noticeably less precise. The test for the presence of a difference between sequences may thus lack power (a substantial true difference may yield a non-significant result with high probability), so that an invalid experiment is liable to go undetected. A common source of differences between the treatment sequences is *carry over*, the residual effect of the treatment applied first lasting over into the second period. The two level data structure means that carry over may be present to a substantial degree and yet remain undetected. Crossover trials are only suitable for comparing treatments whose effects are known to be transient, and the experimental periods should be so arranged that the possibility of carry over is minimised.

Two level data arise in a number of circumstances, both experimental and observational. One of these is the study of laboratory performance, when it is usually necessary to distinguish between the within-batch and between-batch precision of a particular techni-

que. Another is the situation in which some physiological quantity varies both between and within subjects. In this context it is often necessary to distinguish not only between within-subject and between-subject variation* in a single quantity, but also between the within-subject and between-subject regressions relating one quantity to another. Both these common situations are often misunderstood and I hope to return to them later in this series. On a larger scale, multicentre clinical trials provide two level data, with the centres as main units and the subjects as subunits. Such trials are almost always analysed with no account taken of the possibility that the average effect of treatment may differ from one centre to another. Recent statistical developments with associated software are making it possible to undertake more thorough analyses of data like these.

Repeated measures

A superficially similar data structure of very frequent occurrence is the repeated measures design in which an outcome measure is made repeatedly on several occasions on each subject—an example is a glucose tolerance test in which blood glucose is measured at regular intervals after glucose intake. The distinctive feature of this structure is that the sequence of observations cannot be randomised. In a split unit trial, randomising the allocation of the treatments to the subunits within each main unit ensures the validity of the standard analysis; with repeated measures, error discrepancies between occasions far apart in time are liable to be larger than those between neighbouring occasions and the standard split unit analysis does not allow for this. Many textbooks and computer programs do not recognise this distinction and present a split unit analysis under a repeated measures heading. Use of these 'standard repeated measures' analyses is liable to produce erroneous results.

One approach to repeated measures data consists in analysing each of the measurement occasions separately. This cannot be said to be incorrect, but it is liable to fail to extract much of the information from the data and can be quite misleading. The different occasions will not be statistically independent so that it will be difficult to combine the evidence that they individually provide. Perhaps more importantly, it will frequently be the case that the curve

*The terms intrasubject and intersubject variation are often used, but the distinction is important enough to be worth a difference of more than two letters.

joining the occasion means will not agree with the corresponding curve for any of the individual subjects—this will be so when the individual curves differ mainly by way of expansion or contraction of the time scale (so called tempo effects). This last point is illustrated by fig 1, which shows concentration/time curves for drug concentrations in five patients. The heavy line is that joining the means for each time of the five curves. This last line answers such questions as 'What is the mean concentration at 12 minutes?', but it is of a different shape from the individual curves and it fails to answer more interesting questions such as 'What is the mean of the maximum concentrations?'.

A better approach is to try to obtain a small number of derived quantities for each subject which adequately summarise the time course of the observations. If for example the readings increased steadily with time, the slopes for the individual subjects can be calculated. If the response rises to a peak and then falls back as in fig 1, the height of the peak or the area under the curve can be used, perhaps together with the time at which the peak occurs. Each summary measure represents a single variate with a reading for each subject and these readings can be analysed in the usual way. A fuller treatment of this important topic is given by J N S Matthews, D Altman, M J Campbell and P Royston, *Analysis of serial measurements in medical research*, *BMJ* 1990;300:230–5. This paper also addresses the difficult issue of how data of this kind should be presented. The usual technique of displaying simply the occasion means with error bars of some kind, the outcome of separate analyses for each occasion, has been seen above to be inadequate if not actively misleading.

When repeated measures data occur in an experimental context, a special consideration often arises. The subjects or other experimental units will be (or at least should be) allocated to the different treatment groups by a process of randomisation. This means that we know for a fact that at time zero the treatment groups constitute random samples from a single population. This in turn implies that the population means of the different treatments at time zero are all equal. A statistical adjustment,†

Figure 1 Concentration v time for drug concentrations in five patients. The heavy line shows the mean concentration at each time.

†The appropriate technique is known as the analysis of covariance—see a forthcoming note on regression in this series.

compensating for initial differences in the sample means, can sometimes lead to substantial increases in precision and hence to more conclusive results.

Another aspect of the same topic, the fact that the true mean time curves of the different treatments all start at the same point, is the so called race track effect. Suppose for simplicity that the time course of the readings follows a straight line (as with short term growth phenomena). Then the mean treatment differences, which are initially zero, will increase steadily with time (fig 2) and will be evident both in the average heights of each line above the axis and also in the slopes of the lines. In race track terms, the horses that run faster are also the ones that go farther. If the raw data are subjected to a 'standard repeated measures' analysis of variance, the differences will show up in both parts of the analysis, as a main effect of treatment and an interaction between treatment and time, a thoroughly confusing description of what is in fact a straightforward state of affairs.

Treatment structures—factorial trials

Returning to our example involving the effects of different drugs on the responses to an allergen, in the simplest case there will be two allergen concentrations and two drug treatments. The trial as a whole then includes four treatments:

- Drug A—low concentration
- high concentration
- Drug B—low concentration
- high concentration

and these treatments now have a structure. Specifically, the four treatments consist of all the combinations generated by two *factors* (here, drug and allergen concentration) each at two *levels* (A and B, low and high respectively). Such a trial is called a 2×2 factorial arrangement.

The factorial treatment structure needs to be taken into account in the analysis of the results of the trial. Comparisons between the four treatment means give rise to three degrees of freedom

Figure 2 Response v time for three treatments in an experiment with repeated measures.

Figure 3 Treatment means from a 2×2 factorial trial. The difference between the drugs is greater at the high concentration (RS) than at the low concentration (PQ).

and these can be split into three individual contrasts:

- (1) The difference between drug A and drug B, averaged over the two concentrations;
- (2) The difference between the high and low concentrations, averaged over the two drugs;
- (3) The *interaction* contrast, the extent to which the difference between drugs is influenced by concentration.

The first two of these are known as the *main effects* of the two factors.

The concept of interaction is an important

one. It can be illustrated in fig 3. The treatment differences between the drugs at the two concentrations are given by PQ and RS, and interaction is present when these are unequal. If PQ is equal to RS, so that the two lines in the diagram are parallel, the two factors do not interact and they are said to be *additive* in their effects. It should be noted that when the two factors do interact, the main effects are unlikely to be of any interest; one cannot talk of *the* difference between the drugs when this difference depends upon which concentration is being used. This means that, when the results of a factorial trial are being considered, consideration of the possibility of interaction is logically the first step to take.

Factorial trials are by no means confined to the two level data structure, and the possibility of incorporating an extra factor in a contemplated trial is always worth considering. This is because the situation has an agreeable 'heads I win' aspect. There are two possibilities. On the one hand, the two factors may turn out to interact. In this case, knowledge of the interaction is essential to the interpretation of the results and this can only be obtained with the factorial treatment structure. Alternatively, the two factors may act additively. In this case, when measuring the effect of either one of the factors all the subjects in the trial can be used—one has, as it were, two trials (one on each of the factors) for the price of one.

Breast feeding and urinary infection

The debate about the significance of the infection risk to bottle fed babies compared with breast fed has gone on for many years and shows no sign of abating. A recent study from Naples (Alfredo Pisacane and colleagues, *Journal of Pediatrics* 1992;120:87–9) points to a protective effect of breast feeding against urinary tract infection. Over a 14 year period (1976–89) they diagnosed urinary tract infection (more than 100 000 colonies per ml of a single bacterial species in bag urines) in 128 babies aged 6 months or less. For each baby a control of the same age and sex was chosen from among ward admissions for acute illness other than urinary infection, diarrhoea, or respiratory infection. Sixty four (50%) of the cases and 93 (73%) of the controls had ever been breast fed and 16 (12%) of cases and 56 (44%) controls were breast fed at the time of admission to hospital. Babies who had been breast fed at any time had a relative risk of urinary tract infection of 0.38 (95% confidence intervals 0.22 to 0.65) compared with those who had never been breast fed. Comparing those breast or bottle fed at the time of admission the relative risk was 0.18 (0.009 to 0.36).

I'm always suspicious of trials which go on for many years but this does seem to be the first real evidence that breast feeding may protect against urinary tract infection in young babies.

ARCHIVIST