STATISTICS FROM THE INSIDE

5. Data structures

M J R Healy

**17 Moreton Avenue,
Harpenden, Herts
AL5 2EU**

Correspondence to:
Professor Healy.

No reprints available.

The appropriate statistical analysis and interpretation of a set of data depend critically upon its *structure* in terms of statistical dependence and independence. If for instance there are two populations to be compared, the samples from them may be drawn independently or they may be paired or matched. In the latter case a reading in one of the samples tells us something about the value of its mate in the other sample, and the two samples cannot be considered independent.

The data structure will depend mainly upon the manner in which the data have been collected. It is not possible to provide an adequate statistical analysis of a body of data unless the method of collection is thoroughly understood—one of several arguments for invoking statistical thinking, professional or otherwise, *before* the data collection takes place.

Paired and unpaired data

The most familiar structural distinction is that outlined above—when it is required to compare two populations, the sample items may either be matched or independent. The distinction is usually obvious, but it can be checked by seeing whether the items in one sample can be sensibly written down in a different sequence while keeping the other sample fixed; if this is not so, some kind of matching must be involved.

In both these two structures (assuming a continuous measurement-type variate is in question) the quantity of interest is the difference between the population means. The usual methods of analysis are the paired and unpaired *t* analyses. With matched pairs of items, the two populations give rise to a single population consisting of differences between pairs of items. The difference between the means of the original populations is equal to the mean of this population of differences. The mean and standard deviation of this latter population can be estimated using the differences obtained from the sample. The estimated standard error of the mean difference can thus be calculated, and the analysis proceeds in a straightforward manner along the lines described in note 4 to produce a significance probability and confidence limits.

Note that the presentation of the results from paired samples should relate to the population of differences, so that it is essential to present the sample mean difference with its standard error; the standard deviations of the two original matched samples are of little relevance and can usually be omitted. If the original data are to be plotted in a figure, the members of each pair must be joined by a line; alternatively, a plot of the differences can be given.

When the two samples are not matched, the tactic of forming differences is not available. Instead, recourse can be had to the only piece of mathematical statistics that is worth committing to memory:

> *The variance of the difference between two quantities which are statistically independent is the sum of their individual variances.*

(The simplicity of this result explains why so many statistical techniques are more easily explained in terms of variances than of standard deviations.)

Because the two samples have been drawn independently from their populations, the two sample means are statistically independent. Accordingly, the variance of their difference is the sum of their individual variances and this can be written as

$$\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}$$

where σ_1^2, σ_2^2 are the variances of the two populations and n_1, n_2 are the sample sizes. The standard error of the difference will be the square root of this.

At this point it is usual to make the further assumption that the two population variances are equal (even if the means are not). This is plausible enough in the context of significance testing at least; it seems a bit far fetched to assume as a null hypothesis that the two populations have identical means but different standard deviations. It is possible to make a formal significance test of the assumption by calculating the ratio between the sample variances. This ratio is always referred to as *F* (for Fisher) and tables are widely available from which the significance probability corresponding to a given value of *F* (and the two sets of degrees of freedom) can be obtained. This is actually not at all a good idea. There are three reasons for this:

(1) Logically, it is a misuse of a significance test. The test may throw doubt upon the null hypothesis of equal variances if the significance probability is small, but it does not establish the null hypothesis in the contrary situation—'not significant' *never* implies 'non-existent'. Furthermore, the *F* test in this context (though the best test available) is not very powerful unless the samples are large. This means that it is likely to have a fairly high type II error rate—even when the true variances are noticeably different, the probability of not obtaining a significant value of *F* may be quite substantial.

(2) Unlike the *t* test, the *F* test for the

equality of variances relies heavily upon the Normality of the two distributions. With data from quite mildly non-Normal distributions, the actual significance probability may be quite far from that in the tables. Usually in practice, the results will appear to be more significant than they actually are.

(3) The assumption is anyway not one of critical importance. The t test is ***robust*** to the assumption of equal variances, in the sense that even when the assumption is false, the tabulated significance probabilities will not usually be far from the truth.

In summary, unless the sample variances are very different (by a factor of 5 or more, say) the assumption of equal variances can usually be made safely. If the variances do differ markedly, a different method of analysis will be required. This may involve transformation of the data, perhaps to a logarithmic scale, and this will be the subject of a subsequent note in this series.

The point of assuming a common population variance is that both sample variances are now estimating the same thing. They can thus be combined into a single estimate which will be more precise than either one taken separately. The method of doing this is quite simple; the sum of squares of deviations from the mean and the degrees of freedom are calculated for each of the samples, and the total of the two sums of squares is divided by the total of the two degrees of freedom (the divisor (n_1+n_2-2) which appears in many textbooks is actually $(n_1-1)+(n_2-1)$). This procedure is known as ***pooling*** the two estimates. Note that it is equivalent to forming a ***weighted mean*** of the two sample variances using the degree of freedom as weights.

Given the pooled estimates of variance, s^2 say, the analysis of unpaired data is straightforward. The variance of the difference between the means is given by $\sigma^2(\frac{1}{n_1}+\frac{1}{n_2})$. This is estimated by $s^2(\frac{1}{n_1}+\frac{1}{n_2})$. and the square root of this quantity provides the estimated standard error of the difference. This standard error can be used along with values of t having $(n_1-1)+(n_2-1)$ degrees of freedom to provide significance probabilities and confidence intervals for the difference between the population means.

In presenting the results from unpaired data, the individual means and standard errors of the two samples are now of more relevance as the difference of the means and its standard error can be obtained from them, but it will help the reader if these derived quantities are shown explicitly. Thus part of a results table might read

	Means (SEs)	
Treatment (n=16)	Control (n=25)	Difference (T−C)
2·54 (0·25)	2·87 (0·20)	−0·33 (0·32)

Analysis of variance

The notions of paired and unpaired structures extend readily to comparisons involving three or more populations. Such comparisons may be based upon independent samples, or they may utilise matched sets or blocks of three or more items as appropriate. The related methods of analysis are respectively one way and two way analysis of variance. These methods are not as complex as is sometimes thought, and they are widely available in computer packages. It is again important to use the method which is appropriate to the actual structure of the data. If carefully matched data are analysed using a one way analysis of variance (which ignores the matching), the resulting estimate of error is likely to be too large; the work involved in the matching is wasted and the conclusions from the data will be unnecessarily imprecise.

In the present context, analyses of variance are best thought of simply as arithmetic devices for providing a single estimated error variance which can be used when comparing the means of the different groups. They do also provide an F value which is the ratio between the variance derived from the group means and the error variance. This tests the null hypothesis that *all* the population group means are equal. The usefulness of this is limited by the fact that there is no associated estimation procedure—a significant value of F suggests to us that the means are different without giving any information about the sizes or pattern of the differences. For interpretation, the analysis of variance table almost always needs supplementing with suitable tables of means and standard errors. The view sometimes put forward that t tests between individual pairs of means must not be made unless the overall value of F is significant at some level is far too simplistic to be useful.

One way analysis of variance is especially simple as it involves little more than pooling the sample variances of the different groups in exactly the same way as when two populations are concerned. Pooling may be more important, though, as variance estimates from several small samples, each one too imprecise to be very useful on its own, can be pooled to provide a single estimate of reasonable precision. Data with this structure are sometimes analysed using unpaired t tests on every pair of means. This necessarily involves the assumption that all the population variances are equal, so that it is wasteful of information not to use the single pooled estimate. As in the case of two populations, departures from the assumption of equal variance are usually not of great practical importance, unless they are systematic; if, consistently, groups with the larger means also tend to have the larger variances, transformation of the data may be advisable.

Analyses of variance can of course be calculated when there are only two populations in the dataset, the situation more usually handled by calculating a value of t. It is not always appreciated that the two analyses are exactly equivalent. The numerator of the F ratio provided by the analysis of variance has 1 degree of freedom, and the value of F is then *exactly* equal to the square of the corresponding value of t from the alternative type of analysis. The implication of this is that F ratios with 1 degree of freedom in the numerator should almost never be quoted; in the background there must be a mean with its estimated standard error, and quoting these instead will be far more informative.

A very difficult problem that arises when there are more than two populations to be compared stems from the number of possible comparisons that can be made. Suppose, for example, that there are four groups in an experiment, labelled A, B, C, and D. These give rise to six possible comparisons between pairs (AB, AC, AD, BC, BD, and CD), and there are further comparisons that may be of interest—if group A are controls, for example, we may wish to compare the mean of group A with the mean of the other three groups taken together. For each individual comparison we can control the type I error rate at (say) 0·05 by using an appropriate value of *t* in the significance tests, but the probability that at least one of the six or more comparisons will reach significance by chance alone will be considerably greater than 0·05. This is the problem of ***multiplicity*** and in my view has no logically satisfactory solution. There exist a number of ***multiple comparison*** procedures which aim to reduce the type I error probability to a specified level for all the possible comparisons taken together, but the only way to do this is to increase the value of *t* to some extent so that the type I error rate for any single comparison is less than 0·05. The inevitable consequence is that the type II error rate is increased. In a sense, the more groups there are in the experiment, the less likely is the detection of a real difference between the groups. It has, I think, to be faced that the more statements you make, the more likely it is that you will be mistaken; or, put another way, the only people that make no mistakes are those who never say anything.