STATISTICS FROM THE INSIDE

2. Significance tests

M J R Healy

To many people, significance testing is the be-all and end-all of statistical analysis and expertise in statistics is synonymous with the ability to choose a significance test suitable for a particular problem. Nothing could be further from the truth. The level of significance of a body of data (let alone the mere fact that this level does or does not surpass an arbitrary limit) at best conveys only a part of the information that the data contain, and in many problems the assessment of significance is simply a waste of time. For all this, significance testing plays such a major part in most elementary statistical teaching that it is of great importance to come to terms with the rather tricky concepts that are involved.

The argument involved in significance testing closely follows the popperian approach to scientific reasoning, though its use predates Popper's work by a considerable period. In the briefest of summaries, Popper asserts that science progresses through a sequence of conjectures and refutations. First a conjectural hypothesis is put forward to explain the situation in some branch of science. Data are then collected and the scientist decides whether these are of such a kind as to refute the conjecture. If this is so, a fresh conjecture must be made; if not, then further data, perhaps of a different kind, are required.

Suppose then that we administer a new treatment and a placebo control in random order to a number of asthma patients, measure the forced vital capacities under each condition and form the difference (treatment−control) for each of the patients. Statistically speaking, we can envisage the data as coming from a population of patients with a variate* given by the difference in forced vital capacity between the two conditions. A question we might wish to answer is 'Does the mean of the population—the average effect of the treatment over and above that of the control—differ from zero?' The *sample* average is almost certainly not equal to zero, but this may be a chance effect in the sense that, if we took further samples from the same population, we might get sample averages which were as likely to be positive as negative. How plausible is such a situation in the light of the data we have obtained?

Make the provisional conjecture that the population average difference is truly equal to zero—this conjecture is known as the *null hypothesis*. It is now possible, perhaps making further assumptions about the distribution of differences in the population, to calculate mathematically the probability of obtaining a sample mean as large as or greater than the one actually observed. This probability is the *significance level* of the data in relation to the question at issue.

Suppose next that the significance level is very small, say 0·001 or 1 in 1000. There are two possible explanations for this. Either: (a) something very improbable has happened or (b) the null hypothesis is false.

It is natural to accept the second of these and to assert that, in the light of the data, the average difference in the population is not equal to zero. Obviously this assertion may be mistaken, but if it is we have observed the sort of data which would only occur once per 1000 samples in the long run.

If on the other hand the significance level is quite large, 0·40 say, we can no longer choose alternative (b) with any confidence. The data are of a kind which would arise quite commonly if the null hypothesis were true; we are not in a position to disprove it, even provisionally. Notice that we are not in a position to accept it, to say with any certainty that the null hypothesis is true—many other hypotheses will be consistent with our data. We are at best in a 'not proven' situation.

A result which has a small significance probability is often said to be statistically *significant*. Two remarks are immediately called for. First, the word 'significant' is used as a technical term and has nothing whatever to do with its standard dictionary meaning of 'important'. The significance testing argument as outlined above says nothing at all about the actual size of the population mean. In statistical jargon, 'significant' is roughly equivalent to 'consistent' or more roughly still, and with several reservations, to 'probably real'. It is good practice to make these translations when reading the results of a statistical analysis—it shows up nonsense phrases like 'almost significant' or 'probably significant'. On the same lines, the opposite of 'significant' is 'non-significant', never 'insignificant'.

The second issue is where to draw the line between 'significant' and 'non-significant'. No such line exists and any distinction of this kind is completely arbitrary. A convention has grown up which places the dividing line at a significance level of 0·05. The only merit of this choice is that it relates closely to a deviation in a Normal

17 Moreton Avenue, Harpenden, Herts AL5 2EU

Correspondence to: Professor Healy.

No reprints available.

*A variate is an attribute—here the numerical value of the difference—associated with each member of a population; see the first article of this series (*Arch Dis Child* 1991;**66**:1355–6).

distribution equal to two standard deviations, and thus lends itself readily to mental arithmetic. The convention is harmless only if the temptation can be resisted of assuming that results with a significance level of 0·04 can be regarded as firmly established, whereas those with a significance level of 0·06 can be labelled NS and totally ignored. Some authors do not even report such findings numerically, a most pernicious practice. It is usually far better to quote actual significance levels, whatever their size.

When one of the two possible assertions (a) and (b) listed above is made, it may of course be mistaken. Two types of error can thus be distinguished. We may assert that a null hypothesis is false (that a non-zero population mean difference exists) when in fact it is true (the actual population mean difference is equal to zero). This is a false positive and is called an error of type I. Equally, and at least as importantly, we may fail to assert the falsity of a null hypothesis which is false in reality—we fail to establish the existence of a non-zero population mean difference when one actually exists. This is a sort of false negative and is called an error of type II.

The two types of error and their probabilities of occurrence can be used to throw light upon the difficult statistical question 'How large should my study be?' Suppose that, when planning a study, we decide that we can tolerate a type I error probability no larger than some small quantity α. This means that we shall consider a true non-zero mean difference as being established by our data if the significance level when we obtain them is less than or equal to α. Let us also select some non-zero value x of the population mean difference that we are anxious not to overlook. If the true difference is as big as x we would like the probability of a type II error (failing to achieve a result significant at level α) to be less than some small quantity β. The quantities α and $(1-\beta)$ are called the *size* and the *power* of the contemplated test (note that the power depends upon the choice of the value x). With the aid of subsidiary assumptions of Normality, etc, it is then possible to calculate the required number of subjects in the study. Given this number, one can (and should) also make a plot of the *power curve* which relates $(1-\beta)$ to x and shows the probability of establishing as significant at level α a difference of any particular size. Notice that the concepts of size and power are of interest at the planning stage of a study and before the data have been acquired. At the end of the study, a suitable analysis of the data can show the significance level actually achieved and also what values x are 'reasonably consistent' with the data in a sense to be described in a subsequent note.

The selection of a significance test which is appropriate to a particular set of data is a problem with two aspects. Firstly, the test must take proper account of the way the data were collected and the resulting pattern of dependencies between them. If the data collection has involved some kind of pairing or matching of observations, this should be allowed for in the statistical model for the data that underlies the test. Secondly, different tests may make different assumptions—for example, some tests for continuous data assume Normal distributions while others (known as 'distribution-free' or 'non-parametric') avoid this assumption. The choice between competing tests is not entirely straightforward and I hope to return to the subject in a later article.

Finally, let us return to the basic question that underlies any significance test, 'Does the population quantity of interest (for example, the population mean) *differ from* zero?'. In most applications it is appropriate to add the words 'in either direction' at the end of the question, so that we ask, for instance, 'Is the treatment *different from* the placebo in its effect on forced vital capacity?' It is sometimes tempting to specialise the question to indicate one direction only—'Is the mean (treatment−placebo) difference in forced vital capacity *greater than* zero?', that is, 'Is the treatment *better* than the placebo?' for example. This question leads to a *one tailed* significance test. Typically the significance level of a particular sample result is only half that of the more usual two tailed test so that a more significant result is obtained.

The possible use of a one tailed test is of theoretical interest because it opens up the idea of the *plausibility* of different hypotheses—essentially it says that hypothetical departures from the null situation in the 'wrong' direction are so implausible that they can be ignored. Once such an idea is accepted it is a short step to querying the plausibility of the null hypothesis itself. In many studies, is it really plausible that the effect of a new treatment is *exactly* equal to that of a control (note that 'roughly equal' or even 'closely equal' will not do at all)? Such a question not only casts doubt upon the usefulness of significance testing as such, but also points towards the quite different bayesian or subjectivist approach to statistical analysis, where the data are used to modify the subjective probabilities of the researcher or reader. This is currently the subject of brisk controversy among statisticians.