STATISTICS FROM THE INSIDE

8. Size, power, and confidence

M J R Healy

This note is intended to bring together and enlarge upon some of the points I have made earlier in the series on significance testing and estimation, as I have found that there are some misunderstandings that are quite widespread. At the risk of repetitiveness, let me begin by recalling the way that significance testing works. Suppose clinicians are considering doing a clinical trial of a drug treatment which is intended to increase forced expiratory volume in one second (FEV_1) in asthmatic children. Passing over many details that will be of importance in practice, they will start by contemplating a *null hypothesis*, in this case an assumption that the drug may have zero average effect (either absolutely or over and above that of a placebo) in a defined population of patients. From one viewpoint (not necessarily the most informative one) the trial may be regarded as providing an opportunity for the null hypothesis to be disproved.

When in due course the results of the trial become available and have to be interpreted, there are two kinds of mistaken conclusion that can be arrived at.

(1) The null hypothesis may be true, but it may be concluded that a real non-zero average treatment effect exists. In the jargon, the null hypothesis is *rejected* when it is actually true. This is a kind of false positive error and is known as an error of type I.

(2) The null hypothesis may be false—the treatment may have a true non-zero average effect—but the trial may fail to establish its falsity. Some statisticians would say that the null hypothesis was to be *accepted*, but this is unnecessary; a not proven verdict is enough. This is a kind of false negative error and is known as an error of type II.

Statistical significance testing is basically a methodology for controlling the rate of occurrence of type I errors. The null hypothesis asserts that the population mean effect of the treatment is zero, but almost certainly the sample mean effect will differ from zero in either a positive or negative direction—there will be a *discrepancy* between the sample mean and the hypothetical population mean. Statistical theory plus some suitable calculations enable the clinicians to assess the probability of getting, by chance alone, a discrepancy as large as the one that has actually been observed in the trial results. This probability is the *significance probability* of the trial data. If it is small (1 in 100, say), there are two possible explanations: (1) something very improbable may have occurred or (2) the null hypothesis (on the truth of which the calculation of the significance probability was based) may actually have been false.

If the significance probability is small enough, it will be natural to adopt the second of these explanations, thereby rejecting the null hypothesis. In actual fact, the null hypothesis may be true, and the decision to reject it may be mistaken; this is a type I error. The probability of making such an error is precisely the significance probability.

Suppose, though, that the significance probability of the observed discrepancy is not very small: 20–30%, say. The clinicians may then feel that they cannot reject the null hypothesis with an adequate degree of certainty. This does not of course mean that the null hypothesis is true; it may be false (the treatment may have a real non-zero average effect), and in this case a type II error will have occurred.

The clinicians considering the design of their trial will undoubtedly want to know the probability of making a type II error, and statistical theory will help to provide this. The calculation will require three pieces of information.

(1) The significance level demanded at or beyond which the null hypothesis will be rejected. This is called the *size* of the test and is usually denoted by α . The smaller the value of α (the greater the degree of protection against making a type I error), the larger is the probability of making a type II error.

(2) The precision of the sample estimate of the treatment effect. Note that this will depend upon the standard deviation of the observations and also upon the sample sizes. The worse the precision (the greater the standard deviation or the smaller the samples), the larger is the probability of a type II error.

(3) The true departure from the null hypothesis. The smaller the true departure (the smaller the true treatment effect), the larger the probability of a type II error.

The probability of making a type II error is usually denoted by β , and $(1 - \beta)$ (the probability of *not* making a type II error) is called the *power* of the test.

It will be seen that the power is quite a complicated quantity. Let us, to save trouble and thought, follow convention by choosing $\alpha = 0.05$. To find the power we still need to specify both the true discrepancy away from the null hypothesis (that is, the size of the true treatment effect), and also the data precision. This latter will involve both the inherent variability of the material (which is measured by the standard deviation) and also the sample sizes. For any particular level of precision, we can calculate the power for a whole range of possible discrepancies and plot the results, obtaining in this way a *power curve*. A set of power curves is illustrated in the figure. These

Power curves for a trial comparing two independent samples of equal size, using an unpaired *t* test. The parameter *n* is the number in each of the groups.

relate to an unpaired situation for the comparison of the mean of a treated group with that of an independent control group of equal size. The points to notice are: (1) there is a power curve for each value of *n*, the size of each of the samples, and (2) the discrepancy axis is graduated in standard deviation units.

A set of curves such as those in the figure enable the clinicians to make an instructed choice of sample size. Suppose for example that a true mean treatment effect equal to 1 standard deviation was of clinical importance and that the probability of missing such a difference—of failing to reach the 5% level of significance when the trial is performed—is to be at most 0.10 (a power of 90%). Then the figure shows that the number of patients in each group should be at least 25 or so. It also shows what power would be achieved by this sample size for other values of the discrepancy.

It will be seen again that there are three quantities to be specified before the above argument can be used: the size α , the power required $1-\beta$, and the discrepancy to be detected. What is more, the discrepancy must be expressed in standard deviation units—enough must be known about the subjects in the trial for a good guess at the standard deviation to be made.

Just as α is usually taken to be 0.05 (for no very good reason), so also values of β of 0.1 or even 0.2 are commonly used in designing a trial. The tacit implication is that it is more important to avoid recommending an ineffective treatment than it is to fail to spot one that is actually effective. It is not at all clear that such an attitude is always, or even commonly, appropriate.

It is of great importance to realise that the

whole of the above discussion, including the concepts of size and power, has related to the stage of planning the trial, before any results have been obtained. Obviously the possibility of type II errors has to be taken very seriously later on when the results are to hand and have been analysed. However, the clinicians' position *vis-à-vis* the results is now different—they are known and their significance level is no longer subject to uncertainty. The interesting problem now is to specify what hypothetical situations might plausibly have given rise to the data. We no longer have to ask 'How large a true effect might we fail to detect in our proposed trial?'; instead the question becomes 'How large a treatment effect must we still consider plausible, now that our trial results are known?'. The concept of power is replaced by that of *confidence*. Suppose that the findings do not reach significance at some preagreed level; then the original null hypothesis is to that extent consistent with them and provides one possible explanation for them. But it is of course possible to go much farther than this. It is possible to work out a whole range of hypothetical true values of the treatment effect which are in this sense 'plausibly consistent' with the actual data. This range is a confidence interval for the true treatment effect. As well as indicating the range of 'plausible' hypotheses, the width of the interval incidentally provides evidence that the design of the trial has (or has not) been successful in achieving the anticipated level of precision. Notice that a mere report of the significance level achieved—let alone the bare statement that this level did or did not fall short of the magic 5%—is quite inadequate from both viewpoints.

It follows from the above discussion that

considering the power of a proposed trial during the design phase is mandatory, for economic reasons and often for ethical reasons as well. It is inexcusable to embark upon an investigation, especially one involving animal or human subjects, which does not have a good chance of detecting an effect of the size which is considered important. This is why many statisticians demand that the power calculations made when designing the trial should be reported along with the trial results. But it must be emphasised that these calculations rank with assertions that the trial has been approved by the appropriate ethical committee, and are not of direct interest to the readers when they come to interpret the trial results. Their principal requirement at this stage will be a confidence interval for the true treatment effect, so that they can decide for themselves whether this effect may be large enough to be of practical importance in their own particular circumstances.

Confusion between power and confidence leads to difficulties when it is desired to establish a negative conclusion—say, that a treatment has a zero true effect in terms of some undesirable outcome. An argument which is sometimes put forward runs as follows:

(1) The trial was designed to have a small type II error rate β for a given level of significance α ;

(2) The results of the trial were not in the event significant at level α ;

(3) Therefore the null hypothesis of no real treatment effect can safely be accepted.

This argument fails to stand up for several reasons:

(1) The type II error rate β is the probability of failing to achieve significance for a particular non-zero size of treatment effect. At best, only the absence of an effect of this size could be asserted.

(2) The power calculations were based upon assumptions about the underlying variability of the data, which had yet to be obtained. These assumptions may not have been met—the data all too often turn out to be more variable than had been expected.

(3) The design type II error rate may have been as high as 0.1 or 0.2. Again at best, the desired assertion could only be made with this limited degree of certainty.

As I have explained, the consideration of a confidence interval for the true treatment effect makes an approach of this kind unnecessary. The appropriate confidence limit shows the size of the largest true treatment effect that is plausibly consistent with the data. Only if this limit is close to zero can the absence of an appreciable real treatment effect be confidently asserted.