Statistics in Divided Doses

May 2002 Number 4

Variability, probability and power

Contents

Random variation

What are the sources of variability in response to treatments?

Age, state of health, medical history and gender are among the many personal variables known to have potential effects on treatment response. However, even when such factors are similar, unexplained differences between individuals' response may still occur. This is known as *random variation,* and may distort the effects of drug treatment.

What is the relevance of random variation to clinical trials?

There are several factors that can be incorporated into a study design to minimise the potential effects of random variation:

- Obtain an adequate sample size. When inter-subject variability is known to be large (as with blood pressure measurements) it is necessary to obtain as large a sample as possible.
- Match samples of patients for factors known to influence outcome e.g. age.

Random variation imposes the need to determine probability levels when assessing the outcome of a clinical trial. Trial results are usually qualified by use of a probability calculation (e.g. a P value or, preferably, a confidence interval) to take account of random variation (see below).

The null hypothesis

What is meant by a null hypothesis?

This concept is a frequent source of confusion. Many clinical trials involve comparisons between drug treatments in different groups of patients and the numerical value relating to this comparison is commonly referred to as the *effect*. The null hypothesis is a prediction that this effect will be zero. Therefore, within a trial comparing two anti-hypertensive drugs, the null hypothesis would state that, the effects of two drug treatments on blood pressure are similar. Alternatively, it would imply there is no difference between treatments. Statistical hypothesis testing assesses the likelihood of an apparent difference in effect when no difference exists i.e. when the null hypothesis is true.

How does the null hypothesis differ from experimental hypotheses?

When investigators design a clinical trial to assess a new drug, their experimental hypothesis is, usually, that it may offer advantages over existing treatments. The results of the trial, if favourable, are interpreted as confirmation of this hypothesis. The statistician takes an opposing stance in assuming that the new drug has no advantages (i.e. a null hypothesis) and requires the strength of this assumption be tested. However, statistical analysis cannot be used to **prove** the apparent outcome of a clinical trial; its primary function is to set a limit on the uncertainty surrounding such an outcome. Absolute certainty is an elusive objective in clinical trials.

The 'P' value

What is meant by probability?

Probability describes the likelihood of an event occurring. It is usually measured on a scale of 0 to 1 where an impossible event is given a value of 0 and an event that is certain to occur is given a value of 1. Most events have a probability value between 0 and 1 since absolute certainty or uncertainty is virtually impossible to achieve. The probability is usually denoted by a 'P' value.

What does the P value mean?

The P value is the probability of observing the results of a trial when the *null hypothesis* is true. It is not, as is commonly stated, the probability that the observed result is due to chance.

Produced for healthcare professionals by North West Medicines Information Service,
The Pharmacy Practice Unit, 70 Pembroke Place, Liverpool, L69 3GF.
Editor: Frank Leach. Telephone: 0151 794 8117 E-mail: druginfo@liv.ac.uk

How is a P value normally expressed?

A P value is normally expressed as a decimal fraction. It may sometimes be expressed as "lower than" values e.g. < 0.05. Since the advent of computerised systems there is now a trend to express more exact values e.g. P = 0.037

How does the P value relate to statistical significance?

A P value of less than 0.05 (<0.05), i.e. 1 in 20 chance, is commonly regarded as statistically significant but this is a purely arbitrary limit. It is erroneous and potentially hazardous to assume that a P value of less than 0.05 confirms the results of a clinical trial. This can lead to the absurd situation in which P < 0.055 is regarded as insignificant and P < 0.045 as significant. A P value of < 0.05 merely implies that there is less than a 0.05 probability of observing the results of the trial, or more extreme effects, when the null hypothesis is true i.e. when there is no real difference in effect.

How does a P value relate to confidence intervals?

The two are closely related. The P value will be significant (< 0.05) only when the 95% confidence interval excludes either zero or the value specified in the null hypothesis (see *Statistics in Divided Doses issue 3* for a more detailed explanation of confidence intervals).

What limitations should we place on the interpretation of P values?

The outcome of a clinical trial can be affected by many factors including bias. The trial sample may be unrepresentative, hidden factors may affect outcome, the sample size may be too small or an inappropriate statistical test may have been used e.g. the use of a parametric test on unsuitable data. (See *Statistics in Divided Doses issue 3* for a more detailed explanation of parametric tests). As the P value provides no information about the size of a treatment effect, its value is limited.

Should trials with a statistically insignificant P value always be discarded?

No. Failure to publish trials with an apparently negative outcome is the source of so-called positive publication bias. Furthermore, a potentially meaningful difference in treatment outcome may be concealed.

What is meant by a negative trial?

A negative trial is one that fails to detect a difference in treatment outcomes. However, to describe a trial as negative solely on the basis of P values, cannot be justified; use of confidence intervals and careful assessment of the data provide a clearer picture of trial outcome.

The following interpretations can be used to explain a negative outcome:

- The null hypothesis is true and a difference in treatment effects is unlikely. This conclusion can only be reached when the trial has a sufficiently large sample size or is adequately powered
- There is insufficient evidence to reject the null hypothesis i.e. it is impossible to say for sure there is no difference between treatments. This is because the sample size is too small to allow rejection of the hypothesis. In this situation the trial is said to be under-powered (see below).

Power, sample size and errors

What is a power statement?

The power of a trial is the probability that such a trial would detect a real difference in treatment outcomes.

All trials are a compromise between perfection and practicality. Very large trials (megatrials) have proved useful in seeking answers to important treatment issues but problems associated with the use of large samples include cost and availability of patients. How do we know whether a study is sufficiently powered to assess drug effects and provide a reasonably conclusive outcome?

A typical power statement would be “0.9 (90%) at the 0.05 (5%) level of significance to detect a difference of 10mmHg systolic blood pressure….sample size *x* is needed”. This implies that the sample size calculated has a 90% chance of detecting a 10mmHg difference in systolic blood pressure with a probability value of 0.05, or 5% level of significance. It follows that for a study with a power of 0.8 (80%) at the 0.1 (10%) level of significance a much smaller sample size is needed but you will be less sure that a true difference will be detected and hence less able to reject the null hypothesis.

If the sample size obtained in the final results is less than that calculated, the study is under-powered to detect the stated difference between outcomes. However, if no sample size calculation or power statement is included in the report it is difficult to assess whether the study was adequately powered.

How is the sample size of a trial determined?

It is obvious that the power of a trial increases as the sample size increases but there has to be a trade off between the gain in power and the time and cost of testing a large number of subjects. Choosing a sample size is a difficult decision and the primary purpose of power analysis is to guide this choice.

The calculations for power depend on the size of the effect in the population. The size of the effect can be estimated from the results of previously published studies. By using standard deviation information from similar studies, together with formulae, tables or graphical methods a suitable sample size can be calculated to produce the minimum effect size that is considered important or clinically significant. Power calculations can also be performed for trials in which the data are categorical (binary).

The information contained in this bulletin is the best available from the resources at our disposal at the time of publication. Any opinions reflected are the author’s own and may not necessarily reflect those of the Health Authority

What is a Type I error?

A Type I error occurs when we falsely reject the null hypothesis and obtain an erroneously positive result. Such an error might lead us to conclude, wrongly, that a drug treatment was effective. We might suspect a Type I error when a particular hypothesis test has been inappropriately chosen or applied.

What is a Type II error?

A Type II error occurs when we incorrectly accept the null hypothesis. A clinical trial with insufficient sample size may be too weak to show a difference in the effects of treatments, even though such a difference exists and the null hypothesis is untrue.

We suspect a Type II error, when sample sizes are too small and/or confidence intervals (if provided) are large. As we have seen previously (*Statistics in Divided Doses issue 3*), confidence intervals for means and medians provide a ready means of estimating the reliability of results. A large confidence interval suggests inadequate sample size and/or wide inter-subject variability.

Subgroup analysis

What is a sub-group analysis?

A sub-group analysis is applied to assess the influence of suspected variables on treatment outcome.

For example, in recent studies of certain drugs in irritable bowel syndrome, it appeared that men didn't respond to therapy whereas women did. Therefore it would be reasonable to assess the influence of gender (a binary variable) on outcome. However, beware of studies that perform subgroup analysis, generally in an attempt to show positive outcomes in certain groups of patients. It is highly unlikely that these groups will be of sufficient sample size, and therefore adequately powered, to claim with any degree of certainty the outcomes assessed. Moreover, to be valid, such analysis may require a complex analysis e.g. multiple regression.

Relevance of bias

What is meant by bias?

Bias is any influence that distorts outcome, allowing erroneous conclusions.

There are many potential sources of bias within a clinical trial. These may be deliberate (e.g. preferential selection of certain patients or drug doses, deliberate under-powering) or covert (failure to recognise that gender or race may influence outcome, inadequate selection criteria).

How may the influence of bias be reduced?

Randomised allocation of treatments may reduce bias. This ensures that various treatments are allocated by a process of chance rather than by deliberate selection. Random selection is not a haphazard process. It ensures that each treatment has a known (usually equal) chance of selection.

Randomisation

What are the types of randomisation used?

There are various types of randomisation including *simple* and *block* forms. Simple randomisation involves use of a random numbers table or computer-generated number sequences e.g. even numbers (treatment A) vs. odd numbers (treatment B). A potential disadvantage of simple randomisation is that the balance of trial subjects taking either drug may be heavily distorted. *Block randomisation* avoids this and ensures that if the trial is stopped at any point approximately equal numbers of subjects receive A and B.

How is block randomisation applied?

Trial subjects are randomised in discrete blocks. For example, if we compare the effects of two drugs (A and B), we may decide to randomise in blocks of four. In each block two subjects will receive A and two will receive B. There are six different sequences of treatment allocation which we might choose i.e. AABB, ABBA, BBAA, BAAB, ABAB or BABA. If we allocate a number to each sequence (i.e. 1=AABB, 2=ABBA etc) we can use a random number sequence to decide the order of block selection.

The information contained in this bulletin is the best available from the resources at our disposal at the time of publication. Any opinions reflected are the author's own and may not necessarily reflect those of the Health Authority