■ REVIEW ARTICLE

Statistics in orthopaedic papers

A. Petrie

From UCL Eastman Dental Institute, London, England

Although the importance of sound statistical principles in the design and analysis of data has gained prominence in recent years, biostatistics, the application of statistics to the analysis of biological and medical data, is still a subject which is poorly understood and often mishandled. This review introduces, in the context of orthopaedic research, the terminology and the principles involved in simple data analysis, and outlines areas of medical statistics that have gained prominence in recent years. It also lists and provides an insight into some of the more common errors that occur in published orthopaedic journals and which are frequently encountered at the review stage in papers submitted to the *Journal of Bone and Joint Surgery*.

Proper statistical analysis is essential in papers submitted to the *Journal of Bone and Joint Surgery*, but it is clear to me, as a statistical reviewer, that the authors of a substantial number of these have little understanding of statistical concepts. The marked improvement, accessibility and ease of use of statistical software in recent years has led to a proliferation of errors which can be attributed to a lack of awareness of the consequences of inappropriate design and analysis rather than to mistakes in the techniques used.

This article begins with a brief overview of basic statistical theory, explaining the principles of design, the terms and the choice of technique to use in different circumstances. It goes on to highlight some of the most common and more serious misconceptions and errors found at all stages of an investigation. The concepts introduced in this paper are expounded in a number of texts.[1-4]

Design features

Statistics is not just about data analysis. It is essential to give due consideration to the principles of the design of an investigation. Even if the correct analytical method is used in an appropriate manner, valid results may not be obtained unless the design is correct.

Types of study

Studies may be either observational or experimental.

Observational studies. An **observational study** is one in which the investigator does not intervene in any way, but merely observes outcomes of interest and the factors which may contribute to them; if an attempt is made to assess the relationship between the two, the study is termed **epidemiological**. There are different types of observational study. It may be **cross-sectional**, in which case all observations are made at a single point in time, for example in a census, a particular type of survey. Often, however, the study is **longitudinal** when individuals are followed over a period of time, either prospectively or retrospectively.

A **cohort study** is an example of a prospective observational study. Individuals who do not have a condition which is the outcome of interest, are selected from the population and followed forward in time to see who does and who does not develop the condition within the study period. We collect information on factors of interest, termed risk factors, from these individuals and make comparisons to determine whether a particular one occurs more or less frequently or more or less intensively in those who develop the condition. We commonly use the relative risk to assess the effect of a risk factor in a cohort study. The risk of a condition is the chance of developing the condition; the **relative risk** is the risk in those exposed to a particular risk factor divided by that in those not so exposed. If the relative risk is 1, then the risk is the same in those exposed and those not, implying that there is no association between the risk factor and the disease. If the relative risk is greater than 1, the risk of disease is higher in those with the factor. For example, if

■ A. Petrie, MSc, CStat, ILTM, Head of Biostatistics Unit, Senior Lecturer
UCL Eastman Dental Institute, 256 Grays Inn Road, London WC1X 8LD, UK.

Correspondence should be sent to Dr A. Petrie; e-mail: a.petrie@eastman.ucl.ac.uk

©2006 British Editorial Society of Bone and Joint Surgery
doi:10.1302/0301-620X.88B9.17896 $2.00
J Bone Joint Surg [Br] 2006;88-B:1121-36.

the relative risk is 5, then an individual is five times more likely to develop the disease if the factor is present. However, if the relative risk is below 1, the individual is less likely to develop the condition if the factor is present. Thus, if the relative risk is 0.6, the risk is reduced by 40% if the factor is present. The relative risk must not be considered in isolation, but in relation to the absolute risk. If the absolute risk of developing the condition is very low (say, 0.001%), then, even if the relative risk is large, the risk is still very small if the factor is present. Because a cohort study is prospective, it can be expensive and time-consuming to perform. However, it can provide information on a wide range of outcomes, the time sequence of events can be assessed and exposure to the factor can be measured at different points in time so that changes in exposure can be investigated.

A **case-control study** is an example of a retrospective observational study. Individuals with (the cases) and without (the controls) the condition, the outcome of interest, are identified. We collect information about past exposure to suspected aetiological factors in these individuals by looking at their records or by questioning them. We then compare exposures to the factor(s) in the groups of cases and controls to assess whether any one or a combination of them makes an important contribution to the outcome. However, it is impossible to estimate the relative risk directly in a case-control study because some individuals in the study have the condition at the outset, so we cannot evaluate their risks of developing it. Instead, we use the odds ratio to relate the presence or absence of a risk factor to the condition. The odds in an individual exposed to the risk factor is the probability that an individual exposed to the factor has the condition, divided by the probability that an individual exposed to the factor does not. The **odds ratio (OR)** is then the odds of developing the condition in those exposed to the risk factor divided by the odds in those not so exposed. The interpretation of an odds ratio is analogous to that of a relative risk: an odds ratio of 1 indicates that the odds of the outcome are the same in those exposed and not exposed to the risk factor. The odds ratio is approximately equal to the relative risk if the outcome is rare. A case-control study is generally relatively quick, cheap and easy to perform. However, it is not suitable when exposures to the risk factor are rare, and often suffers from bias. **Bias** occurs when there is a systematic difference between the true results and those that are observed in the study. Bias may arise in a case-control study because of the problems emanating from obtaining information retrospectively, for example incomplete records and individuals having a differential ability to remember certain details about their histories.

Experimental studies. If the study is not observational it is **experimental**, in which the investigator intervenes in some way to effect the outcome. Such studies are longitudinal and prospective; the investigator applies the intervention and observes the outcome some time later. They include some case series, trials to assess a preventative measure (e.g. a vaccine), laboratory experiments and clinical trials to assess treatment. In the experimental setting, a **case series** is a non-comparative study comprising a series of case reports describing the effects of treatment, where each report provides the relevant information on a single patient.

A **clinical trial** is any form of planned experiment on humans which is used to evaluate the effect of a treatment on a clinical outcome. The definition can be extended to include animal experimentation. There are various principles that should be applied when designing a clinical trial to help ensure that the conclusions drawn from it are free from bias and are valid. These principles are described in the following paragraphs.

A clinical trial should be comparative. If we are investigating the effect of a new treatment and do not have any other form of management with which to compare it, we cannot be sure that any effect observed is actually due to the treatment; it may be a consequence of time alone or some other factor influencing the result. In statistical terminology, if a study is comparative, we say that it is **controlled**. An individual in the control group may be a positive control or, if ethically feasible, a negative one. Positive controls receive some form of active treatment. Negative controls receive no active treatment: they may receive nothing, but more usually are given a **placebo**, which is identical in form and appearance to the treatment but does not contain any active ingredients. Its purpose is to separate the effect of treatment from the effect of receiving it.

When allocating individuals to different treatments in a clinical trial, we want the groups to comprise individuals who possess similar characteristics at the baseline, the start of the study. If we then find that the average effect of the outcome of interest at the end of the study period is different in the groups being compared, we can attribute this to the effect of treatment rather than to any factor, such as the age of the individual or the stage of disease, which may have influenced the outcome. The way in which we achieve comparable groups at the baseline is to allocate the individuals in the clinical trial to the different treatments using **randomisation** or **random allocation**, a method based on chance. This could be by a mechanical method such as tossing a coin but is invariably based on numbers generated in a random fashion from a computer, possibly contained in a table. For example, individuals entering a trial may be allocated to treatment A if the next number in the sequence of random digits (zero to nine) is odd and to treatment B if that number is even, considering zero as even. Every individual then has an equal chance of receiving either treatment. The investigator does not know in advance which treatment an individual is going to receive as this may, consciously or subconsciously, influence the decision to include that particular individual in the trial. This simple randomisation procedure may be applied when there are three or more treatments in the study and can be modified in a number of ways. For example, in stratified randomisation, individuals are stratified by factors, such as gender, that are known to influence response, and each stratum treated as a sub-population with its own randomisation list; in blocked randomisation, randomisation is performed so that an equal number of individuals is allocated to each treatment.

It is important to design the trial so that it is free from assessment bias which may arise because an individual (the patient, those responsible for administering treatment and/or the assessor) believes that one treatment is superior, which may influence his or her behaviour. Assessment bias may be avoided by introducing **blinding**, otherwise known as **masking**. Ideally the study should be double-blind when the patient, those supervising treatment and the assessors are unaware which treatment each patient is receiving. Sometimes this cannot be achieved, in which case it is important that the patient or the assessor of the response are blind; the study then is said to be single-blind. A negatively controlled study must include a placebo treatment if the study is to be masked.

A clinical trial which is comparative and uses randomisation to allocate individuals to the different treatments is called a **randomised controlled trial** (**RCT**). The most credible type of RCT incorporates masking. These features are included in the CONSORT statement checklist (www.consort-statement.org also available at ww.jbjs.org.uk) which provides a blueprint for all aspects of design and analysis for an optimally reported randomised controlled trial.

Evidence-based medicine and the hierarchy of evidence

Evidence-based medicine has been defined by Straus et al[5] as "the conscientious, explicit and judicious use of current best evidence in making decisions about the care of individual patients". This involves knowing how to translate a clinical problem relating to the care of a particular patient into an answerable clinical question, locate the relevant research and judge its quality. It is important to recognise that different study designs provide varying degrees of evidence relating to the answers obtained from the question posed. This hierarchy of evidence for human investigations is represented in Figure 1, with the type of design most likely to provide the best evidence at the top and the weakest at the bottom. The strongest evidence is obtained from a **meta-analysis**. This is a **systematic review** of the literature which seeks to identify, appraise, select and synthesise all high-quality research relevant to the question of interest and uses quantitative methods to summarise the results. However, the arrangement of the hierarchy depends partly on the problem considered. We would choose to perform a RCT to investigate a novel treatment; but to identify risk factors for a disease outcome, a cohort or case-control study will be more appropriate.

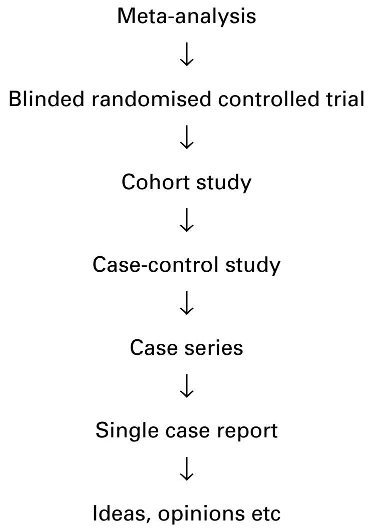

Fig. 1

The hierarchy of evidence (the strongest evidence is provided at the top).

Some common terms

If we are to appreciate fully the benefits and pitfalls associated with statistical analysis, we should be familiar with the terms used and the important concepts which underlie the theory.

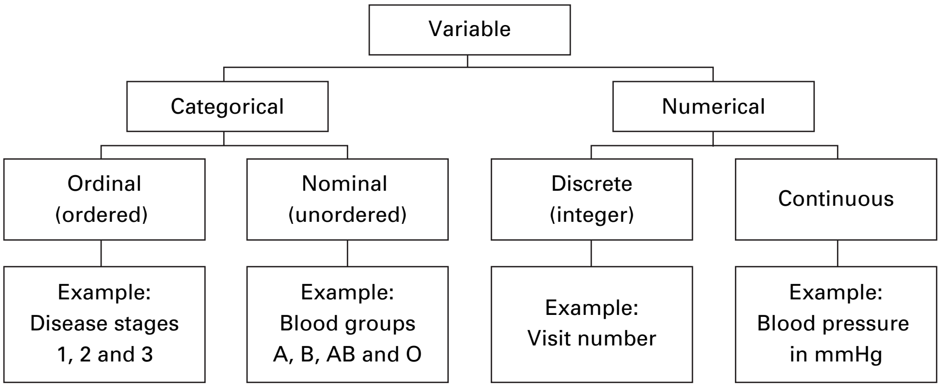

Fig. 2

Categorisation of a variable.

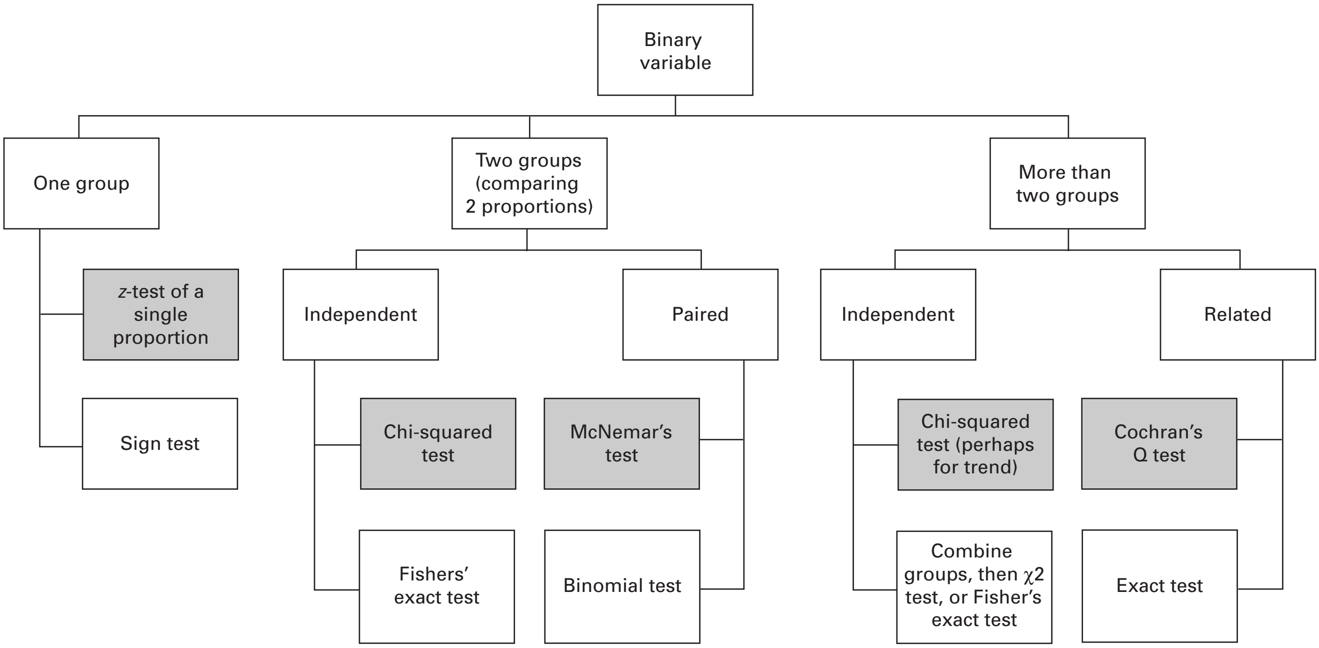

Fig. 3

Flowchart indicating choice of test when the data are binary (tests in the shaded boxes require relevant assumptions to be satisfied).

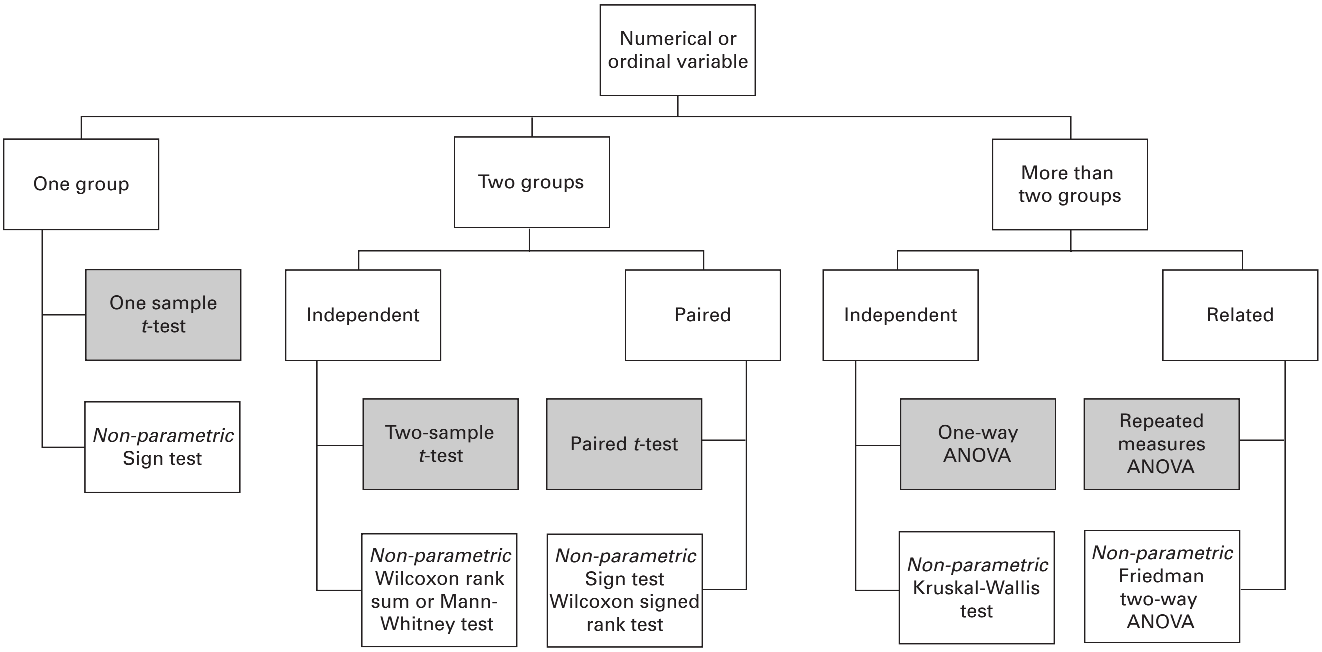
Numerical or ordinal variable
One group
Two groups
More than two groups
One sample *t*-test
Non-parametric Sign test
Independent
Paired
Two-sample *t*-test
Paired *t*-test
Non-parametric Wilcoxon rank sum or Mann-Whitney test
Non-parametric Sign test Wilcoxon signed rank test
Independent
Related
One-way ANOVA
Repeated measures ANOVA
Non-parametric Kruskal-Wallis test
Non-parametric Friedman two-way ANOVA

Fig. 4

Flowchart indicating choice of test when the data are numerical (tests in shaded boxes require relevant assumptions to be satisfied) (ANOVA, analysis of variance).

A **variable** is a quantity that can take various values for different individuals. The variable is either **categorical**, when each individual belongs to one of a number of distinct categories, or **numerical** when the values are discrete or continuous (Fig. 2).

The type of variable determines the form of analysis which we adopt. For example, we typically use a bar chart to illustrate a set of categorical observations, the proportion as the summary measure of interest and a chi-squared test to compare proportions in different groups. In contrast,

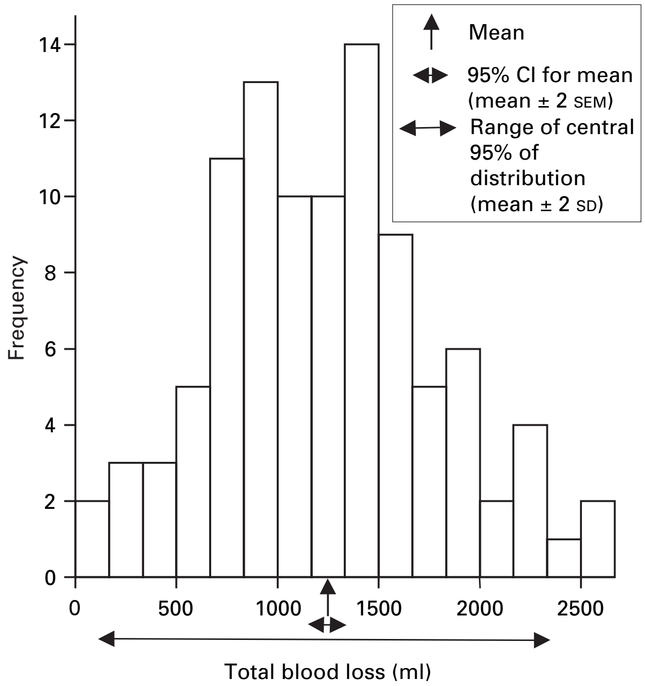

Fig. 5

Histogram showing the distribution of total blood loss in patients undergoing unilateral knee replacement (hypothetical data).

we often use a histogram to illustrate a set of numerical observations, the mean as the summary measure of interest and a *t*-test to compare means in two groups. The terms in this paragraph are explained in the sections which follow, and the different forms of analysis are outlined more fully in the flowcharts in Figures 3 and 4.

The **data** which we collect represent the observations made on one or more variables of interest. The term **statistics** encompasses the methods of collecting, summarising, presenting, analysing and drawing conclusions from such data. These methods may be descriptive or inferential. **Descriptive statistics** are concerned with summarising the dataset, often in a relevant table or diagram, to give a snapshot of the data. However, we have to recognise that it is rarely feasible to study a whole population of individuals. Instead, we select what we hope is a representative sample and use the data to tell us something about the population; this process is called **inferential statistics**. The two inferential processes are estimation of the population parameters which characterise important features of the population, and testing hypotheses relating to them.

Summarising data

Diagrams

It is necessary to condense a large volume of data in order to assess it. A simple approach is to create a table or a diagram to illustrate the **frequency distribution** of the variable. This shows the frequency of occurrence of each observation or group of observations, as appropriate. We commonly display the frequency distribution of a categorical variable in a **bar chart** in which a separate bar is drawn for each category, its length being proportional to the frequency in that category, or in a **pie chart** which is split into sections, one for each category, with the area of each section being proportional to the frequency in that category. We often use a **histogram** to illustrate the frequency distribution of a continuous numerical variable. In order to draw the histogram, we create between 10 and 15 contiguous classes and, for each class, draw a bar whose area is proportional to the frequency in that class (Fig. 5). The histogram is similar to a bar chart but the former has no gaps between the bars. By looking at the shape of the bar chart or histogram, we can assess whether or not the distribution of the variable is symmetrical. It is symmetrical when the left hand side is a mirror image of the right (Fig. 5 is an approximation), skewed to the right when there is a long tail to the right with a few high values, and skewed to the left, when there is a long tail to the left with a few low values.

Summary measures

One way of encapsulating a set of observations is to calculate one or more summary measures which describe important characteristics of the data.

Categorical data. It is common to summarise categorical data by evaluating the number and proportion or percentage of individuals falling into one or more categories of the variable. For example, the gender of an individual is a nominal categorical variable with two categories, male and female. If we know that there are *m* males in our sample of size *n*, then the proportion of males in the sample is $p = m/n$. The percentage of males is obtained by multiplying the proportion by 100, to give 100 *m*/*n*%.

Numerical data. If the variable is numerical (for example, the total blood loss in millimetres in a patient following knee replacement), and we have an average value for a set of observations (a measure of location) with some concept of how widely-scattered the observations are around this average (a measure of spread), then these two quantities alone will enable us to conceptualise the distribution of the data.

The two forms of average which are used most often are the **arithmetic mean**, usually simply called the mean, and the **median**. The arithmetic mean is obtained by adding all the observations and dividing this sum by the number in the dataset. The median is the observation which falls in the middle of the set of observations when they are arranged in increasing order of magnitude. For example, if there are nine observations in the dataset, the median is the 5th, having the same number of observations both below and above it. Provided the frequency distribution is approximately symmetrical, the mean is generally preferred to the median as it has useful statistical properties. If the data are skewed, however, the median is more useful since, unlike the mean, it is not unduly influenced by **outliers**, which are extreme values that do not appear to belong to the main body of the data. If the distribution is skewed to the right, it is possible to create a symmetrical distribution by taking the logarithm of each observation, as long as the values are greater than zero, to any base (typically to base 10 or base *e*) and then plotting the distribution of the logged values. The antilog (i.e. the back-transformed value) of the mean of the logged values is called the geometric mean, and this is a useful summary measure of location for skewed data.

The simplest summary measure of spread is the **range**, which is the difference between the largest and smallest observations in the dataset. However, because it is heavily influenced by outliers, it should be used cautiously. Sometimes, we use a modified range, such as that which omits from the calculation a certain percentage of the observations from both ends of the ordered set. An example is the **interquartile range** which contains the central 50% of the ordered observations, excluding the highest and lowest 25%. An alternative measure of spread which uses every observation in the dataset is the **variance**. In order to calculate the variance from sample data, we subtract each observation from the mean, square these differences, add all the squared differences together and divide this sum by the number of observations in the sample minus one. The variance is approximately equal to the arithmetic mean of the squared differences. In order to obtain a measure which has the same dimension as the original observations, rather than one which is evaluated in squared units, we take the square root of the variance as our measure of spread; this is called the **standard deviation** (**SD**). It can be thought of as an average of the deviations of all the observations from the mean of the observations.

If the data are distributed approximately symmetrically, we can then use our sample values of the mean and standard deviation, together with the properties of a theoretical probability distribution called the **Normal** or **Gaussian** distribution, to provide an interval within which we expect most of the observations in the population to lie. This interval is equal to the mean ± 1.96 × SD. The figure 1.96 is often approximated to 2 so that the interval containing the central 95% of the observations is approximately equal to the mean ± 2 SD. Suppose, for example, the mean total blood loss in 100 patients following unilateral total knee arthroplasty is 1239.7 ml, the SD of this set of observations is 553.5 ml and the distribution of total blood loss is approximately Normal. Then the interval containing the central 95% of the observations is approximately equal to 1239.7 ± 2 × 553.5, i.e. it is from 132.7 ml to 2346.7 ml (Fig. 5).

If observations on the variable of interest are taken on a large number of healthy individuals in the population, this range of values is commonly called the **reference range/interval** or the **normal range**. It is the range of values within which we expect an individual's value for that variable to lie if the individual is healthy. If the value lies outside this range, then it is unlikely (the chance is at most 5%) that this person is healthy with respect to that variable.

Estimating parameters

Estimating the mean

The standard error of the mean. When we have a sample of data, we have to recognise that, although we hope that its values are a reasonable reflection of those in the population from which it is taken, the sample estimates of the parameters of interest may not coincide exactly with the true values in the population. For example, the sample mean is unlikely to be exactly the same as the population mean. Hopefully, the discrepancy between the two will not be too great. In order to assess how close the sample mean is to the population mean and, therefore, to decide whether we have a 'good' or precise estimate, we calculate the **standard error of the mean** (**SEM**); the smaller the SEM, the better the estimate.

The SEM is equal to the SD divided by the square root of the sample size, i.e. SEM = SD/√*n*. Thus the SEM is directly proportional to the SD and inversely proportional to the sample size, and large samples which exhibit little variability provide more precise estimates of the mean than small diverse samples. Thus, if the standard deviation of the total blood loss measurements following surgery on 100 patients is 553.5 ml, the SEM is equal to 553.5/√100 = 55.35 ml.

95% confidence interval for the mean. Although the SEM provides an indication of the precision of the estimate, it is not a quantity that is intuitively helpful. In the example, although we know that the mean total blood loss is estimated to be 1239.7 ml and the SEM is 55.35 ml, it is difficult to assess whether or not the estimated mean is precise. However, we may assess the adequacy of our estimate by creating an interval which contains the true population mean with a prescribed degree (usually 95%) of certainty. This is the usual interpretation of the **95% confidence interval** (**CI**) for the mean. Strictly, the 95% CI for the mean is the range of values within which the true mean would lie on 95% of occasions if we were to repeat the sampling procedure many times. The approximate 95% CI for the mean is obtained by adding and subtracting 2 × SEM to and from the sample mean, i.e. it is the mean ± 2 × SEM. If the confidence interval is wide, then the estimate of the mean is imprecise but if it is narrow this is precise. It is also useful to look at the upper and lower limits of the 95% CI and assess the clinical or biological importance of the result if the true mean were to take either of these values. Thus, in the example of blood loss, the approximate 95% CI for the true mean total blood loss is equal to

1239.7 ± 2 × 55.35 = 1129.0 ml to 1350.4 ml, as shown in Figure 5.

The 95% CI is an extremely useful way of assessing an estimate and, should always be supplied whenever an estimate is provided in the results section of a paper. It is certainly more helpful to present the CI rather than the standard error, which is used to create the confidence interval, or the SD, which is a measure of spread rather than of precision.

Estimating the proportion

The standard error of the proportion. Just as the SEM provides a measure of the sampling error associated with the process of estimating the population mean by the sample mean, the **standard error of the proportion,** SE(p), provides a measure of the sampling error associated with the process of estimating the population proportion by the sample proportion. It is estimated in the sample by SE(p) = $\sqrt{\{p(1-p)/n\}}$. Note, that if the proportion, p, is replaced by a percentage, then the '1' in the formula is replaced by '100'. So for example, if 49 patients out of 100 (i.e., 49%) undergoing unilateral total knee replacement need a blood transfusion, the estimated standard error of this percentage is

$$\sqrt{\{49\ (100 - 49)/100\}} = 5.0\%.$$

95% confidence interval for the proportion. Analogous to the 95% CI for the mean, the 95% CI for the proportion may be interpreted loosely as the range of values which contains the true population proportion with 95% certainty. It is derived similarly to the CI for the mean in that it is approximately the sample estimate ± 2 times the standard error of the estimate, i.e. $p \pm 2 \times$ SE(p). Thus, the approximate 95% CI for the true percentage requiring transfusion in the example is 49 ± 2 × 5.0 = 29% to 59%.

Testing hypotheses

The process

Statistical testing of a hypothesis is an inferential process in that we use our sample data to draw conclusions about one or more parameters of interest in the population. We usually embark on a study with a hypothesis about the population in mind. We may be evaluating a new treatment because we believe that it is more effective than the standard management of a given condition. Our hypothesis then is that the new treatment is better in some defined way than the standard. We use the tools of statistics to learn whether this is likely to be true by deciding whether we have enough evidence in our sample to reject the **null hypothesis, H_0**, that there is no treatment effect in the population, i.e. that the two treatments are equally effective. We calculate a test statistic, a formula specific to the test under investigation into which the sample values are substituted, and relate it to the relevant theoretical statistical distribution (e.g. the Normal, t, F, or chi-squared) to obtain a **p-value**. The p-value is the probability of obtaining the sample values, or values more extreme than those observed, if the null hypothesis about the population is true. The p-value, which ranges from zero to one, links the sample values to the population.

If the p-value is small, then there is a very small chance of getting the sample values if H_0 is true. Since the sample values exist and we cannot change them, the implication is that H_0 is unlikely to be true, and we say that we have evidence to reject H_0 and the test result is statistically significant.

If the p-value is large, then there is a good chance of getting the sample values if H_0 is true. Because the sample values exist, the implication is that H_0 is likely to be true, and we say that we do not have evidence to reject it. This is not the same as providing evidence that the null hypothesis is true, only that we do not have evidence to reject it. The result of the test is then said to be statistically not significant.

The problem is to specify the **level of significance** which distinguishes between large and small p-values. In fact, an arbitrary value of 0.05 was selected many years ago to provide a cut-off value for the p-value. If $p < 0.05$, we reject H_0; otherwise we say that we have no evidence to reject it. If we reject H_0, we reject it in favour of the **alternative hypothesis**. Generally this is non-directional so that if H_0 is that two population means are equal, the alternative hypothesis is that they are not equal. The test is then said to be two-tailed in that there are two possibilities; either the mean of treatment A is greater than that of treatment B, or *vice versa*. One-tailed alternatives are rarely used because we have to be absolutely certain for biological or clinical reasons, in advance of collecting the data, that if H_0 is not true the direction of the difference is known (e.g. that the mean of treatment A is greater than that of B), and this is rarely possible.

The tests

One of the hardest tasks facing the non-statistician is to decide which hypothesis test to use in a given situation. With the advent of personal computers and, in particular, the introduction of the Windows environment, it is relatively easy to produce a test statistic and a p-value once the test has been chosen. How is the test chosen? The flowcharts in Figures 3 and 4 should aid in this decision when dealing with a single variable of interest. A series of questions must be answered in order to progress down a flowchart and arrive at a suitable test:

1. Is the variable categorical or numerical? This will determine whether to use the flowchart in Figure 3 (for binary categorical data to compare two or more proportions) or that in Figure 4 (for numerical data to compare two or more means or medians).
2. How many groups are being compared?
3. Are the groups independent when each group comprises individuals who are unrelated to those in the other group(s) or are they related with the same individual providing measurements on different occasions?
4. Are the assumptions underlying the proposed test satisfied? If not, is there an alternative test which can be used which does not rely on these assumptions? Tests which make no assumption about distribution of the data are called **non-parametric** or **distribution free** tests.

Hypothesis tests: example Total knee replacement (TKR) is associated with major post-operative blood loss and blood transfusion is frequently required. With the increased concern about the risks of blood transfusion, various methods of conservation of blood have been studied. The most appropriate solution is to reduce the loss of blood during and after an operation. Benoni and Fredin[6] performed a prospective double-blind randomised controlled trial designed to evaluate the effect of a fibrinolytic inhibitor, tranexamic acid, on blood loss in patients managed with TKR. Patients who were scheduled to have a unilateral TKR were randomly divided into two independent groups. The 43 patients in the treated group were given 10 mg/kg body-weight tranexamic acid intravenously shortly before the release of the tourniquet and repeated three hours later, in addition to the standard procedures used to control bleeding. The 43 patients in the control group were treated similarly but received a placebo (physiological saline) intravenously instead of tranexamic acid. The PFC total knee prosthesis was used for all patients, and the treated and control groups were found to be comparable with respect to factors likely to influence outcome, such as gender, age, height, weight, whether a cemented or uncemented arthroplasty was used, tourniquet time, and time to drain removal. Blood loss at the end of surgery was recorded by measuring the volume in the suction apparatus and estimating the loss in the swabs. Post-operative blood loss was recorded from the drain bottles at 1, 4, 8 and 24 hours post-operatively and on drain removal at 24 to 33 hours. The null hypothesis for the two-sample *t*-test was that there was no difference in the mean total blood loss after the operation in the two groups in the population. The two-tailed alternative was that these means were different. The mean total blood loss (95% confidence interval, CI) in the treated and control groups, was 730 ml (644 to 816 ml) and 1410 ml (1262 to 1558 ml), respectively. The difference between these means (95% CI) was 680 ml (511 to 849 ml), test statistic 5 8.0, $p < 0.001$. Hence there was strong evidence to reject the null hypothesis, suggesting that tranexamic acid should be used to reduce total blood loss. In addition, the investigators found that 24 patients (55.8%) in the control group required a blood transfusion compared with only eight (18.6%) in the treated group. The difference in these percentages (95% CI) was 37.2% (18.4% to 56.1%); the test statistic from a chi-squared test (testing the null hypothesis that the two percentages were equal in the population) was 11.2, giving $p < 0.001$ so there was strong evidence to suggest that treatment with tranexamic acid reduced the risk of transfusion in such patients.

Errors in testing a hypothesis

When we make a decision to reject or not reject a hypothesis based on the magnitude of the p-value we have to realise that it may be wrong. There are two possible mistakes that we can make:

Type I error - when we incorrectly reject the null hypothesis.

Type II error - when we are wrong in not rejecting the null hypothesis.

The chance or probability of making a Type I error is actually the p-value that we obtain from the test. The maximum chance of a Type I error is the significance level, the cut-off that we use to determine significance, typically 0.05. Then if $p < 0.05$, we reject the null hypothesis, and if $p \geq 0.05$, we say that we do not have evidence to reject the null hypothesis. If we do not reject the null hypothesis, we cannot be making a Type I error since this error only occurs when we incorrectly reject the null hypothesis. When we define our level of significance at the outset, before we collect the data, we are limiting the probability of a Type I error to be no greater than the level chosen.

Instead of concentrating on the probability of a Type II error, the chance of not rejecting the null hypothesis when it is false, we usually focus on the chance of rejecting the null hypothesis when it is false. This is called the **power** of the test. It may be expressed as a probability taking values from 0 to 1, or, more usually, as a percentage. We should design our study in the knowledge that we have adequate power (i.e. ≥ 80%) to detect, as significant, a treatment effect of a given size. If the power is low, then we may fail to detect the effect as significant when there really is a difference between the treatments. We will have wasted all our resources and the study could be ethically unacceptable.

Sample size estimation

How large should the study be? The sample should be large enough to be able to detect a significant treatment effect if it exists, but not too large to be wasting resources and have an unnecessary number of patients receiving the inferior treatment. In order to determine the sample size, we need to have some idea of the results which will be obtained from the study before collecting the data. We must have an appreciation of the variability of our numerical data in order to perform a significance test to investigate a hypothesis of interest (e.g. the null hypothesis that two means are equal). We will need a larger sample if the data are very variable than when they are less so in order to detect, with a specified power, a treatment effect of a given size. Consideration must be given to the significance level and the power of the test, which also affect sample size; the lower the significance level and the higher the power specification, the harder it is to obtain a significant result, and the greater the required sample size. In addition, we have to think carefully about the effect of treatment (e.g. the difference in two means) which we regard as clinically or biologically important. It is easier to detect a large effect than a small one so a smaller sample size is required if the specified effect is larger. All these factors must have numerical values attached to them at the design stage of the study in order to determine the sample size. Sometimes a **pilot study** is needed to estimate the numerical values. This will be necessary if we require an estimate of the standard deviation to describe the variability of the data and there is no published or other material which provides that information. Then, having decided on the test which we believe will be appropriate for the hypothesis of interest, we incorporate the values for these factors into the relevant statistical process to determine the optimal sample size. In order to justify this choice we should provide, in the study protocol and the final paper, a power statement in which we specify the values of all these factors.

There are various approaches to determining sample size. There are specialist computer programs such as nQuery Advisor 6.0 (Statistical Solutions Ltd., Cork, Ireland), books of tables such as Machin et al[7] and a diagram called Altman's nomogram.[8] There are also formulae as provided by Kirkwood and Sterne[3] and simplified formulae devised by Lehr.[9]

As an illustration of the process, let us consider the method which Lehr[9] used to provide crude estimates of sample size when the power of the proposed hypothesis test was fixed at 80% and the level of significance of the two-tailed test set at 5%, so the null hypothesis would be rejected if $p < 0.05$. He showed that for a two-sample comparison, the optimal sample size in each group, assuming n observations in each group, is equal to:

n = 16/(standardised difference)^2
where the standardised difference = d/s
and:

1. For the two-sample t-test comparing two means:
d = important difference in means
s = standard deviation of the observations in each of the two groups, assuming they are equal

2. For the chi-squared test comparing two proportions:
d = important difference in proportions, $p_1 - p_2$

$$s = \sqrt{\bar{p}(1-\bar{p})} \text{ where } \bar{p} = \frac{p_1 + p_2}{2}$$

We can modify the numbers if we require unequal sample sizes or if we expect dropouts during the study. If we are investigating a number of variables of similar importance, we can calculate the optimal sample size for each and base the study on the largest of these.

Sample size estimation: example Consider the example which was used to illustrate the two-sample t-test.[6] At the design stage of the study to compare the mean total blood loss in two groups of patients scheduled to have a total knee replacement, the investigators need to decide on the number of patients to have in each group, assuming equal numbers in the two groups. Both groups will be managed with the standard method of haemostasis, but, in addition, the treated group of patients will be given tranexamic acid intravenously whilst the control group will be given a saline placebo intravenously instead. Let us suppose that the investigators believe that if the mean total blood loss after operation differed by at least 250 ml in the two groups this would be an important difference. They want to know how many patients would be required in order to have an 80% chance of detecting such a difference at the 5% level of significance, if they believe the standard deviation of the observations in each group is around 410 ml. Hence, the standardised difference is 250/410 = 0.61, and using Lehr's formula, the optimal sample size in each group is 16/0.61^2 = 43.

Relationships between variables

The statistical methods described up to this point have all related to a single variable. We are often interested in investigating the relationships between two or more variables. In this case, we generally proceed by estimating the parameters of a suitable **regression model** which describes the relationship in mathematical terms.

Two variables

Univariable linear regression. In univariable or simple linear regression we are concerned with investigating the linear relationship between a numerical variable, y, and a second numerical or ordinal variable, x.

Every individual in the sample has a single measurement on each variable. We start by plotting the pair of values for each individual on a **scatter diagram**. This is a two-dimensional plot with the values of one variable, typically y, represented on the vertical axis, and the values of the other variable, typically x, on the horizontal axis. When we look at the scatter of points in the diagram, it is often possible to visualise a straight line which passes through the midst of the points. We then say that there is a linear relationship between the two variables. We can describe this relationship by formulating a linear regression equation in which one of these variables, x, commonly known as the explanatory, predictor or independent variable may be used to predit the second variable, y, usually called the outcome, response or dependent variable. This regression equation is defined by two parameters, the intercept (the value of y when x is 0) and the slope or gradient, commonly called the regression coefficient, representing the mean change in y for a unit change in x. Because we usually have sample data, we can only estimate the intercept and slope, and so each should be accompanied by some measure of precision, preferably its associated confidence interval. In addition, we often attach a p-value to the slope. This is derived when we test the null hypothesis that the true slope of the line is 0; if significant (typically if $p < 0.05$), we conclude that there is a linear relationship between the two variables.

Because statistical software is generally relatively simple to use, it is easy to fall into the trap of performing a regression analysis without understanding the underlying theory and implications resulting from an inappropriate analysis. It is essential to check the assumptions underlying the regression model to ensure that the results are valid. This is most easily achieved by using the residuals; for a specific individual, a residual is, for a given value of the explanatory variable, x, the difference between the observed value of the dependent variable, y, and its value predicted by the regression equation (i.e. the value of y on the line for that value of x). If the regression assumptions are satisfied, these residuals will be Normally distributed, and a random scatter of points will be obtained both when they are plotted against values of the explanatory variable, indicating that the linearity assumption is satisfied, and also when they are plotted against their corresponding predicted values, indicating that the variability of the observations is constant along the length of the fitted line. If the assumptions are not satisfied, we may be able to find an appropriate transformation of x or y; we then repeat the whole process (i.e. determine the line and check its assumptions) using the transformed data.

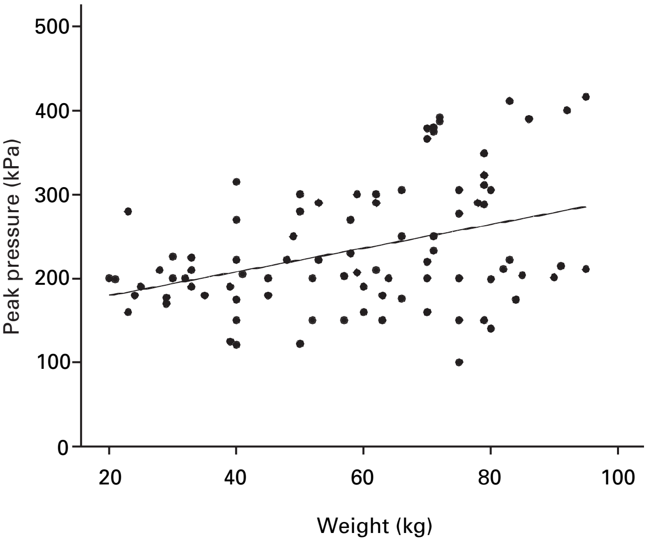

Fig. 6

Scatter diagram showing the linear relationship between peak pressure of the great toe in the left foot and weight in 90 subjects.

Linear correlation. A measure of the linear relationship between the variables is provided by the **Pearson correlation coefficient** generally called simply the correlation coefficient and denoted in the sample by r. It is a dimensionless quantity which has values ranging from -1 to +1. If it is

Linear regression and correlation: example To demonstrate the importance of the toes during walking, Hughes, Clark and Klenerman[10] used a dynamic pedobarograph to examine the weight-bearing function of the foot in a large number of people without foot problems. The subjects were aged between five and 78 years with an equal number of males and females in each five-year age group until aged 30 years and in each ten-year age group after that. Figure 6 is a scatter diagram showing the relationship between the weight (kg) of a subject and the peak pressure (kPa) under his/her great toe (which, of all the toes, bears the greatest peak pressure) in the left foot (these are hypothetical data from 90 subjects based on the results of Hughes et al[10]). After confirming that the underlying assumptions are satisfied, a linear regression analysis of these data gives an estimated linear regression line of:

$$\text{Peak pressure} = 151.5 + 1.41 \times \text{Weight}.$$

The regression coefficient or slope of the line, estimated by 1.41 (95% CI 0.67 to 2.15) kPa per kg of weight, implies that, on average, as a subject's weight increases by 1 kg, the peak pressure under the great toe in the left foot increases by about 1.4 kPa; the slope is significantly different from 0 ($p < 0.001$) so that the subject's weight is an important predictor of the peak pressure under the great toe. In addition, the Pearson correlation coefficient between the peak pressure under the great toe in the left foot and the subject's weight in these data is 0.38 (95% CI 0.18 to 0.54), a positive value indicating that as a subject's weight increases there is a tendency for the peak pressure under the great toe of the left foot to increase as well. When we square the value of the correlation coefficient, we see that only about 14% of the variation in the great toe's peak pressure can be explained by its linear relationship with weight. The remaining 86% is unexplained variation, suggesting that the linear regression line is a poor fit to the data, even though the regression coefficient is significantly different from 0.

It may be of interest to note that Hughes et al[10] found that the toes are in contact for about three-quarters of the walking cycle and exert pressures similar to those from the metatarsal heads. They concluded that the toes play an important part in increasing the weight-bearing area during walking and suggested that, consequently, every effort should be made to preserve their function.

equal to either of its extreme values, then there is a perfect linear relationship between the two variables, and the regression line will have every point lying on it. This line will slope upwards if $r = +1$, with one variable increasing in value as the other variable increases, and the line will slope downwards if $r = -1$, with one variable decreasing in value as the other variable increases. In practice, however, it is more usual to find that there is some degree of scatter about the linear regression line and this will be reflected to some extent by the magnitude of the correlation coefficient, its value being closer to one of the extremes when there is less scatter about the line. If the correlation coefficient is 0, then there is no linear relationship between the variables. We can test the null hypothesis that the true correlation coefficient is 0; the resulting p-value will be identical to that obtained from the test of the hypothesis that the true slope of the regression line is 0 because of the mathematical relationship between the slope and the intercept.

The significance of the correlation coefficient is highly dependent on the number of observations in the sample: the greater the sample size, the smaller the p-value associated with a correlation coefficient of a given magnitude. Therefore, to fully assess the linear association between two variables, we should calculate r^2, as well as r and its p-value. The square of the correlation coefficient describes the proportion of the total variation in one variable (say, y) which can be explained by its linear relationship with the other variable (x). Thus if we estimate the correlation coefficient to be 0.90 and we have 10 pairs of observations so that $p = 0.0004$, we can say that approximately 81% of the variation in y can be attributed to its linear relationship with x. This suggests that there is not too much scatter about the best fitting line and the line is a good fit. However, if we estimate the correlation coefficient to be 0.50 and our sample size is 20 so that $p = 0.03$, then only 25% of the variation in y can be attributed to its linear relationship with x. In this latter situation, in spite of having a correlation coefficient which is significantly different from 0, 75% of the variation in y is unexplained by its linear relationship with x, indicating that there is substantial scatter of points around the best fitting line, and the line is a poor fit.

If one or both of the variables are measured on an ordinal scale, or if we are concerned about the Normality of the numerical variable(s), we can calculate the non-parametric **Spearman correlation coefficient** which also takes values ranging from -1 to +1. Its interpretation is similar to that of the Pearson correlation coefficient although it provides an assessment of association rather than linear association and its square is not a useful measure of goodness-of-fit.

More than two variables

Multivariable linear regression. Multivariable linear regression, also called **multiple linear regression**, is an extension of univariable linear regression. Both univariable and multivariable linear regression models have a single numerical dependent variable, y, but the multivariable model has a number (k, say) of explanatory variables, $x_1, x_2, x_3, \ldots.., x_k$, which are linearly related to y, instead of just one explanatory variable, x. However, if we have too many explanatory variables (also called **covariates**) in the model, we cannot draw useful conclusions from it. As a rough guide, there should be at least ten times as many sets of observations (e.g. individuals) as explanatory variables. So, for example, if our sample comprises 100 individuals, we should have no more than ten explanatory variables in the model.

As an illustration, we might extend the simple linear regression example which investigated the relationship between the peak pressure under the big toe (i.e. the dependent variable) and a subject's weight (i.e. the single explanatory variable in the univariable regression) by including some additional explanatory variables (e.g. the subject's age and gender) in the model. The regression coefficient associated with a particular explanatory variable, x, say, represents the amount by which y changes on average as we increase x_1 by one unit, after adjusting for all the other covariates in the equation. Thus, by considering the magnitude and sign of the estimated regression coefficients, we can determine the extent to which one or more of the explanatory variables may be linearly related to the dependent variable, after adjusting for other covariates in the equation. In addition, the p-values associated with the regression coefficients enable us to identify explanatory variables that are significantly associated with the dependent variable in order to promote understanding. Sometimes we also use the equation to predict the value of the dependent variable from the explanatory variables. We should always check the assumptions underlying the multivariable regression model. We do this by determining the residuals, and, as with a univariable linear regression analysis, plot them to check for Normality, constant variance and linearity associated with each of the explanatory variables.

Other multivariable regression models. The dependent variable in a multiple linear regression equation is a numerical variable. If we have a binary response variable, such as success or failure, then multiple regression analysis is inappropriate. Instead, we can perform a **linear logistic regression analysis** which relies on a model in which the explanatory variables are related in a linear fashion to a particular transformation, called the **logistic transformation**, of the probability of one of the two possible outcomes (say, a success). Back transformation of an estimated regression coefficient in the model provides an estimate of the odds ratio for a unit increase in the predictor. So, for example, if one of the explanatory variables represents treatment (coded as 1 for the new treatment and 0 for the control treatment), the odds ratio for this variable represents the odds of success on the new treatment compared with (i.e. divided by) the odds of success on the control treatment, after adjusting for the other explanatory variables in the model. Computer output for a logistic regression analysis will usually include, for each explanatory variable, an estimate of the odds ratio, its associated confidence interval and a p-value resulting from the test of the hypothesis that the odds ratio is 1.

Logistic regression: example There is a high risk of deep-vein thrombosis (DVT) when patients are immobilised following trauma. Fuchs et al[11] randomised 227 patients who had suffered trauma to the spine, pelvis, tibia or ankle to receive treatment either with the Arthroflow device (Ormed, Freiburg, Germany) and low-molecular-weight heparin (LMWH) or only LMWH. The Arthroflow device allows continuous passive movement of the ankle joint with maximal extension and plantar flexion at a frequency of 30 excursions per minute, giving compression of the crural compartments. Patients with trauma to the ankle did not receive the Arthroflow device. Those patients who showed evidence of DVT when assessed weekly, by venous occlusion plethysmography, compression ultrasonography and continuous wave Doppler, underwent venography for confirmation. Logistic regression analysis, in which the presence or absence of DVT defined the outcome, was used to identify significant risk factors for DVT. The authors found that having an operation (OR 4.1: 95% CI 1.1 to 15.1) significantly increased the odds of a DVT when the remaining covariates in the model were taken into account. Thus the odds of a DVT was over four times greater in those patients who had undergone an operation than in those who had not, after adjusting for the remaining explanatory variables in the model. Other factors which were marginally significant were age > 40 years (OR 2.8: 95% CI 1.0 to 7.8) and obesity (OR 2.2: 95% CI 1.0 to 5.1). A crucial finding was that the odds of a DVT was significantly reduced by approximately 89% if the Arthroflow was used with LMWH compared with when only LMWH was administered (OR 0.11: 95% CI 0.04 to 0.33).

We should not use logistic regression analysis when we have a binary outcome variable (success/failure) and individuals are followed for different lengths of time, because such an analysis does not take time into account. We often use **Poisson regression analysis** in these circumstances, provided we can assume that the rate of the event of interest (e.g. success) is constant over the time period. The explanatory variables in the Poisson model are related in a linear fashion to the logarithm of the rate of success. Thus, if we take the antilog of an estimated regression coefficient associated with a particular explanatory variable, we obtain an estimate of the relative rate for a unit increase in that variable. The interpretation of the coefficients (replacing 'odds' by 'rate') and the computer output arising from a Poission regression analysis are comparable to those of a logistic regression analysis.

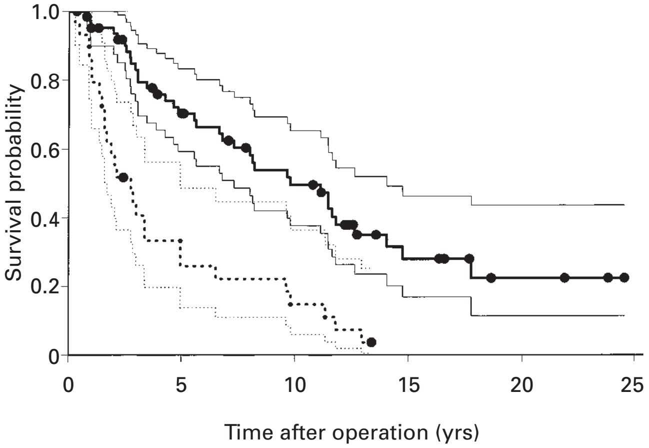

Fig. 7

Kaplan-Meier survivorship curves for flexion osteotomy (heavy solid line, n = 63) and rotational osteotomy (heavy dotted line, n = 29) in patients with avascular necrosis of the hip: faint lines represent 95% confidence intervals for each curve. Censored data (patients lost to follow-up or remaining unrevised at the end of follow-up) are indicated by solid black circles; in the flexion osteotomy group, 12 patients were lost to follow-up and 15 remained unrevised; after rotational osteotomy, one was lost to follow-up and one remained unrevised.

We generally use **survival analysis** when, in addition to having a binary endpoint of interest (e.g. dead/alive; failure/no failure; recurrence/no recurrence) and individuals being followed for varying lengths of time after some critical event (e.g. an operation, start of treatment, time of diagnosis), we have censored data (e.g. an individual leaves the study before the end so that his/her outcome is indeterminate or he/she has not experienced the outcome when the study period is over). In particular, if we wish to investigate the effect on survival of several explanatory variables at the same time, we can perform a **Cox proportional hazards regression analysis** to assess the independent effect of each variable on the hazard ratio, also called the relative hazard. The **hazard** in survival analysis represents the instantaneous risk or probability of reaching the endpoint (e.g. death) at a particular time, and the hazard ratio is interpreted in a similar way to the odds ratio in logistic regression and the relative rate in Poisson regression, replacing 'odds' or 'rate' by 'hazard', as appropriate. As the name suggests, we assume proportional hazards in a Cox regression analysis. This implies that the hazard ratio for a given explanatory variable is constant at all times. So, if a variable affects the hazard in a particular way (e.g. the hazard is twice that for males as it is for females), the hazard ratio for this variable is the same at every time. We often use the simpler **Kaplan-Meier survival analysis**[12] when investigating the effect of a single variable on survival, using the estimated median survival time or cumulative probability of survival at a particular time (e.g. the five-year survival probability) to summarise our findings. We should not calculate the mean survival time as it does not take into account the censored data or the fact that individuals are followed for varying lengths of time. A survival curve (Fig. 7) is a useful graphical summary of the results; this is drawn in a diagram in which the horizontal axis represents the time that has elapsed from the critical event and the vertical axis usually represents the cumulative probability of survival (i.e. of not reaching the endpoint). We

Survival analysis: example Schneider et al[13] compared different types of intertrochanteric osteotomy in the treatment of avascular necrosis of the hip and evaluated their performance in the light of improving outcome after total hip replacement (THR). Flexion osteotomy was undertaken in 63 hips and rotational osteotomy in 29 hips; the risk factors such as alcohol abuse, hyperlipidaemia, smoking, obesity, steroid medication and age, were similarly distributed in two groups of patients. Both groups were predominantly stage III with depression of the articular surface, but without pronounced narrowing of the joint space. Figure 7 shows the Kaplan-Meier estimated survivorship curves, with revision for any reason taken as the end-point, for the two groups. The survival probability was found to be significantly higher for hips having flexion osteotomy than for those having rotational osteotomy ($p < 0.001$), using the log-rank test to compare the survival curves. For the flexion and rotational osteotomy groups, respectively, the five-year survival probabilities were 0.70 (95% CI 0.59 to 0.83) and 0.26 (95% CI 0.14 to 0.49), and the ten-year survival probabilities were 0.50 (95% CI 0.38 to 0.65) and 0.15 (95% CI 0.06 to 0.36). Although flexion osteotomy was found to be superior to rotational osteotomy, neither procedure gave the patients a functional solution.

often use the non-parametric **log-rank test** to compare two or more survival curves in a Kaplan-Meier survival analysis.

Other common forms of analysis

Diagnostic tests

The result of a simple **diagnostic test** indicates whether an individual does or does not have the disease or condition being investigated. It may be used to supplement a clinical examination in order to diagnose or exclude a particular disorder in a patient. Alternatively, it may be used as a **screening** device to ascertain whether or not an apparently healthy individual is likely to have a particular condition.

It is unlikely that every individual undergoing a diagnostic test will be correctly identified as diseased or disease-free, and it is therefore important to be able to quantify the extent to which the test classifies individuals appropriately. To this end, it is helpful to know the estimated values of the **sensitivity**, i.e. the percentage of the diseased individuals who have a positive test result, and the **specificity**, i.e. the percentage of those who are disease-free who have a negative test result, with their associated confidence intervals. The sensitivity and specificity can only be evaluated in a study in which the true disease status of the individuals is known. They do not help the clinician decide whether his/her particular patient has the condition in question. The clinician requires the **positive predictive value** of the test (the percentage of those with a positive test result who actually have the condition) and/or its **negative predictive value** (the percentage of those with a negative test result who do not have the disease), as these values indicate how likely it is that the patient does or does not have the condition when there is a positive or negative test result.

Diagnostic test: example The article by Bhandari et al[14] provides a useful guide to the use of diagnostic tests in the context of surgical patients. In it they provide an example in which the **C-reactive protein (CRP)** test is used to give an indication of whether or not a particular patient has an infection of the hip joint following THR. It is important for the surgeon to be able to assess the validity of the CRP test and decide how likely it is that the patient has or does not have an infection if the result of the test is 'positive' or 'negative', respectively. Table I (derived from Spangehl et al[15]) shows the results for 142 patients who underwent measurement of the CRP level; a value in excess of 10 mg/l was taken as an indication of infection.

Since the sensitivity of the CRP test is 96%, it is very good at identifying patients as having an infection if they truly are infected (i.e. true positives); the specificity is also high at 92%, so the test is also good at identifying true negatives. However, the positive predictive value of the test indicates that if the patient tests positive, the chance of actually having an infection is only 74%, and further testing (e.g. hip aspiration) would be indicated to resolve the uncertainty associated with the test result. Alternatively, if the patient tests negative with, for example, a CRP test of 8 mg/l, the chance of not having an infection is 99% (this is the negative predictive value) so that the surgeon would be unlikely to conduct further tests for infection. Using a Bayesian approach,[16] additional information, such as whether the patient has hip pain or overt signs of infection, can be used to modify the post-test probability of infection.

One way of evaluating the positive or negative predictive value of a test is to use a **Bayesian approach**[16] in which the

Table I. Table of frequencies for the evaluation of a diagnostic test (derived from Spangehl et al[15])

	Peri-prosthetic infection		
C-reactive protein test	**Yes**	**No**	**Total**
Positive (> 10 mg/l)	25	9	34
Negative (≤ 10 mg/l)	1	107	108
Total	26	118	142

Sensitivity = 100 x 25/26 = 96% (95% CI 89% to 100%)
Specificity = 100 x 107/116 = 92% (95% CI 87% to 97%)
Prevalence = 100 x 26/142 = 18% (95% CI 12% to 24%)
Positive predictive value = 100 x 25/34 = 74% (95% CI 59% to 88%)
Negative predictive value = 100 x 107/108 = 99% (95% CI 97% to 100%)

investigator does not use the information provided by the test (i.e. the test result and the sensitivity and specificity of the test) in isolation, but uses it to modify the previous view held before the result of the test was known. Expressing the Bayesian approach in general terms and using statistical terminology, the prior probability of disease is updated by combining it with the current evidence, expressed in terms of a likelihood ratio, to provide an estimate of the posterior probability. In the absence of additional knowledge, the prior probability in the context of a diagnostic test is often taken as the prevalence of the disease in the population of interest. The positive and negative predictive value of a test depend very strongly on the disease prevalence so that, for example, the positive predictive value will be larger if the prevalence is greater. Formulae and diagrams are available so that the clinician can use the Bayesian approach to evaluate the positive predictive value if the patient tests positive for the disease, or the negative predictive value if the patient has a negative test result, provided the sensitivity and specificity of the test are known.

Reliability studies

Measurements exhibit a certain amount of variation when taken by different observers or at different times or using different equipment. Reliability studies are concerned with quantifying this variation or 'measurement error'. The easiest way to proceed is to take pairs of measurements. For example, in a repeatability study, the same observer may repeat the measurement using the same equipment within a short period of time on each of 20 patients to assess intra-observer variation, or in a reproducibility study, two observers may take one measurement each on all 20 patients to assess inter-observer variation. We use a different approach to the analysis according to whether the data are categorical or numerical.

Categorical data. If the data are categorical, for example when two observers each categorise every patient as belonging to one of a number of categories (e.g. disease stage 1, 2 or 3), we can create a two-way contingency table of frequencies. The rows and columns of this table both show the different classifications, but the rows relate to the results of one observer and the columns relate to the results of the second observer. The entries in the table are frequencies. There will be no inter-observer variation if the two observers agree on the classification of every patient, and so all the entries in the table in these circumstances will be along the diagonal. The **kappa statistic (κ)** provides a measure of the proportion of times that the observers agree, modified to take into account the agreement that would occur by chance alone. Its maximum value of 1 indicates perfect agreement between the two observers and its minimum value of 0 indicates that the agreement between the observers is no better than would be expected by chance. Values in between these extremes are generally classified as very good if $0.81 \le \kappa \le 1.00$, good if $0.61 \le \kappa \le 0.80$, moderate if $0.41 \le \kappa \le 0.60$; fair if $0.21 \le \kappa \le 0.40$ and poor if $\kappa < 0.20$.[17] However, kappa depends to some extent on the number of categories. It is easier to get agreement between observers when there are fewer categories in which to classify the patients. A weighted kappa can be evaluated which takes into account the extent to which the observers disagree if the classifications are ordered in some way. Although it is possible to estimate a standard error for kappa and use it to calculate a confidence interval, we do not usually test the hypothesis that kappa is 0 because total lack of agreement is not the real issue in a reliability study.

Kappa statistic: example Chung et al[18] carried out a morphometric analysis of acetabular dysplasia in patients with cerebral palsy requiring hip surgery using three-dimensional CT. In order to assess intra-observer reliability, four orthopaedic surgeons (with 20, five and four years experience and a trainee with two years experience) each examined 54 randomly-ordered 3D reconstructed lateral images of the acetabulum without the femoral head on two occasions. There was an interval of one month between the first and the second examinations. The location of the acetabular defect was classified in each instance as either anterior, global (mixed) or posterior. The kappa statistics assessing intra-observer reliability ranged from 0.81 (very good) for the most experienced surgeon to 0.5 (moderate) for the trainee, with the two surgeons with five and four years experience having kappa values of 0.77 and 0.79 (both good), respectively.

Numerical data. We adopt a different approach when the data are numerical. The most simple procedure is to obtain pairs of observations, either duplicate readings on a number of patients to assess intra-observer agreement, or one reading per patient from each of two observers to assess inter-observer variation. A non-significant result (when $p \ge 0.05$) from a paired *t*-test on these data suggests that there is no evidence of a systematic difference between the pairs of readings; this implies that there is no bias if one set of readings may be taken as the true readings. Then the **Bland-Altman approach**[19] is generally advocated in which we plot the difference between each pair on the vertical axis of a graph against their mean on the horizontal axis (Fig. 8). If we observe a random scatter of points, as opposed to a funnel or cone effect when the differences appear more (or less) disparate as the magnitude of the observations increases, then a single measure of repeatability or reproducibility is sensible. We usually add to this graph the upper and lower limits of agreement. These are obtained by adding and subtracting 1.96 (often approximated to 2) times the SD of the differences to and from the mean of the differences, respectively. The interval between them will contain approximately 95% of the differences. Furthermore, 1.96 times the SD of the differences, the **British Standards repeatability/reproducibility coefficient**, gives an indication of the maximum likely difference that occurs between a pair of readings. This measure should be assessed subjectively, and

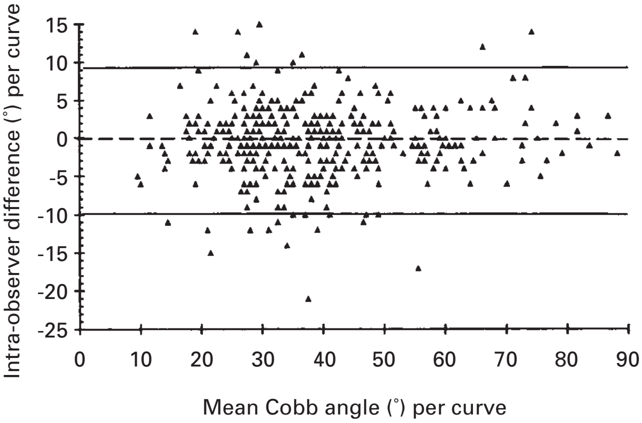

Fig. 8

Bland-Altman[19] diagram depicting the difference between the successive measurements of the Cobb angle plotted against their mean for 67 scoliosis lines for six observers. The dashed line represents the mean value of all differences between the pairs of measurements, and the lines above and below it represent the 95% limits of agreement.

evaluated in the light of the study, using clinical judgment, to determine whether or not the agreement between the pairs of readings is acceptable. Further details of the methods used to measure agreement are described by Streiner and Norman[20] and Lin.[21]

The Bland Altman method: example The variability in measurement of angles in congenital scoliosis is not known, but it has been postulated that it tends to be larger than that in adolescent idiopathic scoliosis due to skeletal immaturity, incomplete ossification and anomalous development of the end-vertebrae. To determine this variability, Loder et al[22] selected radiographs of adequate quality showing 67 scoliotic curves from children with congenital scoliosis. The end-vertebrae were pre-selected to reduce variability. Each of six observers drew endplate lines to determine the Cobb angle for every curve on two separate occasions, at least three weeks apart, using the same goniometer and marker. The investigators found that there was no significant difference between the means of the intra-observer differences among the six observers and, consequently, did not distinguish between the observers when assessing the intra-observer variability. Figure 8 shows the Bland-Altman plot of the difference between the two measurements for an observer plotted against their mean. Since no funnel effect was observed in this diagram, it was reasonable to determine a single measure of intra-observer repeatability, and this was calculated as 9.6°. In addition, the investigators found that the inter-observer reproducibility coefficient was 11.8°, and that the two coefficients were substantially larger than their counterparts obtained in a comparable study of children with adolescent idiopathic scoliosis.

Common errors in statistical analysis

Errors in the use of statistics may occur at all stages of an investigation, in its design, analysis, presentation and interpretation. An error at any of these stages may invalidate any conclusions drawn. Some of the more common errors are summarised below. Further details may be obtained from papers by Altman,[23,24] Buyse et al,[25] Matthews and Altman,[26] Pocock et al,[27] Seigel[28] and Tu et al.[29]

Design

- Failure to follow the basic principles of design, because of an inappropriate or no control group, no randomisation in an experimental study, and/or no blinding.
- No justification, using power analysis, of the sample size used.
- Inadequate response rate.

Analysis

- No checking of underlying assumptions (e.g. of Normality of data, assumptions underlying regression analysis).
- Inappropriate use of the arithmetic mean to summarise skewed data.
- Failure to use (or report) intention-to-treat analysis in a randomised study when those individuals who violate the protocol (e.g. they switch or stop taking treatment) should be analysed as if they were still in the treatment groups to which they were randomised to prevent bias.
- Failure to recognise dependencies in data such as when measurements are taken on two limbs of a patient and the results are treated as independent, as if they came from different patients.
- Failure to use the correct unit of analysis.
- Comparison of p-values. A p-value depends on both the magnitude of the effect (e.g. the difference in two means) and on the precision of its estimate (i.e. its standard error), so differences in p-values may arise from differences in standard errors.
- Inappropriate methods for assessing repeatability.
- Inappropriate analyses associated with the multiplicity of data; e.g. with repeated measures when dependencies in the data are ignored, multiple testing when no adjustments are made to the p-values so that spuriously significant results arise, and with analyses of subgroups which may result in ‘data dredging’, multiple testing and potentially small sample sizes with low power.
- Inappropriate analysis of variance, of which there are many different types.
- Performing statistical tests and presenting p-values on baseline characteristics to compare groups in a randomised study. This is inappropriate because all baseline differences are due to chance if randomisation has been used.
- Distorting associations by ignoring the effect of mathematical coupling, when one variable contains the whole or part of the other, e.g. when relating changes from baseline to baseline values.

Presentation

- Not specifying the primary aim of the study.
- Not providing an adequate description of the randomisation process in an experimental study.
- Not providing an adequate description of the choice of subjects included in a study (e.g. were they consecutive patients within a certain time interval?).
- Not reporting exact p-values but simply stating, for example, $p < 0.05$ or NS.
- Not providing measures of precision and, in particular, too little use of confidence intervals.
- Not providing a measure of the effect of interest, such as the difference in means when comparing two means.
- Giving too many significant figures, particularly for p-values; they should be reported so the number of digits is commensurate with scientific relevance.
- Poor diagrams with inadequate labelling, use of scale changes or breaks in graphs, and the use of bar charts for continuous data.
- Vague descriptions of statistical methods (e.g. of particular types of analysis of variance).
- Omitting key information, such as the number of subjects per group.
- Describing data as non-parametric. It is the statistical methods which make no assumptions about the distribution of the data that are non-parametric.
- Using the ± sign without explaining what the figures are before and after it. It is better to avoid using the ± sign since there is a tendency for the reader to believe, incorrectly, that the interval so created is a 95% confidence interval for the parameter of interest, whereas it is approximately a 67% confidence interval if the second figure is a standard error. Instead, the 95% confidence interval for the parameter should be provided.

Interpretation

- The conclusions go beyond those which the data warrants.
- The conclusions are not a reasonable reflection of the data presented.
- Failure to differentiate between statistical and clinical significance. The use of confidence intervals goes some way to alleviate this problem.
- Placing too much emphasis on a non-significant result which may have arisen because of low power. For example, a non-significant result from a two-sample *t*-test does not imply that the two means are equal, only that there is no evidence to show that they are different. In this context, it is useful to remember that 'absence of evidence is not evidence of absence'.[30]

The main reason for the plethora of statistical errors in publication is that most statistical analyses are performed by non-statisticians who have an inadequate understanding of statistical theory. The ease of use of statistical software has contributed to this problem, as has peer reviewing in some journals by non-statisticians whose knowledge of statistics is limited (Altman[23]). The solution is not straightforward, but the provision of guidelines for reporting trials, such as those in the CONSORT statement's checklist and in Altman et al,[31] goes some way to alleviate the problem.

Supplementary Material

The CONSORT statement and a glossary of terms are available as supplementary material at www.jbjs.org.uk

References

1. **Altman DG.** *Practical statistics for medical research.* London: Chapman & Hall, 1991.
2. **Bland M.** *An introduction to medical statistics.* Third ed. Oxford: Oxford University Press, 2000.
3. **Kirkwood B, Sterne J.** *Medical statistics.* Second ed. Oxford: Blackwell Science, 2003.
4. **Petrie A, Sabin C.** *Medical statistics at a glance.* Second ed. Oxford: Blackwell Publishing Ltd, 2000.
5. **Straus S, Richardson WS, Glasziou P, Haynes RB.** *Evidence-based medicine: how to practice and teach EBM.* Third ed. London: Churchill Livingstone, 2005.
6. **Benoni G, Fredin H.** Fibrinolytic inhibition with tranexamic acid reduces blood loss and blood transfusion after knee arthroplasty: a prospective, randomised, double-blind study of 86 patients. *J Bone Joint Surg [Br]* 1996;78-B:434-40.
7. **Machin D, Campbell M, Fayers P, Pinol A.** *Sample size tables for clinical studies.* Second ed. Oxford: Blackwell Science,1997.
8. **Altman DG.** How large a sample? In: Gore SM, Altman DG, eds. *Statistics in practice.* London: British Medical Association, 1982.
9. **Lehr R.** Sixteen S-squared over D-squared: a relation for crude sample size estimates. *Stat Med* 1992;11:1099-102.
10. **Hughes J, Clark P, Klenerman L.** The importance of the toes in walking. *J Bone Joint Surg [Br]* 1990;72-B:245-51.
11. **Fuchs S, Heyse T, Rudofsky G, Gosheger G, Chylarecki C.** Continuous passive motion in the prevention of deep-vein thrombosis: a randomised comparison in trauma patients. *J Bone Joint Surg [Br]* 2005;87-B:1117-22.
12. **Kaplan EL, Meier P.** Non parametric estimation from incomplete observations. *J Am Statist Ass* 1958;5:457-81.
13. **Schneider W, Aigner N, Pinggera O, Knahr K.** Intertrochanteric osteotomy for avascular necrosis of the head of the femur: survival probability of two different methods. *J Bone Joint Surg [Br]* 2002;84-B:817-24.
14. **Bhandari M, Montori VM, Swiotkowski MF, Guyatt GH.** User's guide to the surgical literature: how to use an article about a diagnostic test. *J Bone Joint Surg [Am]* 2003;85-A:1133-40.
15. **Spangehl MJ, Masri B, O'Connell JX, Duncan CP.** Prospective analysis of preoperative and intraoperative investigations for the diagnosis of infection at the sites of two hundred and two revision total hip arthroplasties. *J Bone Joint Surg [Am]* 1999;81-A:672-83.
16. **Bayes T.** An essay towards solving a problem in the doctrine of chances, 1763. *MD Computing* 1991:8;157-71.
17. **Landis JR, Koch GG.** The measurement of observer agreement for categorical data. *Biometrics* 1977;33:159-74.
18. **Chung CY, Park MS, Choi IH, et al.** Morphometric analysis of acetabular dysplasia in cerebral palsy. *J Bone Joint Surg [Br]* 2006;88-B:243-7.
19. **Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307-10.
20. **Streiner DR, Norman GL.** *Health measurement scales: a practical guide to their development and use.* Third ed. Oxford: Oxford University Press, 2003.
21. **Lin LK.** A concordance correlation coefficient to evaluate reproducibility. *Biometrics* 1989;45:255-68.
22. **Loder RT, Urquhart A, Steen H, et al.** Variability in Cobb angle measurements in children with congenital scoliosis. *J Bone Joint Surg [Br]* 1995;77-B:768-70.
23. **Altman DG.** Statistical reviewing for medical journals. *Stat Med* 1998;17:2661-74.
24. **Altman DG.** Statistics in medical journals: some recent trends. *Stat Med* 2000;19:3275-89.
25. **Buyse M, George SL, Evans S, et al.** The role of biostatistics in the prevention, detection and treatment of fraud in clinical trials. *Stat Med* 1999;18:3435-51.
26. **Matthews JNS, Altman DG.** Interaction 2: compare effect sizes not p-values. *BMJ* 1996;313:808.
27. **Pocock SJ, Assmann SF, Enos LE, Kasten LE.** Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917-30.
28. **Seigel D.** Clinical trials, epidemiology, and public confidence. *Stat Med* 2003;22:3419-25.
29. **Tu YK, Maddick IH, Griffiths GS, Gilthorpe MS.** Mathematical coupling can undermine the statistical assessment of clinical research: illustration from the treatment of guided tissue regeneration. *J Dent* 2004;32:133-42.
30. **Altman DG, Bland JM.** Absence of evidence is not evidence of absence. *BMJ* 1995;311:485.
31. **Altman DG, Machin D, Bryant TN, Gardner MJ, eds.** *Statistics with confidence.* Second ed. London: BMJ Books, 2000.