Statistics in medicine

Ian Kestin

Abstract

This article covers the basic principles of statistics in medicine. Topics covered include types of data, descriptive statistics (mean, median, mode, percentiles), the normal distribution, confidence intervals and the standard error of the mean, hypothesis testing and the choice of statistical tests, type I and II errors, contingency tables, correlation and regression, and meta-analysis.

Keywords Clinical trials; correlation; hypothesis tests; normal distribution; regression; statistics

Royal College of Anaesthetists CPD matrix: 3J03

Introduction

All clinicians should understand the correct use of research data, and that statistics are the tools used to describe and analyse numbers. The complete data set from a study may comprise many thousands of observations, and it is not practical to give the full results in a published paper. Descriptive statistics are used to summarize this numerical information. We also use statistics to infer properties about a wider population of subjects beyond those actually studied, and this is called inferential statistics. Uncertainty, probability and error are crucial concepts for understanding the limitations of statistics.

Types of data

The types of data obtained in a research project determine the methods used to describe and analyse the data. There are three main types.

- **Categorical (nominal) data:** each of the subjects in the study is allocated to one of two or more mutually exclusive categories, for example sex (male/female), blood group (A, B, AB, O) or social class. The categories have no ranking or numerical relationship to each other.
- **Ordinal data (ordered categories or ranked data):** each of the subjects in the study is allocated to one of several mutually exclusive categories, and these categories have an intrinsic ranking or ordering. Examples would be grades of oedema (mild/moderate/severe), or American Association of Anaesthetists (ASA) scores (1, 2, 3, 4 or 5). The categories may be numbered, but this numbering only defines the ordering of the categories, and does not ‘scale’ the relative magnitude of one category to another. For example, head-injured patients are allocated using the Glasgow Coma Score (GCS) to one of thirteen possible categories denoted by a whole number between 3 and 15. A patient with a GCS of 4 is worse than a patient with a GCS of 8, but is not ‘twice as bad’, whereas a patient weighing 80 kg is exactly twice as heavy as one weighing 40 kg. Misusing ordinal data by treating the numbers as if real measurements had been made is a common mistake.
- **Numerical data:** this type of data describes actual numerical properties of the subjects. The measurements can be either discrete or continuously variable. Discrete numerical data can only take certain values, usually integers (e.g. number of children, hospital deaths per year); continuously variable data can theoretically take any numerical value, but usually occur within a certain range (e.g. heart rate, weight).

Descriptive statistics

Descriptive statistics are required to summarize large data sets. Categorical data are easily described by histograms or pie charts; a visual illustration of the data clearly shows the frequency of the categories (Figure 1).

Two essential properties describe ordinal or numerical data:

- the central location – where the bulk of the observations lie
- the variability – how closely the observations are clustered about the central location.

The central location of a series of observations is usually described using the mean, median or mode (Table 1).

Misuse of the mean is a common error, which properly should only be used with continuously variable numerical data. For ordinal data the median or mode must be used (e.g. it is quite wrong to quote a mean GCS of 7.5).

The variability can be described by the range, percentiles or the standard deviation. The range gives the maximum and minimum values of the observations and is useful if there is some particular interest in the maximum or minimum response (e.g. the lowest respiratory rate recorded would be of clinical importance in patients given opiates). However, the range can give a misleading impression of the variability if there are single extreme results in the data.

A percentile is that observation which is greater than the appropriate percentage of all the observations in the data set, so the 10th percentile is the observation that is greater than 10% of all the observations; the median is the 50th percentile; and the 90th percentile is the observation that is greater than 90% of the observations. Commonly used percentiles are the interquartile range (the 25th to 75th percentile) and the 2.5th to 97.5th percentile (containing 95% of the observations). Percentiles can be used for any type of data, but the standard deviation is only applicable to data that are continuously variable and normally distributed (see later). A common graphical way of summarizing information is the ‘box and whisker’ plot (Figure 2).

Frequency distribution curves

A graph showing the probability of obtaining any particular observation is called a frequency distribution (Figure 3).

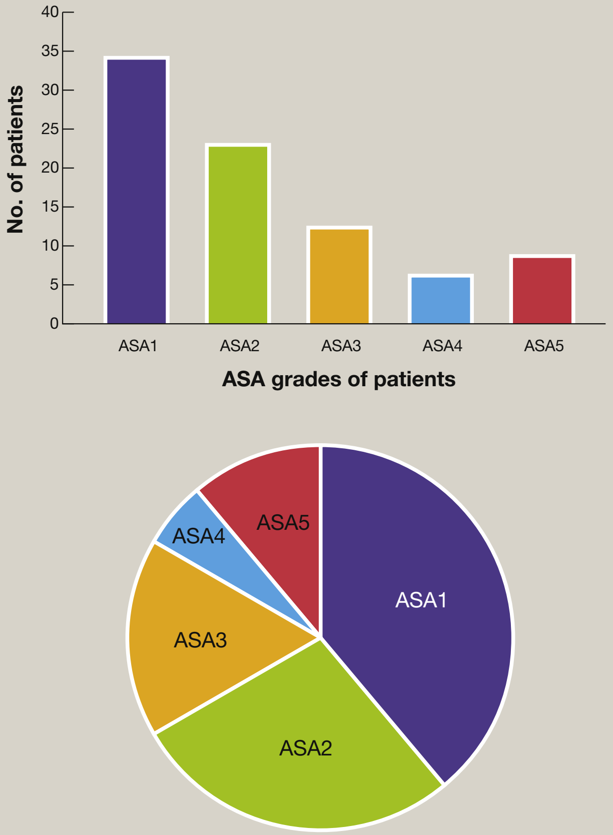
The normal distribution is a specific frequency distribution pattern that is common in biological data for which many statistical tests have been designed (e.g. *t*-test, analysis of variance).

The central location can be described by the mean (which is the same as the mode and median), and the variability is

***Ian Kestin** FRCA is Consultant Anaesthetist at the Western Infirmary, Glasgow, UK. Conflicts of interest: none declared.*

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005

A pie chart and histogram, two ways of illustrating the frequency distribution of categorical or ordinal data



ASA, American Society of Anesthesiologists.

Figure 1

Common measures of central location

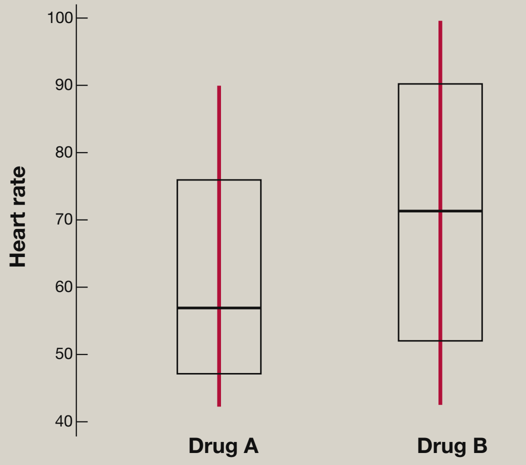
Measure of central location	Type of data	Definition
Mean	Continuously variable	Sum of all observations/number of observations
Median	Ordinal and numerical	The observation with half the observations above and half below, i.e. 50th percentile
Mode	Ordinal and numerical	The most commonly occurring observation

Table 1

described by the standard deviation. Multiples of the standard deviation about the mean always contain the same proportion of the observations (Figure 4).

Not all symmetrical frequency distributions are normal (Figure 5).

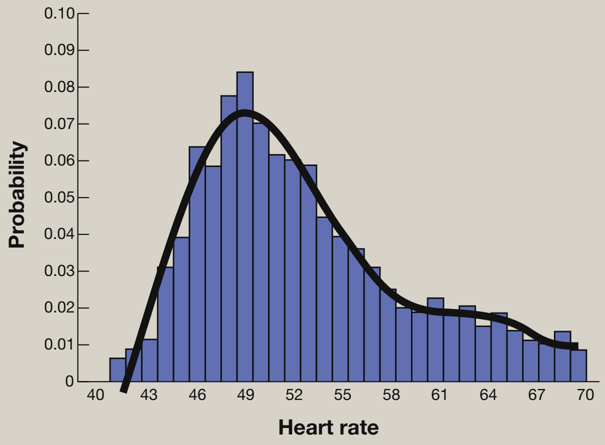
Box and whisker plot of heart rates after two different drugs



The horizontal line shows the median, the ‘box’ shows percentiles, commonly the 2.5th to 97.5th, and the vertical line shows the range of the sample data.

Figure 2

Frequency distribution of heart rates



The probability of any given heart rate is shown by the histogram. The histogram can be replaced by a continuous curve as the intervals on the *x*-axis become smaller.

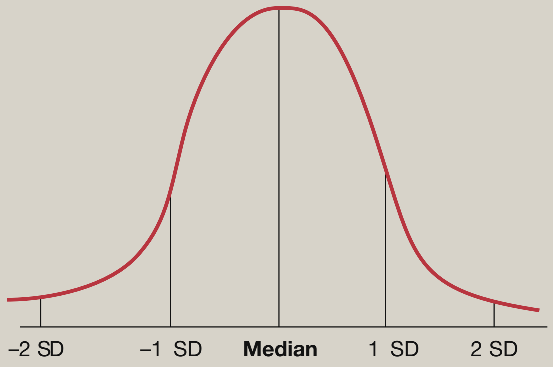
Figure 3

Skewed distributions are a common pattern in biological data, when the frequency distribution curve is not symmetrical (Figure 6).

The frequency of hospital stay after an operation is commonly skewed (see Figure 6); most patients have similar lengths of stay, but some have complications and stay much longer. This is an example of positively skewed data; negatively skewed data is the reverse pattern and is less common. The mean, median and mode have different values if the data are skewed. If the single

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005

Characteristics of a normal distribution



Calculation of sample parameters

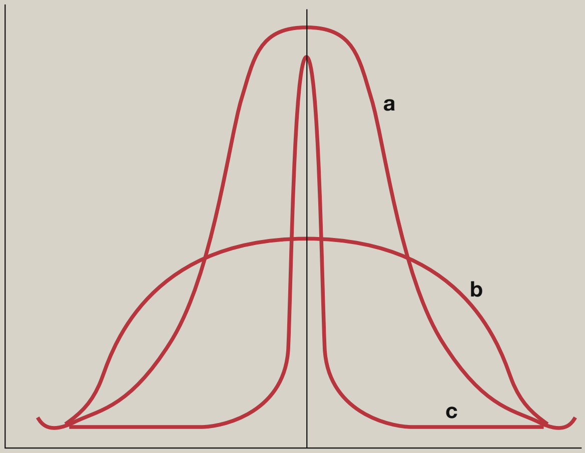
Sample mean = $\sum x / n$

Sample standard deviation (SD) = $\sqrt{\dfrac{\sum (x-\bar{x})^2}{n-1}}$

Mean ± 1 SD	67% of observations
Mean ± 2 SD	95% of observations
Mean ± 3 SD	99% of observations

Figure 4

Kurtosis

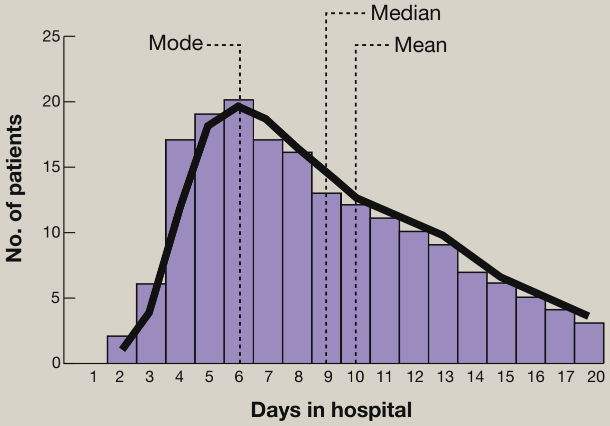


Not all symmetrical frequency distributions are normal; the curve may be flatter or more peaked than the normal distribution. **a** Normal distribution. **b** and **c** Symmetrical distributions that have a broader (platykurtic) or more narrow (leptokurtic) distribution than the normal distribution. A bimodal distribution is an extreme example of a platykurtic distribution.

Figure 5

largest observation in the sample is increased, the mean and range will increase, but the median and mode are unchanged; the median or mode are generally better indicators of the central location of a skewed distribution. If the standard deviation of a sample is more than half the mean, then the data are probably skewed.

Length of stay in hospital after surgery, an example of a positively skewed distribution



The mode is 6 days, the median is 9 days and the mean is 10 days

Figure 6

Skewed data can often be mathematically transformed to conform to a normal distribution. Taking logarithms of a sample that is positively skewed will usually produce a data set that is approximately normally distributed, and then techniques designed for the normal distribution can be used on the transformed data. The alternative is to use statistical tests designed for any data. If the data are transformed, care must be taken when doing any reverse transformations back to the original units — confidence intervals can be reverse transformed, but not the standard deviation (the confidence limits will not be symmetrical about the mean in skewed data). There are methods of testing whether data conform to a normal distribution, for example the Shapiro—Wilk W test, and these ought to be done before using statistics designed for the normal distribution. Statistical techniques that use assumptions about the underlying distribution of the data (nearly always assuming a normal distribution) are called parametric statistics. Techniques to describe and analyse data that make no assumptions about the underlying distribution of the sample or population data are called non-parametric statistics, and can be used on any type of data. Most of these tests are called rank sum tests: all the sample data are sorted in order and assigned a ‘rank’, and then the significance test compares the ranks of the data from different groups.

In practice, parametric statistics are reasonably reliable if used for continuously variable data that are not normally distributed provided the deviation is not too extreme. Parametric statistics should never be used for ordinal data, but non-parametric statistics can be used for all data, including the normal distribution. When the data are normally distributed, it is better to use the parametric tests specifically designed for the normal distribution.

Inferential statistics

The subjects actually studied are a sample of a wider parent population, and we wish to use the results of the sample to infer the likely properties of the parent population. It is never known

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005

how representative this sample is, so these inferences are always made with some uncertainty. This uncertainty is measured by probabilities, and these probabilities measure the degree of confidence of our conclusions about the parent population.

The central limit theorem is the basis for much of inferential statistics. This states that, if several samples are taken from a population, the means of these samples are distributed normally around the true population mean. The standard deviation of this normal distribution of the sample means is called the standard error of the mean (SEM). This is true even if the variable is not distributed normally in the population, provided that the samples are sufficiently large (Figure 7).

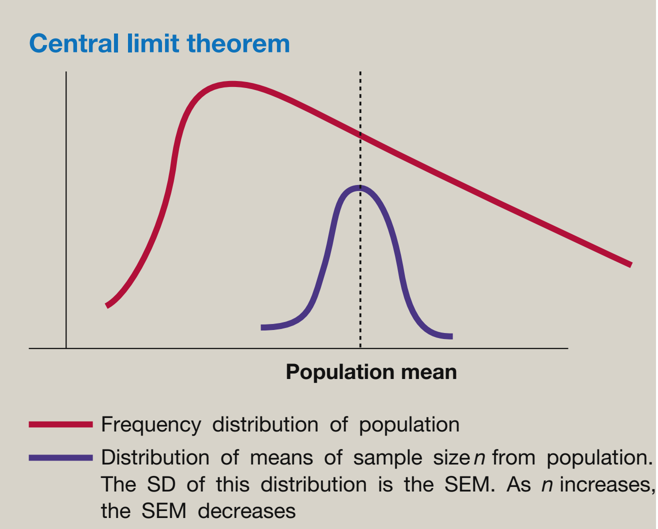
If the variable is distributed normally within the population (unlike Figure 7, in which the population data are positively skewed), then we can further obtain an estimate of the SEM from the sample standard deviations (Figure 8).

Using the properties of the normal distribution, an estimate of the true population mean can then be obtained from the sample mean. We can be 95% confident that the true population mean will lie within the range:

$$\bar{x} \pm (2 \times \text{SEM})$$

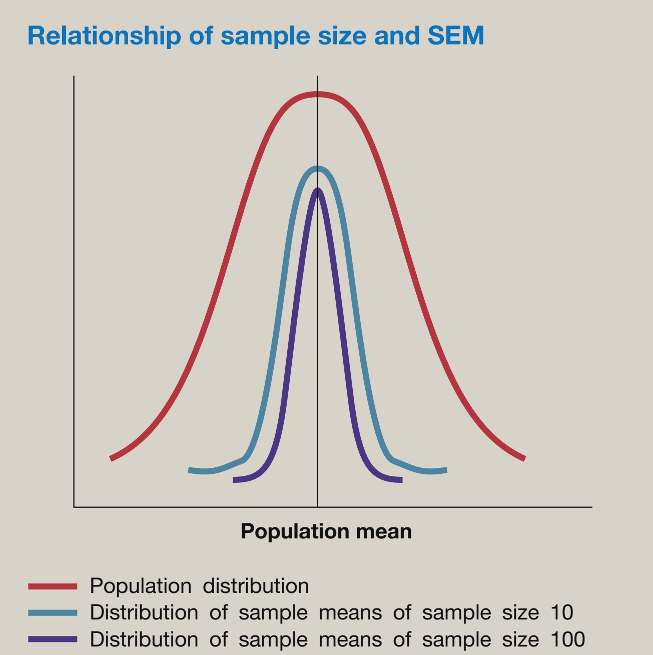
where $\bar{x}$ is the sample mean. This range is called the 95% confidence interval for the true population mean. If we had 100 different samples, we could obtain 100 different estimates of this range; in about 95 of these, the true population mean would be within this range, and in about five of these estimates the true population mean would not lie within this range. In practice, we usually have only one sample, and we do not know whether this is one of the 5% of occasions when the true population mean is outside the calculated range.

As n increases, the SEM decreases, and the 95% confidence intervals for the true population mean are narrower. Intuitively, large samples will be more representative of the whole population, and the sample means of large samples will be more closely clustered about the true population mean than those of small samples.



The mean of samples taken from the population is distributed normally about the population mean. SD, standard deviation; SEM. standard error of the mean.

Figure 7



SEM = SD/√*n*, where SD is the standard deviation of the sample, SEM is the standard error of the mean,*n* and is the number of subjects in the sample.

Figure 8

We have therefore used our sample and the central limit theorem to infer properties about the unknown parent population from which the sample was obtained. This introduces the concepts of 'assumptions' (the assumptions that allow us to use the central limit theorem), of 'probabilities' (the true population mean is 95% probable to lie within the calculated range) and 'inferential errors' (we make conclusions based on probability not certainty). Our application of statistics may be completely correct, but our conclusions can still be wrong!

Hypothesis testing

One of the principal uses of statistics in medicine is to decide whether the data from a clinical trial represent a real difference between treatments or could our results have arisen by chance and the treatments are actually indistinguishable. This is called hypothesis testing. A simple example would be if a clinical trial has been completed and the heart rates have been recorded in two groups of patients treated with drugs A and B. Our samples are drawn from two hypothetical populations: population A, all patients (similar to those in the study given drug A) who are or could be taking drug A, and population B, all patients (similar to those in the study given drug B) who are or could be taking drug B. We do not know how representative our samples are of these two populations. All statistical tests calculate a probability or confidence limits that the sample results could have been obtained if populations A and B did not differ. The logical stages in hypothesis testing are:

1 **Form a null hypothesis.** This states that the frequency distribution of heart rates in population A is the same as that in population B. The alternative hypothesis is

that the populations A and B have different frequency distributions.

2 **Choose the appropriate statistical test.** In this case, we have continuously variable data that are usually normally distributed, so we can use a *t*-test.

3 **Obtain the *p*-value.** The statistical test uses the actual heart rates of our samples to calculate the probability (the *p*-value) that we could have obtained our sample data if the null hypothesis were true, that is if the samples had been obtained randomly from a single population.

4 **Accept or reject the null hypothesis.** Conventionally, a probability of 0.05 (5%) is chosen as the cut-off for the *p*-value as sufficiently unlikely that we can reject the null hypothesis. This value is called the level of statistical significance, and can be chosen at a lower value (e.g. 0.01) if we wish to make our conclusions less likely to be wrong. A value higher than 0.05 is very rarely used. If the *p*-value is less than 0.05, we conclude that drugs A and B have different effects on the heart rate. If $p > 0.05$, then there is a reasonable probability that we could have obtained these results because populations A and B are the same, and we conclude that drug A and drug B do not have different effects. Both of these conclusions may be wrong. This pattern of reasoning is common to all clinical trials using inferential statistical tests for hypothesis testing for differences between groups in the study (Table 2).

Choice of statistical test

The choice of statistical test is determined by the type of data and the number of groups; the most common statistical tests used in medical research are shown in Table 3.

The logical steps in hypothesis testing, and possible errors

Hypothesis testing	Possible errors
Form the null hypothesis	
Choose a statistical test	Incorrect test chosen, e.g. one that is not applicable to this type of data
Obtain a *p*-value	
Reject the null hypothesis if $p < 0.05$	1 By definition, this decision will be incorrect on 5% of occasions (a type I error) 2 Poor study design. Our samples do show a real difference, but this is caused by bias; either because of poor study design or by chance, our samples differ in some important confounding factor
Accept the null hypothesis if $p > 0.05$, and conclude the treatments do not have different effects	There is a real difference in the treatments, but our study has failed to demonstrate this difference. This is a type II error, and is much more common than a type I error. The most common cause of a type II error is insufficient numbers of patients in the study

Table 2

In some studies, each subject in one group is uniquely paired with one in the other group(s). For example, if a variable is measured in the same individual before and after an intervention, then these two observations are 'paired'. There are appropriate statistical tests that should be used for this type of data.

Multiple significance testing

If there are three or more groups in a study, it is tempting to test all the possible paired combinations to determine the differences between the various treatments, that is A versus B, B versus C, A versus C, etc. If all these combinations are tested with a statistical significance level of 5%, the risk of finding spurious differences by chance (a type I error) increases considerably. The correct procedure is to use the appropriate statistical test for three or more groups (Table 3). If we reject the null hypothesis, then we conclude there are differences between the three or more treatments used in the study, but we do not know which treatments differ significantly.

There two ways around this problem. Special multiple comparison techniques (e.g. Scheffe F test, Duncan's test) can be used to determine the differences between the groups that limit the **overall** risk of a type I error to 5%. Alternatively, a Bonferroni technique can be used. A decision is made on how many paired comparisons between groups are required, and each of the chosen paired comparisons is tested with a lower level of statistical significance (usually 0.05 divided by the number of comparisons). This works well if there fewer than five comparisons, but, above this, the technique becomes very conservative, and differences between groups do not achieve statistical significance even when the differences are large.

Contingency tables

Contingency tables are used to analyse categorical data. Usually, the rows of the table are the different groups and the columns are the different categories to which the patients are allocated. Each cell in a table is the number of subjects from that group that have been allocated to that category (Table 4).

Contingency tables can be used for any number of groups and any number of observations about the groups; for example, a study of the GCS of patients in 10 cities would be a 10 × 13 contingency table (10 cities, 13 possible GCS). There are problems of using large contingency tables. In the above example, there may be only one city that is different and the other nine are similar; the analysis may not detect this single difference. The logical process of hypothesis testing is exactly as before, that is, there is a null and alternative hypothesis and a *p*-value is obtained from the sample data. The test commonly used for contingency tables is called the chi-squared (χ^2) test. The number of expected patients if the null hypothesis were true is calculated for each cell and the difference between the observed and expected in each cell is used to obtain the *p*-value.

There are a number of conditions that must be observed when using contingency tables.

- The entries in each cell of the table must be the actual number of patients, not percentages.
- The χ^2 test cannot be used if more than 20% of the cells have an expected value of less than 5; Fisher's exact test is an alternative.

Choice of statistical test is determined by the type of data and the number of groups

Type of data	Two groups	More than two groups
Categorical data (e.g. blood group)	Contingency tables	Contingency tables
Ordinal data (e.g. Glasgow coma score)	Unpaired: Mann—Whitney test Paired: Wilcoxon rank sum test	Unpaired: Kruskall—Wallis test Paired: Friedman's test
Continuously variable data, normally distributed (e.g. weight)	Unpaired: *t*-test Paired: paired *t*-test	Unpaired: ANOVA Paired: paired ANOVA
Continuously variable data, not normally distributed (e.g. duration of hospital stay)	As for ordinal data, or transform the data to a normal distribution	As for ordinal data, or transform the data to a normal distribution

Table 3

A 2 × 2 contingency table, comparing the incidence of nausea and vomiting after two different analgesic drugs

	Nausea and vomiting	No nausea and vomiting
Morphine	14	9
Ketorolac	3	17

Table 4

- There are particular problems in 2 × 2 tables with small sample sizes. If the total in the table is less than 50, a modification of the χ^2, called Yates' continuity correction should be used.

Repeated measurements

Many studies in anaesthesia involve making a series of observations with time on a single subject; for example, the heart rate and blood pressure are measured in each patient for some time after administering a drug. This violates one of the assumptions of statistical tests that all the observations are independent of the others. This is clearly not the case in this experiment; if the heart rate is high, subsequent measurements are also more likely to be high. This type of study can also generate enormous amounts of data, and multiple statistical comparisons are often made searching for statistical differences, increasing the risk of type I errors.

There are two methods of analysing this type of data.

- The simpler method is to use a 'summary measure' for the observations; the changes with time that have been measured in each patient are summarized in a single value used for the analysis. For example, if the systolic blood pressure and heart rate after induction of anaesthesia have been recorded, then examples of suitable summary measures for each patient could be: the lowest systolic arterial pressure, the highest heart rate, the time to lowest blood pressure or the mean heart rate. Which summary measures should be chosen would depend on which clinical factors were thought to be most important. Summary measures can usually be analysed using simple methods.
- The alternative method is to use statistical tests designed for these data, such as analysis of variance for repeated measures.

Correlation and regression

These techniques are used to measure the relationships between two or more variables; for example, do they rise and fall together, are they inversely related (i.e. one decreases while the other increases), is the association linear or is there no relationship at all?

There are important differences between correlation and regression:

- correlation measures the degree of relationship between two **independent** variables
- regression mathematically expresses the dependence of one **dependent** variable on another **independent** variable; regression is used to predict the dependent variable from the independent variable.

The common statistical techniques for correlation and regression measure linear relationships, but non-linear relationships can be expressed mathematically.

Linear regression is expressed by the equation of a straight line:

$$y = mx + c$$

where y is the dependent variable, x is the independent variable, m is the slope of the line and c is the intercept on the y axis.

Statistical programmes will calculate m and c, the statistical significance of this linear relationship, and confidence limits for the true population values for m and c, provided certain assumptions are met. These are:

- the possible values for y in the population for any given value of x should be normally distributed
- the variability of this normal distribution for y should be the same for all values of x
- the relationship is linear.

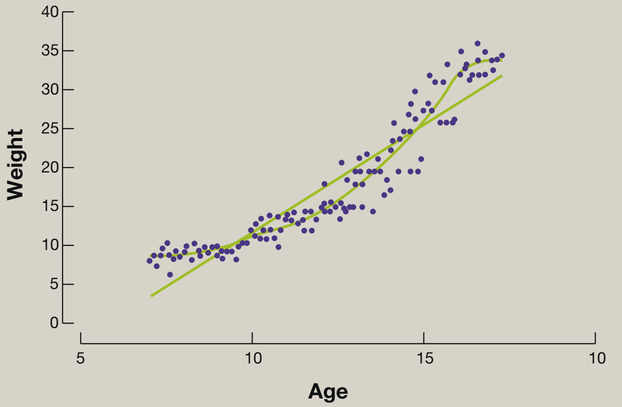
These assumptions can be tested mathematically, but it is usually sufficient to visually check the scattergram for any obvious deviations. If there is no scattergram, the reader cannot check whether linear analysis is appropriate (Figure 9).

Correlation measures the degree of linear association of two independent variables. A common mistake is to assume that, if two variables are correlated, there must be a causal relationship. It is important to inspect the scattergram before using computer programs to obtain a correlation coefficient (Figure 10).

Correlation is most commonly expressed by the Pearson correlation coefficient r. This number can vary between −1 and +1.

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005

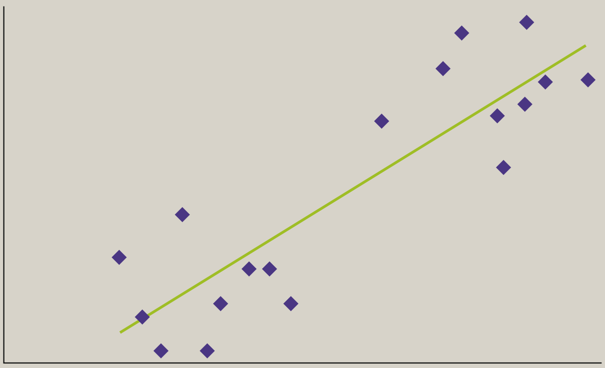
Non-linear relationship



The ages and weights of 130 boys have been plotted and show a non-linear relationship because of the growth increase at puberty. A computer program would derive a highly statistically significant linear regression, although the relationship is better described as sigmoid. Accepting a linear regression equation just because it is statistically significant would miss important information in the data.

Figure 9

Errors in correlation



A statistically significant linear correlation in the whole sample can be calculated, but the data actually have a bimodal pattern, and within each subgroup there is no correlation of the two variables.

Figure 10

The + or − signs convey whether the relationship is positive (i.e. both variables increase together) or inverse (one increases as the other decreases), respectively. The numerical magnitude of r depends on the scatter of the points about the line of best fit. If all the data points were to lie exactly on a straight line, then $r = +1$ or -1. As the amount of scatter about the line of best fit increases, then r approaches 0. The value of r^2 is the variability of one variable that is associated with change in the other variable and is called the coefficient of determination. The value of $(1 - r^2)$ is the amount of change in one of the variables that is **not** associated with changes in the other variable and must be associated with other factors.

We can calculate confidence limits for the true population value of r using either parametric or non-parametric methods. The Pearson correlation coefficient can be used for data that are continuously variable and approximately normally distributed in the population. If this assumption is not correct then the alternative non-parametric method, the Spearman rank correlation coefficient, should be used. The Spearman non-parametric technique has two advantages: the strength of non-linear associations can be measured and the association between ordered categories can be measured.

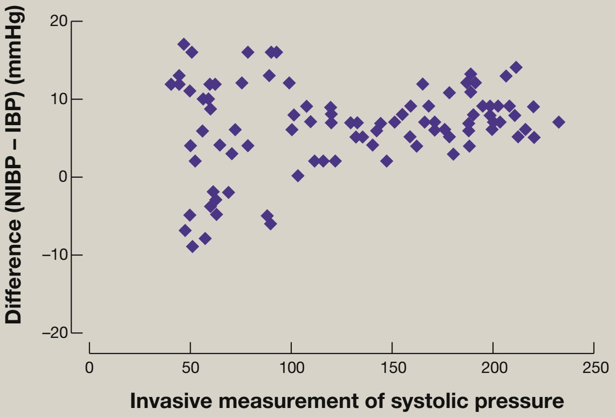
Method comparison studies

Any new method of measurement needs to be compared with the standard techniques. This is done by measuring the variable using both the standard and new technique, and statistically comparing the set of paired measurements. The correlation coefficient has been commonly used for this purpose, but this is incorrect. If a new method is being compared with a standard method, and there is a constant bias to the new method compared with the standard, then the linear correlation coefficient will be +1 because there is perfect correlation, but complete disagreement between the two measurements. A better technique is to use the Bland—Altman plot (Figure 11).

Multiple regression analysis

The dependent variable can be expressed mathematically as a combination of any number of independent variables, either linearly or non-linearly, and this technique is 'multiple regression'. Logistic regression allows us to extend regression techniques to both dependent or independent variables with categorical states, for example smokers/non-smokers, survived/dead.

The Bland–Altman technique for a method comparison study



A new non-invasive method of measuring the arterial blood pressure (NIBP) is being calibrated using simultaneous intra-arterial measurements. The Bland–Altman plot has the systolic pressure measured invasively (IBP) as the *x*-axis and the difference between the two measurements (NIBP − IBP) as the *y*-axis. The mean difference of (NIBP − IBP) is the **bias**, and the 95% confidence interval for the difference is the **precision** or limits of agreement for the two methods.

Figure 11

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005

It is common, especially in epidemiological studies, to find many possible factors associated with the prevalence of a disease. Some of these are independent causative risks and some are associated through other common factors. In the 1980s, smoking was found to be strongly associated with cervical cancer and proposed as a causative agent along with many other possibilities, including parity and alcohol consumption. It is now generally accepted that cervical cancer is commonly caused by a sexually transmitted virus, and these other correlated factors are more commonly present in women likely to acquire the virus. Stepwise logistic regression is a technique that mathematically includes those factors independently associated with the condition and removes from the equation those factors associated through some other common feature.

Meta-analysis and systematic reviews

Meta-analysis is a statistical technique used to combine data from several studies of the same topic to reach a single definitive conclusion. There are varying mathematical techniques for combining the data from different studies, and there is no single ‘correct’ method. Like all statistical methods, useful results are obtained only if the data and methods are used correctly, and there have been some serious errors (Box 1).

If all the studies used the same protocol, the data could be combined with confidence. However, this is rarely the case, and studies differ in selection of patients, treatment schemes, etc. There are statistical techniques that can estimate the amount of variability between the results of different studies that would occur by chance, and if these are the only differences between the studies, they are termed homogeneous. Any greater variability between the studies could be caused by important differences between the methodologies of the studies, and is termed heterogeneity. The amount of heterogeneity permitted in a reliable meta-analysis is not known.

Example of publication bias leading to an incorrect conclusion from a meta-analysis.

Several studies in the 1980s demonstrated that intravenous magnesium reduced mortality after acute myocardial infarction, and a meta-analysis confirmed an important reduction in mortality. Subsequently a large single-centre study showed no beneficial effect of magnesium. A re-examination of the meta-analysis suggested a strong publication bias. Small studies showing a positive beneficial effect of magnesium had been published in journals, but small studies which showed no effect had not been published, probably because they had been rejected by the editors for insufficient power. The meta-analysis, therefore, included a biased selection of all the clinical trials of magnesium that were done and the conclusion was incorrect. To help avoid this problem in the future, a central register of all clinical trials has been suggested. This will enable reviewers to locate all the trials on a subject, published or not.

Box 1

In a systematic review, the methods used to find all the published papers on the topic, the criteria used to assess the quality of the papers and the techniques for combining and analysing the data are all decided in advance and reported in detail in the published paper. Readers ought to be able to repeat the methods used by the authors with the same results. The two principal problems of conventional reviews are failure to find all the papers published on a topic (especially those in a foreign language) and bias by the reviewers. Some conventional reviews and editorials can be little more than disguised polemics. ◆

FURTHER READING

Altman DG. Practical statistics for medical research. London: Chapman & Hall, 1991.

Please cite this article in press as: Kestin I, Statistics in medicine, Anaesthesia and intensive care medicine (2017), https://doi.org/10.1016/j.mpaic.2017.12.005