

Editorial

Introducing the *Critical Care Forum's* ongoing review of medical statistics

Elise Whitley* and Jonathan Ball†

*Lecturer in Medical Statistics, University of Bristol, Bristol, UK

†Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial office, *Critical Care Forum*, editorial@ccforum.com

Published online: 29 January 2002

Critical Care 2002, **6**:3

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

Statistics is increasingly used in all fields of medicine but is often poorly understood and incorrectly applied. *Critical Care* is therefore launching a series of articles aimed at providing a simple introduction or refresher to some of the more commonly used statistical tools and ideas. This series does not aim to be an exhaustive review of medical statistics but rather a starting point to inform readers and stimulate more thought and investigation as to the most appropriate statistical methods to use and the theory and assumptions behind them.

Keywords data analysis, medical statistics

The science of statistics is increasingly employed in all fields of medicine. Statistical techniques are used not only by academics and clinicians directly involved in medical research but also by advocates of evidence-based medicine, who must synthesise results from many different sources to reach useful conclusions. Because of this widespread use, it is important that all those involved in research or the management of patients have a sound grasp of at least the basics of statistical methods. Unfortunately, in practice this is often not true, with many relying on distant memories of poorly understood lectures from undergraduate courses.

In response to this, *Critical Care* is launching a series of articles aimed at providing a simple introduction and/or refresher to some of the more common tools and ideas used in medical statistics. The articles are aimed at a non-specialist audience and will keep algebra and technical language to a minimum. Although some of the topics covered in this series will probably be familiar, it is hoped that there will still be useful lessons to be learned, for example the underlying assumptions of a hypothesis test that were not fully appreciated, or some previously unrecognised confusion between terms.

The first article, presented in this issue, covers the presentation and summary of data. It is unlikely that the material covered by this article will be entirely new to any reader but it is included as a simple introduction to some of the ideas and philosophies

that will be built upon in subsequent articles. Topics to be covered in the series include: standard errors and confidence intervals; hypothesis testing and errors; power calculations; measures of disease; parametric and non-parametric tests; simple regression; and analysis of survival data. Ideally the series will evolve to meet the needs of *Critical Care* readers, and you are encouraged to suggest additional topics that you would like to see covered in the future.

It is vital that the quality of medical research continues to improve and that readers develop a critical eye when considering evidence from published reports. The conduct of badly designed, under-powered and inappropriately analysed studies is not only an indefensible waste of precious resources but is also highly unethical. Unfortunately such research is all too common, and every effort should be made to prevent these situations from arising. Statistical statements can enlighten or mislead depending on how well they are understood, and individuals have a responsibility to ensure that their knowledge is sufficient for their needs. It is hoped that this series will inform readers but also that it will stimulate more thought and investigation as to the most appropriate statistical methods to use and the theory and assumptions behind them.

This series does not claim to be a complete course in medical statistics. There are many useful introductory texts

that cover the ideas presented in this series, and more, in considerably greater detail [1–4]. However, even these might frequently not be sufficient and it is vital that researchers recognise their own limitations and seek professional advice whenever it is needed, if only for reassurance. Medical statistics is a scientific discipline in its own right and a medical statistician fully achieves that role only after years of training and practical experience. Most academic departments, and also many clinical departments, include properly qualified medical statisticians and they should be consulted as early as possible in the research process.

Competing interests

None declared.

References

1. Altman DG: *Practical Statistics for Medical Research*. London: Chapman & Hall; 1991.
2. Bland M: *An Introduction to Medical Statistics*, edn 3. Oxford: Oxford University Press; 2001.
3. Campbell MJ, Machin D: *Medical Statistics: A Commonsense Approach*, edn 2. London: John Wiley & Sons Ltd; 1993.
4. Kirkwood BR: *Essentials of Medical Statistics*. London: Blackwell Science Ltd; 1988.

Review

Statistics review 1: Presenting and summarising data

Elise Whitley* and Jonathan Ball†

*Lecturer in Medical Statistics, University of Bristol, Bristol, UK

†Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 29 November 2001

Critical Care 2002, 6:66-71

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review is the first in an ongoing guide to medical statistics, using specific examples from intensive care. The first step in any analysis is to describe and summarize the data. As well as becoming familiar with the data, this is also an opportunity to look for unusually high or low values (outliers), to check the assumptions required for statistical tests, and to decide the best way to categorize the data if this is necessary. In addition to tables and graphs, summary values are a convenient way to summarize large amounts of information. This review introduces some of these measures. It describes and gives examples of qualitative data (unordered and ordered) and quantitative data (discrete and continuous); how these types of data can be represented figuratively; the two important features of a quantitative dataset (location and variability); the measures of location (mean, median and mode); the measures of variability (range, interquartile range, standard deviation and variance); common distributions of clinical data; and simple transformations of positively skewed data.

Keywords interquartile range, mean, median, range, standard deviation, transformations, unimodal distributions

Data description is a vital part of any research project and should not be ignored in the rush to start testing hypotheses. There are many reasons for this important process, such as gaining familiarity with the data, looking for unusually high or low values (outliers) and checking the assumptions required for statistical testing. The two most common types of data are qualitative and quantitative (Fig. 1). Qualitative data fall into two categories: unordered qualitative data, such as ventilatory support (none, non-invasive, intermittent positive-pressure ventilation, oscillatory); and ordered qualitative data, such as severity of disease (mild, moderate, severe). Quantitative data are numerical and fall into two categories: discrete quantitative data, such as the number of days spent in intensive care; and continuous quantitative data, such as blood pressure or haemoglobin concentrations. Tables are a useful way of describing both qualitative and grouped quantitative data and there are also many types of graph that provide a convenient summary. Qualitative data are commonly described using bar or pie charts, whereas quantitative data can be represented using histograms or box and whisker plots.

Tables and graphs provide a convenient simple picture of a set of data (dataset), but it is often necessary to further summarize quantitative data, for example for hypothesis testing. The two most important elements of a dataset are its location (where on average the data lie) and its variability (the extent to which individual data values deviate from the location). There are several different measures of location and variability that can be calculated, and the choice of which to use depends on individual circumstances.

Measuring location

Mean

The mean is the most well known average value. It is calculated by summing all of the values in a dataset and dividing them by the total number of values. The algebraic notation for the mean of a set of n values (X_1, X_2, \dots, X_n) is:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (1)$$

where $\sum_{i=1}^n X_i$ is the mathematical notation for the sum of all values (X_1, X_2, \dots, X_n). In other words:

$$\sum_{i=1}^n X_i = X_1 + X_2 + \dots + X_n$$

Of all the measures of location, the mean is the most commonly used because it is easily understood and has useful mathematical properties that make it convenient for use in many statistical contexts. It is strongly influenced by extreme values (outliers), however, and is most representative when the data are symmetrically distributed (see below).

Median

The median is the central value when all observations are sorted in order. If there is an odd number of observations then it is simply the middle value; if there is an even number of observations then it is the average of the middle two. The median does not have the beneficial mathematical properties of the mean. However, it is not generally influenced by extreme values (outliers), and as a result it is particularly useful in situations where there are unusually low or high values that would render the mean unrepresentative of the data.

Mode

The mode is simply the most commonly occurring value in the data. It is not generally used because it is often not representative of the data, particularly when the dataset is small.

Example of calculating location

To see how these quantities are calculated in practise, consider the data shown in Table 1. These are haemoglobin concentration measurements taken from 48 patients on admission to an intensive care unit, listed here in ascending order.

The first step in exploring these data is to construct a histogram to illustrate the shape of the distribution. Rather than plot the frequency of each value separately (e.g. one patient with haemoglobin 5.4 g/dl, two patients with haemoglobin 6.4 g/dl, one patient with haemoglobin 7.0 g/dl, and so on), continuous data are generally grouped or categorized before

Table 1

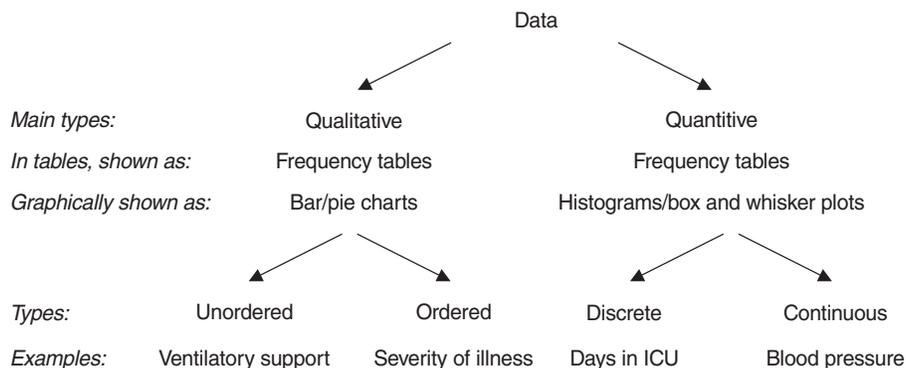
Haemoglobin (g/dl) from 48 intensive care patients					
5.4	8.2	9.3	9.9	10.5	11.9
6.4	8.3	9.4	9.9	10.5	12.3
6.4	8.3	9.4	9.9	10.6	12.6
7.0	8.6	9.4	10.1	10.8	12.7
7.1	8.8	9.4	10.3	10.8	13.0
7.3	8.9	9.5	10.3	11.3	13.3
7.7	9.1	9.7	10.4	11.7	14.0
8.1	9.3	9.7	10.4	11.7	14.1

plotting (e.g. one patient with haemoglobin between 5.0 and 5.9 g/dl, two patients with haemoglobin between 6.0 and 6.9 g/dl, four patients with haemoglobin between 7.0 and 7.9 g/dl, and so on). These categories can be defined in any way and need not necessarily be of the same width, although it is generally more convenient to have equally sized groups. However, the categories must be exhaustive (the categories must cover the full range of values in the dataset) and exclusive (there should be no overlap between categories). Therefore, if one category ends with 6.9 g/dl then the next must begin with 7.0 g/dl rather than 6.9 g/dl. Fig. 2 shows the data in Table 1 grouped into 1 g/dl categories (5.0–5.9, 6.0–6.9, ..., 14.0–14.9 g/dl).

Fig. 2 shows that the data are roughly symmetrically distributed; more common values are clustered around a peak in the middle of the distribution, with decreasing numbers of smaller and larger values on either side. The mean, median and mode of these data are shown in Table 2.

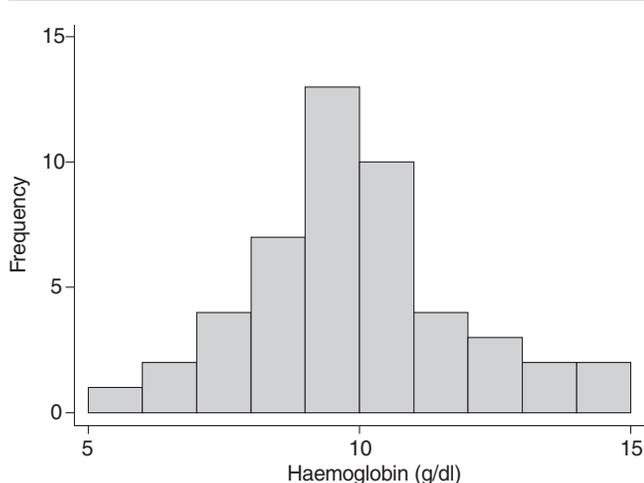
Notice that the mean and the median are similar. This is because the data are approximately symmetrical. In

Figure 1



Types of data. ICU = intensive care unit.

Figure 2



Histogram of admission haemoglobin measurements from 48 intensive care patients.

general, the mean, median and mode will be similar in a dataset that has a symmetrical distribution with a single peak, such as that shown in Fig. 2. However, the dataset presented here is rather small and so the mode is not such a good measure of location.

Measuring variability

Range

As with location, there are a number of different measures of variability. The simplest of these is probably the range, which is the difference between the largest and smallest observation in the dataset. The disadvantage of this measure is that it is based on only two of the observations and may not be representative of the whole dataset, particularly if there are outliers. In addition, it gives no information regarding how the data are distributed between the two extremes.

Interquartile range

An alternative to the range is the interquartile range. Quartiles are calculated in a similar way to the median; the median splits a dataset into two equally sized groups, tertiles split the data into three (approximately) equally sized groups, quartiles into four, quintiles into five, and so on. The interquartile range is the range between the bottom and top quartiles, and indicates where the middle 50% of the data lie. Like the median, the interquartile range is not influenced by unusually high or low values and may be particularly useful when data are not symmetrically distributed. Ranges based on alternative subdivisions of the data can also be calculated; for example, if the data are split into deciles, 80% of the data will lie between the bottom and top deciles and so on.

Standard deviation

The standard deviation is a measure of the degree to which individual observations in a dataset deviate from the mean

Table 2

Mean, median and mode of haemoglobin measurements from 48 intensive care patients listed in Table 1

Measure	Calculation
Mean	The mean is the sum of the observations divided by the number of observations, in this case $\frac{5.4 + 6.4 + \dots + 14.1}{48} = 9.9 \text{ g/dl}$
Median	There are 48 observations in this dataset and so the median is the average of the 24th and 25th (i.e. the average of 9.7 and 9.9 = 9.8 g/dl)
Mode	Several values appear twice in this dataset, 9.9 appears three times and 9.4 appears four times. No value appears more than four times and so the mode is 9.4 g/dl

value. Broadly, it is the average deviation from the mean across all observations. It is calculated by squaring the difference of each individual observation from the mean (squared to remove any negative differences), adding them together, dividing by the total number of observations minus 1, and taking the square root of the result.

Algebraically the standard deviation for a set of n values $\{X_1, X_2, \dots, X_n\}$ is written as follows:

$$SD = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)}} \tag{2}$$

where $\sum_{i=1}^n (X_i - \bar{X})^2 = (X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \dots + (X_n - \bar{X})^2$

and \bar{X} is the mean described above (Eqn 1). It can be seen from this expression that if individual observations are all close to the mean then the standard deviation will be small (at the extreme, if all observations were equal to the mean then the standard deviation would be zero). Conversely, if the observations vary widely then the standard deviation will be substantially larger. The standard deviation summarizes a great deal of information in one number and, like the mean, has useful mathematical properties.

Variance

Another measure of variability that may be encountered is the variance. This is simply the square of the standard deviation:

$$Variance = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(n-1)} \tag{3}$$

The variance is not generally used in data description but is central to analysis of variance (covered in a subsequent review in this series).

Table 3**Range, interquartile range and standard deviation of haemoglobin measurements from 48 intensive care patients listed in Table 1**

Measure	Calculation
Range	The values in this dataset range from 5.4 to 14.1 g/dl
Interquartile range	The median calculated in Table 2 splits the data into two equally sized groups. The lower and upper quartiles split the data into four equally sized groups (4 × 12) and are therefore most easily defined as the average of the 12th and 13th observations for the lower quartile and of the 36th and 37th observations for the upper quartile. In other words, the lower and upper quartiles are 8.7 and 10.8 g/dl, respectively. (There are more complicated methods for calculating the interquartile range [1], but these will not generally give markedly different results.)
Standard deviation (SD)	Using the formula given above: $\begin{aligned} \text{SD} &= \sqrt{\frac{\sum_{i=1}^n (5.4 - 9.9)^2 + (6.4 - 9.9)^2 + \dots + (14.1 - 9.9)^2}{(48 - 1)}} \\ &= \sqrt{\frac{20.25 + 12.25 + \dots + 17.64}{47}} \\ &= 2.0 \text{ g/dL.} \end{aligned}$

Example of calculating variability

Table 3 shows the calculation of the range, interquartile range and standard deviation of the data shown in Table 1. The range, from 5.4 to 14.1 g/dl, indicates the full extent of the data, but does not give any information regarding how the remaining observations are distributed between these extremes. For example, it may be that the lower value of 5.4 g/dl is an outlier and the remainder of the observations are all over 10.0 g/dl, or that most values lie at the lower end of the range with substantially fewer at the other extreme. It is impossible to tell this from the range alone.

The interquartile range (which contains the central 50% of the data) gives a better indication of the general shape of the distribution, and indicates that 50% of all observations fall in a rather narrower range (from 8.7 to 10.8 g/dl). In addition, the median and mean both fall approximately in the centre of the interquartile range, which suggests that the distribution is reasonably symmetrical.

The standard deviation in isolation does not provide a great deal of information, although it is sometimes expressed as a percentage of the mean, known as the coefficient of variation. However, it is often used to calculate another extremely useful quantity known as the reference range; this will be covered in more detail in the next article.

Common distributions and simple transformations

Quantitative clinical data follow a wide variety of distributions, but the majority are unimodal, meaning that the data has a single (modal) peak with a tail on either side. The most common of these unimodal distributions are symmetrical, as

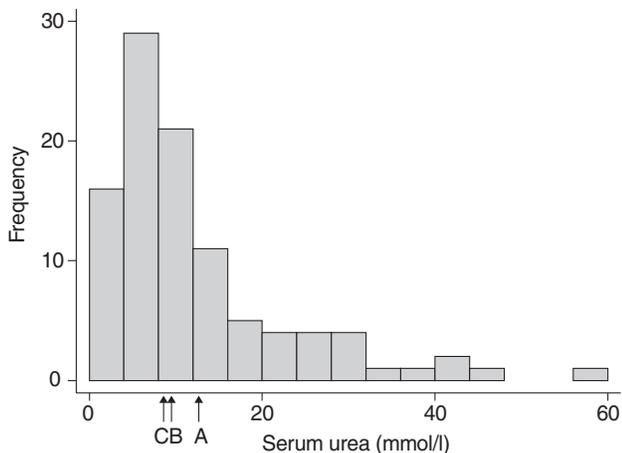
shown in Fig. 2, with a peak in the centre of the data and evenly balanced tails on the right and left.

However, not all unimodal distributions are symmetrical; some are skewed with a substantially longer tail on one side. The type of skew is determined by which tail is longer. A positively skewed distribution has a longer tail on the right; in other words the majority of values are relatively low with a smaller number of extreme high values. Fig. 3 shows the admission serum urea levels of 100 intensive care patients. The majority have a serum urea level below 20 mmol/l, with a peak between 4.0 and 7.9 mmol/l. However, an important minority of patients have levels above 20 mmol/l and some have levels as high as 60 mmol/l.

The mean of these data is 12.25 mmol/l (A) and the median is 9 mmol/l (B), as indicated in Fig. 3. In a positively skewed distribution the median will always be smaller than the mean because the mean is strongly influenced by the extreme values in the right-hand tail, and may therefore be less representative of the data as a whole. However, it is possible to transform data of this type in order to obtain a more representative mean value. This type of transformation is also useful when statistical tests require data to be more symmetrically distributed (see subsequent reviews in this series for details). There is a wide range of transformations that can be used in this context [2], but the most commonly used with positively skewed data is the logarithmic transformation.

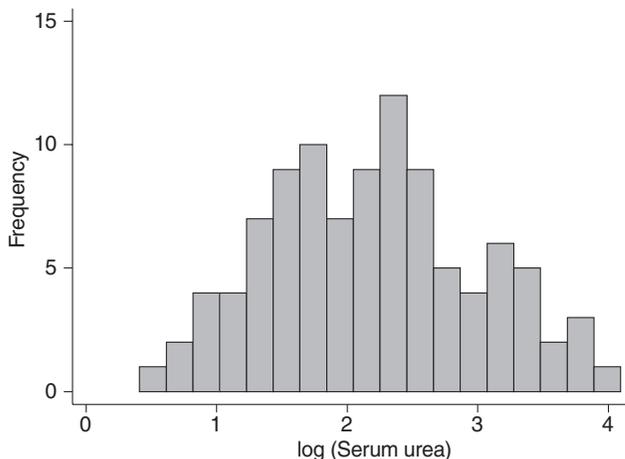
In a logarithmic transformation, every value in the dataset is replaced by its logarithm. Logarithms are defined to a base, the most common being base e (natural logarithms) or base 10. The end result of a logarithmic transformation is indepen-

Figure 3



Histogram of admission serum urea levels from 100 intensive care patients. A = mean; B = median; C = geometric mean.

Figure 4



Logarithmically transformed admission serum urea levels from 100 intensive care patients.

Table 4

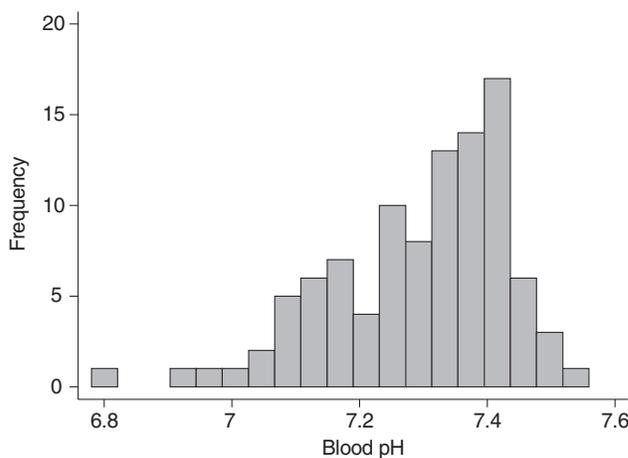
Raw and logarithmically transformed serum urea levels

Raw values	Transformed values
5	1.61
6	1.79
7	1.95
⋮	⋮
55	4.01
56	4.03
57	4.04

dent of the base chosen, although the same base must be used throughout. As an example, consider the data shown in Fig. 3. Although the majority of values are below 20, there is also an important number of values above this. Table 4 shows a sample of the raw numbers along with their logarithmically transformed values (to base e).

Notice that the differences between the raw values are always the same (1), whereas the differences in the transformed values are larger at the lower end of the scale (0.18 and 0.16) than at the upper end (0.02 and 0.01). The logarithmic transformation stretches out the lower end and compresses the upper end of a distribution, with the result that positively skewed data will tend to become more symmetrical in shape. The transformed data from Fig. 3 are shown in Fig. 4, in which it can be seen that there is a single peak at around 2.4 with similar tails to the right and left.

Figure 5



Admission arterial blood pH from 100 intensive care patients.

Calculations and statistical tests can now be carried out on the transformed data before converting the results back to the original scale. For example, the mean of the transformed serum urea data is 2.19. To transform this value back to the original scale, the antilog (or exponential in the case of natural, base e logarithms) is applied. This gives a 'geometric mean' of 8.94 mmol/l on the original scale (C in Fig. 3), the term 'geometric' indicating that calculations have been carried out on the logarithmic scale. This is in contrast to the standard (arithmetic) mean value (calculated on the original scale) of 12.25 mmol/l (A in Fig. 3). Looking at Fig. 3, it is clear that the geometric mean is more representative of the data than the arithmetic mean.

Similarly, a negatively skewed distribution has a longer tail to the left; in other words, the extreme values are at the lower end of the scale. Fig. 5 shows a negatively skewed distribution of admission arterial blood pH from 100 intensive care patients. In this case the mean will be unduly influenced by the extreme low values and the median (which is always greater than the mean in this setting) may be a more representative measure. However, as in the positively skewed case it is possible to transform this type of data in order to make it more symmetrical, although the function used in this setting is not the logarithm (for more details, see Kirkwood [2]).

Finally, it is possible that data may arise with more than one (modal) peak. These data can be difficult to manage and it may be the case that neither the mean nor the median is a representative measure. However, such distributions are rare and may well be artefactual. For example, a (bimodal) distribution with two peaks may actually be a combination of two unimodal distributions (such as hormone levels in men and women). Alternatively, a (multimodal) distribution with multiple peaks may be due to digit preference (rounding observations up or down) during data collection, where peaks appear at round numbers, for example peaks in systolic blood pressure at 90, 100, 110, 120 mmHg, and so on. In such cases appropriate subdivision, categorization, or even recollection of the data may be required to eliminate the problem.

Competing interests

None declared.

References

1. Altman DG: *Practical Statistics for Medical Research*. London: Chapman & Hall; 1991.
2. Kirkwood BR: *Essentials of medical Statistics*. London: Blackwell Science Ltd; 1988.

Review

Statistics review 2: Samples and populations

Elise Whitley* and Jonathan Ball†

*Lecturer in Medical Statistics, University of Bristol, UK

†Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 7 February 2002

Critical Care 2002, **6**:143-148

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The previous review in this series introduced the notion of data description and outlined some of the more common summary measures used to describe a dataset. However, a dataset is typically only of interest for the information it provides regarding the population from which it was drawn. The present review focuses on estimation of population values from a sample.

Keywords confidence interval, normal distribution, reference range, standard error

In medical (and other) research there is generally some population that is ultimately of interest to the investigator (e.g. intensive care unit [ICU] patients, patients with acute respiratory distress syndrome, or patients who receive renal replacement therapy). It is seldom possible to obtain information from every individual in the population, however, and attention is more commonly restricted to a sample drawn from it. The question of how best to obtain such a sample is a subject worthy of discussion in its own right and is not covered here. Nevertheless, it is essential that any sample is as representative as possible of the population from which it is drawn, and the best means of obtaining such a sample is generally through random sampling. (For more details see Bland [1].)

Once a (representative) sample has been obtained it is important to describe the data using the methods described in Statistics review 1. However, interest is rarely focused on the sample itself, but more often on the information that the sample can provide regarding the population of interest.

The Normal distribution

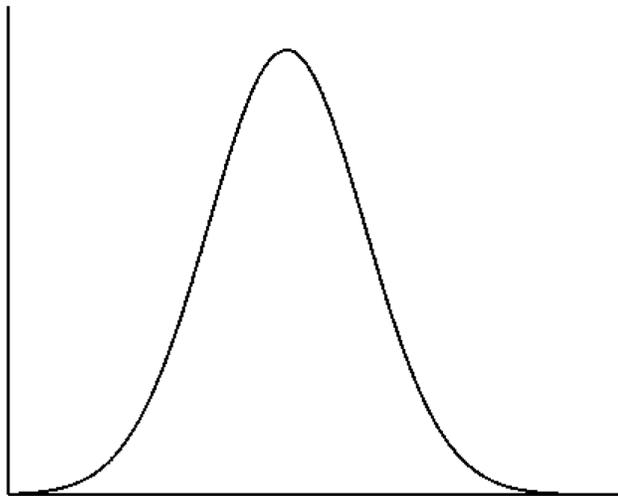
Quantitative clinical data follow a wide range of distributions. By far the most common of these is symmetrical and unimodal, with a single peak in the middle and equal tails at either side. This distinctive bell-shaped distribution is known as 'Normal' or 'Gaussian'. Note that Normal in this context (written with an upper case 'N') has no implications in terms of clinical normality, and is used purely to describe the shape of the distribution.

Strictly speaking, the theoretical Normal distribution is continuous, as shown in Fig. 1. However, data such as those shown in Fig. 2, which presents admission haemoglobin concentrations from intensive care patients, often provide an excellent approximation in practice.

There are many other theoretical distributions that may be encountered in medical data, for example Binary or Poisson [2], but the Normal distribution is the most common. It is additionally important because it has many useful properties and is central to many statistical techniques. In fact, it is not uncommon for other distributions to tend toward the Normal distribution as the sample size increases, meaning that it is often possible to use a Normal approximation. This is the case with both the Binary and Poisson distributions.

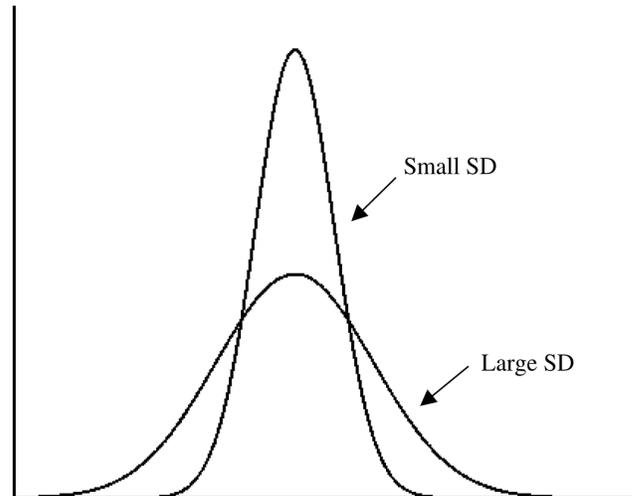
One of the most important features of the Normal distribution is that it is entirely defined by two quantities: its mean and its standard deviation (SD). The mean determines where the peak occurs and the SD determines the shape of the curve. For example, Fig. 3 shows two Normal curves. Both have the same mean and therefore have their peak at the same value. However, one curve has a large SD, reflecting a large amount of deviation from the mean, which is reflected in its short, wide shape. The other has a small SD, indicating that individual values generally lie close to the mean, and this is reflected in the tall, narrow distribution.

Figure 1



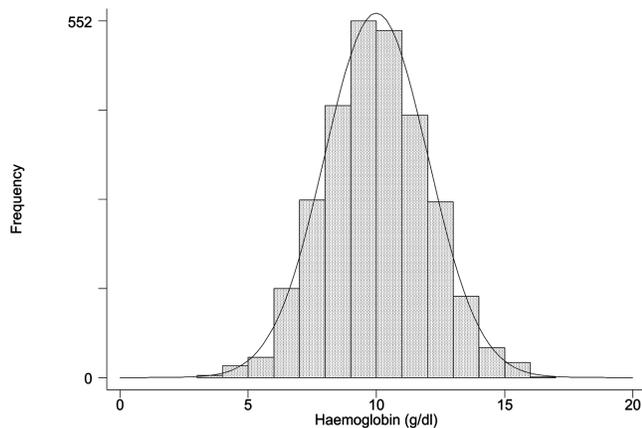
The Normal distribution.

Figure 3



Normal curves with small and large standard deviations (SDs).

Figure 2



Admission haemoglobin concentrations from 2849 intensive care patients.

It is possible to write down the equation for a Normal curve and, from this, to calculate the area underneath that falls between any two values. Because the Normal curve is defined entirely by its mean and SD, the following rules (represented by parts a–c of Fig. 4) will always apply regardless of the specific values of these quantities: (a) 68.3% of the distribution falls within 1 SD of the mean (i.e. between mean – SD and mean + SD); (b) 95.4% of the distribution falls between mean – 2 SD and mean + 2 SD; (c) 99.7% of the distribution falls between mean – 3 SD and mean + 3 SD; and so on.

The proportion of the Normal curve that falls between other ranges (not necessarily symmetrical, as here) and, alternatively, the range that contains a particular proportion of the Normal curve can both be calculated from tabulated values [3]. However, one proportion and range of particular interest is as follows (represented by part d of Fig. 4); 95% of the distribution falls between mean – 1.96 SD and mean + 1.96 SD.

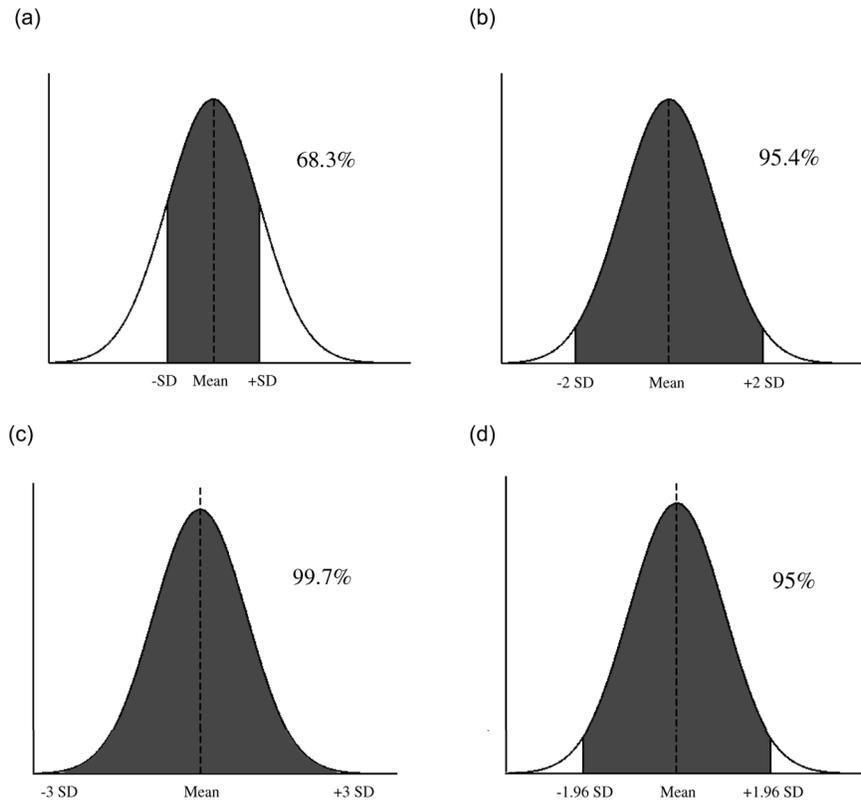
The standard deviation and reference range

The properties of the Normal distribution described above lead to another useful measure of variability in a dataset. Rather than using the SD in isolation, the 95% reference range can be calculated as (mean – 1.96 SD) to (mean + 1.96 SD), provided that the data are (approximately) Normally distributed. This range will contain approximately 95% of the data. It is also possible to define a 90% reference range, a 99% reference range and so on in the same way, but conventionally the 95% reference range is the most commonly used.

For example, consider admission haemoglobin concentrations from a sample of 48 intensive care patients (see Statistics review 1 for details). The mean and SD haemoglobin concentration are 9.9 g/dl and 2.0 g/dl, respectively. The 95% reference range for haemoglobin concentration in these patients is therefore:

$$(9.9 - [1.96 \times 2.0]) \text{ to } (9.9 + [1.96 \times 2.0]) = 5.98 \text{ to } 13.82 \text{ g/dl.}$$

Thus, approximately 95% of all haemoglobin measurements in this dataset should lie between 5.98 and 13.82 g/dl. Comparing this with the measurements recorded in Table 1 of Statistics review 1, there are three observations outside this range. In other words, 94% (45/48) of all observations are within the reference range, as expected.

Figure 4

Areas under the Normal curve. Because the Normal distribution is defined entirely by its mean and standard deviation (SD), the following rules apply: (a) 68.3% of the distribution falls within 1 SD of the mean (i.e. between mean $-$ SD and mean $+$ SD); (b) 95.4% of the distribution falls between mean $-$ 2 SD and mean $+$ 2 SD; (c) 99.7% of the distribution falls between mean $-$ 3 SD and mean $+$ 3 SD; and (d) 95% of the distribution falls between mean $-$ 1.96 SD and mean $+$ 1.96 SD.

Now consider the data shown in Fig. 5. These are blood lactate measurements taken from 99 intensive care patients on admission to the ICU. The mean and SD of these measurements are 2.74 mmol/l and 2.60 mmol/l, respectively, corresponding to a 95% reference range of -2.36 to $+7.84$ mmol/l. Clearly this lower limit is impossible because lactate concentration must be greater than 0, and this arises because the data are not Normally distributed. Calculating reference ranges and other statistical quantities without first checking the distribution of the data is a common mistake and can lead to extremely misleading results and erroneous conclusions. In this case the error was obvious, but this will not always be the case. It is therefore essential that any assumptions underlying statistical calculations are carefully checked before proceeding. In the current example a simple transformation (e.g. logarithmic) may make the data approximately Normal, in which case a reference range could legitimately be calculated before transforming back to the original scale (see Statistics review 1 for details).

Two quantities that are related to the SD and reference range are the standard error (SE) and confidence interval. These

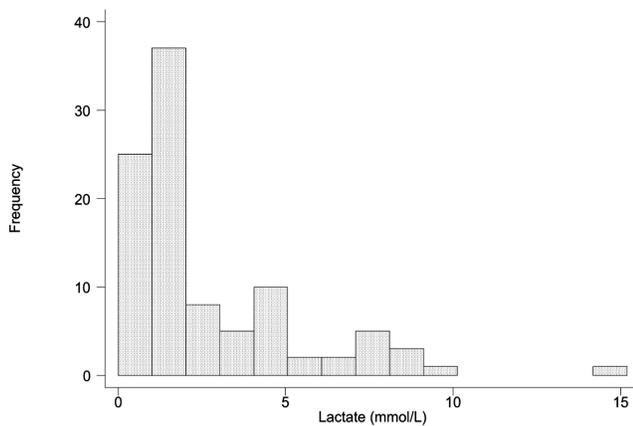
quantities have some similarities but they measure very different things and it is important that they should not be confused.

From sample to population

As mentioned above, a sample is generally collected and calculations performed on it in order to draw inferences regarding the population from which it was drawn. However, this sample is only one of a large number of possible samples that might have been drawn. All of these samples will differ in terms of the individuals and observations that they contain, and so an estimate of a population value from a single sample will not necessarily be representative of the population. It is therefore important to measure the variability that is inherent in the sample estimate. For simplicity, the remainder of the present review concentrates specifically on estimation of a population mean.

Consider all possible samples of fixed size (n) drawn from a population. Each of these samples has its own mean and these means will vary between samples. Because of this variation, the sample means will have a distribution of their own. In fact, if the samples are sufficiently large (greater than

Figure 5



Lactate concentrations in 99 intensive care patients.

approximately 30 in practice) then this distribution of sample means is known to be Normal, regardless of the underlying distribution of the population. This is a very powerful result and is a consequence of what is known as the Central Limit Theorem. Because of this it is possible to calculate the mean and SD of the sample means.

The mean of all the sample means is equal to the population mean (because every possible sample will contain every individual the same number of times). Just as the SD in a sample measures the deviation of individual values from the sample mean, the SD of the sample means measures the deviation of individual sample means from the population mean. In other words it measures the variability in the sample means. In order to distinguish it from the sample SD, it is known as the standard error (SE). Like the SD, a large SE indicates that there is much variation in the sample means and that many lie a long way from the population mean. Similarly, a small SE indicates little variation between the sample means. The size of the SE depends on the variation between individuals in the population and on the sample size, and is calculated as follows:

$$SE = \sigma/\sqrt{n} \tag{1}$$

where σ is the SD of the population and n is the sample size. In practice, σ is unknown but the sample SD will generally provide a good estimate and so the SE is estimated by the following equation:

$$SE = \text{Sample SD}/\sqrt{n} \tag{2}$$

It can be seen from this that the SE will always be considerably smaller than the SD in a sample. This is because there is less variability between the sample means than between individual values. For example, an individual admission haemoglo-

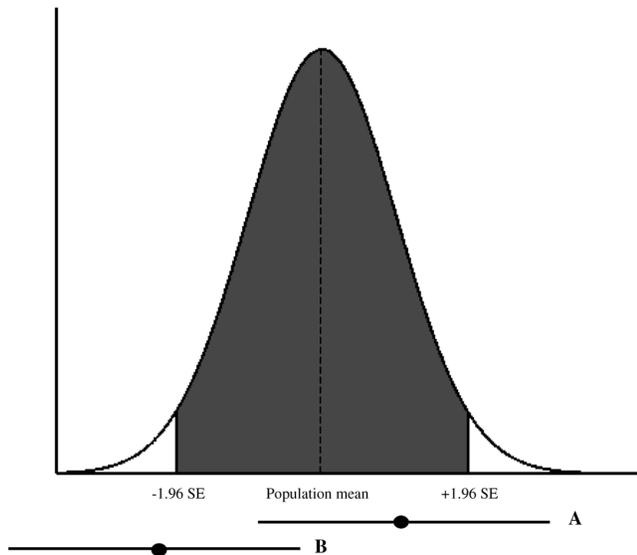
bin level of 8 g/dl is not uncommon, but to obtain a sample of 100 patients with a mean haemoglobin level of 8 g/dl would require the majority to have scores well below average, and this is unlikely to occur in practice if the sample is truly representative of the ICU patient population.

It is also clear that larger sample sizes lead to smaller standard errors (because the denominator, \sqrt{n} , is larger). In other words, large sample sizes produce more precise estimates of the population value in question. This is an important point to bear in mind when deciding on the size of sample required for a particular study, and will be covered in greater detail in a subsequent review on sample size calculations.

The standard error and confidence interval

Because sample means are Normally distributed, it should be possible to use the same theory as for the reference range to calculate a range of values in which 95% of sample means lie. In practice, the population mean (the mean of all sample means) is unknown but there is an extremely useful quantity, known as the 95% confidence interval, which can be obtained in the same way. The 95% confidence interval is invaluable in estimation because it provides a range of values within which the true population mean is likely to lie. The 95% confidence interval is calculated from a single sample using the mean and SE (derived from the SD, as described above). It is defined as follows: (sample mean - 1.96 SE) to (sample mean + 1.96 SE).

To appreciate the value of the 95% confidence interval, consider Fig. 6. This shows the (hypothetical) distribution of sample means centred around the population mean. Because the SE is the SD of the distribution of all sample means, approximately 95% of all sample means will lie within 1.96 SEs of the (unknown) population mean, as indicated by the shaded area. A 95% confidence interval calculated from a sample with a mean that lies within this shaded area (e.g. confidence interval A in Fig. 6) will contain the true population mean. Conversely, a 95% confidence interval based on a sample with a mean outside this area (e.g. confidence interval B in Fig. 6) will not include the population mean. In practice it is impossible to know whether a sample falls into the first or second category; however, because 95% of all sample means fall into the shaded area, a confidence interval that is based on a single sample is likely to contain the true population mean 95% of the time. In other words, given a 95% confidence interval based on a single sample, the investigator can be 95% confident that the true population mean (i.e. the real measurement of interest) lies somewhere within that range. Equally important is that 5% of such intervals will not contain the true population value. However, the choice of 95% is purely arbitrary, and using a 99% confidence interval (calculated as mean \pm 2.56 SE) instead will make it more likely that the true value is contained within the range. However, the cost of this change is that the range will be wider and therefore less precise.

Figure 6

The distribution of sample means. The shaded area represents the range of values in which 95% of sample means lie. Confidence interval A is calculated from a sample with a mean that lies within this shaded area, and contains the true population mean. Confidence interval B, however, is calculated from a sample with a mean that falls outside the shaded area, and does not contain the population mean. SE=standard error.

As an example, consider the sample of 48 intensive care patients whose admission haemoglobin concentrations are described above. The mean and SD of that dataset are 9.9 g/dl and 2.0 g/dl, respectively, which corresponds to a 95% reference range of 5.98 to 13.82 g/dl. Calculation of the 95% confidence interval relies on the SE, which in this case is $2.0/\sqrt{48} = 0.29$. The 95% confidence interval is then:

$$(9.9 - [1.96 \times 0.29]) \text{ to } (9.9 + [1.96 \times 0.29]) = 9.33 \text{ to } 10.47 \text{ g/dl}$$

So, given this sample, it is likely that the population mean haemoglobin concentration is between 9.33 and 10.47 g/dl. Note that this range is substantially narrower than the corresponding 95% reference range (i.e. 5.98 to 13.82 g/dl; see above). If the sample were based on 480 patients rather than just 48, then the SE would be considerably smaller ($SE = 2.0/\sqrt{480} = 0.09$) and the 95% confidence interval (9.72 to 10.08 g/dl) would be correspondingly narrower.

Of course a confidence interval can only be interpreted in the context of the population from which the sample was drawn. For example, a confidence interval for the admission haemoglobin concentrations of a representative sample of postoperative cardiac surgical intensive care patients provides a range of values in which the population mean admission haemoglobin concentration is likely to lie, in postoperative cardiac surgical intensive care patients. It does not provide information

on the likely range of admission haemoglobin concentrations in medical intensive care patients.

Confidence intervals for smaller samples

The calculation of a 95% confidence interval, as described above, relies on two assumptions: that the distribution of sample means is approximately Normal and that the population SD can be approximated by the sample SD. These assumptions, particularly the first, will generally be valid if the sample is sufficiently large. There may be occasions when these assumptions break down, however, and there are alternative methods that can be used in these circumstances. If the population distribution is extremely non-Normal and the sample size is very small then it may be necessary to use non-parametric methods. (These will be discussed in a subsequent review.) However, in most situations the problem can be dealt with using the t-distribution in place of the Normal distribution.

The t-distribution is similar in shape to the Normal distribution, being symmetrical and unimodal, but is generally more spread out with longer tails. The exact shape depends on a quantity known as the 'degrees of freedom', which in this context is equal to the sample size minus 1. The t distribution for a sample size of 5 (degrees of freedom = 4) is shown in comparison to the Normal distribution in Fig. 7, in which the longer tails of the t-distribution are clearly shown. However, the t-distribution tends toward the Normal distribution (i.e. it becomes less spread out) as the degrees of freedom/sample size increase. Fig. 8 shows the t-distribution corresponding to a sample size of 20 (degrees of freedom = 19), and it can be seen that it is already very similar to the corresponding Normal curve.

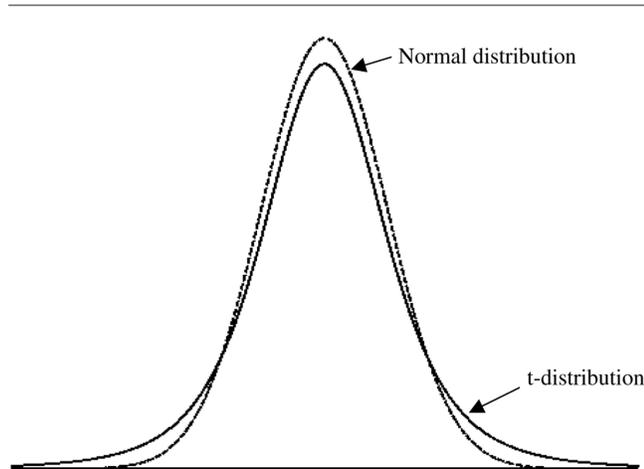
Calculating a confidence interval using the t-distribution is very similar to calculating it using the Normal distribution, as described above. In the case of the Normal distribution, the calculation is based on the fact that 95% of sample means fall within 1.96 SEs of the population mean. The longer tails of the t-distribution mean that it is necessary to go slightly further away from the mean to pick up 95% of all sample means. However, the calculation is similar, with only the figure of 1.96 changing. The alternative multiplication factor depends on the degrees of freedom of the t-distribution in question, and some typical values are presented in Table 1.

As an example, consider the admission haemoglobin concentrations described above. The mean and SD are 9.9 g/dl and 2.0 g/dl, respectively. If the sample were based on 10 patients rather than 48, it would be more appropriate to use the t-distribution to calculate a 95% confidence interval. In this case the 95% confidence interval is given by the following: mean \pm 2.26 SE. The SE based on a sample size of 10 is 0.63, and so the 95% confidence interval is 8.47 to 11.33 g/dl.

Table 1

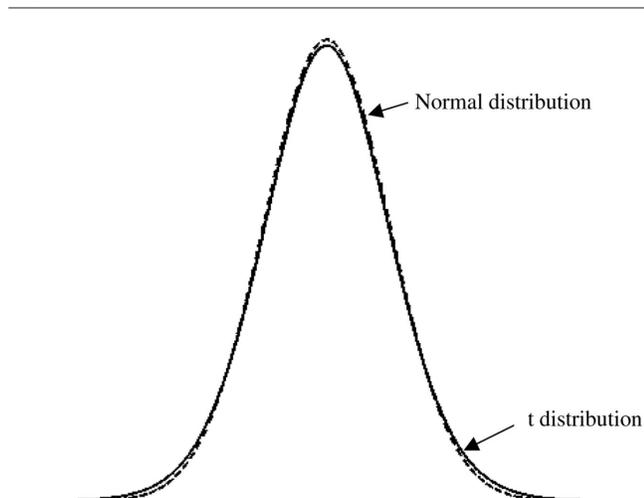
Multiplication factors for confidence intervals based on the t-distribution						
Sample size	10	20	30	40	50	200
Multiplication factor	2.26	2.09	2.05	2.02	2.01	1.97

Figure 7



The Normal and t (with 4 degrees of freedom) distributions.

Figure 8



The Normal and t (with 19 degrees of freedom) distributions.

Note that as the sample sizes increase the multiplication factors shown in Table 1 decrease toward 1.96 (the multiplication factor for an infinite sample size is 1.96). The larger multiplication factors for smaller samples result in a wider confidence interval, and this reflects the uncertainty in the estimate of the population SD by the sample SD. The use of the t-distribution is known to be extremely robust and will

therefore provide a valid confidence interval unless the population distribution is severely non-Normal.

Standard deviation or standard error?

There is often a great deal of confusion between SDs and SEs (and, equivalently, between reference ranges and confidence intervals). The SD (and reference range) describes the amount of variability between individuals within a single sample. The SE (and confidence interval) measures the precision with which a population value (i.e. mean) is estimated by a single sample. The question of which measure to use is well summed up by Campbell and Machin [4] in the following mnemonic: "If the purpose is Descriptive use standard Deviation; if the purpose is Estimation use standard Error."

Confidence intervals are an extremely useful part of any statistical analysis, and are referred to extensively in the remaining reviews in this series. The present review concentrates on calculation of a confidence interval for a single mean. However, the results presented here apply equally to population proportions, rates, differences, ratios and so on. For details on how to calculate appropriate SEs and confidence intervals, refer to Kirkwood [2] and Altman [3].

Key messages

The SD and 95% reference range describe variability within a sample. These quantities are best used when the objective is description.

The SE and 95% confidence interval describe variability between samples, and therefore provide a measure of the precision of a population value estimated from a single sample. In other words, a 95% confidence interval provides a range of values within which the true population value of interest is likely to lie. These quantities are best used when the objective is estimation.

Competing interests

None declared.

References

1. Bland M: *An Introduction to Medical Statistics*. 3rd ed. Oxford, UK: Oxford University Press; 2001.
2. Kirkwood BR: *Essentials of Medical Statistics*. London, UK: Blackwell Science Ltd; 1988.
3. Altman DG: *Practical Statistics for Medical Research*. London, UK: Chapman & Hall; 1991.
4. Campbell MJ, Machin D: *Medical Statistics: a Commonsense Approach*. 2nd ed. Chichester, UK: John Wiley & Sons Ltd; 1993.

Review

Statistics review 3: Hypothesis testing and *P* values

Elise Whitley¹ and Jonathan Ball²

¹Lecturer in Medical Statistics, University of Bristol, Bristol, UK

²Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 18 March 2002

Critical Care 2002, **6**:222-225

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review introduces the general philosophy behind hypothesis (significance) testing and calculation of *P* values. Guidelines for the interpretation of *P* values are also provided in the context of a published example, along with some of the common pitfalls. Examples of specific statistical tests will be covered in future reviews.

Keywords hypothesis testing, null hypothesis, *P* value

The previous review in this series described how to use confidence intervals to draw inferences about a population from a representative sample. A common next step in data analysis is calculation of *P* values, also known as hypothesis testing. Hypothesis testing is generally used when some comparison is to be made. This comparison may be a single observed value versus some hypothesized quantity (e.g. the number of babies born in a single delivery to mothers undergoing fertility treatment as compared with typical singleton birth), or it may be a comparison of two or more groups (e.g. mortality rates in intensive care unit patients who require renal replacement therapy versus those who do not). The choice of which statistical test to use depends on the format of the data and the study design. Examples of some of the more common techniques will be covered in subsequent reviews. However, the philosophy behind these statistical tests and the interpretation of the resulting *P* values are always the same, and it is these ideas that are covered in the present review.

The null hypothesis

A typical research question is most easily expressed in terms of there being some difference between groups. For example, 'In patients with acute myocardial infarction (AMI), does the administration of intravenous nitrate (as compared with none) reduce mortality?' To answer this question, the most appropriate study design would be a randomized controlled trial comparing AMI patients who receive intravenous nitrate with control patients. The challenge then is to interpret the results of that study. Even if there is no real effect of intravenous

nitrate on mortality, sampling variation means that it is extremely unlikely that exactly the same proportion of patients in each group will die. Thus, any observed difference between the two groups may be due to the treatment or it may simply be a coincidence, in other words due to chance. The aim of hypothesis testing is to establish which of these explanations is most likely. Note that statistical analyses can never prove the truth of a hypothesis, but rather merely provide evidence to support or refute it.

To do this, the research question is more formally expressed in terms of there being no difference. This is known as the null hypothesis. In the current example the null hypothesis would be expressed as, 'The administration of intravenous nitrate has no effect on mortality in AMI patients.'

In hypothesis testing any observed differences between two (or more) groups are interpreted within the context of this null hypothesis. More formally, hypothesis testing explores how likely it is that the observed difference would be seen by chance alone if the null hypothesis were true.

What is a *P* value?

There is a wide range of statistical tests available, depending on the nature of the investigation. However, the end result of any statistical test is a *P* value. The '*P*' stands for probability, and measures how likely it is that any observed difference between groups is due to chance. In other words, the *P* value is the probability of seeing the observed difference, or

Table 1**Results from six trials of intravenous nitrates in acute myocardial infarction patients**

Trial	Number dead/randomized		Odds ratio	95% confidence interval	<i>P</i> value
	Intravenous nitrate	Control			
Chiche	3/50	8/45	0.33	(0.09, 1.13)	0.08
Bussman	4/31	12/29	0.24	(0.08, 0.74)	0.01
Flaherty	11/56	11/48	0.83	(0.33, 2.12)	0.70
Jaffe	4/57	2/57	2.04	(0.39, 10.71)	0.40
Lis	5/64	10/76	0.56	(0.19, 1.65)	0.29
Jugdutt	24/154	44/156	0.48	(0.28, 0.82)	0.007

greater, just by chance if the null hypothesis is true. Being a probability, *P* can take any value between 0 and 1. Values close to 0 indicate that the observed difference is unlikely to be due to chance, whereas a *P* value close to 1 suggests there is no difference between groups other than that due to random variation. The interpretation of a *P* value is not always straightforward and several important factors must be taken into account, as outlined below. Put simply, however, the *P* value measures the strength of evidence against the null hypothesis.

Note that the aim of hypothesis testing is not to 'accept' or 'reject' the null hypothesis. Rather, it is simply to gauge how likely it is that the observed difference is genuine if the null hypothesis is true.

Interpreting *P* values

Continuing with the previous example, a number of trials of intravenous nitrates in patients with AMI have been carried out. In 1988 an overview of those that had been conducted at that time was performed in order to synthesize all the available evidence [1]. The results from six trials of intravenous nitrate are given in Table 1.

In the first trial (Chiche), 50 patients were randomly assigned to receive intravenous nitrate and 45 were randomly assigned to the control group. At the end of follow up, three of the 50 patients given intravenous nitrate had died versus eight in the control group. The calculation and interpretation of odds ratios will be covered in a future review. However, the interpretation in this context is that the odds ratio approximately represents the risk of dying in the nitrate group as compared with that in the control group. The odds ratio can take any positive value (above 0); in this context, values less than 1 indicate a protective effect of intravenous nitrate (a reduction in risk of death in patients administered intravenous nitrate), whereas an odds ratio greater than 1 points to a harmful effect (i.e. an increase in risk of death in patients administered intravenous nitrate). An odds ratio close to 1 is consistent with no effect of intravenous nitrate (i.e. no difference between the two groups). Interpretation of the confidence

intervals is just as described in Statistics review 2, with the first confidence interval (Chiche) indicating that the true odds ratio in the population from which the trial subjects were drawn is likely to be between 0.09 and 1.13.

Initially ignoring the confidence intervals, five of the six trials summarized in Table 1 have odds ratios that are consistent with a protective effect of intravenous nitrate (odds ratio <1). These range from a risk reduction of 17% (Flaherty) to one of 76% (Bussman). In other words, in the Bussman trial the risk of dying in the nitrate group is about one-quarter of that in the control group. The remaining trial (Jaffe) has an odds ratio of 2.04, suggesting that the effect of intravenous nitrate might be harmful, with a doubling of risk in patients given this treatment as compared with those in the control group.

The *P* values shown in the final column of Table 1 give an indication of how likely it is that these differences are simply due to chance. The *P* value for the first trial (Chiche) indicates that the probability of observing an odds ratio of 0.33 or more extreme, if the null hypothesis is true, is 0.08. In other words, if there is genuinely no effect of intravenous nitrate on the mortality of patients with AMI, then 8 out of 100 such trials would show a risk reduction of 66% or more just by chance. Equivalently, 2 out of 25 would show such a chance effect. The question of whether this is sufficiently unlikely to suggest that there is a real effect is highly subjective. However, it is unlikely that the management of critically ill patients would be altered on the basis of this evidence alone, and an isolated result such as this would probably be interpreted as being consistent with no effect. Similarly the *P* value for the Bussman trial indicates that 1 in 100 trials would have an odds ratio of 0.24 or more extreme by chance alone; this is a smaller probability than in the previous trial but, in isolation, perhaps still not sufficiently unlikely to alter clinical care in practice. The *P* value of 0.70 in the Flaherty trial suggests that the observed odds ratio of 0.83 is very likely to be a chance finding.

Comparing the *P* values across different trials there are two main features of interest. The first is that the size of the *P* value

is related, to some extent, to the size of the trial (and, in this context, the proportion of deaths). For example, the odds ratios in the Lis and Jugdutt trials are reasonably similar, both of which are consistent with an approximate halving of risk in patients given intravenous nitrate, but the P value for the larger Jugdutt trial is substantially smaller than that for the Lis trial. This pattern tends to be apparent in general, with larger studies giving rise to smaller P values. The second feature relates to how the P values change with the size of the observed effect. The Chiche and Flaherty trials have broadly similar numbers of patients (in fact, the numbers are somewhat higher in the Flaherty trial) but the smaller P value occurs in the Chiche study, which suggests that the effect of intravenous nitrate is much larger than that in the Flaherty study (67% versus 17% reduction in mortality). Again, this pattern will tend to hold in general, with more extreme effects corresponding to smaller P values. Both of these properties are discussed in considerably more detail in the next review, on sample size/power calculations.

There are two additional points to note when interpreting P values. It was common in the past for researchers to classify results as statistically 'significant' or 'non-significant', based on whether the P value was smaller than some prespecified cut point, commonly 0.05. This practice is now becoming increasingly obsolete, and the use of exact P values is much preferred. This is partly for practical reasons, because the increasing use of statistical software renders calculation of exact P values increasingly simple as compared with the past when tabulated values were used. However, there is also a more pragmatic reason for this shift. The use of a cut-off for statistical significance based on a purely arbitrary value such as 0.05 tends to lead to a misleading conclusion of accepting or rejecting the null hypothesis, in other words of concluding that a 'statistically significant' result is real in some sense. Recall that a P value of 0.05 means that one out of 20 studies would result in a difference at least as big as that observed just by chance. Thus, a researcher who accepts a 'significant' result as real will be wrong 5% of the time (this is sometimes known as a type I error). Similarly, dismissing an apparently 'non-significant' finding as a null result may also be incorrect (sometimes known as a type II error), particularly in a small study, in which the lack of statistical significance may simply be due to the small sample size rather than to any real lack of clinical effect (see the next review for details). Both of these scenarios have serious implications in terms of practical identification of risk factors and treatment of disease. The presentation of exact P values allows the researcher to make an educated judgement as to whether the observed effect is likely to be due to chance and this, taken in the context of other available evidence, will result in a far more informed conclusion being reached.

Finally, P values give no indication as to the clinical importance of an observed effect. For example, suppose a new drug for lowering blood pressure is tested against standard treatment, and the resulting P value is extremely small. This

indicates that the difference is unlikely to be due to chance, but decisions on whether to prescribe the new drug will depend on many other factors, including the cost of the new treatment, any potential contraindications or side effects, and so on. In particular, just as a small study may fail to detect a genuine effect, a very large study may result in a very small P value based on a small difference of effect that is unlikely to be important when translated into clinical practice.

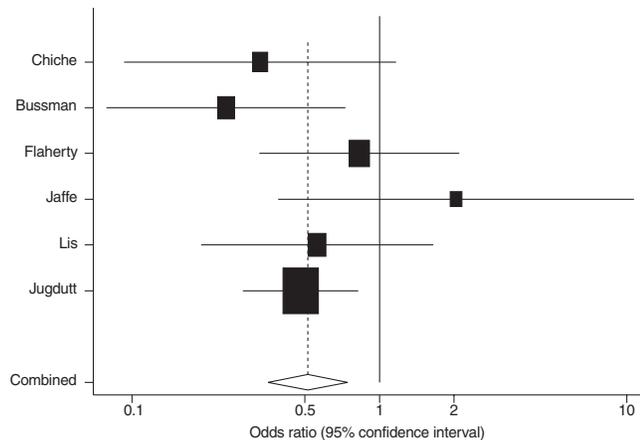
P values and confidence intervals

Although P values provide a measure of the strength of an association, there is a great deal of additional information to be obtained from confidence intervals. Recall that a confidence interval gives a range of values within which it is likely that the true population value lies. Consider the confidence intervals shown in Table 1. The odds ratio for the Chiche study is 0.33, suggesting that the effect of intravenous nitrate is to reduce mortality by two thirds. However, the confidence interval indicates that the true effect is likely to be somewhere between a reduction of 91% and an increase of 13%. The results from that study show that there may be a substantial reduction in mortality due to intravenous nitrate, but equally it is not possible to rule out an important increase in mortality. Clearly, if the latter were the case then it would be extremely dangerous to administer intravenous nitrate to patients with AMI.

The confidence interval for the Bussman study (0.08, 0.74) provides a rather more positive picture. It indicates that, although the reduction in mortality may be as little as 26%, there is little evidence to suggest that the effect of intravenous nitrate may be harmful. Administration of intravenous nitrate therefore appears more reasonable based on the results of that study, although the P value indicates a 1 in 100 probability that this may be a chance finding and so the result in isolation might not be sufficient evidence to change clinical practice.

The overview of those trials was carried out because the results did not appear to be consistent, largely because the individual trials were generally too small to provide reliable estimates of effect. A pooled analysis of the data from all of the nitrate trials shown in Table 1 (and including one other trial with no deaths) was therefore conducted to obtain a more robust estimate of effect (for details of the methods used, see Yusuf *et al.* [1]). The odds ratios and 95% confidence intervals for the individual trials in Table 1 are shown in Fig. 1. The odds ratio for each trial is represented by a box, the size of which is proportional to the amount of statistical information available for that estimate, and the 95% confidence interval is indicated by a horizontal line. The solid vertical line indicates an odds ratio of 1.0; in other words it shows the line of 'no effect'. The combined odds ratio from all six trials is indicated by the dashed vertical line, and its associated 95% confidence interval by the diamond at the bottom.

This pooled analysis resulted in an estimated overall odds ratio of 0.53 with a 95% confidence interval of (0.36, 0.75),

Figure 1

Individual and combined odds ratios and 95% confidence intervals for six intravenous nitrate trials.

suggesting a true reduction in mortality of somewhere between one-quarter and two-thirds. Examination of the confidence intervals from individual studies shows a high degree of overlap with the pooled confidence interval, and so all of the evidence appears to be consistent with this pooled estimate; this includes the evidence from the Jaffe study, which, at first glance, appears to suggest a harmful effect. The P value for the pooled analysis was 0.0002, which indicates that the result is extremely unlikely to have been due to chance.

Note that, since that meta-analysis was reported, treatment of AMI patients has changed dramatically with the introduction of thrombolysis. In addition, the Fourth International Study of Infarct Survival (ISIS-4) [2], which randomized over 58,000 patients with suspected AMI, found no evidence to suggest that mortality was reduced in those given oral nitrates. Thus, in practice the indications for intravenous nitrates in patients with AMI are restricted to symptom and blood pressure control.

Specific methods for comparing two or more means or proportions will be introduced in subsequent reviews. In general, these will tend to focus on the calculation of P values. However, there is still much to be learned from examination of confidence intervals in this context. For example, when comparing the risk for developing secondary infection following trauma in patients with or without a history of chronic alcohol abuse, it may be enlightening to compare the confidence intervals for the two groups and to examine the extent to which they do or do not overlap. Alternatively, it is possible to calculate a confidence interval for the difference in two means or the difference or ratio of proportions directly. This can also give a useful indication of the likely effect of chronic alcohol abuse, in particular by exploring the extent to which the range of likely values includes or excludes 0 or 1, the respective expected values of a difference or ratio if there is

no effect of chronic alcohol abuse, or in other words under the null hypothesis.

Although P values provide a measure of the strength of an association, an estimate of the size of any effect along with an associated confidence interval is always required for meaningful interpretation of results. P values and confidence intervals are frequently calculated using similar quantities (see subsequent reviews for details), and so it is not surprising that the two are closely related. In particular, larger studies will in general result in narrower confidence intervals and smaller P values, and this should be taken into account when interpreting the results from statistical analyses. Both P values and confidence intervals have an important role to play in understanding data analyses, and both should be presented wherever possible.

Key messages

A P value is the probability that an observed effect is simply due to chance; it therefore provides a measure of the strength of an association. A P value does not provide any measure of the size of an effect, and cannot be used in isolation to inform clinical judgement.

P values are affected both by the magnitude of the effect and by the size of the study from which they are derived, and should therefore be interpreted with caution. In particular, a large P value does not always indicate that there is no association and, similarly, a small P value does not necessarily signify an important clinical effect.

Subdividing P values into 'significant' and 'non-significant' is poor statistical practice and should be avoided. Exact P values should always be presented, along with estimates of effect and associated confidence intervals.

Competing interests

None declared.

References

1. Yusuf S, Collins R, MacMahon S, Peto R: **Effect of intravenous nitrates on mortality in acute myocardial infarction: an overview of the randomised trials.** *Lancet* 1988, **1**:1088-1092.
2. Anonymous: **ISIS-4: a randomised factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphate in 58,050 patients with suspected acute myocardial infarction.** *Lancet* 1995, **345**:669-685.
3. Whitley E, Ball J: **Statistics review 1: Presenting and summarising data.** *Crit Care* 202, **6**:66-71.
4. Whitley E, Ball J: **Statistics review 2: Samples and populations.** *Crit Care* 202, **6**:143-148.

This article is the third in an ongoing, educational review series on medical statistics in critical care. Previous articles have covered 'presenting and summarising data' [3] and 'samples and populations' [4]. Future topics to be covered include power calculations, comparison of means, comparison of proportions, and analysis of survival data to name but a few. If there is a medical statistics topic you would like explained, contact us on editorial@ccforum.com.

Review

Statistics review 4: Sample size calculations

Elise Whitley¹ and Jonathan Ball²

¹Lecturer in Medical Statistics, University of Bristol, Bristol, UK

²Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 10 May 2002

Critical Care 2002, **6**:335-341

This article is online at <http://ccforum.com/content/6/4/335>

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review introduces the notion of statistical power and the hazard of under-powered studies. The problem of how to calculate an ideal sample size is also discussed within the context of factors that affect power, and specific methods for the calculation of sample size are presented for two common scenarios, along with extensions to the simplest case.

Keywords statistical power, sample size

Previous reviews in this series introduced confidence intervals and P values. Both of these have been shown to depend strongly on the size of the study sample in question, with larger samples generally resulting in narrower confidence intervals and smaller P values. The question of how large a study should ideally be is therefore an important one, but it is all too often neglected in practice. The present review provides some simple guidelines on how best to choose an appropriate sample size.

Research studies are conducted with many different aims in mind. A study may be conducted to establish the difference, or conversely the similarity, between two groups defined in terms of a particular risk factor or treatment regimen. Alternatively, it may be conducted to estimate some quantity, for example the prevalence of disease, in a specified population with a given degree of precision. Regardless of the motivation for the study, it is essential that it be of an appropriate size to achieve its aims. The most common aim is probably that of determining some difference between two groups, and it is this scenario that will be used as the basis for the remainder of the present review. However, the ideas underlying the methods described are equally applicable to all settings.

Power

The difference between two groups in a study will usually be explored in terms of an estimate of effect, appropriate confidence interval and P value. The confidence interval indicates the likely range of values for the true effect in the population,

while the P value determines how likely it is that the observed effect in the sample is due to chance. A related quantity is the statistical power of the study. Put simply, this is the probability of correctly identifying a difference between the two groups in the study sample when one genuinely exists in the populations from which the samples were drawn.

The ideal study for the researcher is one in which the power is high. This means that the study has a high chance of detecting a difference between groups if one exists; consequently, if the study demonstrates no difference between groups the researcher can be reasonably confident in concluding that none exists in reality. The power of a study depends on several factors (see below), but as a general rule higher power is achieved by increasing the sample size.

It is important to be aware of this because all too often studies are reported that are simply too small to have adequate power to detect the hypothesized effect. In other words, even when a difference exists in reality it may be that too few study subjects have been recruited. The result of this is that P values are higher and confidence intervals wider than would be the case in a larger study, and the erroneous conclusion may be drawn that there is no difference between the groups. This phenomenon is well summed up in the phrase, 'absence of evidence is not evidence of absence'. In other words, an apparently null result that shows no difference between groups may simply be due to lack of statistical power, making it extremely unlikely that a true difference will be correctly identified.

Given the importance of this issue, it is surprising how often researchers fail to perform any systematic sample size calculations before embarking on a study. Instead, it is not uncommon for decisions of this sort to be made arbitrarily on the basis of convenience, available resources, or the number of easily available subjects. A study by Moher and coworkers [1] reviewed 383 randomized controlled trials published in three journals (*Journal of the American Medical Association*, *Lancet* and *New England Journal of Medicine*) in order to examine the level of statistical power in published trials with null results. Out of 102 null trials, those investigators found that only 36% had 80% power to detect a relative difference of 50% between groups and only 16% had 80% power to detect a more modest 25% relative difference. (Note that a smaller difference is more difficult to detect and requires a larger sample size; see below for details.) In addition, only 32% of null trials reported any sample size calculations in the published report. The situation is slowly improving, and many grant giving bodies now require sample size calculations to be provided at the application stage. Many under-powered studies continue to be published, however, and it is important for readers to be aware of the problem.

Finally, although the most common criticism of the size, and hence the power, of a study is that it is too low, it is also worth noting the consequences of having a study that is too large. As well as being a waste of resources, recruiting an excessive number of participants may be unethical, particularly in a randomized controlled trial where an unnecessary doubling of the sample size may result in twice as many patients receiving placebo or potentially inferior care, as is necessary to establish the worth of the new therapy under consideration.

Factors that affect sample size calculations

It is important to consider the probable size of study that will be required to achieve the study aims at the design stage. The calculation of an appropriate sample size relies on a subjective choice of certain factors and sometimes crude estimates of others, and may as a result seem rather artificial. However, it is at worst a well educated guess, and is considerably more useful than a completely arbitrary choice. There are three main factors that must be considered in the calculation of an appropriate sample size, as summarized in Table 1. The choice of each of these factors impacts on the final sample size, and the skill is in combining realistic values for each of these in order to achieve an attainable sample size. The ultimate aim is to conduct a study that is large enough to ensure that an effect of the size expected, if it exists, is sufficiently likely to be identified.

Although, as described in Statistics review 3, it is generally bad practice to choose a cutoff for statistical ‘significance’ based on *P* values, it is a convenient approach in the calculation of sample size. A conservative cutoff for significance, as indicated by a small *P* value, will reduce the risk of incorrectly

Table 1

Factors that affect sample size calculations			
Factor	Magnitude	Impact on identification of effect	Required sample size
<i>P</i> value	Small	Stringent criterion; difficult to achieve ‘significance’	Large
	Large	Relaxed criterion; ‘significance’ easier to attain	Small
Power	Low	Identification unlikely	Small
	High	Identification more probable	Large
Effect	Small	Difficult to identify	Large
	Large	Easy to identify	Small

interpreting a chance finding as genuine. However, in practice this caution is reflected in the need for a larger sample size in order to obtain a sufficiently small *P* value. Similarly, a study with high statistical power will, by definition, make identification of any difference relatively easy, but this can only be achieved in a sufficiently large study. In practice there are conventional choices for both of these factors; the *P* value for significance is most commonly set at 0.05, and power will generally be somewhere between 80% and 95%, depending on the resulting sample size.

The remaining factor that must be considered is the size of the effect to be detected. However, estimation of this quantity is not always straightforward. It is a crucial factor, with a small effect requiring a large sample and *vice versa*, and careful consideration should be given to the choice of value. Ideally, the size of the effect will be based on clinical judgement. It should be large enough to be clinically important but not so large that it is implausible. It may be tempting to err on the side of caution and to choose a small effect; this may well cover all important clinical scenarios but will be at the cost of substantially (and potentially unnecessarily) increasing the sample size. Alternatively, an optimistic estimate of the probable impact of some new therapy may result in a small calculated sample size, but if the true effect is less impressive than expected then the resulting study will be under-powered, and a smaller but still important effect may be missed.

Once these three factors have been established, there are tabulated values [2] and formulae available for calculating the required sample size. Certain outcomes and more complex study designs may require further information, and calculation of the required sample size is best left to someone with appropriate expertise. However, specific methods for two common situations are detailed in the following sections.

Note that the sample sizes obtained from these methods are intended as approximate guides rather than exact numbers. In

other words a calculation indicating a sample size of 100 will generally rule out the need for a study of size 500 but not one of 110; a sample size of 187 can be usefully rounded up to 200, and so on. In addition, the results of a sample size calculation are entirely dependent on estimates of effect, power and significance, as discussed above. Thus, a range of values should be incorporated into any good investigation in order to give a range of suitable sample sizes rather than a single 'magic' number.

Sample size calculation for a difference in means (equal sized groups)

Let us start with the simplest case of two equal sized groups. A recently published trial [3] considered the effect of early goal-directed versus traditional therapy in patients with severe sepsis or septic shock. In addition to mortality (the primary outcome on which the study was originally powered), the investigators also considered a number of secondary outcomes, including mean arterial pressure 6 hours after the start of therapy. Mean arterial pressure was 95 and 81 mmHg in the groups treated with early goal-directed and traditional therapy, respectively, corresponding to a difference of 14 mmHg.

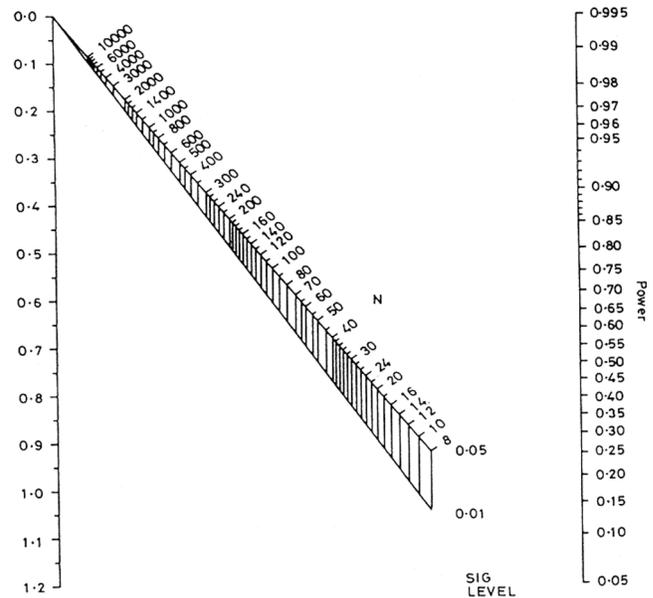
The first step in calculating a sample size for comparing means is to consider this difference in the context of the inherent variability in mean arterial pressure. If the means are based on measurements with a high degree of variation, for example with a standard deviation of 40 mmHg, then a difference of 14 mmHg reflects a relatively small treatment effect compared with the natural spread of the data, and may well be unremarkable. Conversely, if the standard deviation is extremely small, say 3 mmHg, then an absolute difference of 14 mmHg is considerably more important. The target difference is therefore expressed in terms of the standard deviation, known as the standardized difference, and is defined as follows:

$$\text{Standardized difference} = \frac{\text{Target difference}}{\text{Standard deviation}} \quad (1)$$

In practice the standard deviation is unlikely to be known in advance, but it may be possible to estimate it from other similar studies in comparable populations, or perhaps from a pilot study. Again, it is important that this quantity is estimated realistically because an overly conservative estimate at the design stage may ultimately result in an under-powered study.

In the current example the standard deviation for the mean arterial pressure was approximately 18 mmHg, so the standardized difference to be detected, calculated using equation 1, was $14/18 = 0.78$. There are various formulae and tabulated values available for calculating the desired sample size in this situation, but a very straightforward approach is provided by Altman [4] in the form of the nomogram shown in Fig. 1 [5].

Figure 1



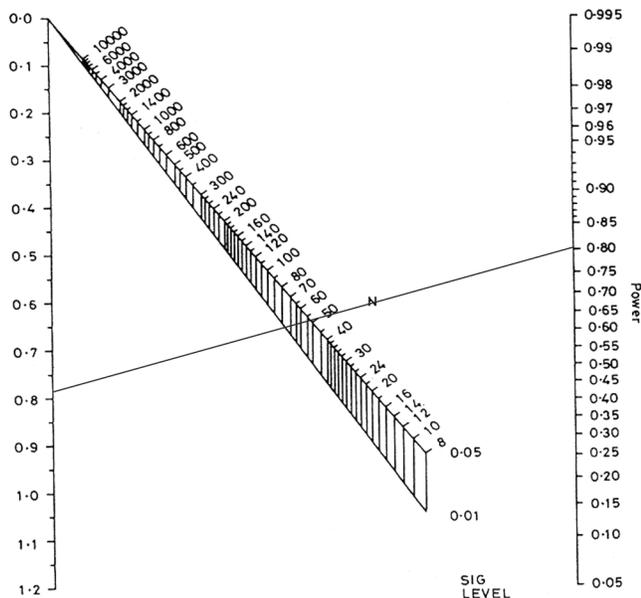
Nomogram for calculating sample size or power. Reproduced from Altman [5], with permission.

The left-hand axis in Fig. 1 shows the standardized difference (as calculated using Eqn 1, above), while the right-hand axis shows the associated power of the study. The total sample size required to detect the standardized difference with the required power is obtained by drawing a straight line between the power on the right-hand axis and the standardized difference on the left-hand axis. The intersection of this line with the upper part of the nomogram gives the sample size required to detect the difference with a *P* value of 0.05, whereas the intersection with the lower part gives the sample size for a *P* value of 0.01. Fig. 2 shows the required sample sizes for a standardized difference of 0.78 and desired power of 0.8, or 80%. The total sample size for a trial that is capable of detecting a 0.78 standardized difference with 80% power using a cutoff for statistical significance of 0.05 is approximately 52; in other words, 26 participants would be required in each arm of the study. If the cutoff for statistical significance were 0.01 rather than 0.05 then a total of approximately 74 participants (37 in each arm) would be required.

The effect of changing from 80% to 95% power is shown in Fig. 3. The sample sizes required to detect the same standardized difference of 0.78 are approximately 86 (43 per arm) and 116 (58 per arm) for *P* values of 0.05 and 0.01, respectively.

The nomogram provides a quick and easy method for determining sample size. An alternative approach that may offer more flexibility is to use a specific sample size formula. An appropriate formula for comparing means in two groups of equal size is as follows:

Figure 2



Nomogram showing sample size calculation for a standardized difference of 0.78 and 80% power.

$$n = \frac{2}{d^2} \times c_{p,power} \tag{2}$$

where n is the number of subjects required in each group, d is the standardized difference and $c_{p,power}$ is a constant defined by the values chosen for the P value and power. Some commonly used values for $c_{p,power}$ are given in Table 2. The number of participants required in each arm of a trial to detect a standardized difference of 0.78 with 80% power using a cutoff for statistical significance of 0.05 is as follows:

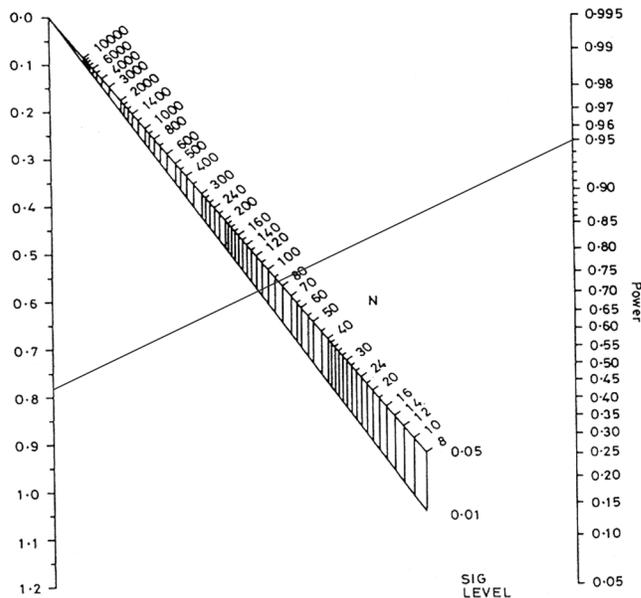
$$\begin{aligned} n &= \frac{2}{0.78^2} \times c_{0.05,80\%} \\ &= \frac{2}{0.6084} \times 7.9 \\ &= 2.39 \times 7.9 \\ &= 26.0 \end{aligned}$$

Thus, 26 participants are required in each arm of the trial, which agrees with the estimate provided by the nomogram.

Sample size calculation for a difference in proportions (equal sized groups)

A similar approach can be used to calculate the sample size required to compare proportions in two equally sized groups.

Figure 3



Nomogram showing sample size calculation for a standardized difference of 0.78 and 95% power.

Table 2

Commonly used values for $c_{p,power}$

P	Power (%)			
	50	80	90	95
0.05	3.8	7.9	10.5	13.0
0.01	6.6	11.7	14.9	17.8

In this case the standardized difference is given by the following equation:

$$\text{Standardized difference} = \frac{(p_1 - p_2)}{\sqrt{[\bar{p}(1 - \bar{p})]}} \tag{3}$$

where p_1 and p_2 are the proportions in the two groups and $\bar{p} = (p_1 + p_2)/2$ is the mean of the two values. Once the standardized difference has been calculated, the nomogram shown in Fig. 1 can be used in exactly the same way to determine the required sample size.

As an example, consider the recently published Acute Respiratory Distress Syndrome Network trial of low versus traditional tidal volume ventilation in patients with acute lung injury and acute respiratory distress syndrome [6]. Mortality rates in the low and traditional volume groups were 31% and 40%, respectively, corresponding to a reduction of 9% in the low

volume group. What sample size would be required to detect this difference with 90% power using a cutoff for statistical significance of 0.05? The mean of the two proportions in this case is 35.5% and the standardized difference is therefore as follows (calculated using Eqn 3).

$$\frac{(0.40 - 0.31)}{\sqrt{[0.355(1 - 0.355)]}} = \frac{0.09}{0.479} = 0.188$$

Fig. 4 shows the required sample size, estimated using the nomogram to be approximately 1200 in total (i.e. 600 in each arm).

Again, there is a formula that can be used directly in these circumstances. Comparison of proportions p_1 and p_2 in two equally sized groups requires the following equation:

$$n = \frac{[p_1(1 - p_1) + p_2(1 - p_2)]}{(p_1 - p_2)^2} \times C_{p,power} \quad (4)$$

where n is the number of subjects required in each group and $C_{p,power}$ is as defined in Table 2. Returning to the example of the Acute Respiratory Distress Syndrome Network trial, the formula indicates that the following number of patients would be required in each arm.

$$\frac{(0.31 \times 0.69) + (0.40 \times 0.60)}{(0.31 - 0.40)^2} \times 10.5 = 588.4$$

This estimate is in accord with that obtained from the nomogram.

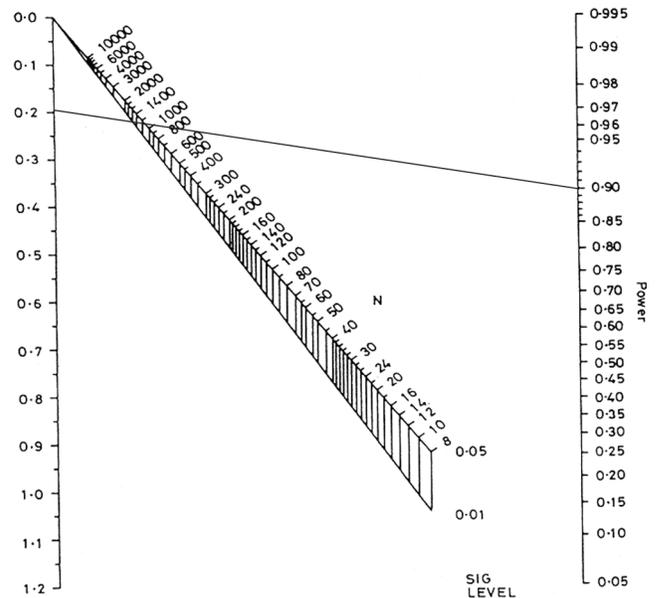
Calculating power

The nomogram can also be used retrospectively in much the same way to calculate the power of a published study. The Acute Respiratory Distress Syndrome Network trial stopped after enrolling 861 patients. What is the power of the published study to detect a standardized difference in mortality of 0.188, assuming a cutoff for statistical significance of 0.05?

The patients were randomized into two approximately equal sized groups (432 and 429 receiving low and traditional tidal volumes, respectively), so the nomogram can be used directly to estimate the power. (For details on how to handle unequally sized groups, see below.) The process is similar to that for determining sample size, with a straight line drawn between the standardized difference and the sample size extended to show the power of the study. This is shown for the current example in Fig. 5, in which a (solid) line is drawn between a standardized difference of 0.188 and an approximate sample size of 861, and is extended (dashed line) to indicate a power of around 79%.

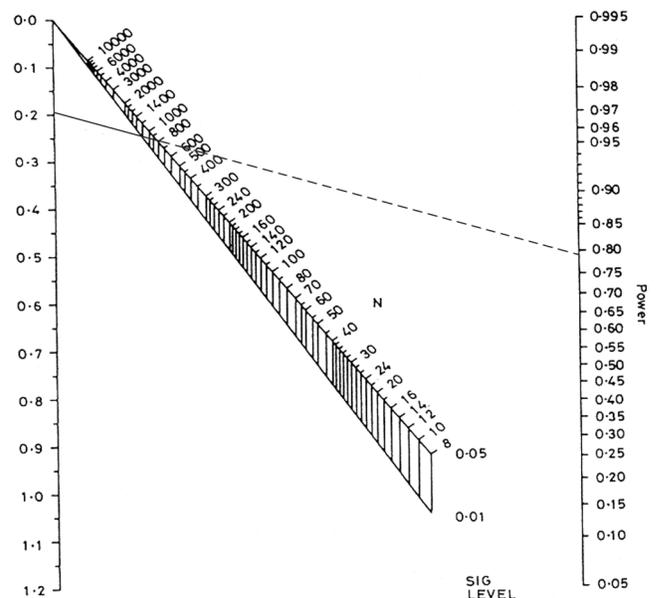
It is also possible to use the nomogram in this way when financial or logistical constraints mean that the ideal sample

Figure 4



Nomogram showing sample size calculation for standardized difference of 0.188 and 90% power.

Figure 5



Nomogram showing the statistical power for a standardized difference of 0.188 and a total sample size of 861.

size cannot be achieved. In this situation, use of the nomogram may enable the investigator to establish what power might be achieved in practice and to judge whether the loss of power is sufficiently modest to warrant continuing with the study.

As an additional example, consider data from a published trial of the effect of prone positioning on the survival of patients with acute respiratory failure [7]. That study recruited a total of 304 patients into the trial and randomized 152 to conventional (supine) positioning and 152 to a prone position for 6 h or more per day. The trial found that patients placed in a prone position had improved oxygenation but that this was not reflected in any significant reduction in survival at 10 days (the primary end-point).

Mortality rates at 10 days were 21% and 25% in the prone and supine groups, respectively. Using equation 3, this corresponds to a standardized difference of the following:

$$\frac{(0.25 - 0.21)}{\sqrt{[0.23(1 - 0.23)]}} = \frac{0.04}{0.421} = 0.095$$

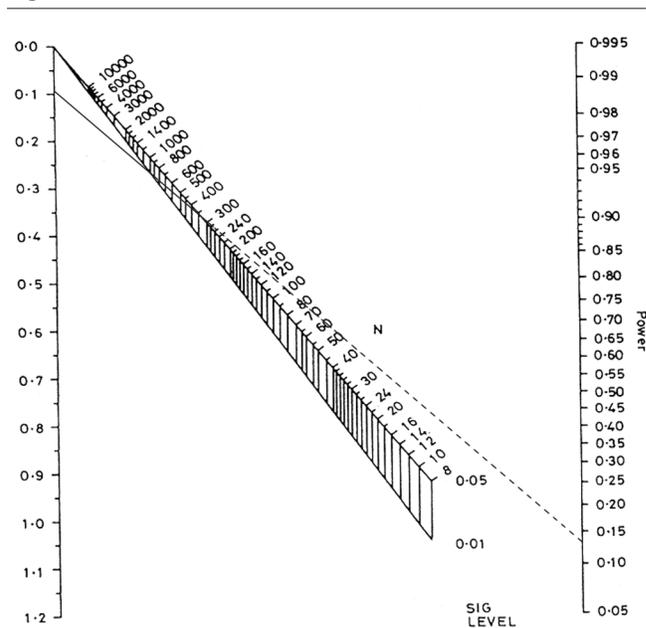
This is comparatively modest and is therefore likely to require a large sample size to detect such a difference in mortality with any confidence. Fig. 6 shows the appropriate nomogram, which indicates that the published study had only approximately 13% power to detect a difference of this size using a cutoff for statistical significance of 0.05. In other words even if, in reality, placing patients in a prone position resulted in an important 4% reduction in mortality, a trial of 304 patients would be unlikely to detect it in practice. It would therefore be dangerous to conclude that positioning has no effect on mortality without corroborating evidence from another, larger trial. A trial to detect a 4% reduction in mortality with 80% power would require a total sample size of around 3500 (i.e. approximately 1745 patients in each arm). However, a sample size of this magnitude may well be impractical. In addition to being dramatically under-powered, that study has been criticized for a number of other methodological/design failings [8,9]. Sadly, despite the enormous effort expended, no reliable conclusions regarding the efficacy of prone positioning in acute respiratory distress syndrome can be drawn from the trial.

Unequal sized groups

The methods described above assume that comparison is to be made across two equal sized groups. However, this may not always be the case in practice, for example in an observational study or in a randomized controlled trial with unequal randomization. In this case it is possible to adjust the numbers to reflect this inequality. The first step is to calculate the total sample size (across both groups) assuming that the groups are equal sized (as described above). This total sample size (N) can then be adjusted according to the actual ratio of the two groups (k) with the revised total sample size (N') equal to the following:

$$N' = \frac{N(1 + k)^2}{4k} \tag{5}$$

Figure 6



Nomogram showing the statistical power for a standardized difference of 0.095 and a total sample size of 304.

and the individual sample sizes in each of the two groups are $N'/(1 + k)$ and $kN'/(1 + k)$.

Returning to the example of the Acute Respiratory Distress Syndrome Network trial, suppose that twice as many patients were to be randomized to the low tidal volume group as to the traditional group, and that this inequality is to be reflected in the study size. Fig. 4 indicates that a total of 1200 patients would be required to detect a standardized difference of 0.188 with 90% power. In order to account for the ratio of low to traditional volume patients ($k=2$), the following number of patients would be required.

$$N' = \frac{1200 \times (1 + 2)^2}{4 \times 2} = \frac{1200 \times 9}{8} = 1350$$

This comprises $1350/3 = 450$ patients randomized to traditional care and $(2 \times 1350)/3 = 900$ to low tidal volume ventilation.

Withdrawals, missing data and losses to follow up

Any sample size calculation is based on the total number of subjects who are needed in the final study. In practice, eligible subjects will not always be willing to take part and it will be necessary to approach more subjects than are needed in the first instance. In addition, even in the very best designed and conducted studies it is unusual to finish with a dataset in which complete data are available in a usable format for every

subject. Subjects may fail or refuse to give valid responses to particular questions, physical measurements may suffer from technical problems, and in studies involving follow up (e.g. trials or cohort studies) there will always be some degree of attrition. It may therefore be necessary to calculate the number of subjects that need to be approached in order to achieve the final desired sample size.

More formally, suppose a total of N subjects is required in the final study but a proportion (q) are expected to refuse to participate or to drop out before the study ends. In this case the following total number of subjects would have to be approached at the outset to ensure that the final sample size is achieved:

$$N'' = \frac{N}{(1 - q)} \quad (6)$$

For example, suppose that 10% of subjects approached in the early goal-directed therapy trial described above are expected to refuse to participate. Then, considering the effect on mean arterial pressure and assuming a P for statistical significance of 0.05 and 80% power, the following total number of eligible subjects would have to be approached in the first instance:

$$N'' = \frac{52}{(1 - 0.1)} = \frac{52}{0.9} = 57.8$$

Thus, around 58 eligible subjects (approximately 29 in each arm) would have to be approached in order to ensure the required final sample size of 52 is achieved.

As with other aspects of sample size calculations, the proportion of eligible subjects who will refuse to participate or provide inadequate information will be unknown at the onset of the study. However, good estimates will often be possible using information from similar studies in comparable populations or from an appropriate pilot study. Note that it is particularly important to account for nonparticipation in the costing of studies in which initial recruitment costs are likely to be high.

Key messages

Studies must be adequately powered to achieve their aims, and appropriate sample size calculations should be carried out at the design stage of any study.

Estimation of the expected size of effect can be difficult and should, wherever possible, be based on existing evidence and clinical expertise. It is important that any estimates be large enough to be clinically important while also remaining plausible.

Many apparently null studies may be under-powered rather than genuinely demonstrating no difference between groups; absence of evidence is not evidence of absence.

This article is the fourth in an ongoing, educational review series on medical statistics in critical care. Previous articles have covered 'presenting and summarizing data', 'samples and populations' and 'hypotheses testing and P values'. Future topics to be covered include comparison of means, comparison of proportions and analysis of survival data, to name but a few. If there is a medical statistics topic you would like explained, contact us on editorial@ccforum.com.

Competing interests

None declared.

References

1. Moher D, Dulberg CS, Wells GA: **Statistical power, sample size, and their reporting in randomized controlled trials.** *JAMA* 1994, **272**:122-124.
2. Machin D, Campbell MJ, Fayers P, Pinol A: *Sample Size Tables for Clinical Studies.* Oxford, UK: Blackwell Science Ltd; 1987.
3. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M: **Early goal-directed therapy in the treatment of severe sepsis and septic shock.** *N Engl J Med* 2001, **345**:1368-1377.
4. Altman DG: *Practical Statistics for Medical Research.* London, UK; Chapman & Hall; 1991.
5. Altman D.G. **How large a sample?** In: Gore SM, Altman DG (eds). *Statistics in Practice.* London, UK: British Medical Association; 1982.
6. Anonymous: **Ventilation with lower tidal volumes as compared with traditional tidal volumes for acute lung injury and the acute respiratory distress syndrome. The Acute Respiratory Distress Syndrome Network.** *N Engl J Med* 2000, **342**:1301-1308.
7. Gattinoni L, Tognoni G, Pesenti A, Taccone P, Mascheroni D, Labarta V, Malacrida R, Di Giulio P, Fumagalli R, Pelosi P, Brazzi L, Latini R; Prone-Supine Study Group: **Effect of prone positioning on the survival of patients with acute respiratory failure.** *N Engl J Med* 2001, **345**:568-573.
8. Zijlstra JG, Ligtenberg JJ, van der Werf TS: **Prone positioning of patients with acute respiratory failure.** *N Engl J Med* 2002, **346**:295-297.
9. Slutsky AS: **The acute respiratory distress syndrome, mechanical ventilation, and the prone position.** *N Engl J Med* 2001, **345**:610-612.

Review

Statistics review 5: Comparison of means

Elise Whitley¹ and Jonathan Ball²

¹Lecturer in Medical Statistics, University of Bristol, Bristol, UK

²Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UK

Correspondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 12 July 2002

Critical Care 2002, **6**:424-428

This article is online at <http://ccforum.com/content/6/5/424>

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review introduces the commonly used t-test, used to compare a single mean with a hypothesized value, two means arising from paired data, or two means arising from unpaired data. The assumptions underlying these tests are also discussed.

Keywords comparison of two means, paired and unpaired data, t test

Previous reviews in this series have introduced the principals behind the calculation of confidence intervals and hypothesis testing. The present review covers the specific case of comparing means in rather more detail. Comparison of means arises in many different formats, and there are various methods available for dealing with each of these. Some of the simpler cases are covered in this review, namely comparison of a single observed mean with some hypothesized value, comparison of two means arising from paired data, and comparison of two means from unpaired data. All of these comparisons can be made using appropriate confidence intervals and t-tests as long as certain assumptions are met (see below). Future reviews will introduce techniques that can be used when the assumptions of the t-test are not valid or when the comparison is between three or more groups.

Of the three cases covered in this review, comparison of means from unpaired data is probably the most common. However, the single mean and paired data cases are introduced first because the t-test in these cases is more straightforward.

Comparison of a single mean with a hypothesized value

This situation is not very common in practice but on occasion it may be desirable to compare a mean value from a sample with some hypothesized value, perhaps from external standards. As an example, consider the data shown in Table 1. These are the haemoglobin concentrations of 15 UK adult

males admitted into an intensive care unit (ICU). The population mean haemoglobin concentration in UK males is 15.0 g/dl. Is there any evidence that critical illness is associated with an acute anaemia?

The mean haemoglobin concentration of these men is 9.7 g/dl, which is lower than the population mean. However, in practice any sample of 15 men would be unlikely to have a mean haemoglobin of exactly 15.0 g/dl, so the question is whether this difference is likely to be a chance finding, due to random variation, or whether it is the result of some systematic difference between the men in the sample and those in the general population. The best way to determine which explanation is most likely is to calculate a confidence interval for the mean and to perform a hypothesis test.

The standard deviation (SD) of these data is 2.2 g/dl, and so a 95% confidence interval for the mean can be calculated using the standard error (SE) in the usual way. The SE in this case is $2.2/\sqrt{15} = 0.56$ and the corresponding 95% confidence interval is as follows.

$$9.7 \pm 2.14 \times 0.56 = 9.7 \pm 1.19 = (8.5, 10.9)$$

Note that the multiplier, in this case 2.14, comes from the t distribution because the sample size is small (for a fuller explanation of this calculation, see Statistics review 2 from this series). This confidence interval gives the range of likely values for the mean haemoglobin concentration in the population

Table 1**Haemoglobin concentrations (g/dl) for 15 UK males admitted into an intensive care unit**

8.1	10.1	12.3
9.7	11.7	11.3
11.9	9.3	13.0
10.5	8.3	8.8
9.4	6.4	5.4

from which these men were drawn. In other words, assuming that this sample is representative, it is likely that the true mean haemoglobin in the population of adult male patients admitted to ICUs is between 8.5 and 10.9 g/dl. The haemoglobin concentration in the general population of adult men in the UK is well outside this range, and so the evidence suggests that men admitted to ICUs may genuinely have haemoglobin concentrations that are lower than the national average.

Exploration of how likely it is that this difference is due to chance requires a hypothesis test, in this case the one sample t-test. The t-test formally examines how far the estimated mean haemoglobin of men admitted to ICU, in this case 9.7 g/dl, lies from the hypothesized value of 15.0 g/dl. The null hypothesis is that the mean haemoglobin concentration of men admitted to ICU is the same as the standard for the adult male UK population, and so the further away the sample mean is from this hypothesized value, the less likely it is that the difference arose by chance.

The t statistic, from which a *P* value is derived, is as follows.

$$t = \frac{\text{sample mean} - \text{hypothesized mean}}{\text{SE of sample mean}} \quad (1)$$

In other words, *t* is the number of SEs that separate the sample mean from the hypothesized value. The associated *P* value is obtained by comparison with the *t* distribution introduced in Statistics review 2, with larger *t* statistics (regardless of sign) corresponding to smaller *P* values. As previously described, the shape of the *t* distribution is determined by the degrees of freedom, which, in the case of the one sample t-test, is equal to the sample size minus 1.

The *t* statistic for the haemoglobin example is as follows.

$$t = \frac{9.7 - 15.0}{0.56} = \frac{-5.3}{0.56} = -9.54$$

In other words, the observed mean haemoglobin concentration is 9.54 SEs below the hypothesized mean. Tabulated

Table 2**Central venous oxygen saturation on admission and 6 h after admission to an intensive care unit**

Subject	Central venous oxygen saturation (%)		
	On admission	6 h after admission	Difference (%)
1	39.7	52.9	13.2
2	59.1	56.7	-2.4
3	56.1	61.9	5.8
4	57.7	71.4	13.7
5	60.6	67.7	7.1
6	37.8	50.0	12.2
7	58.2	60.7	2.5
8	33.6	51.3	17.7
9	56.0	59.5	3.5
10	65.3	59.8	-5.5
Mean	52.4	59.2	6.8

values indicate how likely this is to occur in practice, and for a sample size of 15 (corresponding to 14 degrees of freedom) the *P* value is less than 0.0001. In other words, it is extremely unlikely that the mean haemoglobin in this sample would differ from that in the general population to this extent by chance alone. This may indicate that there is a genuine difference in haemoglobin concentrations in men admitted to the ICU, but as always it is vital that this result be interpreted in context. For example, it is important to know how this sample of men was selected and whether they are representative of all UK men admitted to ICUs.

Note that the *P* value gives no indication of the size of any difference; it merely indicates the probability that the difference arose by chance. In order to assess the magnitude of any difference, it is essential also to have the confidence interval calculated above.

Comparison of two means arising from paired data

A special case of the one sample t-test arises when paired data are used. Paired data arise in a number of different situations, such as in a matched case-control study in which individual cases and controls are matched to each other, or in a repeat measures study in which some measurement is made on the same set of individuals on more than one occasion (generally under different circumstances). For example, Table 2 shows central venous oxygen saturation in 10 patients on admission and 6 hours after admission to an ICU.

The mean admission central venous oxygen saturation was 52.4% as compared with a mean of 59.2% after 6 hours, cor-

responding to an increase of 6.8%. Again, the question is whether this difference is likely to reflect a genuine effect of admission and treatment or whether it is simply due to chance. In other words, the null hypothesis is that the mean central venous oxygen saturation on admission is the same as the mean saturation after 6 hours. However, because the data are paired, the two sets of observations are not independent of each other, and it is important to account for this pairing in the analysis. The way to do this is to concentrate on the differences between the pairs of measurements rather than on the measurements themselves.

The differences between the admission and post-admission central venous oxygen saturations are given in the rightmost column of Table 2, and the mean of these differences is 6.8%. In these terms, the null hypothesis is that the mean of the differences in central venous oxygen saturation is zero. The appropriate t-test therefore compares the observed mean of the differences with a hypothesized value of 0. In other words, the paired t-test is simply a special case of the single sample t-test described above.

The t statistic for the paired t-test is as follows.

$$t = \frac{\text{sample mean of differences} - 0}{\text{SE of sample mean of differences}}$$

$$= \frac{\text{sample mean of differences}}{\text{SE of sample mean of differences}} \quad (2)$$

The SD of the differences in the current example is 7.5, and this corresponds to a SE of $7.5/\sqrt{10} = 2.4$. The t statistic is therefore $t = 6.8/2.4 = 2.87$, and this corresponds to a P value of 0.02 (based on a t distribution with $10 - 1 = 9$ degrees of freedom). In other words, there is some evidence to suggest that admission to ICU and subsequent treatment may increase central venous oxygen saturation beyond the level expected by chance.

However, the P value in isolation gives no information about the likely size of any effect. As indicated above, this is rectified by calculating a 95% confidence interval from the mean and SE of the differences. In this case the 95% confidence interval is as follows.

$$6.8 \pm 2.26 \times 2.4 = 6.8 \pm 5.34 = (1.4, 12.2)$$

This indicates that the true increase in central venous oxygen saturation due to ICU admission and treatment in the population is probably between 1.4% and 12.2%. The decision as to whether this difference is likely to be important in practice should be based on the statistical evidence in combination with other relevant clinical factors. However, it is worth noting that the confidence interval excludes 0 (the expected differ-

Table 3

Mean and standard deviation of mean arterial pressure

	Mean arterial pressure (mmHg)	
	Standard therapy	Early goal-directed therapy
Number of patients	119	117
Mean	81	95
Standard deviation	18	19

ence if the null hypothesis were true); thus, although the increase may be small (1.4%), it is unlikely that the effect is to decrease saturation.

Comparison of two means arising from unpaired data

The most common comparison is probably that of two means arising from unpaired data (i.e. comparison of data from two independent groups). For example, consider the results from a recently published trial that compared early goal-directed therapy with standard therapy in the treatment of severe sepsis and septic shock [1]. A total of 263 patients were randomized and 236 completed 6 hours of treatment. The mean arterial pressures after 6 hours of treatment in the standard and early goal-directed therapy groups are shown in Table 3.

Note that the authors of this study also collected information on baseline mean arterial pressure and examined the 6-hour pressures in the context of these (using a method known as analysis of covariance) [1]. In practice this is a more appropriate analysis, but for illustrative purposes the focus here is on 6-hour mean arterial pressures only.

It appears that the mean arterial pressure was 14 mmHg higher in the early goal-directed therapy group. The 95% confidence intervals for the mean arterial pressure in the two groups are as follows.

$$\text{Standard therapy: } 81 \pm 1.96 \times \frac{18}{\sqrt{119}} = 81 \pm 3.23 = (77.8, 84.2)$$

$$\text{Early goal-directed therapy: } 95 \pm 1.96 \times \frac{19}{\sqrt{117}} = 95 \pm 3.44 = (91.6, 98.4)$$

There is no overlap between the two confidence intervals and, because these are the ranges in which the true population values are likely to lie, this supports the notion that there may be a difference between the two groups. However, it is more useful to estimate the size of any difference directly, and this can be done in the usual way. The only difference is in the calculation of the SE.

In the paired case attention is focused on the mean of the differences; in the unpaired case interest is in the difference of the means. Because the sample sizes in the unpaired case may be (and indeed usually are) different, the combined SE takes this into account and gives more weight to the larger sample size because this is likely to be more reliable. The pooled SD for the difference in means is calculated as follows:

$$SD_{\text{difference}} = \sqrt{\frac{(n_1 - 1) \times SD_1^2 + (n_2 - 1) \times SD_2^2}{(n_1 + n_2 - 2)}} \quad (3)$$

where SD_1 and SD_2 are the SDs in the two groups and n_1 and n_2 are the two sample sizes. The pooled SE for the difference in means is then as follows.

$$SE_{\text{difference}} = SD_{\text{difference}} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (4)$$

This SE for the difference in means can now be used to calculate a confidence interval for the difference in means and to perform an unpaired t-test, as above.

The pooled SD in the early goal-directed therapy trial example is:

$$\begin{aligned} SD_{\text{difference}} &= \sqrt{\frac{(119 - 1) \times 18^2 + (117 - 1) \times 19^2}{(119 + 117 - 2)}} \\ &= \sqrt{\frac{38,232 + 41,876}{234}} = \sqrt{342.34} = 18.50 \end{aligned}$$

and the corresponding pooled SE is:

$$\begin{aligned} SE_{\text{difference}} &= 18.50 \times \sqrt{\frac{1}{119} + \frac{1}{117}} = 18.50 \times \sqrt{0.008 + 0.009} \\ &= 18.50 \times 0.13 = 2.41 \end{aligned}$$

The difference in mean arterial pressure between the early goal-directed and standard therapy groups is 14 mmHg, with a corresponding 95% confidence interval of $14 \pm 1.96 \times 2.41 = (9.3, 18.7)$ mmHg. If there were no difference in the mean arterial pressures of patients randomized to early goal-directed and standard therapy then the difference in means would be close to 0. However, the confidence interval excludes this value and suggests that the true difference is likely to be between 9.3 and 18.7 mmHg.

To explore the likely role of chance in explaining this difference, an unpaired t-test can be performed. The null hypothesis in this case is that the means in the two populations are

the same or, in other words, that the difference in the means is 0. As for the previous two cases, a t statistic is calculated.

$$t = \frac{\text{difference in sample means}}{\text{SE of difference in sample means}}$$

A *P* value may be obtained by comparison with the t distribution on $n_1 + n_2 - 2$ degrees of freedom. Again, the larger the t statistic, the smaller the *P* value will be.

In the early goal-directed therapy example $t = 14/2.41 = 5.81$, with a corresponding *P* value less than 0.0001. In other words, it is extremely unlikely that a difference in mean arterial pressure of this magnitude would be observed just by chance. This supports the notion that there may be a genuine difference between the two groups and, assuming that the randomization and conduct of the trial was appropriate, this suggests that early goal-directed therapy may be successful in raising mean arterial pressure by between 9.3 and 18.7 mmHg. As always, it is important to interpret this finding in the context of the study population and, in particular, to consider how readily the results may be generalized to the general population of patients with severe sepsis or septic shock.

Assumptions and limitations

In common with other statistical tests, the t-tests presented here require that certain assumptions be made regarding the format of the data. The one sample t-test requires that the data have an approximately Normal distribution, whereas the paired t-test requires that the distribution of the differences are approximately Normal. The unpaired t-test relies on the assumption that the data from the two samples are both Normally distributed, and has the additional requirement that the SDs from the two samples are approximately equal.

Formal statistical tests exist to examine whether a set of data are Normal or whether two SDs (or, equivalently, two variances) are equal [2], although results from these should always be interpreted in the context of the sample size and associated statistical power in the usual way. However, the t-test is known to be robust to modest departures from these assumptions, and so a more informal investigation of the data may often be sufficient in practice.

If assumptions of Normality are violated, then appropriate transformation of the data (as outlined in Statistics review 1) may be used before performing any calculations. Similarly, transformations may also be useful if the SDs are very different in the unpaired case [3]. However, it may not always be possible to get around these limitations; where this is the case, there are a series of alternative tests that can be used. Known as nonparametric tests, they require very few or very limited assumptions to be made about the format of the data, and can therefore be used in situations where classical methods, such as t-tests, may be inappropriate. These

methods will be the subject of the next review, along with a discussion of the relative merits of parametric and nonparametric approaches.

Finally, the methods presented here are restricted to the case where comparison is to be made between one or two groups. This is probably the most common situation in practice but it is by no means uncommon to want to explore differences in means across three or more groups, for example lung function in nonsmokers, current smokers and ex-smokers. This requires an alternative approach that is known as analysis of variance (ANOVA), and will be the subject of a future review.

This article is the fifth in an ongoing, educational review series on medical statistics in critical care. Previous articles have covered 'presenting and summarizing data', 'samples and populations', 'hypotheses testing and *P* values' and 'sample size calculations'. Future topics to be covered include comparison of proportions, simple regression and analysis of survival data, to name but a few. If there is a medical statistics topic you would like explained, contact us on editorial@ccforum.com.

Competing interests

None declared.

References

1. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M: **Early goal-directed therapy in the treatment of severe sepsis and septic shock.** *N Engl J Med* 2001, **345**:1368-1377.
2. Altman DG: *Practical Statistics for Medical Research.* London, UK: Chapman & Hall, 1991.
3. Kirkwood BR: *Essentials of Medical Statistics.* Oxford, UK: Blackwell Science Ltd, 1988.

Review

Statistics review 6: Nonparametric methodsElise Whitley¹ and Jonathan Ball²¹Lecturer in Medical Statistics, University of Bristol, Bristol, UK²Lecturer in Intensive Care Medicine, St George's Hospital Medical School, London, UKCorrespondence: Editorial Office, *Critical Care*, editorial@ccforum.com

Published online: 13 September 2002

Critical Care 2002, **6** (DOI 10.1186/cc1820)This article is online at <http://ccforum.com/inpress/cc1820>

© 2002 BioMed Central Ltd (Print ISSN 1364-8535; Online ISSN 1466-609X)

Abstract

The present review introduces nonparametric methods. Three of the more common nonparametric methods are described in detail, and the advantages and disadvantages of nonparametric versus parametric methods in general are discussed.

Keywords nonparametric methods, sign test, Wilcoxon signed rank test, Wilcoxon rank sum test

Many statistical methods require assumptions to be made about the format of the data to be analysed. For example, the paired t-test introduced in Statistics review 5 requires that the distribution of the differences be approximately Normal, while the unpaired t-test requires an assumption of Normality to hold separately for both sets of observations. Fortunately, these assumptions are often valid in clinical data, and where they are not true of the raw data it is often possible to apply a suitable transformation. There are situations in which even transformed data may not satisfy the assumptions, however, and in these cases it may be inappropriate to use traditional (parametric) methods of analysis. (Methods such as the t-test are known as 'parametric' because they require estimation of the parameters that define the underlying distribution of the data; in the case of the t-test, for instance, these parameters are the mean and standard deviation that define the Normal distribution.)

Nonparametric methods provide an alternative series of statistical methods that require no or very limited assumptions to be made about the data. There is a wide range of methods that can be used in different circumstances, but some of the more commonly used are the nonparametric alternatives to the t-tests, and it is these that are covered in the present review.

The sign test

The sign test is probably the simplest of all the nonparametric methods. It is used to compare a single sample with some hypothesized value, and it is therefore of use in those situations in which the one-sample or paired t-test might tradition-

ally be applied. For example, Table 1 presents the relative risk of mortality from 16 studies in which the outcome of septic patients who developed acute renal failure as a complication was compared with outcomes in those who did not. The relative risk calculated in each study compares the risk of dying between patients with renal failure and those without. A relative risk of 1.0 is consistent with no effect, whereas relative risks less than and greater than 1.0 are suggestive of a beneficial or detrimental effect of developing acute renal failure in sepsis, respectively. Does the combined evidence from all 16 studies suggest that developing acute renal failure as a complication of sepsis impacts on mortality?

Fig. 1 shows a plot of the 16 relative risks. The distribution of the relative risks is not Normal, and so the main assumption required for the one-sample t-test is not valid in this case. Rather than apply a transformation to these data, it is convenient to use a nonparametric method known as the sign test.

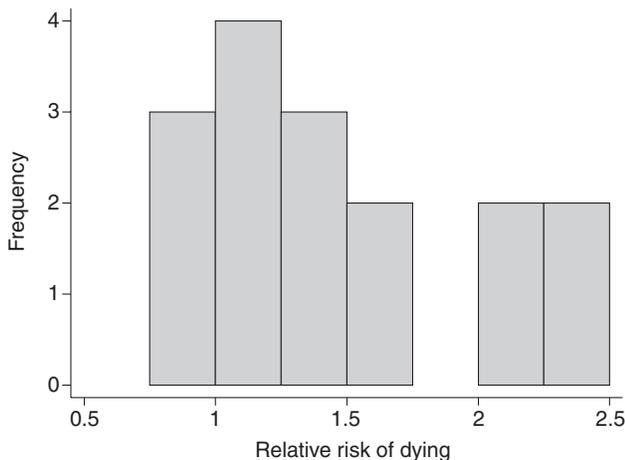
The sign test is so called because it allocates a sign, either positive (+) or negative (-), to each observation according to whether it is greater or less than some hypothesized value, and considers whether this is substantially different from what we would expect by chance. If any observations are exactly equal to the hypothesized value they are ignored and dropped from the sample size. For example, if there were no effect of developing acute renal failure on the outcome from sepsis, around half of the 16 studies shown in Table 1 would be expected to have a relative risk less than 1.0 (a 'negative'

Table 1

Relative risk of mortality associated with developing acute renal failure as a complication of sepsis

Study	Relative risk	Sign
1	0.75	-
2	2.03	+
3	2.29	+
4	2.11	+
5	0.80	-
6	1.50	+
7	0.79	-
8	1.01	+
9	1.23	+
10	1.48	+
11	2.45	+
12	1.02	+
13	1.03	+
14	1.30	+
15	1.54	+
16	1.27	+

Figure 1



Relative risk of mortality associated with developing acute renal failure as a complication of sepsis.

sign) and the remainder would be expected to have a relative risk greater than 1.0 (a 'positive' sign). In this case only three studies had a relative risk of less than 1.0 whereas 13 had a relative risk above this value. It is not unexpected that the number of relative risks less than 1.0 is not exactly 8; the

Table 2

Steps required in performing the sign test

Step	Details
1	State the null hypothesis and, in particular, the hypothesized value for comparison
2	Allocate a sign (+ or -) to each observation according to whether it is greater or less than the hypothesized value. (Observations exactly equal to the hypothesized value are dropped from the analysis)
3	Determine: N_+ = the number of observations greater than the hypothesized value N_- = the number of observations less than the hypothesized value S = the smaller of N_+ and N_-
4	Calculate an appropriate P value

more pertinent question is how unexpected is the value of 3? The sign test gives a formal assessment of this.

Formally the sign test consists of the steps shown in Table 2. In this example the null hypothesis is that there is no increase in mortality when septic patients develop acute renal failure.

Exact P values for the sign test are based on the Binomial distribution (see Kirkwood [1] for a description of how and when the Binomial distribution is used), and many statistical packages provide these directly. However, it is also possible to use tables of critical values (for example [2]) to obtain approximate P values.

The counts of positive and negative signs in the acute renal failure in sepsis example were $N_+ = 13$ and $N_- = 3$, and S (the test statistic) is equal to the smaller of these (i.e. N_-). The critical values for a sample size of 16 are shown in Table 3. S is less than or equal to the critical values for $P = 0.10$ and $P = 0.05$. However, S is strictly greater than the critical value for $P = 0.01$, so the best estimate of P from tabulated values is 0.05. In fact, an exact P value based on the Binomial distribution is 0.02. (Note that the P value from tabulated values is more conservative [i.e. larger] than the exact value.) In other words there is some limited evidence to support the notion that developing acute renal failure in sepsis increases mortality beyond that expected by chance.

Note that the sign test merely explores the role of chance in explaining the relationship; it gives no direct estimate of the size of any effect. Although it is often possible to obtain non-parametric estimates of effect and associated confidence intervals in principal, the methods involved tend to be complex in practice and are not widely available in standard statistical software. This lack of a straightforward effect estimate is an important drawback of nonparametric methods.

Table 3

Critical values for the sign test with a sample size of 16			
<i>P</i> value	0.10	0.05	0.01
Critical value	4	3	2

Table 4

Central venous oxygen saturation on admission and 6 hours after admission

Patient	SvO ₂ (%)		Difference	Sign
	On admission	6 hours		
1	39.7	52.9	13.2	+
2	59.1	56.7	-2.4	-
3	56.1	61.9	5.8	+
4	57.7	71.4	13.7	+
5	60.6	67.7	7.1	+
6	37.8	50.0	12.2	+
7	58.2	60.7	2.5	+
8	33.6	51.3	17.7	+
9	56.0	59.5	3.5	+
10	65.3	59.8	-5.5	-

SvO₂ = central venous oxygen saturation.

The sign test can also be used to explore paired data. Consider the example introduced in Statistics review 5 of central venous oxygen saturation (SvO₂) data from 10 consecutive patients on admission and 6 hours after admission to the intensive care unit (ICU). The paired differences are shown in Table 4. In this example, the null hypothesis is that there is no effect of 6 hours of ICU treatment on SvO₂. In other words, under the null hypothesis, the mean of the differences between SvO₂ at admission and that at 6 hours after admission would be zero. In terms of the sign test, this means that approximately half of the differences would be expected to be below zero (negative), whereas the other half would be above zero (positive).

In practice only 2 differences were less than zero, but the probability of this occurring by chance if the null hypothesis is true is 0.11 (using the Binomial distribution). In other words, it is reasonably likely that this apparent discrepancy has arisen just by chance. Note that the paired t-test carried out in Statistics review 5 resulted in a corresponding *P* value of 0.02, which appears at a first glance to contradict the results of the sign test. It is not necessarily surprising that two tests on the same data produce different results. The apparent discrepancy may be a result of the different assumptions required; in particular, the paired t-test requires that the differences be

Normally distributed, whereas the sign test only requires that they are independent of one another. Alternatively, the discrepancy may be a result of the difference in power provided by the two tests. As a rule, nonparametric methods, particularly when used in small samples, have rather less power (i.e. less chance of detecting a true effect where one exists) than their parametric equivalents, and this is particularly true of the sign test (see Siegel and Castellan [3] for further details).

The Wilcoxon signed rank test

The sign test is intuitive and extremely simple to perform. However, one immediately obvious disadvantage is that it simply allocates a sign to each observation, according to whether it lies above or below some hypothesized value, and does not take the magnitude of the observation into account. Omitting information on the magnitude of the observations is rather inefficient and may reduce the statistical power of the test. An alternative that does account for the magnitude of the observations is the Wilcoxon signed rank test. The Wilcoxon signed rank test consists of five basic steps (Table 5).

To illustrate, consider the SvO₂ example described above. The sign test simply calculated the number of differences above and below zero and compared this with the expected number. In the Wilcoxon rank sum test, the sizes of the differences are also accounted for.

Table 6 shows the SvO₂ at admission and 6 hours after admission for the 10 patients, along with the associated ranking and signs of the observations (allocated according to whether the difference is above or below the hypothesized value of zero). Note that if patient 3 had a difference in admission and 6 hour SvO₂ of 5.5% rather than 5.8%, then that patient and patient 10 would have been given an equal, average rank of 4.5.

Table 5

Steps required in performing the Wilcoxon signed rank test

Step	Details
1	State the null hypothesis and, in particular, the hypothesized value for comparison
2	Rank all observations in increasing order of magnitude, ignoring their sign. Ignore any observations that are equal to the hypothesized value. If two observations have the same magnitude, regardless of sign, then they are given an average ranking
3	Allocate a sign (+ or -) to each observation according to whether it is greater or less than the hypothesized value (as in the sign test)
4	Calculate: R ₊ = sum of all positive ranks R ₋ = sum of all negative ranks R = smaller of R ₊ and R ₋
5	Calculate an appropriate <i>P</i> value

Table 6

Central venous oxygen saturation on admission and 6 hours after admission

Patient	SvO ₂ (%)		Difference	Rank	Sign
	On admission	At 6 hours			
2	59.1	56.7	-2.4	1	-
7	58.2	60.7	2.5	2	+
9	56.0	59.5	3.5	3	+
10	65.3	59.8	-5.5	4	-
3	56.1	61.9	5.8	5	+
5	60.6	67.7	7.1	6	+
6	37.8	50.0	12.2	7	+
1	39.7	52.9	13.2	8	+
4	57.7	71.4	13.7	9	+
8	33.6	51.3	17.7	10	+

Table 7

Critical values for the Wilcoxon signed rank test with a sample size of 10

<i>P</i> value	0.10	0.05	0.01
Critical value	10	8	3

The sums of the positive (R_+) and the negative (R_-) ranks are as follows.

$$R_+ = 2 + 3 + 5 + 6 + 7 + 8 + 9 + 10 = 50$$

$$R_- = 1 + 4 = 5$$

Thus, the smaller of R_+ and R_- (R) is as follows.

$$R = R_- = 5$$

As with the sign test, a *P* value for a small sample size such as this can be obtained from tabulated values such as those shown in Table 7. The calculated value of *R* (i.e. 5) is less than or equal to the critical values for *P* = 0.10 and *P* = 0.05 but greater than that for *P* = 0.01, and so it can be concluded that *P* is between 0.01 and 0.05. In other words, there is some evidence to suggest that there is a difference between admission and 6 hour SvO₂ beyond that expected by chance. Notice that this is consistent with the results from the paired t-test described in Statistics review 5. *P* values for larger sample sizes (greater than 20 or 30, say) can be calculated based on a Normal distribution for the test statistic (see Altman [4] for details). Again, the Wilcoxon signed rank test

gives a *P* value only and provides no straightforward estimate of the magnitude of any effect.

The Wilcoxon rank sum or Mann-Whitney test

The sign test and Wilcoxon signed rank test are useful non-parametric alternatives to the one-sample and paired t-tests. A nonparametric alternative to the unpaired t-test is given by the Wilcoxon rank sum test, which is also known as the Mann-Whitney test. This is used when comparison is made between two independent groups. The approach is similar to that of the Wilcoxon signed rank test and consists of three steps (Table 8).

The data in Table 9 are taken from a pilot study that set out to examine whether protocolizing sedative administration reduced the total dose of propofol given. Patients were divided into groups on the basis of their duration of stay. The data presented here are taken from the group of patients who stayed for 3–5 days in the ICU. The total dose of propofol administered to each patient is ranked by increasing magnitude, regardless of whether the patient was in the protocolized or nonprotocolized group. Note that two patients had total doses of 21.6 g, and these are allocated an equal, average ranking of 7.5. There were a total of 11 nonprotocolized and nine protocolized patients, and the sum of the ranks of the smaller, protocolized group (*S*) is 84.5.

Again, a *P* value for a small sample such as this can be obtained from tabulated values. In this case the two individual sample sizes are used to identify the appropriate critical values, and these are expressed in terms of a range as shown in Table 10. The range in each case represents the sum of the ranks outside which the calculated statistic *S* must fall to reach that level of significance. In other words, for a *P* value below 0.05, *S* must either be less than or equal to 68 or greater than or equal to 121. In this case *S* = 84.5, and so *P* is greater than 0.05. In other words, this test provides no evidence to support the notion that the group who received protocolized sedation received lower total doses of propofol beyond that expected through chance. Again, for larger

Table 8

Steps required in performing the Wilcoxon rank sum (Mann-Whitney) test

Step	Details
1	Rank all observations in increasing order of magnitude, ignoring which group they come from. If two observations have the same magnitude, regardless of group, then they are given an average ranking
2	Add up the ranks in the smaller of the two groups (<i>S</i>). If the two groups are of equal size then either one can be chosen
3	Calculate an appropriate <i>P</i> value

Table 9**Total propofol doses in patients with a 3 to 5 day stay in the intensive care unit**

Nonprotocolized group		Protocolized group	
Dose (g)	Rank	Dose (g)	Rank
7.2	2	5.6	1
15.7	4	14.6	3
19.1	6	18.2	5
21.6	7.5	21.6	7.5
26.8	10	23.1	9
27.4	11	28.3	12
28.5	13	31.7	14
32.8	16	32.4	15
36.3	17	36.8	18
43.2	19		
44.7	20		

S = 84.5

Table 10**Critical values for the Wilcoxon rank sum test with sample sizes of 9 and 11**

<i>P</i> value	0.05	0.01	0.001
Critical value	68–121	61–128	53–136

sample sizes (greater than 20 or 30) *P* values can be calculated using a Normal distribution for *S* [4].

Advantages and disadvantages of nonparametric methods

Inevitably there are advantages and disadvantages to nonparametric versus parametric methods, and the decision regarding which method is most appropriate depends very much on individual circumstances. As a general guide, the following (not exhaustive) guidelines are provided.

This article is the sixth in an ongoing, educational review series on medical statistics in critical care. Previous articles have covered 'presenting and summarizing data', 'samples and populations', 'hypotheses testing and *P* values', 'sample size calculations' and 'comparison of means'. Future topics to be covered include simple regression, comparison of proportions and analysis of survival data, to name but a few. If there is a medical statistics topic you would like explained, contact us on editorial@ccforum.com.

Advantages of nonparametric methods

Nonparametric methods require no or very limited assumptions to be made about the format of the data, and they may therefore be preferable when the assumptions required for parametric methods are not valid.

Nonparametric methods can be useful for dealing with unexpected, outlying observations that might be problematic with a parametric approach.

Nonparametric methods are intuitive and are simple to carry out by hand, for small samples at least.

Nonparametric methods are often useful in the analysis of ordered categorical data in which assignment of scores to individual categories may be inappropriate. For example, nonparametric methods can be used to analyse alcohol consumption directly using the categories never, a few times per year, monthly, weekly, a few times per week, daily and a few times per day. In contrast, parametric methods require scores (i.e. 1–7) to be assigned to each category, with the implicit assumption that the effect of moving from one category to the next is fixed.

Disadvantages of nonparametric methods

Nonparametric methods may lack power as compared with more traditional approaches [3]. This is a particular concern if the sample size is small or if the assumptions for the corresponding parametric method (e.g. Normality of the data) hold.

Nonparametric methods are geared toward hypothesis testing rather than estimation of effects. It is often possible to obtain nonparametric estimates and associated confidence intervals, but this is not generally straightforward.

Tied values can be problematic when these are common, and adjustments to the test statistic may be necessary.

Appropriate computer software for nonparametric methods can be limited, although the situation is improving. In addition, how a software package deals with tied values or how it obtains appropriate *P* values may not always be obvious.

Competing interests

None declared.

References

1. Kirkwood BR: *Essentials of Medical Statistics*. Oxford, UK: Blackwell Science Ltd; 1988.
2. Neave HR: *Elementary Statistics Tables*. London, UK: Routledge; 1981.
3. Siegel S, Castellan NJ: *Non-parametric Statistics for the Behavioural Sciences*, 2nd ed. New York: McGraw-Hill; 1988.
4. Altman DG: *Practical Statistics for Medical Research*. London, UK: Chapman & Hall, 1991.