Scope Papers

(collection of statistical papers first published in Scope)

Edited by J V Freeman

Contents

The visual display of quantitative information

Dr Jenny V. Freeman, Steven A. Julious

Introduction

A knowledge of the appropriate methods for displaying data and results is a valuable skill for researchers and scientists. Appropriate figures are useful as they can be read quickly, and are particularly helpful when presenting information to an audience. In addition, plotting data is an extremely useful first stage to any analysis, as this could show extreme observations (outliers) together with any interesting patterns. In this tutorial, good practice guidelines for presentation of data in figures are given. In addition the basic types of data are described, together with some simple figures appropriate for their display.

Good practice recommendations for construction of figures

Box 1 outlines some basic recommendations for the construction and use of figures to display data. A fundamental principle, for both figures and tables is that they should maximise the amount of information presented for the minimum amount of ink used(Tufte 1983). Good figures have the following four features in common: clarity of message, simplicity of design, clarity of words and integrity of intentions and action(Bigwood and Spore 2003). A figure should have a title explaining what is displayed and axes should be clearly labelled; if it is not immediately obvious how many individuals the figure is based upon, this should also be stated. Gridlines should be kept to a minimum as they act as a distraction and can interrupt the flow of information. When using figures for presentation purposes care must be taken to ensure that they are not misleading; an excellent exposition of the way in which figures can be used to mislead can be found in Huff(Huff 1991).

Box 1 : Guidelines for good practice when constructing figures

1. The amount of information should be maximised for the minimum amount of ink
2. Figures should have a title explaining what is being displayed
3. Axes should be clearly labelled
4. Gridlines should be kept to a minimum
5. Avoid 3-D charts as these can be difficult to read
6. The number of observations should be included

Types of data

In order to appropriately display data, it is first important to understand the different types of data there are as this will determine the best method of displaying them. Briefly, data are either **categorical** or **quantitative**. Data are described as **categorical** when they can be categorised into distinct groups, such as ethnic group or disease severity. Categorical data can be divided into either **nominal** or **ordinal**. Nominal data have no natural ordering and examples include eye colour, marital status and area of residence. **Binary** data is a special subcategory of nominal data, where there are only two possible values, for example (male/female, yes/no, treated/not treated). Ordinal data occurs when there can be said to be a natural ordering of the data values, such as better/same/worse, grades of breast cancer, social class or quality scores.

Quantitative variables can be either discrete or continuous. **Discrete** data are also known as count data and occur when the data can only take whole numbers, such as the number of visits to a GP in a year or the number of children in a family. **Continuous** data are data that can measured and they can take any value on the scale on which they are measured; they are limited only by the scale of measurement and examples include height, weight, blood pressure, area or volume.

Basic charts for categorical data

Categorical data may be displayed using either a **pie chart** or a **bar chart**. Figure 1 shows a pie chart of the distribution of marital status by sex for UK adults at the 2001 census. Each segment of the pie chart represents the proportion of the UK population who are in that category. It is clear from this figure that differences between the sexes exist with respect to marital status; nearly half of all men have never married, whilst this proportion was smaller for women. Interestingly the proportion of women who were widowed was about three times that for men. Figure 2 displays the same data in a bar chart. The different marital status categories are displayed along the horizontal axis whilst on the vertical axis is percentage. Each bar represents the percentage of the total population in that category. For example, examining Figure 2, it can be seen that the percentage of men who are married is about 48%, whilst the percentage of women is closer to 40%. Generally pie charts are to be avoided as they can be difficult to interpret particularly when the number of categories becomes greater than 5. In addition, unless the percentages in the individual categories are displayed (as here) it can be much more difficult to estimate them from a pie chart than from a bar chart. The relative proportions falling in the different categories is much clearer in Figure 2 than in Figure 1. For both chart types it is important to include the number of observations on which it is based, particularly when comparing more than one chart. And finally, neither of these charts should be displayed as 3-D as these are especially difficult to read and interpret.

Figure 1: Pie Chart of marital status for UK,
http://www.statistics.gov.uk/STATBASE/Expodata/Spreadsheets/D7680.xls

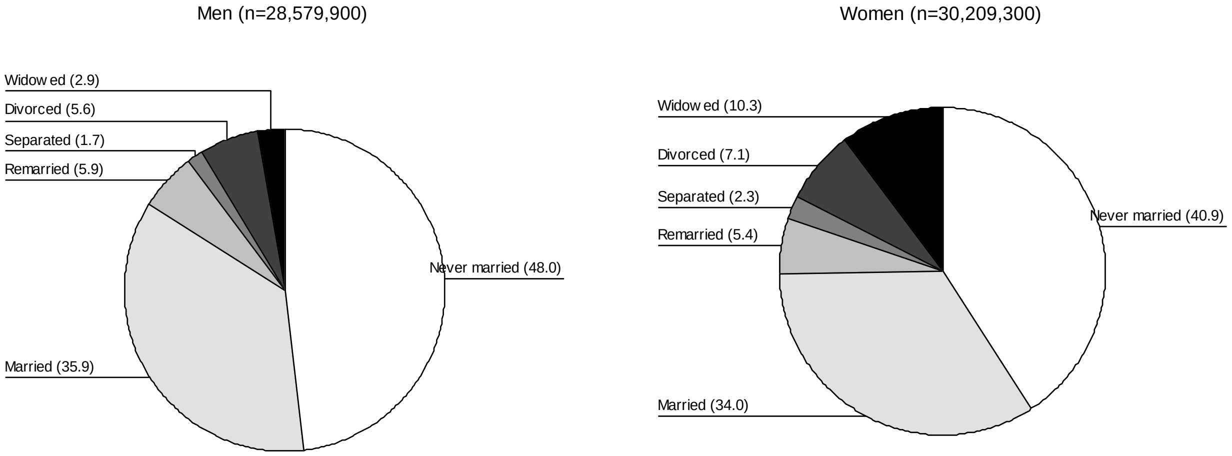
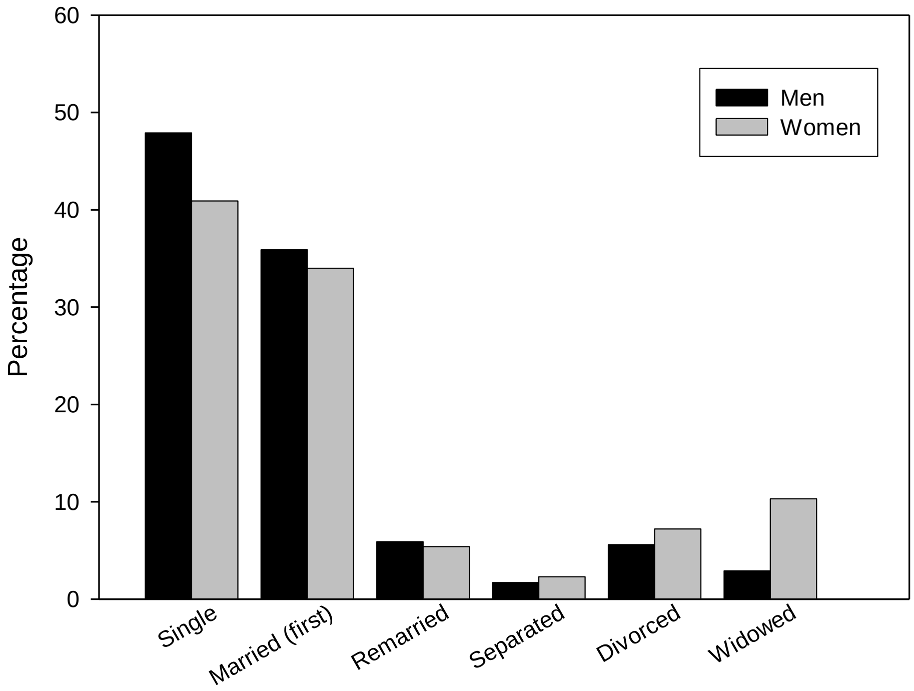


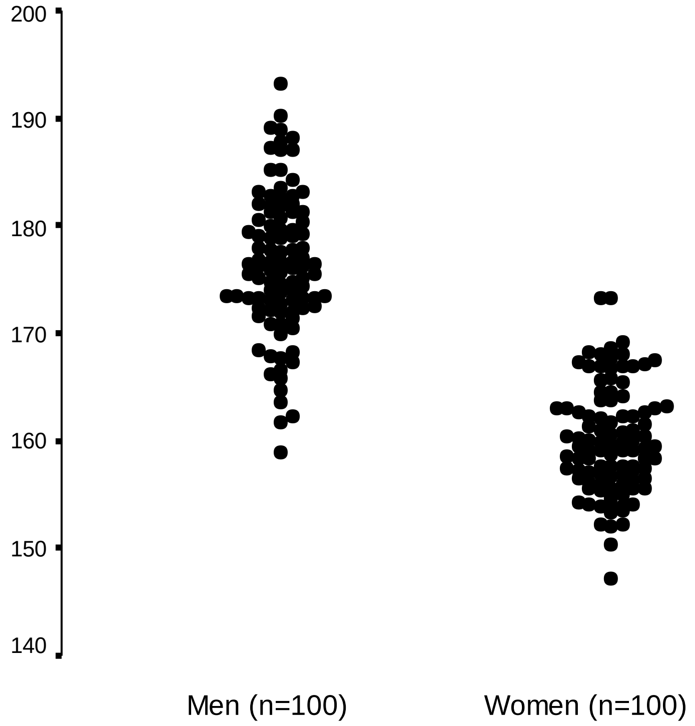
Figure 2: Data in Figure 1 displayed as a Bar Chart



Basic charts for quantitative data

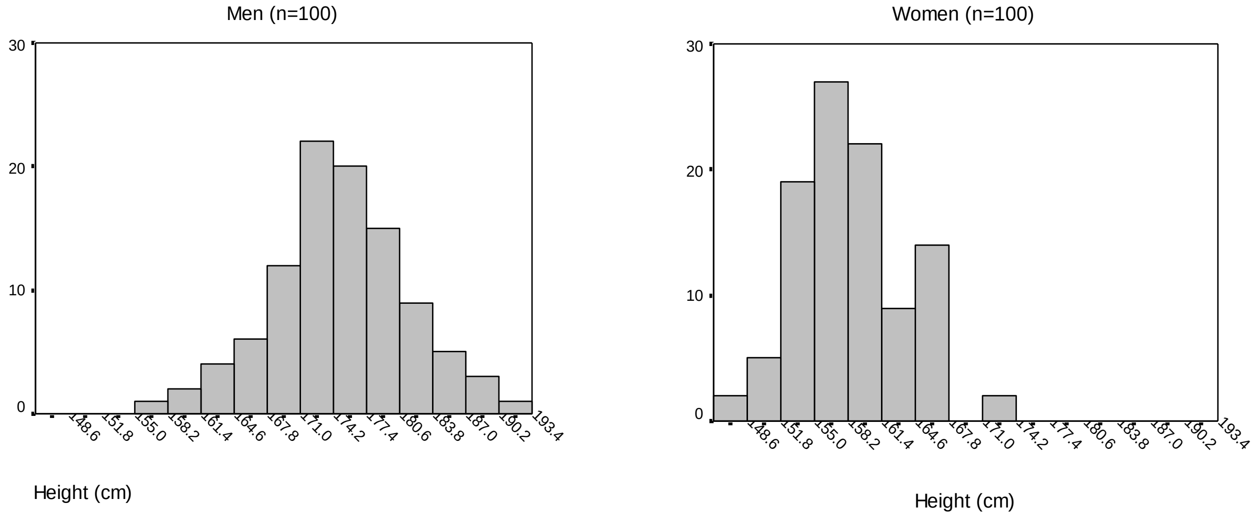
There are several charts that can be used for quantitative data. **Dot plots** are one of the simplest ways of displaying all the data. Figure 3 shows dot plots of the heights for a random sample of 100 couples. Each dot represents the value for an individual and is plotted along a vertical axis, which in this case, represents height in metres. Data for several groups can be plotted alongside each other for comparison; for example, data for the 100 randomly sampled couples are plotted separately by sex in Figure 3 and the differences in height between men and women can be clearly seen.

Figure 3: Dot plot of heights of random sample of 100 couples



A common method for displaying continuous data is a **histogram**. In order to construct a histogram the data range is divided into several non-overlapping equally sized categories and the number of observations falling into each category counted. The categories are then displayed on the horizontal axis and the frequencies displayed on the vertical axis, as in Figure 4. Occasionally the percentages in each category are displayed on the y-axis rather than the frequencies and it is important that if this is done, the total number of observations that the percentages are based upon must be included in the chart. The choice of number of categories is important as too few categories and much important information is lost, too many and any patterns are obscured by too much detail. Usually between 5 and 15 categories will be enough to gain an idea of the distribution of the data.

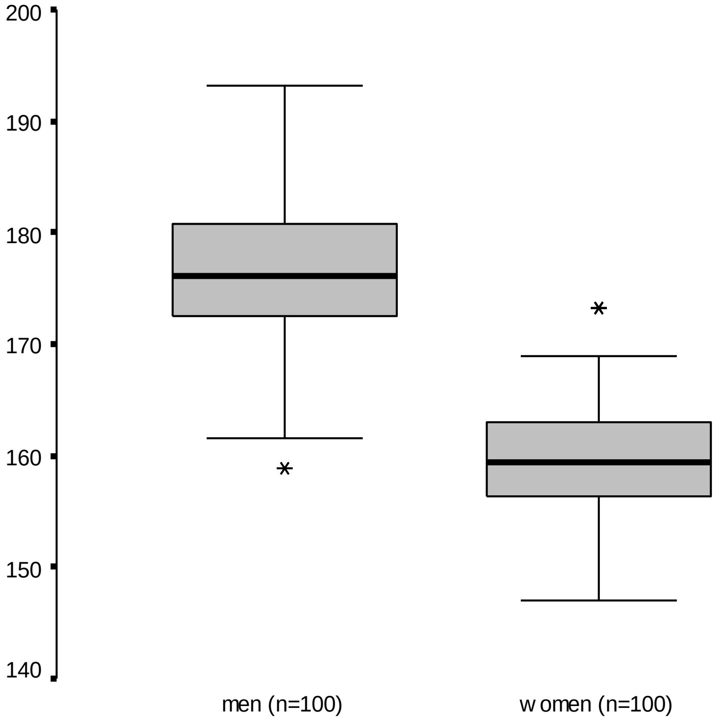
Figure 4: Histograms of height for random sample of 100 couples, by sex



A useful feature of a histogram is that it is possible to assess the distributional form of the data; in particular whether the data are approximately Normal, or are skewed. The histogram of Normally distributed data will have a classic ‘bell’ shape, with a peak in the middle and symmetrical tails, such as that for men in Figure 4a. The **Normal distribution** (sometimes known as the Gaussian distribution) is one of the fundamental distributions of statistics, and its properties, which underpin many statistical methods, will be discussed in a later tutorial. **Skewed** data are data which are not symmetrical, negatively skewed data have a long left-hand tail at lower values, with a peak at higher values, whilst conversely positively skewed data have a peak at lower values and a long tail of higher values.

Another extremely useful method of plotting continuous data is **a box-and-whisker** or **box plot** (Figure 5). Box plots can be particularly useful for comparing the distribution of the data across several groups. The box contains the middle 50% of the data, with lowest 25% of the data lying below it and the highest 25% of the data lying above it. In fact the upper and lower edges represent a particular quantity called the interquartile range. The horizontal line in the middle of the box represents the median value, the value such that half of the observations lie below this value and half lie above it. The whiskers extend to the largest and smallest values excluding the outlying values. The outlying values are those values more than 1.5 box lengths from the upper or lower edges, and are represented as the dots outside the whiskers. Figure 5 shows box plots of the heights of the men and women. As with the dot plots, the gender differences in height are immediately obvious from this plot and this illustrates the main advantage of the box plot over histograms when looking at multiple groups. Differences in the distributions of data between groups are much easier to spot with box plots than with histograms.

Figure 5: Boxplot of height

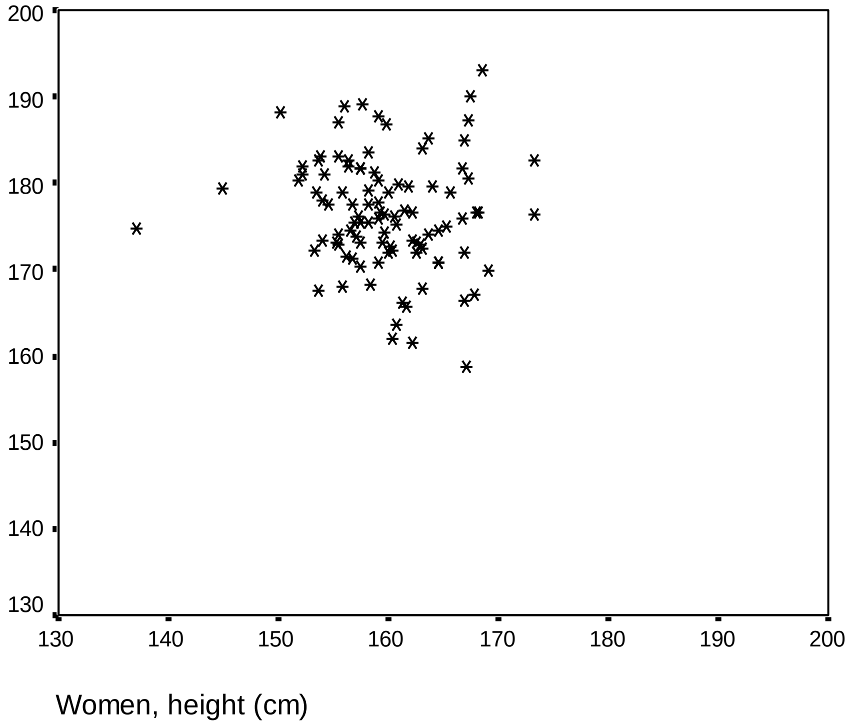


The association between two continuous variables can be examined visually by constructing a **scatterplot**. The values of one variable are plotted on the horizontal axis (known as the X-axis) and the values of another are plotted on the vertical axis (Y-axis). If it is known (or suspected) that the value of one variable (independent) influences the value of the other variable (dependent), it is usual to plot the independent variable on the horizontal axis and the dependent variable on the vertical axis. Although it is not always obvious, it is often clear which variables to place on the X- and Y-axes. Experimentally the X-axis would be something that the experimenter controls while the Y-axis would be the response to the X-axis. Variables that tend to go on the X-axis are variables such age, time (hours, months, years etc) or temperature. This will be discussed again in greater detail in a later tutorial examining the techniques of regression and correlation. Figure 6 shows the scatter plot of women's height against their partner's height and each dot represents the two height values for a couple. Note, here variables could have been placed interchangeably on the X-or Y-axes.

Summary

This tutorial has given basic good practice guidelines for producing figures and highlighted some of the simple figures available for displaying data. The list of figures described is not exhaustive and the work presented here will be revisited and extended in subsequent tutorials. And finally, it is worth considering that often the simplest plots convey the strongest message.

Figure 6: Scatter plot of heights for 100 couples, women on horizontal axis, men on vertical axis



Describing and summarising data

Jenny V Freeman and Steven A Julious

Introduction

A recent tutorial in SCOPE gave some good practice recommendations for the visual display of quantitative data and highlighted some simple figures available for displaying data(Freeman and Julious 2005c). This tutorial is concerned with ways of describing and summarising data. In addition, the presentation of numbers and use of tables will also be covered and good practice guidelines for communicating quantitative information will be outlined.

Describing categorical data

An initial step when describing categorical data is to count the number of observations in each category and express them as percentages of the total sample size. For example, Table 1 shows the marital status of the UK population taken from the 2002 census by sex(Anon 2005). The data are categorised in two ways, by marital status and gender, enabling the distribution of marital status to be compared between the two sexes; Table 1 is an example of a *contingency table* with 6 rows (representing marital status) and 2 columns (representing gender) and marital status is said to have been *cross-tabulated* with study group. When presenting data in this way (as percentages), it is important to include the *denominator* for each group (total sample size), as giving percentages alone can be misleading if the groups contained very different numbers(Altman and Bland 1996).

Table 1: Marital status for UK population, 2001 census

	Men (n=28,579,900)	Women (n=30,209,300)
Never married	48.0	40.9
Married	35.9	34.0
Divorced	5.6	7.1
Remarried	5.9	5.4
Separated	1.7	2.3
Widowed	2.9	10.3

Describing quantitative data

As it can be difficult to make sense of a large amount of quantitative data, an initial approach, in addition to examining the data visually, is to calculate summary measures, to describe the *location* (a measure of the 'middle value') and the *spread* (a measure of the dispersion of the values) for each variable. These are of great interest, particularly if a comparison between groups is to be made or the results of the study are to be generalised to a larger group.

Measures of location

There are several measures of location, as summarised in Box 1. The simplest is the *mode*. This is simply the most common observation and is the highest bar of the histogram. Looking at the histogram for height of a random sample of 100 men, the modal height around 171.9cm as this is the height category with the highest bar on the histogram (Figure 1a). However, the mode is rarely used since its value depends upon the accuracy of the measurement. If, for example, the number of height bands on the histogram were increased from 14 to 19, the mode would change to 176cm (Figure 1b). In addition, it can be difficult to determine if there is more than one distinct peak – for example two peaks would indicate bi-modal data. The most practical use for a mode is when the variable is either ordinal or discrete (with a finite number of categories) where one category dominates.

Box 1: Measures of location

Mode Most common observation

Median Middle observation, when the data are arranged in order of increasing value. If there is an even number of observations the median is calculated as the average of the middle two observations

Mean $= \dfrac{\text{Sum of all observations}}{\text{Number of observations}} = \bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

where $\bar{x}$ is the sample mean, x_i is the ith observation, n is the sample size and the notation $\sum_{i=1}^{n}$ represents the addition or summing up of all the observations from the first (i = 1) to the last (n).

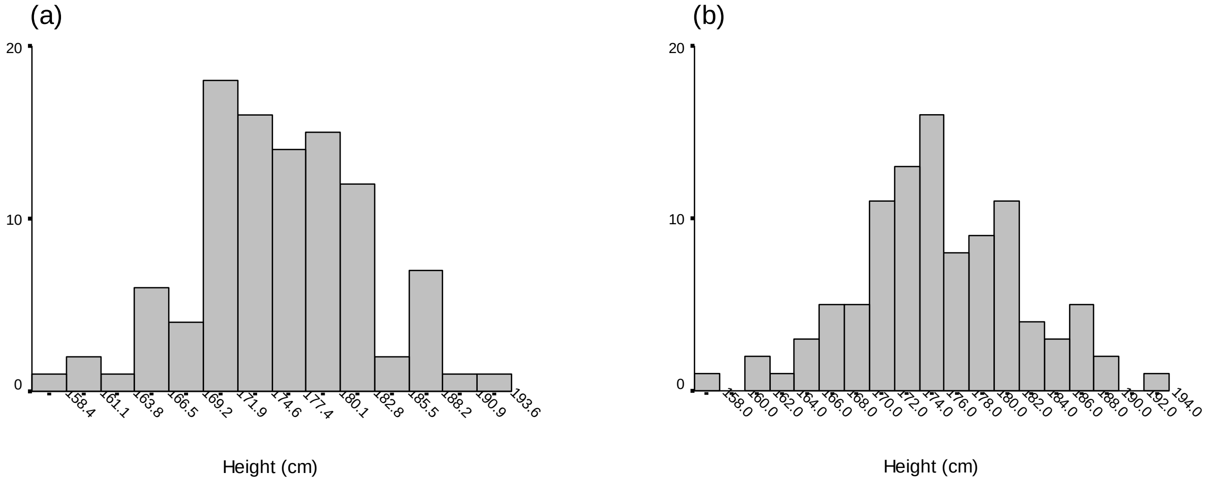
For example, consider the ages (in years) of five individuals: 42,32, 41, 45 and 38.

The most common observation is: 42, 32, 41, 45 or 38. Unfortunately, multiple modes exist in this example, so there is no unique **mode**.

The five ages in ascending order are: 32, 38, 41, 42, and 45. The **median** is the middle or 3rd value of the ranked or ordered ages i.e. 41 years.

The **mean** is: 82 + 72 + 81 + 85 + 58 = 198 divided by the number of observations, 5, i.e. 39.6 years.

Figure 1. Histograms of height for a random sample of 100 men, (a) with 14 bins (categories/bars) and (b) with 19 bins. Note the change in the value of the highest bar

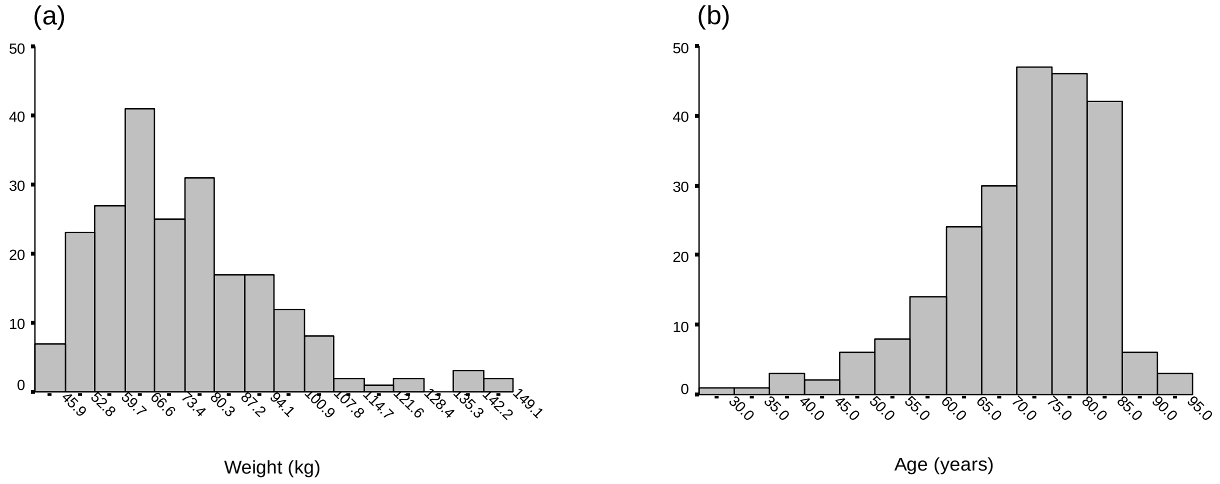


Two other more useful measures are the *median* and the *mean*. The median is the middle observation, when the data are arranged in order of increasing value. It is the value that divides the data into two equal halves. If there is an even number of observations then the median is calculated as the average of the two middle observations. For example, if there are 11 observations the median is simply the 6th observation, but if there are 10 observations the median is the (5th + 6th observation)/2. The median is not sensitive to the behaviour of outlying data, thus if the smallest value was even smaller, or the largest value even bigger it would have no impact on the value of the median. The median height for the 100 men is 176.3cm.

Probably the most commonly used measure of the central value of a set of data is the *mean*. It is calculated as the sum of all observations divided by the total number of observations. Each observation makes a contribution to the mean value and thus it is sensitive to the behaviour of outlying data; as the largest value increases this causes the mean value to increase and conversely, as the value of the smallest observation becomes smaller the value of the mean decreases. The mean height for the 100 men is 176.5cm.

Both the mean and median can be useful, but they can give very different impressions when distribution of the data is *skewed*, because of the relative contributions (or lack of, in the case of the median) of the extreme values. Skewed data are data that are not symmetrical and this is best illustrated by examining the histograms for such data, as in Figures 2a and 2b. Data, such as that in Figure 2a, which have a long right-hand tail of higher values, but where the majority of observations are clustered at lower values are called right skewed, or positively skewed (conversely data where the observations are clustered at higher values but with a long left-hand tail of lower values such as that in Figure 2b are called left skewed or negatively skewed).

Figure 2: Distribution of data displaying positive (a) and negative (b) skew. Data are taken from a randomised controlled trial examining the cost effectiveness of community leg ulcer clinics (n=233)(Morrell et al. 1998).



There are no firm rules about which to use, but when the distribution is not skew it is usual to use the mean; it is worth noting that if the data are symmetrically distributed the mean and median will be similar, as can be seen from the mean and median of the height data described earlier (176.3cm and 176.5cm respectively). However, if data are skew then it is better to use the median, as this is not influenced by the extreme values and may not be as misleading as the mean; an extreme example would be the median of the sample 1, 2, 3, 4 and 100,000, which would be 3, whereas the mean is 20,002. One example of where medians have been used in preference to means is in reporting salaries. Due to a relatively few outlying high-income earners the vast majority of workers were earning much less than the mean wage, thus nowadays, medians are produced and quoted(Bird 2004).

Measures of Spread

In addition to finding measures to describe the location of a dataset, it is also necessary to be able to describe its spread. Just as with the measures of location, there are both simple and more complex possibilities (as summarised in Box 2). The simplest is the *range* of the data, from the smallest to the largest observation. The range of height for the random sample of 100 men is 159 to 193cm (or 35 cm as a single number). The advantage of the range is that it is easily calculated, but its drawback is that it is vulnerable to *outliers*, extremely large and extremely small observations. A more useful measure is to take the median value as discussed above and further divide the two data halves into halves again. These values are called the *quartiles* and the difference between the bottom (or 25% percentile) and top quartile (or 75th percentile) is referred to as the *inter-quartile range* (IQR). This is the observation below which the bottom 25% of the data lie and the observation above which the top 25% lie: the middle 50% of the data lie between these limits. Unlike the range it is not as sensitive to the extreme values. The IQR for the height of the random sample of 100 men is 172.5 to 181 cm (or 8.5cm as a single number). Strictly speaking the range and IQR are single numbers but frequently the two values, minimum and maximum, or the 25% and 75% percentiles respectively, are all reported as this can be more informative.

Box 2: Measures of spread

Range Minimum observation to the maximum observation

Interquartile range Observation below which the bottom 25% of data lie and the observation above which the top 25% of data lie. If the value falls between two observations, e.g. if 25th centile falls between 5th and 6th observations then the value is calculated as the average of the two observation (this is the same principle as for the median

Standard deviation = SD = $\sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

where $\bar{x}$ is the sample mean, x_i is the ith observation, n is the sample size and the notation $\sum_{i=1}^{n}$ represents the addition or summing up of all the squared deviations from the sample mean from the first (i = 1) to the last (nth) observation.

For example, consider the ages (in years) of the five individuals abovel: 32, 38, 41, 42 and 45.

The **range** of the data is from 32 to 45 years or 13 years.

The five ages in ascending order are: 32, 38, 41, 42 and 45. The bottom 25% of data, falls somewhere between the 1st and 2nd ordered observations, i.e. 32 and 38, so we can take the average of these two observations 32 + 38 = 70/2 = 35 years. The top 25% of data, falls somewhere between the 4th and 5th ordered observations, i.e. 42 and 45. So the 75th percentile is the average of the two observations 42 + 45 = 87/2 = 43.5. Hence the **interquartile** range is 35.0 to 43.5 years or 8.5 years**.**

The **standard deviation** is calculated by first working out the squared deviation of each observation from the sample mean of 39.6 years i.e.
$(32 – 39.6)^2 + (38 – 39.6)^2 + (41 – 39.6)^2 + (42 – 39.6)^2 + (45 – 39.6)^2 = 92.7$ years^2.
This result is divided by the number in the sample minus one (i.e. 5 - 1= 4) i.e. 97.2/4 = 24.3 years^2. Finally, we take the square root of this number to give us a **standard deviation** of 4.93 years.

The most common measure of the spread of the data is the standard deviation (see Box 2) and it is used in conjunction with the mean. It provides a summary of the differences of each observation from the mean value. The standard deviation (SD) has units on the same scale as the original measurement (e.g. cm if height is being measured in cm). For the sample of 100 men, the SD for height is 6.6cm and provides a measure of average deviation for an observation from the sample mean.

As with the measures of location, when deciding which measure of spread to present it is necessary to know whether the data are skewed or not; this will also have a bearing on how the data will be analysed subsequently, as will be seen in the following chapter. When the distribution is not skewed it is usual to use the standard deviation. However, if data are skewed then it is better to use the range or inter-quartile range.

Presentation of numbers

As with charts, there are a few basic rules of good presentation, both within the text of a document or presentation, and within tables, as outlined in Box 3.

Box 3: Guidelines for good practice when presenting numbers:

1. The amount of information should be maximised for the minimum amount of ink.
2. Numerical precision should be consistent throughout a paper or presentation, as far as possible.
3. Avoid spurious accuracy. Numbers should be rounded to two effective digits.
4. Quantitative data should be summarised using either the mean and SD (for symmetrically distributed data) or the median and IQR or range (for skewed data). The number of observations on which these summary measures are based should be included.
5. Categorical data should be summarised as frequencies and percentages. As with quantitative data, the number of observations should be included.
6. Tables should have a title explaining what is being displayed and columns and rows should be clearly labelled.
7. Gridlines in tables should be kept to a minimum.
8. Rows and columns should be ordered by size.

A fundamental principle is that the amount of information should be maximised for the minimum amount of ink(Tufte 1983). For summarising numerical data, the mean and standard deviation should be used, or if the data have a skewed distribution the median and range or inter-quartile range should be used. However, for all of these calculated quantities it is important to state the total number of observations on which they are based. When summarising categorical data, both frequencies and percentages can be used, but if percentages are reported it is important that the denominator (i.e. total number of observations) is given. Numerical precision should be consistent throughout and summary statistics such as means and standard deviations should not have more than one extra decimal place (or significant figure) compared to the raw data. Spurious precision should be avoided, although when certain measures are to be used for further calculations greater precision may sometimes be appropriate(Altman & Bland 1996).

Tables, including the column and row headings, should be clearly labelled and a brief summary of the contents of a table should always be given in words, either as part of the title or in the main body of the text. Gridlines can be used to separate labels and summary measures from the main body of the data in a table. However, their use should be kept to a minimum, particularly vertical gridlines, as they can interrupt eye movements, and thus the flow of information. Elsewhere white space can be used to separate data, for example, different variables from each other. The information in tables is easier to comprehend if the columns (rather than the rows) contain like information, such as means and standard deviations, as it is easier to scan down a column than across a row(Ehrenberg 2000). Where there is no natural ordering of the rows (or indeed columns), such as marital status in Table 1, they should be ordered by size as this helps the reader to scan for patterns and exceptions in the data(Ehrenberg 2000).

Table or Chart?

Some basic charts for displaying data were described in the previous tutorial. Plotting data is a useful first stage to any analysis and will show extreme observations together with any discernible patterns. Charts are useful as they can be read quickly, and are particularly helpful when presenting information to an audience such as in a seminar or conference presentation. Although there are no hard and fast rules about when to use a chart and when to use a table, when presenting the results in a report or a paper it is often best to use tables so that the reader can scrutinise the numbers directly. Tables can be useful for displaying information about many variables at once, whilst charts can be useful for showing multiple observations on groups or individuals.

Summary

The essence of any attempt to present data and results, either in a presentation or on paper is to communicate with an audience and it is hoped that by following the basic rules outlined here that task will be made simpler. This tutorial has covered some basic measures for describing and summarising data. It has also outlined some good practice guidelines for communicating quantitative information. The next tutorial will examine the Normal distribution and the Central Limit Theorem.

The Normal distribution

Jenny V Freeman, Steven A Julious

Introduction

The first two tutorials in this series have focussed on displaying data and simple methods for describing and summarising data. There has been little discussion of statistical theory. In this note we will start to explore some of the basic concepts underlying much statistical methodology. We will describe the basic theory underlying the Normal distribution and the link between empirical frequency distributions (the observed distribution of data in a sample) and theoretical probability distributions (the theoretical distribution of data in a population). In addition we will introduce the idea of a confidence interval.

Theoretical probability distributions

Since it is rarely possible to collect information on an entire population, the aim of many statistical analyses is to use information from a sample to draw conclusions (or '**make inferences**') about the population of interest. These inferences are facilitated by making assumptions about the underlying distribution of the measurement of interest in the population as a whole, by applying an appropriate theoretical model to describe how the measurement behaves in the population[1]. In the context of this note the population is a theoretical concept used for describing an entire group and one way of describing the distribution of a measurement in a population is by use of a suitable theoretical probability distribution. Probability distributions can be used to calculate the probability of different values occurring and they exist for both continuous and categorical measurements.

In addition to the Normal distribution (described later in this note), there are many other theoretical distributions, including the Chi-squared, Binomial and the Poisson distributions (these will be discussed in later tutorials). Each of these theoretical distributions is described by a particular mathematical expression (formally referred to as a model) and for each model there exist summary measures, known as **parameters** which completely describe that particular distribution. In practice, parameters are usually estimated by quantities calculated from the sample, and these are known as **statistics** i.e. a statistic is a quantity calculated from a sample in order to estimate an unknown parameter in a population. For example, the Normal distribution is completely characterised by the population mean (μ) and population standard deviation (σ) and these are estimated by the sample mean ($\bar{x}$) and sample standard deviation (s) respectively.

[1] Note that prior to any analysis it is usual to make assumptions about the underlying distribution of the measurement being studied. These assumptions can then be investigated through various plots and figures for the observed data – for example a histogram for continuous data. These investigations are referred to as diagnostics and will be discussed throughout subsequent notes.

The Normal distribution

The Normal, or Gaussian distribution (named in honour of the German mathematician C.F.Gauss, 1777-1855) is the most important theoretical probability distribution in statistics. At this point it is important to stress that in this context the word ‘Normal’ is a statistical term and is not used in the dictionary or clinical sense of conforming to what would be expected. Thus, in order to distinguish between the two, statistical and dictionary ‘normal’, it is conventional to use a capital letter when referring to the Normal distribution.

Box 1: Properties of the Normal distribution

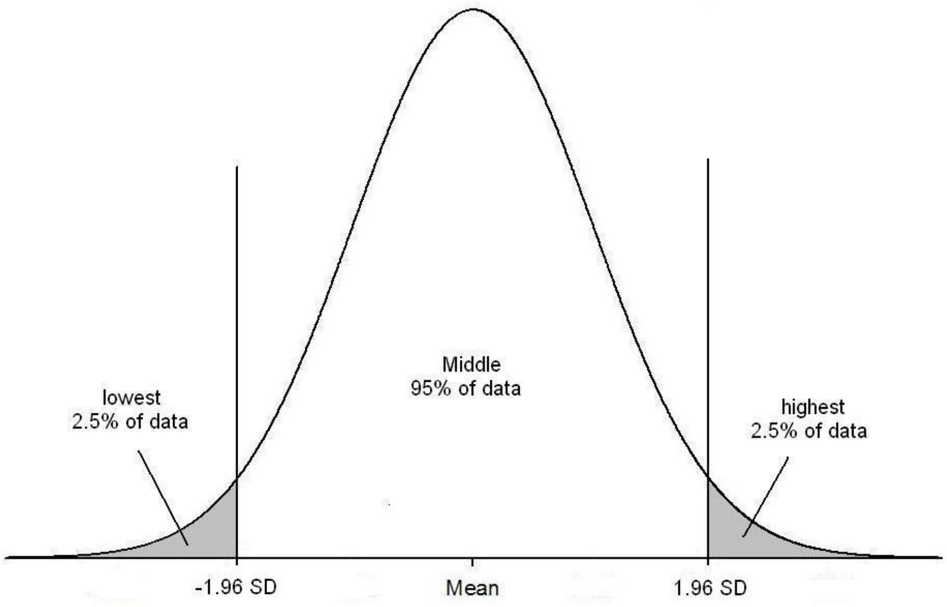
1. It is bell shaped and has a single peak (unimodal)
2. Symmetrical about the mean
3. Uniquely defined by two parameters, the mean (μ) and standard deviation (σ)
4. The mean, median and mode all coincide
5. The probability that a Normally distributed random variable, x, with mean, μ, and standard deviation, σ, lies between the limits ($\mu - 1.96\sigma$) and ($\mu + 1.96\sigma$) is 0.95
i.e. 95% of the data for a Normally distributed random variable will lie between the limits ($\mu - 1.96\sigma$) and ($\mu + 1.96\sigma$)*
6. The probability that a Normally distributed random variable, x, with mean, μ, and standard deviation, σ, lies between the limits ($\mu - 2.44\sigma$) and ($\mu + 2.44\sigma$) is 0.99
7. Any position on the horizontal axis of Figure 1 can be expressed as a number of standard deviations away from the mean value

*This fact is used for calculating the 95% confidence interval for Normally distributed data

The basic properties of the Normal distribution are outlined in box 1. The distribution curve of data which are Normally distributed has a characteristic shape; it is bell-shaped, and symmetrical about a single peak (figure 1). For any given value of the mean, populations with a small standard deviation have a distribution clustered close to the mean (μ), whilst those with a large standard deviation have a distribution that is widely spread along the measurement axis and the peak is more flattened.

As mentioned earlier the Normal distribution is described completely by two parameters, the mean (μ) and the standard deviation (σ). This means that for any Normally distributed variable, once the mean and variance (σ^2) are known (or estimated), it is possible to calculate the probability distribution for that population.

Figure 1: The Normal Distribution



An important feature of a Normal distribution is that 95% of the data fall within 1.96 standard deviations of the mean – the unshaded area in the middle of the curve on figure 1. A summary measure for a sample often quoted is the two values associated with the mean +/- 1.96 x standard deviation ($\bar{x}$ +/-1.96s). These two values are termed the **Normal range** and represent the range within which 95% of the data are expected to lie. Note 68.7% of data lie within 1 standard deviation of the mean whilst virtually all of the data (99.7%) will lie within 3 standard deviations (95.5% will lie within 2). The Normal distribution is important as it underpins much of the subsequent statistical theory outlined both in this and later tutorials, such as the calculation of confidence intervals and linear modelling techniques.

The Central Limit Theorem (or the law of large numbers)

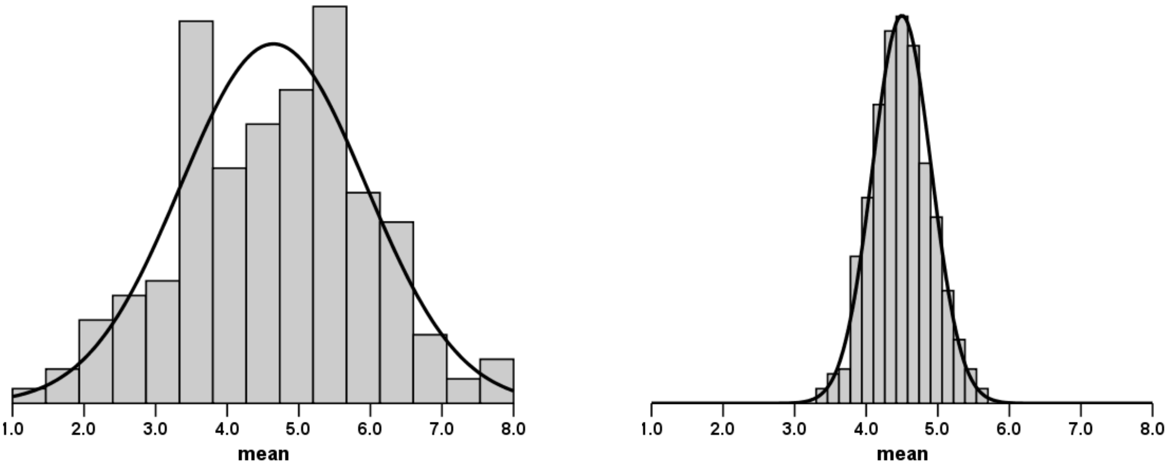
The Central Limit Theorem states that given any series of independent, identically distributed random variables, their means will tend to a Normal distribution as the number of variables increases. Put another way, the distribution of sample means drawn from a population will be Normally distributed whatever the distribution of the actual data in the population as long as the samples are large enough.

In order to illustrate this, consider the random numbers 0 to 9. The distribution of these numbers in a random numbers table would be uniform. That is to say that each number has an equal probability of being selected and the shape of the theoretical distribution is represented by a rectangle. According to the Central Limit Theorem, if you were to select repeated random samples of the same size from this distribution, and then calculate the means of these different samples, the distribution of these sample means would be approximately Normal and this approximation would improve as the size of each sample increased. Figure 2a represents the distribution of the sample means for 500 samples of size 5. Even with such a small sample size the approximation to the Normal is remarkable, whilst repeating the experiment with samples of size 50, improves the fit to the Normal distribution (Figure 2b). The other noteworthy feature of these two figures is that as the size of the samples increases (from 5 to 50), the spread of the means is decreased.

Figure 2: Distribution of means from 500 samples

(a) Samples of size 5, mean=4.64, sd=1.29

(b) Samples of size 50, mean=4.50, sd=0.41



Each mean estimated from a sample is an unbiased estimate of the true population mean and by repeating the sampling many times we can obtain a sample of plausible values for the true population mean. Using the Central Limit Theorem we can infer that 95% of sample means will lie within 1.96 standard deviations of the population mean. As we do not usually know the true population mean the more important inference is that with the sample mean we are 95% confident that the population mean will fall within 1.96 standard deviations of the sample mean. In reality, as we usually only take a single sample, we can use the Central Limit Theorem to construct an interval within which we are reasonably confident the true population mean will lie. This range of plausible values is known as the **confidence interval** and the formula for the confidence interval for the mean is given in Box 2. Technically speaking, the 95% confidence interval is the range of values within which the true population mean will lie 95% of the time if a study was repeated many times. Crudely speaking, the confidence interval gives a range of plausible values for the true population mean. We will discuss confidence intervals further in subsequent notes in context with hypothesis tests and P-values.

Box 2: Formula for the confidence interval for a mean

$$\bar{x} - 1.96 \times \frac{s}{\sqrt{n}} \text{ to } \bar{x} - 1.96 \times \frac{s}{\sqrt{n}}$$

s = sample standard deviation and *n* = number of individuals in the sample

In order to calculate the confidence interval we need to be able to estimate the standard deviation of the sample mean. It is defined as the sample standard deviation, *s*, divided by the square root of the number of individuals in the sample, $s/\sqrt{n}$ and is usually referred to as the **standard error**. In order to avoid confusion, it is worth remembering that using the standard deviation (of all individuals in the sample) you can make inferences about the spread of the measurement within the population for individuals whilst using the standard error you can make inference about the spread of the means: the standard **d**eviation is for **d**escribing (the spread of data) whilst the standard **e**rror is for **e**stimating (how precisely the mean has been pinpointed).

Summary

In this tutorial we have outlined the basic properties of the Normal distribution and discussed the Central Limit Theorem and outlined its importance to statistical theory. The Normal distribution is fundamental to many of the tests of statistical significance outlined in subsequent tutorials, whilst the principles of the Central Limit Theorem enable us to calculate confidence intervals and make inference about the population from which the sample is taken.

Hypothesis testing and Estimation

Jenny V Freeman, Steven A Julious

Introduction

In the previous tutorial we outlined the basic properties of the Normal distribution and discussed the Central Limit Theorem(Freeman and Julious 2005b). The Normal distribution is fundamental to many of the tests of statistical significance covered in subsequent tutorials. As a result of the principles of the Central Limit Theorem the Normal distribution enables us to calculate confidence intervals and make inference about the population from which the sample is taken. In this tutorial we explain the basic principles of **hypothesis testing** (using P-values) and **estimation** (using confidence intervals). By the end of the tutorial you will know of the processes involved and have an awareness of what a P-value is and what it is not, and what is meant by the phrase 'statistical significance'.

Statistical Analysis

It is rarely possible to obtain information on an entire population and usually data or information are collected on a sample of individuals from the population of interest. Therefore one of the main aims of statistical analysis is to use this information from the sample to draw conclusions ('**make inferences**') about the population of interest.

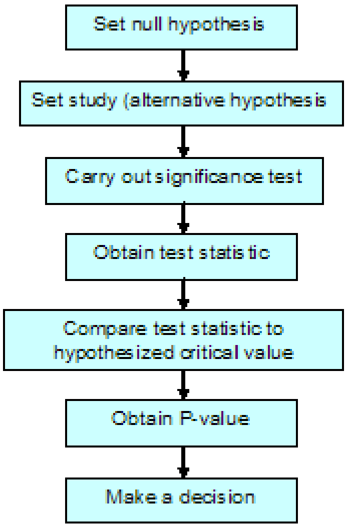
Consider the hypothetical example of a study designed to examine the effectiveness of two treatments for migraine. In the study patients were randomly allocated to two groups corresponding to either treatment A or treatment B. It may be that the primary objective of the trial is to investigate whether there is a difference between the two groups with respect to migraine outcome; in this case we could carry out a significance test and calculate a P-value (hypothesis testing). Alternatively it may be that the primary objective is to quantify the difference between treatments together with a corresponding range of plausible values for the difference; in this case we would calculate the difference in migraine response for the two treatments and the associated confidence interval for this difference (estimation).

Hypothesis Testing (using P-values)

Figure 1 describes the steps in the process of hypothesis testing. At the outset it is important to have a clear research question and know what the outcome variable to be compared is. Once the research question has been stated, the null and alternative hypotheses can be formulated. The null hypothesis ($\boldsymbol{H_0}$) usually assumes that there is no difference in the outcome of interest between the study groups. The study or alternative hypothesis ($\boldsymbol{H_1}$) usually states that there is a difference between the study groups.

In lay terms the null hypothesis is what we are investigating whilst the alternative is what we often wish to show. For example when comparing a new migraine therapy against control we are investigating whether there is no difference between treatments. We wish to prove that this null hypothesis is false and demonstrate that there is a difference at a given level of significance.

Figure 1: Hypothesis testing: the main steps


Set null hypothesis
Set study (alternative hypothesis
Carry out significance test
Obtain test statistic
Compare test statistic to hypothesized critical value
Obtain P-value
Make a decision

In general, the direction of the difference (for example: that treatment A is better than treatment B) is not specified, and this is known as a two-sided (or two-tailed) test. By specifying no direction we investigate both the possibility that A is better than B and the possibility that B is better than A. If a direction is specified this is referred to as a one-sided test (one-tailed) and we would be evaluating only whether A is better then B as the possibility of B being better than A is of no interest. It is rare to do a one-sided test as they have no power to detect a difference if it is in the opposite direction to the one being evaluated. We will not dwell further on the difference between two-sided and one-sided tests other than to state that the convention for one-sided tests is to use a level of significance of 2.5% - half that for a two-sided test. Usually in studies it is two-sided tests that are done.

A common misunderstanding about the null and alternative hypotheses, is that when carrying out a statistical test, it is the alternative hypothesis (that there is a difference) that is being tested. This is not the case – what is being examined is the null hypothesis, that there is no difference between the study groups; we conduct a hypothesis test in order to establish how likely (in terms of probability) it is that we have obtained the results that we have obtained, if there truly is no difference in the population.

For the migraine trial, the research question of interest is:

> ‘*For patients with chronic migraines which treatment for migraine is the most effective?*’

There may be several outcomes for this study, such as the frequency of migraine attacks, the duration of individual attacks or the total duration of attacks. Assuming we are interested in reducing the frequency of attacks, then the null hypothesis, H_o, for this research question is:

> ‘*There is no difference in the frequency of attacks between treatment A and treatment B groups’*

and the alternative hypothesis, H_1, is:

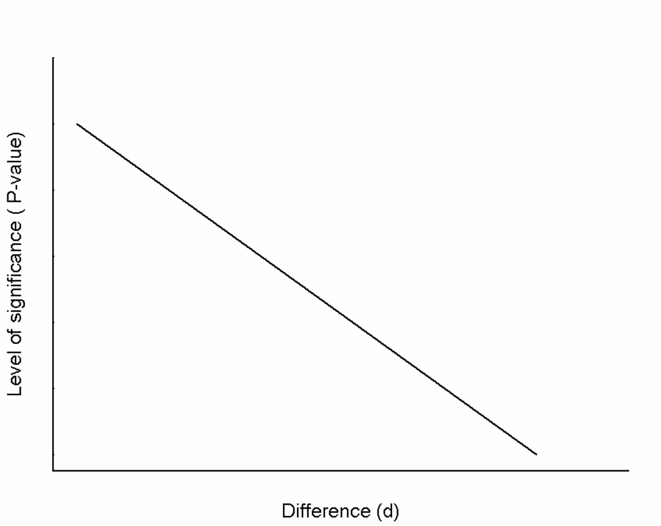
> '*There is a difference in the frequency of attacks between the two treatment groups*'.

Having set the null and alternative hypotheses the next stage is to carry out a significance test. This is done by first calculating a **test statistic** using the study data. This test statistic is then compared to a theoretical value under the null hypothesis in order to obtain a **P-value**. The final and most crucial stage of hypothesis testing is to make a decision, based upon the P-value. In order to do this it is necessary to understand first what a P-value is and what it is not, and then understand how to use it to make a decision about whether to reject or not reject the null hypothesis.

So what does a P-value mean? **A P-value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true**. Common misinterpretations of the P-value are that it is either the probability of the data having arisen by chance or the probability that the observed effect is not a real one. The distinction between these incorrect definitions and the true definition is the absence of the phrase ***when the null hypothesis is true***. The omission of 'when the null hypothesis is true' leads to the incorrect belief that it is possible to evaluate the probability of the observed effect being a real one. The observed effect in the sample is genuine, but what is true in the population is not known. All that can be known with a P-value is, if there truly is no difference in the population, how likely is the result obtained (from the sample). Thus a small P-value indicates that difference we have obtained is unlikely if there genuinely was no difference in the population – it gives the probability of obtaining the study results (or results more extreme) (difference between the two study samples) if there actually is no difference in the population.

In practice, what happens in a trial is that the null hypothesis that two treatments are the same is stated i.e. A=B or A-B=0. The trial is then conducted and a particular difference, d, is observed where A-B=d. Due to pure randomness even if the two treatments are the same you would seldom observe A-B=0. Now if d is small (say a 1% difference in the frequency of attacks) then the probability of seeing this difference under the null hypothesis is very high say P=0.995. If a larger difference is observed then the probability of seeing this difference by chance is reduced, say d=0.05 then the P-value could be P=0.562. As the difference increases therefore so the P-value falls such that a d=0.20 may equate to a P=0.021. This relationship is illustrated in Figure 2: as d increases then the P-value (under the null hypothesis) falls.

Figure 2. Illustration of the relationship between the observed difference and the P-value under the null hypothesis



It is important to remember that a P-value is a probability and its value can vary between 0 and 1. A ‘small’ P-value, say close to zero, indicates that the results obtained are unlikely when the null hypothesis is true and the null hypothesis is rejected. Alternatively, if the P-value is ‘large’, then the results obtained are likely when the null hypothesis is true and the null hypothesis is not rejected. ***But how small is small?*** Conventionally the cut-off value or **two-sided significance level** for declaring that a particular result is **statistically significant** is set at 0.05 (or 5%). Thus if the P-value is less than this value the null hypothesis (of no difference) is rejected and the result is said to be statistically significant at the 5% or 0.05 level (Box 1). For the example above, if the P-value associated with the mean difference in the number of attacks was 0.01, as this is less than the cut-off value of 0.05 we would say that there was a statistically significant difference in the number of attacks between the two groups at the 5% level.

Box 1: Statistical Significance

We say that our results are statistically significant if the P-value is less than the significance level (α), usually set at 5%

	P < 0.05	P≥0.05
Result is	Statistically signific nt	No statistically significant
Decide	That t re is sufficient evidence to reject the null hypothesis and accept the alternative hypothesis	That there is insufficient evidence to reject the null hypothesis

We cannot say that the null hypothesis is true, only that there is not enough evidence to reject it

The choice of 5% is somewhat arbitrary and though it is commonly used as a standard level for statistical significance its use is not universal. Even where it is, one study that is statistically significant at the 5% level is not usually enough to change practice; replication is required. For example to get a license for a new drug usually two statistically significant studies are required at the 5% level which equates to a single study at the 0.00125 significance level. It is for this reason that larger 'super' studies are conducted to get significance levels that would change practice i.e. a lot less than 5%.

Where the setting of a level of statistical significance at 5% comes from is not really known. Much of what we refer to as statistical inference is based on the work of R.A. Fisher (1890-1962) who first used 5% as a level of statistical significance acceptable to reject the null hypothesis. One theory is that 5% was used because Fisher published some statistical tables with different levels of statistical significance and 5% was the middle column. An exercise we do with students in order to demonstrate empirically that 5% is a reasonable level for statistical significance is to toss a coin and tell the students whether we've observed a head or a tails. We keep saying heads. After around 6 tosses we ask the students when they stopped believing we were telling the truth. Usually about half would say after 4 tosses and half after 5. The probability of getting 4 heads in a row is 0.063 and the probability of getting five heads in a row is 0.031; hence 5% is a figure about which most people would intuitively start to disbelieve an hypothesis!

Although the decision to reject or not reject the null hypothesis may seem clear-cut, it is possible that a mistake may be made, as can be seen from the shaded cells of Box 2. For example a 5% significance level means that we would only expect to see the observed difference (or one greater) 5% of the time under the null hypothesis. Alternatively we can rephrase this to state that even if the two treatments are the same 5% of the time we will conclude that they are not and we will make a Type I error. Therefore, whatever is decided, this decision may correctly reflect what is true in the population: the null hypothesis is rejected, when it is fact false or the null hypothesis is not rejected, when in fact it is true. Alternatively, it may not reflect what is true in the population: the null hypothesis may be rejected, when in fact it is true which would lead us to a **false positive** and making a **Type I error**, (**α**); or the null hypothesis may not be rejected, when in fact it is false. This would lead to a **false negative**, and making a **Type II error**, (**β**). Acceptable levels of the Type I and Type II error rates are set before the study is conducted. As mentioned above the usual level for declaring a result to be statistically significant is set at a two sided level of 0.05 prior to an analysis i.e. the type I error rate (α) is set at 0.05 or 5%. In doing this we are stating that the maximum acceptable probability of rejecting the null when it is in fact true (committing a type 1 error, α error rate) is 0.05. The P-value that is then obtained from our analysis of the data gives us the probability of committing a Type I error (making a false positive error).

Box 2: Making a decision

	The null hypothesis is actually:	
Decide to:	**False**	**True**
Reject the null hypothesis	Correct	Type 1 Error (α)
Not reject the null hypothesis	Type 2 Error (β)	Correct

This represents a **well powered study** – one that is able to detect a difference when there truly is a difference

The P value. This is the probability of concluding that there is a difference, when in fact there is no difference, i.e. the probability of

The probability that a study will be able to detect a difference, of a given size, if one truly exists is called the **Power** of the study and is the probability of rejecting the null hypothesis when it is actually false (probability of making a Type II error, β). It is usually expressed in percentages, so for a study which has 90% power, there is a probability of 0.9 of being able to detect a difference, of a given size, if there genuinely is a difference in the population. An underpowered study is one which lacks the ability, i.e. has very low power, to detect a difference when there truly is a difference. The concepts of power and Type I and II errors will be dealt with further in a later tutorial on sample size, as these are important components of sample size calculation.

Estimation (using confidence intervals)

Statistical significance does not necessarily mean the result obtained is clinically significant or of any practical importance. A P-value will only indicate how likely the results obtained are when the null hypothesis is true. It can only be used to decide whether the results are statistically significant or not, it does not give any information about the likely size of the clinical difference. Much more information, such as whether the result is likely to be of clinical importance can be gained by calculating a **confidence interval**. Although in the previous tutorial we talked about the 95% confidence interval for the mean, it is possible to calculate a confidence interval for any estimated quantity (from the sample data), such as the mean, median, proportion, or even a difference. It is a measure of the precision (accuracy) with which the quantity of interest is estimated (in the case of the migraine trial, the quantity of interest is the mean difference in the number of migraine attacks).

Technically, the 95% confidence interval is the range of values within which the true population quantity would fall 95% of the time if the study were to be repeated many times. Crudely speaking, the confidence interval gives a range of plausible values for the quantity estimated; although not strictly correct it is usually interpreted as the range of values within which there is 95% certainty that the true value in the population lies. For the migraine example, let us assume that the quantity estimated, the mean difference in the number of attacks between the groups, was 3 attacks per month and the 95% confidence interval for this difference was 1.2 to 4.8 attacks per month. Thus, whilst the best available estimate of the mean difference was 3 attacks per month, it could be as low as 1.2 or as high as 4.8 attacks per month, with 95% certainty. As the confidence interval excludes 0 we can infer from the observed trial that it is unlikely that there is no difference between treatments. In fact as we have calculated a 95% confidence interval we can deduce that the statistical significance is less than 5%. The actual P-value associated with this difference was 0.01 and given that it is less than 5% we can conclude that the difference is statistically significant at the 5% level.

As confidence intervals are so informative and from them we can infer statistical significance as well as quantify plausible values for the population effect there is a growing consensus that only confidence intervals should be reported for studies. However, it is unlikely that P-values will ever be eliminated as a way to quantify differences.

Statistical and Clinical Significance

So far in this tutorial we have dealt with hypothesis testing and estimation. However, in addition to statistical significance, it is useful to consider the concept of clinical significance. Whilst a result may be statistically significant, it may not be clinically significant (relevant/important) and conversely an estimated difference that is clinically important may not be statistically significant. For example consider a large study comparing two treatments for high blood pressure; the results suggest that there is a statistically significant difference (P=0.023) in the amount by which blood pressure is lowered. This P-value relates to a difference of 3mmHg between the two treatments. Whilst the difference is statistically significant, it could be argued that a difference of 3mmHg is not clinically important. This is supported but the 95% confidence interval of 2.3 to 3.7mmHg. Hence, although there is a statistically significant difference this difference may not be sufficiently large enough to convince anyone that there is a truly important clinical difference.

This is not simply a trivial point. Often in presentations or papers P-values alone are quoted and inferences about differences between groups are made based on this one statistic. Statistically significant P-values may be masking differences that have little clinical importance. Conversely it may be possible to have a P-value greater than the magic 5% but for there to be a genuine difference between groups: absence of evidence does not equate to evidence of absence.

Summary

In this tutorial we have outlined the basic principles of **hypothesis testing** (using P-values) and **estimation** (using confidence intervals). In subsequent tutorials we will be applying this knowledge when performing statistical significance testing in order to make decisions about the results of analyses.

Randomisation in Clinical Investigations

Steven A. Julious and Jenny V. Freeman

Introduction

In previous notes we have outlined methods for describing and summarising data, and the principles of hypothesis testing and estimation(Freeman and Julious 2005a;Freeman and Julious 2006b). In this note we will describe the basic concepts of randomisation in investigations. We will begin by describing the background, followed by describing the rationale for randomisation and then finally will move on to some of the more advanced topics of randomisation pertinent to imaging investigations.

Background

Allocation at random has been a central tenet of clinical trial design since the first reported modern clinical trial was conducted to investigate the effect of streptomycin and bed rest compared to bed rest alone in the treatment of tuberculosis(Bradford Hill 1990;Julious and Zariffa 2002;Medical Research Council 1948;Yoshioka 1998). Randomisation is important as it ensures that the regimen groups being investigated are objectively the same for any demographic or prognostic factors. Randomisation achieves this by ensuring that each subject has a known chance of receiving a given treatment in an allocation that can not be predicted(Altman and Bland 1999). This lack of predictability is important as an investigator should remain masked to the order of the treatments in order to reduce the potential for bias; only finding out what regimen a patient is to be assigned after recruiting a patient into the trial(Day and Altman 2000).

Note the concept of randomisation originally came from clinical trials hence the reference to treatments. As we will describe in this note though, randomisation is an important consideration for all types of clinical investigation. The problem of not allocating at random is evidenced by the following example(Julious and Mullee 1994). An historical trial was undertaken to compare the success of a new treatment [percutaneous nephrolithomy] with an existing treatment [open surgery] in the removal of kidney stones. Examining table 1a it appears that the new treatment is superior with an 83% success rate compared to only 78% on the old treatment. However, when we break the table down into small (table 1b) and large stones (table 1c) in each table the direction of the effect first observed is reversed. The old treatment is superior for both sizes of stones. The only reason why the old treatment seemed inferior to start with was that treatment is confounded with stone. This reversal effect is known as Simpson's Paradox(Simpson 1951;Williams 1949)).

Table 1. A comparison of the success rates of percutaneous nephrolithotomy (New) compared to open surgery (Old) in the removal of Kidney stones a) overall b) for stones <2cm and c) stones ≥2cm

a) Overall

		Success		
		Yes	No	Total
Treatment	New	289 (83%)	61 (17%)	350
	Old	273 (78%)	77 (22%)	350
	Total	562	138	700

b) Stones <2cm

		Success		
		Yes	No	Total
Treatment	New	234 (83%)	36 (17%)	270
	Old	81 (93%)	6 (7%)	87
	Total	315	42	357

c) Stones ≥2cm

		Success		
		Yes	No	Total
Treatment	New	55 (69%)	25 (31%)	80
	Old	192 (73%)	71 (27%)	263
	Total	247	96	343

Confounding is a statistical term for when there is a strong relationship between a third factor and both the outcome and comparison of interest. In the table people who had percutaneous nephrolithomy were also more likely to have small stones and the smaller the stone the better the prognosis. Hence, we would say treatment is confounded with stone size.

Obviously a bias of the magnitude observed with instances of Simpson's Paradox is rare but randomisation protects the investigator from confounding with known and unknown prognostic factors. Therefore, wherever possible, subjects should be assigned to investigations at random. If there are known factors that could effect the outcome, such as centre, age, sex, or baseline risk then the study should be stratified to allow for these and a block size (see below) should be set that provides balance within each strata (see below). If there is to be a constraint in the randomisation, such as unequal allocation then this should be allowed for in the block size and appropriate adjustment made to the sample size. Block size and strata are described below.

Mechanics of Randomisation

Parallel Group Trials

A *parallel group trial* is one in which there are to be at least two arms to be investigated and subjects are to be randomised to each of these arms. It is beyond the scope of this note to describe in detail how to undertake a randomisation however we will give some general hints and tips (used in this context arm is a generic term to describe the groupings in trials. Subjects may be assigned to two

different arms where these arms could be: treatments; assessors or imaging protocols).
When randomising subjects to the different arms in the trial an important consideration is to maintain balance for the interventions to which subjects are being randomised. This is particularly important in small studies where by chance there can easily be an imbalance in the number of subjects on the respective arms. One tool to ensure that groups are balanced is to do introduce "blocks" into the randomisation. Basically a block is a sample size after which there is balance in the randomisation. It is best to illustrate this through a simple worked example.

Consider the case of two groups. We wish to randomly allocate individuals to either group A or group B. In this example we could toss a coin and record either heads (H) or tails (T), so that we can then use the order to allocate individuals to groups (i.e. if heads then group A, if tails then group B. If we set the block size to be 4 we need to ensure that after every four tosses there are two heads and two tails. Thus:

Block 1: T T (H H)
Block 2: T T (H H)
Block 3: T H T (H)
Block 4: T H H (T)

The terms in brackets are not from tosses but entries we were forced to enter to ensure balance. For example in Block 1 the first two tosses were tails. We thus made the next 2 heads so that after "4 tosses" we had a balance. Notice after "16 tosses" by blocking we have 8 heads and 8 tails.

Another important consideration is stratification. Stratification is similar to blocking but here as well as ensuring balance after a requisite block size we also ensure balance by strata. These strata are usually clinical important sub groups such as sex and age.

Again it is best to illustrate by example. Suppose we are doing the same coin tossing to create a randomisation list. For this randomisation we wish to ensure balance for a two level stratification factor. Operationally this would be the same as doing the coin tossing exercise twice: once for each stratum.

Stratum 1
    Block 1: T T (H H)
    Block 2: T T (H H)

Stratum 2
    Block 1: T H T (H)
    Block 2: T H H (T)

Now after "16 tosses" we have balance both in terms of heads and tails and also for heads and tails by strata.

A final consideration, as discussed earlier in the note, is the withholding of the randomisation until the actual allocation of subjects. Even for completely open studies it is preferable to mask the randomisation so that investigators only find out what regimen a patient is to be assigned after the patient has been recruited. In practice this could be done by putting the randomisation in envelopes which are opened only after a subject has been enrolled.

Cross-over Trials

The distinction between *parallel group designs* and *crossover designs* is that in parallel group designs subjects are assigned at random to receive only one investigation, and as a result of the randomisation the groups are the same in all respects other than the investigation made. However, with a cross-over trial all subjects receive all the investigations but it is the order that subjects receive the investigations which is randomised. The big assumption here is that prior to starting each investigation all subjects return to baseline and that the order in which subjects have their investigation does not affect their response to the investigation.

Two Period

Two period cross over trials are the easiest to explain. In the simplest case, for a two arm investigation (comparing A with B say) subjects will be randomised to either A followed by B (AB) or to B followed by A (BA). AB and BA are called sequences and represent the order in which subjects receive the investigations. In practice, subjects are randomly assigned to the either the sequence AB or the sequence BA, and to ensure balance blocking can still be used.

Note that even for retrospective investigations, randomisation should be considered. For example, in a study to investigate the agreement between two image analysts the analysts could have the images assessed randomly with the analysts reading the images in random order much like a AB/BA design.

Multi-period

All investigations are made on all subjects

Imaging comparisons can be complicated as there are often a finite number of subjects on whom a number of investigations are made, such as:

- A fMRI investigation where subjects will receive a number of challenges.
- A comparison of different imaging protocols within the same subject.
- An assessment of new technology such as a comparison of 2D, 3D and 2D and 3D combined SPECT.
- A comparison of several readers on the same subjects to look at agreement.
- A comparison of different therapies or different doses of the same therapy within a subject.

It is quite easy therefore for 4 or 5 investigations to be made on the same subject. If 4 investigations are made on the same subject, that would result in 24 different ways of assigning subjects to these four investigations and hence 24 sequences. This is all very good but what if we have only 12 subjects in the trial?

Actually, for multi period investigations we do not necessarily need to use all possible sequences but can form special sequences to be randomised called Williams Squares(Williams 1949)

It is again best to illustrate through example. In order to investigate an even number of investigations we can build a Williams square from the following sequence:

0, 1, t, 2, t-1, 3, t-2…etc

where t is the number of interventions minus 1. If we were to conduct 4 investigations then t=3 and our sequences would include 0, 1, 2, 3. We build the sequences by forming the first row from the result above. We then form the second by adding 1 to this first row, but where the number is 3 the new number becomes 0 (we are adding in base 3). The calculation is simpler than the explanation

0	1	3	2
1	2	0	3
2	3	1	0
3	0	2	1

This is known as a Latin Square: each investigation appears in every row and column. The columns here would reflect different imaging sessions. A Williams Square is special form of Latin Square such that as well as being balanced for rows and columns each investigation is preceded by each other investigation at least once e.g. 1 is preceded by 0, 2, and 3. Here we are saying that as well the order of investigations being important the effect of preceding investigations is too. Hence we ensure balance for the immediately preceding investigation. This is known as first order balance.

If we were conducting a trial where we are to undertake four different investigations on 12 subjects we would randomise the 4 sequences above so each sequences appears 3 times.

For an odd number of investigations we need to build 2 Latin squares with starting sequences

0, 1, t, 2, t-1, 3, t-2… etc

and

…t-2, 3, t-1, t, 1, 0.

With 5 investigations, t=4 and we would therefore have

0	1	4	2	3
1	2	0	3	4
2	3	1	4	0
3	4	2	0	1
4	0	3	1	2

And

3	2	4	1	0
4	3	0	2	1
0	4	2	3	2
1	0	3	2	3
2	1	3	3	4

Not all investigations are made on all subjects

In imaging investigations there are logistical, practical and safety considerations to be taken into account. For example we may wish to investigate 4 different imaging protocols but these must all be done in one day for each subject and for practical reasons we can only schedule 3 scans in a day. Similar we may wish to look at 4 different protocols but for safety reasons we may only be able to do 3 scans in the 24 hours we have each subject. Although we can still construct Latin Squares we need to construct a special type of these known as a Balanced Incomplete Block. Again we will illustrate by example.

If we could have 3 sessions for each subject but we have 4 investigations, then taking the sequences derived previously and removing the first column

0	1	3	2
1	2	0	3
2	3	1	0
3	0	2	1

and final column

0	1	3	2
1	2	0	3
2	3	1	0
3	0	2	1

would give us 8 sequences as follows

1	3	2
2	0	3
3	1	0
0	2	1

0	1	3
1	2	0
2	3	1
3	0	2

We would hence have balance both for rows and columns as well as for first order balance within 8 sequences.

For an odd number of sequences we use a similar procedure. Using our previous example of have 5 investigations and assuming that we can only do 3 sessions then we could delete the last 2 columns off the first 5 sequences and the first 2 columns off then next i.e.

0	1	4	2	3
1	2	0	3	4
2	3	1	4	0
3	4	2	0	1
4	0	3	1	2

3	2	4	1	0
4	3	0	2	1
0	4	2	3	2
1	0	3	2	3
2	1	3	3	4

Discussion

This note introduced the basic concepts of randomisation including the importance of stratification and blocking. We then described issues pertinent to imaging investigations where we may wish to perform multiple investigations on each subject or the special case where the number of investigations is greater than the number of sessions.

Basic tests for continuous data

Jenny V Freeman, Steven A Julious

As it is rarely possible to study an entire population, data are usually collected from a sample of individuals in order to make inferences, or draw conclusions, about the population of interest. This can be done through a process known as hypothesis testing, the basic principles of which have been outlined in a previous tutorial(Freeman & Julious 2006b). At the outset it is important to have a clear research question and know what the outcome variable to be compared is. Once the research question has been stated, the null and alternative hypotheses can be formulated. The null hypothesis (H_0) assumes that there is no difference in the outcome of interest between the study groups. The study or alternative hypothesis (H_1) states that there is a difference between the study groups. Next the appropriate statistical test must be selected and conducted to obtain a P-value. This P-value will then be used to make a decision about whether the results are statistically significant and whether the null hypothesis can be rejected.

This tutorial will provide a concrete example of how the process of setting and testing a hypothesis is implemented in practice. It will focus on some elementary methods for analysing continuous data: the paired and unpaired t-tests and their non-parametric equivalents. Continuous data are data that can be measured and can take any value on the scale on which they are measured; examples include height, weight, blood pressure and area coverage.

Choosing the statistical method

What type of statistical analysis depends on the answer to five key questions (Box 1) and given answers to these, an appropriate approach to the statistical analysis of the data collected can be decided upon. The type of statistical analysis depends fundamentally on what the main purpose of the study is. In particular, what is the main question to be answered? The data type for the outcome variable will also govern how it is to be analysed, as an analysis appropriate to continuous data would be completely inappropriate for binary data. In addition to what type of data the outcome variable is, its distribution is also important, as is the summary measure to be used. Highly skewed data require a different analysis compared to data which are *Normally* distributed.

Box 1: Five key questions to ask:

1. What are the aims and objectives?
2. What is the hypothesis to be tested?
3. What type of data is the outcome data?
4. How is the outcome data distributed?
5. What is the summary measure for the outcome data?

Before beginning any analysis it is important to examine the data, using the techniques described in the first two tutorials in this series(Freeman & Julious 2005a;Freeman & Julious 2005c); adequate description of the data should precede and complement the formal statistical analysis. For most studies and for RCTs in particular, it is good practice to produce a table or tables that describe the initial or baseline characteristics of the sample.

Comparison of two independent groups

Independent samples t-test

The independent samples *t*-test is used to test for a difference in the mean value of a continuous variable between two independent groups. For example, as part of a randomised clinical trial of two treatments for venous leg ulcers one of the main questions of interest was whether there was a difference in the number of ulcer free weeks between the control and the clinic groups(Morrell, Walters, Dixon, Collins, Brereton, Peters, & Brooker 1998). As the number of ulcer free weeks is continuous data and there are two independent groups, assuming the data are Normally distributed in each of the two groups, then the most appropriate summary measure for the data is the sample mean and the best comparative summary measure is the difference in the mean number of ulcer free weeks between the two groups. When conducting any statistical analysis it is important to check that the assumptions which underpin the chosen method are valid. The assumptions underlying the two-sample *t*-test are outlined in Box 2. The assumption of Normality can be checked by plotting two histograms, one for each sample; these do not need to be perfect, just roughly symmetrical. The two standard deviations should also be calculated and as a rule of thumb, one should be no more than twice the other.

Box 2: The assumptions underlying the use of the independent samples *t*-test:

1. The groups are independent.
2. The variables of interest are continuous.
3. The data in both groups have similar standard deviations.
4. The data is Normally distributed in both groups.

The test statistic for the independent samples t-test, t, is calculated as follows:

$$t = \frac{\left(\bar{x}_1 - \bar{x}_2\right)}{\sqrt{\left(\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}\right)}}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the means of the two groups and *n₁* and *n₂* are the numbers in the two groups and s_p^2 = $\frac{\left(n_1 - 1\right)s_1^2 + \left(n_2 - 1\right)s_2^2}{n_1 + n_2 - 2}$ and is an estimate of the pooled variance. Once this test statistics has been calculated it can be compared to values for the t distribution on $n_1 + n_2 - 2$ degrees of freedom. These can either be found in books of statistical tables, or using the TDIST function in EXCEL. This function requires 3 arguments, X = the t statistic obtained above, deg_freedom = $n_1 + n_2 - 2$ and tails, where 1 indicates a one sided test and 2 indicates a two sided test. It is recommended that you always use 2 here to ensure that you test is two sided. Note that t can take negative as well as positive values and as the number of degrees of freedom gets larger the t distribution approaches the Normal distribution.

For the leg ulcer data, there were 120 patients in the clinic group and their mean number of ulcer free weeks for was 20.1. There were 113 patients in the control group and they had a mean number of ulcer free weeks was 14.2. The pooled estimate of variance, s_p^2, was 5.57. Putting these into the formula above gives a t statistic of 2.49 on 231 degrees of freedom and this in turn gives a P-value of 0.014. The final and most crucial stage of hypothesis testing is to make a decision, based upon this P value. *A P value is the probability of obtaining the study results (or results more extreme) if the null hypothesis is true*. It tells you how likely is the result obtained (from the study data), if there truly is no difference in the population. A ‘small’ P value, say close to zero, indicates that the results obtained are unlikely when the null hypothesis is true and the null hypothesis is rejected. Alternatively, if the P value is ‘large’, then the results obtained are likely when the null hypothesis is true and the null hypothesis is not rejected. Conventionally the cut-off value or *significance level* for declaring that a particular result is *statistically significant* is set at 0.05 (or 5%). Thus if the P value is less than this value the null hypothesis (of no difference) is rejected and the result is said to be statistically significant at the 5% or 0.05 level. For the example above, of the difference in the number of ulcer free weeks, the P value is 0.014. As this is less than the cut-off value of 0.05 there is said to be a statistically significant difference in the number of ulcer-free weeks between the two groups at the 5% level.

Mann-Whitney U test

There are several possible approaches when at least one of the requirements for the *t*-test is not met. The data may be transformed (e.g. the logarithm transformation can be useful particularly when the variances are not equal) or a *non-parametric method* can be used. Non-parametric or distribution free methods do not involve distributional assumptions i.e. making assumptions about the manner in which the data are distributed (for example that the data are Normally distributed). An important point to note is that it is the test that is parametric or non-parametric, not the data.

When the assumptions underlying the *t*-test are not met, then the non-parametric equivalent, the Mann-Whitney *U* test, may be used. Whilst the independent samples *t*-test is specifically a test of the null hypothesis that the groups have the same mean value, the Mann-Whitney *U* test is a more general test of the null hypothesis that the distribution of the outcome variable in the two groups is the same; it is possible for the outcome data in the two groups to have similar measures of central tendency or location, such as mean and medians, but different distributions.

The Mann-Whitney *U* test requires all the observations to be ranked as if they were from a single sample. From this the statistic *U* is calculated; it is the number of all possible pairs of observations comprising one from each sample for which the value in the first group precedes a value in the second group. This test statistic is then used to obtain a P value. The principle is best illustrated with a simple example. Consider the following two samples of size 6 X=(0,6,5,1,1,6) and Y=(9,4,7,8,3,5). Rank these in order as if they were from the same sample:

0 1 1 3 4 5 **5 6 6** 7 8 9

Having ranked the samples, choose one sample, for example X and count the number of observations from Y that are below each of the X observations. The first X value is 0 and there are 0 Y observations below this. The next X value is 1 and there are 0 Y values below this. Where there are ties, these can be disregarded and do not contribute to the total. Continue for each member of X and when this has been done the numbers of preceding Ys can be added to obtain the U statistic, U_x=0+0+0+2+3+3=8. This procedure is repeated for the second sample and then the smaller of the two numbers is used obtain a P-value. As with the t statistic above this value is compared to tabulated critical values under the null hypothesis to obtain a P-value. For the data above, the P-value obtained from the Mann-Whitney U test is 0.12. As this is greater than 0.05 the result is not statistically significant, and there that there is insufficient evidence to reject the null that the two groups differ in terms of their location.

Two groups of paired observations

When there is more than one group of observations it is vital to distinguish the case where the data are paired from that where the groups are independent. Paired data may arise when the same individuals are studied more than once, usually in different circumstances, or when individuals are paired as in a case-control study. As part of leg ulcer trial the researchers were interested in assessing whether there was a change in health related quality of life (HRQoL) between baseline and 3 months for those individuals with a healed leg ulcer (irrespective of study group). HRQoL at baseline and 3 months are both continuous variables and the data are paired as measurements are made on the same individuals at baseline and 3 months; therefore, interest is in the mean of the differences not the difference between the two means.

Paired t-test

If we assume that the paired differences are Normally distributed, then the best comparative summary measure is the mean of the paired difference in HRQoL between baseline and 3 months. Given the null hypothesis (H_0) that there is no difference (or change) in mean HRQoL at baseline and 3 months follow-up in patients whose leg ulcer had healed by 3 months, the most appropriate test is the paired *t*-test. There were 36 patients with a healed leg ulcer at 3 months.

The test statistic for the paired t test is again *t* and is calculated as $t = \frac{\bar{d}}{se(\bar{d})}$ where $\bar{d}$ is the mean of the paired differences and $se(\bar{d})$ is the standard error of $\bar{d}$ and is estimated as $\frac{sd(\bar{d})}{\sqrt{n}}$ and *n* is the number of paired differences. As with the unpaired case this t statistic can then be compared to values for the t distribution on *n*-1 degrees of freedom. The mean change (3 months – baseline) in HRQoL for the 36 patients with healed ulcers was -7.33 with an SD of 16.5 and using these in the formulae above gives a t value of -2.661, which in turn gives a P-value of 0.012. As this is less than the nominal level usually set for statistical significance of 0.05 (or 5%) we can conclude that there is a statistically significant difference in HRQoL between baseline and 3 months. It is worth noting that as the mean change is negative HRQoL actually declined for these patients over the 3 months!

Box 3: The assumptions underlying the use of the paired *t*-test

1. The differences are plausibly Normally distributed (it is not essential for the original observations to be Normally distributed).
2. The differences are independent of each other.

Wilcoxon signed rank sum test

The assumptions underlying the use of the paired *t*-test are outlined in Box 3. If these are not met a non-parametric alternative, the *Wilcoxon signed rank sum test*, can be used. This test is based upon the ranks of the paired differences and test the null hypothesis that the median of these differences is 0 i.e that there is no tendency for the outcome in one group (or under one condition) to be higher or lower than in the other group (or condition). For the leg ulcer trial the null hypothesis would state that there is no tendency for HRQol life at baseline to be better or worse than HRQoL at 3 months. An explanation of how to carry out a Wilcoxon signed rank sum test in practice can be found in Swinscow and Campbell(Swinscow and Campbell 2002). For the leg ulcer data the P-value is 0.012, very similar to that for the paired t-test above. As it is less than 0.05 we would conclude that there is sufficient evidence to reject the null and conclude that the median difference is not equal to 0.

Summary

Outlined above are some simple methods for comparing two groups of continuous data. However, it is important to bear in mind that s*tatistical significance does not necessarily mean the result obtained is clinically significant or of any practical importance***.** A P value will only indicate how likely the results obtained are when the null hypothesis is true. Much more information, such as whether the result is likely to be of clinical importance can be gained by calculating a *confidence interval*, as this a range of plausible values for the estimated quantity. In the next tutorial we will extend the methods outlined above to cover more than two groups.

Basic tests for continuous data: Mann-Whitney U and Wilcoxon signed rank sum tests

Jenny V Freeman, Michael J Campbell

The most recent tutorial examined how the process of setting and testing a hypothesis could be implemented in practice(Freeman and Julious 2006a). It focussed on some elementary methods for analysing continuous data: the paired and unpaired t-tests. However, these tests make particular assumptions about the distribution of the data. Most importantly that the standard deviations are similar (for the independent groups t-test) and that the data to be analysed are approximately Normally distributed (both tests).

This tutorial will discuss some alternative methods that can be used when these assumptions are violated. They are part of a group of statistical tests known as non-parametric or distribution-free tests; distribution-free tests do not involve making any assumptions about how the data are distributed (for example that the data are Normally distributed). An important point to note is that it is the test that is parametric or non-parametric, not the data.

Mann-Whitney U test

When the assumptions underlying the independent samples *t*-test are not met, then the non-parametric equivalent, the Mann-Whitney *U* test, may be used. Whilst the independent samples *t*-test is specifically a test of the null hypothesis that the groups have the same mean value, the Mann-Whitney *U* test is not a test for a difference in medians, as is commonly thought. It is a more general test of the null hypothesis that the distribution of the outcome variable in the two groups is the same; it is possible for the outcome data in the two groups to have similar measures of central tendency or location, such as mean and medians, but different distributions. Consider for example two groups of size 50; group A has 48 observations with value 0 and 2 with value 1 whilst group B has 26 observations with value 0 and 24 with a value of 2. Both groups have a median value of 0 but the p-value from the Mann-Whitney U test is < 0.001, indicating that the distribution of data in two groups is different.

The Mann-Whitney *U* test requires all the observations (for both groups combined) to be ranked as if they were from a single sample. From this the test statistic *U* is calculated; it is the number of all possible pairs of observations comprising one observation from each sample for which the rank of value in the first group precedes the rank of the value in the second group. This test statistic is then used to obtain a P value.

The principle is best illustrated with a simple example. Consider the following two samples of size six X=(0,6,5,1,1,6) and Y=(9,4,7,8,3,5). These are then ranked in order as if they were from the same sample (the values for sample X are given in bold):

Values	**0**	**1**	**1**	3	4	5	**5**	**6**	**6**	7	8	9
Ranks	**1**	**2.5**	**2.5**	4	5	6.5	6**.5**	**8.5**	**8.5**	10	11	12

Having ranked the values altogether, these ranks are then added up separately for each sample to get two separate totals (U statistics), U_x=29.5 and U_y=48.5. A useful check is that the sum of the ranks should add to n(n+1)/2. In this case n(n+1)=12(12+1)/5=78. The smaller of the two U statistics is used obtain a P-value; thus the value of U used for this example is 29.5. As with the t statistic above this value is compared to tabulated critical values under the null hypothesis (table 1) to obtain a P-value. Rank totals greater than the tabulated critical values are not significant. In this case n_1 and n_2 are both 6 and the tabulated critical value is 26. As the value of 29.5 is greater than this, the results do not reach statistical significance at the 5% level, and there that there is insufficient evidence to reject the null that the two groups differ in terms of the distribution of their data.

Table 1: Mann-Whitney test on unpaired samples: 5% levels of P (taken from Swinscow and Campbell(Swinscow & Campbell 2002))

n_1→ n_2 ↓	2	3	4	5	6	7	8	9	10	11	12	13	14	15
4			10											
5		6	11	17										
6		7	12	18	26									
7		7	13	20	27	36								
8	3	8	14	21	29	38	49							
9	3	8	15	22	31	40	51	63						
10	3	9	15	23	32	42	53	65	78					
11	4	9	16	24	34	44	55	68	81	96				
12	4	10	17	26	35	46	58	71	85	99	115			
13	4	10	18	27	37	48	60	73	88	103	119	137		
14	4	11	19	28	38	50	63	76	91	106	123	141	160	
15	4	11	20	29	40	52	65	79	94	110	127	145	164	185
16	4	12	21	31	42	54	67	82	97	114	131	150	169	
17	5	12	21	32	43	56	70	84	100	117	135	154		
18	5	13	22	33	45	58	72	87	103	121	139			
19	5	13	23	34	46	60	74	90	107	124				
20	5	14	24	35	48	62	77	93	110					
21	6	14	25	37	50	64	79	95						
22	6	15	26	38	51	66	82							
23	6	15	27	39	53	68								
24	6	16	28	40	55									
25	6	16	28	42										
26	7	17	29											
27	7	17												
28	7													

The previous tutorial illustrated the use of the independent samples t-test with some data taken from a community leg ulcer trial(Morrell, Walters, Dixon, Collins, Brereton, Peters, & Brooker 1998). For the leg ulcer data, there were 120 patients in the clinic group and their mean number of ulcer free weeks for was 20.1. There were 113 patients in the control group and they had a mean number of ulcer free weeks of 14.2. It was demonstrated that there was a statistically significant difference in the number of ulcer free weeks between the two groups (P=0.014). However, if the number of ulcer free weeks in each group is plotted it can be seen that the data are highly skewed and are not Normally distributed (Figure 1a and 1b).

Figure 1a: Ulcer-free time for clinic group

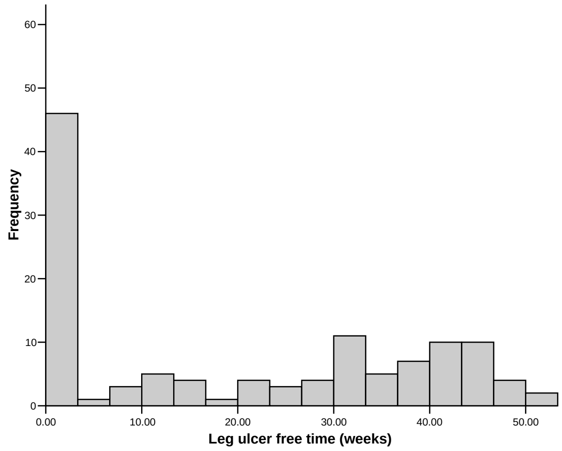
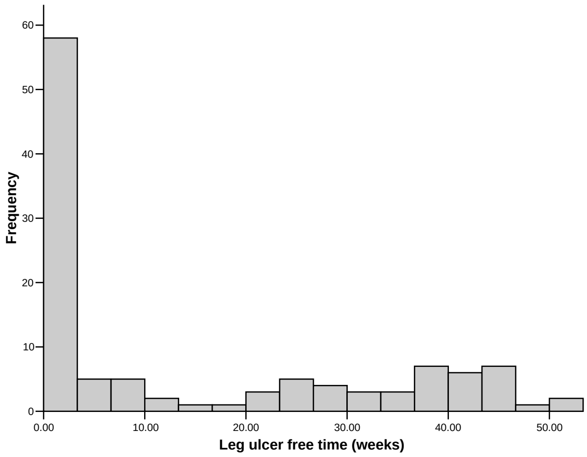


Figure 1b: Ulcer-free time for home group



If, instead of the independent samples t-test a Mann-Whitney U test were carried out on these data the P-value obtained would be 0.017, a value that is remarkably similar to that obtained from the t-test. In fact, the t-test and the Mann-Whitney U test will tend to give similar P-values when the samples are large and approximately equal in size). As this is less than the nominal level usually set for statistical significance of 0.05 we can reject the null hypothesis (that the distribution of the data in the two groups are the same). We conclude that the result is statistically significant and there is evidence that the distribution of ulcer free weeks is different between the two groups. However, we are unable to state what the difference might be, only that there is a difference, if we only consider the P-value.

Two groups of paired observations

When there is more than one group of observations it is vital to distinguish the case where the data are paired from that where the groups are independent. Paired data may arise when the same individuals are studied more than once, usually in different circumstances, or when individuals are paired as in a case-control study. For example, as part of the leg ulcer trial, data were collected on health related quality of life (HRQoL) at baseline, 3 months and 12 months follow-up. The previous tutorial described a method for analysing paired continuous data, the paired t-test.

If the assumptions underlying the use of the paired *t*-test are not met a non-parametric alternative, the *Wilcoxon signed rank sum test*, can be used, This test is based upon the ranks of the paired differences and the null hypothesis is that there is no tendency for the outcome in one group (or under one condition) to be higher or lower than in the other group (or condition). It assumes that (a) the paired differences were independent of each other and (b) the differences come from a symmetrical distribution (this can be checked by eye). As with the Mann-Whitney U test outlined above the Wilcoxon signed rank sum test is most easily illustrated using an example. Swinscow and Campbell (Swinscow & Campbell 2002) give details of a study of foetal movements before and after chorionic villus sampling. The data are shown in table 2:

Table 2 Wilcoxon test on percentage of time foetus spent moving before and after chononic villus sampling for ten pregnant women (Boogert et al. 1987)

Patient no	Before Sampling (2)	After Sampling (3)	Difference (before-after) (4)	Rank (5)	Signed rank (6)
1	25	18	7	9	9
2	24	27	-3	5.5	-5.5
3	28	25	3	5.5	5.5
4	15	20	-5	8	-8
5	20	17	3	5.5	5.5
6	23	24	-1	1.5	-1.5
7	21	24	-3	5.5	-5.5
8	20	22	-2	3	-3
9	20	19	1	1.5	1.5
10	27	19	8	10	10

The differences between before and after sampling are calculated (colum 4) and these are then ranked by size irrespective of sign (column 5; zero values omitted). When two or more differences are identical each is allotted the point half way between the ranks they would fill if distinct, irrespective of the plus or minus sign. For instance, the differences of –1 (patient 6) and +1 (patient 9) fill ranks 1 and 2. As (1 + 2)/2 = 1.5, they are allotted rank 1.5. In column (6) the ranks are repeated for column (5), but to each is attached the sign of the difference from column (4). A useful check is that the sum of the ranks must add to n(n + 1)/2. In this case 10(10 + 1)/2 = 55.

The numbers representing the positive ranks and the negative ranks in column (6) are added up separately and only the smaller of the two totals (irrespective of its sign) is used to obtain a P-value from tabulated critical values under the null hypothesis (Table 3). As with the Mann-Whitney U test rank totals greater then the tabulated critical value are non-significant at the 5% level. In this case the smaller of the two ranks is 23.5 and as this is larger than the number given for ten pairs in table 3 the result is not statistically significant. There is insufficient evidence to reject the null that the median difference in foetal movements before and after sampling is zero. We can conclude that we have little evidence that chorionic villus sampling alters the movement of the foetus.

Table 3: Wilcoxon test on paired samples: 5% and 1% levels of P (taken from Swinscow and Campbell(Swinscow & Campbell 2002))

Number of pairs	5% level	1% level
7	2	0
8	2	0
9	6	2
10	8	3
11	11	5
12	14	7
13	17	10
14	21	13
15	25	16
16	30	19

Note, perhaps contrary to intuition, that the Wilcoxon test, although a test based on the ranks of the data values, may give a different value if the data are transformed, say by taking logarithms. Thus it may be worth plotting the distribution of the differences for a number of transformations to see if they make the distribution appear more symmetrical.

Summary

Outlined above are some non-parametric methods for comparing two groups of continuous data when the assumptions underlying the t-test (paired and unpaired) are not met. However, as stated in the previous tutorial s*tatistical significance does not necessarily mean the result obtained is clinically significant or of any practical importance***.** A P value will only indicate how likely the results obtained are when the null hypothesis is true. Much more information, such as whether the result is likely to be of clinical importance can be gained by calculating a *confidence interval*, as this a range of plausible values for the estimated quantity. Details of how to do this can be found in Statistics with Confidence(Altman et al. 2000)

The analysis of categorical data

Jenny V Freeman, Steven A Julious

Introduction

In the previous two tutorials we have outlined several methods for analysing potential differences between two independent groups for a continuous outcome variable (Freeman and Campbell 2006;Freeman & Julious 2006a). In this tutorial we will discuss several methods for analysing differences between two independent groups when the outcome of interest is binary; a comparison of the two proportions using the Normal approximation to the Binomial and the Chi-squared test. If you recall from the earliest tutorial(Freeman & Julious 2005c), binary data are data which can take only two possible values such as healed / not healed or pregnant/ not pregnant.

The type of statistical test depends upon the answers to five key questions, as outlined in tutorial six (Freeman & Julious 2006a). Briefly, these are:

- The aims and objectives of the study
- The hypothesis to be tested
- The type of outcome data
- The distribution of the outcome data
- The appropriate summary measure for the outcome data.

Example data

The data in this tutorial have come from a randomised controlled trial of community leg ulcer clinics(Morrell, Walters, Dixon, Collins, Brereton, Peters, & Brooker 1998). Patients were allocated to one of two groups; either usual care by the district nursing team (control), or care in a specialist leg ulcer clinic (clinic). The aim of the study was to compare the treatments. One of the outcomes from the trial was whether the leg ulcer was healed at 12 weeks and thus the null hypothesis is that there is no difference between the two groups with respect to leg ulcer healing. The alternative hypothesis is that there is a difference between the two groups.

In this case the outcome is binary (healed/ not healed) and the data are considered to have a binomial distribution. The comparison is between two independent groups (control and clinic groups). Of the 120 patients in the clinic group, 22 (18%) had a healed leg ulcer at 12 weeks and of the 113 patients in the control group, 17 (15%) had a healed leg ulcer at 12 weeks.

These data consist of the frequencies in each group with or without healed ulcers and the simplest way to present them is in the form of a table such as table 1 below. This is known as a 2 by 2 contingency table, because there are two rows and two columns and it is said to have 4 cells (2 x 2). Generally a contingency with r rows and c columns is known as an r by c contingency table and has r x c cells. For example, if there were three treatment groups (rather than the two here) Table 1 would have 3 columns for the groups rather than two and thus the table would be a 2 by 3 contingency table (with 6 cells).

Table 1: 2 x 2 contingency table of Treatment (clinic/home) by Outcome (ulcer healed / not healed) for the Leg ulcer study data[2]

	Treatment		
	Clinic	Home	Total
Outcome:			
Healed	22 (18%)	17 (15%)	39
Not healed	98 (82%)	96 (85%)	194
Total	120 (100%)	113 (100%)	233

Comparison of two proportions

There are several approaches to analyse these data. One of the simplest is a comparison of the proportion healed between the two groups (technically this is known as the Normal approximation to the binomial distribution). In this case the hypothesis test assumes that there is a common proportion (of healed ulcers), π, which is estimated by p where p is the proportion of the total for both groups with a healed leg ulcer. Consider the general form of the table above:

Table 2: General form of 2 x 2 table

	Treatment group		
	1	2	Total
Outcome:			
1	a (=n_1p_1)	c (=n_2p_2)	a+c (=np)
2	b (=$n_1(1-p_1)$)	d (=$n_2(1-p_2)$)	b+d (=n(1-p))
Total	a+b (=n_1)	c+d (=n_2)	a+b+c+d (= n)

The common proportion p = a+c / (a+b+c+d)

If we let

$p_1 = a / (a+b)$ and $p_2 = c / (c+d)$

the common proportion, p, can also be written as:

$$\frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

and the standard error for the difference in proportions is estimated by:

$$se(p_1 - p_2) = \sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

From this we can compute the test statistic

$$z = \frac{p_1 - p_2}{se(p_1 - p_2)},$$

[2] When organising data such as this is it good practice to arrange the table with the grouping variable forming the columns and the outcome variable forming the rows.

which, under the null hypothesis (of no difference) is assumed to be Normally distributed, with a mean of 0 and a standard deviation of 1 (i.e. a standard Normal distribution). The reason we can make this Normal approximation is due to the Central Limit theorem discussed in a previous note(Freeman & Julious 2005b). We can then compare this value to what would be expected under the null hypothesis of no difference, in order to get a P-value.

For the example above

$$p = \frac{(120 * 0.183) + (113 * 0.150)}{120 + 113} = \frac{38.91}{233} = 0.167$$

$$se(p_1 - p_2) = \sqrt{0.167(1 - 0.167)\left(\frac{1}{120} + \frac{1}{113}\right)} = 0.049$$

and thus $z = \frac{0.033}{0.049} = 0.673$

Comparing this z statistic value to the value expected under the null hypothesis gives a P-value of 0.502 (i.e. the probability of observing z=0.67 or a value more extreme if the null hypothesis is true is 0.502). As this is greater than 0.05, we are unable to reject to null and we would conclude that there is no reliable evidence of a difference in leg ulcer healing rates at 12 weeks between the clinic and control groups. In addition to carrying out a significance test we could also calculate a confidence interval for the difference in proportions.

The 95% confidence interval for the difference in proportions is given by

$$(p_1 - p_2) \pm \left[1.96 \times se(p_1 - p_2)\right].$$

Using the data from the example above, the 95% confidence interval for the true difference in leg ulcer healing rates between the two groups is thus 0.063 to 0.129. Therefore we are 95% confident that the true population difference in the proportion of leg ulcers healed, at 12 weeks, between the clinic and control groups lies between -0.063 and 0.128 and the best estimate is 0.033. These may also be expressed as percentages, such that the 95% confidence interval is given by -0.63% and 12.8%.

This approach is only valid when the sample is large enough for the Normal approximation to the binomial to apply; as a rule of thumb both np and n(1-p) should exceed 5 where n=total number of individuals in both groups, p=proportion of individuals with the outcome of interest (irrespective of group) and (1-p) is the proportion of individuals without the outcome of interest (irrespective of group) (see table 3). In addition, thinking of it as a difference in two proportions only makes sense for 2x2 tables i.e. where there are only two groups and two outcomes.

Table 3: Assumptions for the Normal approximation to the Binomial to be valid:

Both np and n(1-p) must be > 5

Where n = total number of individuals in both samples
p = proportion of individuals with condition (irrespective of group)
(1-p) = proportion of individuals without condition (irrespective of group)

Chi-squared test

An alternative approach to analysing the data contained in table 1 and by far the most common, is to apply the Chi-squared test (χ^2 test). This is a more general test and may be used when there are two independent unordered categorical variables that form an r x c contingency table (n.b the current example is a 2 x 2 table, where both r and c = 2). It is valid when at least 80% of expected cell frequencies are greater than 5 and all expected cell frequencies ≥1.

Table 4: Assumptions for the Chi-squared test to be valid:

- The two variables are independent
- At leat 80% of expected cell frequencies are > 5
- All expected cell frequencies are ≥1

The null hypothesis for the chi-squared test is that there is no relationship between the row variable and the column variable, i.e. that being in a particular column does not influence whether you will be in a particular row, and visa versa. For the example above the null hypothesis is that both treatments have the same effect i.e. that being in a particular treatment group (column) is unrelated to whether the leg ulcer healed (row).

If the null hypothesis were true we would expect both treatments to have the same effect and the same the proportion of leg ulcers healed in each group. We estimate this proportion based on the overall proportion healed in both groups. Thus the best estimate of the common “ulcer healing” rate at 12 weeks is given as the common proportion 39 /233 = 16.7% and we use this to calculate the expected number healed for each group. Under the null hypothesis, if there was no relationship between the study group and outcome, we would expect 16.7% of leg ulcers to be healed in each group. For example there are 120 patients in the clinic group and we would expect 20.1 of them to have a healed ulcer by 12 weeks (16.7% of 120). We can calculate the expected number (frequency) in each of the four cells as the (column total*row total)/ overall total. The steps to calculate the Chi-squared statistic are outlined in the table below

Steps to calculate the Chi-squared test statistic

1. Calculate the expected value for each of the four cells in the table
2. Calculate the difference between the observed value and the expected value for each cell
3. Square each difference and divide the resultant quantity by the expected value
4. Add all of these values to get a single number, the χ^2 statistic
5. Calculate the degrees of freedom (df): (number of rows – 1) x (number of columns – 1)
6. Compare this number with tabulated values of the χ^2 distribution with the same degrees of freedom as calculated above

For the leg ulcer example the χ^2 statistic is 0.445 and there is 1 degree of freedom. Comparing this to tabulated values of the χ^2 distribution with 1df gives a P-value of 0.502. As this is greater than the nominal significance level of 0.05 the result is said to be not statistically significant. From this there is insufficient evidence to reject the null and we can conclude that there is no reliable evidence of a difference in leg ulcer healing rates between the two groups.

Chi-squared with continuity correction

In 2 x 2 tables, even when the expected cell counts are bigger than 5, the mathematical approximations for the test statistics are sub-optimal and the null hypothesis is rejected too often on average. In order to overcome this, a modification has been suggested to the formula for the chi-squared test, known as Yates' continuity correction and it is recommended for all 2 x 2 tables (Altman 1991). This continuity correction involves adding on 0.5 to each individual cell's contribution to the overall chi-squared such that the formula for the overall chi-squared may be written

$$\chi^2_{cc} = \sum \frac{(|O-E|-0.5)^2}{E}$$

and this can again be compared with tables for the chi-squared distribution on 1 df.

For the leg ulcer data:

	O	E	|O-E| - 0.5	(|O-E| - 0.5)^2	|O-E| - 0.5^2 / E
Healed / clinic	22	20.1	1.4	1.96	0.98
Not healed / clinic	98	99.9	1.4	1.96	0.020
Healed / control	17	18.9	1.4	1.96	0.104
Not healed / control	96	94.1	1.4	1.96	0.021
Total	233	233			0.243

Thus the χ^2_{cc} = 0.243 and under the null hypothesis of no association between the rows and columns the probability of observing this value of the test statistic or more extreme is about 0.62. Note that this P-value is greater than that obtained without the continuity correction, as this test corrects for the fact that on average the null hypothesis will be rejected to often. This test is said to be more conservative than the original one. The advantage of the continuity correction is that it is easy to implement. It was more commonly used in the past than today where computer intensive methods such as Fisher's Exact test can be readily applied.

Fisher's exact test

In a 2 x 2 table when the expected cell counts are smaller than 5 or any are less than 1 even Yates' correction does not work. In this case Fisher's Exact test, proposed by RA Fisher, can be applied. The test is based upon calculating the probability of rejecting the null hypothesis directly, using all possible tables that could have been observed. This will not be dealt with in more detail here as it will from the basis of a later note. It is mentioned here merely as an alternative test when the assumption underlying the chi-squared test are found not to be valid.

Summary

This tutorial has dealt with some simple methods for analysing binary data. It has outlined how such data can be tabulated using contingency tables and how these can be analysed. Provided the number of observations is large enough, the proportions in the two groups can be compared directly. An alternative is to use a more general test, called the chi-squared test, and this can be extended to more than two groups, or more than two possible outcomes. However, when there are only two groups and two outcomes it is recommended that Yates' continuity correction is used as the standard, or where the assumptions are not valid, Fisher's Exact test.

The analysis of categorical data: Fisher’s Exact test

Jenny V Freeman, Michael J Campbell

Introduction

In the previous tutorial we have outlined some simple methods for analysing binary data, including the comparison of two proportions using the Normal approximation to the binomial and the Chi-squared test((Freeman and Julious 2007)). However, these methods are only approximations, although the approximations are good when the sample size is large. When the sample size is small we can evaluate all possible combinations of the data and compute what are known as exact P-values.

Fisher’s Exact test

When one of the expected values (note: not the observed values) in a 2x2 table is less than 5, and especially when it is less than 1, then Yates’ correction can be improved upon. In this case Fisher’s Exact test, proposed in the mid-1930s almost simultaneously by Fisher, Irwin and Yates(Armitage et al. 2002), can be applied. The null hypothesis for the test is that there is no association between the rows and columns of the 2x2 table, such that the probability of a subject being in a particular row is not influenced by being in a particular column. If the columns represented the study group and the rows represented the outcome, then the null hypothesis could be interpreted as the probability of having a particular outcome is not influenced by the study group, and the test evaluates whether the two study groups differ in the proportions with each outcome.

An important assumption for all of the methods outlined, including Fisher’s exact test, is that the binary data are independent. If the proportions are correlated, then more advanced techniques should be applied. For example in the leg ulcer example of the previous tutorial(Freeman & Julious 2007), if there were more than one leg ulcer per patient, we could not treat the outcomes as independent.

The test is based upon calculating directly the probability of obtaining the results that we have obtained (or results more extreme) if the null hypothesis is actually true, using all possible 2x2 tables that could have been observed, for the same row and column totals as the observed data. These row and column totals are also known as marginal totals. What we are trying to establish is how extreme our particular table (combination of cell frequencies) is in relation all the possible ones that could have occurred given the marginal totals.

This is best explained by a simple worked example. The data below come from an RCT comparing intra-muscular magnesium injections with placebo for the treatment of chronic fatigue syndrome(Cox et al. 1991). Of the 15 patients who had the intra-muscular magnesium injections 12 felt better (80%), whereas, of the 17 on placebo, only 3 felt better (18%).

Table 1: Results of the study to examine whether intramuscular magnesium is better than placebo for the treatment of chronic fatigue syndrome[3]

	Magnesium	Placebo	Total
Felt better	12	3	15
Did not feel better	3	14	17
Total	15	17	32

There are 16 different ways of rearranging the cell frequencies for the above table, whilst keeping the marginal totals the same, as illustrated below in figure 1. The result that corresponds to our observed cell frequencies is (xiii):

Figure 1: Illustration of all the different ways of rearranging cell frequencies in table 1, but with the marginal totals remaining the same

(i)

0	15
15	2

(ii)

1	14
14	3

(iii)

2	13
13	4

(iv)

3	12
12	5

(v)

4	11
11	6

(vi)

5	10
10	7

(vii)

6	9
9	8

(viii)

7	8
8	9

(ix)

8	7
7	10

(x)

9	6
6	11

(xi)

10	5
5	12

(xii)

11	4
4	13

(xiii)

12	3
3	14

(xiv)

13	2
2	15

(xv)

14	1
1	16

(xvi)

15	0
0	17

The general form of table 1 is given in table 2 and under the null hypothesis of no association Fisher showed that the probability of obtaining the frequencies, a, b, c and d in table 2 is

$$\frac{(a+b)!(c+d)!(a+c)!(b+d)!}{(a+b+c+d)!a!b!c!d!} \qquad (1)$$

where x! is the product of all the integers between 1 and x, e.g. 5!=1x2x3x4x5=120 (note that for the purpose of this calculation, we define 0! as 1). Thus for each of the results (i) to (xvi) the exact probability of obtaining that result can be calculated (table 3). For example, the probability of obtaining table (i) in figure 1 is $\frac{15!17!15!17!}{32!0!15!15!2!}$=0.0000002

[3] When organising data such as this is it good practice to arrange the table with the grouping variable forming the columns and the outcome variable forming the rows.

Table2: General form of table 1

	Column 1	Column 2	Total
Row 1	a	b	a+b
Row 2	c	d	c+d
Total	a+c	b+d	a+b+c+d

Table 3: Probabilities associated with each of the frequency tables above, calculated using formula 1

Table	a	b	c	d	P-value
i	0	15	15	2	0.0000002
ii	1	14	14	3	0.0000180
iii	2	13	13	4	0.0004417
iv	3	12	12	5	0.0049769
v	4	11	11	6	0.0298613
vi	5	10	10	7	0.1032349
vii	6	9	9	8	0.2150728
viii	7	8	8	9	0.2765221
ix	8	7	7	10	0.2212177
x	9	6	6	11	0.1094916
xi	10	5	5	12	0.0328475
xii	11	4	4	13	0.0057426
xiii	12	3	3	14	0.0005469
xiv	13	2	2	15	0.0000252
xv	14	1	1	16	0.0000005
xvi	15	0	0	17	0.0000000

From table 3 we can see that the probability of obtaining the observed frequencies for our data is that which corresponds with (xiii), which gives p=0.0005469 and the probability of obtaining our results or results more extreme (a difference that is at least as large) is the sum of the probabilities for (xiii) to (xvi) = 0.000573. This gives the one-sided P-value or obtaining our results or results more extreme, and in order to obtain the two-sided p-value there are several approaches. The first is to simply double this value, which gives p=0.0001146. A second approach is to add together all the probabilities that are the same size or smaller than the one for our particular result, in this case, all probabilities that are less than or equal to 0.0005469, which are tables (i), (ii), (iii), (xiii), (xiv), (xv) and (xvi). This gives a two-sided value of p=0.001033. Generally the difference is not great, though the first approach will always give a value greater than the second. A third approach, which is recommended by Swinscow and Campbell (Swinscow & Campbell 2002) is a compromise and is known as the mid-p method. All the values more extreme than the observed p-value are added up and these are added to one half of the observed value. This gives p=0.000759.

The criticism of the first two methods is that they are too conservative, i.e. is the null hypothesis was true, over repeated studies they would reject the null hypothesis less often than 5%. They are conditional on both sets of marginal totals being fixed, i.e. exactly 15 people being treated with magnesium and 15 feeling better. However, if the study were repeated, even with 15 and 17 in the magnesium and placebo groups respectively, we would not necessarily expect exactly 15 to feel better. The mid-p value method is less conservative, and gives approximately the correct rate of type I errors (false positives).

In either case, for our example, the P-value is less than 0.05, the nominal level for statistical significance and we can conclude that there is evidence of a statistically significant difference in the proportions feeling better between the two treatment groups. However, in common with other non-parametric tests, Fisher's exact test is simply a hypothesis test. It will merely tell you whether a difference is likely, given the null hypothesis (of no difference). It gives you no information about the likely size of the difference, and so whilst we can conclude that there is a significant difference between the two treatments with respect to feeling better or not, we can draw no conclusions about the possible size of the difference.

Example data from last week

Table 1 shows the data from the previous tutorial. It is from a randomised controlled trial of community leg ulcer clinics(Morrell, Walters, Dixon, Collins, Brereton, Peters, & Brooker 1998), comparing the cost-effectiveness of community leg ulcer clinics with standard nursing care. The columns represent the two treatment groups, specialist leg ulcer clinic (clinic) and standard care (home), and the rows represent the outcome variable, in this case whether the leg ulcer has healed or not.

Table 1: 2 x 2 contingency table of Treatment (clinic/home) by Outcome (ulcer healed / not healed) for the Leg ulcer study

	Treatment		
	Clinic	Home	Total
Outcome:			
Healed	22 (18%)	17 (15%)	39
Not healed	98 (82%)	96 (85%)	194
Total	120 (100%)	113 (100%)	233

For this example the two-sided p-value from Fisher's exact test is 0.599 two-sided and in this case we would not reject the null and would conclude that there is insufficient evidence to

Summary

This tutorial has described in detail Fisher's exact test, for analysing simple 2x2 contingency tables when the assumptions for the chi-squared test are not met. It is tedious to do by hand, but nowadays is easily computed by most statistical packages.

Use of Statistical Tables

Lucy Radford, Jenny V Freeman, Stephen J Walters

Introduction

Previous tutorials have looked at hypothesis testing(Freeman & Julious 2006b)and basic statistical tests(Freeman & Campbell 2006;Freeman & Julious 2006a;Freeman & Julious 2007). As part of the process of statistical hypothesis testing, a test statistic is calculated and compared to a hypothesized critical value and this is used to obtain a P-value. This P-value is then used to decide whether the study results are statistically significant or not. This tutorial will explain how statistical tables are used to link test statistics to P-values. It introduces tables for three important statistical distributions: the standard Normal, t and chi-squared distributions and explains how to use them with the help of some simple examples.

Standard Normal Distribution

The Normal distribution is widely used in Statistics and has been discussed in detail previously(Freeman & Julious 2005b). As the mean of a Normally distributed variable can take any value (-∞ to ∞) and the standard deviation any positive value (0 to ∞), there are an infinite number of possible Normal distributions. It is therefore not feasible to print tables for each Normal distribution; however it is possible to convert any Normal distribution to the standard Normal distribution, for which tables are available. The standard Normal distribution has a mean of 0 and standard deviation of 1.

Any value X from a Normal distribution with mean μ and standard deviation σ can be transformed to the standard Normal distribution using the following formula:

(1) $$z = \frac{X - \mu}{\sigma}$$

This transformed X-value, often called z or z-score, is also known as the standard Normal deviate, or Normal score. If an average, rather than a single value, is used the standard deviation should be divided by the square root of the sample size, n, as shown in equation (2).

(2) $$z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

For example, the exam results for the first year of a medical degree are known to be approximately Normally distributed with mean 72 and standard deviation 8. To find the probability that a student will score 89 or more we first need to convert this value to a standard Normal deviate. In this instance, as we have a single value we use equation (1):

$$z = \frac{89 - 72}{8} = 2.13$$

If we wished to find the probability that an average of 10 scores is 75 or more we would use equation (2) to convert to the standard Normal distribution:

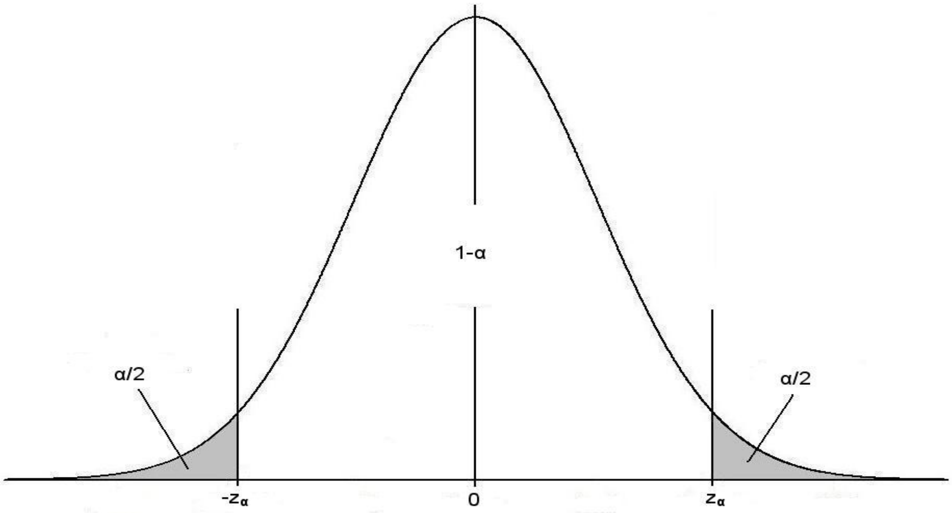
$$z = \frac{75-72}{8/\sqrt{10}} = 1.19$$

We then use the standard Normal table to find the probabilities of observing these z values, or values more extreme given that the population mean and standard deviation are 72 and 8 respectively.

Standard Normal tables can be either one-tailed or two-tailed. In the majority of hypothesis tests the direction of the difference is not specified, leading to a two-sided (or two-tailed) test(Freeman & Julious 2006b). The standard Normal table shown in Table 1 is two-sided[4]. In this two-sided table the value tabulated is the probability, α, that a random variable, Normally distributed with mean zero and standard deviation one, will be either greater than z or less than –z (as shown in the diagram at the top of the table). The total area under the curve represents the total probability space for the standard Normal distribution and sums to 1, and the shaded areas at either end are equal to α/2. A one-tailed probability can be calculated by halving the tabulated probabilities in Table 1. As the Normal distribution is symmetrical it is not necessary for tables to include the probabilities for both positive and negative z values.

[4] A simple trick for seeing whether a particular table is one-tailed or two-tailed is to look at the value that corresponds to a cut-off of 1.96. If the tabulated P-value is 0.05 then the table is for two-tailed p-values.

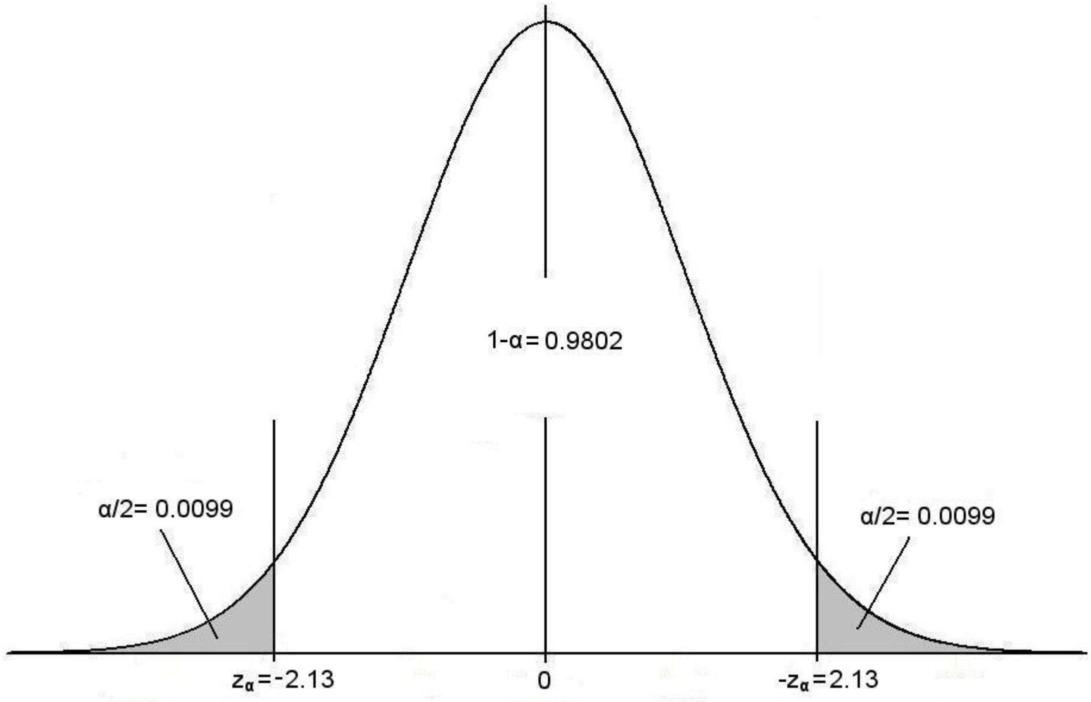
Table 1. Extract from two-tailed standard Normal table. Values tabulated are P-values corresponding to particular cut-offs and are for z values calculated to two decimal places.



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	1.0000	0.9920	0.9840	0.9761	0.9681	0.9601	0.9522	0.9442	0.9362	0.9283
0.10	0.9203	0.9124	0.9045	0.8966	0.8887	0.8808	0.8729	0.8650	0.8572	0.8493
0.20	0.8415	0.8337	0.8259	0.8181	0.8103	0.8206	0.7949	0.7872	0.7795	0.7718
0.30	0.7642	0.7566	0.7490	0.7414	0.7339	0.7263	0.7188	0.7114	0.7039	0.6965
0.40	0.6892	0.6818	0.6745	0.6672	0.6599	0.6527	0.6455	0.6384	0.6312	0.6241
0.50	0.6171	0.6101	0.6031	0.5961	0.5892	0.5823	0.5755	0.5687	0.5619	0.5552
0.60	0.5485	0.5419	0.5353	0.5287	0.5222	0.5157	0.5093	0.5029	0.4965	0.4902
0.70	0.4839	0.4777	0.4715	0.4654	0.4593	0.4533	0.4473	0.4413	0.4354	0.4295
0.80	0.4237	0.4179	0.4122	0.4065	0.4009	0.3953	0.3898	0.3843	0.3789	0.3735
0.90	0.3681	0.3628	0.3576	0.3524	0.3472	0.3421	0.3371	0.3320	0.3271	0.3222
1.00	0.3173	0.3125	0.3077	0.3030	0.2983	0.2837	0.2891	0.2846	0.2801	0.2757
1.10	0.2713	0.2670	0.2627	0.2585	0.2543	0.2501	0.2460	0.2420	0.2380	0.2340
1.20	0.2301	0.2263	0.2225	0.2187	0.2150	0.2113	0.2077	0.2041	0.2005	0.1971
1.30	0.1936	0.1902	0.1868	0.1835	0.1802	0.1770	0.1738	0.1707	0.1676	0.1645
1.40	0.1615	0.1585	0.1556	0.1527	0.1499	0.1471	0.1443	0.1416	0.1389	0.1362
1.50	0.1336	0.1310	0.1285	0.1260	0.1236	0.1211	0.1188	0.1164	0.1141	0.1118
1.60	0.1096	0.1074	0.1052	0.1031	0.1010	0.0989	0.0969	0.0949	0.0930	0.0910
1.70	0.0891	0.0873	0.0854	0.0836	0.0819	0.0801	0.0784	0.0767	0.0751	0.0735
1.80	0.0719	0.0703	0.0688	0.0672	0.0658	0.0643	0.0629	0.0615	0.0601	0.0588
1.90	0.0574	0.0561	0.0549	0.0536	0.0524	0.0512	0.0500	0.0488	0.0477	0.0466
2.00	0.0455	0.0444	0.0434	0.0424	0.0414	0.0404	0.0394	0.0385	0.0375	0.0366
2.10	0.0357	0.0349	0.0340	0.0332	0.0324	0.0316	0.0308	0.0300	0.0293	0.0285
2.20	0.0278	0.0271	0.0264	0.0257	0.0251	0.0244	0.0238	0.0232	0.0226	0.0220
2.30	0.0214	0.0209	0.0203	0.0198	0.0193	0.0188	0.0183	0.0178	0.0173	0.0168
2.40	0.0164	0.0160	0.0155	0.0151	0.0147	0.0143	0.0139	0.0135	0.0131	0.0128
2.50	0.0124	0.0121	0.0117	0.0114	0.0111	0.0108	0.0105	0.0102	0.0099	0.0096
2.60	0.0093	0.0091	0.0088	0.0085	0.0083	0.0080	0.0078	0.0076	0.0074	0.0071
2.70	0.0069	0.0067	0.0065	0.0063	0.0061	0.0060	0.0058	0.0056	0.0054	0.0053
2.80	0.0051	0.0050	0.0048	0.0047	0.0045	0.0044	0.0042	0.0041	0.0040	0.0039
2.90	0.0037	0.0036	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028
3.00	0.0027	0.0026	0.0025	0.0024	0.0024	0.0023	0.0022	0.0021	0.0021	0.0020

From our first example above we want to know what the probability is that a student chosen at random will have a test score of 89, given a population mean of 72 and standard deviation of 8. The z-score calculated above is 2.13. In order to obtain the P-value that corresponds to this z-score we first look at the row in the table that corresponds to a z-score of 2.1. We then need to look down the column that is headed 0.03. The corresponding P-value is 0.0198. However, this is a two-sided probability and corresponds to probability that a z-score is either -2.13 or 2.13 (see figure 1). To get the probability that a student chosen at random will have a test score of at least 89 we need to halve the tabulated P-value. This gives a P-value of 0.0099.

Figure 1: Normal curve showing the Z values and corresponding P-values for the data in example 1.



In a previous tutorial we used the Normal approximation to the binomial to examine whether there were significant differences in the proportion of patients with healed leg ulcers at 12 weeks, between standard treatment and treatment in a specialised leg ulcer clinic(Freeman & Julious 2007). The null hypothesis was that there was no difference in healing rates between the two groups. From this test we obtained a z score of 0.673. Looking this up in Table 1 we can see that it corresponds to a two-sided P-value of 0.503. Thus we cannot reject the null, and we conclude that there is no reliable evidence of a difference in ulcer healing rates at 12 weeks between the two groups.

Student's t-Distribution

The t-test is used for continuous data to compare differences in means between two groups (either paired or unpaired)(Freeman & Julious 2006a). It is based on Student's t-distribution (sometimes referred to as just the t-distribution). This distribution is particularly important when we wish to estimate the mean (or mean difference between groups) of a Normally distributed population but have only a small sample. This is because the t-test, based on the t-distribution, offers more precise estimates for small sample sizes than the tests associated with the Normal distribution. It is closely related to the Normal distribution and as the sample size tends towards infinity the probabilities of the t-distribution approach those of the standard Normal distribution.

The main difference between the t-distribution and the Normal distribution is that the t depends only on one parameter, *v*, the degrees of freedom (d.f.), not on the mean or standard deviation. The degrees of freedom are based on the sample size, n, and are equal to $n - 1$. If the t statistic calculated in the test is greater than the critical value for the chosen level of statistical significance (usually P = 0.05) the null hypothesis for the particular test being carried out is rejected in favour of the alternative. The critical value that is compared to the t statistic is taken from the table of probabilities for the t-distribution, an extract of which is shown in Table 2.

Unlike the table for the Normal distribution described above the tabulated values relate to particular levels of statistical significance, rather than the actual P-values. Each of the columns represents the cut-off points for declaring statistical significance for a given level of (two-sided) significance. For example, the column headed 0.05 in Table 2 gives the values which a calculated t-statistic must be above in order for a result to be statistically significant at the two-sided 5% level. Each row represents the cut-offs for different degrees of freedom. Any test which results in a t statistic less than the tabulated value will not be statistically significant at that level and the P-value will be greater than the value indicated in the column heading. As the t-distribution is symmetrical about the mean, it is not necessary for tables to include the probabilities for both positive and negative t statistics.

Consider for example, a t-test from which a t value of 2.66 on 30 d.f was obtained. Looking at the row corresponding to 30 d.f. in Table 2 this value falls between the tabulated values for 0.02 (=2.457) and 0.01 (=2.75). Thus, the P-value that corresponds with this particular t value will be less than 0.02, but greater than 0.01. In fact the actual (two-tailed) P-value is 0.012.

Table 2. Distribution of t (two-tailed) taken from Swinscow & Campbell(Swinscow & Campbell 2002)

	Probability					
d.f.	**0.5**	**0.1**	**0.05**	**0.02**	**0.01**	**0.00l**
1	l.000	6.314	12.706	3l.821	63.657	636.6l9
2	0.816	2.920	4.303	6.965	9.925	31.598
3	0.765	2.353	3.182	4.541	5.841	12.941
4	0.741	2.132	2.776	3.747	4.604	8.610
5	0.727	2.015	2.571	3.365	4.032	6.859
6	0.718	1.943	2.447	3.l43	3.707	5.959
7	0.711	1.895	2.365	2.998	3.499	5.405
8	0.706	l.860	2.306	2.896	3.355	5.04l
9	0.703	l.833	2.262	2.82l	3.250	4.78l
10	0.700	l.812	2.228	2.764	3.169	4.587
11	0.697	1.796	2.201	2.718	3.l06	4.437
12	0.695	1.782	2.179	2.681	3.055	4.3l8
13	0.694	1.771	2.160	2.650	3.012	4.221
14	0.692	1.76l	2.145	2.624	2.977	4.l40
15	0.69l	l.753	2.13l	2.602	2.947	4.073
16	0.690	1.746	2.120	2.583	2.92l	4.015
17	0.689	1.740	2.110	2.567	2.898	3.965
18	0.688	1.734	2.101	2.552	2.878	3.922
19	0.688	l.729	2.093	2.539	2.861	3.883
20	0.687	1.725	2.086	2.528	2.845	3.850
21	0.686	1.721	2.080	2.518	2.831	3.8l9
22	0.686	1.717	2.074	2.508	2.819	3.792
23	0.685	1.714	2.069	2.500	2.807	3.767
24	0.685	1.711	2.064	2.492	2.797	3.745
25	0.684	1.708	2.060	2.485	2.787	3.725
26	0.684	1.706	2.056	2.479	2.779	3.707
27	0.684	1.703	2.052	2.473	2.771	3.690
28	0.683	1.701	2.048	2.467	2.763	3.674
29	0.683	1.699	2.045	2.462	2.756	3.659
30	0.683	l.697	2.042	2.457	2.750	3.646
40	0.681	l.684	2.021	2.423	2.704	3.551
60	0.679	1.671	2.000	2.390	2.660	3.460
120	0.677	1.658	l.980	2.358	2.617	3.373
∞	0.674	1.645	1.960	2.326	2.576	3.291

Chi-squared Distribution

The final statistical table being considered in this tutorial is that of the chi-squared distribution. There are a wide range of statistical tests that lead to use of the chi-squared distribution, the most common of which is the chi-squared test described in a previous tutorial(Freeman & Julious 2007). Like the t-distribution the chi-squared distribution has only one parameter, the degrees of freedom, *k*. A section of the chi-squared distribution is shown in Table 3. Like the table for the t distribution described above the tabulated values are the chi-squared values that relate to particular levels of statistical significance, rather than actual P-values. Each of the columns represents the cut-off points for declaring statistical significance for a given level of significance. For example, the column headed 0.05 in Table 3 gives the values above which a calculated chi-squared statistic must be in order for a result to be statistically significant at the two-sided 5% level, for degrees of freedom ranging from 1 to 30. Any test which results in a chi-squared statistic less than the tabulated value will not be statistically significant at that level and the P-value will be greater than the value at the top of the column. Consider, for example, a chi-squared value of 4.2 on 1 d.f. Looking at the row corresponding to 1 d.f. in Table 3 this value falls between the tabulated values for 0.05 (=3.841) and 0.02 (=5.412). Thus, the P-value that corresponds with this particular chi-squared statistic will be less than 0.05, but greater than 0.02.

As a second example consider the results of a chi-squared test that was used to assess whether leg ulcer healing rates differed between two different treatment groups (group 1: standard care; treatment 2: specialised leg ulcer clinic)(Freeman & Julious 2007). From this significance test a chi-squared value of 0.243 with 1 d.f. was obtained. Looking at the 1 d.f. row in Table 3 it can be seen that all the values are greater than this value, including the value that corresponds with a P-value of 0.5, 0.455. Thus we can conclude that the P-value corresponding to a chi-squared value of 0.243 is greater than 0.5; in fact the exact value is 0.62.

Table 3. Distribution of χ^2 taken from Swinscow & Campbell(Swinscow & Campbell 2002)

	Probability*					
d.f.	**0.5**	**0.10**	**0.05**	**0.02**	**0.01**	**0.00l**
1	0.455	2.706	3.841	5.412	6.635	10.827
2	1.386	4.605	5.991	7.824	9.210	13.815
3	2.366	6.251	7.815	9.837	11.345	16.268
4	3.357	7.779	9.488	11.668	13.277	18.465
5	4.351	9.236	11.070	13.388	15.086	20.517
6	5.348	10.645	12.592	15.033	16.812	22.457
7	6.346	12.017	14.067	16.622	18.475	24.322
8	7.344	l3.362	15.507	18.168	20.090	26.125
9	8.343	l4.684	16.919	19.679	21.666	27.877
10	9.342	l5.987	18.307	21.161	23.209	29.588
11	10.341	17.275	19.675	22.618	24.725	31.264
12	11.340	18.549	21.026	24.054	26.217	32.909
13	12.340	19.812	22.362	25.472	27.688	34.528
14	13.339	21.064	23.685	26.873	29.141	36.123
15	14.339	22.307	24.996	28.259	30.578	37.697
16	15.338	23.542	26.296	29.633	32.000	39.252
17	16.338	24.769	27.587	30.995	33.409	40.790
18	17.338	25.989	28.869	32.346	34.805	42.312
19	18.338	27.204	30.144	33.687	36.191	43.820
20	19.337	28.412	31.410	35.020	37.566	45.315
21	20.337	29.615	32.671	36.343	38.932	46.797
22	21.337	30.813	33.924	37.659	40.289	48.268
23	22.337	32.007	35.172	38.968	41.638	49.728
24	23.337	33.196	36.415	40.270	42.980	51.745
25	24.337	34.382	37.652	41.566	44.314	52.620
26	25.336	35.563	38.885	42.479	45.642	54.707
27	26.336	36.741	40.113	44.140	45.963	55.476
28	27.336	37.916	41.337	45.419	48.278	56.893
29	28.336	39.087	42.557	46.693	49.588	58.302
30	29.336	40.256	43.773	47.962	50.892	59.703

***: these are two-sided P-values**

Summary

In this tutorial we have shown how to use statistical tables to obtain P-values for the standard Normal, t and chi-squared distributions, and given examples to show how the values from these tables are used to make decisions in a variety of basic statistical tests.

One-way Analysis of Variance

Jenny V Freeman, Michael J Campbell

In a previous tutorial we described the unpaired t-test for comparing two independent groups when the data are Normally distributed(Freeman & Julious 2005a). In this tutorial we will explain how this can be generalised to comparing more than two groups using a method called the one-way analysis of variance (ANOVA). Examples of such comparisons include:

1. Pain score between groups given different analgesics in a clinical trial
2. Birthweights between different methods of delivery for women in a particular hospital
3. Heights of children with different ethnic backgrounds on entry to primary school

Whilst it is possible to compare individual pairs of groups using the t-test this would increase the probability of committing a type 1 error (false positive error) and is not an efficient use of the data. The one-way ANOVA is a single global test of whether the means differ in any of the groups. However, it is worth noting that if there are only two groups then the one-way ANOVA is exactly equivalent to the t-test and will give the same p-value.

There are several assumptions that need to be satisfied for the one-way ANOVA to be valid as outlined in Box 1 and these should be checked before performing the test. The assumption of equality of variances can be tested by examining the standard deviations for each group, as a rule of thumb, no single SD should be greater than twice any of the others. The assumption of normality can be examined by looking at either dot plots of the data in groups or histograms, if the numbers are large. If the assumptions are not met then the data can be transformed, for example by taking logarithms, or by using the non-parametric equivalent of the ANOVA, the Krukall-Wallis test. This test uses exactly the same methodology as the one-way ANOVA, except that the data are ranked (ignoring grouping) and the test is performed on the ranks.

Box 1: Assumptions underlying one-way ANOVA

1. The data are independent
2. The data are Normally distributed in each group
3. The variance is the same in each group

Description of technique

In brief, the one-way ANOVA is based upon the idea that you can partition the variability in a set of data into different sources, for example into random variability between individuals *within groups* (sometimes called the residual or unexplained variability) and variability due to systematic difference *between groups*. Under the null hypothesis that the means are the same, the within and between variances are expected to be the same. However, if there are systematic differences between groups then it would be expected that the between groups variance would be greater than that within groups and a test can be constructed that is based upon the ratio of these two variances. This ratio is known as the F statistic and critical values for a significance test can be obtained from tables of the F-distribution, but in order to do this you need to know the degrees of freedom (df), of which there are two types. There are those due to the variability between groups (df= number of groups -1) and those due to the variability within groups (df= total number of observations – number of groups).

In order the calculate the F-statistics, for each group you need to count the number of observations (n), the mean for the variable of interest, $\bar{y}$, the sum of the observations(T), and the sum of the observations squared(S). Then sum each of these quantities across the groups. Assume, for example, that you have k groups:

Group	1	2 …	k	All groups combined
Number of observations	n_1	n_2 …	n_k	$N = \sum_{i=1}^{k} n_i$
Sum of observations	T_1	T_2 …	T_k	$T = \sum_{i=1}^{k} T_i$
Mean of observations	$\bar{y}_1$	$\bar{y}_2$ …	$\bar{y}_3$	$\bar{y} = T/N$
Sum of squared observations	S_1	S_2 …	S_k	$S = \sum_{i=1}^{k} S_i$

Once the above quantities have been calculated you can then construct an analysis of variance table to obtain the F statistic (Box 2):

Box 2: Analysis of Variance Table:

	Sum of squares (SS)	Degrees of freedom (df)	Mean Square (MS)	F statistic (variance ratio)
Between groups	$\sum_{i=1}^{k} T_i^{\,2}/n_i - T^2/N$	k-1	SS/df	$MS_{between}/MS_{within}$
Within groups	$S - \sum_{i=1}^{k} T_i^2/n_i$	N-k	SS/df	
Total	$S - T^2/N$	N-1		

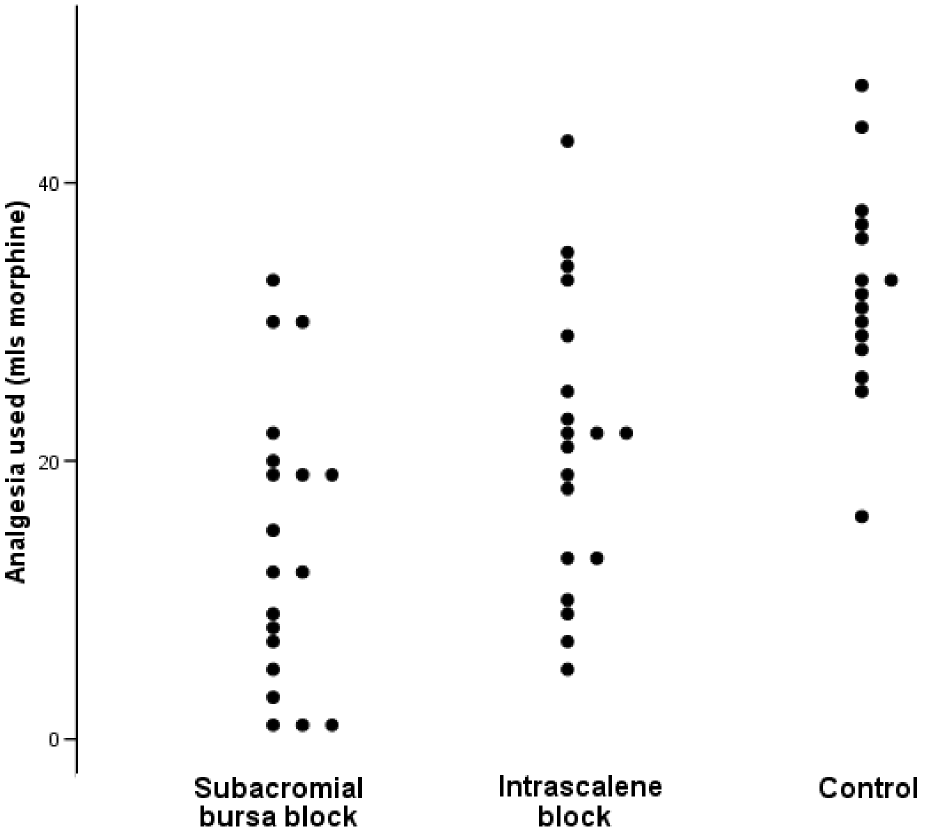
This is then compared to tabulated critical values of an F statistic on k-1 and N-k degrees of freedom to obtain a P-value.

Further details of the mathematics of the technique can be found in Chapter 7 of Statistical Methods for Medical Research by Armitage et al(Armitage, Berry, & Matthews 2002).

Example

As part of a study looking at the post-operative pain experience of patients following shoulder surgery, two different methods of providing surgical analgesia, subacromial bursa block (SBB) and intrascalene block (ISB) were compared with a control procedure(Nisar et al. 2007). There were 19 patients in each of the analgesia groups and 9 in the control group and the outcome of interest was the amount of morphine used by the patient in the first 24 hours after surgery. The data are displayed Figure 1. Examining this figure, it appears that there are systematic differences between the study groups

Figure 1: Amount of patient controlled analgesia (mls morphine) used in first 24 hours after surgery by study group.



The null hypothesis is that there are no differences between the means of the three groups, and this can be tested against the alternative that there are differences between the means. Using the method outlined above the data can be summarised as in table 1:

Table 1: Construction of the components for the ANOVA table for the analgesia trial data

	SBB	ISB	Control	Total
n_i	19	19	15	53
T_i	266	403	485	1154
$\bar{x} = T_i / n_i$	14	21	32	22
T_i^2 / n_i	3724	8548	15682	27954
S_i	5600	10485	16499	32584

The formulae in the ANOVA table above can then be used to obtain the between and within groups sum of squares (Table 1).

Between groups sum of squares = 27954 – (1154*1154)/53 = 2827

Within groups sum of squares = 32584 – 27954 = 4630

These can be used to construct an ANOVA table (Table 2) and calculate the F statistic. This F statistic can then be used to obtain a P-value. For the current example, for an F statistic with 2 and 50 degrees of freedom the P-value is < 0.001. This is highly significant and indicates that there is sufficient evidence to reject the null hypothesis, that the group means are the same and accept the alternative. We would conclude that there are systematic differences between the groups.

Table 2: ANOVA table for analgesia trial data

	Sum of squares	Degrees of freedom	Mean square	F statistic (variance ratio)
Between groups	2827	2	1413.5	15.26
Within groups	4630	50	92.6	
Total	7457	52		

Comparing different groups

Having decided that there are differences between the groups, it might also be of interest to test for contrasts between the groups i.e. compare the groups with each other. One of the advantages of the one-way ANOVA over the t-test is that the standard error for the difference between pairs of groups is based upon the within group mean square. As this has more degrees of freedom than a standard error based upon the two groups alone, the test has greater power to find a difference if one exists. Many different methods for making post hoc comparisons between groups have been proposed, all of which are designed to ensure that the overall type 1 error rate stays below 5%. This will be discussed further in a future tutorial, but briefly, if the numbers in each group are the same then the Student-Newman Keuls test is recommended, otherwise use Scheffé(Armitage, Berry, & Matthews 2002). If there is a single control group that you want to test against all other groups (but not the other groups against each other) then Dunnett's test can be used. These are all readily obtained from the major statistical packages.

Presentation of the results of an ANOVA

When presenting the results of an ANOVA it is good practice to report the group means and the numbers in each group, together with the F statistic, its df and the P-value associated with this statistic. If any post hoc analyses are carried out, the procedure used for the pairwise comparisons (e.g. Scheffé) should be stated and the mean differences between groups together with the associated 95% confidence intervals for these differences should be presented(Lang and Secic 1997). Thus for the analgesia trial, where it was of interest to compare the difference between the two block groups, the results are presented in table 3.

Table 3: Presentation of results of ANOVA

	Mean (95% CI)	P-value
Study group		
ISB (n=19)	14.0 (9.1 to 18.9)	< 0.001*
SBB (n=19)	21.2(16.2 to 26.2)	
Control (n=15)	32.3(28.1 to 36.6)	
Difference (SBB – ISB)	7.2 (-0.7 to 15.1)	0.079**

* ANOVA, $F_{2,50} = 15.26$

** Scheffé method used for post hoc comparison

Sample Size Calculations for Clinical Trials: parallel, two-group clinical trial with a Normally distributed outcome variable.

S. A. Julious, J. V Freeman

Introduction

When planning any study, it is important to have a good idea of how many individuals are needed and this is one of the questions that statisticians get asked most frequently. The justification for the chosen sample size can range from a formal calculation based upon a clinical outcome, to what is feasible. Even for the latter, a sample size justification can be provided as it is possible to determine, for a fixed sample size, what difference could be detected for a given level of power.

This tutorial describes the simplest case, when calculating the sample size needed for a two group comparison with a Normally distributed outcome variable, designed to assess superiority (i.e. whether one group is better than the other), with equal numbers allocated between groups (i.e. an allocation ratio of 1:1).

General principles

In order to calculate the size of the sample required several quantities need to be known or estimated *a priori*, including the number of groups to be compared, the outcome variable (and its distribution), the anticipated size of the difference between groups and an estimate of the population variability. In addition the researcher must decide in advance the maximum acceptable values for the Type I and Type II error rates (see tutorial 4 (Freeman & Julious 2006b)).

Null and Alternative Hypotheses for a Superiority Trial

A superiority trial is designed to determine whether there is evidence of a statistically significant difference between treatment groups for the outcome of interest, compared to the null hypothesis that the treatments are the same. The purpose of the sample size calculation for this type of study is to provide sufficient evidence to reject the null hypothesis when in fact some alternative hypothesis is true.

The null (H_0) and alternative (H_1) hypotheses take the form:

H_0: The two treatments are not different with respect to the mean response ($\mu_A = \mu_B$).

H_1: The two treatments are different with respect to the mean response ($\mu_A \neq \mu_B$).

In the definition of the null and alternative hypotheses μ_A and μ_B refer to the mean response for treatments A and B respectively.

Estimate of effect size

In an ideal world the anticipated size of the treatment difference is estimated from pilot data and this is one of the main reasons for conducting a pilot study. Where no pilot data exist it may be possible to estimate the effect size from the published literature. In terms of superiority an estimate of effect size could be based upon an assessment of what is a clinically meaningful difference. However, it is important that if this is the case, the difference should be realistic. It is no good setting an effect size for the sample size calculation that is unrealistically large, as, if the observed difference is smaller than this the study may well fail to reject the null hypothesis to an appropriate level of statistical significance.

Estimate of population variance

Once the estimated effect size has been determined we need to have an estimate of the of the population variability about this effect size, σ^2. The smaller the variability the fewer individuals are required for (all other things being equal)

.

Type I and type II error rates

In any investigation, we are attempting to find out what is true, although in reality the best we can hope for is to decide what is most likely. However, it is possible that in making a decision we make an error:

- We could reject the null hypothesis (in the case of a superiority trial: that there is no difference between treatments) even though it is true. This is known as a Type I error or a false positive.
- We could fail to reject the null hypothesis even though it is false - there genuinely is a difference between treatments. This is known as a Type II error or false negative error(Freeman & Julious 2006b).

In order to limit the possibility of committing either error, we set the maximum acceptable level for each *a priori*. The probability of committing a Type I or false positive error is given by the greek letter α and we can reduce the risk of committing this type of error by decreasing the value of α of at which a result is declared 'statistically significant'. The significance level is usually set at the "magic" 5% or 0.05. If greater proof is required this could be reduced. For example a significance level of 0.01 would mean we have a Type I error of 1%.

The Type I error is often referred to as a regulatory or society risk as it is upon these that the cost of this error is incurred if a new regimen, in the case of drug trials, were to enter the market needlessly. Even outside of drug trials it is a societal risk as a Type I error may initiate further research in a research area which is actually a dead end.

For a superiority trial there are two chances of rejecting the null hypothesis and thus making a Type I error. The null hypothesis can be rejected if $\mu_\mathrm{A} > \mu_\mathrm{B}$ or if $\mu_\mathrm{A} < \mu_\mathrm{B}$ by a statistically significant amount. As there are two chances of rejecting the null hypothesis the statistical test is referred to as a two tailed test and each tail is allocated an equal amount of the Type I error (=*α*/2). The sum of these tails adds up to the overall Type I error rate (*α*). Thus, for a trial with an overall type I error rate of 5%, the null hypothesis can be rejected if either the test of $\mu_\mathrm{A} > \mu_\mathrm{B}$ is statistically significant at the 2.5% level or the test of $\mu_A < \mu_B$ is statistically significant at the 2.5% level.

The probability of making a Type II or false negative error is given by the greek letter *β*. In general the value of *β* is set between 0.1 to 0.2. The Type II error is often referred to as the sponsor's risk as it is the sponsor that will incur the cost of this error. The acceptable value of *β* is usually greater than that for α as the cost to society from this error is usually lower.

Often instead of referring to the Type II error, reference is made to the power of a study. The power is the probability that we will detect a difference of a specified size, if there is one. That is, it is the probability of *not* making a Type II error i.e. 1- *β*. The power therefore is the probability of rejecting the null hypothesis when it is false. Although we have said above that the Type II error is usually set between 0.1 and 0.2, we would recommend that 0.1 as used as the standard level. Often people talk in terms of power and reducing the power from 0.90 to 0.80 does not seem such a great step to make but in effect we are doubling the type II error for little actual reduction (around 25%) in the sample size.

Table 1: Effect on sample size of changing the size of the constituents used in the sample size calculation

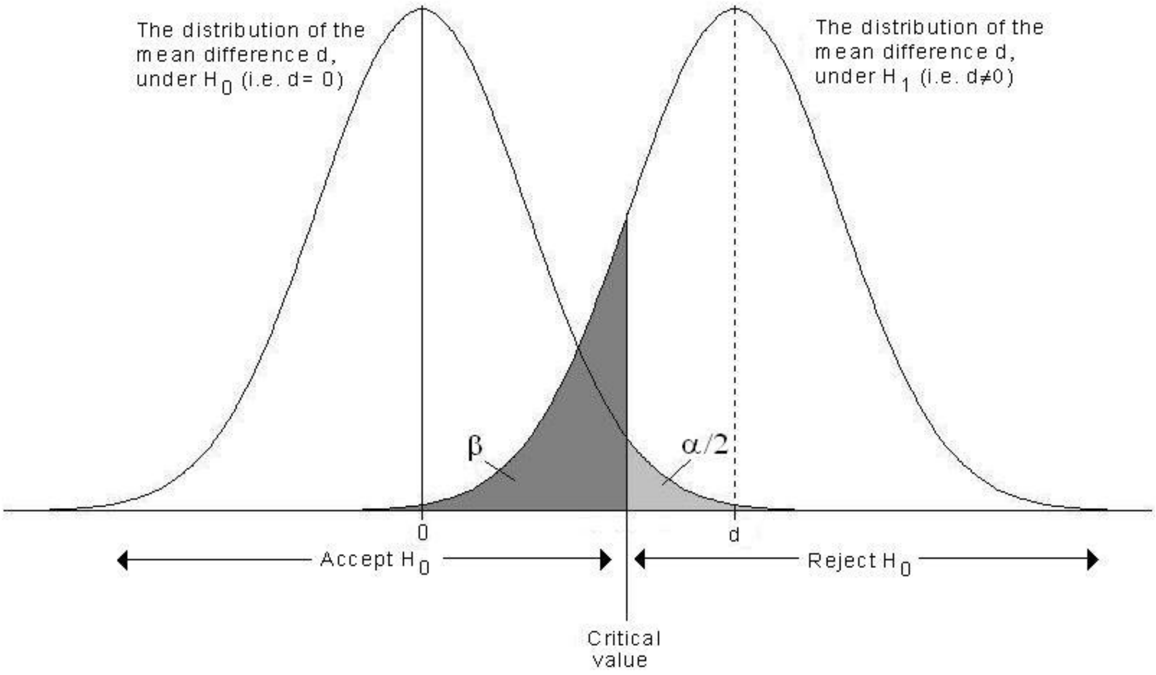
Increase in:	Effect on sample size:
Type I error	Decrease
Type II error	Decrease
Effect size	Decrease
Variance of effect size	Increase

Derivation of the formula for sample size calculation

Figure 1 shows the distribution of the mean difference d, if the null hypothesis is true (i.e. d=0; the left-hand curve) and if the alternative hypothesis is true (right-hand curve). We set a critical value (d_{crit}) for the test such that for any value less than the critical value we would accept the null hypothesis, whilst for any value greater than the critical value we would reject the null hypothesis. For a given critical value:

- If the alternative hypothesis is true, the dark shaded area shows the probability of making a type II error, *β* (i.e. falsely concluding there is no difference even though there is a difference).
- If the null hypothesis is true the lighter shaded area shows the probability of making a type I error α (i.e. falsely concluding that there is a difference even though there is no difference).

Figure 1: Distribution of the mean difference under the null and alternative hypotheses


The distribution of the mean difference d, under H0 (i.e. d= 0)
The distribution of the mean difference d, under H1 (i.e. d≠0)
β
α/2
0
d
Accept H0
Reject H0
Critical value

If we assume that both μ_A and μ_B come from populations with the same variance σ^2, and there are equal numbers in each group (=m) then the standard error of the difference d is equal to $\sigma\sqrt{2/m}$. If the null hypothesis is true we usually allow for the possibility of a Type I error in the either direction and so under the null hypothesis that the mean difference is 0 the critical value is determined by:

(1) $\mathrm{d}_{\mathrm{crit}} = Z_{1-\alpha/2}\sigma\sqrt{\frac{2}{m}}$

On the other hand, if the alternative hypothesis is true, the mean difference is Normally distributed with a mean of d and a standard error of σ√(2/m) and the critical value is determined by:

(2) $\mathrm{d}_{\mathrm{crit}} = d - Z_{1-\beta}\sigma\sqrt{\frac{2}{m}}$

Thus

(3) $Z_{1-\alpha/2}\sigma\sqrt{\frac{2}{m}} = d - Z_{1-\beta}\sigma\sqrt{\frac{2}{m}}$,

and the sample size per treatment group can be estimated by re-arranging (3)

(4) $$m = \frac{2(Z_{1-\alpha/2} + Z_{1-\beta})^2 \sigma^2}{d^2}$$

where σ^2 is the population variance, d is the effect size of interest and *m* is the sample size per group. $Z_{1-\alpha/2}$ and $Z_{1-\beta}$ are Z values obtained from the standard Normal distribution, for the pre-determined values of *α* and *β*. For example, when *α*=0.05, then 1- *α/*2=0.975 and the corresponding Z value is 1.96 and when *β*=0.10 then 1-*β*=0.9 and the corresponding Z value is 1.282.

The result (4) has a number of advantages not least that the calculations are relatively simple and easy to do by hand. However, it is worth noting that it is an approximation and many sample size packages use a slightly different formula based not on the Normal distribution but on the t distribution. When the sample size is small as may often be the case with imaging studies you could add a correction factor of $Z^2_{1-\alpha/2}/4$ to (4) to allow for the approximation (2).

(5) $$m = \frac{2\left(Z_{1-\beta} + Z_{1-\alpha/2}\right)^2 \sigma^2}{d^2} + \frac{Z^2_{1-\alpha/2}}{4}.$$

Table 1 gives sample sizes per group for various standardised differences ($\delta = d/\sigma$) using (5).

Table 4. Sample sizes for per group in a parallel group study for different standardised differences for 90% power and a two sided type I error of 5%

δ	Sample Size
0.05	8407
0.10	2103
0.15	935
0.20	527
0.25	338
0.30	235
0.35	173
0.40	133
0.45	105
0.50	86
0.55	71
0.60	60
0.65	51
0.70	44
0.75	39
0.80	34
0.85	31
0.90	27
0.95	25
1.00	22

For quick calculations the following formula to calculate a sample size with 90% power and a two-sided 5% type I error rate, can be used

(6) $$m = \frac{21\sigma^2}{d^2}.$$

While for 80% power and a two-sided 5% type I error rate we can use

(7) $$m = \frac{16\sigma^2}{d^2},$$

Worked example 1:

Researchers wished to assess which of two methods of managing post-operative pain following shoulder surgery was the most effective in reducing the post-operative experience for patients. The main outcome measure was the amount of patient-administered morphine used in the first 24 hours post-op. The effect size of interest was 7mg morphine and the population standard deviation was estimated as 14mg. The researchers wanted to have 90% power to detect a difference with a two sided significance of 0.05. The calculations are as follows:

$$\beta = 0.1,\ \mathrm{Z}_{1-\beta} = Z_{1-0.1} = Z_{0.9} = 1.282$$

$$\alpha = 0.05\ \mathrm{Z}_{1-\alpha/2} = Z_{1-0.025} = Z_{0.975} = 1.96$$

$$\sigma^2 = 14^2 = 196$$

$$d^2 = 7^2 = 49$$

$$m = \frac{2\left(.282 + 1.96\right)^2 196}{49} + \frac{1.96^2}{4} = 84.08 + 0.96 = 85.04 \approx 86$$

Thus, the number of patients needed per group was 86 and the total number needed in both groups was 2x86 = 172 patients.

Alternatively we can calculate $\delta = d/\sigma$ =7/14=0.5 and from Table 1 we also obtain a sample size estimate of 172.

Accounting for drop-outs:

Described above is a method for calculating the required sample size of evaluable subjects (the minimum number required for the statistical analysis). However, often this is only the first step in estimating the size of a study. Often we can anticipated drop-outs in a study and this needs to be accounted for in the final sample size of patients to be recruited.

To account for drop outs we need to anticipated likely dropout rate. Suppost we anticipate that a proportion p subject to drop out. The recruited sample size (to ensure m evaluable subjects) would then be

(8) $$m_{\mathrm{Re}cruited} = \frac{m}{1-p}$$

Worked example 2:

Revisiting the worked example. suppose the researchers anticipated a drop-out rate of about 20%, the number needed to be recruited in total would need to be 172/(1-0.2)=215.

Sample size based on feasibility

We have just shown how to calculate a sample size for a given effect size. However, it can sometimes be the case that the sample size is fixed by practical considerations or feasibility, even before a study takes place. When this is the case, the trial is not powered to detect any pre-specified effect, but a power analysis can still be used to determine the difference that could be detected for this fixed sample size. In this case, as we now have m, σ^2, β and α, we can re-arrange (1) to calculate the difference (d) that could be detected for this fixed sample size, m, where m is the number per group:

(9) $$d = (Z_{1-\alpha/2} + Z_{1-\beta})\sigma\sqrt{\frac{2}{m}}$$

If sample size is based primarily on feasibility then it should be clearly highlighted as such in any protocol.

Summary

This article describes how to calculate the sample size for a Normally distributed outcome for a two group comparison designed to assess superiority as. A sample size justification should be provided for all clinical studies even where the sample size is based on feasibility

Diagnostic Tests

Michael J Campbell, Jenny V Freeman

In this tutorial we will examine how to evaluate a diagnostic test. Initially we will consider the case when there is a binary measure (two categories: disease present / disease absent). We will then look at how define a suitable cut-off for an ordinal or continuous measurement scale and we will finish with a short discussion contrasting diagnostic tests with screening tests.

When evaluating any diagnostic test one should have a definitive method for deciding whether the disease is present in order to see how well the test performs. For example, to diagnose a cancer one could take a biopsy, to diagnose depression one could ask a psychiatrist to interview a patient, and to diagnose a walking problem one could video a patient and have it viewed by an expert. This is sometimes called the 'gold standard'. Often the gold standard test is expensive and difficult to administer and thus a test is required that is cheaper and easier to use.

Binary situation

Let us consider first the simple binary situation in which both the gold standard and the diagnostic test have either a positive or negative outcome (disease is present or absent). The situation is best summarised by the 2x2 table below (Table 1). In writing this table always put the gold standard on the top, and the results of the test on the side:

Table 1. Standard table for diagnostic tests

		Gold Standard		
		Positive	Negative	
Diagnostic Test	Positive	a	b	a+b
	Negative	c	d	c+d
	Total	a+c	b+d	n

The numbers 'a' and 'd' are the numbers of true positives and true negatives respectively. The number 'b' is the number of false positives, because although the test is positive the patients don't have the disease, and similarly 'c' is the number of false negatives. The *prevalence* of the disease is the proportion of people diagnosed by the gold standard and is given by (a+c)/n, although this is often expressed as a percentage.

In order to assess how good the test is we can calculate the sensitivity and specificity, and the positive and negative predictive values. The sensitivity of the test is the proportion of people with the disease who are correctly identified as having the disease. This is given by a/(a+c) and is usually presented as a percentage. Suppose a test is 100% sensitive. Then the number of false negatives is zero and we would expect Table 2.

Table 2 Results of a diagnostic test with 100% sensitivity

		Gold Standard		
		Positive	Negative	
Diagnostic Test	Positive	a	b	a+b
	Negative	0	d	d
	Total	a	b+d	n

From Table 2 we can see that if a patient has a negative test result we can be certain that the patient *does not* have the disease. Sackett et al(2007) refer to this as SnNout i.e. for a test with a high sensitivity (Sn), a Negative result rules *out* the disease.

The specificity of a test is the proportion of people without the disease who are correctly identified as not having the disease. This is given by d/(b+d) and as with sensitivity is usually presented as a percentage. Now suppose a test is 100% specific. Then the number of false positives is zero and we would expect table 3

Table 3 Results of a diagnostic test with 100% specificity

		Gold Standard		
		Positive	Negative	
Diagnostic Test	Positive	a	0	a
	Negative	c	d	c+d
	Total	a+c	d	n

From Table 3 we can see that if a patient has a positive test we can be certain the patient *has* the disease. Sackett et al (x) refer to this as SpPin., i.e. for a test with a high specificity (Sp), a Positive test rules *in* the disease.

Box 1: Useful Mnemonic

SeNsitivity=1-proportion false Negatives (n in each side)
SPecificity=1-proportion false Positives (p in each side)

What patients really want to know, however, is ‘if I have a positive test, what are the chances I have the disease?’ This is given by the *positive predictive value* (PPV) which is a/(a+b). One way of looking at the test is that before the test the chances of having the disease was (a+c)/n. After the test they are either a/(a+b) or c/(c+d) depending on whether the result was positive or negative.

The *negative predictive value* is the proportion of those whose test result is negative who do not have the disease and is given by d/(c+d).

It should be noted that whilst sensitivity and specificity are independent of prevalence, positive and negative predictive values are not. Sensitivity and specificity are characteristics of the *test* and will be valid for different populations with different prevalences. Thus we could use them in populations with high prevalence such as elderly people as well as for low prevalence such as for young people.

However, the PPV is a characteristic of the *population* and so will vary depending on the prevalence.

To show this suppose that in a different population, the prevalence of the disease is double that of the current population (assume the prevalence is low, so that a and c are much smaller than b and d and thus the results for those without the disease are much the same as the earlier table). The situation is given in Table 4

Table 4 Standard situation but with a doubling of the prevalence

		Gold Standard		
		Positive	Negative	
Diagnostic Test	Positive	2a	b	2a+b
	Negative	2c	d	2c+d
	Total	2(a+c)	b+d	n'

The sensitivity is now 2a/(2a+2c)=a/(a+c) as before. The specificity is unchanged. However the positive predictive value is given by 2a/(2a+b) which is greater than the earlier value of a/(a+b).

Likelihood ratio

It is common to prefer a single summary measure, and for a diagnostic test this is given by the likelihood ratio for a positive test (LR(+) as defined below:

$$LR+ = \frac{\text{Probability of positive test } \textit{given} \text{ the disease}}{\text{Probability of positive test } \textit{without} \text{ disease}} = \frac{\text{Sensitivity}}{\text{1-Specificity}} = \frac{a\ (b+d)}{b\ (a+c)}$$

One reason why this is useful is that it can be used to calculate the odds of having the disease given a positive result. The odds of an event are defined as the ratio of the probability of the event occurring to the probability of the event not occurring i.e. p/(1-p) where p is the probability of the event. Before the test is conducted the probability of having the disease is just the prevalence, and the odds are simply [(a+c)/n]/[b+d)/n]= (a+c)/(b+d). The odds of having the disease after a positive test are given by

Odds of disease after positive test=odds of disease before test x LR(+)= a/b

We can also get the odds of disease after a positive test directly from the PPV since the odds of disease after a positive test is PPV/(1-PPV).

Example:

A recent study by Kroenke et al (Kroenke et al. 2007) surveyed 965 people attending primary care centres in the US. They were interested in whether a family practitioner could diagnose Generalized Anxiety Disorder (GAD) by asking two simple questions (the GAD2 questionnaire): ‘Over the last two weeks, how often have you been bothered by the following problems? 1) feeling nervous, anxious or on edge 2) Not able to stop or control worrying. The patients answered each question from ‘not at all’, ‘several days’, ‘more than half’ and ‘nearly every day’, scoring 0,1,2 or 3 respectively. The scores for the two questions were summed and a score of over 3 was considered positive. Two mental health professionals then held structured psychiatric interviews with the subject over the telephone to diagnose GAD. The professionals were ignorant of the result of the GAD2 questionnaire. The results are given in table 2:

Table 4.2 Results from Kroenke et al(Kroenke, Spitzer, Williams, Monahan, & Löwe 2007)

		Diagnosis by mental health worker		
		Positive	Negative	
GAD2	≥3	63	152	215
	< 3	10	740	750
	Total	73	892	965

The prevalence of the disease is given by (a+c)/n=73/965=0.076=7.6%.

The sensitivity of the test is given by a/(a+c)= 63/73=0.86=86%.

The specificity of the test is given by d/(b+d) = 740/892=0.83=83%

The positive predictive value (PPV) which is a/(a+b) = 63/215==0.29=29%.

The negative predictive value is d/(c+d)= 740/750=0.987=98.7%.

Thus before the test the chances of having GAD were 7.6%. After the test they are either 29% or 1.3% (i.e. 100*(1-0.987) depending on the result. Note that even with a positive test the chances of having GAD are still less than 1/3.

For the GAD example we find that LR(+)= 0.86/(1-0.83)=5.06 and the odds=0.29/(1-0.29)=0.41.

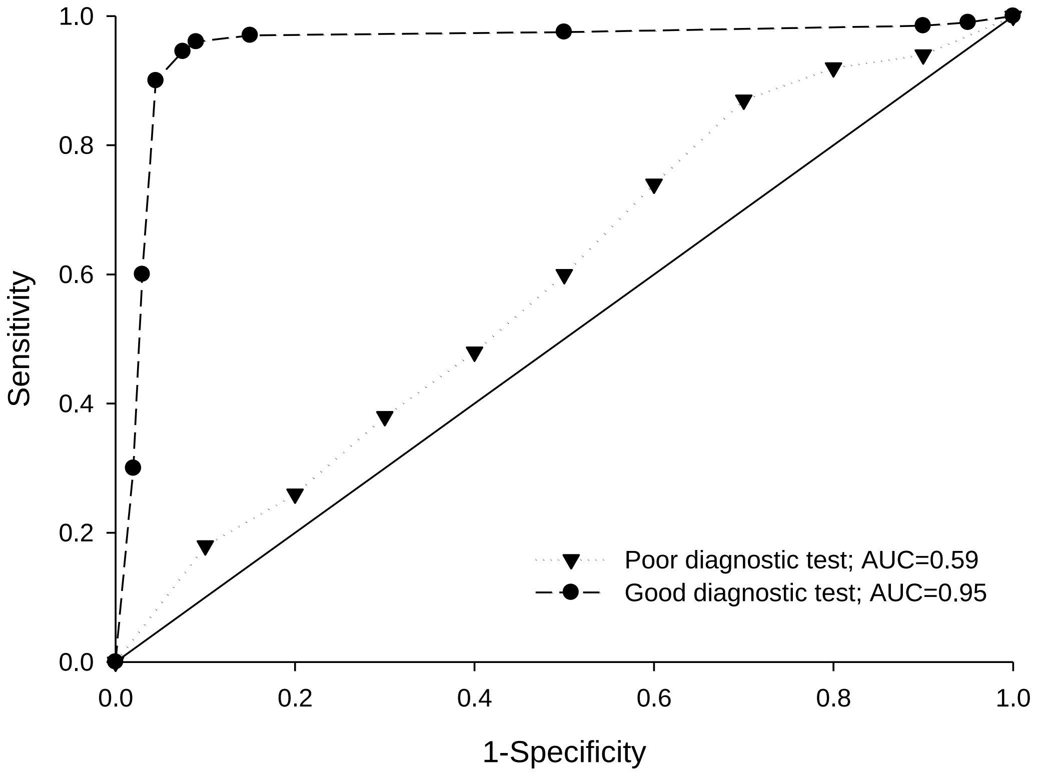
ROC Curves

For a diagnostic test that produces results on a continuous or ordinal measurement scale a convenient cut-off level needs to be selected to calculate the sensitivity and specificity. For example the GAD2 questionnaire has possible values from 0 to 6. Why should one choose the value of 3 as the cut-off? For a cut off of 2 the sensitivity is 0.95, the specificity is 0.64 and the LR(+) is 2.6(Kroenke, Spitzer, Williams, Monahan, & Löwe 2007). One might argue that since a cut-off of 3 has a better LR(+) then one should use it. However, a cut-off of 2 gives a higher sensitivity, which might be important. It should be noted that a sensitivity of 100% is always achievable by stating that everyone has the disease, but this is at the expense of a poor specificity (similarly a 100% specificity can be achieved by stating no-one has the disease. If the prevalence is low, this tactic will have a high accuracy, i.e. it will be right most of the time, but sadly wrong for the important cases). A discussion of the different scenarios for preferring a high specificity or sensitivity is given in the next section.

A simple graphical device for displaying the trade-offs between sensitivity and specificity for tests on a continuous or ordinal scale is a *receiver operating* *characteristics* (ROC) curve (the unusual name originates from electrical engineering). This is a plot of sensitivity versus one minus specificity for different cut-off values of the diagnostic test. ROC curves for two theoretical tests are shown in Fig 1, together with the line of equality which is what we would expect if a test had no power to detect disease. A perfect diagnostic test would be one with no false negatives (i.e. sensitivity or 1) or false positives (i.e. specificity of 1) and would be represented by a line starting at the origin, travelling vertically up the Y-axis to a sensitivity of 1 and then horizontally across to a false positive rate of 1. Any diagnostic test that was reasonable would produce a ROC curve in the upper left-hand triangle of figure 1. The selection of the optimal cut-off will depend upon the relative medical consequences and costs of false positive and false negative errors.

ROC curves are particularly useful for comparing different diagnostic tests and when more than one test is available they can be compared by plotting both on the same plot. A test for which the plot is consistently nearer the left hand side and the top is to be preferred. In addition the area under the curve (AUC) for each plot can be calculated. For the perfect test outlined above the AUC is 1 and represents the total area of the panel (i.e. 1x1). For the two curves displayed it is obvious that the best test is the one with the line represented by the dashed line on the left of the Figure. This has an AUC value of 0.95 compared to the other much poorer fitting line which as an AUC value of 0.59.

Fig 1: Example ROC curves showing also the line of equality



Distinction between diagnosis and screening.

It is important to understand the difference between diagnosing a disease and screening for it. In the former case there are usually some symptoms, and so there may already be a suspicion that something is wrong. If a test is positive some action will be taken. In the latter case there are usually no symptoms and so if the test is negative the person will have no further tests. Recalling Sackett's mnemonics SpPin and SnNout, for diagnosis we want a positive test to rule people in, so we want a high specificity. For screening we want a negative test to rule people out so we want a high sensitivity. Thus mass mammography will have a fairly low threshold of suspicion, to ensure a high sensitivity and reduce the chances of missing someone with breast cancer. The subsequent biopsy of positive results will have a high specificity to ensure that if, say, mastectomy is to be considered, the doctor is almost certain that the patient has breast cancer.

Summary

This tutorial has summarised the methods used for examining the suitability of a particular test for diagnosing disease. In addition it has highlighted the difference between diagnostic and screening tests. In reality the same methods are used to evaluate both diagnostic and screening tests, the important difference being the emphasis that is placed on the sensitivity and specificity. Further details are given in Campbell, Machin and Walters (Campbell et al. 2007) Chapter 4.

Correlation Coefficient

Dr Jenny Freeman, Dr Tracey Young

Many statistical analyses can be undertaken to examine the relationship between two continuous variables within a group of subjects. Two of the main purposes of such analyses are:

- To assess whether the two variables are associated. There is no distinction between the two variables and no causation is implied, simply *association*.
- To enable the value of one variable to be predicted from any known value of the other variable. One variable is regarded as a *response* to the other *predictor (explanatory)* variable and the value of the predictor variable is used to *predict* what the response would be.

For the first of these, the statistical method for assessing the association between two *continuous* variables is known as *correlation*, whilst the technique for the second, prediction of one continuous variable from another is known as *regression*. Correlation and regression are often presented together and it is easy to get the impression that they are inseparable. In fact, they have distinct purposes and it is relatively rare that one is genuinely interested in performing both analyses on the same set of data. However, when preparing to analyse data using either technique it is always important to construct a scatter plot of the values of the two variables against each other. By drawing a scatter plot it is possible to see whether or not there is any visual evidence of a straight line or linear association between the two variables.

This tutorial will deal with correlation, and regression will be the subject of a later tutorial.

Correlation

The correlation coefficient is a measure of the degree of linear association between two continuous variables i.e. when plotted together, how close to a straight line is the scatter of points. No assumptions are made about whether the relationship between the two variables is causal i.e. whether one variable is influencing the value of the other variable; correlation simply measures the degree to which the two vary together. A positive correlation indicates that as the values of one variable increase the values of the other variable increase, whereas a negative correlation indicates that as the values of one variable increase the values of the other variable decrease. The standard method (often ascribed to Pearson) leads to a statistic called *r,* Pearson's correlation coefficient. In essence *r* is a measure of the scatter of the points around an underlying linear trend: the closer the spread of points to a straight line the higher the value of the correlation coefficient; the greater the spread of points the smaller the correlation coefficient. Given a set of *n* pairs of observations (x_1, y_1), (x_2, y_2),..(x_n, y_n) the formula for the Pearson correlation coefficient $\boldsymbol{r}$ is given by:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Certain assumptions need to be met for a correlation coefficient to be valid as outlined in Box 1. Both x and y must both be continuous random variables, (and Normally distributed if the hypothesis test is to be valid).

Pearson's correlation coefficient *r* can only take values between -1 and +1; a value of +1 indicates perfect positive association (Figure 1), a value of -1 indicates perfect negative association (Figure 2), and a value of 0 indicates no linear association (Figure 3).

Figure 1: perfect negative correlation (*r* =-1)

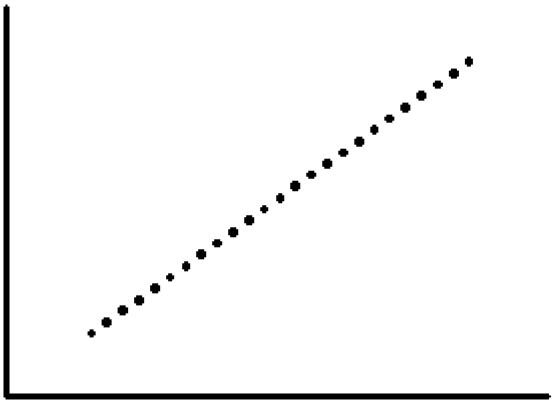


Figure 2: perfect negative correlation (*r* =-1)

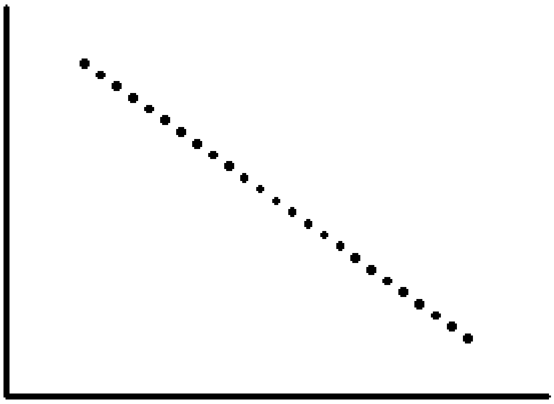
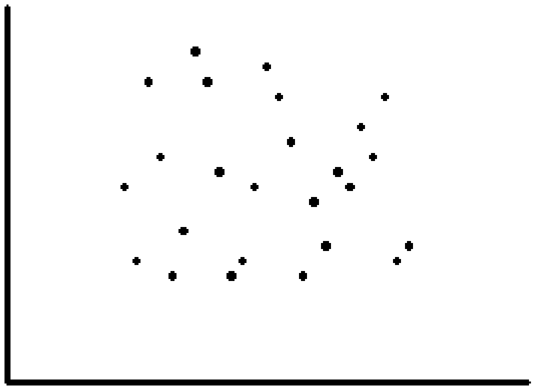


Figure 3: no linear association (*r* =0)



The easiest way to check whether it is valid to calculate a correlation coefficient is to examine the scatterplot of the data. This plot should be produced as a matter of routine when correlation coefficients are calculated, as it will give a good indication of whether the relationship between the two variables is roughly linear and thus whether it is appropriate to calculate a correlation coefficient all. In addition, as the correlation coefficient is highly sensitive to a few abnormal values, a scatterplot will show whether this is the case, as illustrated below in Figures 4 & 5:

Figure 4: The correlation for this plot is 0.8. It is heavily influenced by the extreme cluster of four points away from the main body

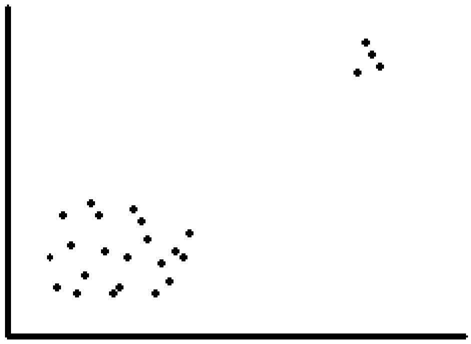
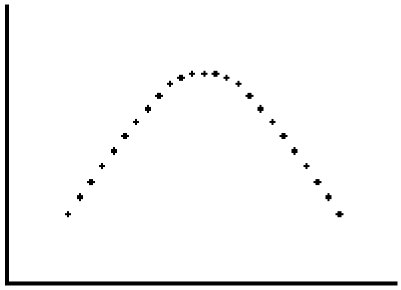


Figure 5: The correlation for this plot is close to 0. Whilst it is clear that the relationship is not linear and so a correlation is not appropriate, it is also clear that there is a strong n shaped relationship between these two variables.



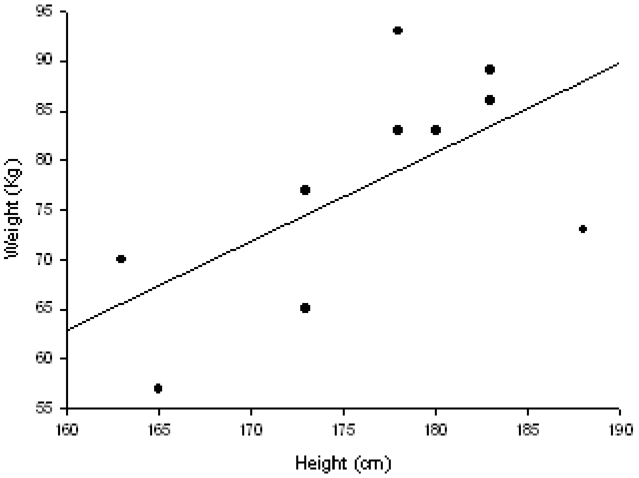
Example

Consider the heights and weights of 10 elderly men:

(173, 65), (165, 57), (173, 77), (183, 89), (178, 93), (188, 73), (180, 83), (183, 86), (163, 70), (178, 83)

Plotting these data indicates that, unsurprisingly, there is a positive linear relationship between height and weight. The shorter a person is the lower their weight and conversely, the taller a person is the greater their weight. In order to examine whether there is an association between these two variables, the *correlation coefficient* can be calculated. In calculating the correlation coefficient, no assumptions are made about whether the relationship is causal i.e. whether one variable is influencing the value of the other variable.

Figure 7: Plot of weight against height for ten elderly men



Subject	x	$x-\overline{x}$	$(x-\overline{x})^2$	y	$y-\overline{y}$	$(y-\overline{y})^2$	$(x-\overline{x})(y-\overline{y})$
1	173	-3.5	12.25	65	-12.5	156.25	43.75
2	165	-11.5	132.25	57	-20.5	420.25	235.75
3	174	-2.5	6.25	77	-0.5	0.25	1.25
4	183	6.5	42.25	89	11.5	132.25	74.75
5	178	1.5	2.25	93	15.5	240.25	23.25
6	188	11.5	132.25	73	-4.5	20.25	-51.75
7	180	3.5	12.25	83	5.5	30.25	19.25
8	182	5.5	30.25	86	8.5	72.25	46.75
9	163	-13.5	182.25	70	-7.5	56.25	101.25
10	179	2.5	6.25	82	4.5	20.25	11.25
Total	1765	0.0	558.50	775	0.0	1148.50	505.50

$\overline{x}$=1765 / 10 = 176.5 cm

$\overline{y}$=775 / 10 = 77.5 kg

r = 505.50 / √(558.50*1148.50) = 0.63

Thus the Pearson correlation coefficient for these data is 0.63, indicating that there is a positive association between height and weight for these 10 men. When calculating the correlation coefficient it is assumed that at least one of the variables is Normally distributed. If the data do not have a Normal distribution, a non-parametric correlation coefficient, Spearman's rho (*r_s)*, can be calculated. This is calculated in the same way as the Pearson correlation coefficient, except that the data are ordered by size and given ranks (from 1 to n, where n represents the total sample size) and the correlation is calculated using the ranks rather than the actual values. For the data above the Spearman correlation coefficient is 0.59.

Subject	Rank *(x)*	$x-\bar{x}$	$(x-\bar{x})^2$	Rank *(y)*	$y-\bar{y}$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	3 (173)	-2.5	6.25	2 (65)	-3.5	12.25	8.75
2	2 (165)	-3.5	12.25	1 (57)	-4.5	20.25	15.75
3	4 (174)	-1.5	2.25	5 (77)	-0.5	0.25	0.75
4	9 (183)	3.5	12.25	9 (89)	3.5	12.25	12.25
5	5 (178)	-0.5	0.25	10 (93)	4.5	20.25	-2.25
6	10	4.5	20.25	4 (73)	-1.5	2.25	-6.75
7	(188)	1.5	2.25	7 (83)	1.5	2.25	2.25
8	7 (180)	2.5	6.25	8 (86)	2.5	6.25	6.25
9	8 (182)	-4.5	20.25	3 (70)	-2.5	6.25	11.25
10	1 (163)	-0.5	0.25	6 (82)	0.5	0.25	0.25
	6 (179)						
Total	55	0.0	82.50	55	0.0	82.50	48.5

$\bar{x}$ (ranks) = 55 / 10 = 5.5

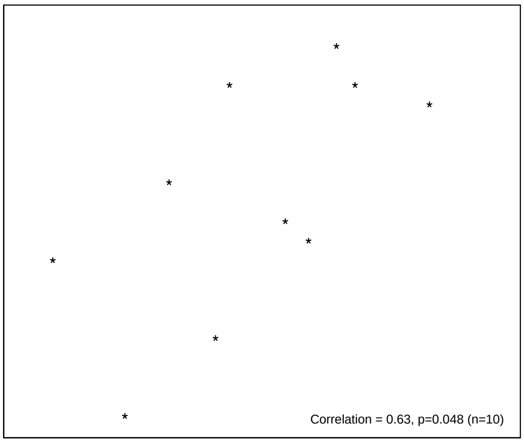
$\bar{y}$ (ranks) = 55 / 10 = 5.5

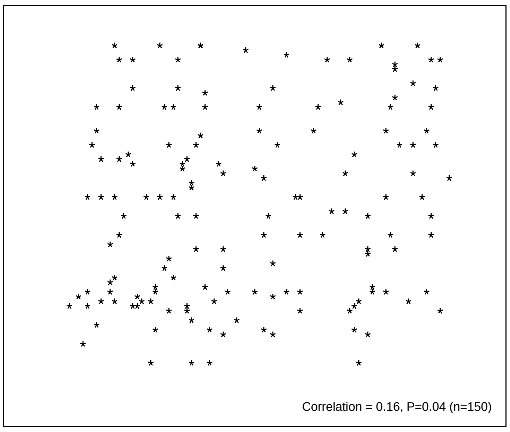
r_s = 48.5 / √(82.5*82.5) = 0.59

The square of the correlation coefficient gives the proportion of the variance of one variable explained by the other. For the example above, the square of the correlation coefficient is 0.398 indicating that about 39.8% of the variance of one variable is explained by the other.

Hypothesis testing

The null hypothesis is that the correlation coefficient is zero. However, its significance level is influenced by the number of observations and so it is worth being cautious when comparing correlations based on different sized samples. Even a very small correlation can be statistically significant if the number of observations is large. For example, with 10 observations, a correlation of 0.63 is significant at the 5% level, whereas with 150 observations, a correlation of 0.16 is significant at the 5% level. Figure 7a& b illustrate this:





The statistical test is based on the test statistic *t* = *r* / se(*r*) which under the null hypothesis follows a Students' *t* distribution on n-2 degrees of freedom and the confidence interval is given by:

The standard error of *r* = $\sqrt{\frac{1-r^2}{n-2}}$

For the Pearson correlation coefficient above the standard error is 0.27, the *t* statistic is 2.30 and the P-value is 0.05.

Box 1: The assumptions underlying the validity of the hypothesis test associated with the correlation coefficient

1. The two variables are observed on a random sample of individuals.
2. The data for at least one of the variables should have a Normal distribution in the population.
3. For the calculation of a valid confidence interval for the correlation coefficient both variables should have a Normal distribution.

When not to use a correlation coefficient

Whilst the correlation coefficient is a useful measure for summarising how two continuous variable are related, there are certain situations when it should not be calculated, as has already been alluded to above. As it measures the linear association between two variables, it should not be used when the relationship is non-linear. Where outliers are present in the data, care should be taken when interpreting its value. It should not be used when the values of one of the variables are fixed in advance, for example when measuring the responses to different doses of a drug. Causation should not be inferred from a correlation coefficient. There are many other criteria that need to be satisfied before causation can be concluded. Finally, just because two variables are correlated at a particular range of values, it should not be assumed that the same relationship holds for a different range.

Summary

This tutorial has outlined how to construct the correlation coefficient between two continuous variables. However, correlation simply quantifies the degree of linear association (or not) between two variables. It is often more useful to *describe* the relationship between the two variables, or even *predict* a value of one variable for a given value of the other and this is done using regression. If it is sensible to assume that one variable may be causing a response in the other then regression analysis should be used. If on the other hand, there is doubt as to which variable is the causal one, it would be most sensible to use correlation to describe the relationship. Regression analysis will be covered in a subsequent tutorial.

Simple Linear Regression

In the previous tutorial we looked at using correlation to assess the strength of the linear relationship between two continuous variables(Freeman and Young 2009). The correlation coefficient simply measures the strength of the linear association as a single number. No distinction is drawn between the two variables and no causation is implied. However, it is often of interest to quantify the relationship between two continuous variables, and given the value of one variable for an individual, to predict the value of the other variable. This is achieved using the technique known as simple linear regression. One variable is regarded as a *response* to the other *predictor (explanatory)* variable and the value of the predictor variable is used to *predict* what the response would be.

Scatter plots

As stated in the previous tutorial, when undertaking either a correlation or simple linear regression analysis it is important to construct a scatter plot of the data. The values of one variable are plotted on the horizontal axis (known as the x-axis) and the values of another are plotted on the vertical axis (y-axis). By drawing a scatter plot it is possible to see whether or not there is any visual evidence of a straight line or linear association between the two variables. It is possible for there to be a relationship between two variables but for that relationship to not be linear. In this case correlation or simple linear regression analysis may not be the most appropriate methods to use. In addition a scatterplot provides a good way of examining the data and checking for outliers or odd values.

If it is known (or suspected) that the value of one variable (known as the independent variable) influences the value of the other variable (known as the dependent variable), it is usual to plot the independent variable on the horizontal axis and the dependent variable on the vertical axis. In the case of height and weight, as height determines weight, to an extent, and not the other way around, a scatterplot of weight against height would be plotted with height on the horizontal axis and weight on the vertical axis.

Simple linear regression

In the technique of simple linear regression a straight-line equation is used to model the relationship between the predictor variable and the response variable. The equation of the regression line is given by:

$$y = a + bx$$

Where :

x = independent / predictor / explanatory variable: variable that is used to predict the values of the response variable. This is plotted on the horizontal axis of a scatter plot

y = dependent / response / outcome variable: variable being predicted by the model. This is plotted on the vertical axis of a scatterplot

a = intercept. This is the point at which the regression line crosses the vertical (Y) axis. Strictly speaking this gives the value of the Y variable (dependent variable) when the X variable (independent variable) is zero.

b = regression coefficient. It is also known as the slope and it shows the average change in the Y variable (outcome) for a unit change in the X variable (predictor/explanatory variable)

a and *b* are calculated as follows:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

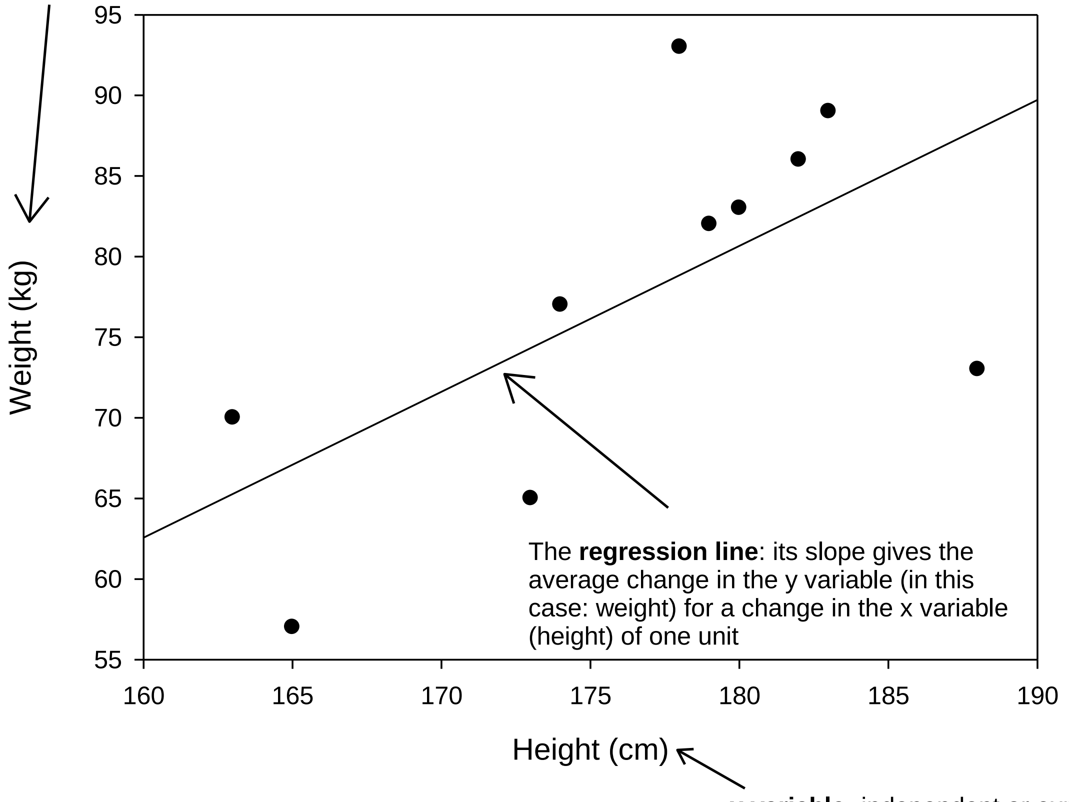
$$a = \bar{y} - b\bar{x}$$

Example: Simple linear regression of weight against height for ten elderly men

The figure below shows the height and weight values for ten elderly men. The data are given in Table 1

Figure 1: Scatter plot of weight against height together with the regression line

y variable: dependent or response variable, plotted on the vertical axis



The **regression line**: its slope gives the average change in the y variable (in this case: weight) for a change in the x variable (height) of one unit

x variable: independent or explanatory variable, plotted on the horizontal axis

Table 1: Calculation of regression equation for regression of weight on height of 10 elderly men

Subject	x	$(x-\bar{x})$	$(x-\bar{x})^2$	y	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
1	173	-3.5	12.25	65	-12.5	156.25	43.75
2	165	-11.5	132.25	57	-20.5	420.25	235.75
3	174	-2.5	6.25	77	-0.5	0.25	1.25
4	183	6.5	42.25	89	11.5	132.25	74.75
5	178	1.5	2.25	93	15.5	240.25	23.25
6	188	11.5	132.25	73	-4.5	20.25	-51.75
7	180	3.5	12.25	83	5.5	30.25	19.25
8	182	5.5	30.25	86	8.5	72.25	46.75
9	163	-13.5	182.25	70	-7.5	56.25	101.25
10	179	2.5	6.25	82	4.5	20.25	11.25
Total	1765	0	558.5	775	0	1148.5	505.5

$\bar{x}$ =1765/10 = 176.5cm

$\bar{y}$ =775/10 = 77.5 kg

b = 505.5 / 558.5 = 0.9051

a = 77.5 – 0.905103*176.5 = -82.25

Thus the regression equation for these data is:

weight = -82.25 +0.9051 * height

From this it can be seen that the slope coefficient was 0.9051, indicating that for every 1cm increase in height there was an increase in weight of 0.9051 kg. Note that the value of the intercept is -82.25. Thus when height is zero weight is -82.25kg. Clearly this is nonsense and illustrates an important principle for regression analyses: they should never be used to predict values outside of the range of observations. However, within the range of the data the regression equation can be used to predict the values of the y variable for particular values of the x variable. For example the estimated weight for an elderly man who was 180cm tall is calculated as follows:

weight = -82.25 +0.9051 * 180 = 80.67kg

Assumptions and model fit

Three important assumptions underlie a simple linear regression analysis as outlined in Box 1 and as with any statistical analysis it is important to check that they are valid and that the model fits the data adequately. The first assumption can be checked by constructing a scatter plot of the data (Figure 1).

Assumptions:

1. The relationship between the two variables should be linear
2. The value of the response variable, y, should have a Normal distribution for each value of the explanatory variable x
3. The variance (or standard deviation) of y should be the same at each value of x i.e. there should be no evidence that as the value of y changes, the spread of the x values changes

The final two assumptions can be checked by examining the residuals from the fitted model. Each y observation has a residual associated with it; this is the difference between the actual observed y value (y_{obs}) and the y value predicted by the model (known as the fitted value (y_{fit})) (see table 2). In Figure 1 for each point the residual is given by the vertical distance between that point and the fitted regression line. For example, for the first observation in Table 1, the actual weight is 65kg and the predicted weight is 74.33kg, thus the residual is given by 65 – 74.33= -9.33kg.

Table 1: Calculation of residuals from the fitted model

Actual height (m)	Actual weight (kg)	Predicted value = -82.25+ height*0.9051	Residual (y_{obs} -y_{fit})
173	65	74.33	-9.33
165	57	67.09	-10.09
174	77	75.24	1.76
183	89	83.38	5.62
178	93	78.86	14.14
188	73	87.91	-14.91
180	83	80.67	2.33
182	86	82.48	3.52
163	70	65.28	4.72
179	83	79.76	2.24

In order for assumption 2 to be valid the residuals should be Normally distributed. This is most easily checked by constructing a histogram of the residuals to check that this is approximately Normal (Figure 2). With only 10 individuals it is difficult to definitively conclude that the residuals are Normally distributed, but given that they are spread out around a central peak it would appear to be reasonable to accept this assumption as being valid. In order to check assumption 3 it is necessary to do a scatter plot of the residuals against the predicted values. This should show a good spread with no obvious patterns (i.e. it looks random) as in Figure 3 below.

Figure 2: Histogram of the residuals from the fitted model

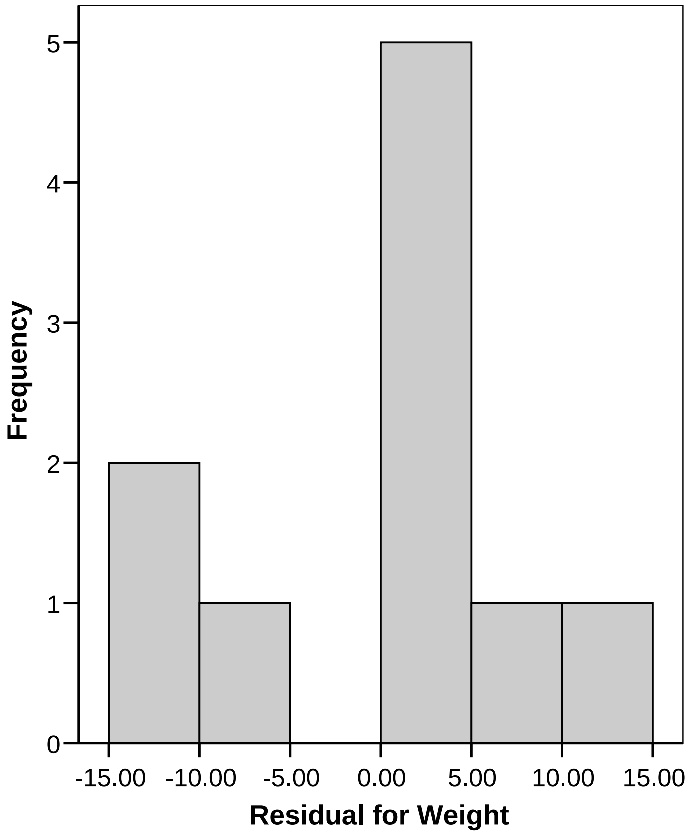
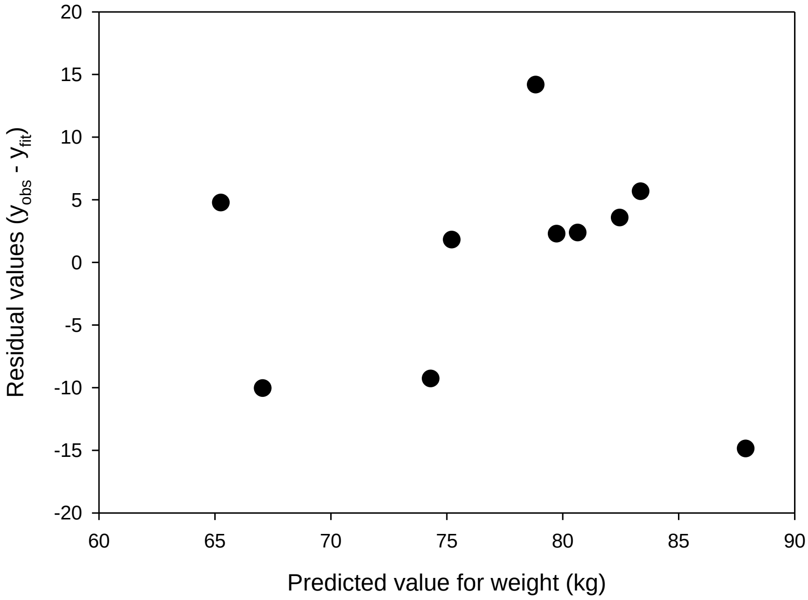


Figure 3: Plot of the residuals from the model against the predicted values



R^2

The value of r^2 is often quoted in published articles and indicates the proportion (sometimes expressed as a percentage) of the total variability of the outcome variable that is explained by the regression model fitted. A well fitting model will have a high r^2 and a badly fitting model will have a low value of R^2. It is calculated as follows:

$$r^2 = \frac{\left(\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}$$

(Note that this is also the square of the correlation coefficient:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}})$$

For the current example the value of r^2 is 0.398. Thus 39.8% of the total variability in weight for the ten men is explained by their heights.

More than one explanatory variable

Simple linear regression as described above involves the investigation of the effect of a single explanatory variable on the outcome of interest. However, there is usually more than one possible explanatory variable influencing the values of the outcome variable and the method of regression can be extended to investigate the influence of more than one explanatory variable on the outcome of interest. In this case it is referred to as multiple regression, and the influence of several explanatory variables can be investigated simultaneously. This is beyond the scope of the current tutorial and will be covered in a subsequent tutorial.

Summary: Regression or Correlation?

Regression and correlation are related methods (note that the r2 coefficient is simply the square of the correlation coefficient!). As they are often presented together it is easy to get the impression that they are inseparable. In fact, they have distinct purposes and it is relatively rare that one is genuinely interested in performing both analyses on the same set of data. Regression is more informative than correlation. Correlation simply quantifies the degree of linear association (or not) between two variables. However, it is often more useful to *describe* the relationship between the two variables, or even *predict* a value of one variable for a given value of the other and this is done using regression. If it is sensible to assume that one variable may be causing a response in the other then regression analysis should be used.

Risk

Jenny Freeman, Dawn Teare

The analysis of data related to risk is important to many fields in medicine, particularly when explaining different treatment options to patients. There are various ways in which risk can be measured and this paper will explain some of the more common measures used including absolute risk, relative risk, odds, odds ratio and number needed to treat.

Risk data

It is often of interest to know about the risks associated with particular events or exposures, for example the risk of developing lung cancer for smokers. At the most basic level risk data are often divided into categories depending on whether individuals are exposed to the hazard of interest or not, and whether they experience the event of interest or not. Data such as these can be organised as follows:

Table 1: 2x 2 table illustrating the calculation of risk

	Exposure:		Total
	Yes	No	
Event:			
Yes	a	b	a+b
No	c	d	c+d
	a+c	b+d	n

where:

a = number of individuals who are exposed and have the event of interest
b = number of individuals not exposed who have the event of interest
c = number of individuals exposed who do not have the event of interest
d = number of individuals not exposed who do not have the event

Risk/Absolute risk

The simplest measure of risk is the absolute risk of an event occurring. This is sometimes simply referred to as the risk and is the number of individuals in the population under study who experience the event of interest within a defined period of time divided by the total number of individuals in the group at the start of the time period.

$$\text{Absolute risk of an event} = \frac{\text{number who have the event of interest}}{\text{number in the group at the start of the follow - up period}}$$

For the data in Table 1:

- Absolute risk of event for the exposed group = a/(a+c)
- Absolute risk of event for the unexposed group = b/(b+d)

The absolute risk is a measure of how likely an event is to occur and is a probability. All probabilities range between 0 and 1, a value of 1 denotes an event that is certain to happen and 0 denotes an event that is never going to happen. It is only possible to obtain the risk of an event occurring with data that are longitudinal in nature as in order to calculate the risk of an event in a given period of time it is necessary to know the total number who were at risk at the start of the time period. Thus it is not possible to compute a risk for data collected as part of a case-control study (a case-control study is one in which a group of subjects (cases) with the disease or condition of interest are compared to a group of subjects (controls) without the disease).

Example 1:

A recent study looked at the effects of the introduction of laparoscopic bariatric surgery in England. The authors examined the 28 day readmission rates by type of procedure. Of the 3191 patients who underwent gastric bypass 308 were readmitted within 28 days, whereas of the 3649 patients who underwent gastric banding 232 were readmitted within 28 days(Burns et al. 2010). These results are reported in the following 2 x 2 table. The columns represent the type of procedure and the rows represent readmission within 28 days.

Table 2: 2x 2 table illustrating the calculation of risk

	Exposure:		Total
	Gastric bypass	Gastric banding	
Event:			
Readmitted within 28 days	308	232	540
Not readmitted	2883	3417	6300
	3191	3649	6840

For these data:

- The absolute risk of readmission (within 28 days) for those patients who underwent a gastric bypass is 308 / 3191 = 0.097
- The absolute risk of readmission (within 28 days) for those patients who underwent gastric banding is 232 / 3649 = 0.064

Absolute risk difference

This is the absolute additional risk of an event due to a particular exposure. It is calculated as the risk in the exposed group minus the risk in the unexposed group (ignoring the sign).

Absolute Risk Difference = |risk in the exposed – risk in the unexposed|[5]

If the risk is harmful, so that the risk is increased by the exposure this difference is called the **absolute risk excess (ARE)** (for example the absolute risk excess for gastric bypass compared to gastric banding is |0.097 – 0.063| = 0.033) and it represents the absolute increase in risk for those exposed compared to the unexposed. If the risk is decreased by the exposure (for example using sunscreen to reduce the risk of melanoma) then this difference is called the **absolute risk reduction (ARR)**. A recent randomised controlled trial looking at the risk of

[5] The symbols | | are mathematical notation for the modulus and indicate that for any quantity within these only the absolute value is to be used, the sign is to be ignored. Thus the modulus of -1 (written as |-1|) is 1

secondary lymphoedema following treatment for breast cancer compared a group who had early physiotherapy and education with a control group who had education alone(Lacomba et al. 2010). At the end of a year's follow-up the two groups were compared for the occurrence of lymphoedema. The data are shown in Table 3:

Table 3: 2x 2 table illustrating the calculation of risk

	Exposure:		Total
	Physiotherapy (n=59)	Control (n=57)	
Event:			18
Lymphoedema	4	14	
No lymphoedema	55	43	98

The risk of lymphoedema for the physiotherapy group was 4/59 = 0.068 and the risk of lymphoedema for the control group was 14/57=0.246. Thus the absolute risk reduction was 0.068-0.246 =| -0.178|=0.178.

Relative risk

The relative risk of a particular event for a given exposure is the ratio of the risk of the event occurring in the exposed group divided by the risk of the event occurring in the unexposed group. Using the terminology in Table 1:

$$\text{Relative risk of an event} = \frac{a/(a+c)}{b/(b+d)} = \frac{a(b+d)}{b(a+c)}$$

For the data in table 2 the relative risk of readmission for those patients undergoing a gastric bypass compared to those undergoing gastric banding is 0.097 / 0.064 = 1.52. Thus patients who undergo gastric bypass are 1.52 times more likely to be readmitted to hospital within 28 days then patients who undergo gastric banding.

RRR / Relative risk reduction

In clinical trials when looking at the benefits of one treatment compared to another, the relative risk reduction can also be calculated. This is the extent to which a treatment reduces a risk in comparison to a group not receiving the treatment of interest. In this context the risk is an adverse outcome or event. It is calculated as follows:

$$\begin{aligned}\text{Relative risk reduction} &= \frac{\text{risk in the control group - risk in the treated group}}{\text{risk in the control group}} \\ &= \frac{\text{risk in the control group}}{\text{risk in the control group}} - \frac{\text{risk in the treated group}}{\text{risk in the control group}} \\ &= 1 - \text{relative risk}\end{aligned}$$

As well as being expressed as a proportion it can also be expressed as a percentage i.e. 100*(1-RR). In the trial of physiotherapy for the prevention of lymphoedema the relative risk reduction is 1-(0.068/0.246) = 0.72, or expressed as a percentage 72%. Thus the relative risk reduction of using physiotherapy to prevent lymphoedema is 72%.

Odds

The odds of an event occurring is the ratio of the probability of the event occurring to the probability of the event not occurring. Using the terminology in Table 1:

Odds of an event given exposure = a/c
Odds of an event given not exposed = b/d

Example

For the data in Table 2 the odds of readmission to hospital within 28 days for patients undergoing gastric bypass was 308/2883 = 0.107 and the odds of readmission to hospital within 28 days for patients undergoing gastric banding was 232/3417 = 0.068

Odds ratio

The odds ratio is the ratio of the odds of an event in the exposed group compared to the unexposed group. Using the terminology of Table 1:

Odds ratio of an event (for exposed compared to not exposed) $= \frac{a/c}{b/d} = \frac{ad}{bc}$

It is used in case-control studies as way to understanding the risk in different groups, as it is not necessary to have a measure of the numbers initially exposed who then develop disease. *Note that when the disease is rare the odds ratio can be interpreted as a relative risk, this is because as a gets smaller (a+c) approaches c and as b gets smaller (b+d) approaches d.* One other useful property of the odds ratio is that it is reversible. i.e. the odds ratio for exposure is the same as the odds ratio for the event of interest. To illustrate this, let us consider the odds of exposure for two groups, one of which experienced the event of interest whilst the other group did not. Using the terminology in Table 1

Odds of an exposure given the event = a/b

Odds of an exposure given no event = c/d

Thus:

Odds ratio of exposure (for event group compared to no event group)

$$= \frac{a/b}{c/d} = \frac{ad}{bc}$$

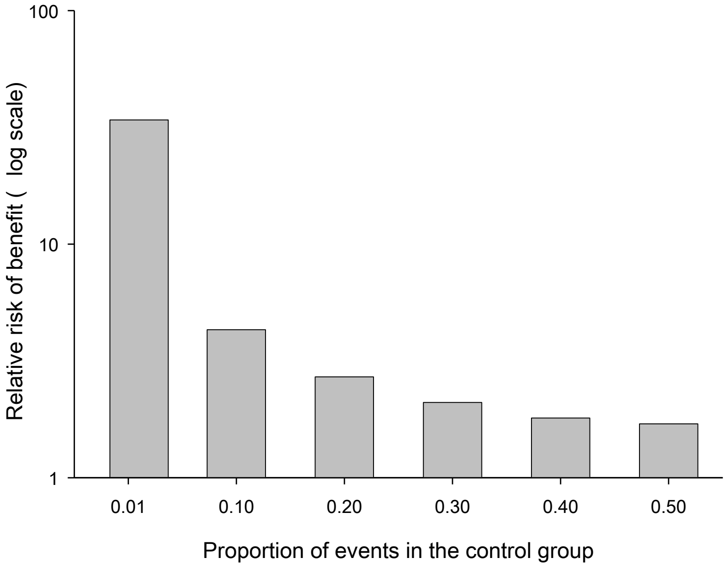
which is the same as the odds ratio of an event. The relative risk does not have this property. For the data in Table 2 the odds ratio of readmission to hospital for patients having gastric bypass compared to patients having gastric banding = 0.107 / 0.068 = 1.57. Note that in this case, as the event of interest (readmission to hospital) is relatively rare, the odds ratio is very similar to the relative risk of 1.52.

Number needed to treat

This is a measure of the impact of a particular risk on patients often used in clinical practice. It is the additional number of people that would need to be given a new treatment in order to cure one extra person compared to the old treatment and is calculated as the reciprocal of the absolute risk reduction = 1/ARR. Alternatively, for a harmful exposure the number needed to treat is referred to as the **number needed to harm**. It represents the additional number of individuals who need to be exposed to the risk in order to have one extra person experience the event of interest, compared to the unexposed group. For the gastric bypass study the number needed to harm is 1/0.033=30.3. As it is usual to round this to the nearest whole number above, the number needed to harm is 31. For the lymphoedema trial the number needed to treat is 1/0.178 = 5.6. Thus the number needed to treat in order to have one addition person without lymphoedema is 6.

However, when calculating the number needed to treat, it is important to know what the absolute risks that it is based upon are. Even though the risk in the control group can change dramatically, giving very different relative risks, the number needed to treat can stay constant, as illustrated by the Figure below. Each bar represents the results of a fictional study. For all of these studies the number needed to treat is 3. The figure shows how the relative risk changes for different values of the risk in the control group. When the risk in the control group is 0.01 (i.e. 1% risk of an event) the relative risk is 34.3 whereas when the risk in the control group has increased to 0.5 (50% risk of an event) the relative risk has decreased to 1.67.

Figure 1: Figure illustrating what happens to the relative risk for different control event rates, when the number needed to treat is fixed at 3.



Risk ladders

A risk ladder is a visual way of quantifying different risks, in comparison to each other. It is often used in clinical practice when explaining individual risks to patients and enables patients to quantify their particular risk in relation to other risks. One of the most well-known is the Calman Chart below(Calman 1996):

Table : Descriptions of risk in relation to the risk of an individual dying (D) in any one year or developing an adverse response (A) in one year(Calman 1996)

Term used	Risk range	Example	Risk
High	= 1: 100	Transmission to susceptible household contacts of measles and chickenpox (A)	1:1 to 1:2
		Transmission of HIV from mother to child (Europe) (A)	1:6
		Gastrointestinal effects of antibiotics (A)	1:10 to 1:20
Moderate	1:100 to 1:1,000	Smoking 10 cigarettes a day (D)	1:200
		All natural causes, age 40 (D)	1:850
Low	1:1,000 to 1:10,000	All kinds of violence and poisoning (D)	1:3,300
		Influenza (D)	1:5,000
		Accident on road (D)	1:8,000
Very low	1:10,000 to 1: 100,000	Leukaemia (D)	1:12,000
		Playing football (D)	1:25,000
		Accident at home (D)	1:26,000
		Accident at work (D)	1:43,000
		Homicide (D)	1:100,000
Minimal	1:100,000 to 1:1,000,000	Accident on railway (D)	1:500,000
		Vaccination associated polio (A)	1:1,000,000
Negligible	<= 1:1,000,000	Hit by lightening (D)	1:10,000,000
		Release of radiation by nuclear power station	1:10,000,000

Points to consider when communicating risk

Individuals who do not deal with numbers and data regularly can often struggle to understand measures of risk, and this case it can be useful to express risks in terms of natural frequencies rather than percentages. Thus if we assume that the success rate following a single cycle of IVF is about 33% then it is more easily understood by stating that of 100 women undergoing treatment 33 will become pregnant. In addition, how a risk is perceived will depend upon how it is presented. Relative risks are often presented, but these cannot but properly understood without reference to the baseline risks involved. Whilst a relative risk of 2 might sound large, if the underlying baseline risk is 1 in 10,000 and this increases to 2 in 10,000, then this will represent a very different risk to an individual than if the baseline risk were 1 in 10 compared to 2 in 10. They both have the same relative risk, but the risk to an individual is very different. When presented with a risk expressed in relative terms it is always useful to know what the baseline risk is. A good further description of both risk ladders and the communication of risk can be found in an article by Edwards et al(Edwards et al. 2002).

It is also worth bearing in mind that in all the data that have been presented in this tutorial, no other factors have been taken into account. This is particularly important when considering the gastric bypass/gastric band data. As these are data from an observational cohort study and not from a randomised controlled trial, it may be that the patients who underwent gastric bypass were different to the patients who underwent gastric banding and these other differences could explain the difference in risk of readmission by 28 days, rather than any underlying risk associated with the actual procedure.

References

References

Marital status of the UK population, by sex, 2001. www.statistics.gov.uk/STATBASE/Expodata/Spreadsheets/D7680.xls . 17-6-2005.
Ref Type: Electronic Citation

Altman, D.G. 1991. *Practical Statistics for Medical Research* London, Chapman & Hall.

Altman, D.G. & Bland, J.M. 1996. Presentation of numerical data. *British Medical Journal*, 312, 572

Altman, D.G. & Bland, J.M. 1999. Treatment allocation in controlled trials: why randomise. *British Medical Journal*, 318, 1209

Altman, D.G., Machin, D., Bryant, T., & Gardner, M.J. 2000. *Statistics with Confidence*, 2nd ed. London, BMJ Books.

Armitage, P., Berry, P.J., & Matthews, J.N.S. 2002. *Statistical Methods in Medical Research*, 4 ed. Oxford, Blackwells.

Bigwood, S. & Spore, M. 2003. *Presenting Numbers, Tables and Charts* Oxford, Oxford University Press.

Bird, D. Methodology for the 2004 annual survey of hours and earnings. Labour market trends. http://www.statistics.gov.uk/articles/nojournal/ASHEMethod_article.pdf , 457-464. 2004. Office for National Statistics. 17-6-2004.
Ref Type: Electronic Citation

Boogert, A., Manhigh, A., & Visser, G.H.A. 1987. The immediate effects of chorionic villus sampling on fetal movements. *American Journal of Obstetrics and Gynaecology*, 157, 137-139

Bradford Hill, A. 1990. Memories of the British streptomycin trial: the first randomised clinical trial. *Controlled Clinical Trials*, 11, 77-79

Burns, E.M., Naseem, H., Bottle, A., Lazzarino, A.I., Aylin, P., Darzi, A., Moorthy, K., & Faiz, O. 2010. Introduction of laparoscopic bariatric surgery in England: observational population cohort study. *British Medical Journal*, 341, c4296

Calman, K. 1996. Cancer: science and society and the communication of risk. *British Medical Journal*, 313, 799-802

Campbell, M.J., Machin, D., & Walters, S.J. 2007. *Medical Statistics: a textbook for the health sciences*, 4 ed. Chichester, Wiley.

Cox, I.M., Campbell, M.J., & Dowson, D. 1991. Red blood cell magnesium and chronic fatigue syndrome. *Lancet*, 337, 757-760

Day, S.J. & Altman, D.G. 2000. Blinding in clinical trials and other studies. *British Medical Journal*, 321, 504

Edwards, A., Elwyn, G., & Mulley, A. 2002. Explaining risks: turning numerical data into meaningful pictures. *British Medical Journal*, 324, 827-830

Ehrenberg, A.S.C. 2000. *A primer in data reduction* Chichester, John Wiley & Sons.

Freeman, J.V. & Campbell, M.J. 2006. Basic test for continuous data: Mann-Whitney U and Wilcoxon signed rank sum tests. *Scope*, 15, (4)

Freeman, J.V. & Julious, S.A. 2005a. Describing and summarising data. *Scope*, 14, (3)

Freeman, J.V. & Julious, S.A. 2005b. The Normal Distribution. *Scope*, 14, (4)

Freeman, J.V. & Julious, S.A. 2005c. The visual display of quantitative information. *Scope*, 14, (2) 11-15

Freeman, J.V. & Julious, S.A. 2006a. Basic tests for continuous Normally distributed data. *Scope*, 15, (3)

Freeman, J.V. & Julious, S.A. 2006b. Hypothesis testing and estimation. *Scope*, 15, (1)

Freeman, J.V. & Julious, S.A. 2007. The analysis of categorical data. *Scope*, 16, (1) 18-21

Freeman, J.V. & Young, T.A. 2009. Correlation coefficient: association between two continuous variables. *Scope* 31-33

Huff, D. 1991. *How to lie with statistics* London, Penguin Books.

Julious, S.A. & Mullee, M.A. 1994. Confounding and Simpson's paradox. *British Medical Journal*, 308, 1408-1481

Julious, S.A. & Zariffa, N. 2002. ABC of pharmaceutical trial design. *Phamaceutical Statistics*, 1, (1) 45-54

Kroenke, K., Spitzer, R.L., Williams, J.B., Monahan, P.O., & Löwe, B. 2007. Anxiety disorders in primary care: prevalence, impairment, comorbidity and detection. *Annals of Internal Medicine*, 146, (317) 325

Lacomba, M.T., Sanchez, M.J., Goni, A.Z., Merino, D.P., Moral, O., Tellez, E.C., & Mogollon, E.M. 2010. Effectiveness of early physiotherapy to prevent lymphoedema after surgery for breast cancer: randomised, single blinded, clinical trial. *British Medical Journal*, 340, b5396

Lang, T.A. & Secic, M. 1997. *How to report statistics in medicine* Philadelphia, American College of Physicians.

Medical Research Council 1948. Streptomycin treatment of pulmonary tuberculosis. *British Medical Journal*, 2, (769) 782

Morrell, C.J., Walters, S.J., Dixon, S., Collins, K., Brereton, L.M.L., Peters, J., & Brooker, C.G.D. 1998. Cost effectiveness of community leg ulcer clinic: randomised controlled trial. *British Medical Journal*, 316, 1487-1491

Nisar, A., Morris, M.W.J., Freeman, J.V., Cort, J.M., Rayner, P.R., & Shahane, S.A. 2007. Subacromial bursa block is an effective alternative to Interscalene block for postoperative pain control after arthroscopic subacromial decompression - a randomised trial. *Journal of Shoulder and Elbow Surgery*, In press,

Simpson, E.H. 1951. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B* (2) 238-241

Swinscow, T.D.V. & Campbell, M.J. 2002. *Statistics at square one*, 10 ed. London, BMJ Books.

Tufte, E.R. 1983. *The visual display of quantitative information* Cheshire, Connecticut, Graphics Press.

Williams, E.J. 1949. Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research, Series A* (2) 149-168

Yoshioka, A. 1998. Use of randomisation in the Medical Research Council's clinical trials of streptomycin in pulmonary tuberculosis in the 1940s. *British Medical Journal*, 317, (1220) 1223