NHS
National Institute for Health Research

The NIHR Research Design Service for the East Midlands

The NIHR Research Design Service for Yorkshire & the Humber

Using Statistics in Research

Author

Stephen J Walters

This Resource Pack is one of a series produced by The NIHR RDS for the East Midlands / The NIHR RDS for Yorkshire and the Humber. This series has been funded by The NIHR RDS EM / YH.

Reference as:
Walters S.J. Using Statistics in Research. The NIHR RDS for the East Midlands / Yorkshire & the Humber, 2007.

Stephen J Walters
The NIHR RDS for the East Midlands / Yorkshire & the Humber
School of Health and Related Research (ScHARR)
University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA

Last updated: 2009

The NIHR RDS for the East Midlands
Division of Primary Care,
14th Floor, Tower building
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: 0115 823 0500

www.rds-eastmidlands.nihr.ac.uk

Leicester: enquiries-LNR@rds-eastmidlands.org.uk

Nottingham: enquiries-NDL@rds-eastmidlands.org.uk

The NIHR RDS for Yorkshire & the Humber
ScHARR
The University of Sheffield
Regent Court
30 Regent Street
Sheffield
S1 4DA
Tel: 0114 222 0828

www.rds-yh.nihr.ac.uk

Sheffield: rds-yh@sheffield.ac.uk
Leeds: rds-yh@leeds.ac.uk
York: rds-yh@york.ac.uk

Table of Contents

1. Introduction

Whatever type of research we undertake (for example, naturalistic observation, case study, surveys or experiments) our efforts can usually generate a considerable amount of data: numbers that represent our research findings and provide the basis for our conclusions. Statistical analyses are the methods most often used to summarise and interpret data. Very often it is the fear of using statistics that prevents us from fully realising the potential of our data. In reality, statistics can be a straightforward and enjoyable process! Knowledge of complex mathematical theories and formulae are not a pre-requisite, just confidence in the data and in your own ability to analyse and interpret it.

This pack is set out to achieve a single aim, to guide you, the researcher, to select the most appropriate statistical test for your data in order to evaluate its significance to your research aims and hypotheses. It is not meant to be an 'all inclusive' guide to statistical testing but rather a starting point for those who are newcomers to statistical analysis. No mathematical background is assumed or familiarity with research methodology and design, see The NIHR RDS EM / YH Resource Pack: '*Experimental Designs'*. However, as you become more experienced and confident in the use of statistics you may want to use it as a manual to be 'dipped into' at the appropriate place for your current research project. It should also be noted that although the orientation of this pack will be on 'quantitative data', other types of data (for example, qualitative) are extensively used in research.

2. Selecting your participants: Populations and samples

How you select the participants (alternatively you can use the term 'subjects') for your study is crucial. It can make the difference between a well designed experimental study which lends itself to appropriate statistical analysis or it can produce a poorly designed study which, among other things, can affect the statistical analysis used and ultimately the generalisability of the results. It is therefore important that the participants who will take part in your study are the most appropriate for your research question. As a researcher you must ensure that, as a group, your participants are not atypical or unrepresentative of the population you want to generalise to. This advice applies to all types of participant groups, whether they are patients, clients, chiropodists, speech therapists, or some similar group. In order to achieve this you need to consider the distinction between a population and a sample.

2.1 Populations

In the statistical sense a population is a theoretical concept used to describe an entire group of individuals in whom we are interested. Examples are the population of all patients with diabetes mellitus, or the population of all middle-aged men. Parameters are quantities used to describe characteristics of such populations. Thus the proportion of diabetic patients with nephropathy, or the mean blood pressure of middle-aged men, are characteristics describing the two populations. Generally, it is costly and labour intensive to study the entire population. Therefore we collect data on a sample of individuals from the population who we believe are representative of that population, that is, they have similar characteristics to the individuals in the population. We then use them to draw conclusions, technically make inferences, about the population as a whole. The process is represented schematically in Figure 1. So samples are taken from populations to provide estimates of population parameters. Some common population parameters and their corresponding sample statistics or estimates are described in Table 1.

Figure 1: Population and sample

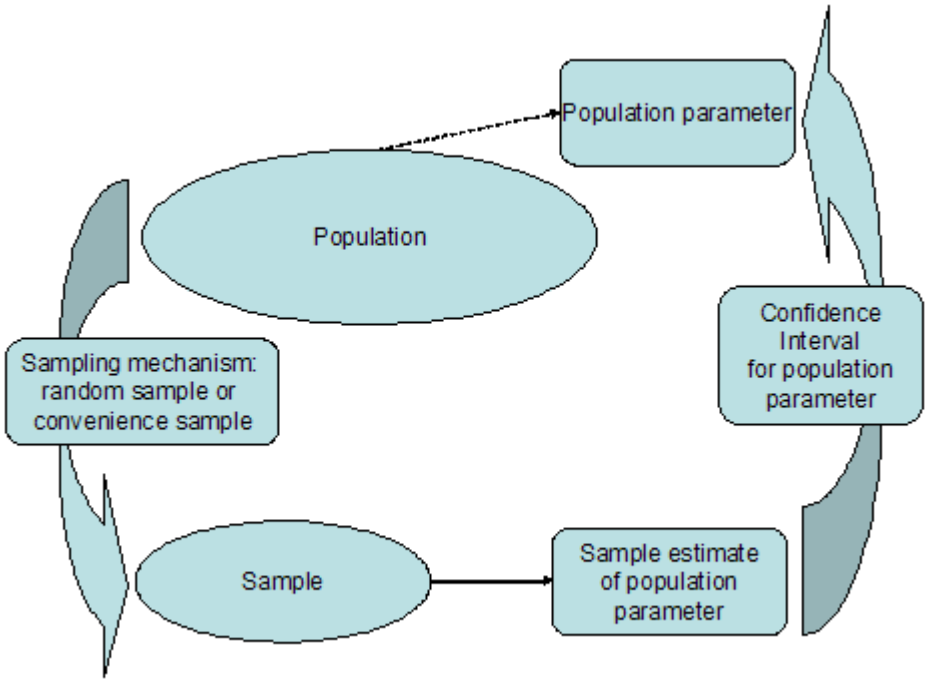


Table 1: Population parameters and sample statistics

	Population parameter	Sample statistic
Mean	μ	$\bar{x}$
Standard deviation	•	s
Proportion	π	p
Rate	λ	r

It is important to note that although the study populations are unique, samples are not, as we could take more than one sample from the target population if we wished. Thus for middle-aged men there is only one normal range for blood pressure. However, one investigator taking a random sample from a population of middle-aged men and measuring their blood pressure may obtain a different normal range from another investigator who takes a different random sample from the same population of such men. By studying only some of the population we have introduced a sampling error.

2.2 Samples

In some circumstances the sample may consist of all the members of a specifically defined population. For practical reasons, this is only likely to be the case if the population of interest is not too large. If all members of the population can be assessed, then the estimate of the parameter concerned is derived from information obtained on all members and so its value will be the population parameter itself. In this idealised situation we know all about the population, as we have examined all its members and the parameter is estimated with no bias. The dotted arrow in Figure 1 connecting the population ellipse to the population parameter box illustrates this. However, this situation will rarely be the case, so in practice we take a sample which is often much smaller in size than the population under study.

Ideally we should aim for a random sample. A list of all individuals from the population is drawn up (the sampling frame), and individuals are selected randomly from this list, that is, every possible sample of a given size in the population has an equal chance of being chosen. Sometimes, there may be difficulty in constructing this list or we may have to 'make-do' with those subjects who happen to be available or what is termed a convenience sample. Essentially if we take a random sample then we obtain an unbiased estimate of the corresponding population parameter, whereas a convenience sample may provide a biased estimate but by how much we will not know. Different types of samples, including random samples, are further described in The NIHR RDS EM / YH Resource Pack: '*Sampling*.'

Sample size

It is much easier to produce a biased sample when sample sizes are small, although in some types of clinical research the subject area dictates the use of small numbers. In the field of clinical neuropsychology, for example, you may be interested in evaluating specific everyday memory deficits in patients with acquired temporal lobe deficits. The rarity of this condition may make it possible for you to gain access to only one or two patients. As a general rule, however, you should aim for the largest sample size possible for your type of data. This issue becomes especially relevant when using a research design that involves collecting data from such methods as questionnaires (structured and semi-structured) and surveys, which if not produced in large enough quantities are more likely to produce a biased response based upon an unrepresentative, small sample size. It is also important to remember that the larger the data set the greater the reliability of the statistical analysis. Importantly, statistical analysis will help you reduce your chance of obtaining random errors in two main ways: first, by ensuring accurate and consistent description of a sample from a target population and second, by providing a consistent basis for the inference of characteristics of your target population, based on the characteristics of only a sample.

In summary, it is important that from the very onset of your research study you carefully address the issues of composition and size of your sample. As a

researcher you must avoid working with samples that are not representative of your target population. You should also try to ensure that your study size is sufficient, whenever possible to give a decisive answer, one way or another, to your basic research question. Following these procedures will increase the likelihood of you being able to draw valid conclusions from your research design and analysis and reduce the risk of misinterpretation due to biased sampling and chance error.

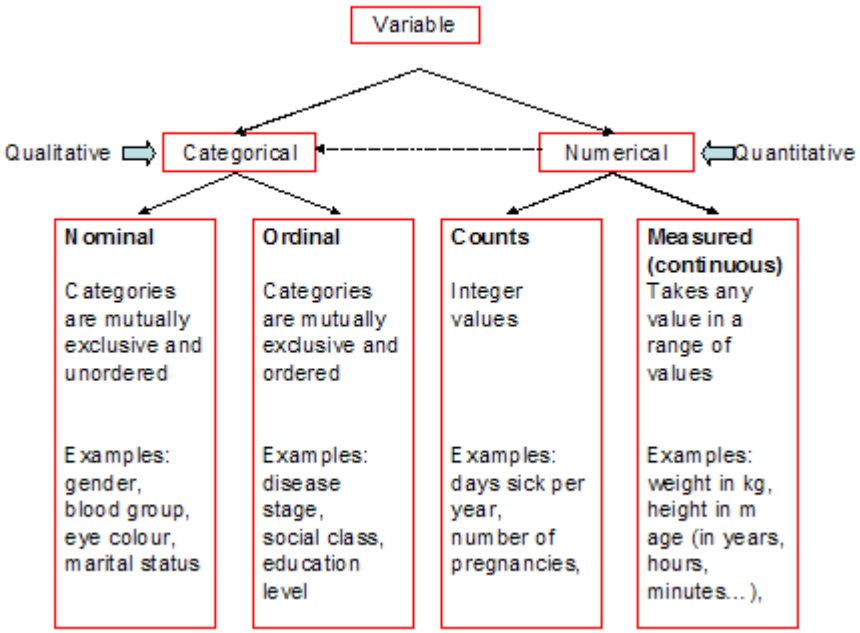
3. The research hypothesis

An investigator conducting a study usually has a theory in mind: for example, patients with diabetes have raised blood pressure, or oral contraceptives may cause breast cancer. This theory is known as the study or research hypothesis. However, it is impossible to prove most hypotheses, one can always think of circumstances which have not yet arisen under which a particular hypothesis may or may not hold. Thus one might hold a theory that all Chinese children have black hair. Unfortunately, having observed 1,000 or even 1,000,000 Chinese children and checked that they all have black hair would not have proved the hypothesis. On the other hand, if only one fair-haired Chinese child is seen, the theory is disproved. Thus there is a simpler logical setting for disproving hypotheses than for proving them. The converse of the study hypothesis is the null hypothesis. Examples are: diabetic patients do not have raised blood pressure, or oral contraceptives do not cause breast cancer. Such a hypothesis is usually phrased in the negative and that is why it is termed null.

4. Describing and summarising data

By the time you start this stage, you have selected your participant sample, identified your research hypothesis, and used the appropriate study design, see The NIHR RDS EM / YH Resource Pack: '*Experimental Design'* to test the research hypothesis. You now have, in your possession, some actual raw data! By data, we generally mean the reduction of the actual results of a study into numerical format. This is the first, preliminary step in the statistical analysis of your data. However, what is particularly crucial at this stage is that you address the issue of what type of data you have. Figure 2 shows a basic summary of data types, although some data do not fit neatly into these categories.

Figure 2: Broad classification of the different types of data with examples



4.1 Categorical or Qualitative Data

Qualitative data is also a term sometimes used (more so in the past) for what we would now call categorical or nominal data.

Nominal Categorical Data

Nominal or categorical data are data that one can name and put into categories. They are not measured but simply counted. They often consist of unordered 'either-or' type observations which have two categories and are often know as binary or dichotomous data. For example: Dead or Alive; Male or Female; Cured or Not Cured; Pregnant or Not Pregnant. However, nominal

categorical data often can have more that two categories, for example: blood group O, A, B, AB; country of origin; ethnic group.

Ordinal Data

If there are more than two categories of classification it may be possible to order them in some way. For example, education may be classified into three categories: none or elementary school, middle school, college and above. Thus someone who has been to middle school has more education than someone from elementary school but less than someone from college. However, without further knowledge it would be wrong to ascribe a numerical quantity to position: one cannot say that someone who had middle school education is twice as educated as someone who had only elementary school education. This type of data is also known as ordered categorical data.

Ranks

In some studies it may be appropriate to assign ranks. For example, patients with rheumatoid arthritis may be asked to order their preference for four dressing aids. Here although numerical values from 1 to 4 may be assigned to each aid one cannot treat them as numerical values. They are in fact only codes for best, second best, third choice and worst. Ordered categories can also be thought of as ranks with a high proportion of tied values, this means that often the same methods can be applied to ordinal data and to ranks.

4.2 Numerical or Quantitative Data

Count or discrete quantitative data

Figure 2 examples includes “number of pregnancies” and this is termed count data. Other examples are often counts per unit of time such as the number of deaths in a hospital per year, or the number of attacks of asthma a person has per month.

Measured or Numerical Continuous

Such data are measurements that can, in theory at least, take any value within a given range. These data contain the most information, and are the ones most commonly used in statistics. Examples of continuous data in Figure 2 are: “age”, “height” and “weight”.

However, for simplicity, it is often the case in medicine that continuous data are dichotomised to make nominal data. Thus diastolic blood pressure, which is continuous, is converted into hypertension (> 90 mmHg) and normotension (≤ 90 mmHg). This clearly leads to a loss of information. There are two main reasons for doing this. It is easier to describe a population by the proportion of people affected (for example, the proportion of people in the population with hypertension is 10%). Further one often has to make a decision: if a person has hypertension, then they will get treatment, and this too is easier if the population is grouped.

Interval and Ratio Scales

One can distinguish between interval and ratio scales. In an interval scale, such as body temperature or calendar dates, a difference between two measurements has meaning, but their ratio does not. Consider measuring temperature (in degrees centigrade) then we cannot say that a temperature of $20^{o}C$ is twice as hot as a temperature of $10^{o}C$. In a ratio scale, such as bodyweight, a 10% increase implies the same weight increase whether expressed in kilograms or pounds. The crucial difference is that in a ratio scale, the value of zero has real meaning, whereas in an interval scale, the position of zero is arbitrary.

One difficulty with giving ranks to ordered categorical data is that one cannot assume that the scale is interval. Thus, as we have indicated when discussing ordinal data, one cannot assume that risk of cancer for an individual educated to middle school level, relative to one educated only to primary school level is the same as the risk for someone educated to college level, relative to someone educated to middle school level. Were we simply to score the three levels of education as 1, 2, 3 in the subsequent analysis, then this would imply in some way the intervals have equal weight.

EXERCISE 1

Classify the following data as numerical (quantitative) or categorical (qualitative):

Temperature
Marital status
Severity of arthritic pain
Blood pressure
Number of visits to their GP per year

4.3 Organising your data: descriptive statistics

Summarising Categorical Data

Binary data are the simplest type of data. Each individual has a label which takes one of two values. A simple summary would be to count the different types of label. However, a raw count is rarely useful. Furness et al (2003), in a study of 567 car crashes, reported more accidents to white cars, 145, than to any other colour car in Auckland, New Zealand over a one-year period. As a consequence, a New Zealander may think twice about buying a white car! However, it turns out that there are simply more white cars on the Auckland roads than any other colour. It is only when this number is expressed as a *proportion* that it becomes useful. When Furness et al (2003) looked at the proportion of white cars that had accidents compared to the proportion of all cars that had accidents, they found the proportions very similar and so white cars are not more dangerous than other colours. Hence the first step to analysing categorical data is to count the number of observations in each category and express them as proportions of the total sample size. Proportions are a special example of a *ratio*. When time is also involved (as in counts per year) then it is known as a *rate*. These distinctions are given below.

***Ratios, proportions, percentages, risk and rates*.**

A *ratio* is simply one number divided by another. If we measure how far a car travels in a given time then the ratio of the distance travelled to the time taken to cover this distance is the *speed.*

Proportions are ratios of counts where the numerator (the top number) is a subset of the denominator (the bottom number). Thus in a study of 50 patients, 20 are depressed, so the proportion is 20/50 or 0.4. It is usually easier to express this as a percentage, so we multiply the proportion by 100, and state that 40% of the patients are depressed. A proportion is known a *risk* if the numerator counts events which happen prospectively. Hence if 300 students start medical school and 15 drop out before their final examinations, the *risk* of dropping out is 15/300 = 0.05 or 5%

Rates always have a time period attached. If 600,000 people in the UK die in one year, out of a population of 60,000,000, the death *rate* is 600,000/60,0000,000 or 0.01 deaths per person per year. This is known as the *crude death rate* (crude because it makes no allowance for important factors such as age). Crude death rates are often expressed as deaths per thousand per year, so the crude death rate is 10 deaths per thousand per year, since it is much easier to imagine 1000 people, of whom 10 die, than it is 0.01 deaths per person!

Illustrative Example – Special Care Baby Unit

Simpson (2004) describes a prospective study, in which 98 preterm infants were given a series of tests shortly after they were born, in an attempt to predict their outcome after one year. We will use this example in this section and in the next section where we discuss quantitative data. One categorical variable recorded was the type of delivery in five categories as displayed in Table 2. The first column shows category names, whilst the second shows the number of individuals in each category together with its percentage contribution to the total.

Table 2: Type of delivery for 98 babies admitted to a special care baby (Simpson, 2004)

Type of delivery	Frequency	Percent
Standard vaginal delivery	38	39%
Assisted vaginal delivery	10	10%
Elective Caesarean Section	8	8%
Emergency Caesarean Section	13	13%
Emergency Caesarean section/not in labour	29	30%
Total	**98**	**100%**

In addition to tabulating each variable separately, we might be interested in whether the type of delivery is related to the gender of the baby. Table 3 shows the distribution of type of delivery by gender; in this case it can be said that delivery type has been *cross-tabulated* with gender. Table 3 is an example of a *contingency* table with 5 rows (representing type of delivery) and 2 columns (gender). Note that we are interested in the distribution of modes of delivery within gender, and so the percentages add to 100 down each column, rather than across the rows.

Table 3: Type of delivery and gender of 98 babies admitted to a special care baby unit (Simpson, 2004)

Type of Delivery	Gender	
	Male	**Female**
	n (%)	*n* (%)
Standard vaginal delivery	15 (33)	23 (43)
Assisted vaginal delivery	4 (9)	6 (11)
Elective Caesarean section	4 (9)	4 (8)
Emergency Caesarean section	6 (13)	7 (13)
Emergency Caesarean section/not in labour	16 (36)	13
Total	**45 (100)**	**53 (100)**

Comparing outcomes for binary data

Many studies involve a comparison of two groups. We may wish to combine simple summary measures to give a summary measure which in some way shows how the groups differ. Given two proportions one can either subtract one from the other, or divide one by the other.

Suppose the results of a clinical trial, with a binary categorical outcome (positive or negative), to compare two treatments (a new test treatment versus a control) are summarised in a 2 by 2 contingency table as in Table 4. Then the results of this trial can be summarised in a number of ways.

Table 4: Example of 2 by 2 contingency table with a binary outcome and two groups of subjects

	Treatment Group	
Outcome	**Test**	**Control**
Positive	*a*	*.b*
Negative	*c*	*.d*
	a + c	*b + d*

The ways of summarizing the data presented in Table 4 are given below.

Summarising comparative binary data - Differences in Proportions, and Relative Risk

From Table 4, the proportion of subjects with a positive outcome under the under the active or test, treatment is $p_{Test} = \frac{a}{a+c}$ and under the control treatment is

$$p_{Control} = \frac{b}{b+d}.$$

The difference in proportions is given by

$$d_{prop} = p_{Test} - p_{Control}.$$

In prospective studies the proportion is also known as a risk. When one ignores the sign, the above quantity is also known as the *absolute risk difference* (*ARD*), that is.

$$ARD = \left| p_{Control} - p_{Test} \right|,$$

where the symbols |.| mean to take the absolute value.

If we anticipate that the treatment to reduce some bad outcome (such as deaths) then it may be known as the *absolute risk reduction* (ARR).

The risk ratio, or relative risk (*RR*), is

$$RR = p_{Test} / p_{Control}.$$

A further summary measure, used only in clinical trials is the *number needed to treat (NNT)/harm*. This is defined as the inverse of the ARD. NNTs are discussed in Campbell et al (2007) or Altman et al (2000).

Each of the above measures summarises the study outcomes, and the one chosen may depend on how the test treatment behaves relative to the control. Commonly, one may chose an absolute risk difference for a clinical trial and a relative risk for a prospective study. In general the relative risk is independent of how common the risk factor is. Smoking increases ones risk of lung cancer by a factor of 10, and this is true in countries with a high smoking prevalence and countries with a low smoking prevalence. However, in a clinical trial, we may be interested in what reduction in the proportion of people with poor outcome a new treatment will make.

Summarising binary data – Odds and Odds Ratios

A further method of summarising the results is to use the odds of an event rather than the probability. The odds of an event are defined as the ratio of the probability of occurrence of the event to the probability of non-occurrence, that is**,** $p/(1 - p)$.

The odds ratio (OR) is the ratio of odds for test group to the odds for control group

$$\frac{p_{Test} \big/ (1 - p_{Test})}{p_{Control} \big/ (1 - p_{Control})}.$$

Using the notation of Table 4 we can see that the odds of an outcome for the test group to the odds of an outcome for control group is:

$$OR_{Test/Control} = \frac{a}{c} \Big/ \frac{b}{d} = \frac{ad}{bc}.$$

When the probability of an event happening is rare, the odds and probabilities are close, because then *a* is much smaller than *c* and so *a/(a+c)* is approximately *a/c* and *b* is much smaller than *d* and so *b/(b+d)* is approximately *b/d.* Thus the *OR* approximates the *RR* when the successes are rare (say with a maximum incidence less than 10% of either p_{Test} or p_{Control} Sometimes the odds ratio is referred to as 'the approximate relative risk'.

EXERCISE 2

Ninety-nine pregnant women, with dystocia (difficult childbirth or labour), were allocated at random to receive immersion in water in a birth pool (Intervention group: Labour in water 49 women) or standard augmentation for dystocia (control group: Augmentation 50 women) in a randomised-controlled trial to evaluate the impact of labouring in water during the first stage of labour (Cluett et al 2004). The main outcome was use of epidural analgesia at any stage of labour. The results are shown in Table E2 below.

Table E2: Epidural analgesia data from a randomised controlled trial of labouring in water compared with standard augmentation for management of dystocia in first stage of labour (Cluett et al, 2004).

	Intervention	**Control**
Epidural Analgesia at any stage of labour	**(Labour in water)**	**(Augmentation)**
Yes	23	33
No	26	17
Total	49	50

i) What is the proportion of women who had an epidural in each of the two groups?

ii) What is the relative risk of the use of an epidural for the labour in water women compared with the augmentation women?

iii) Calculate the Odds ratio of epidural for the Labour in water women compared with Augmentation women. Compare this estimated OR with the RR estimate from part ii: what do you notice?

iv) Find the absolute risk difference for the use of an epidural for labour in water compared to augmentation.

Summarising continuous data

A quantitative measurement contains more information than a categorical one, and so summarizing these data is more complex. One chooses summary statistics to condense a large amount of information into a few intelligible numbers, the sort that could be communicated verbally. The two most important pieces of information about a quantitative measurement are 'where is it?' and 'how variable is it?' These are categorised as measures of location (or sometimes 'central tendency') and measures of spread or variability.

Measures of Location

Mean or Average

The arithmetic mean or average of *n* observations $\bar{x}$ (pronounced *x* bar) is simply the sum of the observations divided by their number, thus:

$$\bar{x} = \frac{\text{Sum of all sample values}}{\text{Size of sample}} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

In the above equation, x_i represents the individual sample values and $\sum_{i=1}^{n} x_i$ their sum. The Greek letter 'Σ' (sigma) is the Greek capital 'S' and stands for 'sum' and simply means "add up the *n* observations x_i from the 1st to the last (n^{th})".

Example – Calculation of the mean – Birth weights

Consider the following five birth weights in kilograms recorded to 1 decimal place selected randomly from the Simpson (2004) study of low birth weight babies.

1.2, 1.3, 1.4, 1.5, 2.1

The sum of these observations is (1.2 + 1.3 + 1.4 + 1.5+ 2.1) = 7.5.

Thus the mean $\bar{x}$ = 7.5/5 = 1.50 kg. It is usual to quote 1 more decimal place for the mean than the data recorded.

The major advantage of the mean is that it uses all the data values and is, in a statistical sense, efficient. The mean also characterises some important statistical distributions to be discussed later on. The main disadvantage of the mean is that it is vulnerable to what are known as outliers. Outliers are single observations which, if excluded from the calculations, have noticeable influence on the results. For example if we had entered '21' instead of '2.1' in the calculation of the mean, we would find the mean changed from 1.50 kg to 7.98 kg. It does not necessarily follow, however, that outliers should be excluded from the final data summary, or that they result from an erroneous measurement.

Median

The median is estimated by first ordering the data from smallest to largest, and then counting upwards for half the observations. The estimate of the median is either the observation at the centre of the ordering in the case of an odd number of observations, or the simple average of the middle two observations if the total number of observations is even.

Example – Calculation of the median – Birth weights

Consider the following 5 birth weights in kilograms selected randomly from the Simpson (2004) study.

Rank order	Weight (kg)	
1	1.2	
2	1.3	
3	1.4	← **median**
4	1.5	
5	2.1	

If we had observed an additional value of 3.5 kg in the birth weight the median would be the average of the 3rd and the 4th observation in the ranking, namely the average of 1.4 and 1.5, which is 1.45kg.

The median has the advantage that it is not affected by outliers, so for example the median in the data would be unaffected by replacing '2.1' with '21'. However, it is not statistically efficient, as it does not make use of all the individual data values.

Mode

A third measure of location is termed the mode. This is the value that occurs most frequently or if the data are grouped, the grouping with the highest frequency. It is not used much in statistical analysis, since its value depends on the accuracy with which the data are measured, although it may be useful for categorical data to describe the most frequent category. However, the expression 'bimodal' distribution is used to describe a distribution with two peaks in it. This can be caused by mixing two or more populations together. For example height might appear to have a bimodal distribution if one had men and women in the population. Some illnesses may raise a biochemical measure, so in a population containing healthy individuals and those who are ill one might expect a bimodal distribution. However, some illnesses are *defined* by the measure of, say obesity or high blood pressure, and in these cases the distributions are usually unimodal with those above a given value regarded as *ill*.

Measures of Dispersion or Variability

Range and Interquartile Range

The range is given as the smallest and largest observations. This is the simplest measure of variability. For some data it is very useful, because one would want to know these numbers. For example in a sample, the age of the youngest and oldest participant. However, if outliers are present it may give a distorted impression of the variability of the data, since only two of the data points are included in making the estimate.

Quartiles

The quartiles, namely the lower quartile, the median and the upper quartile, divide the data into four equal parts ie there will be approximately equal numbers of observations in the four sections (and exactly equal if the sample size is divisible by four and the measures are all distinct). The quartiles are calculated in a similar way to the median; first order the data and then count the appropriate number from the bottom. The interquartile range is a useful measure of variability and is given by the difference of the lower and upper quartiles. The interquartile range is not vulnerable to outliers, and whatever the distribution of the data, we know that 50% of them lie within the interquartile range.

Illustrative example - Calculation of the range, quartiles and inter-quartile range

Suppose we had 10 birth weights arranged in increasing order from the Simpson (2004) study.

Order	Birth weight (kg)
1	1.51
2	1.55
3	1.79
4	2.10
5	2.18
6	2.22
7	2.37
8	2.40
9	2.81
10	2.85

Lower quartile (25th percentile)

Median (50th percentile)

Upper quartile (75th percentile)

Inter quartile range

The range of birth weights in these data is from 1.51 kg to 2.85 kg (simply the smallest and largest birth weights). The median is the average of the 5th and 6th observations (2.18 + 2.22)/2 = 2.20kg. The first half of the data has 5 observations so the first quartile is the 3rd ranked observation, namely 1.79kg, and similarly the third quartile would be the 8th ranked observation, namely 2.40 kg. So the interquartile range is from 1.79 to 2.40 kg.

Standard Deviation and Variance

The standard deviation (*SD or s*) is calculated as follows:

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

The expression $\sum_{i=1}^{n}(x_i - \bar{x})^2$ may look complicated, but it is easier to understand when thought of in stages. From each *x* value subtract the mean $\bar{x}$, square this difference, then add each of the *n* squared differences. This sum is then divided by (*n* - 1). This expression is known as the *variance*. The

variance is expressed in square units, so we take the square root to return to the original units, which gives the standard deviation, *s*. Examining this expression it can be seen that if all the *x*'s were the same, then they would equal $\bar{x}$ and so *s* would be zero. If the *x*'s were widely scattered about $\bar{x}$, then *s* would be large. In this way *s* reflects the variability in the data. The standard deviation is vulnerable to outliers, so if the 2.1 was replaced by 21 we would get a very different result.

Illustrative example - Calculation of the standard deviation

Consider the five birth weights (in kg): 1.2, 1.3, 1.4, 1.5, 2.1. The calculations to work out the standard deviation are given in the following table.

	Weight (kg)	**Mean Weight (kg)**	**Differences from Mean**	**Square of differences from mean**
Subject	x_i	$\bar{x}$	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	1.2	1.5	-0.30	0.09
2	1.3	1.5	-0.20	0.04
3	1.4	1.5	-0.10	0.01
4	1.5	1.5	0.00	0.00
5	2.1	1.5	0.60	0.36
Totals (Sum)	**7.5**		**0**	**0.50 kg^2**
Mean	**1.50 kg**		**Variance**	**0.13 kg^2**
n	**5**		**Standard**	
n **- 1**	**4**		**Deviation**	**0.35 kg**

Variance = 0.50/4

SD = square root of the Variance

We first find the mean to be 1.5 kg, then subtract this from each of the five observations to get the 'Differences from the Mean'. Note the sum of this column is zero. This will always be the case: the positive deviations from the mean cancel the negative ones.

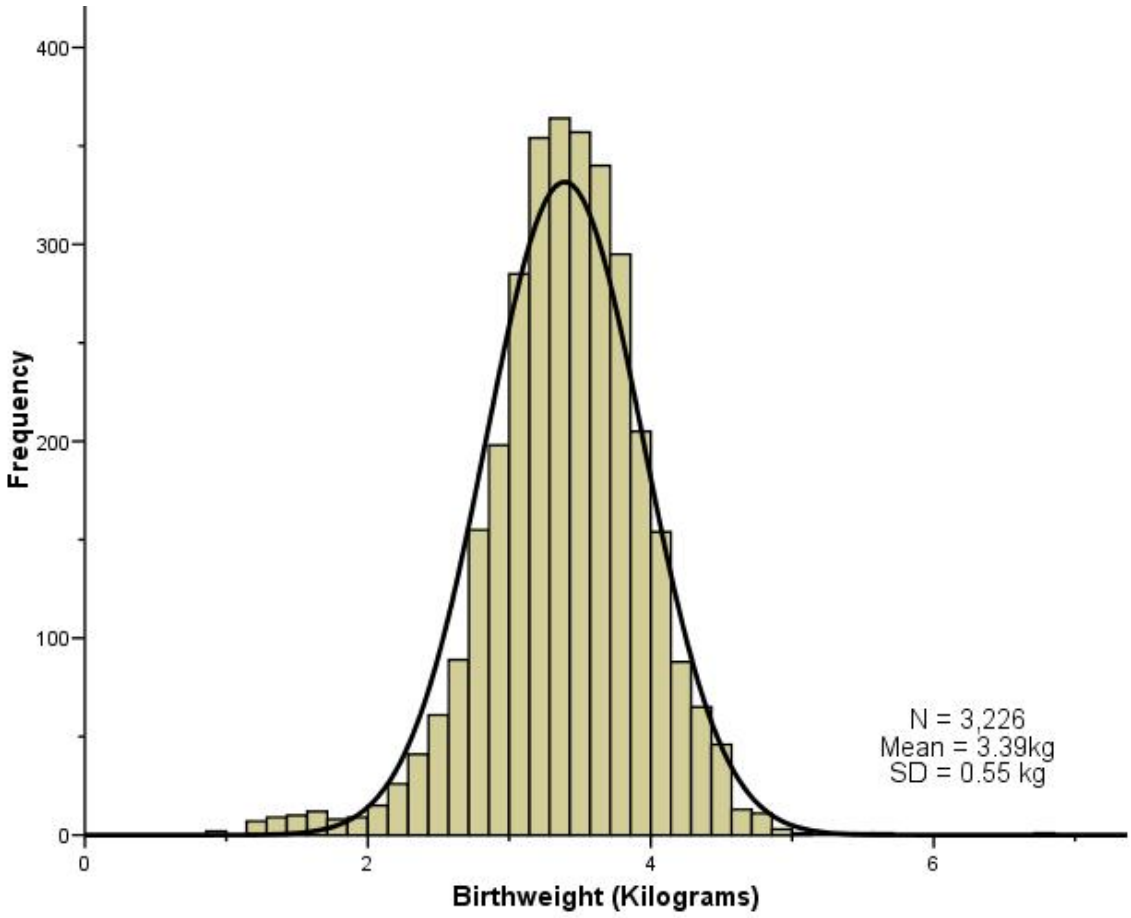
A convenient method of removing the negative signs is by squaring the deviations, which is given in the next column, which is then summed to get 0.50 kg^2. Note that the bulk of this sum (72%) is contributed by one observation, the value 2.1 from subject 5, which is the observation furthest from the mean. This illustrates that much of the value of an *SD* is derived from the outlying observations. We now need to find the average squared deviation. Common sense would suggest dividing by *n*, but it turns out that this actually gives an estimate of the population variance which is too small. This is because we use the estimated mean $\bar{x}$ in the calculation in place of the true population mean. In fact we seldom know the population mean so there is

little choice but for us to use its estimated value, $\bar{x}$, in the calculation. The consequence is that it is then better to divide by what are known as the *degrees of freedom,* which in this case is *n* – 1, to obtain the *SD*.

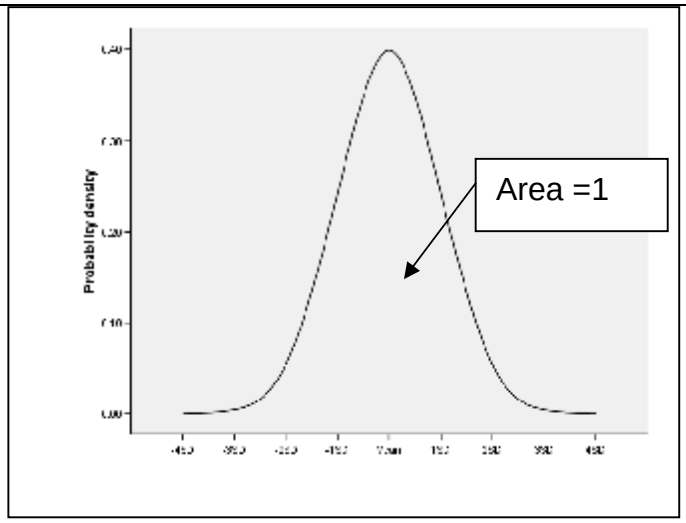
Normal distribution

It is often the case with medical data that the histogram of a continuous variable obtained from a single measurement on different subjects will have a symmetric 'bell-shaped' distribution. One such example is the histogram of the birth weight (in kilograms) of the 3,226 newborn babies shown in Figure 3. This symmetric 'bell-shaped' distribution mentioned above is known as the Normal distribution and is one of the most important distributions in statistics.

Figure 3 Distribution of birth weight in 3,226 newborn babies with superimposed population Normal distribution curve (data from O'Cathain et al 2002)



The histogram of the sample data is an estimate of the population distribution of birth weights in newborn babies. This population distribution can be estimated by the superimposed smooth 'bell-shaped' curve or 'Normal' distribution shown. We presume that if we were able to look at the entire population of newborn babies then the distribution of birth weight would have exactly the Normal shape. The Normal distribution has the following properties.

Properties of the Normal distribution	Figure 4 The Normal probability distribution
Total area under the curve = 1 (or 100%). Bell shaped and symmetrical about its mean. The peak of the curve lies above the mean. Any position along the horizontal axis can be expressed as a number of SDs away from the mean. The mean and median coincide.	

The Normal distribution, Figure 4, is completely described by two parameters: one, μ, represents the population mean or centre of the distribution and the other, σ, the population standard deviation. There are infinitely many Normal distributions depending on the values of m and s. The Standard Normal distribution has a mean of zero and a variance (and standard deviation) of one and a shape as shown in Figure 4, and you can convert any other Normal distributions to this standard form.

EXERCISE 3

Table E3: Systolic blood pressure levels (mmHg) in 16 middle aged men before and after a standard exercise (from Altman *et al* 2000).

Subject	***Systolic blood pressure (mmHg)***		***DIFFERENCE***
Number	***BEFORE***	***AFTER***	***After-Before***
1	148	152	4
2	142	152	10
3	136	134	-2
4	134	148	14
5	138	144	6
6	140	136	-4
7	132	144	12
8	144	150	6
9	128	146	18
10	170	174	4
11	162	162	0
12	150	162	12
13	138	146	8
14	154	156	2
15	126	132	6
16	116	126	10

i) Using Table E3, calculate the following measures of location for the blood pressure BEFORE exercise:
(a) mode
(b) median
(c) mean

ii) Using Table E3, calculate the following measures of spread for the blood pressure BEFORE exercise data:
(a) range
(b) interquartile range
(c) standard deviation

5. Statistical Analysis

The following section introduces the concept of hypothesis testing and describes some basic methods for testing hypotheses. In order to provide continuity, data from the same study will be used to illustrate key concepts. The study, briefly described in Box 1, is a randomised controlled trial of standard care versus specialist community leg ulcer clinics for treatment of venous leg ulcers. Patients were assessed at entry to the study, after 3 months and 1 year and outcome measures included ulcer healing rates, health-related quality of life, satisfaction with the service and treatment cost.

It is rarely possible to obtain information on an entire population and usually data are collected on a sample of individuals from the population of interest. The main aim of statistical analysis is to use the information from the sample to draw conclusions (make inferences) about the population of interest. For example, the leg ulcer trial (Box 1) was conducted as an randomized controlled trial as it was not possible to study all individuals with venous leg ulcers and so instead a sample of individuals with venous leg ulcers in the Trent region was studied in order to estimate the potential cost effectiveness of specialist clinics compared to standard care. The two main approaches to statistical analysis, hypothesis testing and estimation, are outlined in the following sections.

Box 1 Case study of a randomised controlled trial

Morrell CJ *et al* (1998) Cost effectiveness of community leg ulcer clinics: randomised controlled trial. *British Medical Journal* 316: 1487-1491.

This study comprised a randomised controlled trial looking at the effectiveness of community leg ulcer clinics compared with usual care. Two hundred and thirty-three patients with venous leg ulcers were randomly allocated to either usual care at home by district nursing team (control group, n=113)) or weekly treatment with four layer bandaging in a specialist leg ulcer clinic (intervention group, n=120) and followed-up for a year. Outcomes included time to complete ulcer healing, ulcer-free weeks, patient health status, recurrence of ulcers, satisfaction with care and use of services.

At the end of 12 months the mean time (in weeks) that each patient was free from ulcers during follow up was 20.1 and 14.2 in the clinic and control groups, respectively. On average, patients in the clinic group had 5.9 more ulcer-free weeks (95% confidence interval 1.2 to 10.6 weeks; P=0.014) than the control patients.

5.1 Hypothesis Testing (using P values)

Before examining the different techniques available for analysing data, it is first essential to understand the process of hypothesis testing and its key principles, such as what a P value is and what is meant by the phrase 'statistical significance'. Figure 5 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 (H_0) assumes that there is no difference in the outcome of interest between the study groups. The study or alternative hypothesis (H_A) states that there is a difference between the study groups. In general, the direction of the difference (for example: that treatment A is better than treatment B) is not specified. For the leg ulcer trial, the research question of interest was:

For patients with leg ulcers does specialist treatment at a leg ulcer clinic affect the number of ulcer-free weeks over the 12 month follow-up compared to district nursing care at home?

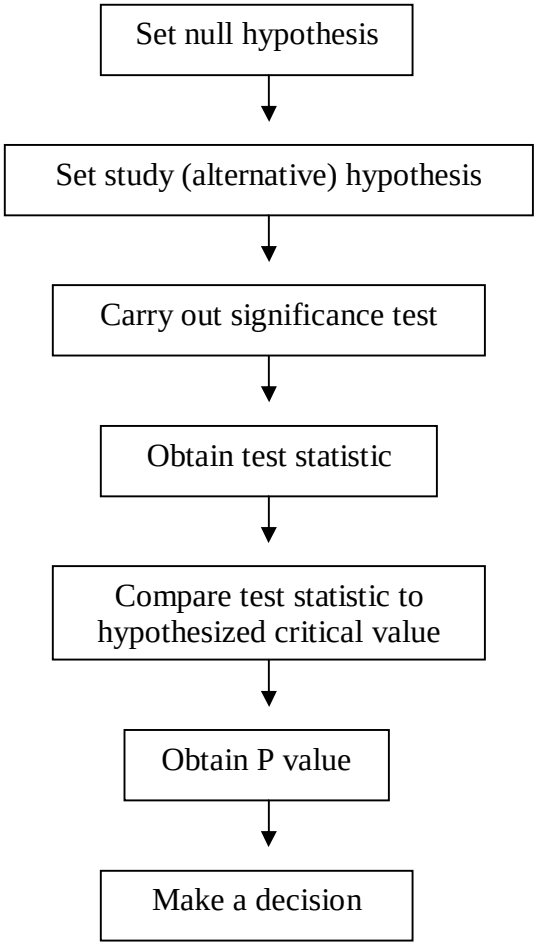
The null hypothesis, **H_0**, is:

There is no difference in ulcer-free weeks between the control (home) and intervention (clinic) group.

and the alternative hypothesis, **H_A**, is:

There is a difference in ulcer-free weeks between the control and intervention groups.

Figure 5: Hypothesis testing: the main steps



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 used to obtain a P value. For the comparison above, patients in the clinic group had, on average, 5.9 more ulcer-free weeks after 12 months than the control (home) group and the P value associated with this difference was 0.014. 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 study data).

Box 2 Statistical significance

We say that our results are statistically significant if the P value is less than the significance level (α) set at 5% or 0.05.

	P≤ 0.05	**P> 0.05**
Result is	Statistically significant	Not statistically significant
Decide	That there 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.

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 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 2).

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.

Though 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 Table 5. Whatever is decided, this decision may correctly reflect what is true in the population: the null hypothesis is rejected, when it is in 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 is rejected, when it is in fact true (false positive or Type I error, α); or the null hypothesis is not rejected, when in fact it is false (false negative, Type II error, β).

Table 5 Making a decision

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

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. It is usually expressed in percentages, so for a study which has 80% power, there is a likelihood of 80% of being able to detect a difference, of a given size, if there genuinely is a difference in the population.

5.2 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 effect size. Much more information, such as whether the result is likely to be of clinical importance can be gained by calculating a confidence interval. A confidence interval may be calculated for any estimated quantity (from the sample data), such as the mean, median, proportion, or even a difference, for example the mean difference in ulcer-free weeks between two groups. It is a measure of the precision (accuracy) with which the quantity of interest is estimated (in this case the mean difference in the number of ulcer-free weeks). Full details of how to calculate confidence intervals are given in Altman et al (2000).

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 leg ulcer example above, the quantity estimated was the mean difference in the number of ulcer-free weeks between the groups, 5.9 weeks. The 95% confidence interval for this difference was 1.2 to 10.5 weeks. Thus, whilst the best available estimate of the mean difference was 5.9 weeks, it could be as low as 1.2 weeks or as high as 10.5 weeks, with 95% certainty. The P value associated with this difference was 0.014 and in the previous section it was concluded that this difference was statistically significant at the 5% level. Whilst the P value will give an indication of whether the result obtained is statistically significant it gives no other information. The confidence interval is more informative as it gives a range of plausible values for the estimated quantity. Provided this range does not include the value for no difference (in this case 0) it can be concluded that there is a difference between the groups being compared.

EXERCISE 4

Two hundred and thirty-three patients with venous leg ulcers were allocated at random to intervention (120 patients) or control (113 patients) groups in a randomised-controlled trial to establish the effectiveness of community leg ulcer clinics that use four layer compression bandaging versus usual care provided by district nurses. (Morrell et al 1998, BMJ, 316, 1487-1491). The results section of the paper says:

"*The cumulative percentages of leg ulcers healed at 12 weeks were 34% in the intervention group and 24% in the control group (difference 10%, 95% Confidence Interval –2% to 22%, p = 0.11).*"

i) What is meant by "p = 0.11"? Do these data suggest that the intervention alters the healing rates of leg ulcers at 12 weeks? Comment on the results of this hypothesis test.

ii) What is meant by a 95% confidence interval? Discuss whether the confidence interval for the difference in healing rates between the intervention and control groups suggests that patients in the intervention group might have a higher rate of ulcer healing at 12 weeks follow-up than patients in the control group.

6. Dealing with data: how to choose the right statistical test

In this section we will now be putting some of the theory into practice and looking at some of the more basic statistical tests that you will come across in the literature and in your own research, the choice of method of analysis for a problem depends on the comparison to be made and the data to be used. There are often several different approaches to even a simple problem. The methods described here and recommended for particular types of question may not be the only methods, and may not be universally agreed as the best method. However, these would usually be considered as valid and satisfactory methods for the purposes for which they are suggested here.

6.1 The use of computers in statistical analysis

With today's computer technology there is virtually no need to perform statistical analyses by hand. Although all the tests described below can be performed manually, the dramatic rise in the availability of computer statistical packages over the past 15 years (eg Stata; SPSS for Windows; Minitab etc.) has meant that the majority of statistical analysis can now be performed by computer. It is often much simpler to let a computer programme do the work and many of the current packages have the added bonuses of being easy to use, versatile and widely available. For this reason, it is assumed in this section that you have access to a computer package that will allow you to store, process and analyse data. (Warning: be aware that computers will analyse anything! It is up to the researcher to check that data and results are sensible). For this reason, the reader, whether novice or more experienced researcher, is strongly recommended to refer to The NIHR RDS EM / YH Resource Pack: '*Using SPSS'*. Full details of the mathematics behind the various tests and how to perform them manually, using a calculator, are given in Altman (1991) and Campbell et al (2007).

6.2 What type of statistical test?

Five key questions to ask:

1. What are the aims and objectives of the study?
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?

Given the answers to these five key questions, 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 categorical 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.

6.3 Choosing the statistical method

The choice of method of analysis for a problem depends on the comparison to be made and the data to be used. This section outlines the methods appropriate for three common problems in statistical inference as outlined below:

1. Comparison of two independent groups, e.g. groups of patients given different treatments.
2. Comparison of the response of one group under different conditions as in a cross-over trial or of matched pairs of subjects.
3. Investigation of the relationship between two variables measured on the same sample of subjects.

Before beginning any analysis it is important to examine the data, using the techniques described previously; adequate description of the data should precede and complement the formal statistical analysis. For most studies and for randomised controlled trials in particular, it is good practice to produce a table that describes the initial or baseline characteristics of the sample.

6.4 Different approaches

There are often several different approaches to even a simple problem. The methods described (Figures 6 to 8) here and recommended for particular types of question may not be the only methods, and may not universally agreed as the best method. Statisticians are at least as prone to disagree as clinicians! However, these would usually be considered as valid and satisfactory methods for the purposes for which they are suggested here.

Example - Choosing the right statistical test - comparing ulcer-free weeks from the Leg ulcer study

The research question in this example is asking is there a difference in ulcer-free weeks between the Intervention & Control Groups? So in this case we are comparing two independent groups, e.g. two groups of patients given different treatments, so we can use the flow diagram of Figure 6.

Ulcer-free weeks is continuous data, and if the data is symmetric or Normally distributed then the best summary measure or the data is the sample mean and the best comparative summary measure is the mean difference in ulcer-free weeks between the two groups. In these circumstances, the flow diagram of Figure 6 suggests that the most appropriate hypothesis test is the two independent samples *t*-

test. Alternatively, if the data are not Normally distributed we can use the Mann-Whitney U test.

Figure 6: Methods for comparing two independent groups or samples

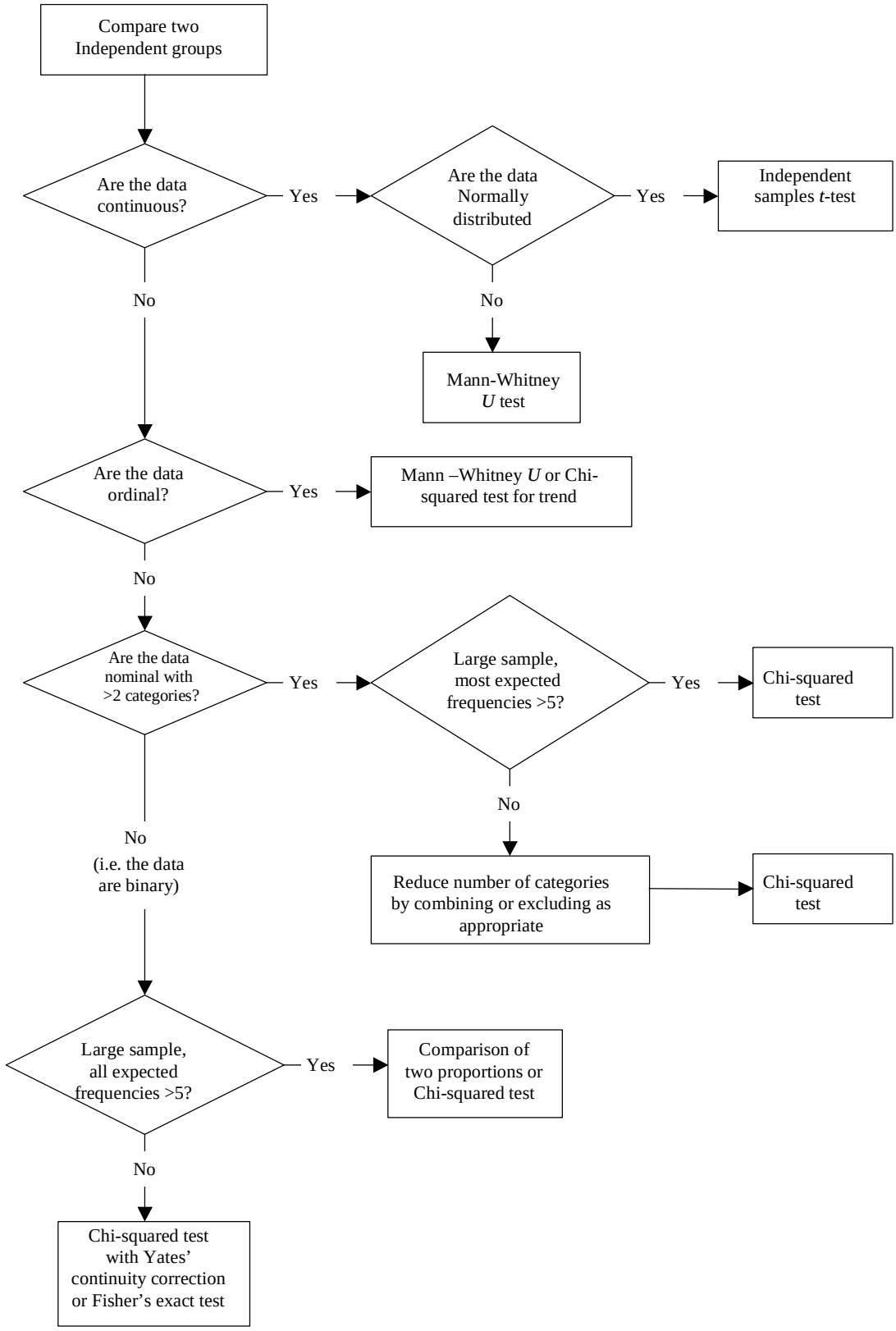


Figure 7: Methods for differences or paired samples

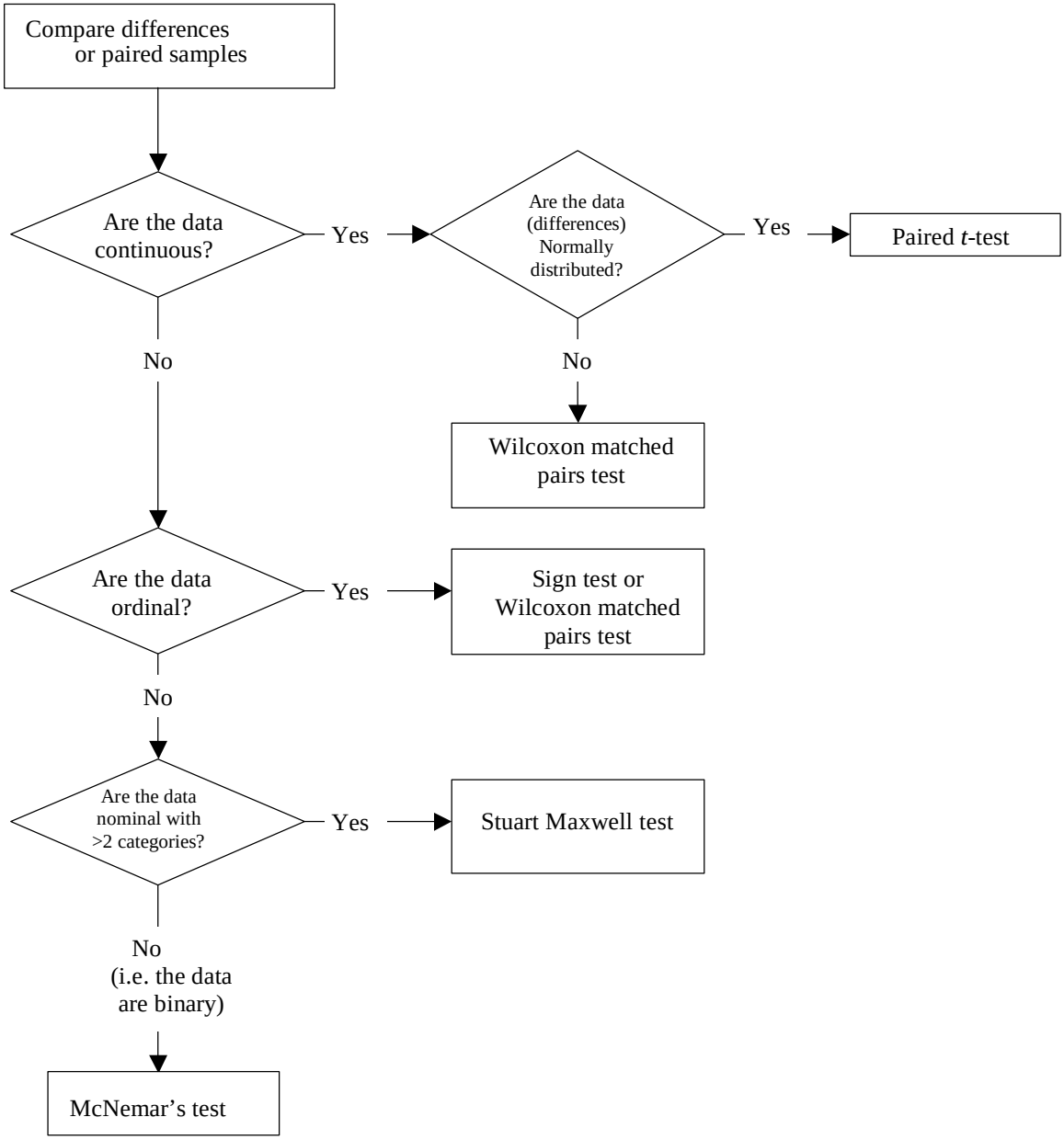


Figure 8: Methods for relationships between two variables

	Continuous, Normal	***Continuous, non-Normal***	***Ordinal***	***Nominal***	***Binary***
Continuous	Regression Correlation: (Pearson's *r*)	Regression Rank Correlation: (Spearman's r_s)	Rank Correlation: (Spearman's r_s)	One-way Analysis of Variance	Independent samples *t*-test
Continuous, non-Normal		Regression Rank Correlation: (Spearman's r_s)	Rank Correlation: (Spearman's r_s)	Kruskall-Wallis test	Mann-Whitney *U* test
Ordinal			Rank Correlation (Spearman's r_s)	Kruskall-Wallis test	Mann-Whitney *U* test Chi-squared test for trend
Nominal				Chi-squared test	Chi-squared test
Binary					Chi-squared test Fisher's exact test

EXERCISE 5

1. Table E2 in Exercise 2, shows the epidural analgesia data from a randomised controlled trial of labouring in water (Intervention) compared with standard augmentation (Control) for management of dystocia in first stage of labour (Cluett et al, 2004). How would you compare the epidural analgesia rates (the proportions of women in each group having epidural analgesia during labour) between the Intervention and Control group women? Using Figures 6 to 9, stating any assumptions that you make, identify the most appropriate statistical test for comparing epidural analgesia rates between the two groups.

2. Table E3 in Exercise 3 shows the blood pressure (in mmHg) of 16 middle-aged men before and after a standard exercise programme and the difference. How would you compare the blood pressure difference (before – after) exercise? Using Figures 6 to 9, stating any assumptions that you make, identify the most appropriate statistical test for comparing blood pressure before and after exercise.

7. Interpreting data

As mentioned previously, it is also important to distinguish between statistical significance and clinical significance, which refers to whether a result is of any clinical or practical importance. Statistical significance does not necessarily mean the result obtained is clinically significant or of any practical importance. You need to remember that a statistical test can only tell you the probability of observing the data or more extreme data if the null hypothesis of no difference is true. A P value will only indicate how likely the results obtained are when the null hypothesis is true. It does not tell you how large the difference is, nor whether the difference is great enough to be of clinical importance. One of our main goals as researchers should be to focus on the practical implications of our work as well as its statistical significance.

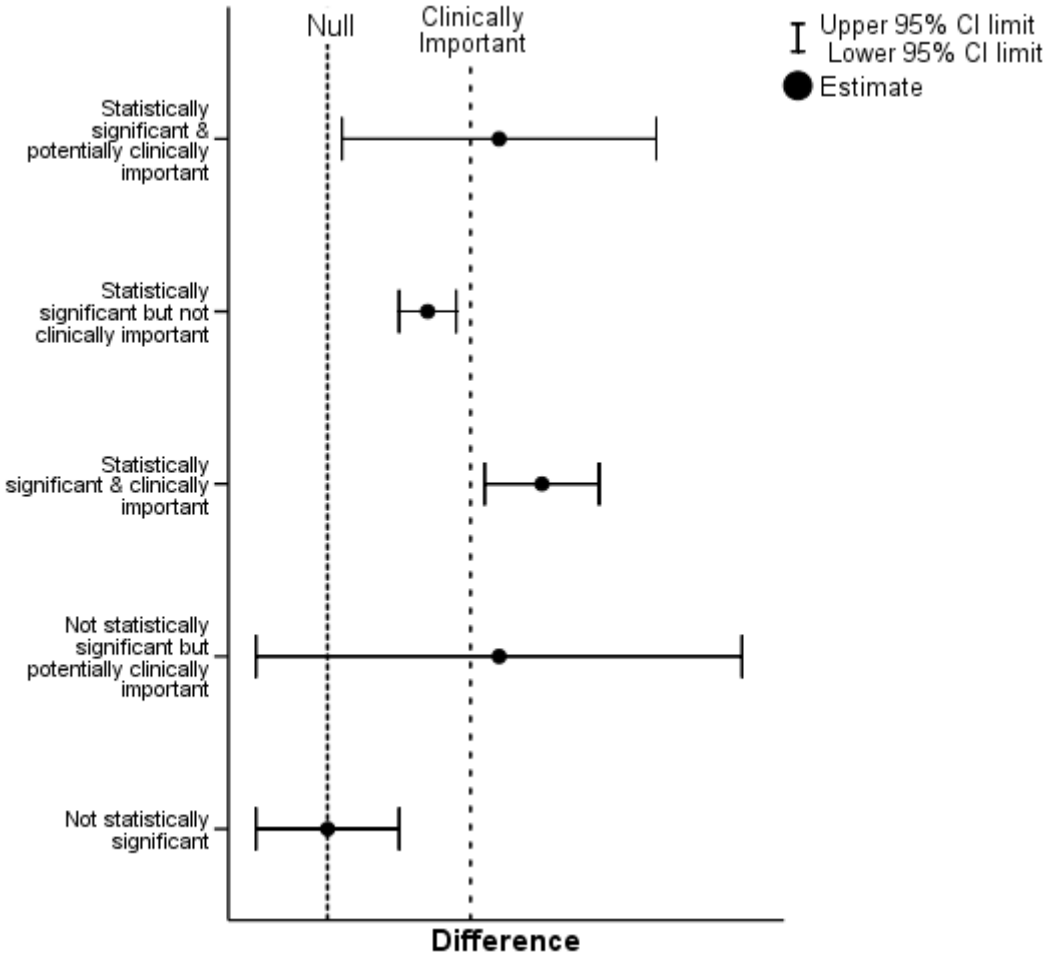
7.1 Confidence Intervals rather than P Values

All that we know from a hypothesis test is, for example, that there is a difference in ulcer-free weeks between the Intervention and Control group in the leg ulcer study. It does not tell us what the difference is or how large the difference is. To answer this we need to supplement the hypothesis test with a confidence interval which will give us a range of values in which we are confident the true population mean difference will lie.

Simple statements in a study report such as '$P < 0.05$' or 'P = NS' do not describe the results of a study well, and create an artificial dichotomy between significant and non-significant results. Statistical significance does not necessarily mean the result is clinically significant. The *P* value does not relate to the clinical importance of a finding, as it depends to a large extent on the size of the study. Thus a large study may find small, unimportant, differences that are highly significant and a small study may fail to find important differences.

Supplementing the hypothesis test with a confidence interval will indicate the magnitude of the result and this will aid the investigators to decide whether the difference is of interest clinically (see Figure 9). The confidence interval gives an estimate of the precision with which a statistic estimates a population value, which is useful information for the reader. This does not mean that one should not carry out statistical tests and quote *P* values, rather that these results should supplement an estimate of an effect and a confidence interval. Many medical journals now require papers to contain confidence intervals where appropriate and not just P values.

Figure 9 Use of confidence intervals to help distinguish statistical significance from clinical importance



Example - ***Clinical Importance*** - ***ulcer-free weeks from the Leg ulcer study***

At the end of 12 months the mean time (in weeks) that each patient was free from ulcers during follow up was 20.1 and 14.2 in the clinic and control groups, respectively. On average, patients in the clinic group had 5.9 more ulcer-free weeks (95% confidence interval 1.2 to 10.6 weeks; P=0.014) than the control patients.

If we regard the Difference in Figure 9 as clinically important if a mean difference in ulcer-free weeks of 12 or more is observed, then the above result is not *clinically important* although it is *statistically significant*. Hence this situation corresponds to the second confidence interval down in Figure 9.

7.2 Relationship between confidence intervals and statistical significance

Different though hypothesis testing and confidence intervals may appear, there is in fact a close relationship between them. If the 95% CI does not include zero (or, more generally the value specified in the null hypothesis) then a hypothesis test will return a statistically significant result. If the 95% CI does include zero then the hypothesis test will return a non-significant result. The confidence interval shows the magnitude of the difference and the uncertainty or lack of precision in the estimate of interest. Thus the confidence interval conveys more useful information than a P values. For example, whether a clinician will use a new treatment that reduces blood pressure or not will depend on the amount of that reduction and how consistent the effect is across patients. So, the presentation of both the P value and the confidence interval is desirable - but if only one is to be presented the P value would be omitted. Presenting a 95% CI indicates whether the result is statistically significant at the 5% level.

Conclusion

The aim of this pack has been to provide you, the researcher, with an overview of the different types of statistical tests used in quantitative data analysis and their application to research. As stated in the introduction to this pack, it has never been the aim to make it an all-inclusive guide to statistical testing but rather an introduction for you to expand upon as you become more experienced in research design and methodology. This pack should also be used in conjunction with the other packs in this series which complement the material presented here.

Answers to exercises

Exercise 1
Classify the following data as quantitative or qualitative:

Temperature	CONTINUOUS
Marital status	NOMINAL
Severity of arthritic pain	ORDERED CATEGORICAL
Blood pressure	CONTINUOUS
Number of visits to their GP per year	DISCRETE

Exercise 2
i) Proportion of women with epidural:
Labour in water is 23/49 = 0.47 or 47%
Augmentation is 33/50 = 0.66 or 66%.

ii) Relative risk of epidural for the Labour in water women compared with those having Augmentation is 0.47/0.66 = 0.71.

iii) Odds of epidural: Labour in water = 0.47/(1 - 0.47) = 0.89.
Odds of epidural: Augmentation is 0.66/(1 - 0.66) = 1.94.
OR of epidural for the Labour in water women compared with Augmentation is 0.89/1.94 = 0.46.
The OR is less than the RR and they differ because the event is quite common.

iv) 'Risk' of epidural on Labour in water is 23/49 = 0.47
'Risk' of epidural on Augmentation is 33/50 = 0.66.
The Absolute Risk Difference for Labour in water is ARR = |0.47 - 0.66| = 0.19.

Exercise 3
i)
(a) mode = 138 mmHg
(b) median = 139 mmHg
(c) mean = 1.41 mmHg

ii)
(a) range 116 to 170 mmHg
(b) inter quartile range IQR is 133 to 149 mmHg
(c) standard deviation = 13.62 mmHg

Exercise 4
i) A P value is the probability of observing the data (test statistic) or more extreme data if the null hypothesis (of no difference in ulcer healing rates between the two groups is true). The P value is greater than 0.05 i.e. not statistical significant, which implies we cannot reject the null hypothesis of no difference in ulcer healing rates between the groups. The interpretation – no reliable statistical evidence to suggest that leg ulcer healing rates at 12 weeks differ between the clinic and home treated groups.

ii) Definition of a 95% confidence interval (CI), something on the lines of, if we repeated the study 100 times, and calculated 100 CIs, 95 out of the 100 CI would contain the true population value of the difference in ulcer healing rates between the two groups. Conversely five of the studies/intervals would not contain the true value. Will also accept the layman's definition – 95% certain that the true difference in healing rates lies somewhere between –2% and 22% and the best estimate we have is 10%. The 95% CI includes zero which is consistent with the observed P value being greater than 0.05 and suggests that there is no difference in ulcer healing rates between the groups. The interpretation – no reliable statistical evidence to suggest that leg ulcer healing rates at 12 weeks differ between the clinic and home treated groups.

Exercise 5

1. We are interested in comparing the epidural analgesia rates (the proportions of women in each group having epidural analgesia during labour) between the Intervention and Control group women? Therefore Figure 6 is appropriate – methods for comparing two independent groups or samples. The data is binary; epidural analgesia during labour: yes or no. Therefore if we assume we have a large sample (n=99) and all expected frequencies are greater than five, then in these circumstances, the flow diagram of Figure 6 suggests that the most appropriate hypothesis test is a comparison of two proportions or the *chi-squared* test. Alternatively, if the expected frequencies are not greater than five then we can use the chi-squared test with Yates' continuity correction or Fisher's exact test.

2. The blood pressure difference data is paired (same person measured before and after exercise). Therefore we are interested in methods for difference in paired samples and can use Figure 7. Blood pressure (and the difference in BP) is measured on a continuous scale. If we assume the differences in blood pressure (before–after) are Normally distributed, then in these circumstances, the flow diagram of Figure 7 suggests that the most appropriate hypothesis test is the paired samples *t*-test. Alternatively, if the data (differences) are not Normally distributed we can use the Wilcoxon matched paired test.

References and further reading

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

Altman D.G., Machin D., Bryant T.N., & Gardner M.J. Statistics with Confidence. Confidence intervals and statistical guidelines (2nd Edition). London: British Medical Journal, 2000.

Armitage, P., Berry, G., and Matthews, J.N.S. (2002). Statistical Methods in Medical Research. 4th edition. Blackwell Science, Oxford.

Bland M. (2000). An Introduction to Medical Statistics, 3rd edition. Oxford University Press, Oxford.

Bland M. and Peacock J. (2000). Statistical Questions in Evidence-based Medicine. Oxford University Press, Oxford.

Campbell M.J., Machin D., & Walters S.J (2007). Medical Statistics: A textbook for the Health Sciences. 4th Edition. Wiley, Chichester.

Campbell M.J. (2006). Statistics at Square Two: Understanding Modern Statistical Applications in Medicine. 2nd edition. BMJ, London.

Cluett ER, Pickering RM, Getliffe K and Saunders NJSG (2004). Randomised controlled trial of labouring in water compared with standard augmentation for management of dystocia in first stage of labour. BMJ; 328: 314.

Furness S, Connor J, Robinson E, Norton, R, Ameratunga S and Jackson R (2003). Car colour and risk of car crash injury: population based case control study. BMJ; 327: 1455-1456.

Morrell CJ, Walters SJ, Dixon S, Collins KA, Brereton LM, Peters J and Brooker CG (1998). Cost effectiveness of community leg clinics. BMJ; 316, 1487-1491.

O'Cathain A, Walters SJ, Nicholl JP, Thomas KJ and Kirkham M (2002). Use of evidence based leaflets to promote informed choice in maternity care: randomised controlled trial in everyday practice. BMJ; 324: 643-646.

Simpson AG (2004). A comparison of the ability of cranial ultrasound, neonatal neurological assessment and observation of spontaneous general movements to predict outcome in preterm infants. PhD Thesis, University of Sheffield.

Swinscow T.D.V., Campbell M.J. (2002). Statistics at Square One. 10th Edition. BMJ, London.

Glossary

Alternative hypothesis (H_1 or H_A) The hypothesis against which the *null hypothesis* is tested.

Alpha (α) error The probability of a *type I error*. See also *significance level*.

Average Most often used for the arithmetic mean of a sample of observations, but can also be used for other measures of location, such as the median.

Beta (β) error Synonym for *type II error*.

Bias Deviation of results or inferences from the truth, or processes leading to such deviation.

Clinical vs. statistical significance The distinction between results in terms of their possible clinical or practical importance rather than simply in terms of their statistical significance. For example, with large samples very small differences that have little or no clinical importance may turn out to be statistically significant. The practical implications of any finding in a medical investigation must be judged on clinical as well as statistical grounds.

Confidence interval A range of values, calculated from the sample observations that are believed, with a particular probability, to contain the true parameter value. A 95% confidence interval, for example, implies that, were the estimation process repeated again and again, then 95% of the calculated intervals would be expected to contain the true parameter value. Note that the stated probability level refers to the properties of the interval and not to the parameter itself, which is not considered a *random variable*.

Contingency tables The tables arising when observations on a number of categorical variables are cross-classified. Entries in each cell are the number of individuals with the corresponding combination of variable values.

Data set A general term for observations and measurements collected during any type of scientific investigation.

Degrees of freedom An elusive concept that occurs throughout statistics. Essentially, the term means the number of independent units of information in a sample relevant to the estimation of a parameter or calculation of statistics. In a 2 x 2 contingency table with a given set of marginal totals, only one of the four cell frequencies is free, and the table therefore has a single degree of freedom.

Descriptive statistics A general term for methods of summarising and tabulating data that make their main features clearer; for example, calculating the means and standard deviations.

Estimate Either a single number (*point estimate*) or range of numbers (*interval estimate*) which are inferred to be plausible values for some parameter of interest.

Frequency distribution The division of a sample of observations into a number of classes, together with the number of observations in each class. Acts as a useful summary of the main features of the data, such as the location, shape and spread.

Hypothesis testing A general term for the procedure of assessing whether a sample data is consistent or otherwise with statements made about the population.

Normal distribution A *probability distribution* of a *random variable*, *x*, that is assumed by many statistical methods. The distribution is bell-shaped, as shown in Figures 3 and 4. The Normal distribution, (Figure 3), is completely described by two parameters: one, μ, represents the population mean or centre of the distribution and the other, σ, the population standard deviation. There are infinitely many Normal distributions depending on the values of • and • . The Standard Normal distribution has a mean of zero and a variance (and standard deviation) of one and a shape as shown in Figure 4.

Normality A term used to indicate that some variable of interest has a *Normal distribution*.

Null hypothesis (H_0) The 'no difference' or 'no association' hypothesis to be tested (usually by means of a significance test) against an *alternative hypothesis* that postulates non-zero differences or association.

Paired sample Two samples of observations with the characteristic feature that each observation in one sample has one and only one matching observation in the other sample.

Parameter A numerical characteristic of a population or model. For example, the mean, • , for the Normal distribution model.

Population In statistics this term is used for any finite or infinite collection of 'units', which are often people, but may be, for example, institutions, events etc.

Probability The quantitative expression of the chance that an event will occur. Can be defined in a variety of ways, of which the most common is still that involving a long term relative frequency, ie

$$P(A) = \frac{\text{number of times A occurs}}{\text{number of time A could occur}}$$

For example, if you toss a two-sided coin, 100 times and observe 51 heads the probability of a head is 51/100 or 0.51.

Probability distribution A mathematical formula, that gives the probability of each value of the variable for discrete random variables. For continuous random variables, it is the curve described by a mathematical formula which specifies, by way of areas under the curve, the probability that the variable falls within a particular interval. The Normal distribution is an example of a probability distribution for a continuous random variable.

Power The probability of rejecting the null hypothesis when it is false. Power gives a method of discriminating between competing tests of the same hypothesis, the test with the higher power being preferred. It is also the basis of procedures for estimating the sample size needed to detect an effect of a particular magnitude.

P value The probability of the observed data (or data showing more extreme departure from the *null hypothesis*) when the null hypothesis is true.

Quartiles The values that divide a *frequency distribution* or *probability distribution* into four equal parts.

Random Governed by chance; not completely determined by other factors.

Random sample A sample of *n* individuals selected from a population in such a way that each sample of the sample size is equally likely.

Random variable	A variable, the values of which occur according to some specified *probability distribution*.
Sample	A selected subset of a population, chosen by some process, usually with the objective of investigating particular properties of the parent population.
Significance test	A statistical procedure that, when applied to a set of observations results in a *P value* relative to some hypothesis. Examples include the *t*-test, chi-squared test and Mann-Whitney test.
Significance level	The level of probability at which it is agreed that the *null hypothesis* will be rejected. Conventionally set at 0.05.
Standard Deviation or SD or sd or s	The standard deviation of a sample of observations. It is a measure of their variability around the mean.
Statistic	A numerical characteristic of a sample. For example, the sample mean and sample variance. See also parameter.
Test statistic	A statistic used to assess a particular hypothesis in relation to some population. The essential requirement of such a statistic is a known distribution when the *null hypothesis* is true.
Type I error	The error that results when the *null hypothesis* is falsely rejected.
Type II error	The error that results when the *null hypothesis* is falsely accepted.
Variable	Some characteristic that differs from subject to subject or from time to time.