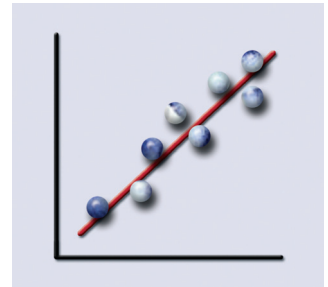


IN BRIEF

- The analysis of data from any research project seeks to answer the research question which was set out at the beginning of the project. As discussed in previous articles in this series the research question can be descriptive, may concern relationships, or may explore differences between groups.
- This article identifies the steps involved in the statistical analysis of both quantitative and qualitative data, including advice on software packages which can be used to support data analysis.
- For quantitative research a guide is given to choosing the right statistic for your data.

Research in primary dental care Part 6: Data analysis

A. C. Williams¹, E. J. Bower² and J. T. Newton³



The analysis of data from a research project seeks to answer the research question which the investigators set at the outset of the study. This article provides information on data analysis for both quantitative and qualitative data, always referring the reader back to their initial research question. Advice is also given on software which can be used for the analysis of data. A guide is provided to the choice of appropriate statistics for studies involving quantitative data.

RESEARCH IN PRIMARY DENTAL CARE

1. Setting the scene
2. Developing a research question
3. Designing your study
4. Measures
5. Devising a proposal, obtaining funding and ethical considerations
- 6. Data analysis**
7. Writing up your research

In this article we will discuss the analysis of data. The data analysis phase of any research project can be the most baffling, however the principle is simple – the data analysis answers the question which you stated at the start of the research.

It is beyond the scope of this series of articles to discuss data analysis in detail. Thus the article will introduce software packages before focusing on two important aspects of quantitative data analysis – how to describe data, and how to select the appropriate statistical test based on the research question. We will also briefly introduce qualitative data analysis. Throughout this article the importance of seeking advice on data analysis from an experienced researcher and/or statistician is emphasised.

SOFTWARE PACKAGES

There are several computer programs available to help you analyse your data. These can be divided into those designed for use with quantitative data and those designed for analysing qualitative data. These programs are designed to cover a very wide range of statistical techniques and will have many features which you are unlikely to use. It is therefore best to work with someone who is familiar with the program. Listed below are some examples of computer programs for data analysis.

Programs for analysing quantitative data

Epi-info: Very good, simple, program which covers most of the statistical analyses which researchers would want to use.

User-friendly screens, and guides to selecting and interpreting the statistics available. This can be downloaded from the internet

www.cdc.gov/epiinfo

Minitab: Simple program which covers all the basic statistical analyses. Not as easy to use as Epi-info.

SPSS: Commonly used comprehensive statistical package. Popular amongst social scientists. Complex to use.

STATA: A very powerful program with many sophisticated features. Popular amongst statisticians and epidemiologists. Complex to use.

Programs for analysing qualitative data

NU*DIST: Powerful program that stores data, allows the researcher to assign codes to data and analyses the codes numerically. It relies on the user to define the codes. An updated version, **NVIVO**, is now available.

Ethnograph: A similar program to **NU*DIST**. Allows the researcher to assign codes to data and analyse the frequency with which codes occur.

QUANTITATIVE DATA ANALYSIS

Data cleaning

Before you commence data analysis it is necessary to spend some time checking and 'cleaning' the data. You should also check that subjects who do not meet your original inclusion criteria

¹Consultant Senior Lecturer in Orthodontics, Department of Child Dental Health, University of Bristol Dental School, Lower Maudlin St, Bristol BS1 2LY;

²GDP, Staff Dental Service, Eastman Dental Hospital, 256, Gray's Inn Road, London WC1X 8LD; ³Professor of Psychology as Applied to Dentistry, Department of Dental Public Health & Oral Health Services Research, GKT Dental Institute, Caldecot Road, London SE5 9RW

*Correspondence to: Prof. J. T. Newton
Email: Tim.Newton@kcl.ac.uk

Refereed Paper

doi:10.1038/sj.bdj.4811467
© British Dental Journal 2004; 197:
67-73

are excluded at this stage, but remember to report how many are excluded and why. Given that the data are now on a spreadsheet or in a database it should be relatively easy to check for out-of-range values. For example, the variable gender may have three possible values (1 = Male, 2 = Female, 9 = Missing). If we examine the data and find a value of 3 then this is an error that will need to be corrected. For interval (continuous) data such as age, the easiest way to check for errors is to plot the data in a histogram and identify the outliers. It is also important in the data-cleaning process to identify nonsensical responses, for example, a subject who is recorded as being both male and pregnant! The easiest way to check for these sorts of errors is to cross-tabulate the relevant variables (see later) and check to see if there are subjects in the wrong cell of the table.

Describing the participants

The first step in the analysis is to describe the characteristics of the participants and then to

compare the sample with the population from which it was taken. In describing the characteristics of the sample, there is typically a mix of data types. Simple descriptive statistics will include:

- Frequency counts
- Proportions
- Measures of central tendency (mean, median, mode)
- Measures of dispersion (standard deviation, inter-quartile range etc).

Describing nominal and ordinal (categorical) data

Frequency counts are simply the recording of the number of times a particular value occurs in the data set. So, for example, the sample may consist of 100 males and 50 females. Frequencies can be readily expressed as proportions (for example, the sample consisted of 67% males, 33% females) and can be displayed in pie charts or bar charts.

You may also wish to describe the ‘average’ participant. The arithmetic mean is not suitable for use with nominal or ordinal data. For nominal (unordered categorical) variables the *mode* is often used, this takes the value of the most frequently occurring value (in the example given above, the modal value of gender is ‘Male’). For ordinal (ordered categorical) data we can use the mode, or the *median* value. The median is the middle value when all the values are ordered – half the sample will be below the median, and half above the median.

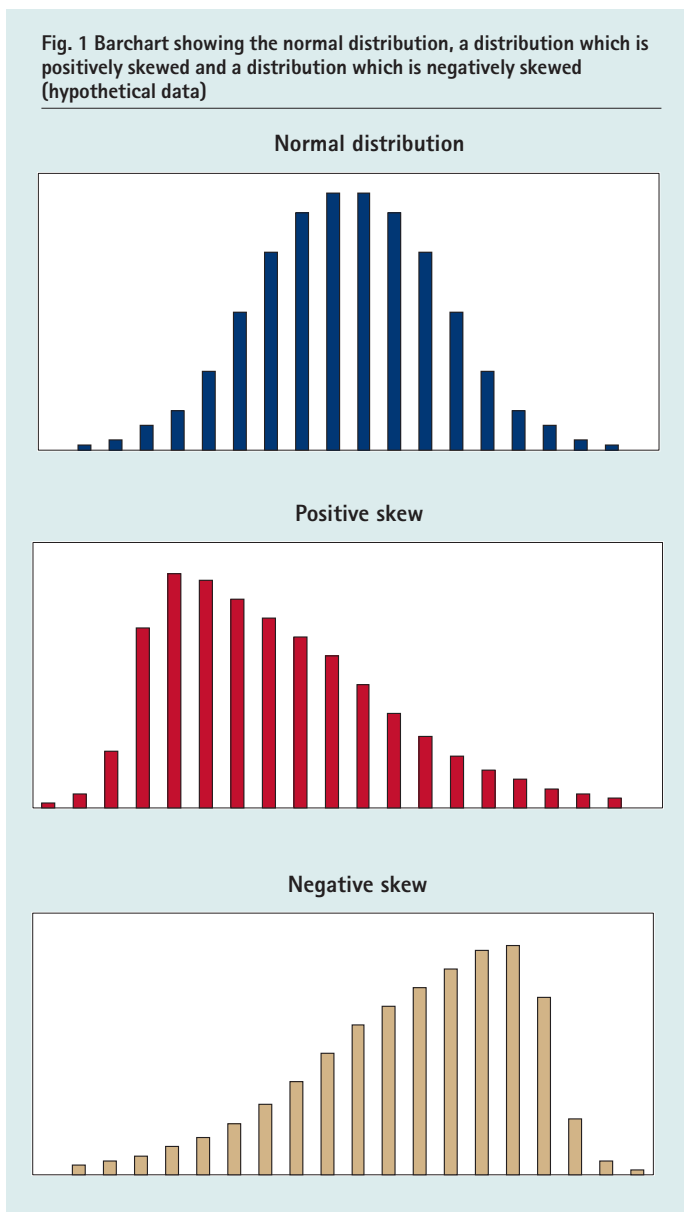
Describing interval (continuous) data

When describing interval data (for example, the age of the sample) we are concerned with three aspects: first, how the data are distributed; second, measuring the ‘average’ value, and third, exploring the spread of the data.

The easiest way to explore the distribution of data is to plot a frequency distribution – this plots the values of the variable on the x axis and the number of subjects with that value on the y axis. A line joining these points will show the distribution of the data.

There are many possible distributions of data. A central ‘hump’ (the top of which is the average case for that variable) with two equal ‘tails’ is known as the *normal distribution* (Fig. 1). Many data in nature, for example, the height of all the men in the UK, follow an approximately normal distribution. Most simple statistical tests assume that the data are normally distributed. If the data are not normally distributed, they may be *skewed*, that is the bulk of the curve is shifted to the left or right creating a tail that is longer on one side (Fig. 1). Many dental data are skewed. For example, data on the impact of oral health measured using the Oral Health Impact Profile (OHIP) in the *Adult Dental Health Survey*¹ were positively skewed (see Fig. 2).

Other data distributions are possible, for example, the *bimodal distribution*. This has



two 'humps' (see Fig. 3). This may represent the case where two distinct populations have been measured on the same variable. For example, if we measured the heights of everyone in the UK we might expect to get a bimodal distribution – with the two humps representing the average heights of men, and women respectively. Thus, a bimodal distribution can represent the merging of two different normal distributions.

There are three methods for determining the average value of interval data. The median and mode can be used as discussed previously. The *arithmetic average* (or *mean*) can also be calculated. The mean is calculated by summing all the values and then dividing by the number of values. If data are normally distributed, the median, mode and mean are all equal. When data are positively skewed the mean is lower than the median. If the mean is greater than the median then the data are negatively skewed.

The simplest measure of the spread or dispersion of data is the *range*. This demonstrates the spread from the lowest to the highest value. This is a useful measure but is heavily influenced by rare extreme values. Other measures try to overcome this disadvantage by describing the range of values for a proportion of the sample. For example, the *inter-quartile* range describes the spread between the lowest and highest 25% of the sample.

The *variance* is another measure of spread, calculated by averaging the square of the distance of each value from the mean. It is not expressed in the same units as the original data, which makes its interpretation difficult. In order to obtain a measure of spread in the same units as the original data, the square root of the variance is taken, to give the *standard deviation*. If your data are normally distributed then we can assume that 68.26% of your data will be within ± 1 standard deviation of the mean. Similarly, 95.44% of your data will be within ± 2 standard deviations of the mean. Figure 4 summarises some measures of dispersion.

The second task when describing the sample is to compare the sample with the population from which it was drawn and determine whether the sample is representative of this population. For example, you may have selected a sample of patients for your study and want to know how they compare with all the patients who are registered with your practice so that you can generalise your results to your total practice population. These comparisons will be limited by the information that you have available about your total population. It is likely that you will know some basic demographic details about your patients, for example, their age and sex. You may also know some basic treatment details such as the proportion of regular attenders. The analyses here are all based on the comparison of groups: the group formed by the sample and the group formed by the population. Descriptive analysis of the characteristics of

Fig. 2 Graph showing the distribution of OHIP scores in the most recent *Adult Dental Health Survey*. Note positive skew. (Reproduced from Nuttall N M, Steele J G, Pine C M, White D, Pitts N B. The impact of oral health on people in the UK in 1998. *Br Dent J* 2001; 190: 121–126)

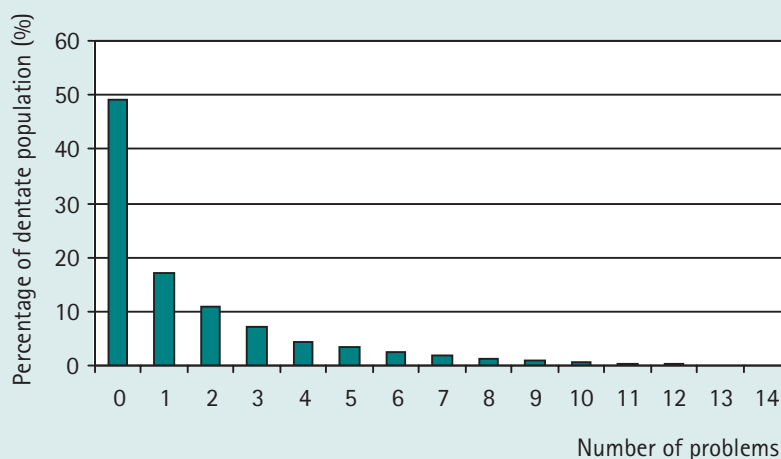


Fig. 3 A bimodal distribution (hypothetical data)

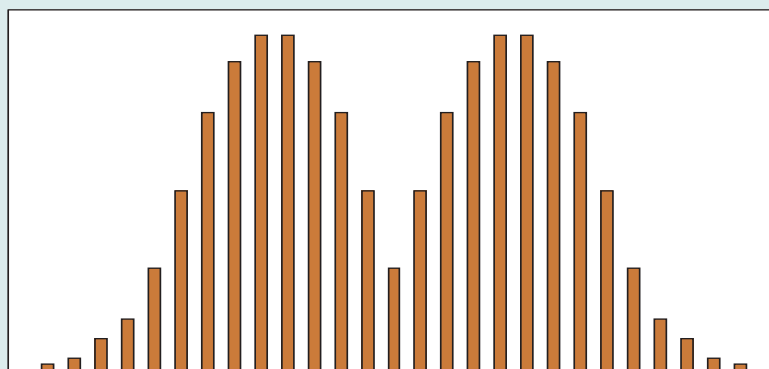
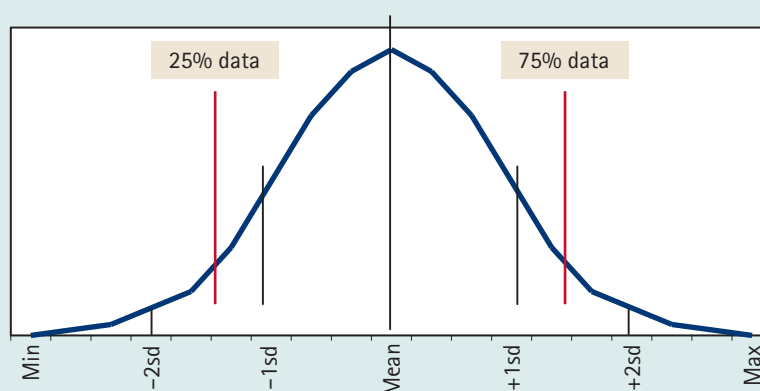


Fig. 4 Measures of dispersion and the normal distribution (hypothetical data)



Range = (Max – Min)

Inter-Quartile range = (Point of 75% of data – Point of 25% of data)

68.26% of the data fall within the range of ± 1 standard deviation of the mean

95.44% of the data fall within the range of ± 2 standard deviations of the mean

Table 1 Example of crosstabulation. Hypothetical example looking at the association between gender and referral to hygienist

	Referral to hygienist		Total
	No	Yes	
Male participants	60	40	100
Female participants	35	65	100
Total	95	105	200

Chi-square = 12.5, $P < 0.05$

your sample and your total population will give you a quick indication of whether your sample is representative or not. A more formal way of checking interval (continuous) data is to calculate 95% *confidence intervals*. A 95% confidence interval gives the range of values within which one can be 95% confident that the true population mean lies. It is calculated as follows:²

$$95\% \text{ confidence intervals} = \text{sample mean} \pm 1.96 \times \text{SE (standard error)}$$

If you are dealing with categorical data divided into two groups (for example, sex) then you can use the proportion (p) of subjects with a characteristic expressed as a percentage to calculate your confidence intervals:

$$95\% \text{ confidence intervals} = p \pm 1.96 \times \sqrt{\{p(1-p)\}/n}$$

where n = sample size

Choosing the appropriate statistical test

Throughout this series great emphasis has been placed on the importance of the research question. The research question drives the design of the study, the choice of methods and will now determine the appropriate data analysis strategy. The following types of research questions were identified:

- Descriptive questions
- Questions of relationship or association
- Questions of difference between groups – comparative studies

These different question types require different types of statistical analysis. We will examine each type of question in turn and illustrate the statistics for that type of question.

Descriptive questions

Most research will start with some description. The statistics required are described above.

Questions of relationship or association

We will explore two types of relationship:

- Association
- Correlation

Association

Association refers to the extent to which two variables tend to occur together. For example,

we may hypothesise that women tend to be referred to the hygienist more frequently than men. That is, we think that the variables ‘Gender’ and ‘Referral to hygienist’ are associated – they are not completely independent. Note that both these variables are nominal data – Gender is ‘Male’ or ‘Female’, and Referral to Hygienists can take the value ‘Yes’ or ‘No’.

To look at the extent to which these variables are associated we first need to crosstabulate them – that is we examine each case and determine which of the four possible combinations of these two variables they demonstrate. We then repeat this for all cases (see Table 1).

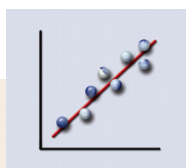
We can then calculate the association between the two variables using the chi-squared (χ^2) statistic. If the chi-squared value is significant then the two variables are associated. For the hypothetical example given in Table 1 we can see that a higher proportion of women were referred to the hygienist than men. The value of chi-squared is 12.5 which is statistically significant at the $P \leq 0.05$ level. There is therefore an association between gender and referral pattern in this sample which, in 95% of cases, is unlikely to have occurred purely by chance.

Correlation

The correlation between variables refers to the extent to which they ‘go together’, that is, one changes as the other changes. The simplest form of correlation is between two variables. They may be positively correlated that is, as one increases so does the other, for example, height and weight. In general, taller people are heavier. Note that there is not a perfect correlation – some shorter people may be heavier than some taller people, but there is a positive correlation. If we plotted one against the other we could summarise the data by a straight line running from the bottom left to the top right of a graph (Fig. 5). Alternatively, two variables may be negatively correlated, that is, as one increase the other decreases. For example, as sugar intake increases we might expect the number of sound and healthy teeth to decrease. We can summarise this as a line running from the top left of the graph to the bottom right (Fig. 5).

The correlation between two variables can be summarised using the Correlation Coefficient, values of which range from -1 to $+1$. A correlation of 0 means that the two variables are not correlated at all. A value of -1 represents a perfect negative correlation, $+1$ a perfect positive correlation.

There are many statistical techniques that use correlation. The simplest involve just two variables, though it is possible to correlate three or more variables. For the two variable case we can use two statistics. The Pearson correlation is used when both variables are interval variables, and both are normally distributed. If either or both variables are not normally distributed then we should use a similar statistic called Spearman’s rho. Spearman’s rho is an example of a ‘non-parametric’ statistic. Non-parametric statistics are



The research question drives the design of the study, the choice of methods and will now determine the appropriate data analysis strategy

a group of statistical techniques which can be used when your data are not normally distributed (for example, the data may be skewed), or where you have data that are ranked ordinal rather than interval. They are also used when the variances of the samples are significantly different, which is more likely if the sample sizes are dissimilar. This is particularly relevant to unrelated samples. A good introduction to those techniques is given in Sprent & Smeeton.³

It is also possible to correlate more than two variables. To do this we need to use regression techniques, which are complex correlational techniques. These techniques allow us to demonstrate the relationship between an outcome variable (say DMFT) and several predictor variables (for example, age, gender and social class). We can identify the extent to which each predictor variable influences the outcome separately, and we can build a model which shows how all the predictors together relate to the outcome variable. These are complex statistical tests, the results of which can be difficult to interpret. We strongly advise you to consult a statistician at an early stage if you are interested in using these techniques.

The most important thing to remember about all correlational techniques is that they only demonstrate that two things are related. They do not demonstrate that one causes the other. Correlation is not the same as causation.

Questions of difference between groups – comparative studies

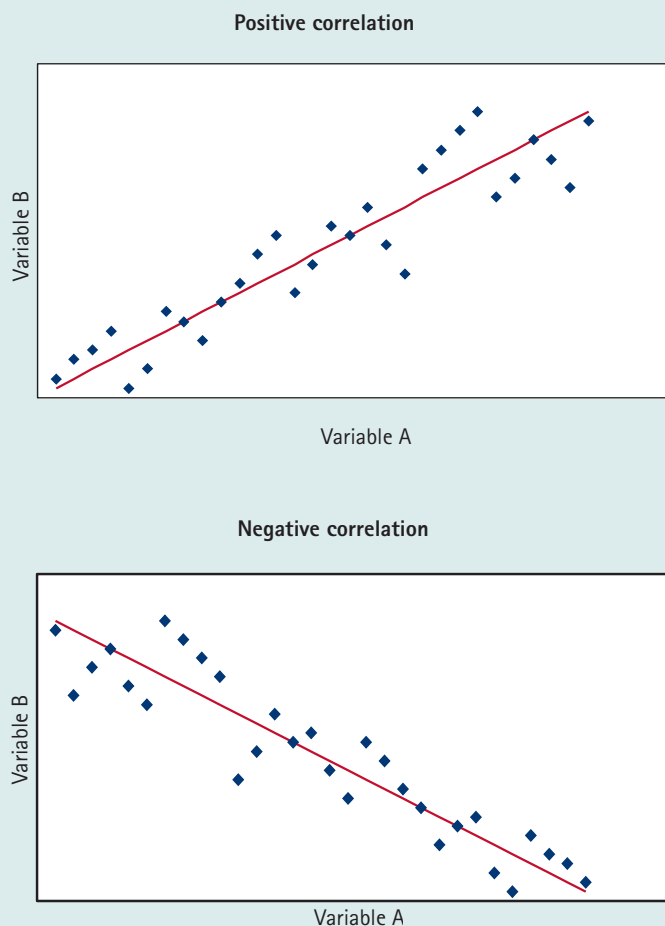
Two groups

Your research question may require you to compare two groups on a variable which is interval (continuous) data. The choice of statistical test to compare the groups will depend on the type of data that you are comparing, the number of groups, whether the groups are related (or paired) and whether the data are normally distributed or not. For example, we may wish to compare percentage plaque scores for two groups where one group was given oral hygiene instruction by a dental practitioner, and the other group was taught by a dental hygienist. The outcome measure is interval data. If the data are normally distributed, we can compare the two means using the Student's *t*-test statistic. If the data are not normally distributed then we need to use a non-parametric test called the Mann-Whitney *U* test. If the two groups are related (for example, plaque scores in one side of the mouth compared with the other side of the mouth in the same patient) then you should use the paired *t*-test or its non-parametric equivalent, the Wilcoxon signed rank test.

Three or more groups

You may wish to compare outcomes between three or more groups. For example, DMFT scores for patients within several age groups. To compare the mean DMFT for the different age groups a statistical test called the Analysis of Variance (or ANOVA) is used. Alternatively, if the data are

Fig. 5 Graph showing a positive correlation and a negative correlation (hypothetical data)



not normally distributed, then you should use the Kruskal-Wallis test which is a non-parametric version of the ANOVA.

Before and after changes (related samples)

Alternatively, you may wish to assess the impact of a treatment by comparing the same group of subjects before and after the intervention. In the simple case of comparing one group before and after treatment we treat the data as paired data and use the paired *t*-test or the Wilcoxon signed rank test as previously. In cases where you wish to compare a single group on more than two occasions a test called the repeated measures ANOVA is used. This is a complex statistical technique and it would be best to ask advice from a statistician if you wish to use such a technique.

Mixed designs

Some experimental designs will combine a comparison between groups with before and after components. So, for example, you may wish to compare two groups of participants before and after treatment, one group receiving an active treatment, the other receiving a placebo. An ANOVA based technique can be used to analyse such designs but is probably best carried out by a statistician.

Fig. 6 Choosing the right statistic for your research question and your data

Research question is <i>Descriptive</i>		
Data are <i>Nominal</i> (categorical)	Data are <i>Ordinal</i> (ordered categorical data)	Data are <i>Interval</i> (continuous data)
<ul style="list-style-type: none"> • Frequencies • Proportions • Mode 	<ul style="list-style-type: none"> • Frequencies • Proportions • Median • Mode 	<ul style="list-style-type: none"> • Examine distribution • Measures of central tendency <ul style="list-style-type: none"> Mean Median Mode • Measures of dispersion <ul style="list-style-type: none"> Range Standard deviation
Research question is about <i>Relationships</i>		
Association between two variables	Correlation between two variables	
Data should be nominal or ordinal (categorical data)	Data should be ordinal or interval (ordinal data are subject to non-parametric tests)	
<ul style="list-style-type: none"> • Crosstabulate • Chi-square 	<ul style="list-style-type: none"> • Pearson's correlation coefficient (parametric test) • Spearman's rho (non-parametric test) 	
Research question is about <i>Comparing groups</i>		
	Outcome is a nominal or ordinal variable (categorical data)	Outcome is an ordinal or interval variable (ordinal data are subject to non-parametric tests)
One group measured on two occasions (related data)		<ul style="list-style-type: none"> • Within subjects <i>t</i>-test (parametric) • Wilcoxon Signed Rank test (non-parametric)
One group measured on three or more occasions (related data)		<ul style="list-style-type: none"> • Repeated measures ANOVA (parametric) • Friedman test (non-parametric)
Two separate groups (independent data)	<ul style="list-style-type: none"> • Crosstabulate • Chi-square 	<ul style="list-style-type: none"> • Student's <i>t</i>-test (parametric) • Mann-Whitney <i>U</i> test (non-parametric)
Three or more separate groups (independent data)	<ul style="list-style-type: none"> • Crosstabulate • Chi-square 	<ul style="list-style-type: none"> • Analysis of Variance (ANOVA) (parametric) • Kruskal-Wallis test (non-parametric)

Fig. 7 Checklist for choosing the appropriate statistical test

1. Does the hypothesis (generated from the research question) predict a **relationship** between variables or a **comparison (difference)** between groups?
2. Consider the **type** of data (nominal, ordinal, interval).
3. Are **parametric test assumptions** met? (Parametric tests are used when data are at interval level, the samples are drawn from a normally distributed population and the variances of the samples do not differ significantly.)
4. Are the data **related** (measurements repeated on the same subjects under different conditions) or **independent** (measurements taken on different subjects)?
5. Are there **more than two** variables or groups, and are measurements taken in the same subjects **more than twice**?

In summary, the choice of statistical method that you use for quantitative data analysis will depend on the nature of the research question and the type of data that you have collected. Figure 6 summarises the types of data analyses you may wish to undertake and the appropriate statistical test. Figure 7 provides a checklist of considerations when choosing which test to use.

ANALYSING QUALITATIVE RESEARCH

There are two basic approaches to the analysis of qualitative data: the inductive approach and the deductive approach. There are a number of specific methods for analysing qualitative data which differ in the extent to which they emphasise inductive techniques over deductive techniques or vice versa. You should seek the help of an experienced researcher when analysing qualitative data.

Deductive techniques of qualitative data analysis

Deductive techniques commence with an explicit structure and then use that structure to analyse the data, for example, if a series of interviews are carried out exploring patients' reasons for making a complaint. The researcher might carry out a series of in-depth qualitative interviews with patients that have made complaints. A schedule could then be drawn up listing the major reasons for patients' complaints (for example, trauma following treatment, difficulties of communication, unexpected adverse

outcomes ... etc). The data analysis would then consist of examining each interview to determine how many patients had complaints of each type and the extent to which complaints of each type co-occur. The key point is that in the deductive approach the researcher imposes their own structure on the data and then uses this to analyse the interviews. Deductive approaches are useful for research questions where the researcher is confident that they know what the full range of answers will be. These answers might be based on previous empirical or theoretical knowledge. An example of a deductive approach is content analysis.⁴

Inductive techniques of qualitative data analysis

Inductive techniques use the data themselves to derive the structure of the analysis. The researcher starts by assuming that the categories which can be used to summarise the data are a theoretical 'blank sheet'. Specific techniques are then used to determine the categories which are used to analyse the data. Once these categories have been identified they can be validated against previous research and theoretical knowledge. For example, an inductive approach to the analysis of the complaints data we mentioned above would examine each example of a complaint and identify a category to which it belonged (for example, communication errors). The next complaint would then be compared with the previous one to determine whether it was a similar type. If not then a new category is devised (for example, poor outcome) and so on for all the interviews. Grounded theory is an example of an inductive approach to qualitative data analysis.⁵

Combining inductive and deductive approaches

In practice most qualitative researchers use

Table 2 Further reading

General guide to quantitative data analysis:

1. Bulman J S, Osborn J F. *Statistics in Dentistry*. London: British Dental Association, 1989.
2. Petrie A, Bulman J S, Osborn J F. *Further statistics in Dentistry*. London: BDJ Books, 2002.
3. Armitage P, Berry G, Matthews J N S. *Statistical methods in medical research*. Oxford: Blackwell Science, 2002.
4. Banerjee A. *Medical statistics made clear: an introduction to basic concepts*. London: Royal Society of Medicine Press Ltd., 2003.

Readers may also wish to consult the series of articles on statistics published in the *British Dental Journal* by Petrie A, Bulman J S & Osborn J F, from October 2002.

Information on analysis of qualitative data:

1. Silverman D. *Interpreting qualitative data*. London: Sage, 2001.
2. Richie J, Spencer L. Qualitative data analysis for applied policy research. In Bryman A, Burgess R (ed) *Analyzing Qualitative Data*. London: Routledge, 1994.

(either implicitly or explicitly) a combination of inductive and deductive approaches. An example is Interpretive Phenomenological Analysis.⁶ This approach adopts an inductive approach to the categorisation of data, but then uses these categories deductively to give an idea of how commonly categories occur and the nature of their relationships.

1. Kelly M, Steele J, Nuttall N, Bradnock G, Morris J, Pine C, Pitts N, Treasure E, White D. *Adult dental health survey: Oral health in the United Kingdom 1998*. London: The Stationary Office, 2000.
2. Banerjee A. *Medical statistics made clear: an introduction to basic concepts*. London: Royal Society of Medicine Press Ltd., 2003.
3. Sprent P, Smeeton N C. *Applied nonparametric statistical methods*. London: Chapman & Hall/CRC, 2000.
4. Weber R P. *Basic content analysis (Quantitative applications in the social sciences)*. London: Sage, 1990.
5. Glaser B G, Strauss A L. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New York: Aldine Publishing Company, 1967.
6. Smith J A, Harre R M, Van Langenhove L. *Rethinking methods in Psychology*. London: Sage, 1995.