CHAPTER 17

The analysis and interpretation of quantitative data

This chapter:

- stresses the advantages of using a software package when analysing quantitative data and your likely need for help and advice when doing this;
- shows how to create a data set for entry into a computer;
- distinguishes between exploratory and confirmatory data analysis;
- explains statistical significance and discusses its controversial status;
- advocates greater reliance on measures of effect sizes;
- suggests how to explore, display, and summarize the data;
- discusses ways of analysing relationships between various types of data and a range of statistical tests that might be used;
- does the same thing for analysing differences between data; and
- considers issues specific to the analysis of quasi-experiments, single-case experiments, and non-experimental fixed designs.

Introduction

You would have to work quite hard in a research project not to have at least some data in the form of numbers, or which could be sensibly turned into numbers of some kind. Hence, techniques for dealing with such quantitative data are an essential feature of your professional toolkit. Their analysis covers a wide range of things, from simple organization of the data to complex statistical analysis. This chapter does not attempt a comprehensive treatment of all aspects of quantitative data analysis. Its main aim is to help you appreciate some of the issues involved so that you have a feeling for the questions you need to ask when deciding on an appropriate kind of analysis.

Some assumptions

1. *Everyone doing real world research needs to understand how to summarize and display quantitative data*. This applies not only to those using fixed and multi-strategy designs, but also to users of flexible designs where their data are essentially qualitative. Even die-hard qualitative researchers will often collect small amounts of numerical data or find advantage in turning some qualitative data into numbers for summary or display purposes. This does not necessarily call for the use of statistical tests. Simple techniques may be all you need to interpret your data.

2. *For relatively simple statistical tests specialist statistical software is not essential*. If you only have a very small amount of quantitative data, it may be appropriate for you to carry out analyses by 'hand' (or with the help of an electronic calculator). However, the drudgery and potential for error in such calculation, and the ease with which the computer can perform such mundane chores for you, suggest strongly that you make use of the new technology if at all possible. For such tasks, and for simple statistical tests, spreadsheet software such as Excel may be all that you need. 'Analyse-it' (http://www.analyse-it.com) is a straightforward package which can be used with Excel to produce most of the commonly used statistics and charts. Appendix A gives details. It has been used for several of the figures showing the results of different statistical analyses in this chapter.

3. *If you need to carry out complex statistical tests you will need to use a specialist statistical computer package. A range of commonly used statistical packages is discussed in Appendix A*. SPSS (the Statistical Package for the Social Sciences) is the market leader by some margin but other packages are well worth considering, particularly if you wish to follow the exploratory data analysis (EDA) approach highlighted in the chapter. Facility in the use of at least one specialist statistical package is a useful transferable skill for the real world researcher.

4. *You have some prior acquaintance with the basic concepts and language of statistical analysis*. If not, you are recommended to spend some time with one of the many texts covering this at an introductory level (e.g. Graham, 2013; Robson, 1994; Rowntree, 2000).

5. *You will seek help and advice in carrying out statistical analyses*. The field of statistical analysis is complex and specialized and it is unreasonable to expect everyone carrying out real world research to be a statistical specialist. It is, unfortunately, a field where it is not at all difficult to carry out an analysis which is simply wrong, or inappropriate, for your data or your purposes. And the negative side of readily available specialist statistical software is that it becomes that much easier to generate elegantly presented rubbish (remember GIGO – Garbage In, Garbage Out).

Preferably, such advice should come from an experienced statistician sympathetic to the particular difficulties involved in applied social research. Repeating the advice once more – it should be sought at the earliest possible stage in the *design* of your project. Inexperienced non-numerate researchers often have a touching faith that research is a linear process in which they first collect the data and then the statistician shows them the analysis to carry out. It is, however, all too easy to end up with unanalysable data, which, if they had been collected in a somewhat different way, would have been readily analysable. In the absence of personal statistical support, you should be able to use this chapter to get an introduction to the kind of approach you might take. The references provided should then help with more detailed coverage.

Organization of the chapter

The chapter first covers the creation of a 'data set' as a necessary precursor to data analysis. Suggestions are then made about how you might carry out various types of data analysis appropriate for different research designs and tasks.

Creating a data set

If you are to make use of a computer to help with analysis, then the data must be entered into the computer in the form required by the software you are using. This may be done in different ways:

1. *Direct automatic entry*. It may be feasible for the data to be generated in such a way that entry is automatic. For example, you may be using a structured observation schedule with some data collection device (either a specialized instrument or a laptop computer) so that the data as collected can be directly usable by the analysis software.
2. *Creation of a computer file which is then 'imported' to the analysis software*. It may be easier for your data to be entered into a computer after collection. For example, a survey might use questionnaires which are 'optically readable'. Respondents, or the person carrying out the survey, fill in boxes on the form corresponding to particular answers. The computer can directly transform this response into data which it can use. Such data form a computer 'file' which is then 'imported' into the particular analysis software being used. This is feasible with most statistical packages although you may need assistance to ensure that the transfer takes place satisfactorily.
3. *Direct 'keying' of data into analysis software*. For much small-scale research, automatic reading or conversion of the data into a computer file will either not be possible or not be economically justifiable. There is then the requirement for manual entry of data into the analysis software. The discussion below assumes that you will be entering the data in this way.

   Whichever approach is used, the same principle applies. Try at the design stage to capture your data in a form which is going to simplify this entry process. Avoid intermediate systems where the original response has to be further categorized. The more times that data are transferred between coding systems, the greater the chance of error. *Single-transfer coding* (i.e. where the response is already in the form which has to be entered into the computer) is often possible with attitude and other scales, multiple-choice tests, inventories, checklists, and many questionnaires. In a postal or similar survey questionnaire, you will have to weigh up whether it is more important to simplify the task of the respondent or the task of the person transferring the code to the computer. Box 17.1 shows possible alternatives.

   The conventions on coding are essentially common sense. Suggestions were made in Chapter 11 (p. 272) about how this might be dealt with in relation to questionnaires. Note that it is helpful to include the coding boxes on the questionnaire itself, conventionally in a column on the right-hand side of each page.

   The data sets obtained from other types of project will be very various. However, it is almost always possible to have some sensible arrangement of the data into *rows* and *columns*. Typically each row corresponds to a *record* or *case*. This might be all of the data obtained from a particular

BOX 17.1

Question formats requiring (a) single-transfer coding and (b) double-transfer coding

(a) How many children are there in your school?

under	40	40–49	50–59	60–69	70–79	80–89	90–100	over 100
code	1	2	3	4	5	6	7	8

enter code ( )

(b) How many children are there in your school?
(please circle)

under   40   40–49   50–59   60–69   70–79   80–89   90–100   over 100

(response has then to be translated into appropriate code)

respondent. A record consists of *cells* which contain data. The cells in a column contain the data for a particular *variable*. Figure 17.1 presents a simple example derived from a survey-type study.

Student	Faculty	Sex	Entry points	Degree class	Income
1	A	F	14	2.1	14,120
2	EN	M	6	2.2	15,900
3	EN	M	5	Fail	11,200
4	ED	F	10	2.2	21,640
5	S	M	4	2.1	25,000
6	B	F	13	2.1	11,180
7	A	F	16	2.1	12,600
8	EN	M	6	3	9,300
9	ED	M	5	3	2,200
10	EN	M	*	2.2	17,880

Key: A = Arts; B = Business; Ed = Education; EN = Engineering; S = Sciences;
M = Male; F = Female; * = missing data

Note: data are fictitious, but modelled on those in Linsell and Robson, 1987

Figure 17.1 Faculty, entry points, degree classification, and income two years after graduating of a sample of students.

A similar matrix would be obtained from a simple experiment where, say, the columns represent scores obtained under different experimental conditions.

Entering the data into the computer

The details of the procedure for entering this data set into the computer vary according to the particular software you are using. With early versions of software, this was quite complex but later versions are straightforward to use, particularly if you are familiar with the operation of spreadsheets.

Missing data

'The most acceptable solution to the problem of missing information is not to have any' (Youngman, 1979, p. 21). While this is obviously a counsel of perfection, it highlights the problem that there is no really satisfactory way of dealing with missing data. It may well be that the reason why data are missing is in some way related to the question being investigated. Those who avoid filling in the evaluation questionnaire, or who are not present at a session, may well have different views from those who responded. So it is well worth spending considerable time, effort and ingenuity in seeking to ensure a full response. Software normally has one or more ways of dealing with missing data when performing analyses and it may be necessary to investigate this further as different approaches can have substantially different effects on the results obtained.

Technically there is no particular problem in coding data as missing. There simply needs to be a signal code which is used for missing data, and only for missing data. Don't get in the habit of using 0 (zero) to code for missing data as this can cause confusion if the variable in question could have a zero value or if any analytic procedure treats it as a value of zero (99 or –1 are frequently used). Software packages should show the value that you have specified as missing data and deal with it intelligently (e.g. by computing averages based only on the data present).

It is worth noting that a distinction may need to be made between missing data where there is no response from someone, and a 'don't know' or 'not applicable' response, particularly if you have catered for possible responses of this type by including them as one of the alternatives.

Cleaning the data set after entry

Just as one needs to proof-read text for errors, so a computer data set needs to be checked for errors made while 'keying in' the data. One of the best ways of doing this is for the data to be entered twice, independently, by two people. Any discrepancies can then be resolved. This is time-consuming but may well be worthwhile, particularly if substantial data analysis is likely.

A valuable tip is to make use of 'categorical' variables whenever feasible. So, in the data set of Box 17.1 'degree class' has the categories 'first, 'upper second', etc. The advantage is that the software will clearly show where you have entered an invalid value.

While this eliminates several potential mistakes, it is, of course, still possible to enter the wrong class for an individual. The direct equivalent of proof-reading can be carried out by checking the computer data set carefully against the original set. Simple *frequency analyses* (see below) on each of the columns are helpful. This will throw up whether 'illegal', or highly unlikely, codes have been entered. For continuous variables *box plots* can be drawn, and potential 'outliers' highlighted (see p. 420).

Cross-tabulation

This involves counting the codes from one variable that occur for each code in a second variable. It can show up more subtle errors. Suppose that the two variables are 'withdrew before completing degree' and 'class of final degree'. *Cross-tabulation* might throw up one or two students who appeared to have withdrawn before completion but were nevertheless awarded a classified degree. These should then be checked, as while this might be legitimate (perhaps they returned), it could well be a miscoding. Cross-tabulation is easy when the variables have only a few values, as is the case with most categorical variables. However, it becomes very tedious when continuous variables such as age or income, which can take on many values, are involved. In this circumstance, *scattergrams/scatter plots* (see below) provide a useful tool. These are graphs in which corresponding codes from two variables give the horizontal and vertical scale values of points representing each record. 'Deviant' points which stand out from the general pattern can be followed up to see whether they are genuine or miscoded.

> *The 'cleaned' data set is an important resource for your subsequent analyses. It is prudent to keep a couple of copies, with one of the copies being at a separate physical location from the others. You will be likely to modify the set in various ways during analysis (e.g. by combining codes); however, you should always retain copies of the original data set.*

Starting data analysis

Now that you have a data set entered into the computer you are no doubt itching to do something with it. Data analysis is commonly divided into two broad types: exploratory and confirmatory. As the terms suggest, exploratory analysis explores the data trying to find out what they tell you. Confirmatory analysis seeks to establish whether you have actually got what you expected to find (for example on the basis of theory, such as predicting the operation of particular mechanisms).

With all data sets, and whatever type of research design, there is much to be said for having an initial exploration of the data. Try to get a feeling for what you have got and what it is trying to tell you. Play about with it. Draw up tables. Simple graphical displays help: charts, histograms, graphs, pie-charts, etc. Get summaries in the form of means and measures of the amount of variability, etc. (Details on what is meant by these terms, and how to do it, are presented later in the chapter.) Acquiring this working knowledge is particularly useful when you are going on to use various statistical tests with a software package. Packages will cheerfully and quickly produce complex nonsense if you ask them the wrong question or misunderstand how you enter the data. A good common-sense understanding of the data set will sensitize you against this.

Exploratory approaches of various kinds have been advocated at several points during this book. They are central to much flexible design research. While these designs mainly generate qualitative data, strategies such as case study commonly also result in quantitative data which we need to explore to see what has been found and to help direct later stages of data collection.

Much fixed design research is exclusively quantitative. The degree of pre-specification of design and of pre-thought about possible analyses called for in fixed design research means that the major task in data analysis is confirmatory; i.e. we are seeking to establish whether our

predictions or hypotheses have been confirmed by the data. Such *confirmatory data analysis (CDA)* is the mainstream approach in statistical analysis.

However, there is an influential approach to quantitative analysis known as *exploratory data analysis (EDA)* advocated by Tukey (1977) – see also Myatt and Johnson (2014). Tukey's approach and influence come in at two levels. First, he has proposed several ingenious ways of displaying data diagrammatically. These devices, such as 'box plots', are non-controversial, deserve wider recognition, and are discussed below. The more revolutionary aspect of the EDA movement is the centrality it places on an informal, pictorial approach to data. EDA is criticized for implying that the pictures are all that you need; that the usual formal statistical procedures involving tests, *significance* levels, etc. are unnecessary. Tukey (1977) does acknowledge the need for CDA; in his view it complements EDA and provides a way of formally testing the relatively risky inductions made through EDA.

To a large extent, EDA simply regularizes the very common process whereby researchers make inferences about relationships between variables after data collection which their study was not designed to test formally – or which they had not expected prior to the research – and provides helpful tools for that task. It mirrors the suggestion made in Chapter 6 that while in fixed design research strong pre-specification is essential and you have clear expectations of what the results will show (i.e. the task of analysis is primarily confirmatory), this does not preclude additional exploration. Using EDA approaches, with a particular focus on graphical display, has been advocated by Connolly (2006) as a means of avoiding the ecological fallacy of making inferences about individuals from the group data provided from summary statistics.

In practice the EDA/CDA distinction isn't clear cut. As de Leeuw puts it (in Van de Geer, 1993), the view that:

> The scientist does all kinds of dirty things to his or her data . . . and at the end of this thoroughly unrespectable phase he or she comes up (miraculously) with a theory, model, or hypothesis. This hypothesis is then tested with the proper confirmatory statistical methods. [This] is a complete travesty of what *actually* goes on in all sciences some of the time and in some sciences all of the time. There are no two phases that can easily be distinguished (emphasis in original).

> *The treatment in this chapter is influenced by EDA and seeks to follow its spirit. However, there is no attempt to make a rigid demarcation between 'exploring' and 'confirming' aspects.*

A note on 'levels' of measurement

A classic paper by Stevens (1946) suggested that there were four 'levels' of measurement ('nominal', 'ordinal', 'interval' and 'ratio'). Nominal refers to a set of categories used for classification purposes (e.g. marital status); ordinal also refers to a set of categories where they can be ordered in some meaningful way (e.g. social class); interval refers to a set of categories which are not only ordered but also have equal intervals on some *measurement scale* (e.g. calendar time); ratio is the same as interval level, but with a real or true zero (e.g. income).

Although very widely referred to in texts dealing with the analysis of quantitative data, the value of this typology has been queried by statisticians (e.g. Velleman and Wilkinson, 1993). Gorard (2006) considers it unnecessary and confusing. He claims that there is little practical

difference between interval and ratio scales and points out that the same statistical procedures are traditionally suggested for both. Also that:

> So-called 'nominal' measures are, in fact, not numbers at all but categories of things that can be counted. The sex of an individual would, in traditional texts, be a nominal measure. But sex is clearly not a number . . . The only measure involved here is the frequency of individuals in each category of the variable 'sex' – i.e. how many females and how many males (p. 61).

Such frequencies are, of course, 'real numbers' and can be added, subtracted, multiplied and divided like other numbers. 'Ordinal' measures are also categories of things that can be counted and can be treated in exactly the same way. The only difference is in the possibility of ordering which can be used when describing and displaying frequencies.

He highlights the fact that a major problem arises when ordinal categories are treated as real numbers. For example examination grades, A, B, C, D and E may be given points scores, say that A is 10 points, B is 8 points, etc. As such points scores are essentially arbitrary; attempts to treat them as real numbers, for example by working out average points scores, lead to arbitrary results. Gorard's advice is to:

> . . . use your common sense but ignore the idea of 'levels' of measurement. If something is a real number then you can add it. If it is not a real number then it is not really any kind of number at all (p. 63).

Our advice is to take note of this advice but not to let it inhibit you from carrying out any of the statistical analyses (particularly the simple ones) covered in the chapter – providing you understand what you are doing, and it seems likely to shed light on what the data are trying to tell you. The notion that specific measurement scales are requirements for the use of particular statistical procedures, put forward by Stevens (1946), followed up in influential statistics text-books (e.g. Siegel, 1959), and still commonly found, is rejected by many mathematical statisticians (see Gaito, 1980; Binder, 1984). There is nothing to stop you carrying out any analysis on quantitative data *on statistical grounds*. As Lord (1953) trenchantly put it in an early response to Stevens, 'the numbers do not know where they came from' (p. 751). The important thing is the *interpretation* of the results of the statistical analysis. It is here that the provenance of the numbers has to be considered, as well as other matters including the design of the study.

Exploring the data set

Frequency distributions and graphical displays

A simple means of exploring many data sets is to recast them in a way which counts the *frequency* (i.e. the number of times) that certain things happen and to find ways of displaying that information. For example, we could look at the number of students achieving different degree classifications. Some progress can be made by drawing up a *frequency distribution* as in Figure 17.2. This table can, alternatively, be presented as a *bar chart* (Figure 17.3).

Degree class	First	Upper second	Lower second	Third	Pass	Fail	Total
Frequency	9	64	37	30	7	3	150
Percentage	6	42.7	24.7	20	4.7	2	100

Note: 'Frequency' is the number of students with that degree class.

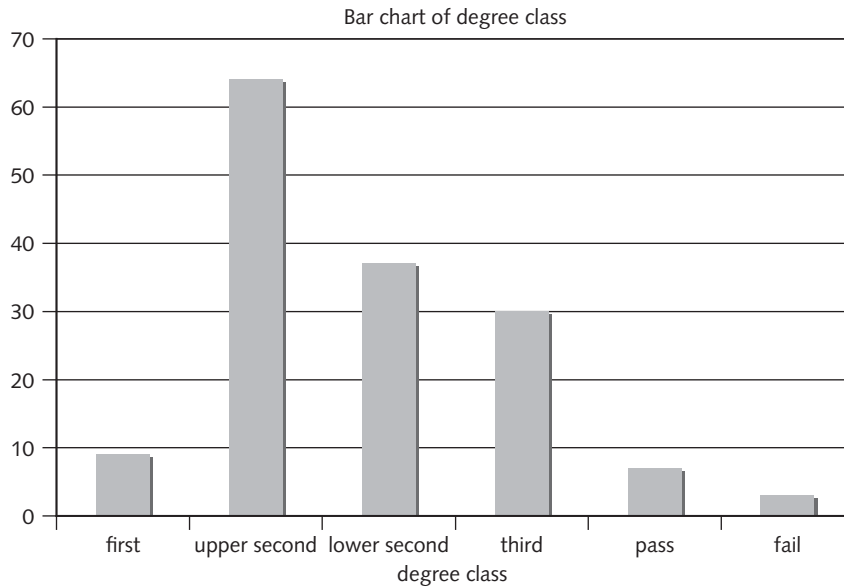**Figure 17.2** Frequency distribution of students across 'degree class'.

Figure 17.3 Bar chart showing distribution of students across 'degree class'.

The chart can be shown with either frequencies or percentages on the vertical axis; be sure to indicate which you have used. The classes of degree are ordered (here shown from first class 'downward' going from left to right). For some other variables (e.g. for faculties) the ordering is arbitrary. A distinction is sometimes made between histograms and bar charts. A bar chart is a histogram where the bars are separated from each other, rather than being joined together. The convention has been that histograms are only used for continuous variables (i.e. where the bar can take on any numerical value and is not, for example, limited to whole number values).

Pie charts provide an alternative way of displaying this kind of information (see Figure 17.4). Bar charts, *histograms* and pie charts are probably preferable ways of summarizing data to the corresponding tables of frequency distributions. It is claimed they are more quickly and easily understood by a variety of audiences – see Spence and Lewandowsky (1990) for a review of relevant empirical studies. Note, however, that with continuous variables (i.e. ones which can take on any numerical value, not simply whole numbers) both frequency tables and histograms may lose considerable detailed information. This is because of the need to group together a
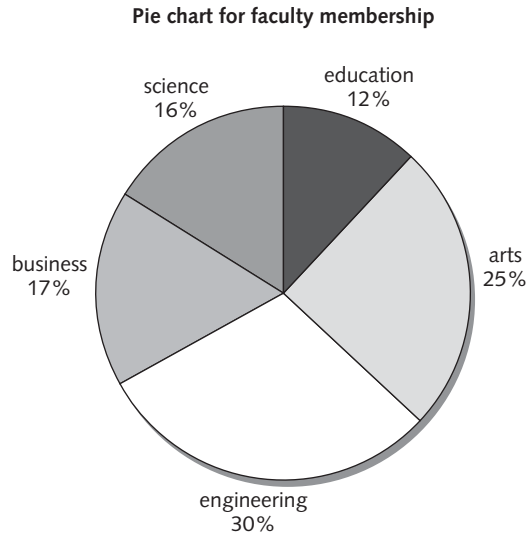
Pie chart for faculty membership

Figure 17.4 Pie chart showing relative numbers of students in different faculties.

range of values for a particular row of the frequency table or bar of the histogram. In all cases there will be a trade-off between decreasing the complexity of the display and losing information. An alternative EDA approach to displaying the data is the *box plot* (see p. 420).

Graphs (line charts) are well-known ways of displaying data. Excel, and statistical packages, provide ways of generating and displaying them although the quality of output many not be high enough for some needs. Specialized graphics packages (e.g. DeltaGraph, available from http://www.redrocksw.com) have a range of such displays available. Increasingly, professional standard displays are expected in presenting the results of projects, and apart from assisting communication, can help in getting over messages about the quality of the work. It is a matter of judgement whether or not any package to which you have access provides output of a quality adequate for presentation to a particular audience.

Marsh and Elliott (2008) give detailed, helpful and down-to-earth suggestions for producing numerical material clearly, in a section on 'Good Table Manners' (pp. 126–9). Tufte (2001) provides a fascinating compendium for anyone who needs to take graphical display seriously.

Summary or descriptive statistics

Summary statistics (also commonly known as *descriptive statistics*) are ways of representing some important aspect of a set of data by a single number. The two aspects most commonly dealt with in this way are the *level* of the distribution and its *spread* (otherwise known as *dispersion*). Statistics summarizing the level are known as *measures of central tendency*. Those summarizing the spread are called *measures of variability*. The *skewness* (asymmetricality), and other aspects of the shape of the distribution which are also sometimes summarized, are considered below in the context of the normal distribution (see p. 424).

Measures of central tendency

The notion here is to get a single figure which best represents the level of the distribution. The most common such measure to the layperson is the 'average', calculated by adding all of the scores together and then dividing by the number of scores. In statistical parlance, the figure obtained by carrying out this procedure is referred to as the *arithmetic mean*. This is because average, as a term in common use, suffers from being imprecise – some other more-or-less mid-value might also be referred to as average. There are, however, several other measures of central tendency in use, some appropriate for special purposes. Box 17.2 covers some of them.

Measures of variability

The extent to which the data values in a set of scores are tightly clustered or relatively widely spread out is a second important feature of a distribution for which several indices are in use. Box 17.3 gives details of the most commonly used measures. Several of them involve calculating *deviations* which are simply the difference between an individual score and the mean. Some individual scores are above the mean (positive deviations) and others below (negative deviations). It is an arithmetical feature of the mean that the sum of positive deviations is the same as the sum of negative deviations. Hence the *mean deviation* is calculated by ignoring the sign of the deviations, so that a non-zero total is obtained. The *standard deviation* and *variance* are probably the most widely used measures of variability, mainly because of their relationship to popular statistical tests such as the *t*-test and analysis of variance (discussed later in the chapter). However, Gorard (2006, pp. 17–19 and 63–73) makes a strong case for using the mean absolute deviation (i.e. ignoring the sign of the difference) rather than standard deviation, as it is simpler to compute, has a clear everyday meaning, and does not overemphasize extreme scores. This is part of his campaign in favour of 'using everyday numbers effectively in research'.

BOX 17.2

Measures of 'central tendency'

The most commonly used are:

- *Mean* (strictly speaking this should be referred to as the *arithmetic mean* as there are other, rarely used, kinds of mean) – this is the average, obtained by adding all the scores together and dividing by the number of scores.
- *Median* – this is the central value when all the scores are arranged in order of size (i.e. for 11 scores it is the sixth). It is also referred to as the '50th percentile' (i.e. it has 50 per cent of the scores below it, and 50 per cent above it).
- *Mode* – the most frequently occurring value.

   Note: Statistics texts give formulae and further explanation.

BOX 17.3

Measures of variability

Some commonly used measures are:

- *Range* – difference between the highest and the lowest score.
- *Midspread* or *inter-quartile range* – difference between the score which has one quarter of the scores below it (known as the 'first *quartile*', or '25th percentile') and that which has three-quarters of the scores below it (known as the 'third quartile', or '75th percentile').
- *Mean deviation* – the average of the deviations of individual scores from the mean (ignoring the sign or direction of the deviation).
- *Variance* – the average of the squared deviations of individual scores from the mean.
- *Standard deviation* – square root of the variance.
- *Standard error (SE)* – the standard deviation of the *mean* score.

  Note: Statistics texts give formulae and further explanation.

Statistics packages provide a very wide range of summary statistics, usually in the form of an optional menu of ways of summarizing any column within your data table.

Further graphical displays for single variables

It is possible to incorporate summary statistics into graphical displays in various ways.

Standard deviation error bars

A standard deviation error bar is a display showing the mean value as a dot, which has extending above and below it an 'error bar'. This represents one standard deviation unit above and below the mean. Typically, about two-thirds of the observed values will fall between these two limits (see the discussion of the normal distribution below).

This is often a useful way of displaying the relative performance of sub-groups, and more generally of making comparisons. A similar-looking display is used to show the *confidence intervals* for the mean. These are limits within which we can be (probabilistically) sure that the *mean* value of the population from which our sample is drawn lies: 95 per cent limits (i.e. limits within which we can be 95 per cent sure) are commonly used, but others can be obtained. Figure 17.5 shows both error bar charts for one standard deviation and 95 per cent confidence intervals.

Box plots and whiskers

Figure 17.6 shows the general meaning of the box and its upper and lower 'whiskers'. Note that the plot is based on medians and other percentiles, rather than on means and standard deviations.