Analysis or interpretation?

The traditional, and still widely used, terminology is to refer to the 'analysis' of data, whether quantitative or qualitative. Taken literally, analysis is a 'breaking up' of something complex into smaller parts and explaining the whole in terms of the properties of, and relations between, these parts. Not only is this, necessarily, a reductionist process but it is also seen by many as necessarily reliant on the particular form of statistical reasoning where hypotheses are based on probability theory applied to sampling distributions. This approach, discussed in Chapter 16, p. 446, has an important role when dealing with quantitative data from some experimental and other fixed designs. However, in real world research which generates quantitative data, it is rare to find that the rather restrictive design assumptions for the approach are met. The major research traditions in flexible design research are incompatible with the approach.

Interpretation carries very different conceptual baggage. Whereas the purpose of analysis is often seen as a search for causes (usually in the positivistic 'successionist' sense discussed in Chapter 2, p. 32), interpretation is considered to be about shedding light on meaning. This is a well-established view of the task when dealing with qualitative data, but Byrne (2002) makes a persuasive case for also focusing on the interpretation of quantitative data.

Quantitative and qualitative data – and their integration in multi-strategy designs

The following two chapters focus on the analysis and interpretation of quantitative and qualitative data, respectively. Multi-strategy (mixed methods) designs will have substantial amounts of both types of data for which the techniques and approaches in the chapters can be used. They can make separate contributions to the findings of the study but there is also a possibility of their integration to take full advantage of the opportunity provided by this type of design. The final section of Chapter 17 (p. 492) discusses some of the issues involved.

CHAPTER 16

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 generate at least some data in the form of numbers or which could not be sensibly turned into numbers of some kind. Hence, techniques for dealing with such quantitative data are an essential feature of your professional tool-kit. 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' (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 (p. 515). SPSS (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. 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. 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

The point has already been made several times that you should be thinking about how your data are to be analysed at the design stage of your project. This is important not only to ensure that what you collect is analysable but also to simplify as much as possible the actual process of analysis.

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.

BOX 16.1

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)

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 16.1 shows possible alternatives.

The conventions on coding are essentially common sense. Suggestions were made in Chapter 10 (p. 266) 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 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 respondent. A record consists of *cells* which contain data. The cells in a column contain the data for a particular *variable*. Figure 16.1 presents a simple example derived from a survey-type study. 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.

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

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 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 Figure 16.1 '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, p. 421) 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. 425).

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, *scatter plots* (see below, p. 433) 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 (2007). 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 (p. 425). 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 5 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 (e.g. Blaikie, 2003), 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 points out 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).

My 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 textbooks (e.g. Siegel, 1959), and still commonly found, is rejected by many mathematical statisticians (see Binder, 1984; Gaito, 1980). 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 16.2. This table can, alternatively, be presented as a *bar chart* (Figure 16.3).

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,

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 16.2: Frequency distribution of students across 'degree class'.

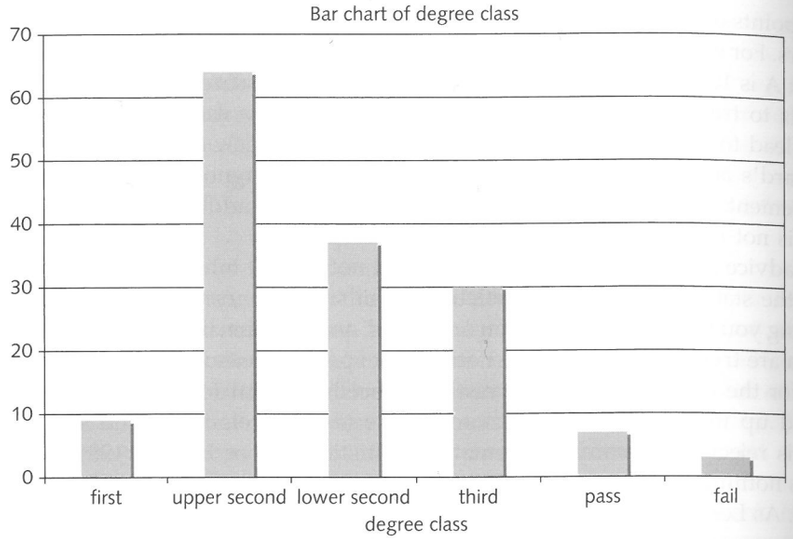

Figure 16.3: Bar chart showing distribution of students across 'degree class'

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 16.4). 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. My personal experience is that there are individual differences, with some people finding tables easier to understand. 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 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. 425).

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 www.redrocksw.com/deltagraph) 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.

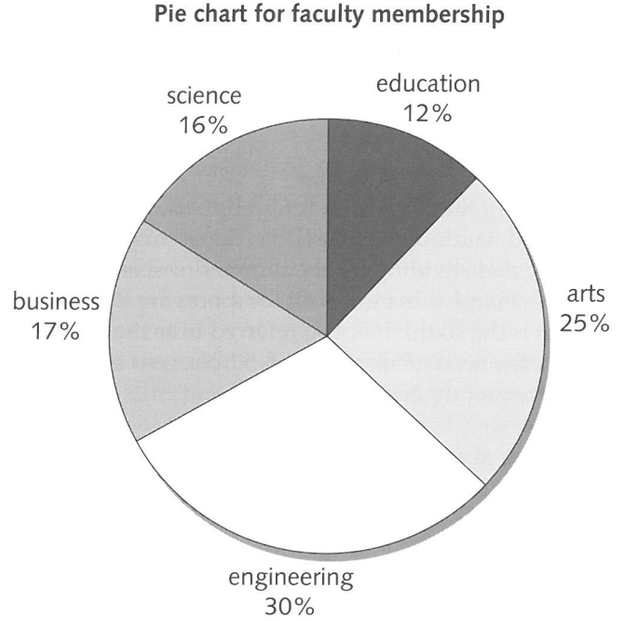

Figure 16.4: Pie chart showing relative numbers of students in different faculties.

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. 429).

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 lay person is the 'average', calculated by adding all of the scores together and then dividing by the number of scores. In statistical

BOX 16.2

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.

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 16.2 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 16.3 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 – see p. 449 and p. 452). However, Gorard (2006, pp. 17–19 and 63–73) makes a strong case for using the mean deviation rather than standard deviation, as it is simpler to compute, has a clear everyday meaning, and does not over-emphasize extreme scores. This is part of Gorard's campaign in favour of 'using everyday numbers effectively in research'.

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.

BOX 16.3

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* – the standard deviation of the *mean* score.

Note: Statistics texts give formulae and further explanation.

Further graphical displays for single variables

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

Standard deviation error bars

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, p. 429).

This is often a useful way of displaying the relative performance of subgroups, 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 16.5 shows both error bar charts for one standard deviation and 95 per cent confidence intervals.

Box plots and whiskers

Figure 16.6 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.

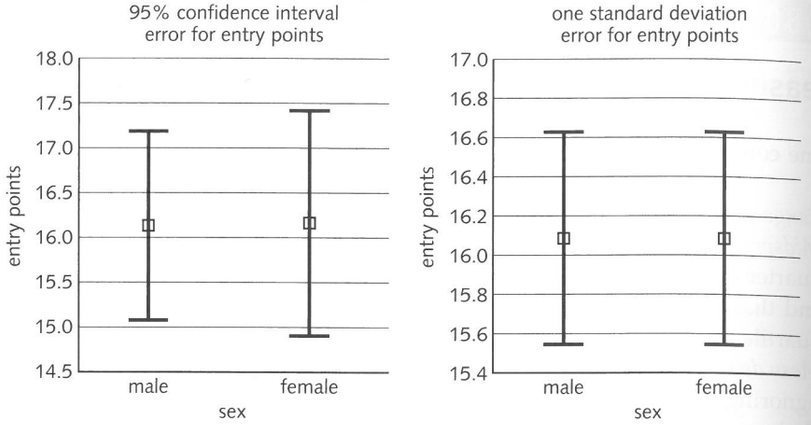

Figure 16.5: Display of error bar charts.

Outliers

The term *outlier* is commonly used for a value which is a lot higher, or a lot lower than the main body of the data. Marsh and Elliott (2008, pp. 168–71) suggest as a rule of thumb that values which are more than one and a half times the inter-quartile range ($Q_U - Q_L$) above the upper quartile, or more than one and a half times the inter-quartile range below the lower quartile, can be considered outliers. They term points as *far outliers* if they are more than three times the inter-quartile range above or below.

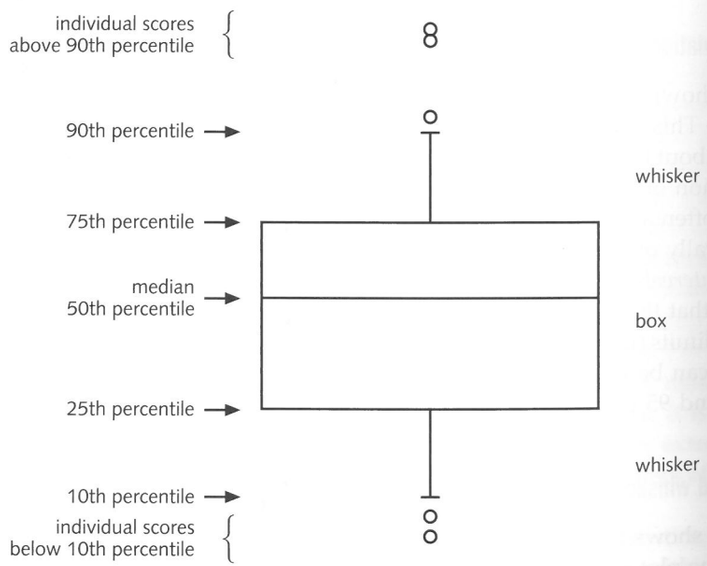

Figure 16.6: The 'box and whisker' plot.

Outliers call for special attention. They can arise for a variety of reasons. For example, an error might be made in entering the data set where separate entries of '2' and '7' get entered together as '27'. If no objective reason of this kind can be unearthed, then its treatment is problematic. Many statistical procedures are very sensitive to the presence of outliers. For example, one advantage of the median over the mean as a measure of central tendency is its lack of such sensitivity. EDA has been much interested in outliers, both in their own right, and in the study of measures which are robust (i.e. relatively unaffected) in their presence.

Manipulating the data

Marsh and Elliott (2008, p. 57) point out that 'data are produced, not given'. This stance rejects our classical heritage in the sense that the derivation of the word data is 'things given'.[1] The 'produced not given' point is important. Many of the data that we collect are actually produced during the research itself. They tend not to be things lying around that we pick up. We often have a very active hand, not only in what is collected, but in how it is collected. The actual numbers that we subject to analysis are very much formed by a process of selection and choice – at a very simple level, for example, do we use grams, kilograms, ounces, pounds, tons?

This basic choice will have been made at the time that the data are collected. In the example, this would probably now be metric in most countries, with the specific unit chosen to avoid very large, or very small, numbers (e.g. 5 grams rather than 0.005 kilograms; 2.3 kilograms rather than 2300 grams). There is still the possibility of manipulating the data subsequently, so that it is easier to analyse, or so that attention can be focused on features of interest, or so that it is easier to compare two or more sets of data. As in so many aspects of research, this process is driven by your research questions. Are there things that you can do with your data that can help give clearer answers to these questions?

It perhaps needs saying that this is nothing to do with 'How to Lie with Statistics' (Huff, 1991). 'Massaging' the data to give a biased or downright untruthful message should have no place in the kind of research covered in this book. The prime safeguard is your own honesty and integrity but this should be supported by detailed reporting of what you have done. Sufficient detail should be included to enable the sceptical reader to follow the trail from the collected data, through whatever you do to it, to the interpretation and conclusion.

Scaling data

The earlier section on descriptive statistics emphasized two aspects of a set of data: its *level* and its *spread*. The two simplest ways of scaling data involve these aspects directly.

[1] Also, in terms of its derivation the word is plural – one datum; two or more data. However, many people now use data as a singular noun. In a field where the term is used frequently, such as research reports, you may be perceived as ignorant of the 'correct' usage if you follow the popular trend. Not wanting to put you in that position, I propose to play the pedant and stick to the plural use.

Adding or subtracting a constant

A straightforward way of focusing attention on a particular aspect of the data is to add or subtract a particular constant amount from each of the measurements. The most common tactic is to subtract the arithmetic mean from each score. As discussed above in connection with measures of variability, scores transformed in this way are referred to as *deviations*. A similar tactic can be used when the median or some other measure of central tendency has been employed.

Multiplying by a constant

This is sometimes referred to as scaling or rescaling the variable. It is what you do when changing from weight in imperial measure (pounds, ounces, etc.) to metric (kilograms, grams). This tactic is particularly useful in comparing different sets of data which have initially been measured on different scales. For example, the prices of goods or services in the UK and other European countries could be better compared by transforming them all into the standard 'euro'.

Other transformations

There are many other things that you can do. *Taking logarithms* or *taking a power* (e.g. square, square root, reciprocal) are tactics commonly used when the distribution of the scores is asymmetrical or in some other way inappropriate for the type of statistical analysis proposed. Details are given in Marsh and Elliott (2008, Chapter 10).

Standardizing data

One way of manipulating data is very commonly used. It involves combining the two approaches covered above, i.e. subtracting a measure of level (central tendency) from an individual score, and then dividing by an appropriate measure of variability. The mean and standard deviation (or mean deviation) or median and mid-spread could be used. Distributions of scores that have been standardized in this way are much easier to compare, and in some circumstances combine, than unstandardized ones.

The normal distribution

The so-called *normal* (or *Gaussian*) distribution is a theoretical distribution of scores for which the shape is completely determined once the mean and standard deviation (SD) are known. Its shape is shown as Figure 16.7. Many distributions of scores obtained in practice are reasonable approximations to the normal distribution. To find if this is the case for a particular set of scores, they are first standardized as shown above and then scrutinized to see whether the proportion of cases falling at different distances from the mean are as predicted from tables showing the theoretical distribution.

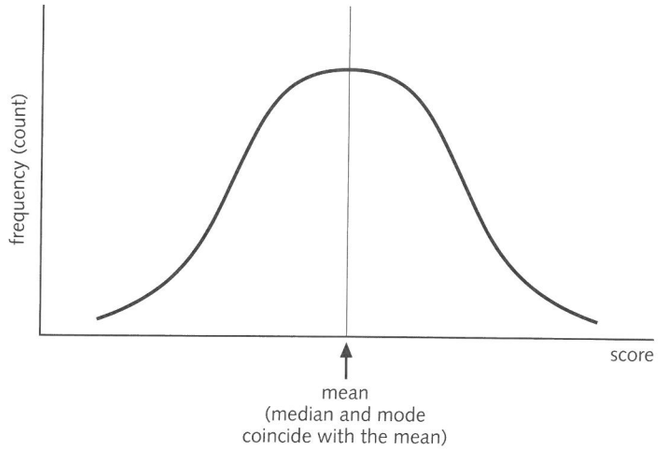

Figure 16.7: The theoretical 'normal' distribution.

For example, the expectation is that:

- 68 per cent of cases are within one SD of the mean;
- 95 per cent of cases are within two SDs of the mean; and
- 99.7 per cent are within three SDs of the mean.

Further details and appropriate tables are in many statistics texts (Robson, 1994, provides a simple account). It is possible to test the 'goodness of fit' of your data to the normal distribution by using a version of the chi-square test (see below, p. 431).

Whether or not a distribution of scores can reasonably be represented as normal is then of value in describing, summarizing and comparing data. However, don't fall into the trap of thinking that 'only "normal" is normal'. Data won't necessarily fall into this pattern. This is no major disaster; your job is to seek to understand what you have got, playing about with the scale if this seems to help. Such transformations may bring the distribution closer to normal but in itself that may not further your understanding.

The normal distribution also has a part to play if one wants to go on to carry out formal statistical tests on the data. Many of the more commonly used tests are based on the assumption that a normal distribution is involved. Often these tests are *robust* in the sense that deviations from normality do not appear to have much effect on the outcome of the test. However, there are 'distribution free' tests (commonly called 'non-parametric' tests) available (Higgins, 2003; Pett, 1997; Sprent and Smeeton, 2007) which do not make assumptions about the shape of the distributions involved.

Skewness

As can be seen from Figure 16.7, the normal distribution is symmetrical about its centre (which is where the mean, median and mode coincide). In practice, a distribution may be 'skewed' as shown in Figure 16.8. 'Negative' skew suggests that the majority of extreme

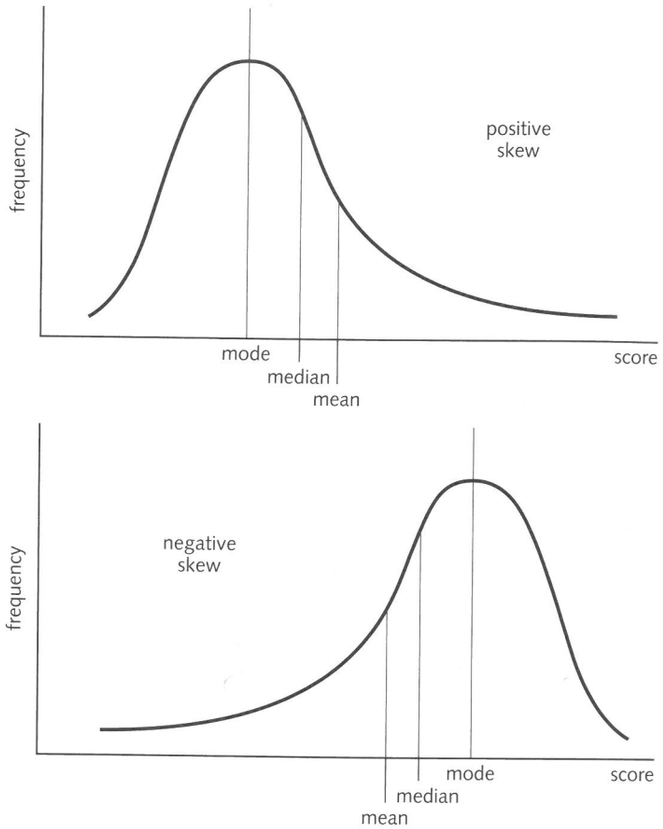

Figure 16.8: Positively and negatively skewed distributions.

observed values are less than the mean; 'positive' skew that the majority of extreme observed values are above the mean. A simple indication of this can be obtained by comparing the mean and median values. If the median is less than the mean, this suggests that over 50 per cent of the values are below the mean, and hence, to compensate, the right-hand or upper tail of the distribution must extend further – indicating positive skew. Statistical packages usually provide a measure of the skewness of a distribution. A normal distribution (being symmetrical) has a value of 0; positive values indicate a distribution with a long right tail, negative values a distribution with a long left tail.

Exploring relationships between two variables

Having considered how one might deal with individual variables, let us switch the focus to one of the main concerns in carrying out quantitative social research – looking for

relationships between variables. Here we will limit ourselves to relations between two variables.

To say that there is a relationship between two variables means that the distribution of scores or values on one of the variables is in some way linked to the distribution of values on the second variable – that, say, higher scores on one variable for that case (person, perhaps) tend to occur when there are higher scores on the second variable for that case. An example would be the relationship between smoking and lung cancer; those who smoke are more likely to develop lung cancer.

Cross-tabulation

Cross-tabulation is a simple and frequently used way of displaying a relationship between two variables. It is an extension of the use of frequency tables as discussed in connection with the analysis of single variables. Take once more the data on student intake presented in Figure 16.1. Let us say that we are interested in the relationship between faculty and the relative number of male and female students, i.e. between the variables 'faculty' and 'sex'. Figure 16.9 shows how these data could be presented in a 'contingency table'. There are five faculties (five levels of the variable 'faculty') and two sexes (two values of the variable 'sex') and hence 10 (five times two) possible combinations of levels of the variables. The boxes in the table, corresponding to each of these combinations, are referred to as *cells*. The total for each row and each column is given at the end or margin of the row or column. These totals are called the *row marginals* and *column marginals*, respectively.

The row total presentation shows the way in which females (and males) are distributed across the faculties labelled as 'counts'. The column total presentation shows the relative percentages (or proportions) of males and females in different faculties (e.g. the proportion of males in the science faculty). The contingency table, viewed in terms of percentages, helps to highlight any relationships between the two variables. Here the low percentage of females in the engineering faculty and high proportion in the arts is a striking, though unsurprising, feature.

Chi-square tests

Chi-square, in a contingency table, is a measure of the degree of association or linkage between the two variables. The more that there is a tendency for the relative number of males and females to vary from faculty to faculty, the greater is chi-square. It is based on the differences or discrepancies between the frequencies in the different cells (the

Male	Arts	Engineering	Business	Science	Education	Total
Female	3	29	9	10	4	55
Total	22	1	8	6	8	45
	25	30	17	16	12	100

Figure 16.9: 'Sex' by 'faculty' cross-tabulation.

		Faculty					
		Arts	Engineering	Business	Science	Education	Total
Male	Count	3	29	9	10	4	55
	Expected Count	13.8	16.5	9.4	8.8	6.6	55
Female	Count	22	1	8	6	8	45
	Expected Count	11.3	13.5	7.7	7.2	5.4	45
Total		25	30	17	16	12	100

Chi-square test result

Pearson chi-square 42.39
Degrees of freedom (DF) 4
$p < 0.05$ (exact p is 0.000 to three decimal places)
Note: No cells have an *expected* value of less than 5

Figure 16.10: Results of a chi-square analysis of the 'sex' by 'faculty' cross-tabulation in Figure 16.9.

'counts') and those that you would expect if there was no association at all between the two variables (i.e. the ratio of males to females is the same in all faculties). These latter are known as the 'expected' counts and are shown in Figure 16.10.

You will often see assessments of the *statistical significance* of relationships in contingency tables. This concept, and some of the problems in its use, are discussed later in the chapter on p. 446. It effectively tests the plausibility that a null hypothesis of no relationship is true. If the result you have obtained would be very unlikely if the null hypothesis were true it becomes reasonable to rule out the possibility that purely random factors are involved. If its probability is sufficiently small (conventionally taken as 1 in 20, i.e. $p = 0.05$), the relationship is taken to be due to some non-chance factor. The chi-square (χ^2) test is commonly used to assess the statistical significance of such relationships in contingency tables. The probability in this example is less than 0.0005. This is clearly very much smaller than the conventional 0.05 and hence statistically significant. 'Degrees of freedom (df)', refers to a somewhat esoteric statistical concept linked to the number of cells in the contingency table, which is used when assessing the statistical significance of the value of chi-square.

Statisticians warn against the use of chi-square when one or more *expected* frequencies fall below a particular value, usually taken as 5 in small tables. Fisher's exact test is a substitute which can be used in circumstances where the expected frequencies are too low for chi-square (see Pett, 1997).

A chi-square analysis, if statistically significant as in the present case, indicates that *overall* there is a relationship between the two variables (here 'faculty' and 'sex') which is unlikely to be explained by chance factors. In two-by-two contingency tables (where both variables only have two values) statisticians formerly used a somewhat different formula incorporating a 'correction for continuity' (sometimes referred to as 'Yates' correction') for computing chi-square. This is now considered to be inappropriate (Richardson, 1990). Some statistical packages provide both chi-square and a 'corrected' value when

analysing two-by-two tables producing an appropriately adjusted chi-square. You are recommended to ignore the corrected value.

Using chi-square to test for 'goodness of fit'

Chi-square can also be used to compare frequencies on a single variable to see how closely they 'fit' to those expected or predicted on some theoretical basis. A common theoretical expectation is for all frequencies to be the same; or perhaps it may be desired to test the goodness of fit to the frequencies expected if the data were normally distributed. The difference in terms of computation is that these expected frequencies have to be supplied, rather than being generated automatically from the observed frequencies.

Scattergrams

A *scattergram* (also known as a *scatter plot*) is a graphical representation of the relationship between two variables. It only makes sense when it is possible to order the values for each of the variables in some non-arbitrary manner. Hence in the data set of Figure 16.1 it would be reasonable to draw a scattergram for, say 'degree class' against 'entry points' but not for 'faculty' against 'entry points'. This is because any particular ordering of the faculties along an axis is arbitrary, and the apparent graphical relationship between the variables will vary with the ordering. Figure 16.11 presents a scattergram showing the relationship between 'entry points' and 'income' for a sample of graduates. It shows the position of each person on the two variables. For example, the far right point on the scattergram corresponds to someone who gained 26 entry points and has an income of about £18,000.

Scattergrams are a powerful pictorial device, giving a clear picture of the nature and strength of the relationship between the variables. They have their limitations, however. Many types of data are not readily amenable to display in this way, particularly when there are very few values on one or both of the variables. Nevertheless, unless you have data where the ordering of values is arbitrary, you should always consider the feasibility of drawing a scattergram for two-variable data. It is possible to produce contingency tables from the same data, summarizing by taking appropriate intervals along the variables when they take on many values.

Correlation coefficients

Measures of correlation (i.e. of the co-relationship between two variables) are referred to as *correlation coefficients*. They give an indication of both the strength and the direction of the relationship between the variables. The commonly used coefficients assume that there is a linear relationship between the two variables. Figure 16.12 demonstrates this in the idealized form of the 'perfect' linear correlation. However, perfection is not of this world. Certainly, you are very unlikely to get that degree of 'tightness' in the relationship, with data concerning humans and their doings. Figure 16.13 illustrates the kind of

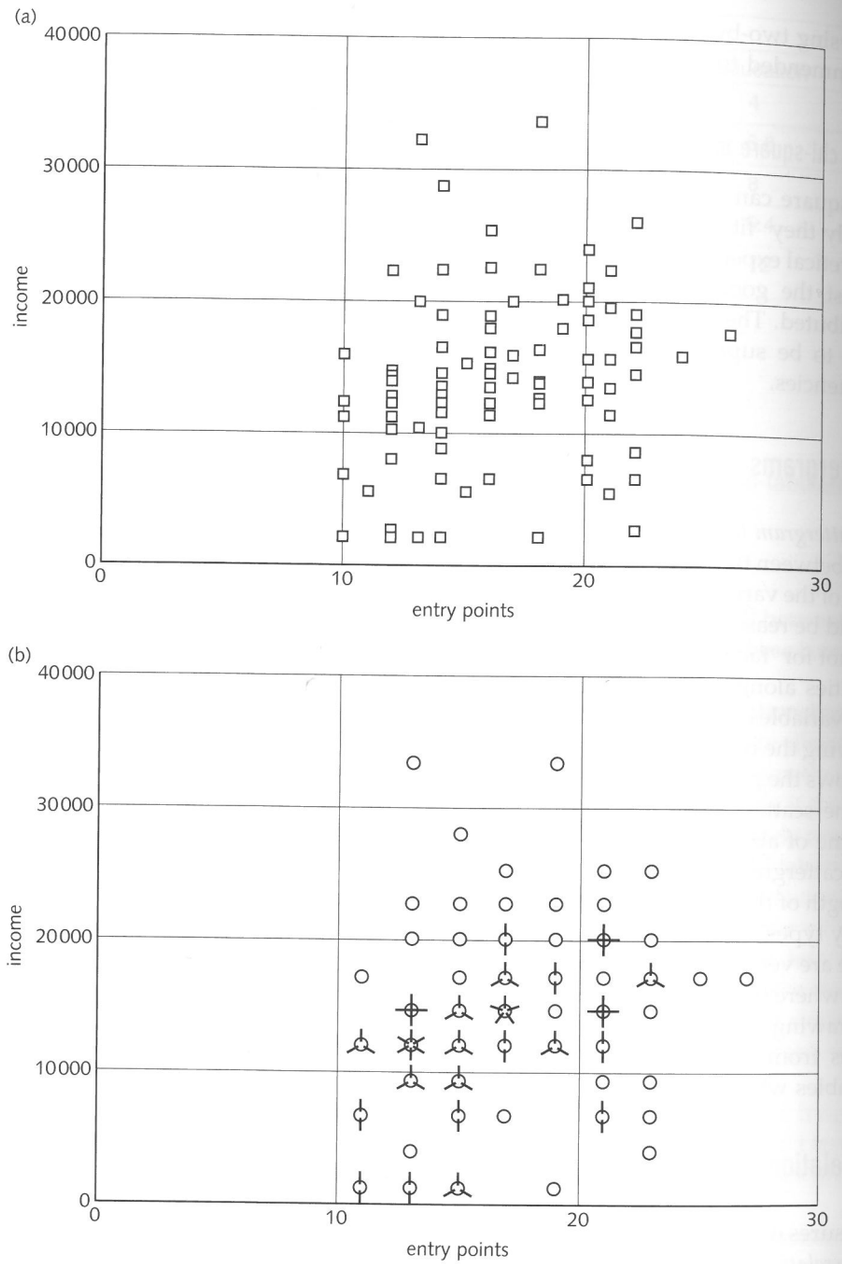

Figure 16.11: Scattergrams of 'entry points by income'.

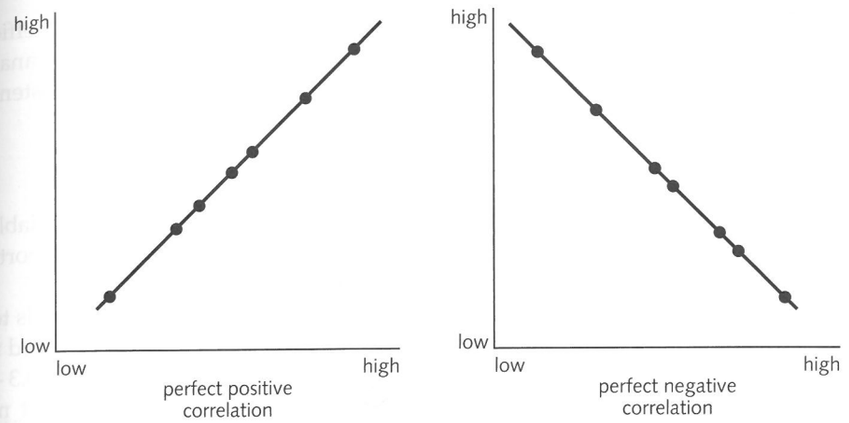

Figure 16.12: A 'perfect' linear correlation.

picture you are likely to see if there is a strong linear correlation. As you can see, the points fall within a cigar-shaped 'envelope'. The thinner the cigar, the stronger the relationship. With weaker correlations, the cigar is fatter; an essentially zero correlation shows no discernible pattern in the scattergram.

Commonly used correlation coefficients include Pearson's correlation coefficient (r), the Spearman rank correlation coefficient (known as Spearman's rho – ρ and Kendall's rank correlation coefficient (known as Kendall's tau – τ). As their labels suggest, the latter two are used with data in the form of ranks, or orderings, of data (what is first, second, etc.). The data may have been collected in this form, perhaps through participants expressing their preferences for different objects or situations, or may have been collected in other forms and subsequently converted into ranks. They do not assume normal distribution of the data and hence may be used when that assumption, on which the Pearson's coefficient is based, is dubious. They are, however, measures of linear

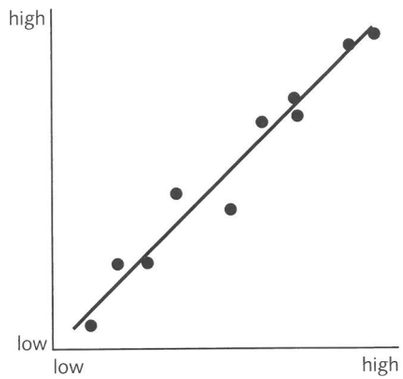

Figure 16.13: Example of a high positive correlation.

correlation (see below). The Spearman coefficient is effectively a Pearson coefficient performed on the ranks and is preferred by some on that ground, but most analysts appear to prefer Kendall's tau, possibly because it deals with ties more consistently.

Proportion of variance explained (PVE)

While the correlation coefficient is a measure of the relationship between the variables, it is difficult to assess the strength of this relationship (real 'significance' or importance rather than statistical significance) from the correlation coefficient.

The square of the correlation coefficient (r^2) is a useful index as it corresponds to the proportion of the variation in values of one of the variables which can be predicted from the variation in the other variable. Broadly speaking, if this is low (say less than 0.3 – but this will depend on circumstances) then it is unlikely to be profitable to exert much further time and effort in investigating the relationship. High values might suggest carrying out a subsequent regression analysis (see below).

Measuring the statistical significance of a correlation

The statistical significance of correlation coefficients is commonly computed. This concept and some of the problems in its use are discussed below (p. 446).

It is important to appreciate that the size of correlation coefficient which reaches a particular statistical significance (conventionally $p = 0.05$ being taken as the largest acceptable probability for this type of significance) is very strongly affected by the size of the sample of data involved. Thus for 20 pairs of scores the value of the Pearson correlation coefficient is 0.44 (two-tailed test – see below, p. 449); for 50 it is 0.28; for 100 less than 0.2; and for 500 less than 0.1. This illustrates the point that statistical significance has little to do with significance as commonly understood. Certainly, with a large sample such as 500, you can achieve statistical significance when less than 1 per cent of the variability in one variable is predictable from variation in the other variable; 99 per cent comes from other sources!

The message is that if the statistical significance of a correlation is to be quoted, make sure that both the size of the correlation (and/or of its square as a measure of the proportion of variance explained) and the size of the sample are also quoted.

Non-linear relationships between variables

It is perfectly possible to have some form of non-linear relationship between two variables. One value of the scattergram is in highlighting such non-linearities, in part because they are likely to call for discussion and explanation. They should also give a warning against using statistical techniques which assume linearity. *Curvilinear relationships* might be found. The envelope, instead of being cigar shaped, might be better represented by a banana or boomerang, as in Figure 16.14.

This is one situation where the data transformations discussed earlier in the chapter (p. 427) may be of value, as the appropriate transformation might convert the relationship in Figure 16.14 to something closely approaching linearity – and hence more

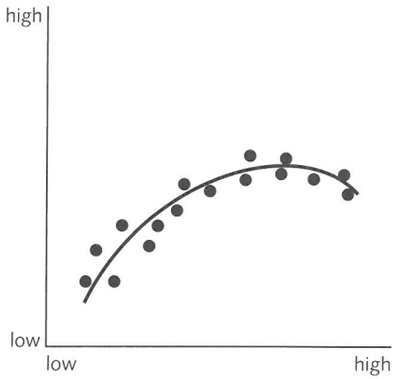

Figure 16.14: Example of a curvilinear relationship.

amenable to statistical analysis. Even if this transformation does 'work' in that sense, there may be consequent problems of interpretation. To know that there is a strong linear correlation between one variable and, say, the square of another variable may be of descriptive and even predictive value, but defy your attempts at understanding. However, finding that a reciprocal transformation works such that a non-linear relationship involving 'time elapsed' as one variable becomes linear when a 'rate' measure (i.e. reciprocal of time) is used, may well be readily interpretable.

Lines of 'best fit'

It is possible to draw a line of best fit on a scattergram. This can be estimated by drawing a line having roughly equal numbers of points above and below it, and making each point as near to the line as possible (using the minimum, i.e. perpendicular, distance from the line in each case).

There are systematic means of drawing such a line, which should be employed if it is to be used in any formal way. One approach which is commonly used is *linear regression*. This involves finding the line for which the squared deviation of individual points from the line (in the vertical, i.e. the Y dimension) is a minimum. This can be routinely performed by many computer packages, including 'analyse-it'. There are alternative ways of deriving these lines (see, for example, Marsh and Elliott, 2008, pp. 196–204, who advocate 'resistant lines'). When data are 'well behaved' (reasonably normal distributions with no problematic 'outliers'), linear regression is probably preferable, if only because of the ease with which the task can be completed.

The 'line of best fit', when obtained by one of the above means, is a powerful and useful way of summarizing the linear relationship between two variables. All straight lines can be expressed by a simple algebraic formula, one form of which is

$$Y = bX + a$$

Where Y and X are the two variables (conventionally, when there are dependent and independent variables, Y is the dependent variable and X the independent variable); and a and b are constants which typify the particular line of best fit. The constant a is known

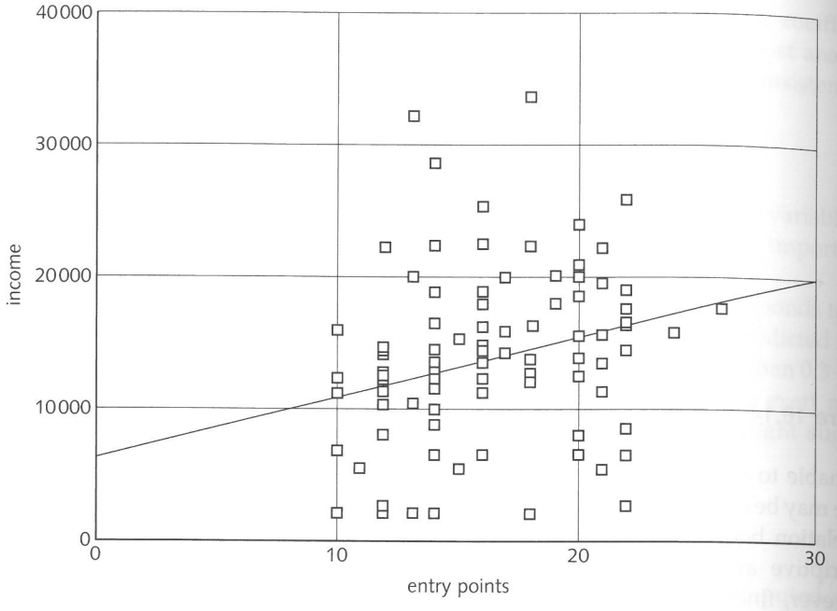

Figure 16.15: Fitting a 'regression line' for relationship between 'entry points' and 'income'.

as the *intercept* and is the point where the line cuts the vertical or *Y* axis; *b* is known as the *slope*. This is shown diagrammatically in Figure 16.15.

In addition to providing an elegant way of summarizing the data, the line of best fit (or the coefficients *a* and *b*, which amount to the same thing) can be used for predictive purposes, for example, to give an estimate of the likely increase in income over a specified number of years.

There is a difficulty with the data in that the amount of variability of the points around the regression line is not constant. It appears to increase with higher values of entry points and income. This not uncommon feature goes by the somewhat fearsome name of *heteroscedasticity* and, strictly, violates one of the assumptions on which Pearson's correlation coefficient is based. Again, this is a situation where possible transformations of the data might be attempted.

Exploring relationships among three or more variables

Research designs often involve more than two variables, calling for different approaches to those covered already. Multivariate techniques which are concerned with the *joint* effects of multiple variables are covered later in the section. We will, however, first consider approaches which look at the effects of taking into account a third variable on the relationship between two variables.

Three-variable contingency tables: the 'elaboration' approach

With non-experimental strategies, it is often essential to explore the effects of other variables when seeking to understand the basis of a two-variable relationship. One set of techniques is known as *elaboration analysis*. It has been widely used in the analysis of data from surveys and other non-experimental designs. Marsh (1982, pp. 84–97) gives a very clear account of the underlying logic. It involves the following steps:

1. Establish a relationship between two variables.
2. Subdivide the data on the basis of the values of a third variable.
3. Review the original two-variable relationship for each of the subgroups.
4. Compare the relationship found in each subgroup with the original relationship.

The third variable is referred to as the *test variable* (or *control*) *variable*. The original relationship between the two variables, where the third variable is not being held constant at a particular value, is called the *zero-order relationship*. The relationship that is found for a particular value of the test variable is known as a *partial relationship*. See de Vaus (2002, pp. 297–317) for an account of the statistics involved.

The pattern of effects of the test variable on the zero-order relationship can help in interpreting and understanding what is going on.

The website gives details of the interpretation of various patterns obtained when exploring relationships among three variables.

This approach to data analysis is simply a somewhat more complex version of the use of contingency tables, which were covered earlier. It provides a way of testing out and possibly modifying the conceptual framework you developed when designing the study (see Chapter 4, p. 72). Or, in other words, identifying the causal links. In realist terms, this amounts to specifying which mechanisms are in operation. The real world is complex and analysis may well not generate clear-cut patterns. In practice it is likely that multiple causation is the norm for many of the phenomena which interest us and that models which allow for multiple independent variables are to be preferred (see below).

It is possible to extend this type of analysis to four or even more variables (i.e. to two or more test variables) but it rapidly becomes unwieldy, particularly when there are several categories on each variable. Unless large amounts of data have been collected, the database for each subgroup becomes very small. The choice of test variables for elaboration analysis is obviously of central importance as it is only possible to include very few of them. They have to be pre-specified at least to the extent that you have collected the necessary data (some of the variables on which data have been collected can, of course, be omitted at the analysis stage). The message for small-scale studies is to keep the conceptual model simple. As repeatedly emphasized in Chapter 5, if you are carrying out fixed design research with a view to understanding and explaining a phenomenon, you don't do this unless and until you have established a clear and simple conceptual framework.

Using partial correlations

Essentially the same type of logical analysis can be carried out using partial correlation coefficients, rather than proportions in contingency tables. This amounts to examining the correlation between two variables and then seeing how, if at all, it changes when one or more other variables are held constant.

In the three-variable case, the correlation matrix is first calculated, which gives each of the three possible correlations between the variables. A partial correlation matrix is then calculated. Interpretation of the relationship between the variables is based on the pattern of correlations and the logical potential link between the test variable and the two original variables (e.g. antecedent or intervening).

The partial correlation approach cannot be used when testing for a 'moderated' relationship (i.e. there is an interaction in the sense that the relationship between them is influenced by a third variable) because this depends on comparing the relationship for different categories of the test variable and the partial correlation effectively gives you a single averaged figure. There are also problems in computing the correlation coefficients if one of the variables has a small number of categories or values.

Multiple regression

Multiple regression is multiple in the sense that it involves a single dependent variable and two or more independent variables (or, in the terminology more commonly used in non-experimental research, a single response variable and more than one explanatory variable). It is a flexible, widely used approach which has been made readily accessible through computer packages.

Taking the simplest possible case for multiple regression of one dependent variable and two independent variables, the regression equation is

$$y = a + b_1x_1 + b_2x_2$$

where y is the dependent variable, x_1 and x_2 are the two independent variables, a is the intercept, and b_1 and b_2 the regression coefficients for the two independent variables. The regression coefficient gives you the change in the dependent variable for each unit change in that independent variable, *with the effect of any of the independent variables controlled* (referred to as 'partialled out').

While multiple regression can be used in the same way as linear regression, to give a line of best fit and to provide predictions through substitutions of different values of x_1 and x_2, its main use is to provide an estimate of the relative importance of the different independent variables in producing changes in the dependent variable. To do this, it is necessary to convert the regression coefficients to allow for the different scales on which they have been measured. When this is done, they are referred to as *standardized regression coefficients* or *beta weights*. They then tell you how many standard deviation units the dependent variable will change for a unit change in that independent variable.

The output from a statistical package will provide various statistics which may include:

- *R-squared*. This is the *multiple coefficient of determination*, a measure of the proportion of the variance in the dependent variable which is explained by the independent variables in the equation. If, for example, R^2 is 0.14, the proportion of variance explained is 14 per cent. An 'adjusted R^2' may also be produced. This will be smaller than R^2 and is adjusted in the sense that it takes into account the number of independent variables involved and would normally be preferred to the unadjusted value.
- *t-value of coefficients*. This presents a test of whether or not the associated beta coefficient is significantly different from zero. A probability value will usually be given in each case.
- *Standard error of coefficients*. This is a measure of the accuracy of the individual regression coefficients. This information is useful in assessing the likely accuracy of predictions based on the regression equation.
- *Analysis of variance (ANOVA) tables*. Discussed later in the chapter, p. 452.

This discussion merely scratches the surface of multiple regression and its possibilities. If a major concern is in *developing* a model effectively in deciding on an appropriate regression equation, then an option known as *stepwise regression* is worth considering. This starts with the simplest possible model and then step by step examines the implications of adding further independent variables to the equation.

If you already have an explicit model which you are testing, *hierarchical* (or *stepwise*) *multiple regression* is preferable. This involves entering the variables into the analysis in an order determined by your model.

You are strongly recommended to seek advice when considering using multiple regression, as not only is it complicated but also it is particularly easy to do something silly and inappropriate with the packages available. It is worth noting, however, that multiple regression can be used with a wide variety of types of data. In particular, it can be used with categorical variables such as 'gender' and 'faculty' in the example we have been using. A difficulty here is that the ordering of categories is essentially arbitrary for such variables, and particularly when there are more than two categories for a variable, the ordering chosen would affect the result obtained. This can be handled by the use of so-called 'dummy variables'. It involves coding particular categories as 'present' (say coded '1') or absent (say coded '0'). Cramer (2003, Chapters 10–12) gives details. Alternatively *logistic regression* may be used for both continuous and categorical data as long as you have (or can produce) a categorical dependent variable (see Field, 2005, Chapter 6, for a thorough discussion).

Multivariate exploratory techniques

Strictly speaking, these involve more than one dependent or response variable and possibly additional explanatory variables (Ryan, 2008). This excludes multiple regression although it is commonly referred to as a multivariate technique. There is a wide variety of different exploratory techniques designed specifically to identify patterns in multivariate data sets of which factor analysis has been the most widely used.

Exploratory factor analysis

Factor analysis is an approach to making sense of a large number of correlations between variables. It has similarities with regression analysis but differs in that the variables all have equal status; no single variable is designated as the dependent or criterion variable. Factor analysis starts with a matrix of correlations.

Matrices of this type, particularly when they contain up to 100 or so variables, are very difficult to interpret. Factor analysis aids in this process by pointing to clusters of variables which are highly intercorrelated. The 'factors' referred to are hypothetical constructs developed to account for the intercorrelations between the variables. Factor analysis seeks to replace a large and unwieldy set of variables with a small and easily understood number of factors. Suppose that your correlation matrix arises from a 50-item questionnaire on aggression and aggressiveness. You find, say, that there are strong intercorrelations between 12 items concerning aggression towards family and friends, and similar intercorrelations between nine items concerning aggressiveness towards people in authority, but no other strong clusters. This then provides good evidence that two factors are important in understanding your results.

The technique is commonly used in the development of tests and scales (see, for example, Loewenthal, 2001). It allows you to assess the extent to which different test items are measuring the same concept (strong intercorrelations) or whether their answers to one set of questions are unrelated to their answers on a second set. Hence we get an assessment of whether the questions are measuring the same concepts or variables.

Factor analysis is typically used as an exploratory tool. There is an alternative version referred to as 'confirmatory factor analysis' – see below. Exploratory factor analysis starts with the correlation matrix. For it to be worthwhile to carry out the analysis, the matrix should show a substantial number of significant correlations (either positive or negative).

The number of respondents should exceed the number of variables. When the interest is not simply to describe the factors summarizing the relations between variables, but to try to get a reliable estimate of these underlying factors, then minima of five times the number of participants to the number of variables have been suggested. There are many versions of factor analysis including canonical, alpha, image and maximum likelihood factoring, but the most commonly used are *principal-components analysis* (strictly speaking a form of regression analysis) and *principal-axis factoring* (sometimes simply referred to as 'factor analysis'). Accounts of the process are found in specialized texts such as Child (2006), Kline (1993) and Loewenthal (2001).

SPSS provides most factor analysis options you may need. Bryman and Cramer (2008) provide details of the procedures to be followed and the kinds of output obtained when SPSS is used to carry out principal-components and principal-axis analyses. Brief details of other multivariate techniques which can be used in an explanatory mode are given below. Further details and examples that help in understanding the situations in which the techniques can be used are given in the Electronic Statistics Textbook, available at www.statsoft.com/textbook/stathome.html. Hill and Lewicki (2006) is the corresponding paper version.


The website includes a short discussion of additional multivariate exploratory techniques.

Model-testing multivariate techniques

Although factor analysis is usually employed primarily as an exploratory tool, it can be used to assess the extent to which the solution obtained matches a hypothesized pattern and is therefore useful when we are testing a model or conceptual structure when it is referred to as *confirmatory factor analysis* (Brown, 2006). Other multivariate approaches are specifically targeted at model testing. They include:

Path analysis

Sometimes referred to as *causal modelling*, the central idea in path analysis is that, if we can safely make assumptions about the chronological order of variables in our research, then we can use partialling of variance techniques from multiple regression to test models about the interrelationship of these variables. Causal models can only be built from what are effectively correlational techniques if other assumptions as well as the temporal assumptions about the effect of the variables apply. While this is a tool of considerable potential, the necessary assumptions are highly restrictive and there are many problems of interpretation (Olobatuyi, 2006).

Structural equation modelling (SEM)

This is a multivariate data analysis technique which combines elements of both multiple regression and factor analysis. Its goal is similar to that of factor analysis through providing a summary of the interrelationships among variables. It is similar to path analysis in that researchers can test models in the form of hypothesized relationships between constructs. A model can be tested statistically in a simultaneous analysis of the entire system of variables to assess its fit to the data. SEM has become increasingly popular in psychology and the social sciences for the analysis of non-experimental data and has also been quite intensively used in applied field such as social work research (e.g. Guo, Perron and Gillespie, 2009).

Issues of interest in real world research are often complex and multidimensional in nature. However, until relatively recently researchers were dissuaded from working with complex research questions because the statistical techniques available did not easily allow for testing of multivariate models. Weston and Gore (2006), in an accessible brief guide to SEM, cite several applied examples where this approach has been valuable. For example, Tylka and Subich (2004) hypothesized that eating-disorder patterns in adult women were a function of personal, socio-cultural and relational factors. Using SEM they tested a multidimensional model about how these factors interact in complex ways to explain symptom severity and showed its close fit to the data. Using a similar approach, Long (1998) successfully tested a model of workplace stress and coping for employed women, which included constructs such as human agency, status, coping, work–environment demand and distress, across levels of employment prestige.

It is important to stress that, just as with techniques discussed above, such as correlation, multiple regression and path analysis, causality cannot be determined by

the results of an SEM analysis – this is a judgement based on the adequacy of your underlying theory and research design. Also that, as Weston and Gore (2006) point out:

> researchers can easily misuse SEM. Just as researchers are free (although not encouraged) to conduct several different multiple regression models until they find a model to their liking, they can also analyze models in SEM, identify and remove weaknesses in the model, and then present the revised model as if it were the originally hypothesized model. Most users would likely agree that SEM's true power lies in the fact that researchers must specify complex relationships a priori and then test whether those relationships are reflected in the sample data. Optimally, researchers will draw these hypothesized relationships from previous research or theory and will present the results with integrity. Should the researcher detect weaknesses in the proposed model, he or she should further explore them using a modified model in a new sample (p. 733).

If you are considering using SEM, Blunch (2008) and Kline (2004) provide accessible introductions. Much of its popularity can be traced to the development of increasingly accessible and reliable software packages which include LISREL, EQS and AMOS. Schumacker and Lomax (2004) provide a clear non-technical introduction covering each of these. However, SEM is complex statistically and calls for a great deal of judgement on the part of the researcher to avoid the misuse and misinterpretation of the results of analysis. It depends on an evaluation of multiple test statistics and indices to determine whether the model fits the data adequately and there is considerable controversy about what constitutes an acceptable fit. Weston and Gore (2006) provide a set of guidelines on how to avoid the pitfalls.

Analysing differences

So far, this chapter has focused on displaying, describing and summarizing quantitative data and on analysing relationships among data. We now turn to what has traditionally been viewed as the major task when analysing quantitative data. Are there *differences* between the scores, values or observations obtained under one condition and those obtained under another condition (or conditions)?

Looking for differences and looking for relationships are really two ways of viewing the same thing. Asking whether there are differences between the performance of three groups taught by different methods is essentially equivalent to asking whether there is a relationship between teaching method and performance.

It is to answer questions of difference that many of the *tests of statistical inference* have been developed. The basic logic behind such tests is not difficult, although its working out in specific tests can be complex. The test is commonly used to make decisions about the state of affairs in some 'population' as compared with the actual sample of scores or observations that we have obtained. For example, suppose we want to find out whether the ratio of men to women in a particular sample is such that we can consider it

representative of a specific population where the ratio of men to women is known. If there is a 50-50 split in the population but there are no women in a randomly selected sample of 20 people from that population, then common sense might be unwilling to regard this as a representative sample and perhaps cast doubts upon the randomness of the procedure used to select the sample. However, even if we decide that the sample was not drawn from the 50-50 population, we could be wrong. A sample of 20 consisting of 20 women is in fact just as likely to occur as any other specific sample (the analogy often drawn is with tossing coins – it is possible to have a sequence of 20 heads in a row and that sequence is just as likely as any other specific sequence such as HTTHTTTHHTHH THTTTTHH). There are, however, many possible ways in which one could end up with, say, 11 males and 9 females, but in fact only one sequence which gives all 20 females. It is then possible to come to the decision (based on probabilities) that the sample didn't come from the population when it in fact did, an error known as a *type one error*.

Statistical tests provide ways of assessing this type of error. This is where the term *statistical significance* rears its head. In situations like the one discussed, it refers to the probability of making a type one error (given the symbol alpha – α). The convention has also been mentioned of setting this at a probability of 0.05 (i.e. 5 per cent or 1 in 20). However, the fact that many computer programs typically generate exact probability figures for the chance of making a type one error, rather than saying that it is 'less than 0.05', means that there is an increasing tendency for such exact probabilities to be quoted.

There is also a *type two error*: that is, the probability of deciding that the sample came from the population when in fact it did not (given the symbol beta – β). There is an inverse relationship between the two types of error, in the sense that we can reduce our chances of making a type one error by setting the significance level at a very low probability (say 0.001, or 1 in 1000). However, setting the decision line at this point produces a corresponding increase in the chances of making a type two error.

The *power* of a statistical test is the probability that the test will reject a false null hypothesis (that it will not make a type two error). As power increases, the chances of a type two error decrease. The probability of a type two error is referred to as the false negative rate (β). Therefore power is equal to ($1 - \beta$). Cohen (1962) called attention to the poor statistical power of much published research – a situation which has not changed radically over 40 years later (see Cashen and Geiger, 2004).

Conventional statistical tests assume that we are dealing with randomly selected samples from known populations. This is rarely the case in experimental research. As discussed in Chapter 5, true experiments pay careful attention to the random *assignment* of a set of participants to different experimental treatments or conditions. However, it is relatively rare that serious attention is given to the random *selection* of these participants to ensure that, probabilistically, they are representative of known populations. Experimenters typically work with people they can persuade to take part, effectively a convenience sample.

This approach is justified by claiming that the interest is in testing for possible differences between the two conditions rather than in generalizing to the populations from which the samples are drawn. However, the theoretical basis to many of the statistics used does assume random selection. The general issue of statistical testing with non-random samples is discussed later in the chapter (p. 457).

Statistical significance

For many statistically oriented social scientists, quantitative analysis is virtually synonymous with significance testing. The whole point and purpose of the exercise is taken to be 'have we got a significant result?' 'Is $p < 0.05$?' This refers to *statistical significance*. (I am grateful to a reviewer, Joe Maxwell, who pointed out the potentially misleading treatment of statistical significance in a draft version of an earlier edition of this text. The following discussion leans heavily on material which he kindly provided.)

The probability that a significance test gives you is *not* that a result is due to chance (as is commonly claimed). What a p value actually tells you is something that sounds very similar to this statement but is in fact quite different. It tells you how likely it *would be* that you would get the difference you did (or one more extreme), by chance alone, if there really is no difference between the categories represented by your groups, in the population from which you drew your sample. This assumption of 'no difference' is referred to as the 'null hypothesis'. In other words, a statistical significance test 'tests' the plausibility that the null hypothesis – no difference between the population means – is true. If your result would be very unlikely *if* the null hypothesis were true, this makes it less plausible that the null hypothesis *is* true.

Thus, the result of a statistical significance test tells you nothing directly about the actual population to which you want to make inferences; it simply helps you rule out one possible validity threat to your result, namely, that the result could be due to random variation in your sample, rather than to real differences in the population. If your p value is small rather than large, this makes it less likely that your result is due to chance variation rather than to a true difference, other things being equal. However, the 'other things being equal' is very important, because the actual likelihood that your result is due to chance is not completely expressed by the p value. Statistical significance tests say *nothing* about all the other possible validity threats to the result, or how likely these are relative to the proposed explanation. For example, suppose you pull a coin out of your pocket, flip it 10 times, and get 10 heads. This is an extremely unlikely occurrence (less than one chance in a thousand, or $p < 0.001$) if it is a fair coin (one that has an equal chance of coming up heads or tails – the null hypothesis). However, my judgement would be that it's still more likely that this particular result is due to chance than it is because the coin is biased. If it came up 50 times, the latter possibility becomes somewhat more plausible. Both explanations are unlikely, and if no other explanations can be put forward, then the more improbable it is that your result could have happened by chance if you have a fair coin. Hence, the more likely it is that the alternative explanation of bias, however implausible, is true.

Statistical significance testing is both deeply entrenched in practice and highly controversial. Meehl (1978) goes so far as to conclude that reliance on statistical significance was one of the 'worst things that ever happened in the history of psychology' (p. 817). Haig (1996) considers that: 'It is a major professional embarrassment that researchers continue to employ such tests in the face of more than three decades of damning criticism'. Perhaps the most swingeing criticism comes from Ziliak and McCloskey (2008) in their text *The Cult of Statistical Significance* who show convincingly, with evidence ranging from agronomy to zoology – taking in psychology, medicine and economics as major miscreants – how wide the disaster is and how bad it has been for the progress of science.

One problem, mentioned earlier in the chapter, is that statistical significance is not related to the size or importance of an effect or relationship, which is in many cases what we are really interested in. The chance of obtaining a statistically significant result increases as the sample size increases, because, for example, you then get a more sensitive test of any difference between the experimental and control groups in an RCT. But there is always likely to be some difference between the two conditions. Hence the common injunction to 'use a larger number of participants' may buy statistical significance at the expense of real life triviality. Paradoxically, if one is relying on statistical significance, there is much to be said for keeping the sample small so that only robust effects are going to be picked up.

Readers who wish to work out their own position on this controversy might review the interestingly titled *What if There Were No Significance Tests?* (Harlow, Mulaik and Steiger, 1997). See also Hagen (1997) and Gigerenzer, Krauss and Vitouch (2004) to get a flavour of the debate.

Measuring effect sizes

It would be helpful to use a statistic which is, unlike statistical significance, independent of sample size. When looking at the difference between the means of two sets of scores, this can be achieved by dividing the difference in means by the standard deviation in the population from which they come. So one obtains a difference expressed in standard deviation units; e.g. the difference in means is 0.6 standard deviations. The effect size is sometimes referred to as the *practical significance* of a result and it is not uncommon to find major discrepancies between this and statistical significance – a large effect size but lack of statistical significance; or the reverse (Alhija and Levy, 2009).

There are some underlying complexities. The population standard deviation is rarely known and a standard deviation estimated from the sample of scores available usually has to be substituted. Details of how this can be done are provided in texts concentrating on statistical analysis (e.g. Clark-Carter, 1997). There is also the issue of what constitutes a large enough difference to be taken note of. Cohen (1988) provides guidelines suggesting that a value of 0.2 is small, 0.5 is medium and 0.8 is large. The use of *confidence intervals*, as discussed on p. 425, is another possibility. Confidence intervals are routinely computed by statistical packages. Effect sizes can be derived from the information about means and standard deviations. Details are given in Dancey and Reidy (2007).

An alternative approach to measuring the size of an effect produced in a study involves evaluating the *proportion of variance explained* (PVE) by means of various statistics based on measures such as the square of the correlation between two variables. Rosnow and Rosenthal (1996) have suggested some convenient procedures for the compilation of both effect sizes and confidence intervals for several statistics.

A third, and more direct, approach to communicating the magnitude of the effect is to simply report the actual differences between the groups (as discussed below, p. 450). This is often more meaningful to practitioners and other non-specialists than the two previous approaches.

Power analysis

As discussed above, the power of a statistical test is the probability that it will correctly lead to the rejection of a false null hypothesis – the probability that it will result in the conclusion that the phenomenon exists. A statistical power analysis can be used either retrospectively (i.e. after the study has been carried out) or prospectively (i.e. before it has been carried out). A prospective analysis is often used to determine a required sample size to achieve a target level of statistical power, while a retrospective analysis computes the statistical power of a test given the sample size and effect size. Cohen (1988) provides a very convenient source for formulas and tables to compute power in a range of common analyses. There are several websites which will calculate these statistics for you (e.g. www.danielsoper.com/statcalc/default.aspx#c17; see 'sample size' for prospective analyses, and 'statistical power' for retrospective analyses).

Although there are no formal standards for power, a value of 0.80 is commonly quoted as acceptable (i.e. at least an 80 per cent chance of rejecting a false null hypothesis). Setting the bar at this level can cause virtually insurmountable problems in some research contexts. For example McDonald and Fuller (1998) illustrate the difficulty in studying black bear cub survival in the wild. Although their data represented over 10 years of data collection, they could not generate a sufficiently large sample size to adequately test a simple hypothesis with the design and analytical methods they used.

A common misconception is that power is a property of a study or experiment. Any statistical result that has a *p*-value has an associated power. Hence if you are carrying out several tests there will be a different level of statistical power associated with each one. Statistical power provides important additional information about the importance of non-significant test results that researchers should consider when drawing their conclusions. A non-significant result coupled with high statistical power to detect an effect of interest to the researcher increases confidence that the effect was not missed. On the other hand, a non-significant result coupled with low statistical power to detect the effect of interests suggests that another study, or more sampling, is required before strong conclusions can be drawn.

It is becoming common for funding agencies and boards reviewing research to request that prospective power analyses are carried out. The argument is that if a study is inadequately powered, there is no point in carrying it out.

> *Because significance testing is expected by many audiences, including sponsors and journal editors, it is advisable to give measures of statistical significance. However, because of the various criticisms, you should not rely solely on them. Providing additional information on the direction and size of the effect or relationship found in a study is highly recommended.*

Depending on the type of study this might be based on differences in means, correlation coefficients and/or regression coefficients (each of these is discussed later in the chapter in connection with different analyses). These are simply ways of summarizing aspects of the data, i.e. summary or descriptive statistics (p. 423). Hence they do not carry with them the positivistic conceptual baggage associated with some uses of significance testing. With a realist approach, statistical analysis is used to confirm the existence of mechanisms whose operation we have predicted in the contexts set up in an experiment or other study. Large effect sizes provide confidence in their existence; hence they are what you are looking for. Significance levels play a subsidiary role, their inclusion perhaps lacking something in logic but sanctioned by convention.

Practical significance indices provide information about the size of observed difference or relationship (e.g. effect size). Clinical significance measures provide data regarding the extent to which the intervention makes a real difference to the quality of life of the participants or to those with whom they interact.

Single-group tests

In most situations we are concerned with comparing the scores or values obtained under one condition with those obtained under another condition during the current project. However, you might want to compare what you have obtained with some expectation arising outside the study to see whether there is a difference.

Chi-square as a test of 'goodness of fit'

This test has already been mentioned (p. 433).

One-group t-test

The t-test is a very widely used method to compare two means. In this version, the comparison is between a mean obtained from the particular sample of scores that you have obtained under some condition, and a hypothesized population mean. Figure 16.16 summarizes the output of a one-group *t*-test.

Probability values for the statistical significance of *t*-test results (and for many other statistical tests) can either be 'one-tailed' or 'two-tailed'. 'Two-tailed' means that you are simply concerned with establishing the probability of a difference between the two

		n	Mean	SE	SD
	Entry points	33	20.9	1.11	6.4
	Hypothesized		18.6		
	Mean	20.9			
	SE	1.11			
	t statistic	2.06			
	DF	32			
	2-tailed *p*	0.0477			

Key: SE = standard error; SD = standard deviation; DF = degrees of freedom
As $p < 0.05$ the difference in means is statistically significant

Note: The output from 'analyse-it' also gives 95% CI (confidence interval) values.

Figure 16.16: Results from a one-group *t*-test.

means. With 'one-tailed' the hypothesis is that the difference will be in a particular direction (hence referred to as a 'directional' hypothesis). 'Two-tailed' probabilities should be selected unless there is a strong *a priori* reason (e.g. from your model or other theory about what is happening) for expecting the difference to be in a particular direction. Details of the means themselves should always be reported.

Two-group tests

Many of the questions we are interested in when carrying out a study producing quantitative data boil down to whether there are differences between the scores obtained under two conditions or by two groups. Do mature students admitted without standard entry qualifications get poorer degrees than 18-year-old standard entrants? Do patients suffering from lower back pain get better more quickly when treated by chiropractic methods than by drugs? And so on.

Two-group t-tests

The *t*-test is very commonly used to compare the means of two groups. It comes in two versions. The *paired two-group t-test* (sometimes called the *dependent samples t-test*) should be used when there are pairs of scores. This would be, for example, if the same person provided a score in each of the conditions. The *unpaired two-group t-test* (otherwise known as the *independent samples t-test*) is where there is no such basis for putting together pairs of scores. Figure 16.17 gives an example of output from an independent samples *t*-test, and Figure 16.18 that from a dependent samples *t*-test.

Gender	n	Mean	SE	SD
Female	12	119.3	6.20	21.5
Male	7	101.0	7.79	20.6
Mean difference	18.3			
SE	10.07			
t statistic	1.82			
DF	17.0			
2-tailed *p*	0.0864			

Key: SE = standard error; SD = standard deviation; DF = degrees of freedom
As $p > 0.05$ the difference in means is not statistically significant

Notes: (i) The output from 'analyse-it' also gives 95% CI (confidence interval) values. (ii) The example shows that it is possible to have unequal numbers in the two conditions. However, statistically, it is preferable to have equal sample sizes.

Figure 16.17: Results from an independent samples *t*-test.

Score	n	Mean	SE	SD
First measure	12	103	3.96	13.7
Second measure	12	107	4.09	14.2
Difference (first – second)	12	−4	2.58	8.9
Mean difference	−4			
SE	2.58			
t statistic	−1.55			
DF	11			
2-tailed *p*	0.1492			

Key: SE = standard error; SD = standard deviation; DF = degrees of freedom
As **p** > 0.05 the difference in means is not statistically significant

Note: The output from 'analyse-it' also gives 95% CI (confidence interval) values

Figure 16.18: Results of a dependent samples *t*-test.

Recall, once again, that it is good practice, when recording the results of such tests, to include not only the *t*-value and its statistical significance (the probability value, which must be lower than 0.05 for conventional statistical significance) but also the means and standard deviations of the two sets of scores. A minus sign for the value of *t* is of no importance and does not affect its significance; it simply indicates that the mean of whatever has been taken as the first set of scores is less than that for the second group of scores. The 'df' which occurs in this and many other printouts, refers to 'degrees of freedom'. It is a statistical concept of some importance but is in many cases, including this, simply related to the size of the sample. Packages also provide a test for equality of variances (the F-test, sometimes called a 'variance-ratio test). As you might expect, this tells you if you have a statistically significant difference in the variances (concerned with the distribution of scores about the mean) of the two groups. If there is no significant difference, you can use the output for 'equal variances assumed' and if there is a significant difference, use the output for 'equal variances not assumed.'

Non-parametric equivalents to the *t*-test

Most statistical packages provide a range of 'non-parametric' tests. Parametric tests (of which the *t*-test is an example) are ones that have been based in their derivation on certain assumptions as to the nature of the distributions from which the data come (usually that they are normal). Non-parametric tests are based on other principles and do not make this kind of assumption. Proponents of parametric tests argue that they are more *efficient* (in the sense that they will detect a significant difference with a smaller sample size than the corresponding non-parametric test – however, this is not always the case, see p. 452); that it is possible to carry out a greater range and variety of tests with them; and that they are *robust* (meaning that violations of the assumptions on which they

are based, e.g. about the normality of the distribution from which the data samples are drawn, have little or no effect on the results they produce).

Advocates of non-parametric tests counter with the arguments that their tests tend to be easier to use and understand and hence less prone to mindless regurgitation; that because of their 'distribution-free' nature (i.e. no assumptions made about the type of distribution of the scores), they are usable in a wider variety of contexts. They have also been proposed as preferable when the assumption of random sampling from known populations, central to most conventional parametric statistics, cannot be assumed. The best of such tests are virtually identical efficiency-wise to parametric ones in situations where the latter can legitimately be used – and obviously preferable in other situations. There is now an adequate range of tests to deal with virtually any situation (Higgins, 2003; Pett, 1997; Sprent and Smeeton, 2007).

A pragmatic approach is suggested, driven mainly by the kind of data you have to deal with. If your data are obviously non-normal, or are in the form of ranks (i.e. first, second, etc.), then a non-parametric test is needed. Otherwise, the range of tests to which you have access through computer packages and the expectations in the field in which you are working are important considerations. Conventional parametric testing is the safe option as it is widely used, and with computer packages, you do not need to worry about the amount of computation required.

The *Mann–Whitney U test* is a non-parametric equivalent of the unpaired two-group *t*-test. The *Wilcoxon signed-rank test* is a non-parametric equivalent of the paired two-group *t*-test. Computation is straightforward and the output in both cases provides 'z'-scores (standardized scores expressed in standard deviation units) and associated probabilities. If there are ties in the scores, a corrected z-score is provided. Strictly, the tests should not be used if there is a substantial proportion of ties.

There are other non-parametric tests provided by some packages which are appropriate for use in the same type of situation as a Mann–Whitney U test. They are not widely used.

Three (or more)-group tests

It is not uncommon in experiments to have three or more conditions. You may wish, for example, to compare the effects of 'high', 'medium' and 'low' stress levels on performance in some situation. It would be possible to take these conditions in pairs and carry out three separate *t*-tests. However, there are techniques which allow this to be done in a single overall test. It is necessary to separate out the 'independent samples' and 'paired samples' designs in the same kind of way as was done with the *t*-test.

Analysis of variance (single factor independent samples)

This situation requires the simplest version of a very widely used, and extremely versatile, technique known as *analysis of variance*. Figure 16.19 shows the format of data referred to. It is commonly referred to as a 'one-Way ANOVA'. Figure 16.20 illustrates the type of output generated for this design. The key finding here is that *there is an overall difference between the means under the different conditions*. This is shown by the F-test result and its associated

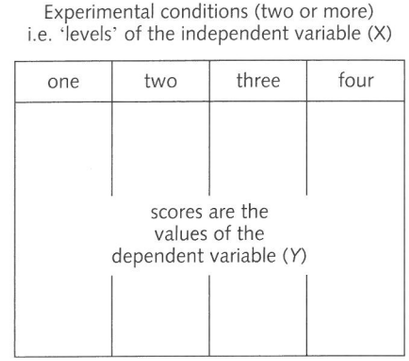

Figure 16.19: Format of single-factor independent samples analysis of variance.

Conditions	n	Mean	SE	Pooled SE	SD
A	5	15.06	0.668	1.069	1.49
B	4	20.55	1.109	1.195	2.22
C	4	16.20	1.046	1.195	2.09
D	5	25.80	1.469	1.069	3.28
Source of variation	Sum squares	DF	Mean square	F statistic	p
Conditions	344.98	3	114.99	20.13	<0.0001
Residual	79.96	14	5.71		
Total	424.95	17			
Tukey					
Contrast	Difference	95% CI			
A v B	−5.49	−10.15	to −0.83	(significant)	
A v C	−1.14	−5.80	to 3.52		
A v D	−10.74	−15.13	to −6.35	(significant)	
B v C	4.35	−0.56	to 9.26		
B v D	−5.25	−9.91	to −0.59	(significant)	
C v D	−9.60	−14.26	to −4.94	(significant)	

Overall, the difference between conditions is statistically significant ($p < 0.0001$)

Note: The Tukey test is one of several available to test for the difference between pairs of conditions. The printout shows if a particular comparison between a pair ('contrast') is statistically significant ($p < 0.05$).

Figure 16.20: Results of a single factor independent samples analysis of variance.

probability, which is smaller than the 0.05 level. This is conventionally reported as: *the difference between groups is statistically significant ($F = 20.13$, $p < 0.0001$; with 3 and 14 df).*

When there is a significant overall difference between the groups, various additional statistics are available helping to pinpoint which of the differences between particular pairs of means are contributing to this overall difference. In Figure 16.20, the *Tukey test* has been used, but others are available. They are alternative ways of dealing with the problem of assessing significance level when a sequence of similar tests is carried out on a data set. Effectively what 'statistically significant at the 5 per cent level' means is that if you carry out, say, 20 tests, you would expect one of the 20 (5 per cent) to be significant even if there is no real effect and only random factors are at work. Any significant difference between two conditions should only be reported if the overall F-test is significant.

As with *t*-tests, it is helpful to report not only the results of the statistical test but also to give the summary statistics, such as the means, standard deviations and confidence intervals under the different conditions. This applies when reporting any analysis of variance findings and helps you, and the reader, to appreciate what the analysis means.

Kruskal–Wallis test

This is a non-parametric equivalent to the above analysis of variance. It is simpler to compute manually but with computer assistance, there seems little reason to prefer it unless the data are in a form for which a parametric test is unsuitable.

Analysis of variance (single factor repeated measures)

The only difference between this design and the previous analysis of variance is that there is a basis for pairing individual scores across the conditions (usually because the same person produces a score under each condition, i.e. there are 'repeated measures'). The format of this design is given as Figure 16.21 and Figure 16.22 gives an example. In the example the same group of 'subjects' are each tested on a task at four-yearly intervals. The use of the term 'subjects' rather than 'participants' is deeply engrained in the psyche of users of analysis of variance – sorry!

Experimental conditions (two or more)
i.e. 'levels' of the independent variable (X)

	one	two	three	four
participant 1				
participant 2		scores are the values of the dependent variable (Y)		
participant 3				
participant 4				

Figure 16.21: Format of single-factor repeated measures analysis of variance.

Score	n	Mean	SE	SD	
1 year	12	103.0	3.96	13.7	
2 years	12	107.0	4.09	14.2	
3 years	12	110.0	3.85	13.3	
4 years	12	112.0	4.26	14.8	
Source of variation	Sum squares	DF	Mean square	F statistic	p
Score	552.0	3	184.0	3.03	0.0432
Subjects	6624.0	11	602.2	-	-
Residual	2006.0	33	60.8		
Total	9182.0	47			

Overall, the difference between the scores in different years is statistically significant ($p < 0.05$).

Figure 16.22: Results of single factor repeated measures analysis of variance.

Friedman test

This is a non-parametric equivalent to the above paired samples analysis of variance. Again, it is relatively simple to compute manually but with computer assistance, there seems little reason to prefer it when the data are in a form for which a parametric test is suitable.

Testing differences when two (or more) independent variables are involved

As discussed in Chapter 5, it is feasible to have two or more independent variables in an experiment, commonly in what are called factorial designs (p. 105). There is a plethora of different analyses for the many different designs. The following account simply tries to introduce some main issues.

Simple two-way independent samples analysis of variance

Figure 16.23 illustrates the form of the data with this design, and Figure 16.24 gives corresponding output. The analysis permits an assessment of the effect of each variable separately (the 'main effect' of variables A and B) and also of any possible 'interaction' (or AB effect) between the two variables. In the case of the example shown in Figure 16.21 significant effects were found for both A and B variables. When the AB interaction is significant this means that the effect of one variable differs at different levels of the second variable. It is then not legitimate to talk about A or B having an overall effect. In this example the interaction is non-significant so the problem does not occur.

The pattern of results should be examined carefully. It helps to display any significant interaction graphically.

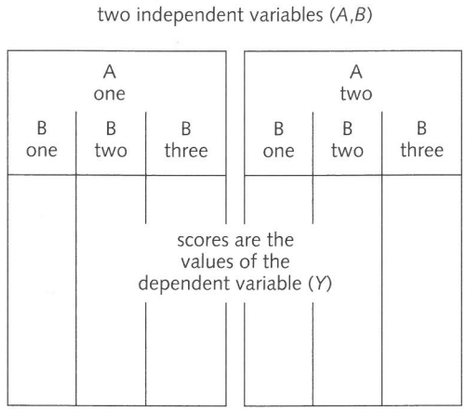

Figure 16.23: Format for simple two-variable analysis of variance.

Two-variable (or factor) analysis of variance with repeated measures

A frequently used complication of the above design is to add repeated measures. Thus, participants may not be simply tested once under a particular combination of levels of the two variables but may be given a series of trials. See Field (2005) for further discussion and examples.

A	n	Mean	SE	Pooled SE	SD
A_1	10	175.8	3.90	4.92	12.3
A_2	10	203.7	4.75	4.92	15.0
A_3	10	197.0	7.07	4.92	22.4
B	n	Mean	SE	Pooled SE	SD
B_1	15	185.5	5.23	4.02	20.2
B_2	15	198.8	4.92	4.02	19.0
Source of variation	Sum squares	DF	Mean square	F statistic	p
A	4242.5	2	2121.2	8.75	0.0014
B	1320.0	1	1320.0	5.45	0.0283
AB	759.3	2	379.6	1.57	0.2294
Residual	5816.4	24	242.4		
Total	12138.2	29			

Figure 16.24: Results of a two-way independent samples analysis of variance.

Testing differences when two (or more) dependent variables are involved

The above analyses are all limited to dealing with a single dependent variable. In studies with more than one, it is possible simply to repeat the analysis for each dependent variable in turn. However, there are advantages in carrying out a single, global, analysis. One version of this is called *multivariate analysis of variance* (MANOVA). The analysis and its interpretation are complex and should not be undertaken without advice. Field (2005) provides more detailed coverage.

Statistical testing and non-random samples

The theoretical rationale of many of the statistical tests discussed in this chapter is based on an assumption of random sampling. However, in practice, much published research in fields covered in this text does not make use of random sampling techniques to select participants. How can this discrepancy be dealt with? Possibilities include:

Use random sampling whenever feasible

In some situations random sampling from known populations, if taken into account at the design stage, can be achieved without major problems. This is the preferred solution for those wishing to continue to use conventional parametric statistical tests.

Use randomization tests

While non-parametric tests, discussed earlier in the chapter, are presented as being 'distribution-free' (i.e. they are not dependent on assumptions about the type of distribution of scores, whereas conventional tests typically assume normal distributions) they can also be derived without assuming random sampling from known population. However, random *assignment* to different conditions (as in true experiments) is assumed. The probability of particular patterns of results can be computed as an exercise in permutations and combinations (sometimes referred to as Monte Carlo techniques, where several thousand random repetitions are carried out to establish a very close approximation to the exact probability of different outcomes). Tests such as Mann–Whitney and Wilcoxon simplify the task by dealing with ranks rather than scores. They give tables of probabilities of different rankings, cutting out the need for complex computations.

The rapid increase in readily available computing power has extended the range and complexity of designs which can be analysed by randomization tests. Sprent and Smeeton (2007) give full coverage. Joe Maxwell (personal communication, 21 July 2010) clarifies this point:

> If an experimental study uses random assignment to conditions, but not random sampling from some population, then the result of a significance test tells you how likely it would be, if the random assignment were done repeatedly with the same

sample, that you would get the difference you did (or a larger one) between the groups, by chance variation in assignment alone, if the intervention really had no overall causal effect on the outcome. However, it would not allow you make any statistical inferences to a larger population than the actual sample; any such inferences would have to be based on other arguments (such as the similarity of the sample to the population of interest – not always a reliable indicator). Box, Hunter, and Hunter (1978) are excellent on this issue, which is actually more complicated than suggested here.

Assume that the discrepancy is not important

This is the current default assumption. As analyses making this assumption continue to be widely published, you are unlikely to suffer if you do likewise. Widely used statistical texts such as Box, Hunter and Hunter (1978) assert that standard tests, such as the *t*-test, are adequate approximations which free one from the random sampling assumption (p. 96).

However, the evidence on this issue is inconsistent. For example, Keele, McConnaughy and White (2008) re-analysed data from two experiments using both non-parametric tests (Kruskal-Wallis and Wilcoxon) and randomization tests and showed that the approximations from classical tests (analyses of variance and *t*-tests) in several cases came to different conclusions about the statistical significance of results. Interestingly, both patterns were found; in some cases the conventional tests gave non-statistically significant results while the non-parametric or randomization test was statistically significant, sometimes the reverse.

If you don't use random sampling but still want to assess the generalizabiliy of your results to some population, you can (with some risk) assume that the distribution is not effectively different from random. In doing this, it seems prudent not to assume a normal distribution, and thus to use non-parametric tests.

In the context of survey research, Pruchno *et al.* (2008) show that respondents recruited using convenience sampling can differ on many variables from those recruited by random means. An unsurprising finding, but worth stressing.

Quantitative analysis and different fixed design research strategies

This chapter has not attempted to make one-to-one correspondences between the different fixed design research strategies and particular techniques of quantitative analysis. Ways of displaying, summarizing and manipulating data are essentially common to all strategies. Many of the ways of exploring relationships among the data are applicable to each of them. There are, admittedly, problems when tests of

significance are used in the absence of random allocation but they do not so much preclude their use in quasi-experiments or non-experimental fixed designs, as require the analyst to think very carefully about interpretation.

There are, of course, strong links between some analytical approaches and particular research strategies, which have been referred to at various points in the chapter. Analysis of variance was developed to deal with the analysis of the true experiment. Correlation matrices and associated techniques such as factor analysis are commonly used in surveys and similar non-experimental settings. However, labelling each test as appropriate only for a specific strategy would be as unnecessarily restrictive as insisting that a particular method of data collection, such as direct observation, should only be used in an experiment – when it could well play a part in a case study, or even in a survey.

It may be helpful, nevertheless, to highlight issues in the analysis of data from two types of experimental study likely to be of real world interest (quasi-experiments and single-case experiments) and the analysis of surveys and other non-experimental designs.

The analysis of quasi-experiments

Three possible quasi-experimental designs were recommended for serious consideration in Chapter 5: the pre-test post-test non-equivalent groups design; the interrupted time-series design; and the regression discontinuity design (p. 117). The three designs require differing approaches to their analysis.

Pre-test post-test non-equivalent groups design

This is a very common design and it is not unusual (though incorrect) for researchers to ignore the fact that it is not a true experiment and not to acknowledge this in either their description or analysis. The non-equivalence of the groups means that there are possible selection effects that may bias the results. Several techniques are available which attempt to separate out the treatment effect (i.e. the effect of the independent variable on the dependent variable) from the effect of selection differences. The most frequently used approaches are:

- simple analysis of variance;
- analysis of variance with blocking or matching of participants;
- analysis of variance based on 'gain' scores (e.g. difference between pre and post scores); and
- analysis of covariance (with either single or multiple covariates).

Reichardt (2005) provides details of each of these approaches and of the difficulties they raise. He concludes that 'any one of these statistical methods could be biased enough so that a useful treatment might look harmful and a harmful treatment could look benign or even beneficial'. His recommendation, in line with the general approach taken in this text, is not that we give up and regard the statistical procedures as worthless but that we give them a relatively minor role – by seeking to eliminate, or at least trying to reduce,

the effect of selection and other threats through the *design* of the study rather than relying on the statistical analysis removing their effects. One interesting approach is to try to 'bracket' the effect of a treatment by using a variety of different but reasonable techniques of analysis (see the evaluation of the *Sesame Street* TV series by Cook *et al.*, 1975, for an example; Fisch and Truglio, 2001, include a wide range of related evaluations).

Interrupted time-series design

The approaches to the analysis of this design which are generally regarded as satisfactory (e.g. the use of autoregressive integrated moving average models – ARIMA models – as developed by Box and Jenkins, 1976) require a minimum sequence of about 50y data points, and preferably over 100 of them. Series of this length tend to be very rare in small-scale studies although, if anyone manages to generate this amount of data, the discussion by Glass, Willson and Gottman (2008) provides useful suggestions.

For smaller amounts of data, statistical analysis by standard methods is not advisable. Shadish, Cook and Campbell (2002, pp. 198–203) suggest approaches which can be used to assist in making causal inferences. They point out that even short interrupted time series before and after an intervention can help in eliminating several threats to internal validity, depending on the pattern of results. They recommend adding design features such as control groups, multiple replications, etc. to enhance the interpretability of short time series. While several different statistical analyses have been used, there is little consensus about their worth. Their conclusion is that the best advice is to use several different statistical approaches. If the results from these converge, and are consistent with visual inspection, confidence in the findings is warranted. Gorman and Allison (1996) summarize the possible statistical tests that could be used.

Similar data patterns to those found in interrupted time-series designs occur with single-case designs and hence the forms of analysis suggested for single-case designs may be appropriate in some cases (see the section below).

Regression discontinuity design

There are strong advocates, not only of the design itself as discussed in Chapter 5, p. 117, but also of its analysis using multiple regression techniques to provide unbiased estimates of treatment effects. Trochim is foremost as a proponent (see, for example, Trochim, 1984, 1990).[2] However, there is a continuing debate about appropriate statistical approaches (Reichardt, Trochim and Cappelleri, 1995; Stanley, 1991). As with interrupted time series (and single-case experiments) visual inspection of the pattern of results may well be adequate to provide a convincing demonstration of an effect – or lack of it. If you feel that statistical analysis is called for, you are recommended to seek advice and/or undertake further reading. The general introduction by Shadish *et al.* (2002, Chapter 7) is accessible and balanced, and includes an appendix on the various statistical approaches to the analysis of regression discontinuity designs (pp. 243–5).

[2] See Trochim's website at www.socialresearchmethods.net/research/rd.htm which gives free access to a full set of his publications on regression discontinuity.

The analysis of single-case experiments

The simple form of a single-case experiment involves a single independent variable and a single dependent variable (traditionally the *rate* of some response, although this is not a necessary feature). It is, effectively, a time-series experiment carried out using a single 'case' (typically a single person; possibly replicated with a small number of other persons). Hence it might have been dealt with earlier together with other two-variable situations. However, there are major differences of ideology about purposes and procedures of analysis between single-case experimenters and others that warrant their separate consideration.

Researchers using single-case designs have traditionally avoided statistical analysis and relied upon 'eyeballing' the data – looking at a comparison, in graphical form, of the participant's performance (sometimes called 'charting') in the different phases of the study. They tend to claim that if statistical techniques were needed to tease out any effects, then the effects were not worth bothering about (see Sidman, 1960 for a clear exposition of the ideas underlying this methodology). This argument has undeniable force, though we will see that, as usual, things are more complicated. It arose in part because of the applied focus of many single-case studies (see the *Journal of Applied Behavior Analysis* for examples), where the distinction is commonly made between *clinical significance* and *statistical significance*. As discussed previously (p. 449), the latter refers to the unlikeliness that a result is due to chance factors; clinical significance means that a treatment has produced a substantial effect such that, for example, a person with problems can now function adequately in society.

However, while Skinner and his fellow experimenters working with rats or pigeons in the laboratory were able to exert sufficient control so that stable baselines could be obtained from which the performance in subsequent phases could be differentiated, this has not surprisingly proved more difficult in applied 'field' studies with humans. This once again illustrates the difference between 'open' and 'closed' systems discussed in Chapter 2. Some researchers with 'applied' interests have questioned the wisdom of ignoring non-dramatic changes, particularly in exploratory work (e.g. Barlow, Nock and Hersen, 2008). They argue that we should determine whether changes are reliable and then subsequently follow them up. It has also been demonstrated that the reliability of 'eye-balling' as a technique can leave a lot to be desired, with different individuals reaching different conclusions as to what the data were telling them. Hagopian *et al.* (1997) have developed structured criteria which increase reliability.

There is an increasing interest in, and use of, statistical tests of significance in single-case studies, although this is still a controversial area (Gorman and Allison, 1996; Kazdin, 1982, Chapter 10). The techniques advocated are themselves the subject of controversy, largely for the same reason encountered in considering time-series quasi-experimental designs, namely, the lack of independence of successive data points obtained from the same individual. Again, the sequence of data points obtained in most real world single-case experiments is insufficient to use standard time-series approaches but they provide a good solution if extensive data are available.

Kazdin (1982, Appendix B) gives details on various possible tests and their application. Todman and Dugard (2001) argue in favour of the use of randomization (exact) tests which

can provide valid statistical analyses for all designs that incorporate a random procedure for assigning treatments to subjects or observation periods, including single-case designs. The tests have not been used widely until recently as they require substantial computational power. They give examples of the analysis of a wide range of single-case designs using Excel, with references to the use of SPSS and other statistical packages.

If it is important that you carry out a statistical analysis of the results of a single-case study (rather than simply performing a visual analysis), there is much to be said for introducing randomization at the design stage so that these tests can be used. Edgington (1996) provides a clear specification of what is required in design terms.

The analysis of surveys

Surveys, and many other non-experimental designs, in their simplest form produce a two-dimensional rows and columns matrix coding data on a range of variables from a set of respondents. As ever, your research questions drive the form of analysis which you choose. If the study is purely descriptive, the techniques covered in the section on 'Exploring the Data Set' (p. 421) may be all that you need. Frequency distributions, graphical displays such as histograms and box plots, summary statistics such as means and standard deviations, will go a long way toward providing the answers required. You may need to go on to analyse relationships using contingency tables and correlation coefficients and/or analysing differences through *t*-tests and other statistics.

With studies seeking to explain or understand what is going on, you should still start with this exploratory, 'know your data', phase. However, when you have some conceptual model, possibly expressed in terms of mechanisms, which provides the basis for the research questions, you will need to take the analysis further following a more confirmatory style.

This may involve the testing of some specific two-variable relationships. The 'elaboration' approach (discussed on p. 439) can help to clarify the meaning of relationships between variables. More sophisticated analyses such as multiple linear regression perform a similar task in more complex situations. And modelling techniques such as SEM (p. 443) provide ways of quantifying and testing the conceptual model.

Causation in surveys and other non-experimental designs

Catherine Marsh (1982) accepts that 'the process of drawing causal inferences from surveys is problematic and indirect' (p. 69). Writing from a realist perspective she provides a very clear account of some ways of going about this difficult task (Chapter 3). Recall that the traditional positivist view of causation, known as 'constant conjunction', was found in Chapter 2 to be seriously wanting. Following realist precepts, we are concerned to establish a causal model which specifies the existence of a number of mechanisms or processes operating in particular contexts.

How can surveys help in doing this? Obviously correlations can be computed but the injunction that 'correlation does not imply causation' should be etched deeply on the cortex of anyone who has followed a course in statistics. The temptation to get a statistical package to cross-tabulate 'ALL' against 'ALL' (i.e. asking it to crunch out relationships between all the variables and then cherry-picking the significant ones) should be resisted. Working with a 5 per cent significance level means that we expect on average 1 in 20 of the results to be significant due to chance factors. While there are ways of dealing with the problem of chance correlations (e.g. by splitting the cases randomly into two equal subsets, exploring the correlations obtained with the first, then checking with the second), you still need to make ad hoc arguments to explain your pattern of findings.

One approach, simple in principle, involves the *elaboration* strategy discussed earlier in the chapter (p. 439). If you have a thought-through proposed causal model involving a small number of possible mechanisms it can be tested in this way. Unfortunately, this process rapidly becomes unwieldy as more variables are included in the model. By going back and forth between model and data, gradually elaborating the model, it may be feasible to build up a more complex model. This process need not be totally 'pure'. While the unprincipled fishing trip is not a good way of developing a causal model, further development of the model based in part on hunches and insights that occur during analysis is well within the spirit of exploratory data analysis. Marsh and Elliott (2008) summarize the task as follows:

> In the absence of an experiment, a statistical effect of one variable on another cannot just be accepted as causal at face value. If we want to show that it is not *spurious*, we have to demonstrate that there are no plausible *prior variables* affecting both the variables of interest. If we want to argue that the *causal mechanism* is fairly direct, we have to control for similar intervening variables (p. 252, emphases added).

Marsh and Elliott (2008, especially Chapters 11 and 12) give an excellent account of the use of elaboration and related techniques to establish causation. It is stressed here as still being of value in small-scale real world study when data on a restricted set of variables have been collected, and the model to be tested is necessarily simple.

More generally, the more powerful multivariate techniques accessible via specialist statistical packages, in particular structural equation modelling (SEM), as discussed on p. 443, have tended to replace the type of sequential logic required when using elaboration.

A note on interactions

Formally, an interaction is when the effect of one variable on a second one depends on the value of a third variable. They were discussed earlier in the chapter in the context of two variable analyses of variance (p. 456). When such interactions exist – and they are extremely common in social research – they mean that any generalizations we seek to make about causal processes are limited. This is a serious obstacle to positivistic researchers who seek universal laws. However, it is entirely consistent with the realist view that mechanisms do not operate universally and that the research task is to specify the contexts in which they work.

A realist reminder on the analysis and interpretation of experiments

The realist view of experimentation, as discussed in Chapter 5, p. 90, is fundamentally different from the traditional positivist one. The main remaining similarity is in the active role taken by the experimenter. It is her responsibility to set up and manipulate the situation so that predicted mechanisms are given the chance to operate. This calls for a considerable degree of prior knowledge and experience of the phenomenon or situation studied, so that a conceptual map or model can be developed with some confidence. This predicts likely mechanisms, the contexts in which they may work, and for whom they will work.

Viewing experiments in this light does not, in general, either preclude or privilege any of the experimental designs covered in Chapter 5, nor any of the analyses covered in this chapter. There may well be situations, as discussed by Pawson and Tilley (1997, pp. 34–54) where the nature of likely causal agents are such that, for example, randomized allocation to different conditions changes the nature of the phenomenon of interest. Hence control group methodology is inappropriate. Careful attention to such possible distortions is needed at the design stage.

Assuming the choice of an appropriate design, the question becomes one of deciding whether we have good evidence for the operation of the predicted mechanisms. In most cases, this boils down to deciding whether an obtained relationship or difference (typically of means) provides that evidence. While it will in many cases be feasible to generate statistical significance values, they need to be taken with even more than the usual pinch of salt (see p. 446). This is because their derivation has been within the positivistic conceptualization of experimental design. While it seems intuitively reasonable that, irrespective of this heritage, a lower probability value for the statistical significance of a difference gives greater confidence in the reality of that difference, it is difficult to go much further than that. Recall, however, that the advice has been throughout to place greater reliance on measures of effect sizes based on statistics such as confidence intervals and standardized differences in means (p. 447). Armed with these and graphical displays, you then have to judge the quality of the evidence.

Further reading

The website gives annotated references to further reading for Chapter 16.

CHAPTER 17

The analysis and interpretation of qualitative data

This chapter:

- stresses the need for a systematic analysis of qualitative data;
- emphasizes the central role of the person doing the analysis, and warns about some deficiencies of the human as analyst;
- discusses the advantages and disadvantages of using specialist computer software;
- explains the Miles and Huberman approach to analysis which concentrates on reducing the bulk of qualitative data to manageable amounts and on displaying them to help draw conclusions;
- suggests thematic coding analysis as a generally useful technique when dealing with qualitative data;
- reviews the widely used grounded theory approach;
- summarizes a range of alternative approaches; and
- finally considers issues involved in integrating qualitative and quantitative data in multi-strategy designs.

Introduction

Qualitative data have been described as an 'attractive nuisance' (Miles, 1979). Their attractiveness is undeniable. Words, which are by far the most common form of qualitative data, are a speciality of humans and their organizations. Narratives, accounts and other collections of words are variously described as 'rich', 'full' and 'real', and