15

Quantitative data analysis

Chapter outline

Chapter guide

In this chapter, some of the basic but nonetheless most frequently used methods for analysing quantitative data analysis will be presented. In order to illustrate the use of the methods of data analysis, a small imaginary set of data based on attendance at a gym is used. It is the kind of small research project that would be feasible for most students doing undergraduate research projects for a dissertation or similar exercise. The chapter explores:

- the importance of *not* leaving considerations of how you will analyse your quantitative data until after you have collected all your data; you should be aware of the ways in which you would like to analyse your data from the earliest stage of your research;
- the distinctions between the different kinds of variable that can be generated in quantitative research; knowing how to distinguish types of variables is crucial so that you appreciate which methods of analysis can be applied when you examine variables and relationships between them;
- methods for analysing a single variable at a time (*univariate analysis*);
- methods for analysing relationships between variables (*bivariate analysis*);
- the analysis of relationships between three variables (*multivariate analysis*).

Introduction

In this chapter, some very basic techniques for analysing quantitative data will be examined. In the next chapter, the ways in which these techniques can be implemented using sophisticated computer software will be introduced. As explained in Chapter 16, this software has been known for years as SPSS, but the version described in the chapter is referred to as PASW Statistics 18. However, I will continue to refer to the software as SPSS, since the name SPSS is to be restored for the next release, when it will be referred to as IBM SPSS. The formulae that underpin the techniques to be discussed will not be presented, since the necessary calculations can easily be carried out by using SPSS. Two chapters cannot do justice to these topics and readers are advised to move as soon as possible on to books that provide more detailed and advanced treatments (e.g. Bryman and Cramer 2011).

Before beginning this exposition of techniques, I would like to give you advance warning of one of the biggest mistakes that people make about quantitative data analysis:

> I don't have to concern myself with how I'm going to analyse my survey data until after I've collected my data. I'll leave thinking about it till then, because it doesn't impinge on how I collect my data.

This is a common error that arises because quantitative data analysis looks like a distinct phase that occurs after the data have been collected (see, for example, Figure 7.1, in which the analysis of quantitative data is depicted as a late step—number 9—in quantitative research). Quantitative data analysis is indeed something that occurs typically at a late stage in the overall process and is also a distinct stage.

However, that does not mean that you should not be considering how you will analyse your data until then. In fact, you should be fully aware of what techniques you will apply at a fairly early stage—for example, when you are designing your questionnaire, observation schedule, coding frame, or whatever. The two main reasons for this are as follows.

1. You cannot apply just any technique to any variable. Techniques have to be appropriately matched to the types of variables that you have created through your research. This means that you must be fully conversant with the ways in which different types of variable are classified.
2. The size and nature of your sample are likely to impose limitations on the kinds of techniques you can use (see the section on 'Kind of analysis' in Chapter 8).

In other words, you need to be aware that decisions that you make at quite an early stage in the research process, such as the kinds of data you collect and the size of your sample, will have implications for the sorts of analysis that you will be able to conduct.

A small research project

The discussion of quantitative data analysis will be based upon an imaginary piece of research carried out by an undergraduate social science student for a dissertation. The student in question is interested in the role of the sport and leisure industry and in particular, because of her own enthusiasm for leisure clubs and gyms, with the ways in which such venues are used and people's reasons for joining them. She has read an article that suggests that participant involvement in adult fitness programmes is associated with their attitudinal loyalty, comprising investment of time and money, social pressure from significant others, and internalization or commitment to the fitness regime (Park 1996). She intends to use this theory as a framework for her findings. The student is also interested in issues relating to gender and body image and she suspects that men and women will differ in their reasons for going to a gym and the kinds of activities in which they engage in the gym. Her final issue of interest relates to the importance of age in determining gym involvement. In particular, she has discovered that previous research has shown that older people tend to show higher levels of attitudinal loyalty to recreational activities more generally and she wants to find out if this finding also applies to involvement in leisure clubs and gyms.

She secures the agreement of a gym close to her home to contact a sample of its members by post. The gym has 1,200 members and she decides to take a simple random sample of 10 per cent of the membership (that is, 120 members). She sends out postal questionnaires to members of the sample with a covering letter testifying to the gym's support of her research. One thing she wants to know is how much time people spend on each of the three main classes of activity in the gym: cardiovascular equipment, weights equipment, and exercises. She defines each of these carefully in the covering letter and asks members of the sample to keep a note of how long they spend on each of the three activities on their next visit. They are then requested to return the questionnaires to her in a prepaid reply envelope. She ends up with a sample of ninety questionnaires—a response rate of 75 per cent.

Part of the questionnaire is presented in Tips and skills 'A completed and processed questionnaire' and has been completed by a respondent and coded by the student. The entire questionnaire runs to four pages, but only twelve of the questions are provided here. Many of the questions (1, 3, 4, 5, 6, 7, 8, and 9) are pre-coded, and the student simply has to circle the code to the far right of the question under the column 'code'. With the remainder of the questions, specific figures are requested, and she simply transfers the relevant figure to the code column.

Tips and skills
A completed and processed questionnaire

Questionnaire		Code
1. Are you male or female (please tick)?		
Male ✓ Female ___		(1) 2
2. How old are you?		
21 years		21
3. Which of the following best describes your *main* reason for going to the gym? (please tick *one* only)		
Relaxation	___	1
Maintain or improve fitness	✓	(2)
Lose weight	___	3
Meet others	___	4
Build strength	___	5
Other (please specify)	___	6

4. When you go to the gym, how often do you use the cardiovascular equipment (jogger, step machine, bike, rower)? (please tick)

Always	✓	(1)
Usually	___	2
Rarely	___	3
Never	___	4

5. When you go to the gym, how often do you use the weights machines (including free weights)? (please tick)

Always	✓	(1)
Usually	___	2
Rarely	___	3
Never	___	4

6. How frequently do you usually go to the gym? (please tick)

Every day	___	1
4–6 days a week	___	2
2 or 3 days a week	✓	(3)
Once a week	___	4
2 or 3 times a month	___	5
Once a month	___	6
Less than once a month	___	7

7. Are you usually accompanied when you go to the gym or do you usually go on your own? (please tick *one* only)

On my own	✓	(1)
With a friend	___	2
With a partner/spouse	___	3

8. Do you have sources of regular exercise other than the gym?

Yes ___ No ✓ 1 (2)

*If you have answered **No** to this question, please proceed to question 10*

9. If you have replied **Yes** to question **8**, please indicate the *main* source of regular exercise in the last six months from this list. (please tick *one* only) (0)

Sport	___	1
Cycling on the road	___	2
Jogging	___	3
Long walks	___	4
Other (please specify)	___	5

10. During your last visit to the gym, how many minutes did you spend on the cardiovascular equipment (jogger, step machine, bike, rower)?

33 minutes 33

11. During your last visit to the gym, how many minutes did you spend on the weights machines (including free weights)?

17 minutes 17

12. During your last visit to the gym, how many minutes did you spend on other activities (e.g. stretching exercises)?

5 minutes 5

Missing data

The data for all ninety respondents are presented in Tips and skills 'Gym survey data'. Each of the twelve questions is known for the time being as a variable number (var00001, etc.). The variable number is a default number that is imposed by SPSS, the statistical package that is described in the next chapter. Each variable number corresponds to the question number in Tips and skills 'A completed and processed questionnaire' (i.e. var00001 is question 1, var00002 is question 2, etc.). An important issue arises in the management of data as to how to handle 'missing data'. Missing data arise when respondents fail to reply to a question—either by accident or because they do not want to answer the question. Thus, respondent 24 has failed to answer question 2, which is concerned with age. This has been coded as a zero (0) and it will be important to ensure that the computer software is notified of this fact, since it needs to be taken into account during the analysis. Also, question 9 has a large number of zeros, because many people did not answer it, because they have been filtered out by the previous question (that is, they do not have other sources of regular exercise). These have also been coded as zero to denote missing data, though strictly speaking their failure to reply is more indicative of the question not being applicable to them. Note also, that there are zeros for var00010, var00011, and var00012. However, these do *not* denote missing data but that the respondent spends zero minutes on the activity in question. Everyone has answered questions 10, 11, and 12, so there are in fact no missing data for these variables. If there had been missing data, it would be necessary to code missing data with a number that could not also be a true figure. For example, nobody has spent 99 minutes on these activities, so this might be an appropriate number, as it is easy to remember and could not be read by the computer as anything other than missing data.

Tips and skills
Gym survey data

var00001	var00002	var00003	var00004	var00005	var00006	var00007	var00008	var00009	var00010	var00011	var00012
1	21	2	1	1	3	1	2	0	33	17	5
2	44	1	3	1	4	3	1	2	10	23	10
2	19	3	1	2	2	1	1	1	27	18	12
2	27	3	2	1	2	1	2	0	30	17	3
1	57	2	1	3	2	3	1	4	22	0	15
2	27	3	1	1	3	1	1	3	34	17	0
1	39	5	2	1	5	1	1	5	17	48	10
2	36	3	1	2	2	2	1	1	25	18	7
1	37	2	1	1	3	1	2	0	34	15	0
2	51	2	2	2	4	3	2	0	16	18	11
1	24	5	2	1	3	1	1	1	0	42	16
2	29	2	1	2	3	1	2	0	34	22	12
1	20	5	1	1	2	1	2	0	22	31	7
2	22	2	1	3	4	2	1	3	37	14	12
2	46	3	1	1	5	2	2	0	26	9	4
2	41	3	1	2	2	3	1	4	22	7	10
1	25	5	1	1	3	1	1	1	21	29	4
2	46	3	1	2	4	2	1	4	18	8	11
1	30	3	1	1	5	1	2	0	23	9	6
1	25	5	2	1	3	1	1	1	23	19	0
2	24	2	1	1	3	2	1	2	20	7	6
2	39	1	2	3	5	1	2	0	17	0	9
1	44	3	1	1	3	2	1	2	22	8	5
1	0	1	2	2	4	2	1	4	15	10	4
2	18	3	1	2	3	1	2	1	18	7	10
1	41	3	1	1	3	1	2	0	34	10	4
2	38	2	1	2	5	3	1	2	24	14	10
1	25	2	1	1	2	1	2	0	48	22	7
1	41	5	2	1	3	1	1	2	17	27	0

var00001	var00002	var00003	var00004	var00005	var00006	var00007	var00008	var00009	var00010	var00011	var00012
2	30	3	1	1	2	2	2	0	32	13	10
2	29	3	1	3	2	1	2	0	31	0	7
2	42	1	2	2	4	2	1	4	17	14	6
1	31	2	1	1	2	1	2	0	49	21	2
2	25	3	1	1	2	3	2	0	30	17	15
1	46	3	1	1	3	1	1	3	32	10	5
1	24	5	2	1	4	1	1	2	0	36	11
2	34	3	1	1	3	2	1	4	27	14	12
2	50	2	1	2	2	3	2	0	28	8	6
1	28	5	1	1	3	2	1	1	26	22	8
2	30	3	1	1	2	1	1	4	21	9	12
1	27	2	1	1	2	1	1	3	64	15	8
2	27	2	1	2	4	2	1	4	22	10	7
1	36	5	1	1	3	2	2	0	21	24	0
2	43	3	1	1	4	1	2	0	25	13	8
1	34	2	1	1	3	2	1	1	45	15	6
2	27	3	1	1	2	1	1	4	33	10	9
2	38	2	1	3	4	2	2	0	23	0	16
1	28	2	1	1	3	3	1	2	38	13	5
1	44	5	1	1	2	1	2	0	27	19	7
2	31	3	1	2	3	2	2	0	32	11	5
2	23	2	1	1	4	2	1	1	33	18	8
1	45	3	1	1	3	1	1	2	26	10	7
2	34	3	1	2	2	3	2	0	36	8	12
1	27	3	1	1	2	3	1	3	42	13	6
2	40	3	1	1	2	2	1	4	26	9	10
2	24	2	1	1	2	1	1	2	22	10	9
1	37	2	1	1	5	2	2	0	21	11	0
1	22	5	1	1	4	1	1	1	23	17	6
2	31	3	1	2	3	1	1	4	40	16	12
1	37	2	1	1	2	3	2	0	54	12	3
2	33	1	2	2	4	2	2	0	17	10	5
1	23	5	1	1	3	1	1	1	41	27	8
1	28	3	1	1	3	3	2	0	27	11	8
2	29	2	1	2	5	2	1	2	24	9	9
2	43	3	1	1	2	1	2	0	36	17	12
1	28	5	1	1	3	1	1	1	22	15	4
1	48	2	1	1	5	1	1	4	25	11	7
2	32	2	2	2	4	2	2	0	27	13	11
1	28	5	1	1	2	2	2	0	15	23	7
2	23	2	1	1	5	1	1	4	14	11	5
2	43	2	1	2	5	1	2	0	18	7	3
1	28	2	1	1	4	3	1	2	34	18	8
2	23	3	1	1	2	1	2	0	37	17	17
2	36	1	2	2	4	2	1	4	18	12	4
1	50	2	1	1	3	1	1	2	28	14	3
1	37	3	1	1	2	2	2	0	26	14	9
2	41	3	1	1	2	1	1	4	24	11	4
1	26	5	2	1	5	1	1	1	23	19	8
2	28	3	1	1	4	1	2	0	27	12	4
2	35	2	1	1	3	1	1	1	28	14	0
1	28	5	1	1	2	1	1	2	20	24	12
2	36	2	1	1	3	2	2	0	26	9	14
2	29	3	1	1	4	1	1	4	23	13	4
1	34	1	2	2	4	2	1	0	24	12	3
1	53	2	1	1	3	3	1	1	32	17	6
2	30	3	1	1	4	1	2	0	24	10	9
1	43	2	1	1	2	1	1	2	24	14	10
2	26	5	2	1	4	1	1	1	16	23	7
2	44	1	1	1	4	2	2	0	27	18	6
1	45	1	2	2	3	3	2	0	20	14	5

Types of variable

One of the things that might strike you when you look at the questions is that the kinds of information that you receive varies by question. Some of the questions call for answers in terms of real numbers: questions 2, 10, 11, and 12. Questions 1 and 8 yield either/or answers and are therefore in the form of dichotomies. The rest of the questions take the form of lists of categories, but there are differences between these too. Some of the questions are in terms of answers that are rank ordered: questions 4, 5, and 6. Thus we can say in the case of question 6 that the category 'every day' implies greater frequency than '4–6 days a week', which in turn implies greater frequency than '2 or 3 days a week', and so on. However, in the case of questions 3, 7, and 9, the categories are *not* capable of being rank ordered. We cannot say in the case of question 3 that 'relaxation' is more of something than 'maintain or improve fitness' or 'lose weight'.

These considerations lead to a classification of the different types of variable that are generated in the course of research. The four main types are:

- *Interval/ratio variables*. These are variables where the distances between the categories are identical across the range of categories. In the case of variables var00010 to var00011, the distance between the categories is 1 minute. Thus, a person may spend 32 minutes on cardiovascular equipment, which is 1 minute more than someone who spends 31 minutes on this equipment. That difference is the same as the difference between someone who spends 8 minutes and another who spends 9 minutes on the equipment. This is the highest level of measurement and a very wide range of techniques of analysis can be applied to interval/ratio variables. There is, in fact, a distinction between interval and ratio variables, in that the latter are interval variables with a fixed zero point. However, since most ratio variables exhibit this quality in social research (for example, income, age, number of employees, revenue), they are not being distinguished here.
- *Ordinal variables*. These are variables whose categories can be rank ordered (as in the case of interval/ratio variables) but the distances between the categories are not equal across the range. Thus, in the case of question 6, the difference between the category 'every day' and '4–6 days a week' is not the same as the difference between '4–6 days a week' and '2 or 3 days a week', and so on. Nonetheless, we can say that 'every day' is more frequent than '4–6 days a week', which is more frequent than '2 or 3 days a week', etc. You should also bear in mind that, if you subsequently group an interval/ratio variable like var00002, which refers to people's ages, into categories (e.g. 20 and under; 21–30; 31–40; 41–50; 51 and over), you are transforming it into an ordinal variable.
- *Nominal variables*. These variables, also known as *categorical variables*, comprise categories that cannot be rank ordered. As noted previously, we cannot say in the case of question 3 that 'relaxation' is more of something than 'maintain or improve fitness' or 'lose weight'.
- *Dichotomous variables*. These variables contain data that have only two categories (for example, gender). Their position in relation to the other types is slightly ambiguous, as they have only one interval. They therefore can be considered as having attributes of the other three types of variable. They look as though they are nominal variables, but because they have only one interval they are sometimes treated as ordinal variables. However, it is probably safest to treat them for most purposes as if they were ordinary nominal variables.

The four main types of variable and illustrations of them from the gym survey are provided in Table 15.1.

Multiple-indicator (or multiple-item) measures of concepts, like Likert scales (see Key concept 7.2), produce strictly speaking ordinal variables. However, many writers argue that they can be treated as though they produce interval/ratio variables, because of the relatively large number of categories they generate. For a brief discussion of this issue, see Bryman and Cramer (2011), who distinguish between 'true' interval/ratio variables and those produced by multiple-indicator measures (2011: 71–3).

Figure 15.1 provides guidance about how to identify variables of each type.

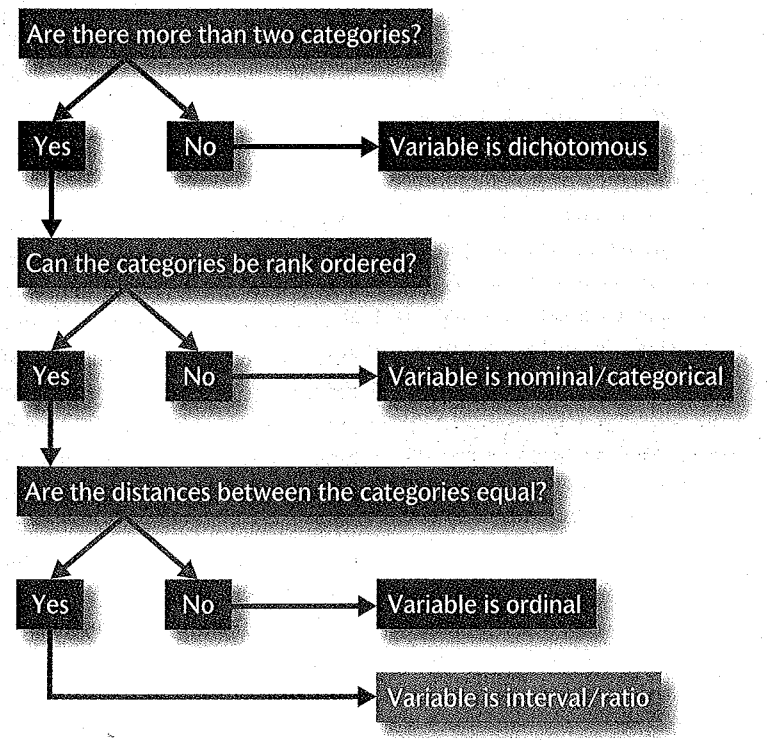
Table 15.1

Types of variable

Type	Description	Examples in gym study	Variable Name in SPSS (see Chapter 16)
Interval/ratio	Variables where the distances between the categories are identical across the range	var00002 var00010 var00011 var00012	age cardmins weimins othmins
Ordinal	Variables whose categories can be rank ordered but the distances between the categories are not equal across the range	var00004 var00005 var00006	carduse weiuse frequent
Nominal	Variables whose categories cannot be rank ordered; also known as *categorical*	var00003 var00007 var00009	reasons accomp exercise
Dichotomous	Variables containing data that have only two categories	var00001 var00008	gender othsourc

Figure 15.1

Deciding how to categorize a variable



Univariate analysis

Univariate analysis refers to the analysis of one variable at a time. In this section, the commonest approaches will be outlined.

Frequency tables

A **frequency table** provides the number of people and the percentage belonging to each of the categories for the variable in question. It can be used in relation to all of the different types of variable. An example of a frequency table is provided for var00003 in Table 15.2. Notice that nobody chose two of the possible choices of answer—'meet others' and 'other'—so these are not included in the table. The table shows, for example, that 33 members of the sample go the gym to lose weight and that they represent 37 per cent (percentages are often rounded up and down in frequency tables) of the entire sample. The procedure for generating a frequency table with SPSS is described on page 361.

If an interval/ratio variable (like people's ages) is to be presented in a frequency table format, it is invariably the case that the categories will need to be grouped. When grouping in this way, take care to ensure that the categories you create do not overlap (for example, like this: 20–30, 30–40, 40–50, etc.). An example of a frequency table for an interval/ratio variable is shown in Table 15.3, which provides a frequency table for var00002, which is concerned with the ages of those visiting the gym. If we did not group people in terms of age ranges, there would be thirty-four different categories, which is too many to take in. By creating five categories, we make the distribution of ages easier to comprehend. Notice that the sample totals 89 and that the percentages are based on a total of 89 rather than 90. This is because this variable contains one missing value (respondent 24). The procedure for grouping respondents with SPSS is described on page 359.

Table 15.2

Frequency table showing reasons for visiting the gym

Reason	*n*	%
Relaxation	9	10
Maintain or improve fitness	31	34
Lose weight	33	37
Build strength	17	19
TOTAL	90	100

Table 15.3

Frequency table showing ages of gym members

Age	*n*	%
20 and under	3	3
21–30	39	44
31–40	23	26
41–50	21	24
51 and over	3	3
TOTAL	89	100

Diagrams

Diagrams are among the most frequently used methods of displaying quantitative data. Their chief advantage is that they are relatively easy to interpret and understand. If you are working with nominal or ordinal variables, the *bar chart* and the *pie chart* are two of the easiest methods to use. A bar chart of the same data presented in Table 15.2 is presented in Figure 15.2. Each bar represents the number of people falling in each category. This figure was produced with SPSS. The procedure for generating a bar chart with SPSS is described on page 363.

Another way of displaying the same data is through a pie chart, like the one in Figure 15.3. This also shows the relative size of the different categories but brings out as well the size of each slice relative to the total sample. The percentage that each slice represents of the whole sample is also given in this diagram, which was also produced with SPSS. The procedure for generating a pie chart with SPSS is described on page 363.

If you are displaying an interval/ratio variable, like var00002, a *histogram* is likely to be employed. Figure 15.4, which was also generated by SPSS, uses the same data and categories as Table 15.3. As with the bar chart, the bars represent the relative size of each of the age bands.

Figure 15.2

Bar chart showing the main reasons for visiting the gym (SPSS output)

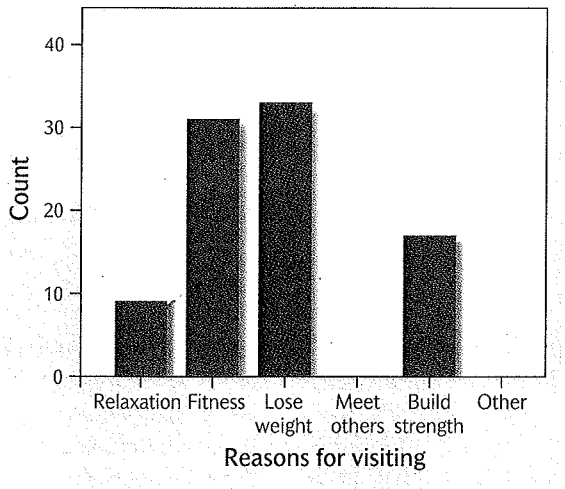


Figure 15.4

Histogram showing the ages of gym visitors (SPSS output)

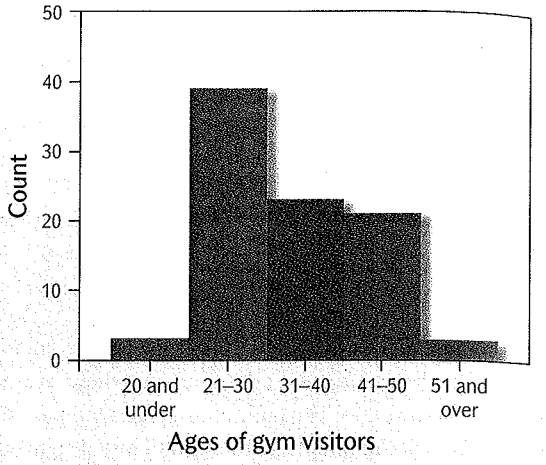
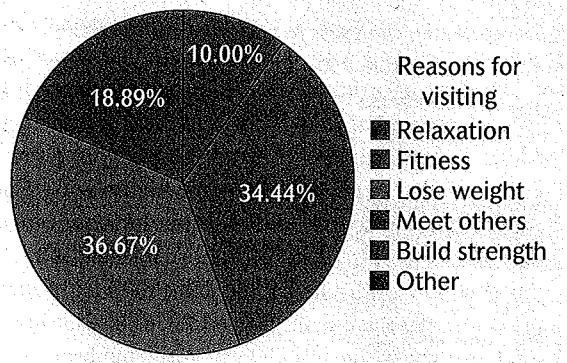


Figure 15.3

Pie chart showing the main reasons for visiting the gym (SPSS output)



However, note that, with the histogram, there is no space between the bars, whereas there is a space between the bars of a bar chart. Histograms are produced for interval/ratio variables, whereas bar charts are produced for nominal and ordinal variables. The procedure for generating a histogram with SPSS is described on page 363.

Measures of central tendency

Measures of central tendency encapsulate in one figure a value that is typical for a **distribution of values**. In effect, we are seeking out an average for a distribution, but, in quantitative data analysis, three different forms of average are recognized.

- *Arithmetic mean*. This is the average as we understand it in everyday use—that is, we sum all the values in a distribution and then divide by the number of values. Thus, the arithmetic mean (or more simply the *mean*) for var00002 is 33.6, meaning that the average age of gym visitors is nearly 34 years of age. The mean should be employed only in relation to interval/ratio variables, though it is not uncommon to see it being used for ordinal variables as well.
- *Median*. The **median** is the mid-point in a distribution of values. Whereas the mean is vulnerable to **outliers** (extreme values at either end of the distribution), which will exert considerable upwards or downwards pressure on the mean, by taking the mid-point of a distribution the median is not affected in this way. The median is derived by arraying all the values in a distribution from the smallest to the largest and then finding the middle point. If there is an even number of values, the median is calculated by taking the mean

of the two middle numbers of the distribution. In the case of var00002, the median is 31. This is slightly lower than the mean, in part because some considerably older members (especially respondents 5 and 10) inflate the mean slightly. The median can be employed in relation to both interval/ratio and ordinal variables.

- *Mode.* The **mode** is the value that occurs most frequently in a distribution. The mode for var00002 is 28. The mode can be employed in relation to all types of variable.

The procedure for generating the mean, median, and mode with SPSS is described on page 363.

Measures of dispersion

The amount of variation in a sample can be just as interesting as providing estimates of the typical value of a distribution. For one thing, it becomes possible to draw contrasts between comparable distributions of values. For example, is there more or less variability in the amount of time spent on cardiovascular equipment as compared to weights machines?

The most obvious way of measuring dispersion is by the **range**. This is simply the difference between the maximum and the minimum value in a distribution of values associated with an interval/ratio variable. We find that the range for the two types of equipment is 64 minutes for the cardiovascular equipment and 48 minutes for the weights machines. This suggests that there is more variability in the amount of time spent on the former. However, like the mean, the range is influenced by outliers, such as respondent 60 in the case of var00010.

Another **measure of dispersion** is the **standard deviation**, which is essentially the average amount of variation around the mean. Although the calculation is somewhat more complicated than this, the standard deviation is calculated by taking the difference between each value in a distribution and the mean and then dividing the total of the differences by the number of values. The standard deviation for var00010 is 9.9 minutes and for var00011 it is 8 minutes. Thus, not only is the average amount of time spent on the cardiovascular equipment higher than for the weights equipment; the standard deviation is greater too. The standard deviation is also affected by outliers, but, unlike the range, their impact is offset by dividing by the number of values in the distribution. The procedure for generating the standard deviation with SPSS is described on page 363.

A type of figure that has become popular for displaying interval/ratio variables is the *boxplot* (see Figure 15.5). This form of display provides an indication of both central tendency (the median) and dispersion (the range). It also indicates whether there are any outliers. Figure 15.5 displays a boxplot for the total number of minutes users spent during their last gym visit. There is an outlier—case number 41, who spent a total of 87 minutes in the gym. The box represents the middle 50 per cent of users. The upper line of the box indicates the greatest use of the gym within the 50 per cent and the lower line of the box represents the least use of the gym within the 50 per cent. The line going across the box indicates the median. The line going upwards from the box goes up to the person whose use of the gym was greater than any other user, other than case number 41. The line going downwards from the box goes down to the person whose use of the gym was lower than that of any other user. Boxplots are useful because they display both central tendency and dispersion. They vary in their shape depending on whether cases tend to be high or low in relation to the median. With Figure 15.5, the box and the median are closer to the bottom end of the distribution, suggesting less variation among gym users below the median. There is more variation above the median. The procedure for generating the standard deviation with SPSS is described on page 363.

Bivariate analysis

Bivariate analysis is concerned with the analysis of two variables at a time in order to uncover whether or not the two variables are related. Exploring relationships between variables means searching for evidence that the variation in one variable coincides with variation in another variable. A variety of techniques is available for examining relationships, but their use depends on the nature of the two variables being analysed. Figure 15.6 attempts to portray the main types of bivariate analysis according to the types of variable involved.

Figure 15.5

A boxplot for the number of minutes spent on the last visit to the gym

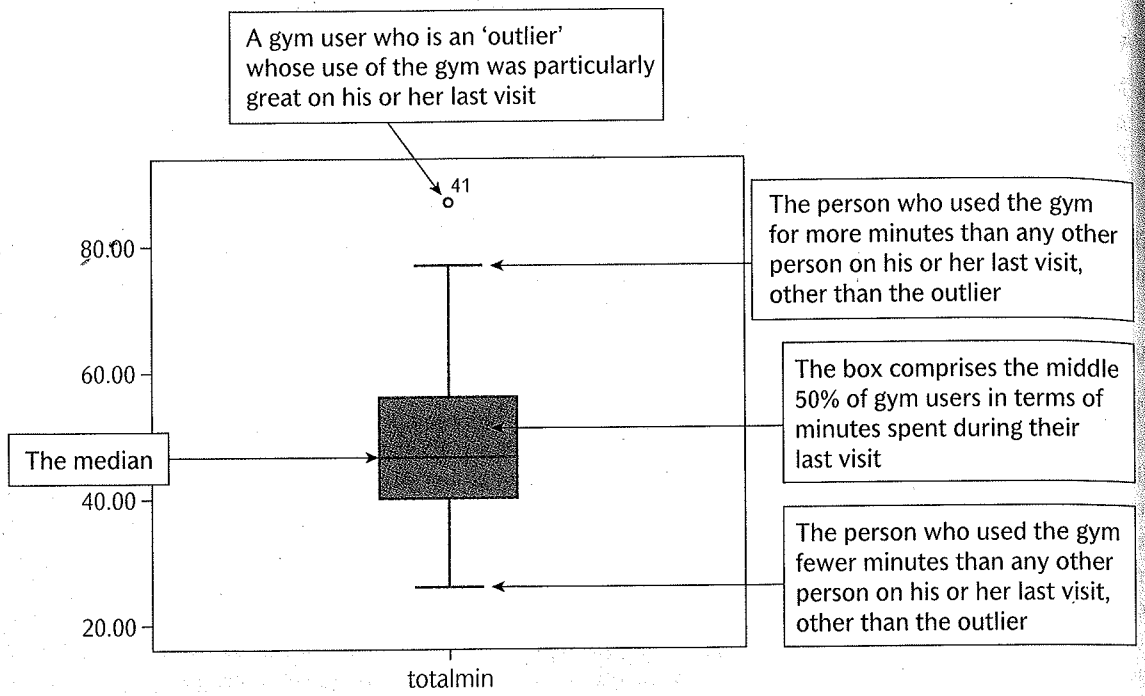


Figure 15.6

Methods of bivariate analysis

	Nominal	Ordinal	Interval/ratio	Dichotomous
Nominal	Contingency table + chi-square (χ^2) + Cramér's *V*	Contingency table + chi-square (χ^2) + Cramér's *V*	Contingency table + chi-square (χ^2) + Cramér's *V* If the interval/ratio variable can be identified as the dependent variable, compare means + eta	Contingency table + chi-square (χ^2) + Cramér's *V*
Ordinal	Contingency table + chi-square (χ^2) + Cramér's *V*	Spearman's rho (ρ)	Spearman's rho (ρ)	Spearman's rho (ρ)
Interval/ratio	Contingency table + chi-square (χ^2) + Cramér's *V* If the interval/ratio variable can be identified as the dependent variable, compare means + eta	Spearman's rho (ρ)	Pearson's *r*	Spearman's rho (ρ)
Dichotomous	Contingency table + chi-square (χ^2) + Cramér's *V*	Spearman's rho (ρ)	Spearman's rho (ρ)	phi (ϕ)

Relationships not causality

An important point to bear in mind about all of the methods for analysing relationships between variables is that it is precisely **relationships** that they uncover. As was noted in Chapter 3 in relation to cross-sectional designs, this means that you cannot infer that one variable causes another. Indeed, there are cases when what appears to be a causal influence working in one direction actually works in the other way. An interesting example of this problem of causal direction will be presented much later in the book in Chapter 27. The example shows that Sutton and Rafaeli (1988) expected to find a causal relationship between the display of positive emotions (for example, smiling, or friendliness on the part of checkout staff) in retail outlets and sales in those outlets. In other words, the display of positive emotions was deemed to have a causal influence on levels of retail sales. In fact, the relationship was found to be the other way round: levels of retail sales exerted a causal influence on the display of emotions (see Research in focus 27.6 for more detailed explanation of this study).

Sometimes, we may feel confident that we can infer a causal direction when a relationship between two variables is discerned—for example, if we find that age and voting behaviour are related. It is impossible for the way people vote to influence their age, so, if we do find the two variables to be related, we can infer with complete confidence that age is the independent variable. It is not uncommon for researchers, when analysing their data, to draw inferences about causal direction based on their assumptions about the likely causal direction among related variables, as Sutton and Rafaeli (1988) did in their study. Although such inferences may be based on sound reasoning, they can only be inferences, and there is the possibility that the real pattern of causal direction is the opposite of that which is anticipated.

Contingency tables

Contingency tables are probably the most flexible of all methods of analysing relationships in that they can be employed in relation to any pair of variables, though they are not the most efficient method for some pairs, which is the reason why the method is not recommended in all the cells in Figure 15.6. A contingency table is like a frequency table but it allows two variables to be simultaneously analysed so that relationships between the two variables can be examined. It is normal for contingency tables to include percentages, since these make the tables easier to interpret. Table 15.4 examines the relationship between two variables from the gym survey: gender and reasons for visiting the gym. The percentages are *column percentages*—that is, they calculate the number in each **cell** as a percentage of the total number in that column. Thus, to take the top left-hand cell, the three men who go to the gym for relaxation are 7 per cent of all 42 men in the sample. Users of contingency tables often present the presumed independent variable (if one can in fact be presumed) as the column variable and the presumed dependent variable as the row variable. In this case, we are presuming that gender influences reasons for going to the gym. In fact, we know that going to the gym cannot influence gender. In such circumstances, it is column rather than row percentages that will be required. The procedure for generating a contingency table with SPSS is described on pages 366–7.

Table 15.4

Contingency table showing the relationship between gender and reasons for visiting the gym

Reasons	Gender			
	Male		Female	
	No.	%	No.	%
Relaxation	3	7	6	13
Fitness	15	36	16	33
Lose weight	8	19	25	52
Build strength	16	38	1	2
TOTAL	42		48	

Note: $\chi^2 = 22.726$ $p < 0.0001$

Contingency tables are generated so that patterns of association can be searched for. In this case, we can see clear gender differences in reasons for visiting the gym. As our student anticipated, females are much more likely than men to go to the gym to lose weight. They are also somewhat more likely to go to the gym for relaxation. By contrast, men are much more likely to go to the gym to build strength. There is little difference between the two genders in terms of fitness as a reason.

Pearson's *r*

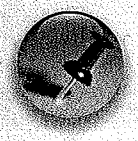
Pearson's *r* is a method for examining relationships between interval/ratio variables. The chief features of this method are as follows:

- the coefficient will almost certainly lie between 0 (zero or no relationship between the two variables) and 1

(a perfect relationship)—this indicates the *strength* of a relationship;

- the closer the coefficient is to 1, the stronger the relationship; the closer it is to 0, the weaker the relationship;
- the coefficient will be either positive or negative—this indicates the *direction* of a relationship.

To illustrate these features consider Tips and skills 'Imaginary data from five variables to show different types of relationship', which gives imaginary data for five variables, and the scatter diagrams in Figures 15.7–15.10, which look at the relationship between pairs of interval/ratio variables. The scatter diagram for variables 1 and 2 is presented in Figure 15.7 and shows a perfect **positive relationship**, which would have a Pearson's r correlation of 1. This means that, as one variable increases, the other variable increases by the same amount and that no other variable is related to either of them. If the correlation was below 1, it would mean that variable 1 is related to at least one other variable as well as to variable 2.



Tips and skills
Imaginary data from five variables to show different types of relationship

Variables 1	2	3	4	5
1	10	50	7	9
2	12	45	13	23
3	14	40	18	7
4	16	35	14	15
5	18	30	16	6
6	20	25	23	22
7	22	20	19	12
8	24	15	24	8
9	26	10	22	18
10	28	5	24	10

The scatter diagram for variables 2 and 3 (see Figure 15.8) shows a perfect **negative relationship**, which would have a Pearson's r correlation of −1. This means that, as one variable increases, the other variable decreases and that no other variable is related to either of them.

If there was no or virtually no correlation between the variables, there would be no apparent pattern to the markers in the scatter diagram. This is the case with the relationship between variables 2 and 5. The correlation is virtually zero at −0.041. This means that the variation in each variable is associated with other variables than the ones present in this analysis. Figure 15.9 shows the appropriate scatter diagram.

If a relationship is strong, a clear patterning to the variables will be evident. This is the case with variables 2 and 4, whose scatter diagram appears in Figure 15.10. There is clearly a positive relationship, and in fact the Pearson's r value is +0.88 (usually, positive correlations are presented without the + sign). This means that the variation in the two variables is very closely connected but that there is some influence of other variables in the extent to which they vary.

Going back to the gym survey, we find that the correlation between age (var00002) and the amount of time spent on weights equipment (var00011) is −0.27 implying a weak negative relationship. This suggests that there is a tendency such that, the older a person is the less likely he or she is to spend much time on such equipment, but that other variables clearly influence the amount of time spent on this activity.

In order to be able to use Pearson's r, the relationship between the two variables must be broadly *linear*—that is, when plotted on a scatter diagram, the values of the

Figure 15.7
Scatter diagram showing a perfect positive relationship

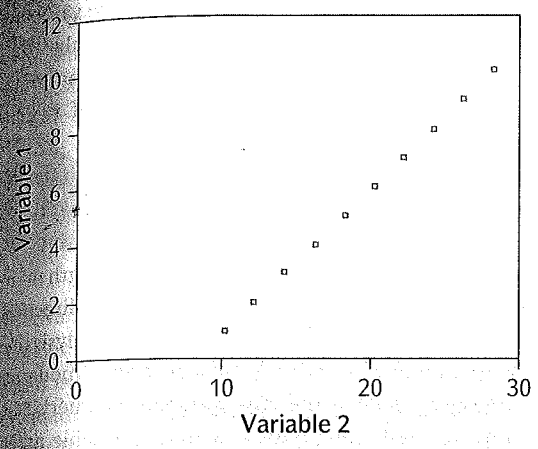


Figure 15.8
Scatter diagram showing a perfect negative relationship

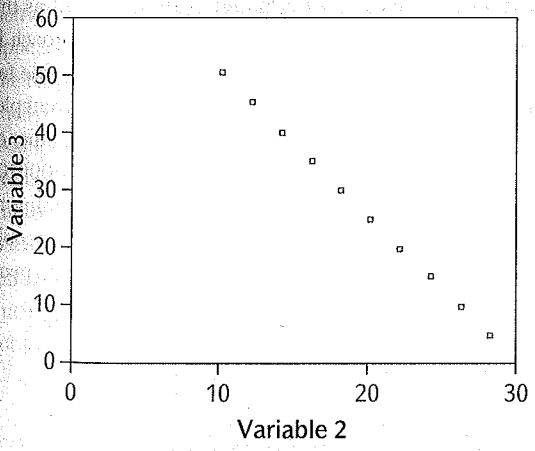


Figure 15.9
Scatter diagram showing two variables that are not related

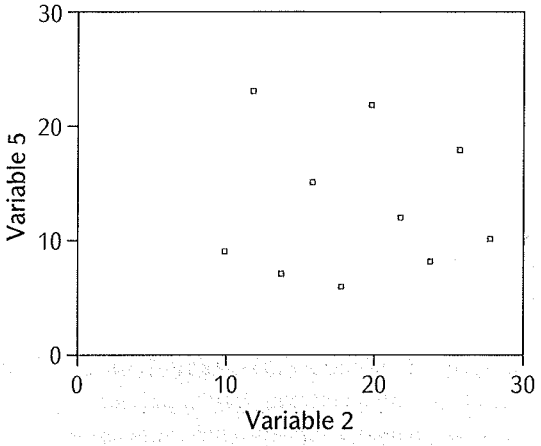
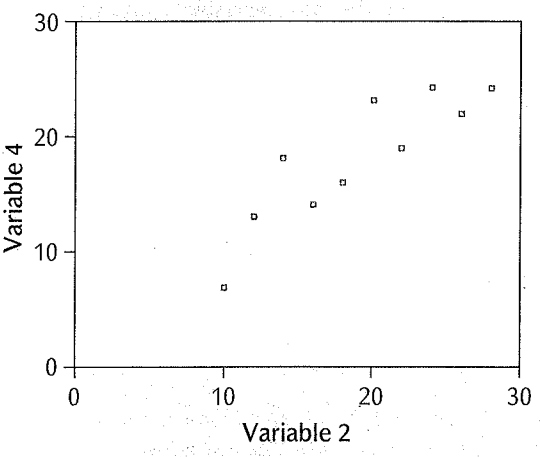


Figure 15.10
Scatter diagram showing a strong positive relationship



two variables approximate to a straight line (even though they may be scattered, as in Figure 15.10) and do not curve. Therefore, plotting a scatter diagram before using Pearson's r is important, in order to determine that the nature of the relationship between a pair of variables does not violate the assumptions being made when this method of correlation is employed.

If you square a value of Pearson's r, you can derive a further useful statistic—namely the *coefficient of determination*, which expresses how much of the variation in one variable is due to the other variable. Thus, if r is -0.27, r^2 is 0.0729. We can then express this as a percentage by multiplying r^2 by 100. The product of this exercise is 7 per cent. This means that just 7 per cent of

the variation in the use of cardiovascular equipment is accounted for by age. The coefficient of determination is a useful adjunct to the interpretation of correlation information.

The procedure for generating Pearson's *r* with SPSS is described on page 368 and the procedure for generating scatter diagrams with SPSS is described on pages 368–72.

Spearman's rho

Spearman's rho, which is often represented with the Greek letter ρ, is designed for the use of pairs of ordinal variables, but is also used, as suggested by Figure 15.6, when one variable is ordinal and the other is interval/ratio. It is exactly the same as Pearson's *r* in terms of the outcome of calculating it, in that the computed value of rho will be either positive or negative and will vary between 0 and 1. If we look at the gym study, there are three ordinal variables: var00004, var00005, and var00006 (see Table 15.1). If we use Spearman's rho to calculate the correlation between the first two variables, we find that the correlation between var00004 and var00005—frequency of use of the cardiovascular and weights equipment—is low at 0.2. A slightly stronger relationship is found between var00006 (frequency of going to the gym) and var00010 (amount of time spent on the cardiovascular equipment), which is 0.4. Note that the latter variable is an interval/ratio variable. When confronted with a situation in which we want to calculate the correlation between an ordinal and an interval/ratio variable, we cannot use Pearson's *r*, because both variables must be at the interval/ratio level of measurement. Instead, we must use Spearman's rho (see Figure 15.6). The procedure for generating Spearman's rho with SPSS is described on page 368.

Phi and Cramér's *V*

Phi (ϕ) and **Cramér's *V*** are two closely related statistics. The phi coefficient is used for the analysis of the relationship between two dichotomous variables. Like Pearson's *r*, it results in a computed statistic that varies between 0 and + or −1. The correlation between var00001 (gender) and var00008 (other sources of regular exercise) is 0.24, implying that males are somewhat more likely than females to have other sources of regular exercise, though the relationship is weak.

Cramér's *V* uses a similar formula to phi and can be employed with nominal variables (see Figure 15.6). However, this statistic can take on only a positive value, so that it can give an indication only of the strength of the relationship between two variables, not of the direction. The value of Cramér's *V* associated with the analysis presented in Table 15.4 is 0.50. This suggests a moderate relationship between the two variables. Cramér's *V* is usually reported along with a contingency table and a **chi-square test** (see below). It is not normally presented on its own. The procedure for generating phi and Cramér's *V* with SPSS is described on pages 366–7.

Comparing means and eta

If you need to examine the relationship between an interval/ratio variable and a nominal variable, and if the latter can be relatively unambiguously identified as the independent variable, a potentially fruitful approach is to compare the means of the interval/ratio variable for each subgroup of the nominal variable. As an example, consider Table 15.5, which presents the mean number of minutes spent on cardiovascular equipment (var00010) for each of the four categories of reasons for going to the gym (var00003). The means suggest that people who go to the gym for fitness or to lose weight spend

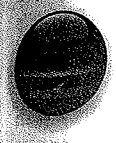
Table 15.5

Comparing subgroup means: time spent on cardiovascular equipment by reasons for going to the gym

Time	Reasons				
	Relaxation	Fitness	Lose weight	Build strength	Total
Mean number of minutes spent on cardiovascular equipment	18.33	30.55	28.36	19.65	26.47
n	9	31	33	17	90

considerably more time on this equipment than people who go to the gym to relax or to build strength.

This procedure is often accompanied by a test of association between variables called **eta**. This statistic expresses the level of association between the two variables and, like Cramér's *V*, will always be positive. The level of eta for the data in Table 15.5 is 0.48. This suggests a moderate relationship between the two variables. Eta-squared expresses the amount of variation in the interval/ratio variable that is due to the nominal variable. In the case of this example, eta-squared is 22 per cent. Eta is a very flexible method for exploring the relationship between two variables, because it can be employed when one variable is nominal and the other interval/ratio. Also, it does not make the assumption that the relationship between variables is linear. The procedure for comparing means and for generating eta with SPSS is described on page 372.



Multivariate analysis

Multivariate analysis entails the simultaneous analysis of three or more variables. This is quite an advanced topic, and it is recommended that readers examine a textbook on quantitative data analysis for an exposition of techniques (e.g. Bryman and Cramer 2011). There are three main contexts within which multivariate analysis might be employed.

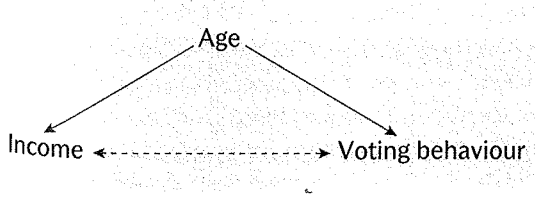
Could the relationship be spurious?

In order for a relationship between two variables to be established, not only must there be evidence that there is a relationship but the relationship must be shown to be *non-spurious*. A **spurious relationship** exists when there appears to be a relationship between two variables, but the relationship is not real: it is being produced because each variable is itself related to a third variable. For example, if we find a relationship between income and voting behaviour, we might ask: could the relationship be an artefact of age (see Figure 15.11)? The older one is, the more likely one is to earn more, while age is known to influence voting behaviour. If age were found to be producing the apparent relationship between income and voting behaviour, we would conclude that the relationship is spurious. In this case, the variable age would be known as a **confounding variable**.

An interesting possible case of a **spurious relationship** was highlighted in a very short report in *The Times* (1 October 1999, p. 2) of some medical findings. The article noted that there is evidence to suggest that women on hormone replacement therapy (HRT) have lower levels of heart disease than those not on this form of therapy. The article cites Swedish findings that suggest that the relationship may be due to the fact that women who choose to start the therapy are 'thinner, richer and healthier' than those who do not. These background factors would seem to affect both the likelihood of taking HRT *and* the likelihood of getting heart disease. A further illustration in connection with a health-related issue comes from another *Times* article (Hawkes 2003), which reports a relationship among men between frequency of shaving and likelihood of a heart attack or stroke. The reason appears to be that each of the variables (frequency of shaving and vulnerability to a heart attack or stroke) is affected by lifestyle and hormonal factors.

Figure 15.11

A spurious relationship



Could there be an intervening variable?

Let us say that we do not find that the relationship is spurious; we might ask *why* there is a relationship between two variables. For example, there have been several studies that have explored the relationship between an organization's market orientation and its business performance. However, the mixed nature of the findings to have emerged from these studies led Piercy, Harris, and Lane (2002) to suggest that there is a more complex relationship between these two variables than previous studies have assumed. In particular, they speculated that higher levels of market orientation are associated with higher

levels of employee motivation, satisfaction, and commitment, which in turn leads to enhanced organizational performance. Employee attitudes are thus an **intervening variable**:

market orientation	→	employee attitudes	→	organizational performance

An intervening variable allows us to answer questions about the bivariate relationship between variables. It suggests that the relationship between the two variables is not a direct one, since the impact of market orientation on organizational performance is viewed as occurring via employee attitudes.

Could a third variable moderate the relationship?

We might ask a question like: does the relationship between two variables hold for men but not for women? If it does, the relationship is said to be moderated by gender. We might ask in the gym study, for example, if the relationship between age and whether visitors have other sources of regular exercise (var00008) is moderated by gender. This would imply that, if we find a pattern relating age to other sources of exercise, that pattern will vary by gender. Table 15.6 shows the relationship between age and other sources of exercise. In this table, age has been broken down into just three age bands to make the table easier to read. The table suggests that the 31–40 age group is less likely to have other sources of regular exercise than the 30 and under and 41 and over age groups. However, Table 15.7, which breaks the relationship down by gender, suggests that the pattern for males and females is somewhat different. Among males, the pattern shown in Table 15.6 is very pronounced, but for females the likelihood of having other sources of exercise declines with gender. It would seem that the relationship between age and other sources of exercise is a **moderated relationship** because it is moderated by gender. This example illustrates the way in which contingency tables can be employed for multivariate analysis. However, there is a wide variety of other techniques (Bryman and Cramer 2011: ch. 10). The procedure for conducting such an analysis with SPSS is described on pages 372–3.

Table 15.6

Contingency table showing the relationship between age and whether or not gym visitors have other sources of regular exercise (%)

Other source of exercise	Age		
	30 and under	31–40	41 and over
Other source	64	43	58
No other source	36	57	42
n	42	23	24

Table 15.7

Contingency table showing the relationship between age and whether or not gym visitors have other sources of regular exercise for males and females (%)

Other source of exercise	Gender					
	Male			Female		
	30 and under	31–40	41 and over	30 and under	31–40	41 and over
Other source	70	33	75	59	50	42
No other source	30	67	25	41	50	58
n	20	9	12	22	14	12

Statistical significance

difficulty with working on data deriving from a sam-
s that there is often the lingering worry that, even
gh you have employed a probability sampling pro-
re (as in the gym survey), your findings will not be
eralizable to the population from which the sample
drawn. As we saw in Chapter 8, there is always the
sibility that **sampling error** (difference between the
ulation and the sample that you have selected) has
urred, even when probability sampling procedures
e been followed. If this happens, the sample will be
epresentative of the wider population and therefore
findings will be invalid. To make matters worse,
re is no feasible way of finding out whether or not
y do in fact apply to the population! What you can do
rovide an indication of how confident you can be in
ur findings. This is where **statistical significance** and
various tests of statistical significance come in.

We need to know how confident we can be that our findings can be generalized to the population from which that sample was selected. Since we cannot be absolutely certain that a finding based on a sample will also be found in the population, we need a technique that allows us to establish how confident we can be that the finding exists in the population and what risk we are taking in inferring that the finding exists in the population. These two elements—confidence and risk—lie at the heart of tests of statistical significance (see Key concept 15.1). However, it is important to appreciate that tests of statistical significance can be employed only in relation to samples that have been drawn using probability sampling. The process of inferring findings from a probability sample to the population from which it was selected is known as **statistical inference**.

Key concept 15.1
What is a test of statistical significance?

A test of statistical significance allows the analyst to estimate how confident he or she can be that the results deriving from a study based on a randomly selected sample are generalizable to the population from which the sample was drawn. When examining statistical significance in relation to the relationship between two variables, it also tells us about the risk of concluding that there is in fact a relationship in the population when there is no such relationship in the population. If an analysis reveals a statistically significant finding, this does not mean that the finding is intrinsically significant or important. The word 'significant' seems to imply importance. However, statistical significance is solely concerned with the confidence researchers can have in their findings. It does not mean that a statistically significant finding is substantively significant.

In Chapter 8 (see Tips and skills 'Generalizing from a random sample to the population'), in the context of the discussion of the standard error of the mean, we began to get an appreciation of the ideas behind statistical significance. For example, we know that the mean age of the gym sample is 33.6. Using the concept of the standard error of the mean, we can calculate that we can be 95 per cent confident that the population mean lies between 31.72 and 35.47. This suggests that we can determine in broad outline the degree of confidence that we can have in a sample mean.

In the rest of this section, we will look at the tests that are available for determining the degree of confidence we can have in our findings when we explore relationships between variables. All of the tests have a common structure.

- *Set up a null hypothesis.* **A null hypothesis** stipulates that two variables are not related in the population—for example, that there is *no* relationship between gender and visiting the gym in the population from which the sample was selected.

- *Establish the level of statistical significance that you find acceptable.* This is essentially a measure of the degree of risk that you might reject the null hypothesis (implying that there *is* a relationship in the population) when you should support it (implying that there is no relationship in the population). Levels of statistical significance are expressed as probability levels—that is, the probability of rejecting the null hypothesis when you should be confirming it. See Key concept 15.2 on this issue. The convention among most social researchers is that the maximum level of statistical significance that is acceptable is $p < 0.05$, which implies that there are fewer than 5 chances in 100 that you could have a sample that shows a relationship when there is not one in the population.
- *Determine the statistical significance of your findings* (that is, use a statistical test like chi-square—see below).
- If your findings are statistically significant at the 0.05 level—so that the risk of getting a relationship as strong as the one you have found, when there is *no* relationship in the population, is no higher than 5 in 100—you would *reject* the null hypothesis. Therefore, you are implying that the results are unlikely to have occurred by *chance*.

Key concept 15.2
What is the level of statistical significance?

The level of statistical significance is the level of risk that you are prepared to take that you are inferring that there is a relationship between two variables in the population from which the sample was taken when in fact no such relationship exists. The maximum level of risk that is conventionally taken in social research is to say that there are up to 5 chances in 100 that we might be falsely concluding that there is a relationship when there is not one in the population from which the sample was taken. This means that, if we drew 100 samples, we are recognizing that as many as 5 of them might exhibit a relationship when there is not one in the population. Our sample might be one of those 5, but the risk is fairly small. This significance level is denoted by $p < 0.05$ (p means probability). If we accepted a significance level of $p < 0.1$, we would be accepting the possibility that as many as 10 in 100 samples might show a relationship where none exists in the population. In this case, there is a greater risk than with $p < 0.05$ that we might have a sample that implies a relationship when there is not one in the population, since the probability of our having such a sample is greater when the risk is 1 in 10 (10 out of 100 when $p < 0.1$) than when the risk is 1 in 20 (5 out of 100 when $p < 0.05$). Therefore, we would have greater confidence when the risk of falsely inferring that there is a relationship between 2 variables is 1 in 20, as against 1 in 10. But, if you want a more stringent test, perhaps because you are worried about the use that might be made of your results, you might choose the $p < 0.01$ level. This means that you are prepared to accept as your level of risk a probability of only 1 in 100 that the results could have arisen by chance (that is, due to sampling error). Therefore, if the results, following administration of a test, show that a relationship is statistically significant at the $p < 0.05$ level, but *not* the $p < 0.01$ level, you would have to confirm the null hypothesis.

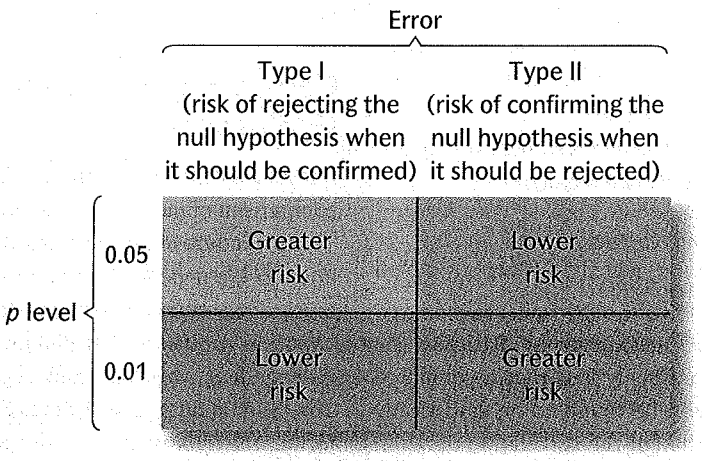
There are in fact two types of error that can be made when inferring statistical significance. These errors are known as Type I and Type II errors (see Figure 15.12). A Type I error occurs when you reject the null hypothesis when it should in fact be confirmed. This means that your results have arisen by chance and you are falsely concluding that there is a relationship in the population when there is not one. Using a $p < 0.05$ level of significance means that we are more likely to make a Type I error than when using a $p < 0.01$ level of significance. This is because with 0.01 there is less chance of falsely rejecting the null hypothesis. However, in doing so, you increase the chance of making a Type II error (accepting the null hypothesis when you should reject it). This is because you are more likely to confirm the null hypothesis when the significance level is 0.01 (1 in 100) than when it is 0.05 (1 in 20).

The chi-square test

The chi-square (χ^2) test is applied to contingency tables like Table 15.4. It allows us to establish how confident we can be that there is a relationship between the two variables in the population. The test works by calculating

Figure 15.12

Type I and Type II errors



for each cell in the table an expected frequency or value—that is, one that would occur on the basis of chance alone. The chi-square value, which in Table 15.4 is 22.726, is produced by calculating the differences between the actual and expected values for each cell in the table and then summing those differences (it is slightly more complicated than this, but the details need not concern us here). The chi-square value means nothing on its own and can be meaningfully interpreted only in relation to its associated level of statistical significance, which in this case is $p < 0.0001$. This means that there is only 1 chance in 10,000 of falsely rejecting the null hypothesis (that is, inferring that there is a relationship in the population when there is no such relationship in the population). You could be extremely confident that there is a relationship between gender and reasons for visiting the gym among all gym members, since the chance that you have obtained a sample that shows a relationship when there is no relationship among all gym members is 1 in 10,000.

Whether or not a chi-square value achieves statistical significance depends not just on its magnitude but also on the number of categories of the two variables being analysed. This latter issue is governed by what is known as the 'degrees of freedom' associated with the table. The number of degrees of freedom is governed by the simple formula:

Number of degrees of freedom
= (number of columns – 1)(number of rows – 1).

In the case of Table 15.4, this will be (2 – 1)(4 – 1)—that is, 3. In other words, the chi-square value that is arrived at is affected by the size of the table, and this is taken into account when deciding whether the chi-square value is statistically significant or not. The procedure for chi-square in conjunction with a contingency table with SPSS is described on pages 366–7.

Correlation and statistical significance

Examining the statistical significance of a computed correlation coefficient, which is based on a randomly selected sample, provides information about the likelihood that the coefficient will be found in the population from which the sample was taken. Thus, if we find a correlation of –0.62, what is the likelihood that a relationship of at least that size exists in the population? This tells us if the relationship could have arisen by chance.

If the correlation coefficient r is –0.62 and the significance level is $p < 0.05$, we can reject the null hypothesis that there is no relationship in the population. We can infer that there are only 5 chances in 100 that a correlation of at least –0.62 could have arisen by chance alone. You *could* have 1 of the 5 samples in 100 that shows a relationship when there is not one in the population, but the degree of risk is reasonably small. If, say, it was found that $r = -0.62$ and $p < 0.1$, there could be as many

as 10 chances in 100 that there is no correlation in the population. This would *not* be an acceptable level of risk for most purposes. It would mean that in as many as 1 sample in 10 we might find a correlation of −0.62 or above when there is not a correlation in the population. If $r = -0.62$ and $p < 0.001$, there is only 1 chance in 1,000 that no correlation exists in the population. There would be a very low level of risk if you inferred that the correlation had not arisen by chance.

Whether a correlation coefficient is statistically significant or not will be affected by two factors:

1. the size of the computed coefficient; and
2. the size of the sample.

This second factor may appear surprising. Basically, the larger a sample, the more likely it is that a computed correlation coefficient will be found to be statistically significant. Thus, even though the correlation between age and the amount of time spent on weights machines in the gym survey was found to be just −0.27, which is a fairly weak relationship, it is statistically significant at the $p < 0.01$ level. This means that there is only 1 chance in 100 that there is no relationship in the population. Because the question of whether or not a correlation coefficient is statistically significant depends so much on the sample size, it is important to realize that you should always examine *both* the correlation coefficient *and* the significance level. You should not examine one at the expense of the other.

This treatment of correlation and statistical significance applies to both Pearson's r and Spearman's rho. A similar interpretation can also be applied to phi and Cramér's V. SPSS automatically produces information regarding statistical significance when Pearson's r, Spearman's rho, phi, and Cramér's V are generated.

Comparing means and statistical significance

A test of statistical significance can also be applied to the comparison of means that was carried out in Table 15.5. This procedure entails treating the total amount of variation in the dependent variable—amount of time spent on cardiovascular equipment—as made up of two types: variation *within* the four subgroups that make up the independent variable, and variation *between* them. The latter is often called the *explained variance* and the former the *error variance*. A test of statistical significance for the comparison of means entails relating the two types of variance to form what is known as the F statistic. This statistic expresses the amount of explained variance in relation to the amount of error variance. In the case of the data in Table 15.5, the resulting F statistic is statistically significant at the $p < 0.001$ level. This finding suggests that there is only 1 chance in 1,000 that there is no relationship between the two variables among all gym members. SPSS produces information regarding the F statistic and its statistical significance if the procedures described on page 372 are followed.

Checklist

Doing and writing up quantitative data analysis

- ○ Have you answered your research questions?
- ○ Have you made sure that you have presented only analyses that are relevant to your research questions?
- ○ Have you made sure that you have taken into account the nature of the variable(s) being analysed when using a particular technique (that is, whether nominal, ordinal, interval/ratio, or dichotomous)?
- ○ Have you used the most appropriate and powerful techniques for answering your research questions?
- ○ If your sample has *not* been randomly selected, have you made sure that you have not made inferences about a population (or at least, if you have done so, have you outlined the limitations of making such an inference?)?
- ○ If your data are based on a cross-sectional design, have you resisted making unsustainable inferences about causality?

- ○ Have you remembered to code any missing data?
- ○ Have you commented on all the analyses you present?
- ○ Have you gone beyond univariate analysis and conducted at least some bivariate analyses?
- ○ If you have used a Likert scale with reversed items, have you remembered to reverse the coding of them?

Key points

- You need to think about your data analysis before you begin designing your research instruments.
- Techniques of data analysis are applicable to some types of variable and not others. You need to know the difference between nominal, ordinal, interval/ratio, and dichotomous variables.
- You need to think about the kinds of data you are collecting and the implications your decisions will have for the sorts of techniques you will be able to employ.
- Become familiar with computer software like SPSS before you begin designing your research instruments, because it is advisable to be aware at an early stage of difficulties you might have in presenting your data in SPSS.
- Make sure you are thoroughly familiar with the techniques introduced in this chapter and when you can and cannot use them.
- The basic message, then, is not to leave these considerations until your data have been collected, tempting though it may be.
- Do not confuse statistical significance with substantive significance.

Questions for review

- At what stage should you begin to think about the kinds of data analysis you need to conduct?
- What are missing data and why do they arise?

Types of variable

- What are the differences between the four types of variable outlined in this chapter: interval/ratio; ordinal; nominal; and dichotomous?
- Why is it important to be able to distinguish between the four types of variable?
- Imagine the kinds of answers you would receive if you administered the following four questions in an interview survey. What kind of variable would each question generate: dichotomous; nominal; ordinal; or interval/ratio?

1. Do you enjoy going shopping?

   Yes ___

   No ___

2. How many times have you shopped in the last month? Please write in the number of occasions below.

   ___

3. For which kinds of items do you most enjoy shopping? Please tick one only.

   Clothes (including shoes) ____
   Food ____
   Things for the house ____
   Presents ____
   Entertainment (CDs, videos, etc.) ____

4. How important is it to you to buy clothes with designer labels?

   Very important ____
   Fairly important ____
   Not very important ____
   Not at all important ____

Univariate analysis

- What is an outlier and why might one have an adverse effect on the mean and the range?
- In conjunction with which measure of central tendency would you expect to report the standard deviation: the mean; the median; or the mode?

Bivariate analysis

- Can you infer causality from bivariate analysis?
- Why are percentages crucial when presenting contingency tables?
- In what circumstances would you use each of the following: Pearson's *r*; Spearman's rho; phi; Cramér's *V*; eta?

Multivariate analysis

- What is a spurious relationship?
- What is an intervening variable?
- What does it mean to say that a relationship is moderated?

Statistical significance

- What does statistical significance mean and how does it differ from substantive significance?
- What is a significance level?
- What does the chi-square test achieve?
- What does it mean to say that a correlation of 0.42 is statistically significant at $p < 0.05$?

Online Resource Centre

www.oxfordtextbooks.co.uk/orc/brymansrm4e/

Visit the Online Resource Centre that accompanies this book to enrich your understanding of quantitative data analysis. Consult web links, test yourself using multiple choice questions, and gain further guidance and inspiration from the Student Researcher's Toolkit.