3 Summing Things up: Graphs, Averages, Standard Deviations, Correlations and so on

The starting point for a statistical analysis is typically data on a number of 'variables' (for example sex, weight, age) relating to a sample of 'cases' (for example people). This chapter looks at how we can summarise the pattern of a single variable, and the relationship between pairs of variables, by means of a few well-chosen numbers or graphs, and how these summaries can be interpreted.

▶ 3.1 Introduction

A few years ago (in 1993) I wanted to find out how much students drank. Is the stereotype of the drunken student accurate? To get some information on this, I asked a short series of questions to all the students in three groups: a full-time business course (FB in Table 3.1), a full-time course on personnel management (FP) and a part-time course on personnel management (PP). These included: How many units of alcohol did you drink last Saturday? There were similar questions for Sunday and Monday. A unit of alcohol is the amount in one glass of wine, or half a pint of beer. (UK government guidelines suggest that the maximum alcohol consumption, from a health point of view, should be 14 units a week for women, and 21 for men.) I also asked each student for their estimate of the average number of cigarettes they smoked each day.

There were 92 students in the three groups. This is too many to show you here, so I have taken a random sample of 20 of the 92, and put this data in Table 3.1. The data from all 92 students is on the web as `drink.xls`, and the group of 20 as `drink20.xls`. The data in Table 3.1 seems to confirm a few stereotypes. The figures for units drunk on Saturday night have an overall average of 5.4, well above the 2 or 3 units a night implied by the government guidelines. The average consumption for males was 13.4 units, and for females 3.3 units: males seem to drink much more than females. Similarly, the average Saturday consumption among the full-time students (7.6 units) is more than the consumption of the part-time students (5.7 units), most of

Table 3.1 Drink data from 20 students

Sex	*Age*	*Course*	*Satunits*	*Sununits*	*Monunits*	*Daycigs*
F	24	FP	0	0	0	0
M	20	FB	26	18	26	20
F	32	FP	1	3	1	0
F	32	PP	5	2	0	25
F	21	FP	3	2	0	0
M	20	FB	3	2	0	5
F	19	FB	1	2	0	5
F	21	FB	2	2	6	0
M	21	FB	6	8	8	0
M	19	FB	4	10	6	0
M	22	FB	19	0	15	7
M	23	FB	15	0	18	0
F	21	FP	5	5	0	0
F	21	FP	0	0	0	0
M	19	FB	24	6	24	0
F	21	FP	4	2	0	0
F	38	PP	0	0	0	0
F	32	PP	12	3	0	0
M	22	FB	10	8	20	0
F	19	FB	7	1	7	0

Satunits refers to the number of units of alcohol drunk on the Saturday, Sununits to the number on Sunday, and Monunits to the number on Monday. Daycigs refers to the average daily number of cigarettes smoked. This format, with short variable names, is convenient for exporting data to SPSS. The data is in *drink20.xls*.

whom had full-time jobs as well as their studies, and so were not 'proper' students. However, this difference is not as much as I had expected.

▶ 3.2 What can a sample tell us about?

As we saw in Section 2.5, samples are useful because they tell you about some wider context. I am not just interested in how much the students in three of my classes drank on three days in 1993. I am also interested in the drinking habits of other students at other times. We can visualise this in terms of buckets and balls. Imagine a bucket with balls in it to represent all 'student-days' which were lived in 1993. We need to define 'student', so let's take all full-time and part-time students in UK higher education in 1993. If there were, say, a million of these, this would mean a total of 365 million student-days. My data file from 92 students includes data on 276 (92 × 3) of these student-days, of which 60 appear in Table 3.1. How accurate a picture of the whole population of 365 million student-days will this data provide**?**

The picture is obviously unlikely to be fully accurate. One problem is the size of the sample. The figures in Section 3.1 about units drunk on Saturday are based on a sample of 20 student-days. The corresponding figures from all 92 students in the sample were slightly different. For example, the sample of 20 gives an average number of units drunk by full-time students of 7.6, whereas the full sample of 92 gives a figure of 5.9. Obviously, if we could get data on *all* full-time student-days in the UK in 1993, the answer would almost certainly be slightly different. We'll look at how we can estimate the likely size of these differences in Chapter 7.

A more serious problem is that I chose the sample in a haphazard way that makes it difficult to feel confident that the pattern in the sample will reflect the wider population. I had one part-time and two full-time classes, but does this reflect the general pattern? Probably not. And as I only had three classes, the sample is likely to be less varied than the whole population. I asked about a Saturday, Sunday and Monday, but perhaps students drink less on other days? And I chose just one week; if this were just before Christmas, for example, the results might be higher than they would otherwise have been. (The best way to choose a sample like this would be to choose randomly from the whole population – see Section 2.5 – but this was impractical, as it often is.)

There are yet more problems. The students who answered my questions may not have been telling the truth; one 19-year-old girl in the sample of 92 said she drank a total of 140 units (70 pints of beer) over the three days! And there is the time problem: to what extent can data gathered in the past tell us about the future? All this means that you should be very cautious about reading too much into this data. All we can do is make a judgement about the extent to which patterns found in the sample can be extrapolated to a wider context. The same problem applies to many other samples which are used as the basis of statistical analysis.

However, having said all this, for the purposes of the rest of this chapter, I'll assume that we can extrapolate the results from the sample of 92 to a wider context. I'll use the word 'students' for this slightly vague, wider context.

▶ 3.3 Variables, cases and units of analysis

The columns in Table 3.1, Sex, Age, Course and so on, represent 'variables'. These are characteristics which *vary* from person to person in the sample.

The first row in Table 3.1 gives the *values* of all the variables for the first person in the sample (F, 24, FP and so on), the second row gives the values for the second person and so on.

There are two important types of variable: 'category' and 'number'. Sex and Course are category variables: the values are categories like Male or Female. Age and the rest of the variables are number variables: their values are numbers. This is a very obvious, but very important distinction. (There are further distinctions mentioned in the Similar concepts section.)

The 'unit of analysis' in this survey is the student. Each student in the survey, represented by a row in Table 3.1, is known as a 'case'. In this survey the cases are people, but in other surveys they could be countries, accidents or days.

Sections 3.4–3.8 are organised according to the number and type of the variables whose values you want to summarise.

▶ 3.4 Summarising a single variable

The information in Table 3.1 is a bit of a mess. It is not easy to see any patterns. In this section I'll look at ways of summing up the pattern of a single variable. This pattern is often referred to as a 'distribution'. If the variable is a category variable, then all you can do is count up the number of people in each category and present the results in a table, bar chart or pie diagram. You can also use this data to estimate probabilities as described in the last chapter. This is too obvious to bother with an example, but we will look at the two variable equivalent in Section 3.5. If the variable is a number variable, there are more possibilities.

3.4.1 Histograms

A histogram is a bar chart to show the frequencies of different values of a number variable. This is a very useful sort of diagram, often overlooked by beginners. Figures 3.1 and 3.2, both based on the sample of 92, illustrate two histograms.

In Figure 3.1, the bars represent the number of students who drank no units, one unit, two units and so on. The word 'frequency' in this context means simply how frequently something happened, in other words, how many of the students each bar applies to. One 20-year-old girl in the sample claimed to have drunk 40 units (20 pints of beer) on the Saturday. This is far more than the rest of the students in the sample and is excluded from Figure 3.1 because I think it isn't true. Drinking this much would be dangerous. Obviously you should consider such 'outliers' carefully, and only exclude them if there is good reason.

Figure 3.1 Histogram of units drunk on Saturday based on sample of 92 (one outlier excluded)

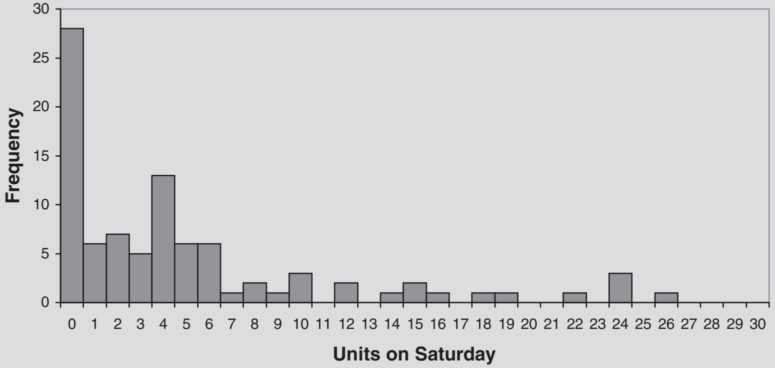


Figure 3.2 Histogram of estimated total weekly units drunk based on sample of 92 (three outliers excluded)

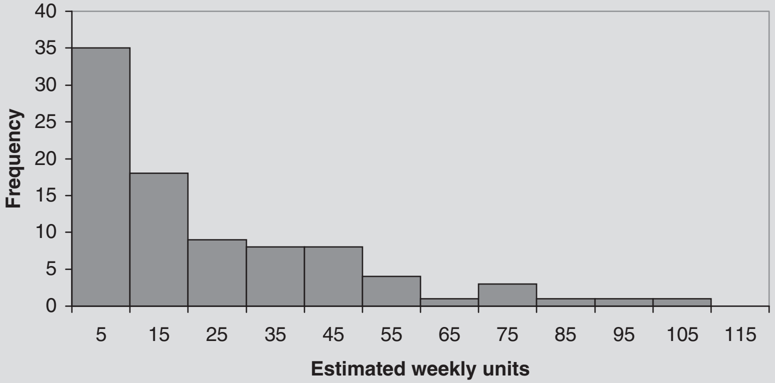


Figure 3.2 shows the estimated weekly consumption based on this sample. For each student the average over the three days is calculated, and then multiplied by seven to give an estimate for the week; this is easy with Excel.[29] Do you think this is likely to be an accurate estimate, or is it likely to be too high or too low**?**

I think it may be too high if students drink more at the weekends than during the rest of the week. But it's the best we can do with this data.

Table 3.2 Frequencies for Figure 3.2

Top of interval	*Middle of interval*	*Frequency*
10	5	35
20	15	18
30	25	9
40	35	8
50	45	8
60	55	4
70	65	1
80	75	3
90	85	1
100	95	1
110	105	1
120	115	0
Above 120 (excluded from Figure 3.2)		3

In Figure 3.2 each bar represents not the frequency of a single number of units, but the frequency of a range of numbers. The first bar, for example, is centred on 5 units and obviously includes all quantities from zero up to 10. Figure 3.2 is based on Table 3.2. This shows the frequencies in each 'interval'. As an example, look at the second row of this table. The top of the interval is 20, and the top of the interval below is 10. The interval in question includes all the numbers above 10 and up to and including 20: eg 11.7, 18.7 and 20, but not 21 or 10 (which is in the interval below). The frequency for this interval is 18: there were 18 students in the sample who drank an amount in this range. The middle points of each interval are used as the labels for each bar on the horizontal axis.

It should be obvious how to draw histograms like Figure 3.2 by hand. Using Excel, you can, of course, automate this process.[30] It's even easier with SPSS.[31] The main purpose of a histogram is to get a feel for the general pattern. Have you any comments on the general pattern of Figure 3.2**?**

The pattern is roughly what I would have expected: more than 50% drink modestly – less than 20 units a week – but there is a 'tail' of bigger drinkers. It is often useful to compare histograms for different groups, for example males versus females or full timers versus part timers. Histograms like Figure 3.2 are far more useful than those which show frequencies of individual values like Figure 3.1 whenever there are a large number of different possible values. The scale in Figure 3.2 goes up to 120, and fractions are also possible. A diagram like Figure 3.1, with a bar for each value, would be messy and uninformative. If you don't believe me, draw it!

3.4.2 Averages: the mean and the median of a number variable

Sometimes it is helpful to sum up a distribution as a single number to show where the centre or 'average' is. There are two commonly used, and useful, measures: the mean and the median. The 'mean' is simply the ordinary average of everyday life. To work it out you simply add up all the values in the sample and divide by the number of values. In the case of the data on units drunk on Saturday night in Table 3.1, these values are the number of units drunk by each individual student, and the number of values is, of course, 20. What is the mean number of units drunk on Saturday night by the sample in Table 3.1**?**

The mean is 7.4 units of alcohol. The mean of the whole sample of 92 students is 5.4 units. In ordinary English, of course, this is simply the average. Excel follows this usage: the function for the mean is average.

The 'median' is simply the middle value when the values are arranged in order of size. What is the median number of units drunk on Saturday night by the sample in Table 3.1**?**

To work this out you need to put the values in order of size. There are three zeros, so they come first: 0, 0, 0, 1, 1, 2, 3, 3, 4, 4, 5, 5 and so on. The median is the middle value in this list. There are, however, an even number of values, 20, so there is no middle value. The middle is between the 10th (4) and the 11th (5), so we take the value midway between these two: the median is 4.5 units of alcohol. Half the values (ten of them) are below 4.5 and the other half (the other ten) are above 4.5, so 4.5 is squarely in the middle.

As with the mean, this is not quite the same as the median of the whole sample of 92, which comes to 3.5 units. Which do you think is the more useful measure, the mean or the median**?**

It depends on the situation and on what you want to know. If you were buying the drinks for a group of students on Saturday night, the mean is a useful measure if you want an idea of how much the drinks will cost. The mean is obtained by adding up all the units drunk and then imagining them shared out equally between the students, each student then gets the mean number of units. If you were buying drinks for a group of 20 students, you would expect them – assuming they are average students – to drink about 20 times this mean, or 108 units of alcohol (using the mean of 5.4 from the whole sample of 92 students, as this is more likely to be representative of students in general). You could then estimate how much this will cost. This won't be exact, but the best sort of average you can use here is the mean.

If, on the other hand, you want some idea of the state of inebriation of a typical student, then the median is more use. The median of 3.5 means that half the students you meet on a Saturday night are likely to have more to drink than 3.5 units and the other half less. (In our sample of 92 students, 46 drank less than 3.5 units on Saturday and 46 more than 3.5.) The idea of quartiles and percentiles (see below) extends this idea further. In practice, the median is an underused measure. It is often more meaningful than the mean. Don't forget it.

In our sample, the median is less than the mean. Why is this**?**

Figures 3.1 and 3.2 show that the distribution is not symmetrical. The values tail off more gradually at the top end than they do at the bottom. The distribution is said to be 'skewed', which is the reason for the difference between the mean and the median. It may be obvious to you why this should be, but if not, imagine working out the mean by taking the numbers in pairs – first the biggest and the smallest: 26 and 0 which have a mean of 13, second the next biggest and next smallest: 24 and 0 (there are two 0s in the data) which have a mean of 12 and so on. You will end up with 10 numbers, none of which are less than 4.5 (the median of the 20 values in Table 3.1), so the overall mean must be more than 4.5.

With some data, only whole numbers make sense. Imagine a country where the mean number of offspring people have is 2.7. Any particular person might have 2 children or 3, but 2.7 children is obviously impossible. Is it meaningful and sensible to talk about a mean of 2.7 here**?**

Yes. If there were a million people, the mean of 2.7 enables you to predict that they will have about 1.35 million children, which helps to predict population trends (remembering that children have two parents). Means are helpful because they enable you to estimate totals, and if the total is useful, then so is the mean. However, the mean obviously does *not* represent a typical family. Needless to say, Excel and SPSS can be used to work out means and medians.[32]

3.4.3 Maximum and minimum, quartiles and percentiles of a number variable

These concepts provide another way of summing up the pattern. The idea is very simple: arrange the numbers in order of size – just as we did for the median – and then read off

- the maximum and minimum
- the value halfway up the list (the median)
- the values a quarter and three-quarters of the way up the list (the 'first quartile' and the 'third quartile')
- the value 10% of the way up the list (the '10th percentile'). Similarly the value 80% of the way up the list is the 80th percentile. The first quartile is the same as the 25th percentile. And so on.

What are the quartiles of the number of units drunk on Saturday night by the sample whose alcohol consumption is shown in Table 3.1**?** What is the 80th percentile**?** What is the maximum**?**

The first step is the arrange the data in order of size, just as we did for the median (0, 0, 0, 1, 1, 2, 3, 3, 4, 4, 5, 5, 6, 7, 10, 12, 15, 19, 24, 26). There are 20 numbers in the list, so a quarter of the 20 is 5 and the other 15, starting from the sixth in the list, are the remaining three-quarters. The fifth number is 1 and the sixth is 2. The obvious thing then is, just like the median, to say that the first quartile is midway between these two, that is, 1.5 units of alcohol. Similarly the third quartile is 11 (midway between the 15th and 16th numbers: 10 and 12), and the 80th percentile is midway between the 16th and the 17th number (12 and 15), that is, 13.5.

You may wonder what happens if we want the 78th percentile, 78% of the way up the list. Where exactly is this? Presumably a bit less than 13.5, but how much less? You may also be a bit confused about why I chose midway between the fifth and the sixth number for the first quartile; I've explained above, but you may not be convinced. If so, please don't worry. These points are not very important. Quartiles and percentiles are particularly useful with large samples, and with large samples it usually makes little difference if you choose one number, or the next, or something between them. Imagine a sample of 1000 students. Strictly, the first quartile is midway between the 250th and 251st number in the list of 1000, but these are very likely to be the same or only very slightly different. So it really doesn't matter much.

Excel[33] and SPSS[34] have exact definitions of quartiles and percentiles built in; you could try analysing the data in Table 3.1 with these packages to see what answers they give. You will probably find that you get *three* different answers, because these packages use more subtle methods than the midway method above.

Quartiles and percentiles are useful because they are easy to interpret. The first quartile is 1.5 units: this means that 25% of students drank less than 1.5 units, and the other 75% drank more. Similarly the 80th percentile consumption (13.5 units) is the level of consumption that is exceeded by 20% of students. And from another survey, it was found that the 95th percentile of the heights of 17-year-old men in the UK in the 1990s was about 190 cm;[35] this is the height of a man towards the top of the spectrum of heights, 95% of the way up, to be precise. Only 5% of 17-year-old men were taller than this.

3.4.4 Measures of the spread of a number variable: range, interquartile range, mean deviation and standard deviation

Histograms are a good way of showing the overall pattern of a distribution. Quartiles and percentiles can also be used to similar effect. Averages, however, just focus on one aspect, the middle, either in the sense of the mean or the median. Whether the values spread out far from this average, above it, below it or in both directions, is impossible to say from the average alone.

Table 3.3 Weights in grams of apples in two bags

Bag 1	130	124	124	120	121	124	128	129
Bag 2	173	192	36	166	119	31	178	129

We often need a measure of 'spread' as well as an average. Let's consider another example. The weights (in grams) of the apples in two bags are in Table 3.3.

The apples in Bag 1 are produced on a factory farm with industrial-style controls: the result is that the apples are all very similar in weight. The 'range' (biggest minus smallest) is only 10 g. By contrast, the apples in Bag 2 are organically grown and very variable: their weights are much more spread out. Some are much bigger than others. The range in this case is 161 g (192 – 31). Note that the mean weights of apples in both bags are very similar: 125 g in Bag 1 and 128 g in Bag 2. The total weight of each bag is obviously eight times the mean: 1000 g and 1024 g. Is this obvious**?**

I hope so. Remember that to work out the mean you find the total weight and then divide by the number of apples. Multiplying the mean by the number of apples just undoes the last step and gets back to the total weight.

I'll explain four ways of measuring spread. We have already met the first. The 'range' is simply the difference between the biggest and smallest value, 10 g for Bag 1 and 161 g for Bag 2, as explained above. This is very simple, but it has a couple of snags. The fact that it depends on just two values in a sample means it tends to be rather erratic. All the information from values in the middle is ignored. The second snag is that the range of big samples is likely to be bigger than the range of smaller samples. Imagine, say, a thousand apples from the Bag 2 source. In this bigger bucket, there is very likely to be a bigger apple than the 192 g apple in Bag 2, and one smaller than 31 g. This means that the range is almost certain to be more. For this reason, it is difficult to generalise the range to a wider context, although you can still compare two samples like Bag 1 and Bag 2. On the other hand, it is very simple and very easy to interpret.

The 'interquartile range' is just what it says it is, the range or difference between the two quartiles. Using the method for quartiles in the last section, what are the two quartiles of the weights in the two bags? And using these, what are the two interquartile ranges**?**

The quartiles for Bag 1 are 122.5 and 128.5, which gives an interquartile range of 128.5–122.5 or 6 g. The corresponding figure for the second bag is 98 g. The bigger figure for Bag 2 indicates that the quartiles are much further apart, which reflects the fact that the weights are more varied.

The 'mean deviation from the mean' is also just what it says it is. Taking Bag 1 as our example, the mean is 125, so the deviation from the mean of the first (130 g) apple is 5 g. Similarly, the deviation of the second is 1 g. The mean deviation is simply the mean of all eight deviations (5, 1, 1, 5, 4, 1, 3, 4) which is 3 g. What is the corresponding mean deviation from the mean for Bag 2**?**

You should have found this is 49.5 g, far bigger than Bag 1's figure, because the weights are more spread out so the deviations from the mean are, on average, larger.

The 'standard deviation' (sd) is defined as the square root of the mean of the squared[36] deviations from the mean. All this means is that we square the deviations above (25, 1, 1, 25, 16, 1, 9, 16), take the mean as before (11.75), and then take the square root of the answer (3.43). This is the standard deviation for the first bag. What is the standard deviation of the weight of apples in the second bag**?**

The standard deviation is 59.18 g, much larger than Bag 1, as we would expect. The sd is similar to the mean deviation from the mean. The only difference is the squaring, and then the square rooting to undo the squaring. Not surprisingly, the two answers tend to be similar: the sd is normally a bit bigger but how much bigger depends on the pattern of the numbers.

You may wonder why we bother with the sd when the mean deviation from the mean seems similar, but more straightforward. The reason is that the sd is necessary for much of the mathematical development of statistics, for example we need it if we are to use the normal distribution (Section 5.6). As this is a book on *non-mathematical* statistics, my initial inclination was to leave it out and concentrate on more user-friendly measures. The reason for including the sd is that it is very widely used: it has cornered the market for measures of spread. You will find references to standard deviations all over the place, so it is essential to understand what it is.

The standard deviation is influenced strongly by outliers – values much larger or smaller than the rest of the distribution. For example, the standard deviation of the results in Figure 3.2 is 24 units of alcohol. Including the three outliers (126, 163, 327) increases the standard deviation to 43, a very substantial difference. This makes the standard deviation a rather erratic measure if there are outlying values. These outliers may be mistakes in the data, which can have a disproportionate effect on the answer. Do you think the mean deviation from the mean and the interquartile range will be as strongly influenced by outliers**?**

No. The mean deviation is increased from 19 to 26 by including the outliers in Figure 3.2. The interquartile range is unchanged at 33. The exact size of outliers makes no difference here: if they were even larger, the third quartile and the interquartile range would be unchanged.

In my view, the popularity of the standard deviation is undeserved. It is complicated and prone to giving odd answers in some situations. I will not use it much in the rest of this book. The mean deviation from the mean and the interquartile range are generally preferable ways of measuring spread. Or look at the percentiles, or draw a histogram.

In SPSS you will find all of this under Analyze – Descriptive statistics. There are also useful Excel functions.[37]

3.4.5 Proportion of values over a critical level

The final way of summing up the pattern of a number variable is to reduce it to a category variable. Taking three units a day as the limit for reasonable drinking, we can define anything over this as over the limit. Table 3.1 shows that the proportion of students over the limit on Saturday night was 12 out of 20, or 60%. What is the corresponding figure for the whole sample of 92 (use Figure 3.1 or Excel[38])**?**

The proportion in the whole sample is 46 out of 92, or exactly 50%. You can, of course, think of this as the probability of a student drinking too much.

▶ 3.5 Summarising the relation between two category variables

We are often interested not so much in a single variable but in how two variables relate to each other. Do students drink more on Saturday night than they do on Sunday nights? Do students who smoke a lot also drink a lot? Do male students drink more than female students? Are male students more likely to smoke than female students? All these questions involve two of the variables in Table 3.1, units on Saturday and units on Sunday, average cigarettes and units drunk, sex and units on Saturday and sex and smoker/non-smoker. How can we analyse these?

If we have two category variables, the obvious thing to do here is to make a table to show how frequently each combination of values (male smoker, male non-smoker, female smoker and female non-smoker) occurs. Table 3.4 presents this information as percentages because it is easier to compare males and females. It is usually a good idea to use percentages in these tables, although you will need to ensure that you choose the right percentages. The percentages in Table 3.4 are row percentages; column percentages (the percentage of non-smokers who are female and so on) are of less obvious interest here. Can you see how Table 3.4 is produced from the data in Table 3.1**?**

There are 12 females listed in Table 3.1, of whom 10 are listed as 0 under Daycigs. Ten out of 12 is 83% and so on. Tables like this are easy to produce

Table 3.4 Percentages of female and male smokers in Table 3.1

Sex	*Non-smoker*	*Smoker*	*Total*
Female	83%	17%	100% (n = 12)
Male	63%	38%	100% (n = 8)
Total	75%	25%	100% (n = 20)

n is the number of students on which the percentages are based.

with SPSS[39] or with Excel.[40] You can also illustrate this table by means of bar charts of various kinds; I'll leave you to explore this if you want to.

▶ 3.6 Summarising the relation between one category and one number variable

Tables of frequencies like Table 3.4 are not really practical if either of the variables is a number variable. (If you can't see why, try setting up a table like Table 3.4 to show the relationship between Satunits and Sex.) The number variable needs to be summarised using the mean or something similar (for example the median). Table 3.5 shows how the male data on units drunk on Saturday night can be compared with the female data.

Check that you can see how this is worked out from Table 3.1. It is good practice to include the number of people that each mean is based on. Statisticians tend to use the symbol *n* for this number. The table is easily produced by SPSS (Analyze – Compare means – Means) or Excel.[41] You could also include other quantities, for example median, standard deviation, in the table.

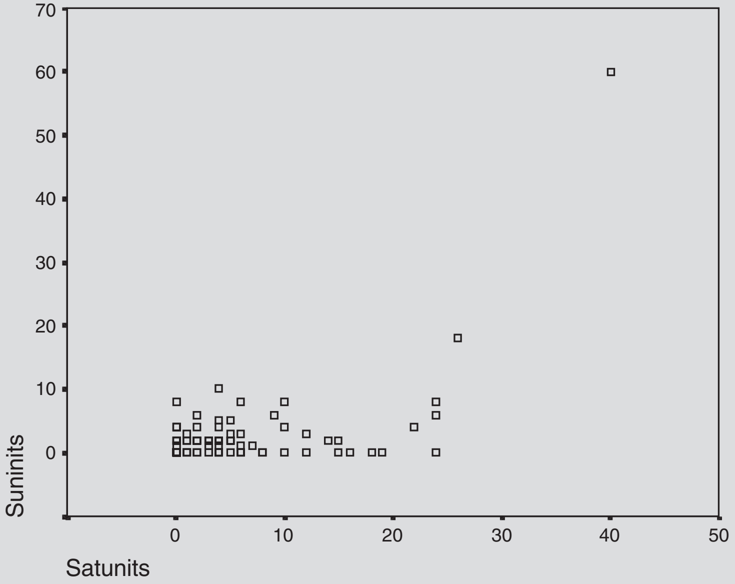
▶ 3.7 Summarising the relation between two number variables

When we have two number variables, tables like Table 3.5 are not feasible. Instead, there are four possibilities: producing a 'scatter diagram', working out the difference between values of the variables, a 'correlation coefficient' between the variables or the 'slope' in a regression model. The next three subsections deal with the first three of these. We'll leave the fourth until Chapter 9 when we look at regression.

Table 3.5 Mean units drank on Saturday night for males and females (based on Table 3.1)

Sex	*Mean units on Saturday*	*n (number of students)*
Female	3	12
Male	13	8
Overall	7	20

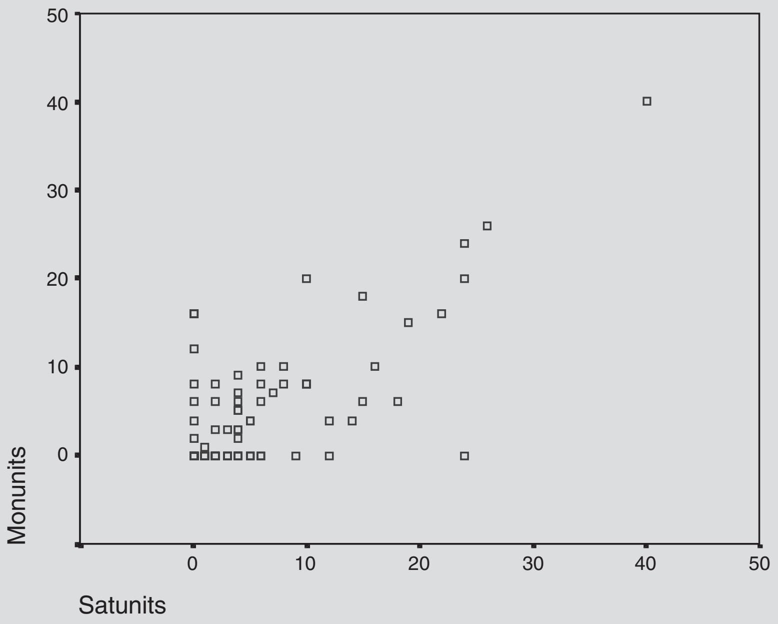
Figure 3.3 Scatter diagram of Satunits and Sununits



3.7.1 Scatter diagrams

Scatter diagrams are very useful for showing the detailed relationship between two number variables. Figures 3.3, 3.4 and 3.5 show three of these, all based on the full set of data from which Table 3.1 is extracted. These diagrams are all produced by SPSS (Graphs – Scatter). They could easily be produced by Excel (use the Chart Wizard), but SPSS offers some useful extras, for example if you want to see *all* the scatter diagrams showing the relation between several variables, use Graph – Scatter – Matrix.

Figure 3.4 Scatter diagram of Satunits and Monunits



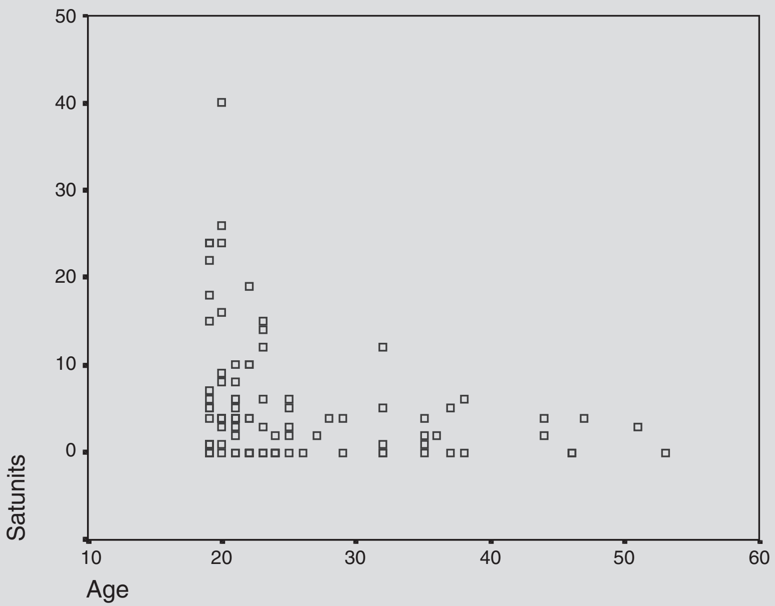
In these diagrams, each point represents a student, for example in Figure 3.3 the top right point represents a student who says she drank 40 units on Saturday and 60 on Sunday. Each diagram has a different pattern which tells its own story. Apart from the two points on the right, there seems little relationship between the amounts drunk on Saturday and Sunday. The pattern of Sunday drinking for those who drunk nothing on Saturday is fairly similar to the pattern for those who drunk 24 units on Saturday. On the other hand, Figure 3.4 shows that there is a slight tendency for students who drank more on Saturday to drink more than average on Monday as well. Why do you think there was this difference between Sunday and Monday**?**

I don't know. One possibility is that some of the heavy drinkers on Saturday had too much of a hangover to drink much on Sunday, but had recovered to return to heavy drinking by Monday.

Figure 3.5 shows the relationship between age and units drunk on Saturday. What do you think this diagram shows**?**

I think it shows that students over 30 drink less than the younger students. There is only one student over 30 who drank more than 10 units on Satur-

Figure 3.5 Scatter diagram of age and Satunits



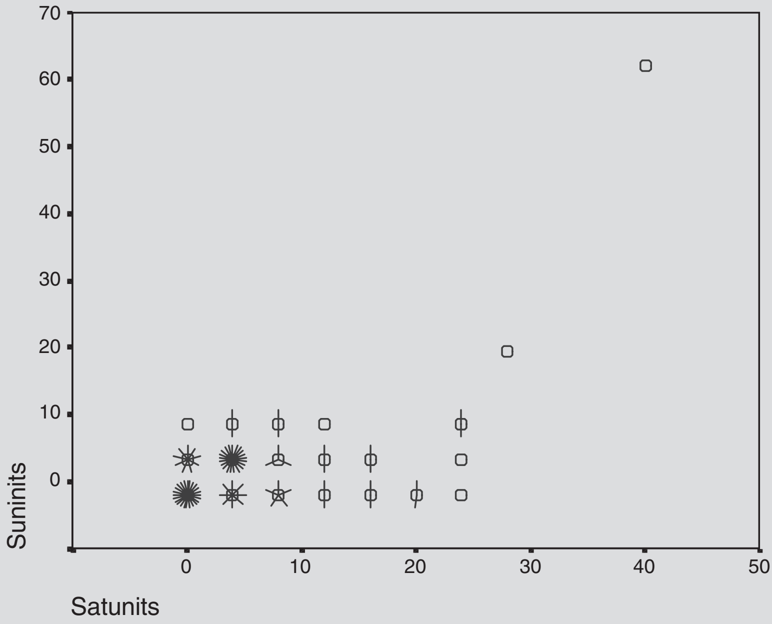
day, and the majority drank much less. On the other hand, among the younger students, there were some very heavy drinkers.

In Figure 3.5, I have put age along the horizontal axis. This is because there is a convention that says you should put the 'independent' variable along the horizontal axis. In this example, age is the independent variable and units drunk on Saturday is the 'dependent' variable, because it is natural to assume that units drunk on Saturday depend, to some extent, upon age, but not vice versa. However, this is often a hazy distinction, and this rule is only a convention. Which is the independent variable in Figure 3.4**?**

You might say that they both have the same status, so neither is the independent variable, so it doesn't matter which one goes on the horizontal axis. Or, you might think that Monday consumption depends on Saturday consumption because Monday comes after Saturday, so Saturday should go on the horizontal axis. I think either attitude is OK.

Figures 3.3, 3.4 and 3.5 are all based on a sample of 92 students. Each student is represented by one point, so there should be 92 points in each diagram. In fact, if you count up, you will find that there are considerably fewer than 92 points in all three diagrams. Why do you think this is**?**

Figure 3.6 SPSS scatter diagram with sunflowers



There were several students who drank nothing on Saturday and Sunday. This means that the point in the bottom left on Figure 3.3 represents not one student but a whole group of students. Similarly, some of the other points will represent several students. Figure 3.6 shows one way of getting round this problem. This is produced by SPSS.[42] What do you think the 'sunflowers' represent**?**

Each petal of the sunflowers represent a single student, for example a three-petalled sunflower represents three students.

3.7.2 Differences between the variables

Sometimes it is useful to summarise the relationship between two variables by means of a single number, rather like an average gives a crude summary of the pattern of a single variable. There are different ways in which this can be done. The first is to focus on the *difference* between the two variables. For example, we could subtract the units drunk on Sunday from the units drunk on Saturday to give a measure of how much more is drunk on Saturday (0, +8, −2, +3, +1 . . . in Table 3.1). We can then work out the average difference. What is this average for the data in Table 3.1**?**

The difference is 3.65 units: the average student drank this much more on Saturday night than on Sunday. The corresponding difference for the whole sample of 92 students is 2.9 units. This would not make sense in relation to the third scatter diagram in Figure 3.5 because the two variables are on completely different scales – one is age and the other alcohol consumption. The difference would not be meaningful.

3.7.3 The correlation between the variables

A second way of summing up the relationship between two numerical variable as a number is to use a 'correlation coefficient'. This works even if the two variables are on completely different scales like Figure 3.5. Correlation coefficients measure the extent to which high values of one variable go with high values of the other. Are the students who drink a lot of Saturday night the same students as those that drink a lot on Monday night. Or do those who drink a lot on Saturday have less on Monday (perhaps they are recovering from a hangover)? The informal answers can be seen in Figure 3.4. There does seem to be a positive relationship between consumption on Saturday and Monday. In general students who drink more on Saturday do seem to drink more (than average) on Monday too. On the other hand, the picture for Saturday and Sunday is not so clear. There does not seem to be much relationship one way or the other. The simplest way to measure correlation is as follows.

Imagine meeting two of the students in Table 3.1 and comparing their drink figures for Saturday and Sunday. If the two students were the first two in the list, the bigger drinker on Saturday (number two in the list with 26 units) would also be the bigger drinker on Sunday (18 units vs 0 units). I'll call this a 'same-direction' observation. This tends to confirm the idea that Saturday and Sunday drinking are correlated. On the other hand, if the two students were the next two in the list, the conclusion would be the opposite. Number four drunk more on Saturday (5 vs 1), yet number three drank more on Sunday (3 vs 2). This is a 'reversed-direction' observation. The question now is whether, if we met lots of pairs of students, we would have predominantly same-direction or reversed-direction observations. Answering this question with all 20 students in Table 3.1 is rather hard work, so I'll start by imagining that we just have the first four. The analysis is in Table 3.6.

Table 3.6 shows that there are six possible pairs of students, of which five give same-direction observations. Five-sixths of the observations are same-direction, and one-sixth are reversed-direction. This provides a crude measure of correlation. If we met any a pair of students from this group, at random, the probability of a same-direction observation would be 5/6, and

Table 3.6 Calculation of a correlation coefficient between Satunits and Sununits

	Second student			
First student	*1*	*2*	*3*	*4*
1				
2	+1			
3	+1	+1		
4	+1	+1	–1	

+1 indicates a same-direction observation, and –1 a reversed-direction one. The numbers 1, 2, 3, 4 represent the students in the first four rows of Table 3.1.

of a reversed-direction observation would be 1/6. This indicates a fairly high correlation. Can you use the first four rows of data in Table 3.1 to do a similar calculation for age and units drunk on Saturday**?**

You should have found two same-direction observations, and three reversed-direction observations. The sixth is inconclusive – neither one thing nor the other – because the ages of the third and fourth students are the same. This suggests a probability of a same-direction (SD) observations of 2/6 or 1/3, and 3/6 or 1/2 for reversed-direction (RD) observations. Notice that you are comparing the ages of the two students, and the amounts they drank on Saturday. You do not need to compare ages with amount drunk, which would, of course, not be a sensible thing to do.

These probabilities can be converted to a measure known as 'Kendall's correlation coefficient' (or sometimes 'Kendall's tau'):

Kendall's correlation coefficient = proportion of SDs – proportion RDs.

In the first example (Table 3.6) this would come to 5/6 – 1/6 or 0.67. What is Kendall's correlation coefficient in the second example (age and Satunits)**?**

One-third minus a half which is –0.17. This is negative, indicating that there are more reversed-direction observations.

Kendall's correlation coefficient is designed to range from +1 to –1. It will be +1 if all the pairs are same-direction and none are reversed, so the formula becomes simply 1 – 0. This will be the case if the scatter diagram shows an 'uphill' straight line. This is shown in the first scatter diagram in Figure 3.7. Take any two crosses you like in this diagram, and you will get a same-direction observation.

Figure 3.7 Scatter diagrams corresponding to correlations of +1, 0 and –1

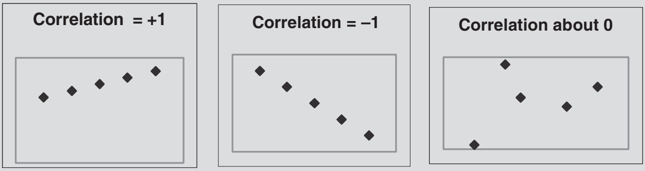


Table 3.7 Relation between Kendall's correlation coefficient and the probability of a same-direction observation

Kendall's correlation coefficient	*Probability of same direction observation*
+1	1
+0.5	0.75
0	0.5
–0.5	0.25
–1	0

This assumes there are no inconclusive pairs.

At the other extreme Kendall's correlation would be –1 if all the pairs were reversed (the formula would be 0 – 1). And if there are approximately equal numbers of same and reversed-direction pairs, the correlation would be about zero. Table 3.7 is intended as a reminder for interpreting correlations, for example a correlation of +0.5 means that the probability of a same-direction observation is 75%.

So far we have just used samples of four. The corresponding correlations for all 92 students (in `drink.xls`) are given in Table 3.8. Not surprisingly, the answers are rather different from the calculations based on the first four students in Table 3.1.

This table shows, for example, that the correlation between age and units drunk on Saturday is –0.22. This indicates a weak negative relationship. What do you think the probability of two students encountered at random being a same-direction pair, that is, the older one drinking more**?**

Using Table 3.7 the answer is obviously between 0.25 and 0.5, say about 0.4. Table 3.7 assumes there are no inconclusive pairs for which Age or Satunits, or both, for the two students are equal. As there are inconclusive pairs

Table 3.8 Kendall's correlation coefficient between variables in Table 3.1 (data from all 92 students)

Age				
–0.22	*Satunits*			
–0.07	0.20	*Sununits*		
–0.26	0.33	0.16	*Monunits*	
–0.12	0.10	0.11	0.09	*Daycigs*

like this, Table 3.7 should be viewed as a rough guide only. The scatter diagram showing the relationship between these two variables is Figure 3.5. This shows a weakish negative relation as the correlation leads us to expect. Similarly the correlation between Satunits and Monunits is positive (0.33) as Figure 3.3 suggests.

Working out Kendall's correlation coefficient by the method of Table 3.6 is hard work for anything but very small samples. You can use SPSS[43] or Excel[44] to do the calculations for you. You will, however, find that the answers may not quite tally. SPSS uses a slightly different version of Kendall's tau. You will also find a function, correl, in Excel, which produces another correlation coefficient (Pearson's), and yet other correlation coefficient, Spearman's, produced by SPSS. I'll sort this out in Section 3.10.

3.8 Summarising the relation between three or more variables

This is obviously more complicated. Table 3.9 relates three of the variables in the drink data (Course, Sex, Satunits). This is an edited version of an SPSS table,[45] but a similar table can be produced using a Pivot Table in Excel (see Appendix A.4).

Table 3.9 shows that the pattern of males drinking more than females holds for the two full-time courses, but not for the part-time course (PP). However, there were only four males on this course, so we should obviously not be too confident that this conclusion can be safely generalised beyond this sample. It is always good practice to include the sample size on which means and other statistics are based so that readers do not jump to firm conclusions based on very small samples. You could also display the means in Table 3.9 as a clustered bar chart.[46] It is possible to include further vari-

Table 3.9 Table showing relation between three variables in drink data

	Satunits for female students				*Satunits for male students*			
Course	*Mean*	*Bottom q'tile*	*Top q'tile*	*n*	*Mean*	*Bottom q'tile*	*Top q'tile*	*n*
FB	5.9	0.6	8.1	19	11.7	3.5	22.5	20
FP	2.7	0.0	4.3	20	4.8	1.0	12.0	10
PP	2.5	0.0	4.0	19	0.3	0.0	–	4

Edited version of SPSS output from Custom tables – Basic tables procedure. *n* is the number in each group. There were only four students in the male PP group who drank only one unit between them, so SPSS has refused to estimate the third quartile.

ables in tables like 3.9, but this may be at the cost of making the presentation too complex to take in. SPSS also offers three-dimensional scatter plots for three numerical variables (Graphs – Scatter), which you may wish to try.

▶ 3.9 Beyond the data: cause, effect and the wider context

My reason for going on and on about the data from the 92 students is not that I expect you to be particularly interested in this group of students. I'm sure you aren't. My intention, and hope, is that this data is helpful for learning about a wider context, in which you may be more interested. There are several different aspects of this wider context. First, I hope you will apply the methods in the sections above to situations of interest to you. Second, even if you are interested in the drinking habits of students, you would certainly want come to conclusions about a wider population of students. The extent to which this is possible and meaningful depends on a number of factors, particularly the source of the sample (see Section 3.2) and its size (see Chapter 7).

There is also a third, slightly more subtle, issue. This is the extent to which we can use the data to come to conclusions about cause and effect. As an example, consider the fact that the average number of units drunk on Saturday night by the full-time students in this sample of 92 was 6.5, whereas the corresponding figure for the part-time students was 2.1. The full-time students drank a lot more than the part-timers, who had full-time jobs to do when they were not studying. What explanation would you suggest for this difference**?**

There are many possible answers, but they fall into four categories. The first type of answer just says it's chance. If you took another sample of students, the answer may be different. In this case, the size of the difference, and the relatively large sample, suggests this is unlikely. But it's certainly possible: we'll look at how the chance explanation can be checked in Chapter 8.

The second category of answer focuses on suggesting a mechanism by which being a full-time student can *cause* students to drink more than they would if they were part-time. It might be excessive amounts of free time, the lack of pressure to turn up to work early in the morning, or a culture of heavy drinking. Whatever the detailed explanation, the common feature of these explanations is that they imply that if a student changes from a full-time to a part-time course, the change is likely to be accompanied by a reduction in the amount drunk. If the drinking of a full-time student is a problem, the cure may be transfer to a part-time course. Was your suggested explanation of this type**?**

If it wasn't, there are two further possibilities. The third is that there might be some essentially accidental feature of the samples which leads to the difference in Saturday drinking habits. The part-time sample contained a greater proportion of female students and a greater average age. Female students tended to drink less (Table 3.5), as did older students (Figure 3.5). This could be the explanation for the difference in drinking patterns between full-time and part-time students. If this explanation is right, would forcing a student from a full-time to a part-time course be likely to cure a heavy drinking problem**?**

No, of course not. This would not change either age or sex. The final type of explanation would involve a drinking habit somehow making people become full-time students. Does this sound plausible**?**

It's definitely possible. Drinkers might be attracted to the (full-time) student lifestyle, or employers might be put off by excessive drinking. To sum up, if D represents being a heavy drinker, F represents being a full-time student, the data shows a tendency for D and F to be associated: D people tend to be F people and vice versa. There are four possible types of explanation:

1. It's chance: another sample may be different.
2. F causes D.
3. Something else causes D and F.
4. D causes F.

So, when you notice a relationship between two variables, be careful about jumping to conclusions. You should always check out these four types of explanation.

3.10 Similar concepts

There are distinctions between different types of number variables: those on ‘continuous’ and ‘discrete’ scales; and ‘ordinal’, ‘interval’ and ‘ratio’ scales. You will find more information in more detailed texts.[47] There are many other ways of summing up data: numerous types of graph and coefficients to measure various characteristics of the data. I’ve been selective in this chapter and left out anything which seems too easy to be worth going on about (pie charts, simple bar charts, the mode as a type of average), or too complicated to be useful. Many things simply don’t justify the effort of getting to know them. You will also find so-called short methods for calculating standard deviations and correlation coefficients. These are a complete waste of time: the best short method is a computer (or calculator). Forget them! However, there are a few other concepts which must be mentioned here because they are widely used: you will probably come across them so you need to know a bit more about them.

The standard measure of correlation – what is meant when people simply refer to the correlation – is ‘Pearson’s product moment correlation coefficient’. This is the quantity calculated by the Excel function correl. It is similar to Kendall’s tau (Section 3.7.3), ranging from +1 for a perfect positive relation to –1 for a perfect negative relation. However, it is calculated in a different way, by means of a relatively complex formula which meshes well with mathematical probability theory, the original reason for its popularity. Obviously, in any given situation, Pearson’s and Kendall’s coefficients will give slightly different answers, but they are generally close enough for you to use what you have learned about Kendall’s when interpreting Pearson’s correlation. Figure 3.7 applies to Pearson’s coefficient as well as Kendall’s. Between the three values in this figure (–1, 0, +1), there is a tendency for Pearson’s coefficient to be larger (in magnitude, ignoring any negative sign) than Kendall’s, for example the Pearson correlation corresponding to the Kendall correlation of –0.22 in Table 3.8 is –0.29.

As well as being arithmetically awkward, the standard deviation is also a nuisance because it comes in two versions. The standard deviation defined in Section 3.4.4 corresponds to the Excel function stdevp, or σ_n on many calculators. The slight problem with this function is that if you use it with the data from a sample to estimate the standard deviation of a wider population, there will be a consistent tendency for the answer to be too small. The other version, stdev (or σ_{n-1}), has a correction built in so that, on average, the answer will be much closer. In practice, as you are usually using a sample to come to conclusions about a wider population, the version with the correction is the one you usually want. For this reason, it is the one produced by SPSS. If you are in doubt about which function is which, remember that stdev – the one you probably want – is the *larger* of the two. However, with large samples, the difference is small and certainly not worth worrying about.

In a similar vein, Kendall's tau (Section 3.7) comes in different versions. SPSS produces tau-b which incorporates a correction for inconclusive observations for which there is a tie on one or both variables. This complicates the interpretation of Kendall's tau, for example Table 3.7 won't be exactly right. Kendall's tau-b attempts to get round this, but at the cost of losing a lot of the transparency, which is the main advantage of Kendall's tau in the first place.

Another measure of the spread of the values of a number variable is the 'variance'. This is simply the square of the standard deviation. It is important for statistical modelling; we will see how in Section 9.3.1.

If you want to see some further possibilities (box and whisker plots, stem and leaf plots, measures of skewness, the variance, trimmed means and so on), explore the output produced by SPSS, especially the menu options under Analyze – Descriptive statistics. But I have tried to cover the essentials here.

▶ 3.11 Exercises

The obvious thing to do is to try out the concepts and methods in the sections above with some data of your own. Get some data on something you care about and explore it.

3.11.1 Histograms

What is the best width of interval for a histogram? If you have Excel, you could try experimenting with different interval (bar) widths for Figure 3.2. What about one unit? And what about 20? I think you will find 10 is about the best, but this is to some extent a matter of opinion. The file `words.xls` contains data on the lengths of words (number of letters) in passages from two different sources: Lewis Caroll's *Alice's Adventures in Wonderland* and an academic sociology article. Draw a histogram to show the pattern of each distribution. Have you any comments on the differences between the two distributions?

Now draw another version of the first histogram with just three intervals (bars): the first comprising one, two and three-letter words; the second only four-letter words, and the third words with five or more letters. You should find that the middle bar is the smallest, apparently suggesting that Lewis Caroll tends to avoid four-letter words. Is this right? (The lesson of this is that when you draw a histogram you should always make the intervals the same width.)

3.11.2 Quartiles, percentiles and so on

In a firm with 2000 employees, the upper quartile of the salary distribution is £25 000, the 95th percentile is £40 000, the mean salary is £19 000 and the standard deviation of the salaries is £8000. Answer as many of the following questions as you can (you may not have the information to do them all):

- How many employees earn more than £40 000?
- How many employees earn £25 000 or less?
- How many employees earn between £25 000 and £40 000?
- How many employees earn less than £19 000?
- How many employees earn less than £11 000?

3.11.3 Experiments with correlations and standard deviations

The marks for a group of four students in two examinations were:

Mathematics:	Ann: 60, Bill: 75, Sue: 85, Dan: 90
English:	Ann: 50, Bill: 48, Sue: 49, Dan: 46

The standard deviations (stdevp) of the marks are 11.5 for maths and 1.5 for English. Pearson's correlation coefficient between the marks in the two subjects is –0.8. Work out Kendall's correlation coefficient. Is it similar to Pearson's coefficient? Assuming that these marks are reasonably typical of all students doing the two examinations, what comments would you make about the difference between the two standard deviations, and about the correlation? Would it be fair to add the marks in the two examination to give students an overall mark?

You should be able to answer the following questions without a computer or calculator:

- What would the standard deviation of the maths marks be if the marks were halved (that is, they became 30, 37.5, 42.5, 45)?
- What if 30 was deducted from each mark?
- What effect would both of these changes have on the correlation? Would it still be –0.8 in both cases?
- Find the standard deviations of these marks: 60, 60, 75, 75, 85, 85, 90, 90 (compare with the maths marks above)
- And the standard deviation of these: 70, 70, 70, 70.

If you have a computer available you should be able to check these.

The final part of this exercise exploits one of the strengths of spreadsheets; the fact that they allow you to experiment: to change the data and see what effect this has on the answer. See if you can find six numbers whose mean is 30 and whose standard deviation is 10. Make the best guess you can, then work out the standard deviation (using the function stdevp), then try and adjust the guess and so on until the mean is 10 and the sd is 30 (both to the nearest whole number). How many steps did you need?

3.11.4 Exploring relationships between variables on the Isle of Fastmoney

The data in the file *iofm.xls* is from a random sample of 100 people on an (imaginary) island. Sexn is a numerical coding for sex, run 10 km is each person's time in minutes in a 10 km race, cycle 10 km is their time in a cycle race, and earn000E is their earnings in thousands of euros. What relationships can you find between the variables? Are they all what you would expect? (We'll return to this example in Chapter 9.)

3.11.5 Risk and return on the stock market

The file *shares.xls* contains daily share prices for three shares traded on a stock market. It also contains a calculation of the daily return: the gain you would make if you bought the share the day before and sold it on the day in question. For example, if you buy a share for £5 and sell it the next day for £4.50, you have lost £0.50 or 10% of your original investment. So the return would be –10%. Produce a histogram of the daily returns of each of the three shares over the time period. (So you need three histograms, one for each share.) Now work out the means and standard deviations of the daily returns for each of the three shares, and the (three) correlation coefficients between the returns for each of the three shares. How would you explain the interpretation of these to someone not familiar with statistics?

▶ 3.12 Summary of main points

There are many ways of summarising things statistically. In this chapter I have explained what I think are the most useful of these. Regardless of the type of summary you use, you should always consider carefully the source of the data you are using, and be cautious about generalising your results or jumping to conclusions about cause and effect.