4 The following table shows data about the time taken (in seconds, to the nearest second) for each one of a series of 75 similar chemical experiments.

Time (s)	50–60	61–65	66–70	71–75	76–86
Number of experiments	4	13	26	22	10

(a) State the type of diagram appropriate for illustrating the data.

(b) A calculation using the data in the table gave an estimate of 69.64 seconds for the mean time of the experiments. Explain why this value is an estimate.

(c) Estimate the median of the times taken for completing the experiment.

(d) It was discovered later that the four experiments in the class 50–60 had actually taken 57, 59, 59 and 60 seconds. State, without more calculation, what effect (if any) there would be on the estimates of the median and mean if this information were taken into account. (OCR, adapted)

5 The standardised marks received by 318 students who took a mechanics examination are summarised in the following grouped frequency table.

Mark	0–29	30–39	40–49	50–59	60–69	70–79	80–89	90–100
Frequency	12	7	13	25	46	78	105	32

(a) Draw a histogram of these data, and describe the skewness of the distribution.

(b) Estimate the mean mechanics mark.

(c) Estimate the median mechanics mark by drawing a cumulative frequency graph.

The same 318 students also took a statistics examination during the same session. The mean and median of those marks were 71.5 and 70.0 respectively. Write a brief comparison of the students' performances in the two examinations.

6 An ordinary dice was thrown 50 times and the resulting scores were summarised in a frequency table. The mean score was calculated to be 3.42. It was later found that the frequencies 12 and 9, of two consecutive scores, had been swapped. What is the correct value of the mean?

7 Three hundred pupils were asked to keep a record of the total time they spent watching television during the final week of their summer holiday. The times, to the nearest $\frac{1}{4}$ hour, are summarised in the following table.

Number of hours	$0\pm4\frac{3}{4}$	$5\pm9\frac{3}{4}$	$10\pm14\frac{3}{4}$	$15\pm19\frac{3}{4}$	$20\pm24\frac{3}{4}$	$25\pm29\frac{3}{4}$	$30\pm34\frac{3}{4}$	$35\pm39\frac{3}{4}$
Frequency	4	21	43	62	90	56	18	6

(a) Estimate the mean viewing time.

(b) State two sources of inaccuracy in your estimate of the mean.

(c) Find an estimate of the median viewing time.

(d) What do the values of the mean and median indicate about the skewness of the data?

8 It is sometimes said that for any set of quantitative data, the median (me), mode (mo) and mean ($\bar{x}$) are such that either $\bar{x} \leqslant \text{me} \leqslant \text{mo}$ or $\text{mo} \leqslant \text{me} \leqslant \bar{x}$. Check that this is true of the distribution in Question 2. Show that the statement is untrue for the following data.

x	1	2	3	4	5
f	2	1	11	9	7

9 The table gives the prices (in dollars) of shares in 10 firms on Monday and Tuesday of a particular week. The Monday price is m, the Tuesday price is t, and $d = t - m$.

Firm	A	B	C	D	E	F	G	H	I	J
m	151	162	200	233	287	302	303	571	936	1394
t	144	179	182	252	273	322	260	544	990	1483
d	−7	17	−18	19	−14	20	−43	−27	54	89

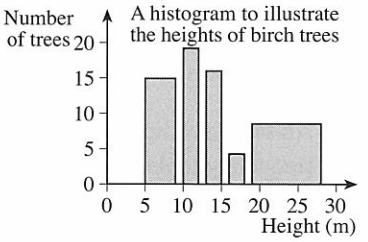
(a) Calculate $\bar{m}$, $\bar{t}$ and $\bar{d}$. Does $\bar{d} = \bar{t} - \bar{m}$?

(b) Calculate the medians of m, t and d. Is it true that $\text{me}_d = \text{me}_t - \text{me}_m$?

10 The height, correct to the nearest metre, was recorded for each of the 59 birch trees in an area of woodland. The heights are summarised in the following table.

Height (m)	5–9	10–12	13–15	16–18	19–28
Number of trees	14	18	15	4	8

(a) A student was asked to draw a histogram to illustrate the data and produced the following diagram. Give two criticisms of this attempt at a histogram.



(b) Using graph paper, draw a correct histogram to illustrate the above data.

(c) Calculate an estimate of the mean height of the birch trees, giving your answer correct to 3 significant figures.

(OCR)

3 Measures of spread

This chapter describes three different measures of spread and their methods of calculation. When you have completed it, you should

- know what the range is, and be able to calculate it
- know what the quartiles are and how to find the interquartile range from them
- be able to construct a box-and-whisker plot from a set of data
- know what the variance and standard deviation are, and be able to calculate them
- be able to select an appropriate measure of spread to use in a given situation.

3.1 Introduction

You saw in Chapter 2 how a set of data could be summarised by choosing an appropriate typical value, or 'measure of location' as it is more correctly known. Three different measures of location, the mean, the median and the mode, were introduced.

Now consider the two sets of data A and B given below.

A:	48	52	60	60	60	68	72
B:	0	10	60	60	60	110	120

For both data sets A and B, mean = median = mode = 60. If you were given nothing but a measure of location for each set you might be tempted to think that the two sets of data were similar. Yet if you look in detail at the two sets of data you can see that they are quite different. The most striking difference between the two data sets is that set B is much more spread out than set A. Measures of location do not give any indication of these differences in spread, so it is necessary to devise some new measures to summarise the spread of data.

3.2 The range

The most obvious method of measuring spread is to calculate the difference between the lowest value and the highest value. This difference is called the **range**.

> The **range** of a set of data values is defined by the equation
>
> range = largest value ± smallest value.

The range of data set A is $72-48=24$, whereas the range of data set B is $120-0=120$. Calculating the ranges shows clearly that data set B is more spread out than data set A.

It is quite common for students to give the range as an interval. This would mean, for instance, that the range of data set A would be given as 48 to 72, *or* 48–72, *or* $48 \rightarrow 72$. *In statistics it is usually much more helpful to give the range as a single value, so the definition above is used.*

If you are going to use the range as a measure of spread it is helpful to realise its limitations. If you consider the two further data sets C and D shown below, you will see that they both have the same range, 8.

C:	2	4	6	8	10
D:	2	6	6	6	10

Although both data sets C and D have the same range, the patterns of their distributions are quite different from one another. Data set C is evenly spread within the interval 2 to 10 whereas data set D has more of its values 'bunched' centrally. Because the range is calculated from extreme values it ignores the pattern of spread for the rest of the values. This is a major criticism of the range as a measure of spread. Although the range is easy to calculate, it ignores the *pattern* of spread and considers only the extreme values.

3.3 The interquartile range

Since the range ignores the internal spread of the values in a data set, an alternative measure is needed. One possibility is to look at the spread between two values which are at some fixed, but interior, position. A sensible choice, which is associated naturally with the median, is to choose the values that are at the positions one-quarter and three-quarters of the way through the data when the values are arranged in order. These points are known as the **lower quartile** and the **upper quartile** respectively, and they are usually denoted by the symbols Q_1 and Q_3 respectively. The difference between these values is called the **interquartile range**.

Interquartile range = upper quartile – lower quartile = $Q_3 - Q_1$.

The interquartile range is really just the range of the middle 50% of the distribution.

Notice that there is also a **middle quartile**, Q_2, which is the median.

To find the position of the quartiles for small data sets there are several possible methods that you might see in textbooks. The one suggested below is fairly easy to apply.

Finding the quartiles

- First arrange the data in ascending order.

Case 1 An even number of data values

- Split the data into their upper half and lower half.
- Then the median of the upper half is Q_3, and the median of the lower half is Q_1.

Case 2 An odd number of data values

- Find the median, Q_2, and delete it from the list.
- Split the remaining data into their upper half and lower half.
- Then the median of the upper half is Q_3, and the median of the lower half is Q_1.

Example 3.3.1
Find the quartiles and the interquartile range for each of the two sets of data below.

(a) 7 9 12 13 8 11

(b) 7 8 22 20 15 18 19 13 11

(a) First, arrange the data in numerical order.

7 8 9 11 12 13

The number of data values is even, so divide the data into its lower and upper halves:

Lower half: 7 8 9 Upper half: 11 12 13

The lower quartile Q_1 is the median of the lower half, which is 8. The upper quartile Q_3 is the median of the upper half, which is 12. So

interquartile range $= Q_3 - Q_1 = 12 - 8 = 4$.

(b) Arrange the data in numerical order.

7 8 11 13 15 18 19 20 22

Since the number of data values (9) is odd, find the median $Q_2 = 15$ and delete it.

7 8 11 13 18 19 20 22

This automatically divides the data into lower and upper halves.

The median of the lower half is the lower quartile, so $Q_1 = \frac{1}{2}(8+11) = 9.5$, and the median of the upper half is the upper quartile, so $Q_3 = \frac{1}{2}(19+20) = 19.5$.

The interquartile range is $Q_3 - Q_1 = 19.5 - 9.5 = 10$.

In Chapter 2 you saw how the median of the heights of female students taken from the 'Brain size' datafile could be found with the aid of the stem-and-leaf diagram in Fig. 2.1. The diagram is reproduced in Fig. 3.1.

15	7 9	(2)
16	0 0 **4 4** 4 4 5 8 8 8 9 9 **9**	(13)
17	**3** 3 4 5 9	(5)

Key: 17 | 3 = 173 cm

Fig. 3.1. Stem-and-leaf diagram of the heights of female students.

Since there are 20 students, the upper and lower halves of the data set will contain 10 values each. The lower quartile is then at the position which is equivalent to the median of the lower half. This is half way between the 5th and 6th values (in ascending order). These are shown in bold type in Fig. 3.1.

Therefore $Q_1 = \frac{1}{2}(164 + 164) = 164$.

Similarly the upper quartile is at the position which is equivalent to the median of the upper half of the data set. This is halfway between the 15th and 16th values (in ascending order). These are also shown in bold type in Fig. 3.1.

Therefore $Q_3 = \frac{1}{2}(169 + 173) = 171$.

The interquartile range is therefore $Q_3 - Q_1 = 171 - 164 = 7$.

It is quite likely that the size of a data set will be much larger than the ones which you have met so far. Larger data sets are usually organised into frequency tables and it is then necessary to think carefully about how to find the position of the quartiles. In Chapter 2 you saw how to find the median of a set of data which referred to the numbers of brothers and sisters of children in a school. This data set was given in Table 2.2. Table 3.2 below reproduces Table 2.2.

Number of brothers and sisters	Frequency	Cumulative frequency
0	36	36
1	94	130
2	48	178
3	15	193
4	7	200
5	3	203
6	1	204
	Total: 204	

Table 3.2. Frequency distribution of the number of brothers and sisters of the children at a school.

There are 204 observations. This means that each half will have 102 data values.

The position of the lower quartile, Q_1, will be halfway between the 51st and 52nd values (in ascending order). From the cumulative frequency column you can see that both values are 1, so $Q_1 = 1$. The position of the upper quartile, Q_3, will be halfway between the $(102 + 51)$th and $(102 + 52)$th values (in ascending order); that is, between the 153rd and 154th values. From the cumulative frequency column you can see that both values are 2, so $Q_3 = 2$.

For continuous variables large data sets are usually grouped and so the individual values are lost. The quartiles are estimated from a cumulative frequency graph using a method similar to that described in Chapter 2 to find the median.

Table 3.3 gives the frequency distribution for the playing times of the selection of CDs which you first met in Table 2.3.

Playing time, x (min)	Class boundaries	Frequency	Cumulative frequency
40±44	$39.5 \leqslant x < 44.5$	1	1
45–49	$44.5 \leqslant x < 49.5$	7	8
50–54	$49.5 \leqslant x < 54.5$	12	20
55–59	$54.5 \leqslant x < 59.5$	24	44
60–64	$59.5 \leqslant x < 64.5$	29	73
65–69	$64.5 \leqslant x < 69.5$	14	87
70–74	$69.5 \leqslant x < 74.5$	5	92
75–79	$74.5 \leqslant x < 79.5$	3	95
		Total: 95	

Table 3.3. Playing times of 95 CDs.

To obtain an estimate of the lower quartile of the playing times you read off the value corresponding to a cumulative frequency equal to one-quarter of the total frequency, which in this case is $\frac{1}{4} \times 95 = 23.75$. From the cumulative frequency graph in Fig. 3.4 you can see that $Q_1 \approx 55$ minutes. Similarly you find an estimate of the upper quartile by reading off the value corresponding to a cumulative frequency equal to three-quarters of the total frequency, which is $\frac{3}{4} \times 95 = 71.25$. From the cumulative frequency graph in Fig. 3.4 this method gives $Q_3 \approx 64$ minutes.

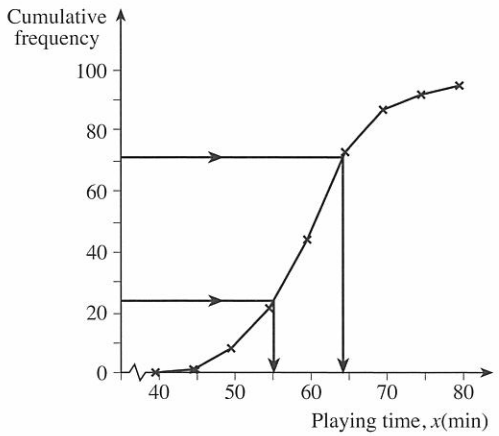


Fig. 3.4. Cumulative frequency graph for the data in Table 3.3.

Then the interquartile range is

$$Q_3 - Q_1 \approx 64 - 55 = 9,$$

so the interquartile range is approximately 9 minutes.

You may be wondering how to interpret the interquartile range for a set of data. For example, is the value of 9 in the example above large or small? The answer is that you cannot tell without more information. Normally you would be comparing the spreads of two or more data sets. You can then make a more sensible comment on whether a particular interquartile range is large or small by comparing its size with the other interquartile ranges. The following example illustrates this idea.

Example 3.3.2

Two people did separate traffic surveys at different locations. Each person noted down the speed of 50 cars which passed their observation point. The results are given in Table 3.5.

Speed, v (km h^{-1})	A frequency	B frequency
$0 \leqslant v < 20$	7	1
$20 \leqslant v < 40$	11	3
$40 \leqslant v < 60$	13	5
$60 \leqslant v < 80$	12	20
$80 \leqslant v < 100$	5	18
$100 \leqslant v < 120$	2	3
	Totals: 50	50

Table 3.5. Speeds of 50 cars at each of two locations.

(a) Draw a cumulative frequency graph for each set of data and use it to estimate the median speed and the interquartile range of speeds at each observation point.

(b) Use your results to part (a) to comment on the locations.

(a)

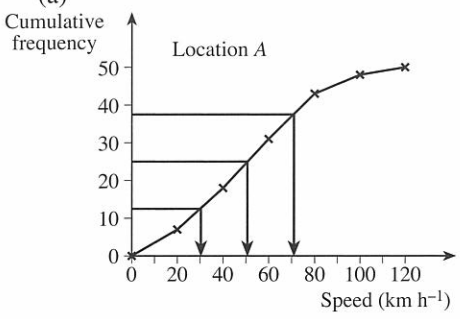


Fig. 3.6. Cumulative frequency graph for the distribution of speeds of 50 cars at location A.

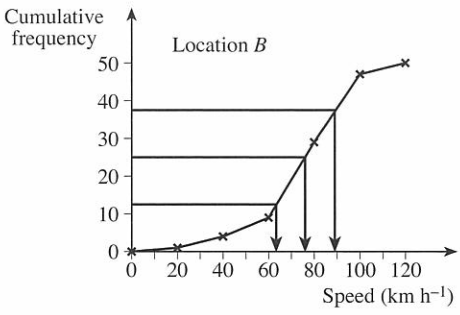


Fig. 3.7. Cumulative frequency graph for the distribution of speeds of 50 cars at location B.

From Fig. 3.6, you can estimate the median and quartiles for the cars at location A. The median corresponds to a cumulative frequency of 25 and, from Fig. 3.6, it is approximately 51. The lower quartile corresponds to a cumulative frequency of 12.5, and it is approximately 30. The upper quartile corresponds to a cumulative frequency of 37.5, and it is approximately 71.

From Fig. 3.7, you can estimate the median and quartiles for the cars at location B. The values of the median and the quartiles are approximately 76, 64 and 89.

(b) You can now compare the medians and the interquartile ranges.

For A the median is 51 and the interquartile range is 41.

For B the median is 76 and the interquartile range is 25.

The median speed at A is lower than the median speed at B and the interquartile range at A is greater than the interquartile range at B. So at location A the cars go more slowly and there is a greater variation in their speeds. Perhaps B is on or near a motorway, and A may be in a town near some point of congestion. You cannot say for certain what types of location A and B are, but the summary values do give you an idea of the type of road at each position.

3.4 The five-number summary

One helpful way of summarising data is to give values which provide essential information about the data set. One such summary is called the **five-number summary**. This summary gives the median, Q_2, the lower quartile Q_1, the upper quartile Q_3, the minimum value and the maximum value.

Example 3.4.1

The data below give the number of fish caught each day over a period of 11 days by an angler. Give a five-number summary of the data.

0 2 5 2 0 4 4 8 9 8 8

Rearranging the data in order gives:

0 0 2 2 4 4 5 8 8 8 9

The median value is $Q_2 = 4$. As the number of data values is odd, deleting the middle one and finding the medians of the lower and upper halves gives

$$Q_1 = 2 \quad \text{and} \quad Q_3 = 8.$$

The five-number summary is then the minimum value, 0, the lower quartile, 2, the median, 4, the upper quartile, 8, and the maximum value, 9.

3.5 Box-and-whisker plots

You can convert a five-number summary into a useful diagram, called a **box-and-whisker plot** or a **boxplot**. To draw a box-and-whisker plot, first draw a scale, preferably using graph paper. You can draw the scale vertically or horizontally, but in this book, the scale and the diagram are always drawn horizontally. Above the scale draw a box (or rectangle) in which the left side is above the point corresponding to the lower quartile and the right side is above the point corresponding to the upper quartile. Then mark a third line inside the box above the point which corresponds to the median value. After this you draw the two whiskers. The left whisker extends from the lower quartile to the minimum value and the right whisker extends from the upper quartile to the maximum. Fig. 3.8 shows the box-and-whisker plot for the data in Example 3.4.1.

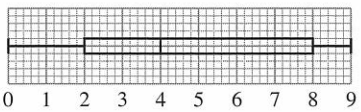


Fig. 3.8. Box-and-whisker plot for the distribution of the numbers of fish caught by an angler over 11 days.

In a box-and-whisker plot the box itself indicates the location of the middle 50% of the data. The whiskers then show how the data is spread overall.

Another important feature of a set of data is its shape when represented as a frequency diagram. The three pictures in Fig. 3.9 show three different shapes which commonly occur when you draw histograms or bar charts.

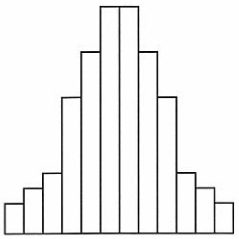


Fig. 3.9a

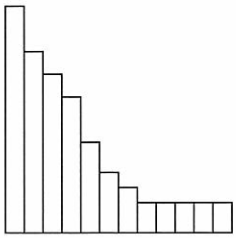


Fig. 3.9b

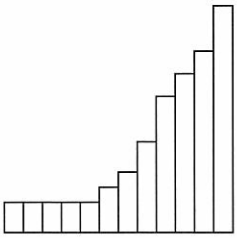


Fig. 3.9c

Fig 3.9. Possible shapes of frequency distributions.

The distribution in Fig. 3.9a is symmetrical. If a distribution is not symmetrical it is said to be **skewed**, or to have **skewness**. The distribution in Fig. 3.9a may therefore be said to have zero skewness. You were briefly introduced to the term 'skewed' in Section 2.9.

The distribution in Fig. 3.9b is certainly not symmetrical; there is a 'tail' which stretches towards the higher values. This distribution is said to have **positive skew**, or to be **skewed positively**.

The distribution in Fig. 3.9c is also not symmetrical; there is a 'tail' which stretches towards the lower values. This distribution is said to have **negative skew**, or to be **skewed negatively**.

Another method of assessing the skewness of a distribution is to use the quartiles Q_1, Q_2 and Q_3. Remember that Q_2 denotes the median, and that Q_1 and Q_3 denote the lower and upper quartiles.

If $Q_3 - Q_2 \approx Q_2 - Q_1$, then the distribution is said to be (almost) symmetrical, and a box-and-whisker plot of such data, as in Fig. 3.10, would show a box in which the line corresponding to the median is in the centre of the box.

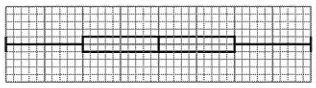


Fig 3.10. Box-and-whisker plot for a set of data in which $Q_3 - Q_2 \approx Q_2 - Q_1$.

If $Q_3 - Q_2 > Q_2 - Q_1$, as in Fig. 3.11, then the distribution is said to have positive skew, and the line representing the median would be nearer to the left side of the box.

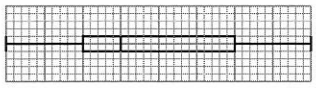


Fig 3.11. Box-and-whisker plot for a set of data in which $Q_3 - Q_2 > Q_2 - Q_1$.

If $Q_3 - Q_2 < Q_2 - Q_1$, as in Fig. 3.12, then the distribution would be said to have negative skew, and the line representing the median would be nearer to the right side of the box.

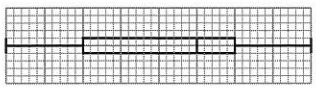


Fig 3.12. Box-and-whisker plot for a set of data in which $Q_3 - Q_2 < Q_2 - Q_1$.

For the data in Example 3.4.1, $Q_1 = 2$, $Q_2 = 4$ and $Q_3 = 8$, so $Q_3 - Q_2 = 8 - 4 = 4$ and $Q_2 - Q_1 = 4 - 2 = 2$. The set of data in Example 3.4.1 therefore has positive skew.

The length of the whiskers can also give some indication of skewness. If the left whisker is shorter than the right whisker, then that would tend to indicate positive skew; whereas if the right whisker is shorter than the left whisker, negative skew would be implied.

It is possible for the data to give different results for skewness depending on what measure you use. For example, it is perfectly possible for a box-and-whisker plot to have $Q_3 - Q_2 > Q_2 - Q_1$, indicating positive skew, but for the left whisker to be longer than the right whisker, which would tend to suggest negative skew. In such cases you must make a judgement about which method of assessing skewness you think is the more important. Fortunately data of this sort do not occur commonly.

3.6* Outliers

The quartiles of a data set can also be used to assess whether the data set has any outliers. Outliers are unusual or 'freak' values which differ greatly in magnitude from the majority of the data values. But just how large or small does a value have to be to be an outlier? There is no simple answer to this question, but one 'rule of thumb' developed by the statistician John Tukey is to use 'fences'.

The upper fence is at a value 1.5 times the interquartile range above the upper quartile:

$$\text{Upper fence} = Q_3 + 1.5(Q_3 - Q_1).$$

The lower fence is at a value 1.5 times the interquartile range below the lower quartile:

$$\text{Lower fence} = Q_1 - 1.5(Q_3 - Q_1).$$

John Tukey then said that any value which is bigger than the upper fence or smaller than the lower fence is considered to be an outlier.

For the data of Example 3.4.1, $Q_1 = 2$ and $Q_3 = 8$, so

$$\text{upper fence} = Q_3 + 1.5(Q_3 - Q_1) = 8 + 1.5 \times (8 - 2) = 17,$$

$$\text{lower fence} = Q_1 - 1.5(Q_3 - Q_1) = 2 - 1.5(8 - 2) = -7.$$

In this case there is no value above 17 or below -7, so this data set does not contain any values which could be said to be outliers.

Exercise 3A

1 Find the range and interquartile range of each of the following data sets.

(a) 7 4 14 9 12 2 19 6 15

(b) 7.6 4.8 1.2 6.9 4.8 7.2 8.1 10.3 4.8 6.7

2 Find the interquartile range of the leaf lengths displayed in Exercise 1A Question 1.

3 The number of times each week that a factory machine broke down was noted over a period of 50 consecutive weeks. The results are given in the following table.

Number of breakdowns	0	1	2	3	4	5	6
Number of weeks	2	12	14	8	8	4	2

Find the interquartile range of the number of breakdowns in a week.

4 For the data in Miscellaneous exercise 1 Question 6, find the lower and upper quartiles of the annual salaries.

5 For the data in Miscellaneous exercise 1 Question 8, find the median and interquartile range for the traffic noise levels in the two streets. Use the statistics to compare the noise levels in the two streets.

6 The audience size in a theatre performing a long-running detective play was monitored over a period of one year. The sizes for Monday and Wednesday nights are summarised in the following table.

Audience size	50–99	100–199	200–299	300–399	400–499	500–599
Number of Mondays	12	20	12	5	3	0
Number of Wednesdays	2	3	20	18	5	4

Compare the audience sizes on Mondays and Wednesdays.

7 The following stem-and-leaf diagrams refer to the datafile 'Cereals' in Chapter 1. They are the ratings of the cereals with fat content 0 and with fat content 1.

Fat content 0

2	9	(1)
3	1 3 5 6	(4)
4	1 1 1 2 2 4 6 7	(8)
5	3 3 3 5 8 9	(6)
6	0 1 3 5 8	(5)
7	3 4	(2)
8		(0)
9	4	(1)

Fat content 1

2	2 3 4 7 8 8 9	(7)
3	0 1 1 2 6 6 6 7 8 9 9 9 9	(13)
4	0 7 9	(3)
5	0 0 2 2 5 9	(6)
6	8	(1)

Key: 4|7 means 47

Compare the two sets of ratings by finding the ranges, medians and quartiles.

8 Draw box-and-whisker plots for data which have the following five-number summaries, and in each case describe the shape of the data.

(a)	6.0 kg	10.2 kg	12.7 kg	13.2 kg	15.7 kg
(b)	±12 °C	±8 °C	±6 °C	3 °C	11 °C
(c)	37 m	48 m	60 m	72 m	82 m

9 State, giving reasons, whether box-and-whisker plots or histograms are better for comparing two distributions.

10 The following figures are the amounts spent on food by a family for 13 weeks.

\$48.25 \$43.70 \$52.83 \$49.24 \$58.28 \$55.47 \$47.29
\$51.82 \$58.42 \$38.73 \$42.76 \$50.42 \$40.85

(a) Obtain a five-number summary of the data.

(b) Construct a box-and-whisker plot of the data.

(c) Describe any skewness of the data.

11* The lower and upper quartiles for a data set are 56 and 84. Decide which of the following data values would be classified as an outlier according to the criteria of Section 3.6.

(a) 140 (b) 10 (c) 100

12* For the data of Exercise 1A Question 3, construct a box-and-whisker plot.

(a) Calculate the inner and outer fences.

(b) State, giving a reason, whether there are any outliers.

(c) Comment on the shape of the distribution.

3.7 Variance and standard deviation

One of the reasons for using the interquartile range in preference to the range as a measure of spread is that it takes some account of how the interior values are spread rather than concentrating solely on the spread of the extreme values. The interquartile range, however, does not take account of the spread of all of the data values and so, in some sense, it is still an inadequate measure. An alternative measure of spread which does take into account the spread of all the values can be devised by finding how far each data value is from the mean. To do this you would calculate the quantities $x_i - \bar{x}$ for each x_i.

The example in Section 2.4 used the playing times, in minutes, of nine CDs.

49 56 55 68 61 57 61 52 63

The mean of these times was found to be 58 minutes. If the mean is subtracted from each of the original data values you get the following values.

−9 −2 −3 10 3 −1 3 −6 5

If you ignore the negative signs, then the resulting values give an idea of the distance of each of the original values from the mean. So these distances would be

9 2 3 10 3 1 3 6 5.

The mean of these distances would be a sensible measure of spread. It would represent the mean distance from the mean.

In this case the mean distance would be

$$\frac{1}{9}(9+2+3+10+3+1+3+6+5) = \frac{1}{9} \times 42 = 4.66\ldots.$$

To represent this method with a simple formula it is necessary to use the modulus symbol $|v|$, which denotes the magnitude, or numerical value, of v. It is now possible to write a precise formula for the mean distance:

$$\text{mean distance} = \frac{1}{n}\sum|x_i - \bar{x}|.$$

Unfortunately a formula involving the modulus sign is awkward to handle algebraically. The modulus sign can be avoided by squaring each of the quantities $x_i - \bar{x}$.

This leads to the expression $\frac{1}{n}\sum(x_i - \bar{x})^2$ as a measure of spread.

This quantity is called the **variance** of the data values. It is the mean of the squared distances from the mean. So, for the data on playing times of CDs,

$$\begin{aligned}\text{variance} &= \frac{1}{n}\sum(x_i - \bar{x})^2 \\ &= \tfrac{1}{9}\left(9^2 + 2^2 + 3^2 + 10^2 + 3^2 + 1^2 + 3^2 + 6^2 + 5^2\right) \\ &= \tfrac{1}{9}(81 + 4 + 9 + 100 + 9 + 1 + 9 + 36 + 25) = \tfrac{274}{9} = 30.4\ldots.\end{aligned}$$

If the data values $x_1, x_2, \ldots, x_n$ have units associated with them, then the variance will be measured in units^2. In the example the data values were measured in minutes and therefore the variance would be measured in minutes^2. This is something which can be avoided by taking the positive square root of the variance. The positive square root of the variance is known as the **standard deviation**, often shortened to 'SD', and it always has the same units as the original data values. The formula for standard deviation is

$$\sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}.$$

So the standard deviation of the playing times of the nine CDs is $\sqrt{30.4\ldots} = 5.52$, correct to 3 significant figures.

The calculation of the variance can be quite tedious, particularly when the mean is not a whole number. Fortunately, there is an alternative formula which is easier to use:

$$\text{variance} = \frac{1}{n}\left(x_1^2 + x_2^2 + \ldots + x_n^2\right) - \bar{x}^2.$$

This can be written in Σ-notation as

$$\text{variance} = \frac{1}{n}\sum x_i^2 - \bar{x}^2.$$

Using this alternative formula with the data on the playing times of CDs gives

$$\begin{aligned}\text{variance} &= \frac{1}{n}\sum x_i^2 - \bar{x}^2 \\ &= \tfrac{1}{9}\left(49^2 + 56^2 + 55^2 + 68^2 + 61^2 + 57^2 + 61^2 + 52^2 + 63^2\right) - 58^2 \\ &= \tfrac{1}{9}(2401 + 3136 + 3025 + 4624 + 3721 + 3249 + 3721 + 2704 + 3969) - 3364 \\ &= \tfrac{1}{9} \times 30\,550 - 3364 = 3394.4\ldots - 3364 = 30.4\ldots.\end{aligned}$$

This is the same value found by using the original formula. This does not, of course, prove that the two formulae are always equivalent to each other. A proof is given in Section 3.8.

The **variance** of a set of data values $x_1, x_2, \ldots, x_n$ whose mean is

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum x_i$$

is given by either of the two alternative formulae

$$\text{variance} = \frac{1}{n}\sum (x_i - \bar{x})^2 \quad \text{or} \quad \text{variance} = \frac{1}{n}\sum x_i^2 - \bar{x}^2. \qquad (3.1), (3.2)$$

The **standard deviation** is the square root of the variance.

Example 3.7.1

The 12 boys and 13 girls, in a class of 25 students, were given a test. The mean mark for the 12 boys was 31 and the standard deviation of the boys' marks was 6.2. The mean mark for the girls was 36 and the standard deviation of the girls' marks was 4.3. Find the mean mark and standard deviation of the marks of the whole class of 25 students.

Let $x_1, x_2, \ldots, x_{12}$ be the marks of the 12 boys in the test and let $y_1, y_2, \ldots, y_{13}$ be the marks of the 13 girls in the test.

Since the mean of the boys' marks is 31, $\frac{\sum x}{12} = 31$, so $\sum x = 12 \times 31 = 372$.

As the standard deviation of the boys' marks is 6.2, the variance is $6.2^2 = 38.44$.

Therefore, using Equation 3.2,

$$38.44 = \frac{\sum x^2}{12} - 31^2, \text{ which gives } \sum x^2 = 12 \times \left(38.44 + 31^2\right) = 11\,993.28.$$

Similarly,

$$\sum y = 13 \times 36 = 468, \text{ and}$$

$$\sum y^2 = 13 \times \left(4.3^2 + 36^2\right) = 17\,088.37.$$

The overall mean is $\frac{\sum x + \sum y}{25} = \frac{372 + 468}{25} = \frac{840}{25} = 33.6.$

The overall variance is $\frac{\sum x^2 + \sum y^2}{25} - 33.6^2 = \frac{11\,993.28 + 17\,088}{25} - 33.6^2$

$$= 34.306.$$

The overall standard deviation is $\sqrt{34.306} = 5.86$, correct to 3 significant figures.

Although the standard deviation makes use of all the data values, it suffers from the same disadvantage as the mean, namely that an outlier can have an undue influence. Either a very low value or a very high value will increase the standard deviation considerably. Consider the data below.

45 46 46 48 49 50 52

It is left as an exercise for you to check the following values for this data set.

Standard deviation = 2.33, correct to 3 significant figures.
Interquartile range = $50 \pm 46 = 4$.

Now consider what would happen if the lowest value were 25 instead of 45.

25 46 46 48 49 50 52

The interquartile range is unchanged, but you can check that the standard deviation is now 8.44, correct to 3 significant figures. This is several times larger than its previous value.

3.8* Proof of the equivalence of the variance formulae

You may omit this section if you wish.

First note that $\bar{x} = \dfrac{x_1 + x_2 + \ldots + x_n}{n}$, so $n\bar{x} = x_1 + x_2 + \ldots + x_n$.

$$\text{Then variance} = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$= \frac{1}{n}\left\{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2\right\}$$

$$= \frac{1}{n}\left\{\left({x_1}^2 - 2x_1\bar{x} + \bar{x}^2\right) + \left({x_2}^2 - 2x_2\bar{x} + \bar{x}^2\right) + \ldots + \left({x_n}^2 - 2x_n\bar{x} + \bar{x}^2\right)\right\}$$

$$= \frac{1}{n}\left\{\left({x_1}^2 + {x_2}^2 + \ldots + {x_n}^2\right) - 2\bar{x}\left(x_1 + x_2 + \ldots + x_n\right) + \left(\overbrace{\bar{x}^2 + \bar{x}^2 + \ldots + \bar{x}^2}^{n \text{ of these}}\right)\right\}$$

$$= \frac{1}{n}\left\{\left({x_1}^2 + {x_2}^2 + \ldots + {x_n}^2\right) - 2\bar{x} \times n\bar{x} + n\bar{x}^2\right\}$$

$$= \frac{1}{n}\left\{\left({x_1}^2 + {x_2}^2 + \ldots + {x_n}^2\right) - 2n\bar{x}^2 + n\bar{x}^2\right\}$$

$$= \frac{1}{n}\left\{\left({x_1}^2 + {x_2}^2 + \ldots + {x_n}^2\right) - n\bar{x}^2\right\}$$

$$= \frac{1}{n}\sum {x_i}^2 - \bar{x}^2.$$

This shows that the two formulae for variance are equivalent.

Exercise 3B

1 State or find the mean of

(a) 1, 2, 3, 4, 5, 6, 7, (b) 4, 12, −2, 7, 0, 9.

Using the formula $\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$, find the standard deviation of each data set.

2 Find the standard deviation of the following data sets, using the formula $\sqrt{\frac{1}{n}\sum(x-\bar{x})^2}$.

(a) 2, 1, 5.3, −4.2, 6.7, 3.1 (b) 15.2, 12.3, 5.7, 4.3, 11.2, 2.5, 8.7

3 The masses, x grams, of the contents of 25 tins of Brand A anchovies are summarised by $\sum x = 1268.2$ and $\sum x^2 = 64\,585.16$. Find the mean and variance of the masses. What is the unit of measurement of the variance?

4 The standard deviation of 10 values of a variable is 2.8. The sum of the squares of the 10 values is 92.8. Find the mean of the 10 values.

5 The mean and standard deviation of the heights of 12 boys in a class are 148.8 cm and 5.4 cm respectively. A boy of height 153.4 cm joins the class. Find the mean and standard deviation of the heights of the 13 boys.

6 The runs made by two batsmen, Anwar and Qasim, in 12 innings during the 1999 cricket season are shown in the following table.

Anwar	23	83	40	0	89	98	71	31	102	48	15	18
Qasim	43	32	61	75	68	92	17	15	25	43	86	12

Giving your reasons, state which batsman you consider to be

(a) better, (b) more consistent.

7 The following stem-and-leaf diagrams are for the masses of 20 female students and 18 male students from the datafile 'Brain size' in Chapter 1.

Females

4	8	(1)
5	2 4 4 5 8 8	(6)
6	1 2 3 3 4 5 6 6 7 9	(10)
7	0 2 9	(3)
8		

Males

4		
5		
6	0 1 5 5 7 9	(6)
7	0 8 8 8	(4)
8	1 1 2 2 4 5 7 7	(8)

Key: 6|1 means 61 kg

Summary: $\sum f = 1246$, $\sum f^2 = 78\,704$,
$\sum m = 1360$, $\sum m^2 = 104\,162$,

where f and m represent the masses of female and male students respectively.

Compare the masses of the females and males by drawing box-and-whisker plots and calculating the means and standard deviations of the masses.

3.9 Calculating variance from a frequency table

Table 3.13 reproduces Table 2.5, which gave the frequency distribution of the numbers of brothers and sisters of children in a school.

Number of brothers and sisters, x_i	Frequency, f_i	$x_i f_i$
0	36	0
1	94	94
2	48	96
3	15	45
4	7	28
5	3	15
6	1	6
Totals:	$\sum f_i = 204$	$\sum x_i f_i = 284$

Table 3.13. Frequency distribution of the number of brothers and sisters of children in a school.

In order to calculate the variance you need first to find the mean. This was done in Section 2.6, and the mean was $\frac{284}{204} = 1.39\ldots$.

Since the 204 values consist of 36 '0's, 94 '1's, 48 '2's and so on,

$$x_1^2 + x_2^2 + \ldots + x_n^2 = \left(\overbrace{0^2 + 0^2 + \ldots + 0^2}^{36 \text{ of these}}\right) + \left(\overbrace{1^2 + 1^2 + \ldots + 1^2}^{94 \text{ of these}}\right) + \ldots$$

$$+ \left(\overbrace{4^2 + 4^2 + \ldots + 4^2}^{7 \text{ of these}}\right) + \left(\overbrace{5^2 + 5^2 + 5^2}^{3 \text{ of these}}\right) + 6^2$$

$$= \left(0^2 \times 36\right) + \left(1^2 \times 94\right) + \left(2^2 \times 48\right)$$

$$+ \left(3^2 \times 15\right) + \left(4^2 \times 7\right) + \left(5^2 \times 3\right) + \left(6^2 \times 1\right)$$

$$= 0 + 94 + 192 + 135 + 112 + 75 + 36 = 644.$$

You can include this calculation in the table by adding a fourth column for $x_i^2 \times f_i$.

Number of brothers and sisters, x_i	Frequency, f_i	$x_i f_i$	$x_i^2 f_i$
0	36	0	0
1	94	94	94
2	48	96	192
3	15	45	135
4	7	28	112
5	3	15	75
6	1	6	36
Totals:	$\sum f_i = 204$	$\sum x_i f_i = 284$	$\sum x_i^2 f_i = 644$

Table 3.14. Calculation of the variance for the data in Table 3.13.

So the variance is $\frac{644}{204} - (1.39\ldots)^2 = 1.218\ldots = 1.22$, correct to 3 significant figures.

The standard deviation is then $\sqrt{1.218\ldots} = 1.10$, correct to 3 significant figures.

To summarise the method used to find the variance:

The variance of data given in a frequency table in which the variable takes the value x_1 with frequency f_1, the value x_2 with frequency f_2 and so on is given by the two formulae

$$\text{variance} = \frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i} \quad \text{or} \quad \text{variance} = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2. \qquad (3.3), (3.4)$$

The second formula is usually easier to use.

If the data are grouped, you need a single value to represent each class. In Section 2.6 you saw that the most reasonable choice was the mid-class value. After you have made this simplifying assumption, then the calculation proceeds in the same way as in Table 3.14.

Example 3.9.1
Calculate an estimate of the variance of the data given in Table 2.6.

Table 3.15 reproduces Table 2.6 with an extra column representing $x_i^2 f_i$.

Playing time, x (min)	Class boundaries	Frequency, f_i	Mid-class value, x_i	$x_i f_i$	$x_i^2 f_i$
40–44	$39.5 \leqslant x < 44.5$	1	42	42	1764
45–49	$44.5 \leqslant x < 49.5$	7	47	329	15 463
50–54	$49.5 \leqslant x < 54.5$	12	52	624	32 448
55–59	$54.5 \leqslant x < 59.5$	24	57	1368	77 976
60–64	$59.5 \leqslant x < 64.5$	29	62	1798	111 476
65–69	$64.5 \leqslant x < 69.5$	14	67	938	62 846
70–74	$69.5 \leqslant x < 74.5$	5	72	360	25 920
75–79	$74.5 \leqslant x < 79.5$	3	77	231	17 787
	Totals:	$\sum f_i = 95$		$\sum x_i f_i = 5\,690$	$\sum x_i^2 f_i = 345\,680$

Table 3.15. Calculation of the variance of the playing times for 95 CDs.

Using the formula $\text{variance} = \frac{\sum x_i^2 f_i}{\sum f_i} - \bar{x}^2$,

$$\text{variance} = \frac{345\,680}{95} - \left(\frac{5690}{95}\right)^2 = 51.4, \text{ correct to 3 significant figures.}$$

The variance is therefore 51.4 minutes^2, correct to 3 significant figures.

You should remember that, just as with the calculation of the mean in Section 2.6, the value of the variance calculated from grouped data is only an estimate. This is because the individual values have been replaced by mid-class values.

3.10 Making the calculation of variance easier

You saw in Section 2.7 that you could simplify the calculation of the mean. To find the mean of 907, 908, 898, 902 and 897 you subtracted 900 from each of the numbers, giving 7, 8, −2, 2 and −3. You then found the mean of these numbers, which was 2.4. To recover the mean of the original five numbers 907, 908, 898, 902 and 897 you simply added 900 to 2.4 to give a mean for the original numbers of 902.4. You can also use this idea of simplifying the numbers when calculating variance, but the details are slightly different.

To find the variance of 907, 908, 898, 902 and 897, subtract 900 as before to obtain 7, 8, −2, 2 and −3. The spread of the two sets of numbers will be identical, so the variance of 907, 908, 898, 902 and 897 will be the same as the variance of 7, 8, −2, 2 and −3.

The variance of 7, 8, −2, 2 and −3 is given by

$$\tfrac{1}{5}\left(7^2+8^2+(-2)^2+2^2+(-3)^2\right)-2.4^2=20.24.$$

This value will also represent the variance of the original numbers. Notice also that the standard deviation will be $\sqrt{20.24}=4.498\ldots$ for both sets of data. After you found the mean of the new numbers it was necessary to add 900 to the new mean to recover the mean of the original set of numbers. You do *not* need to add anything to the variance of the new numbers to recover the variance of the original set of numbers. This is because the amount of spread is the same for both sets of data.

Example 3.10.1

The heights, x cm, of a sample of 80 female students are summarised by the equations

$$\sum(x-160)=240 \quad \text{and} \quad \sum(x-160)^2=8720.$$

Find the standard deviation of the heights of the 80 female students.

Let $y=x-160$ and then $\sum y=240$ and $\sum y^2=8720$.

The variance of the y-values is given by

$$\text{variance}=\frac{\sum y^2}{n}-\bar{y}^2=\frac{8720}{80}-\left(\frac{240}{80}\right)^2=109-3^2=100.$$

Therefore the standard deviation of the y-values is 10.

But the standard deviation of the x-values will be the same as the standard deviation of the y-values. Therefore the standard deviation of the female students' heights will also be 10.

Exercise 3C

1 The number of absences each day among employees in an office was recorded over a period of 96 days, with the following results.

Number of absences	0	1	2	3	4	5
Number of days	54	24	11	4	2	1

Calculate the mean and variance of the number of daily absences, setting out your work in a table similar to Table 3.14.

2 Plates of a certain design are painted by a particular factory employee. At the end of each day the plates are inspected and some are rejected. The table shows the number of plates rejected over a period of 30 days.

Number of rejects	0	1	2	3	4	5	6
Number of days	18	5	3	1	1	1	1

Show that the standard deviation of the daily number of rejects is approximately equal to one-quarter of the range.

3 The times taken in a 20 km race were noted for 80 people. The results are summarised in the following table.

Time (minutes)	60–80	80–100	100–120	120–140	140–160	160–180	180–200
No. of people	1	4	26	24	10	7	8

Estimate the variance of the times of the 80 people in the race.

4 The mass of coffee in each of 80 packets of a certain brand was measured correct to the nearest gram. The results are shown in the following table.

Mass (grams)	244–246	247–249	250–252	253–255	256–258
Number of packets	10	20	24	18	8

Estimate the mean and standard deviation of the masses, setting out your work in a table similar to Table 3.15.

State two ways in which the accuracy of these estimates could be improved.

5 Here are 10 values of a variable x. Find the variance using the formula $u = x - 20$.

18.9 20.7 19.3 20.1 21.3 19.6 20.5 20.9 18.8 20.8

6 The formula $u = x + 20$ is used to find the standard deviation of the values of x given in a frequency table. It is found that $\sum f_i = 40$, $\sum u_i f_i = 112$ and $\sum u_i^2 f_i = 10\,208$. Find the mean and variance of the values of x.

7 At the start of a new school year, the heights of the 100 new pupils entering the school are measured. The results are summarised in the following table. The 10 pupils in the 110– group have heights not less than 110 cm but less than 120 cm.

Height (h cm)	100–	110–	120–	130–	140–	150–	160–
Number of pupils	2	10	22	29	22	12	3

By using the formula $u = h - 135$, obtain estimates of the mean and variance of the heights of the 100 pupils.

8 The ages, in completed years, of the 104 workers in a company are summarised as follows.

Age (years)	16–20	21–25	26–30	31–35	36–40	41–50	51–60	61–70
Frequency	5	12	18	14	25	16	8	6

Estimate the mean and standard deviation of the workers' ages.

In another company, with a similar number of workers, the mean age is 28.4 years and the standard deviation is 9.9 years. Briefly compare the age distribution in the two companies.

3.11 Choosing how to represent data

In the first three chapters of this book you have met a variety of ways of representing data. The purpose of this section is to give an overview of these methods and to discuss the advantages and disadvantages of the different methods in more detail.

When you are faced with a set of raw data for a variable, it is helpful to split it into groups in some way. This will allow you to see how the values of the variable are distributed. A very effective way of doing this for a small data set is to construct a stem-and-leaf diagram. You can then see the shape of the distribution by rotating the diagram through 90° anticlockwise. You can compare two sets of data by drawing a 'back-to-back' stem-and-leaf diagram. Fig. 3.16 shows such a diagram for the data in Exercise 3B Question 7. From it you can see that the masses for women have a greater spread than those for men but that the mass of a woman is, on average, less than that of a man.

Females												**Males**								
(1)									8	4										
(6)					8	8	5	4	4	2	5									
(10)	9	7	6	6	5	4	3	3	2	1	6	0	1	5	5	7	9			(6)
(3)								9	2	0	7	0	8	8	8					(4)
											8	1	1	2	2	4	5	7	7	(8)

Key: 6|1 means 61 kg

Fig. 3.16. Back-to-back stem-and-leaf diagram of the masses in kg of 20 female and 18 male students.

A stem-and-leaf diagram has the advantage that it contains all the original data values and so you can easily find the range, the median and the quartiles from it. You can also

calculate the mean and standard deviation exactly from a stem-and-leaf diagram because information about all the values is there.

For larger data sets, however, a stem-and-leaf diagram would be very tedious to draw, and it can look confusing because it contains so much information. In these cases it may be better to make a frequency table and draw a histogram in order to show the shape of the distribution. Histograms have the advantage over stem-and-leaf diagrams that you can group the data into classes of any width you like and these classes need not have the same widths.

From the scale on the horizontal axis of a histogram you can easily see the range of values that the variable takes. This information is not so easily seen from a stem-and-leaf diagram: you have to use the 'key' to interpret the values.

Some of the information in the original data set is lost when the data are assembled into a grouped frequency table. This is the price which you pay for making it easier to see the distribution of the data. As a result, values for the median, quartiles, mean and standard deviation which are found from a frequency table (or a histogram) are estimates rather than exact values.

Cumulative frequency diagrams are usually drawn in order to estimate the median and quartiles of grouped data. They are also useful for estimating the number of data values that lie below or above a given value of the variable.

Box-and-whisker plots give a compact means of showing the shape of a distribution. They have the advantage over histograms and stem-and-leaf diagrams that they give the lowest and highest values, the median and the quartiles directly. They are particularly useful when you want to compare several sets of related data. Fig. 3.17 shows box-plots of 'sugar' for different shelf numbers for data taken from the datafile 'Cereals' in Chapter 1. At a glance you can see that the distribution for shelf 1 is positively skewed, that for shelf 2 is negatively skewed and that for shelf 3 is symmetrical. More importantly, the sugar content for cereals on shelf 2, which is at 'child height', is higher than for the other two shelves.

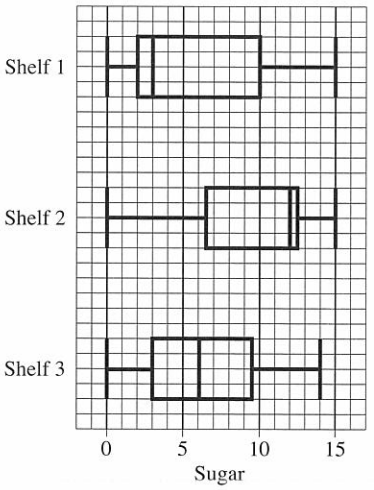


Fig. 3.17. Box-and-whisker plots for 'sugar' for different supermarket shelves.

Box-and-whisker plots have the disadvantage that, unlike histograms and stem-and-leaf diagrams, you cannot find the mean and standard deviation from them. Also, unlike histograms and stem-and-leaf diagrams, they give no indication of the size of the data set. For example, you cannot tell from Fig. 3.17 how many types of cereal there were on each of the shelves.

Once you have an idea of the distribution of a variable you are in a position to decide the most appropriate measures of location and spread for representing it. The location of a

data set can be represented by the mean, the median or the mode. The mean has the advantage that it uses all the data values but has the disadvantage that outliers affect it unduly. The median makes less use of the actual values but has the advantage that outliers do not affect it at all. The mode has the advantage that it can be used for qualitative as well as quantitative variables.

The spread of a data set can be represented by the range, the standard deviation or the interquartile range. The range has the disadvantage that it depends only on the highest and lowest values and so it is very sensitive to the presence of outliers. The standard deviation, like the mean, has the advantage that it uses all the data values, but it too has the disadvantage that outliers affect it unduly. The interquartile range, like the median, makes less use of the actual values, but it has the advantage that outliers do not affect it at all.

3.12 Practical activities

Each of Practical activities 1 to 4 requires you to obtain data for two or more different groups. For Practical activity 5 you are provided with the data.

For Practical activities 1 to 4 analyse the data by

(a) illustrating them in a way which allows you to compare the distributions of the groups;
(b) choosing suitable measures of location and spread for representing the groups and finding their values;
(c) comparing the results of (a) and (b) for the groups and commenting on them.

1 Wastepaper Select a student and ask them to try to throw a ball of paper into a wastepaper bin about 4 metres away with their 'natural' or stronger throwing arm. Ask them to repeat this until they are successful, and record the total number of attempts needed. Repeat this experiment with about 30 students.

Repeat this whole procedure with a second group of students but ask them to use their 'weaker' arm.

2 Age distributions From the internet or the library obtain data giving the age distribution of the population of your town or country for males and for females.

3 Newspapers Does the length of the sentences in a newspaper vary from one paper to another?

Choose two or more newspapers, preferably with different styles of journalism, and for each newspaper count the number of words per sentence for at least 100 sentences. (You should ignore the headlines.)

(Alternatively you could compare the styles of different authors.)

4 Just a minute! How well can people estimate time?

Ask at least 90 people to estimate a time interval of one minute. You will need to decide on a standard procedure for doing this. Record the value of the estimate to the nearest second. Have three distinct groups of at least 30 people, for example children of two different ages and adults.

5 Exam performance The data set below refers to a group of students who started to study A-level mathematics at the same time. There are three variables. These are: 'Gender', where 'F' denotes female and 'M' denotes male; 'Grade' which gives the grade obtained by the student in the International GCSE mathematics exam and 'Mark' which is the percentage which the student obtained in a test taken at the end of the first term of A-level study.

By means of appropriate diagrams and calculations, investigate (a) possible differences in exam performance between genders, (b) a possible relationship between IGCSE grade and test mark.

Gender	Grade	Mark	Gender	Grade	Mark	Gender	Grade	Mark
M	B	86	M	B	88	M	A	48
F	B	41	F	A*	89	M	A	67
M	B	58	F	A	76	F	B	59
F	B	50	M	B	59	M	A	50
F	B	70	M	B	65	F	A*	52
M	A	62	F	B	82	M	A	52
M	B	56	M	A*	100	M	A*	55
M	A*	74	M	A*	59	F	A	93
M	A	43	F	A	74	M	A	71
F	B	48	M	A*	62	F	B	66
M	B	54	F	A	92	F	B	50
M	B	67	F	A*	67	F	A	71
M	A*	54	M	A	63	F	A	84
M	B	69	F	A	83	F	B	67
M	A	64	F	B	72	F	B	83
M	A	73	F	B	75	M	A*	87
M	A*	73	F	A	64	M	A*	82
F	A	51	F	A	65	M	A*	97
M	A	64	F	B	56	M	B	54
M	A	74	F	A	57	F	A	85
F	B	64	F	A*	91	M	A	82

Miscellaneous exercise 3

1 Seven mature robins (*Erithacus rubecula*) were caught and their wingspans were measured. The results, in centimetres, were as follows.

23.1 22.7 22.1 24.2 23.9 20.9 25.2

Here are the corresponding figures for seven mature house sparrows (*Passer domesticos*).

22.6 24.1 23.5 21.8 21.0 24.4 22.8

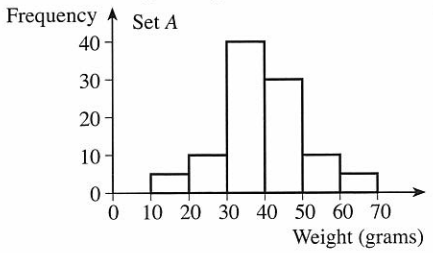
Find the mean and standard deviation of each species' wingspan, and use these statistics to compare the two sets of figures.

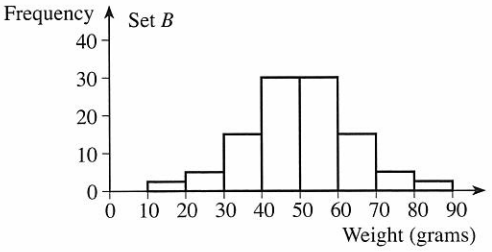
2 The depth of water in a lake was measured at 50 different points on the surface of the lake. The depths, x metres, are summarised by $\sum x = 934.5$ and $\sum x^2 = 19\,275.81$.

(a) Find the mean and variance of the depths.

(b) Some weeks later the water level in the lake rose by 0.23 m. What would be the mean and variance of the depths taken at the same points on the lake as before?

3 The following histograms are of two sets of 100 weights.





(a) Which set has the greater mean?

(b) Which set has the greater variance?

(c) Estimate the mean and variance of Set A.

(d) Why are your answers to (c) estimates?

4 The lengths of 120 nails of nominal length 3 cm were measured, each correct to the nearest 0.05 cm. The results are summarised in the following table.

Length (cm)	2.85	2.90	2.95	3.00	3.05	3.10	3.15
Frequency	1	11	27	41	26	12	2

(a) Draw a box-and-whisker plot of these results, taking the extremes as 2.825 cm and 3.175 cm.

(b) Estimate the standard deviation.

(c) It is claimed that for a roughly symmetrical distribution the statistic obtained by dividing the interquartile range by the standard deviation is approximately 1.3. Calculate the value of this statistic for these data, and comment.

5 Three statistics students, Ali, Les and Sam, spent the day fishing. They caught three different types of fish and recorded the type and mass (correct to the nearest 0.01 kg) of each fish caught. At 4 p.m. they summarised the results as follows.

	Number of fish by type			All fish caught	
	Perch	Tench	Roach	Mean mass (kg)	Standard deviation (kg)
Ali	2	3	7	1.07	0.42
Les	6	2	8	0.76	0.27
Sam	1	0	1	1.00	0

(a) State how you can deduce that the mass of each fish caught by Sam was 1.00 kg.

(b) The winner was the person who had caught the greatest total mass of fish by 4 p.m. Determine who was the winner, showing your working.

(c) Before leaving the waterside, Sam catches one more fish and weighs it. He then announces that if this extra fish is included with the other two fish he caught, the standard deviation is 1.00 kg. Find the mass of this extra fish. (OCR)

6 The heights of 94 policemen based at a city police station were measured, and the results (in metres) are summarised in the following table.

Height (m)	1.65–1.69	1.70–1.74	1.75–1.79	1.80–1.84	1.85–1.89
Frequency	2	4	11	23	38

Height (m)	1.90–1.94	1.95–1.99	2.00–2.04	2.05–2.09
Frequency	9	4	2	1

(a) Draw a cumulative frequency diagram and estimate the median and quartiles.

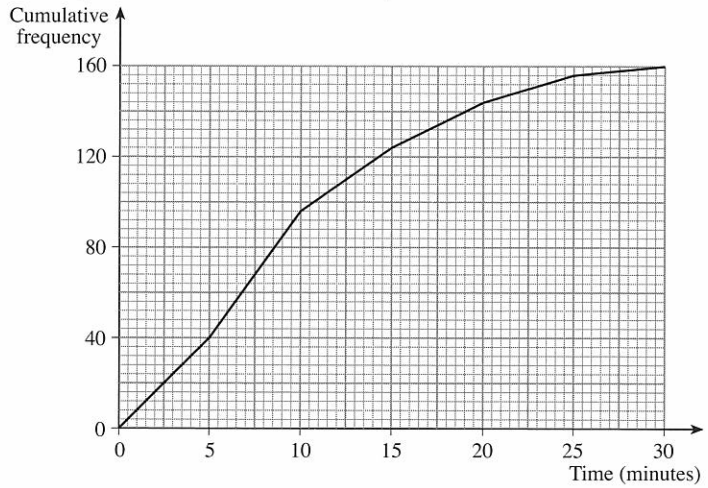
(b) What do the values found in part (a) indicate about the shape of the distribution?

(c) Estimate the mean and standard deviation of the heights.

(d) A possible measure of skewness is $\dfrac{3(\bar{x} - Q_2)}{\text{standard deviation}}$ with the usual notation.

Calculate this number and state how it confirms your answer to part (b).

7 The diagram shows a cumulative frequency curve for the lengths of telephone calls from a house during the first six months of last year.



(a) Find the median and interquartile range.

(b) Construct a histogram with six equal intervals to illustrate the data.

(c) Use the frequency distribution associated with your histogram to estimate the mean length of call.

(d) State whether each of the following is true or false.

(i) The distribution of these call times is skewed.

(ii) The majority of the calls last longer than 6 minutes.

(iii) The majority of the calls last between 5 and 10 minutes.

(iv) The majority of the calls are shorter than the mean length.

8 The following table gives the ages in completed years of the 141 persons in a town involved in road accidents during a particular year.

Age (years)	12–15	16–20	21–25	26–30	31–40	41–50	51–70
Frequency	15	48	28	17	14	7	12

Working in years, and giving your answers to 1 decimal place, calculate estimates of

(a) the mean and standard deviation of the ages,

(b) the median age.

Which do you consider to be the better representative average of the distribution, the mean or the median? Give a reason for your answer.

4 Probability

This chapter is about calculating with probabilities. When you have completed it, you should

- know what a 'sample space' is
- know the difference between an 'outcome' and an 'event' and be able to calculate the probability of an event from the probabilities of the outcomes in the sample space
- be able to use the addition law for mutually exclusive events
- know the multiplication law of conditional probability, and be able to use tree diagrams
- know the multiplication law for independent events.

4.1 Assigning probability

In many situations you may be unsure of the outcome of some activity or experiment, although you know what the possible outcomes are. For example, you do not know what number you will get when you roll a dice, but you do know that you will get $1, 2, 3, 4, 5$ or 6. You know that if you toss a coin twice, then the possible outcomes are (H,H), (H,T), (T,H) and (T,T). If you are testing a transistor to see if it is defective, then you know that the possible outcomes are 'defective' and 'not defective'.

The list of all the possible outcomes is called the **sample space** of the experiment. The list is usually written in curly brackets, $\{\ \}$.

Thus the sample space for rolling the dice is $\{1,2,3,4,5,6\}$, the sample space for tossing a coin twice is $\{(H,H),(H,T),(T,H),(T,T)\}$, and the sample space for testing a transistor is $\{\text{defective, not defective}\}$.

It is conventional when writing pairs of things like H,H to put them in brackets, like coordinates.

Each of the outcomes of an experiment has a probability assigned to it. Sometimes you can assign the probability using symmetry. For example, the sample space for throwing a dice is $\{1,2,3,4,5,6\}$, and you would assign each outcome the probability $\frac{1}{6}$, in the belief that the dice is fair, and that each outcome is equally likely. This is the usual method for calculations about games of chance.

Now suppose that the dice is not fair, so that you cannot use the method of symmetry for assigning probabilities. In this case you will have to carry out an experiment and throw the dice a large number of times. Suppose that you throw the dice 1000 times and the frequencies of the six possible outcomes in the sample space are as in Table 4.1.

Outcome	1	2	3	4	5	6
Frequency	100	216	182	135	170	197

Table 4.1. Frequencies for the outcomes when rolling a dice 1000 times.

You would then assign the probabilities $\frac{100}{1000}$, $\frac{216}{1000}$, $\frac{182}{1000}$, $\frac{135}{1000}$, $\frac{170}{1000}$ and $\frac{197}{1000}$ to the outcomes 1, 2, 3, 4, 5 and 6 respectively. These are called the **relative frequencies** of the outcomes, and you can use them as estimates of the probabilities. You should realise that if you were to roll the dice another 1000 times, the results would probably not be exactly the same, but you would hope that they would not be too different. You could roll the dice more times and hope to improve the relative frequency as an approximation to the probability.

Sometimes you cannot assign a probability by using symmetry or by carrying out an experiment. For example, there is a probability that my house will be struck by lightning next year, and I could insure against this happening. The insurance company will have to have a probability in mind when it calculates the premium I have to pay, but it cannot calculate it by symmetry, or carry out an experiment for a few years. It will assign its probability using its experience of such matters and its records.

When **probabilities** are assigned to the outcomes in a sample space,

- each probability must lie between 0 and 1 inclusive, and
- the sum of all the probabilities assigned must be equal to 1.

Example 4.1.1

How would you assign probabilities to the following experiments or activities?

(a) Choosing a card from a standard pack of playing cards.

(b) The combined experiment of tossing a coin and rolling a dice.

(c) Tossing a drawing pin on to a table to see whether it lands point down or point up.

(d) Four international football teams, Argentina, Cameroon, Nigeria, Turkey (A, C, N and T), play a knockout tournament. Who will be the winner?

(a) The sample space would consist of the list of the 52 playing cards $\{AC, 2C, 3C, \ldots, KS\}$ in some order. (Here A means ace, C means clubs, and so on.) Assuming that these cards are equally likely to be picked, the probability assigned to each of them is $\frac{1}{52}$.

(b) The sample space is $\begin{Bmatrix}(H,1),(H,2),(H,3),(H,4),(H,5),(H,6),\\(T,1),(T,2),(T,3),(T,4),(T,5),(T,6)\end{Bmatrix}$, and each of the outcomes would be assigned a probability of $\frac{1}{12}$.

(c) The sample space is {point down, point up}. You would need to carry out an experiment to assign probabilities.

(d) The sample space is {A wins, C wins, N wins, T wins}. You have to assign probabilities subjectively, according to your knowledge of the teams and the game. The probabilities p_A, p_C, p_N and p_T must all be non-negative and satisfy $p_A + p_C + p_N + p_T = 1$.

4.2 Probabilities of events

Sometimes you may be interested, not in one particular outcome, but in two or three or more of them. For example, suppose you toss a coin twice. You might be interested in whether the result is the same both times. The list of outcomes in which you are interested is called an **event**, and is written in curly brackets. The event that both tosses of the coin give the same result is $\{(H,H),(T,T)\}$. Events are often denoted by capital letters. Thus if A denotes this event, then $A = \{(H,H),(T,T)\}$. An event can be just one outcome, or a list of outcomes or even no outcomes at all.

You can find the probability of an event by looking at the sample space and adding the probabilities of the outcomes which make up the event. For example, if you were tossing a coin twice, the sample space would be $\{(H,H),(H,T),(T,H),(T,T)\}$. There are four outcomes, each equally likely, so they each have probability $\frac{1}{4}$. The event A consists of the two outcomes (H,H) and (T,T), so the probability of A is $\frac{1}{4}+\frac{1}{4}$, or $\frac{1}{2}$.

This is an example of a general rule.

> The probability, $\mathrm{P}(A)$, of an event A is the sum of the probabilities of the outcomes which make up A.

Often a list of outcomes can be constructed in such a way that all of them are equally likely. If all the outcomes are equally likely, then the probability of any event A can be found by finding the number of outcomes which make up event A and dividing by the total number of outcomes. When the outcomes are not equally likely, then the probability of any event has to be found by adding the individual probabilities of all the outcomes which make up event A.

Example 4.2.1

A fair 20-sided dice has eight faces coloured red, ten coloured blue and two coloured green. The dice is rolled.

(a) Find the probability that the bottom face is red.

(b) Let A be the event that the bottom face is not red. Find the probability of A.

(a) Each face has an equal probability of being the bottom face: as there are 20 faces, each of them has a probability of $\frac{1}{20}$. There are eight red faces, each with probability $\frac{1}{20}$, so $\mathrm{P}(\text{red}) = \frac{8}{20} = \frac{2}{5}$.

(b) $\mathrm{P}(A) = \mathrm{P}(\text{blue or green}) = \mathrm{P}(\text{blue}) + \mathrm{P}(\text{green}) = \frac{10}{20} + \frac{2}{20} = \frac{12}{20} = \frac{3}{5}$.

Example 4.2.2

The numbers $1, 2, \ldots, 9$ are written on separate cards. The cards are shuffled and the top one is turned over. Calculate the probability that the number on this card is prime.

The sample space for this activity is $\{1,2,3,4,5,6,7,8,9\}$. As each outcome is equally likely each has probability $\frac{1}{9}$.

Let B be the event that the card turned over is prime. Then $B = \{2,3,5,7\}$.

$\mathrm{P}(B)$ = sum of the probabilities of the outcomes in $B = \frac{4}{9}$.

Example 4.2.3

A circular wheel is divided into three equal sectors, numbered 1, 2 and 3, as shown in Fig. 4.2. The wheel is spun twice. Each time, the score is the number to which the black arrow points. Calculate the probabilities of the following events:

(a) both scores are the same as each other,
(b) neither score is a 2,
(c) at least one of the scores is a 3,
(d) neither score is a 2 and both scores are the same,
(e) neither score is a 2 or both scores are the same.

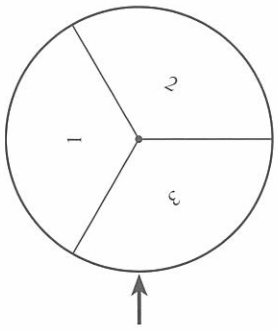


Fig. 4.2

Start by writing down the sample space.

$$\{(1,1),(1,2),(1,3),(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$$

Each outcome has probability $\frac{1}{9}$.

(a) Let A be the event that both scores are the same, so $A = \{(1,1),(2,2),(3,3)\}$.

$\mathrm{P}(A)$ = sum of the probabilities of the outcomes in $A = \frac{3}{9} = \frac{1}{3}$.

(b) Let B be the event that neither score is a 2, so $B = \{(1,1),(1,3),(3,1),(3,3)\}$.

$\mathrm{P}(B)$ = sum of the probabilities of the outcomes in $B = \frac{4}{9}$.

(c) Let C be the event that at least one of the scores is a 3, so $C = \{(1,3),(2,3),(3,1),(3,2),(3,3)\}$. Then $\mathrm{P}(C) = \frac{5}{9}$.

(d) Let D be the event that neither score is a 2 *and* both scores are the same, so $D = \{(1,1),(3,3)\}$. Then $\mathrm{P}(D) = \frac{2}{9}$.

(e) Let E be the event that neither score is a 2 *or* both scores are the same, so $E = \{(1,1),(1,3),(3,1),(3,3),(2,2)\}$. Then $\mathrm{P}(E) = \frac{5}{9}$.

Example 4.2.4

Jafar has three playing cards, two queens and a king. Tandi selects one of the cards at random, and returns it to Jafar, who shuffles the cards. Tandi then selects a second card. Tandi wins if both cards selected are kings. Find the probability that Tandi wins.

Imagine that the queens are different, and call them Q_1 and Q_2, and call the king K. Then the sample space is:

$$\{(Q_1,Q_1),(Q_1,Q_2),(Q_1,K),(Q_2,Q_1),(Q_2,Q_2),(Q_2,K),(K,Q_1),(K,Q_2),(K,K)\}.$$

Each outcome has probability $\frac{1}{9}$.

Let T be the event that Tandi wins. Then $T = \{(K,K)\}$ and

$\mathrm{P}(T)$ = sum of the probabilities of the outcomes in $T = \frac{1}{9}$.

The probability that Tandi wins a prize is $\frac{1}{9}$.

Although the event that Tandi wins is just a single outcome, it is still listed in curly brackets.

Example 4.2.5
A dice with six faces has been made from brass and aluminium, and is not fair. The probability of a 6 is $\frac{1}{4}$, the probabilities of 2, 3, 4, and 5 are each $\frac{1}{6}$, and the probability of 1 is $\frac{1}{12}$. The dice is rolled. Find the probability of rolling (a) 1 or 6, (b) an even number.

(a) $\mathrm{P}(1 \text{ or } 6) = \mathrm{P}(1) + \mathrm{P}(6) = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$.

(b) $\mathrm{P}(\text{an even number}) = \mathrm{P}(2 \text{ or } 4 \text{ or } 6) = \mathrm{P}(2) + \mathrm{P}(4) + \mathrm{P}(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{4} = \frac{7}{12}$.

Sometimes it is worth using a different approach to calculating the probability of an event.

Example 4.2.6
You draw two cards from an ordinary pack. Find the probability that they are not both kings.

The problem is that the sample space has a large number of outcomes. In fact there are 52 ways of picking the first card, and then 51 ways of picking the second, so there $52 \times 51 = 2652$ possibilities. The sample space therefore consists of 2652 outcomes, each of which is assigned a probability $\frac{1}{2652}$.

To avoid counting all the outcomes which are not both kings, it is easier to look at the number of outcomes which *are* both kings.

Writing the first card to be drawn as the first of the pair, these outcomes are (KC, KD), (KD, KC), (KC, KH), (KH, KC), (KC, KS), (KS, KC), (KD, KH), (KH, KD), (KD, KS), (KS, KD), (KH, KS) and (KS, KH).

There are thus 12 outcomes that are both kings. So the number which are not both kings is $2652 - 12 = 2640$. All 2640 of these outcomes have probability $\frac{1}{2652}$, so $\mathrm{P}(\text{not both kings}) = \frac{2640}{2652} = \frac{220}{221}$.

It is always worth watching for this short cut, and it is also useful to have some language to describe it. If A is an event, the event 'not A' is the event consisting of those outcomes in the sample space which are not in A. Since the sum of the probabilities assigned to outcomes in the sample space is 1,

$$\mathrm{P}(A) + \mathrm{P}(\text{not } A) = 1.$$

The event 'not A' is called the **complement** of the event A. The symbol A' is used to denote the complement of A.

If A is an event, then A' is the complement of A, and

$$\mathrm{P}(A) + \mathrm{P}(A') = 1. \qquad (4.1)$$

4.3 Addition of probabilities

Consider a game in which a fair cubical dice with faces numbered 1 to 6 is rolled twice. A prize is won if the total score on the two rolls is 4 or if each individual score is over 4.

You can write the sample space of all possible outcomes as 36 equally likely pairs,

$$\left\{\begin{matrix}(1,1) & (1,2) & (1,3) & (1,4) & (1,5) & (1,6)\\ (2,1) & (2,2) & & \dots & & (2,6)\\ \vdots & \vdots & & & & \vdots\\ (6,1) & (6,2) & & \dots & & (6,6)\end{matrix}\right\},$$

each having probability $\frac{1}{36}$.

Let A be the event that the total score is 4 and let B be the event that each roll of the dice gives a score over 4.

Then $A = \{(1,3),(2,2),(3,1)\}$, and $B = \{(5,5),(5,6),(6,5),(6,6)\}$, so

$$\mathrm{P}(A) = \tfrac{3}{36} = \tfrac{1}{12} \quad \text{and} \quad \mathrm{P}(B) = \tfrac{4}{36} = \tfrac{1}{9}.$$

A prize is won if A happens *or* if B happens, so $\mathrm{P}(\text{a prize is won}) = \mathrm{P}(A \text{ or } B)$.

This means that a prize will be won if any of the outcomes in $\{(1,3),(2,2),(3,1),(5,5),(5,6),(6,5),(6,6)\}$ occurs. Therefore

$$\mathrm{P}(\text{a prize is won}) = \mathrm{P}(A \text{ or } B) = \tfrac{7}{36}.$$

The key point is that $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$. The word 'or' is important. Whenever you see it, it should suggest to you the idea of adding probabilities. Notice, however, that A and B have no outcomes which are common to both events. Two events which have no outcomes common to both are called **mutually exclusive** events. So the result $\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$ is known as the **addition law of mutually exclusive events**.

The addition law of mutually exclusive events can be extended to apply to more than two events if none of the events has any outcome in common with any of the other events.

If $A_1, A_2, \dots, A_n$ are n mutually exclusive events, then

$$\mathrm{P}(A_1 \text{ or } A_2 \text{ or } \dots \text{ or } A_n) = \mathrm{P}(A_1) + \mathrm{P}(A_2) + \dots + \mathrm{P}(A_n). \qquad (4.2)$$

The addition law needs to be modified when the events are not mutually exclusive. Here is an example.

Example 4.3.1

Two fair dice are thrown. A prize is won if the total is 10 or if each individual score is over 4. Find the probability that a prize is won.

The sample space is the same set of 36 pairs listed at the top of this page.

Let C be the event that the total score is 10, so $C = \{(5,5),(4,6),(6,4)\}$.

Let B be the event that each roll of the dice results in a score over 4, as before, so $B = \{(5,5),(5,6),(6,5),(6,6)\}$.

Therefore $\mathrm{P}(C) = \frac{3}{36} = \frac{1}{12}$ and $\mathrm{P}(B) = \frac{4}{36} = \frac{1}{9}$.

A prize is won if B or C occurs, and the possible outcomes which make up this event are $\{(5,5),(4,6),(6,4),(5,6),(6,5),(6,6)\}$.

Therefore $\mathrm{P}(B \text{ or } C) = \frac{6}{36} = \frac{1}{6}$. But $\mathrm{P}(B) + \mathrm{P}(C) = \frac{1}{12} + \frac{1}{9} = \frac{7}{36}$, so in this case $\mathrm{P}(B \text{ or } C) \neq \mathrm{P}(B) + \mathrm{P}(C)$.

For events such as B and C, which are not mutually exclusive, the addition rule given by Equation (4.2) is not valid.

The rule can be modified so that it applies to any two events. This will be studied later in the course. Can you see how to modify the rule?

Exercise 4A

1 A fair dice is thrown once. Find the probabilities that the score is

(a) bigger than 3,
(b) bigger than or equal to 3,
(c) an odd number,
(d) a prime number,
(e) bigger than 3 and a prime number,
(f) bigger than 3 or a prime number or both,
(g) bigger than 3 or a prime number, but not both.

2 A card is chosen at random from an ordinary pack. Find the probability that it is

(a) red,
(b) a picture card (K, Q, J),
(c) an honour ($A, K, Q, J, 10$),
(d) a red honour,
(e) red, or an honour, or both.

3 Two fair dice are thrown simultaneously. Find the probability that

(a) the total is 7,
(b) the total is at least 8,
(c) the total is a prime number,
(d) neither of the scores is a 6,
(e) at least one of the scores is a 6,
(f) exactly one of the scores is a 6,
(g) the two scores are the same,
(h) the difference between the scores is an odd number.

4 A fair dice is thrown twice. If the second score is the same as the first, the second throw does not count, and the dice is thrown again until a different score is obtained. The two different scores are added to give a total.

List the possible outcomes.

Find the probability that

(a) the total is 7,
(b) the total is at least 8,
(c) at least one of the two scores is a 6,
(d) the first score is higher than the last.

5 A bag contains ten counters, of which six are red and four are green. A counter is chosen at random; its colour is noted and it is replaced in the bag. A second counter is then chosen at random. Find the probabilities that

(a) both counters are red,

(b) both counters are green,

(c) just one counter is red,

(d) at least one counter is red,

(e) the second counter is red.

6 Draw a bar chart to illustrate the probabilities of the various total scores when two fair dice are thrown simultaneously.

4.4 Conditional probability

Consider a class of 30 pupils, of whom 17 are girls and 13 are boys. Suppose further that five of the girls and six of the boys are left-handed, and all of the remaining pupils are right-handed. If a pupil is selected at random from the whole class, then the chance that he or she is left-handed is $\frac{6+5}{30} = \frac{11}{30}$. However, suppose now that a pupil is selected at random from the girls in the class. The chance that this girl will be left-handed is $\frac{5}{17}$. So being told that the selected pupil is a girl alters the chance that the pupil will be left-handed. This is an example of **conditional probability**. The probability has been calculated on the basis of an extra 'condition' which you have been given.

There is some notation which is used for conditional probability. Let L be the event that a left-handed person is chosen, and let G be the event that a girl is chosen. The symbol $\mathrm{P}(L \mid G)$ stands for the probability that the pupil chosen is left-handed *given* that the pupil chosen is a girl. So in this case $\mathrm{P}(L \mid G) = \frac{5}{17}$, although $\mathrm{P}(L) = \frac{11}{30}$.

It is useful to find a connection between conditional probabilities (where some extra information is known) and probabilities where you have no extra information. Notice that the probability $\mathrm{P}(L \mid G)$ can be written as

$$\mathrm{P}(L \mid G) = \frac{5}{17} = \frac{5/30}{17/30}.$$

The fraction in the numerator is the probability of choosing a left-handed girl if you are selecting from the whole class, and the denominator is the probability of choosing a girl if you are selecting from the whole class. In symbols this could be written as

$$\mathrm{P}(L \mid G) = \frac{\mathrm{P}(L \text{ and } G)}{\mathrm{P}(G)}.$$

This equation can be generalised to any two events A and B for which $\mathrm{P}(A) > 0$.

If A and B are two events and $P(A) > 0$, then the **conditional probability** of B given A is

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}. \qquad (4.3)$$

Rewriting this equation gives

$$P(A \text{ and } B) = P(A) \times P(B \mid A), \qquad (4.4)$$

which is known as the **multiplication law of probability**.

Suppose a jar contains seven red discs and four white discs. Two discs are selected without replacement. ('Without replacement' means that the first disc is not put back in the jar before the second disc is selected.) Let R_1 be the event {the first disc is red}, let R_2 be the event {the second disc is red}, let W_1 be the event {the first disc is white} and let W_2 be the event {the second disc is white}. To find the probability that both of the discs are red you want to find $P(R_1 \text{ and } R_2)$.

Using the multiplication law, Equation 4.4, to find this probability,

$$P(R_1 \text{ and } R_2) = P(R_1) \times P(R_2 \mid R_1).$$

Now $P(R_1) = \frac{7}{11}$, since there are 7 red discs in the jar and 11 discs altogether. The probability $P(R_2 \mid R_1)$ appears more complicated, but it represents the probability that the second disc selected is red *given* that the first disc was red. To find this just imagine that one red disc has already been removed from the jar. The jar now contains 6 red discs and 4 white discs. The probability *now* of getting a red disc is $P(R_2 \mid R_1) = \frac{6}{10}$.

Therefore, using the multiplication law,

$$P(R_1 \text{ and } R_2) = P(R_1) \times P(R_2 \mid R_1) = \frac{7}{11} \times \frac{6}{10}.$$

You can represent all the possible outcomes when two discs are selected from the jar in a **tree diagram**, as in Fig. 4.3.

Notice that probabilities on the first 'layer' of branches give the chances of getting a red disc or a white disc when the first disc is selected. The probabilities on the second 'layer' are the conditional probabilities. You can use the tree diagram to calculate the probability of any of the four possibilities, R_1 and R_2, R_1 and W_2, W_1 and R_2 and W_1 and W_2.

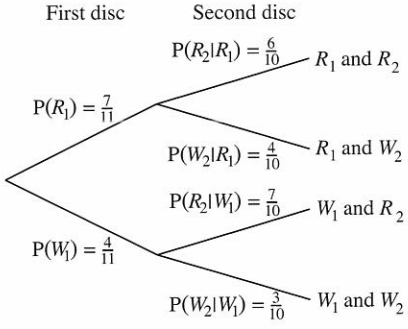


Fig. 4.3. Tree diagram to show the outcomes when two discs are drawn without replacement from a jar.

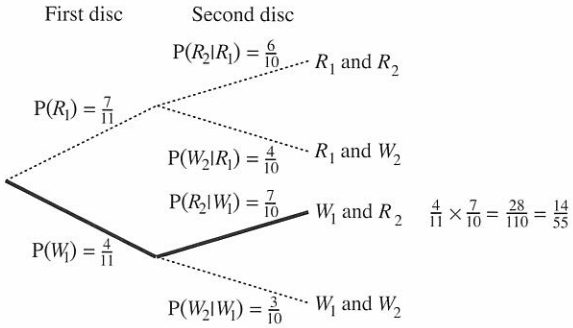


Fig. 4.4. Tree diagram to show the calculation of $P(W_1 \text{ and } R_2)$.

To do this you move along the appropriate route, multiplying the probabilities, as shown in Fig. 4.4. For example, to find the probability of getting a white disc followed by a red disc, $P(W_1 \text{ and } R_2)$, trace that route on the tree diagram and multiply the relevant probabilities.

You could also have found $P(W_1 \text{ and } R_2)$ by using the multiplication law.

$$P(W_1 \text{ and } R_2) = P(W_1) \times P(R_2 \mid W_1) = \frac{4}{11} \times \frac{7}{10} = \frac{14}{55}.$$

You can now use the addition and multiplication laws together to find the probability of more complex events. For example,

$$P(\text{both discs are the same colour}) = P((R_1 \text{ and } R_2) \text{ or } (W_1 \text{ and } W_2)).$$

The event R_1 and R_2 is the event that both discs are red, and the event W_1 and W_2 is the event that both discs are white. These events cannot both be satisfied at the same time, so they must be mutually exclusive. Therefore you can use the addition law, giving

$$\begin{aligned} P((R_1 \text{ and } R_2) \text{ or } (W_1 \text{ and } W_2)) &= P(R_1 \text{ and } R_2) + P(W_1 \text{ and } W_2) \\ &= P(R_1) \times P(R_2 \mid R_1) + P(W_1) \times P(W_2 \mid W_1) \\ &= \frac{7}{11} \times \frac{6}{10} + \frac{4}{11} \times \frac{3}{10} = \frac{42}{110} + \frac{12}{110} = \frac{54}{110} = \frac{27}{55}. \end{aligned}$$

You can also use the tree diagram for this calculation. This time there is more than one route through the tree diagram which satisfies the event whose probability is to be found. As before, you follow the appropriate routes and multiply the probabilities. You then add all the resulting products, as in Fig. 4.5.

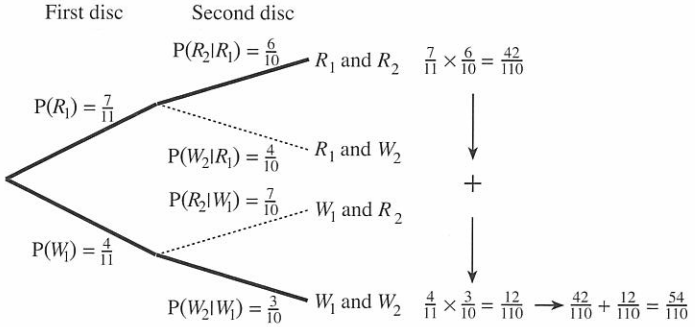


Fig. 4.5. Tree diagram to show the calculation of $P((R_1 \text{ and } R_2) \text{ or } (W_1 \text{ and } W_2))$.