2 Frequency distributions

2.1 The classification of raw data

As was mentioned in section 1.3, the raw data from an investigation usually require classification before patterns can readily be observed in them. Let us look again at the sets of scores obtained by the two groups of students in the hypothetical language teaching experiment discussed briefly in section 1.3, designed to test the effectiveness of the language laboratory as compared with more traditional methods. The data are repeated in table 2.1.

We notice that there seem to be more single-figure marks in the group B column than in the group A column, and that the range of marks for group B is larger than that for group A (5–18 as against 9–19). Otherwise, however, little pattern can be seen at a glance. Some of the scores in each column occur more than once. If we now count the frequency with which each score occurs in a given column, we obtain a *frequency distribution* for each set of scores, as shown in table 2.2. The picture is now clearer: we see that 15 is the most frequent score for group A, the frequencies falling away on either side of this score. The most frequent score for group B is 12, the frequencies again tailing away on either side. The frequency distributions also show clearly the difference in variability between the two sets of scores: the marks for group B are more spread than those for group A. Thus the frequency distribution will give us a rough idea about the *central tendency* of the scores and about their *variability*. Precise measures of these properties will be discussed in chapter 3.

2.2 Grouped data

A distribution giving a frequency for each individual value taken by the variable, as above, works well where there is a small number

Table 2.1 Scores in a language test for two groups taught by different methods

Marks out of 20	
Group A (Language laboratory) (N = 30)	*Group B (Traditional) (N = 30)*
15	11
12	16
11	14
18	18
15	6
15	8
9	9
19	14
14	12
13	12
11	10
12	15
18	12
15	9
16	13
14	16
16	17
17	12
15	8
17	7
13	15
14	5
13	14
15	13
17	13
19	12
17	11
18	13
16	11
14	7

of values (for instance, there are only 15 actual values of the variable in the language test data). Let us now consider what happens if the variable can take a wider range of values. The data in table 2.3 represent the frequency of sentences of particular lengths (in numbers of words) in the first 100 sentences of Iris Murdoch's *The Bell* (Penguin edition, 1962).

Table 2.2 Frequency distributions for scores on language test

Score	*Group A*		*Group B*	
5			/	1
6			/	1
7			//	2
8			//	2
9	/	1	//	2
10			/	1
11	//	2	///	3
12	//	2	~~///~~	5
13	///	3	////	4
14	////	4	///	3
15	~~///~~ /	6	//	2
16	///	3	//	2
17	////	4	/	1
18	///	3	/	1
19	//	2		

Such a distribution is not, by itself, particularly useful, because there are large numbers of values taken by the variable (sentence length), many with very low frequencies. A clearer picture emerges if the sentence length values are grouped in the manner shown in table 2.4. Here, the data have been reclassified so that the total frequencies within the *class intervals* 1–5, 6–10, 11–15 and so on are recorded. Although we obtain a clearer idea of the distribution by grouping in this way, we also lose some of the original information. We know how many sentences have lengths in the range 1–5 words, but we no longer know, from the grouped data, what proportion of these have lengths of 1, 2, 3, 4 and 5 words. For the purpose of later statistical calculations, one of two assumptions can be made: either that the frequencies are evenly spread over the class interval (for example, 3.6 sentences of each of the lengths 6, 7, 8, 9 and 10 for the Iris Murdoch data); or that the total frequency within the class interval is concentrated at its mid-point (3, 8, 13, and so on, for the sentence length data). Which assumption we make depends on just what we want to do with the data, as we shall see later.

Table 2.3 Sentence length (words) distribution for the first 100 sentences of Iris Murdoch's *The Bell* (with hyphenated items treated as single words)

Sentence length (no. words)	*Frequency*	*Sentence length (no. words)*	*Frequency*
3	1	23	4
4		24	2
5	1	25	4
6	2	26	
7	2	27	1
8	8	28	3
9	3	29	3
10	3	30	2
11	5	31	1
12	3	32	2
13	3	33	2
14	8	34	1
15	7	35	
16	3	36	2
17	4	37	1
18	1	38	1
19	6	39	
20	4	40	1
21	2	41	
22	3	42	1

Table 2.4 Grouped data for sentence length distribution

Sentence length (no. words)	*Frequency*
1- 5	2
6-10	18
11-15	26
16-20	18
21-25	15
26-30	9
31-35	6
36-40	5
41-45	1

2.3 Histograms

An even clearer idea of a frequency distribution can be obtained by converting it to a *histogram*. For data which are not grouped, we simply arrange the values taken by the variable on the horizontal axis, and frequency values on the vertical axis, and then draw a box or bar over each value taken by the variable, at a height corresponding to the frequency found for that value. The data for our hypothetical language teaching experiment are presented as histograms in figures 2.1 and 2.2.

If we are dealing with grouped data, the width of a box in the histogram corresponds to the class interval, as in figure 2.3 which shows the sentence length distribution of the data from *The Bell*. The horizontal axis is labelled with the mid-points of the class intervals.

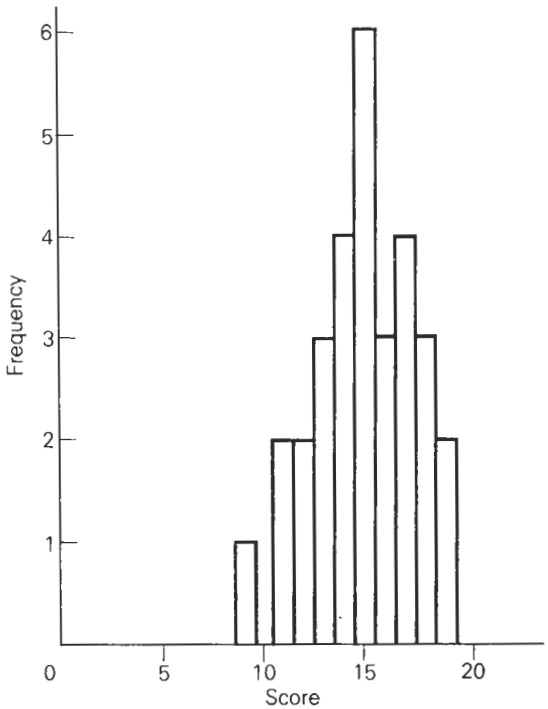

Figure 2.1 Language test scores: group A

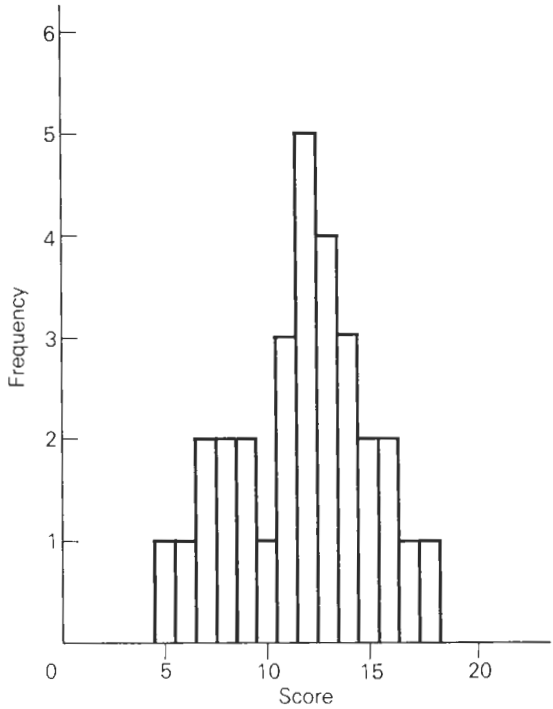

Figure 2.2 **Language test scores: group B**

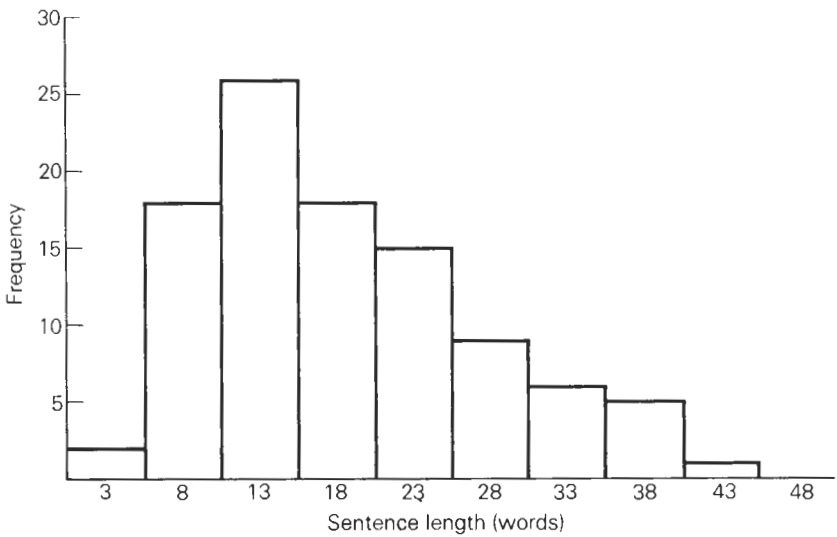

Figure 2.3 **Sentence length distribution for the first 100 sentences of Murdoch's *The Bell***

It is extremely important that histograms and other graphical representations of frequency distributions should be clearly labelled: they should have a title, and the relevant variables should be specified along each axis, together with the unit of measurement where appropriate.

2.4 Frequency polygons

An alternative way of presenting distributions graphically is to draw a *frequency polygon*. Instead of a box, we draw a point over the value of a variable at a height corresponding to the frequency of that value. If the data are grouped, the point is placed over the mid-point of the class interval. The points are then joined by straight lines, as shown in figures 2.4–2.7. Note that the graph is normally taken to zero at the limits of the range of values, where it is sensible to do so. One advantage of frequency polygons is

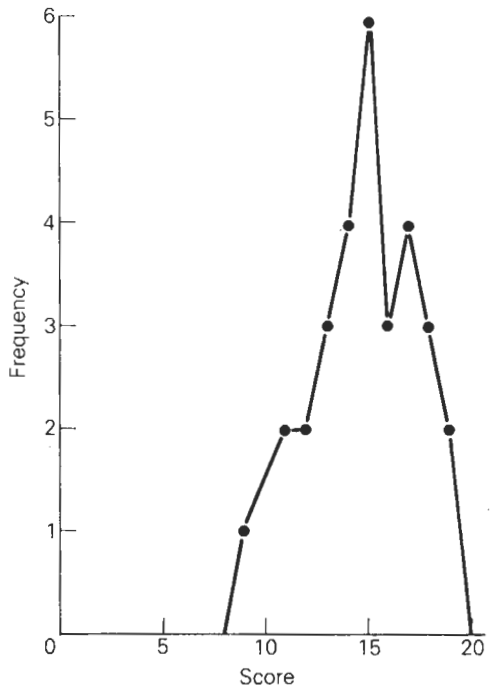

Figure 2.4 Language test scores: group A

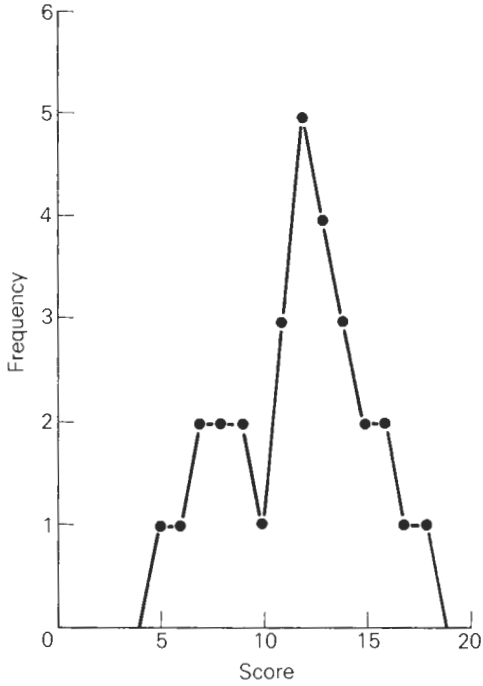

Figure 2.5 **Language test scores: group B**

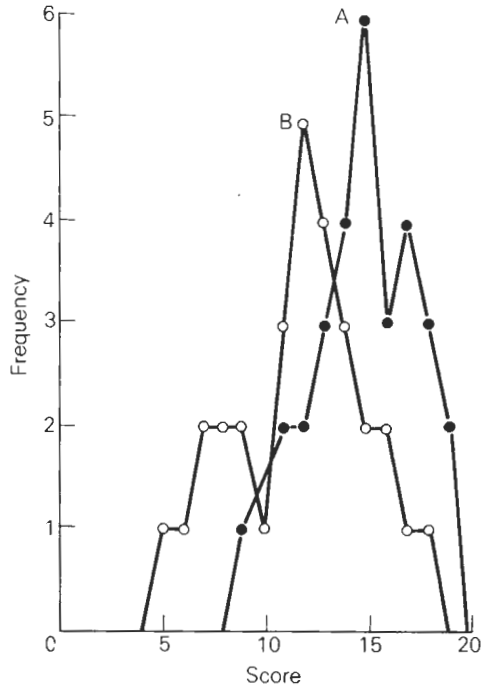

Figure 2.6 **Language test scores: groups A and B**

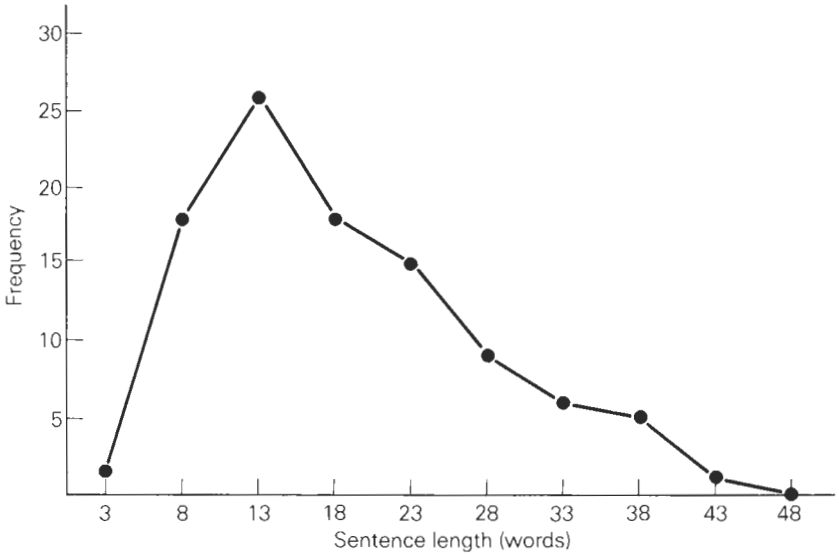

Figure 2.7 **Sentence length distribution for the first 100 sentences of Murdoch's *The Bell***

that they provide an excellent visual means of comparing two distributions, by plotting them on the same graph. This is illustrated by figure 2.6 in which the distributions for the two groups of language learners are superimposed.

2.5 The shapes of frequency distributions

Because they are made up of straight lines, and represent data from a relatively small number of observations, the frequency polygons in figures 2.4–2.7 are irregular. If, however, we were to draw polygons for much larger sets of data, we should find that the irregularities would smooth out, so that we could draw a smooth curve through the points. The shape of the curve is an important property of the distribution.

A particularly important kind of distribution, the so-called *normal distribution*, has a bell-shaped curve, symmetrical about its highest point, as shown in figure 2.8. We shall investigate the properties of the normal distribution in chapter 4. Meanwhile, it does not take too much imagination to see that the distributions given by our language test results approximate to the 'normal'

shape. If a distribution is lopsided rather than symmetrical, it is said to be *skewed*. If the high frequencies correspond to low values of the variable, as in the sentence length distribution in figures 2.3 and 2.7, the distribution is *positively skewed*; if the higher frequencies are at higher values, it is *negatively skewed* (see figure 2.9).

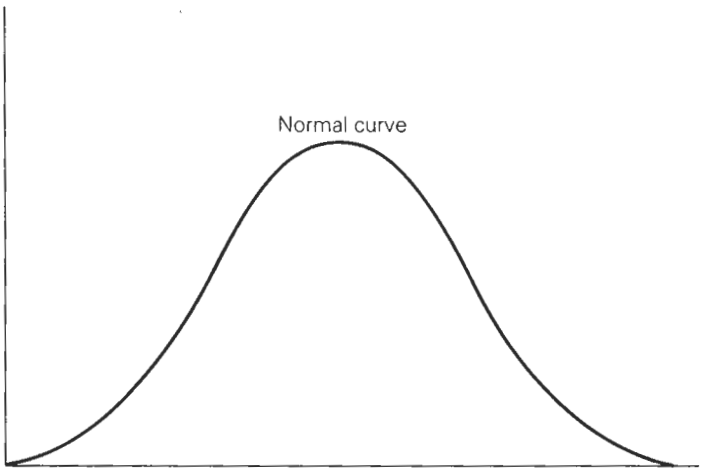

Figure 2.8 Normal distribution curve

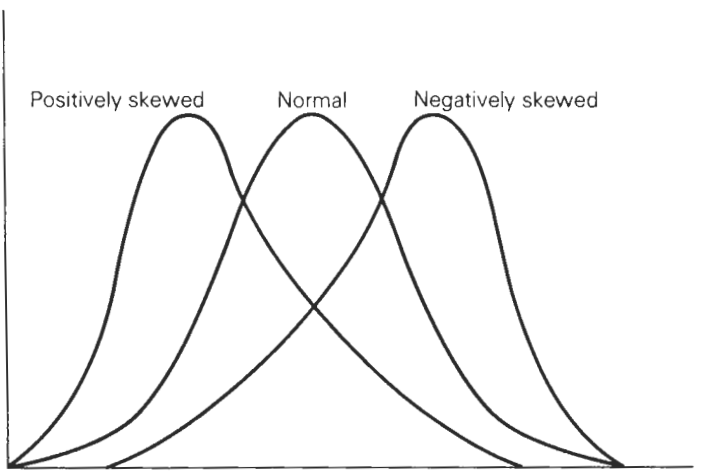

Figure 2.9 Skewed distributions

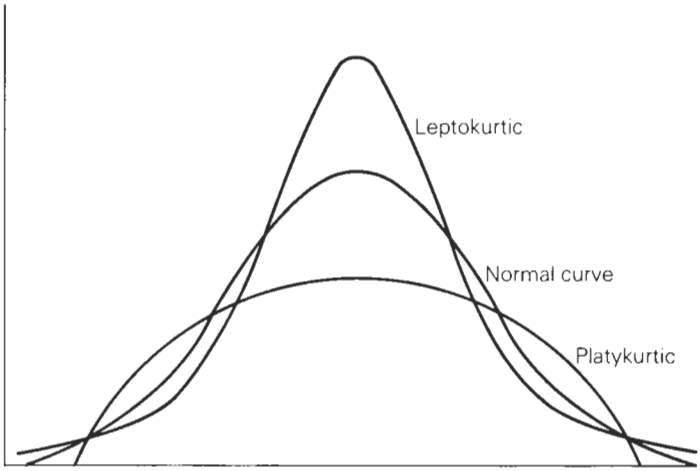

Figure 2.10 Kurtosis

A further property of distribution curves is their *kurtosis*. This refers to the degree of peaking: if a curve is more peaked than the normal distribution, it is said to be *leptokurtic*; if less peaked, it is *platykurtic* (see figure 2.10). Kurtosis is not as important as skewedness in later statistical work on a distribution, and we shall not discuss it further here.

2.6 Conclusion

The first stage in a statistical examination of data is to prepare a frequency distribution table, which can then be converted to a visual representation in the form of a histogram or frequency polygon. The latter have the advantage of greater clarity when comparing two or more superimposed distribution curves. This preliminary work gives the investigator some indication of the most typical value and the spread of data, and also shows the shape of the distribution he is dealing with, a factor of considerable importance in further statistical work.

Exercises

1 Take two texts from different varieties of written English and draw up a frequency distribution for the lengths of the first 200

words in each text, making clear your criteria for defining a word. Plot your distributions (i) as histograms, (ii) as frequency polygons. Comment on the shapes of the distributions, and on any differences you observe.

2 In a study by Crompton, the intensity of stressed and unstressed syllables (in decibels from an arbitrary norm) was measured in a sample of spoken French. The results for the first 100 syllables of each type were as follows:

Stressed

21	30	28	19	21	19	20	22	26	22
26	23	21	30	25	27	26	25	31	26
27	22	16	18	29	23	19	24	24	25
25	25	25	19	24	20	24	20	20	25
22	20	22	22	22	26	27	22	25	30
27	20	25	24	22	21	28	24	23	23
26	29	31	23	29	27	28	31	29	27
16	19	23	23	19	25	23	28	26	25
26	23	31	23	31	27	29	25	30	27
27	22	25	21	24	25	20	22	21	28

Unstressed

25	29	27	23	18	22	24	21	25	14
25	22	25	29	25	19	26	25	28	20
23	25	22	27	27	21	22	22	27	23
21	28	24	21	26	24	18	23	22	25
22	24	21	21	22	16	25	16	23	22
28	20	15	28	25	15	10	14	19	24
25	20	22	20	23	22	7	20	26	21
28	25	23	23	14	28	20	22	28	21
30	28	20	16	18	29	16	25	24	16
25	28	20	19	21	24	26	25	28	14

Group the data using an appropriate interval, and draw frequency polygons to compare the distribution of intensities for stressed and unstressed syllables. Comment on the results.

3 In the same study of French, the length of pause (in units of 1/50 sec) was measured for each tone group boundary which was not sentence-final. The results were as follows:

33	22	28	33	16	2	26	7	22	9
18	26	7	22	25	5	2	13	6	11
5	26	22	30	32	37	14	5	33	36
24	35	31	34	10	27	10	5	8	11
6	7	17	31	9	8	19	0	6	22
33	3	21	2	27	27	24	0	10	34
3	37	21	9	19	4	12	17	24	11
6	4	15	3	33	21	34	40	7	0
3	29	25	25	3	33	10	41	13	0
28	19	14	2	0	2	25	22	22	0
26	4	25	25	0	0	24	20	25	0
7	22	21	10	30	30	10	22	9	0
0	3	16	28	5	6	28	23	10	18
22	30	34	25	23	30	28	25	1	16
7	4	17	5	28	13	25	23	13	0

Group these data using an appropriate interval, and draw a histogram of the grouped frequency distribution. Comment on the results.