4 The normal distribution

4.1 The importance of the normal distribution

We saw in section 2.5 that, if we took a very large population (or indeed very large samples from a population) and drew a frequency polygon to represent the distribution, we should be able to construct a smooth curve through the points. We also mentioned that one particular kind of frequency distribution, the so-called 'normal' distribution, is of special importance in statistical work. There are several reasons for this. Firstly, many properties found in the natural world, and of interest to natural and social scientists, are distributed according to the normal curve. Secondly, the normal distribution has certain special mathematical properties which make it possible to predict what proportion of the population will have values of a normally distributed variable within a given range. Thirdly, some important tests for significant differences between sets of data assume that the variables concerned are normally distributed.

4.2 The properties of the normal curve

As we saw earlier, the normal curve is bell-shaped and symmetrical about its highest point, at which the mean, median and mode of the distribution all coincide, as shown in figure 4.1. Certain important properties of the curve are concerned with the way in which the area under it is cut up by lines drawn vertically from different points on the horizontal axis. In order to understand this, let us return for a moment to the histograms we discussed in section 2.3. The height of each box in a histogram is proportional to the frequency being represented. Since the width of each box is the

same, the area of the box is also proportional to the frequency of observations in the interval (which may be one unit or more: see the discussion of grouped frequency distributions in section 2.2) represented by that box. Now, a curve such as the normal curve is simply the limiting, smoothed-out shape we should get if we took a large number of observations and made our class intervals (that is, the width of our histogram boxes) very small. This is shown diagrammatically in figure 4.2. It follows that the area under the curve between two vertical lines drawn at particular values on the horizontal axis is proportional to the frequency of observations occurring between these two values of the variable. This would be true for any frequency curve, not just for the normal curve. However, the normal curve has special additional properties, which we shall now discuss.

The normal curve is entirely defined by just two properties of the distribution: the mean and the standard deviation. If we know these two values, we can construct the whole curve, by means of a rather complicated formula which will not be discussed here. A particularly important point is that a vertical line drawn at any given (whole or fractional) number of standard deviations from the mean will cut off a 'tail' containing a constant proportion of the total area under the curve, which can be read off from a statistical table such as that given as table A2 of appendix 1. A line

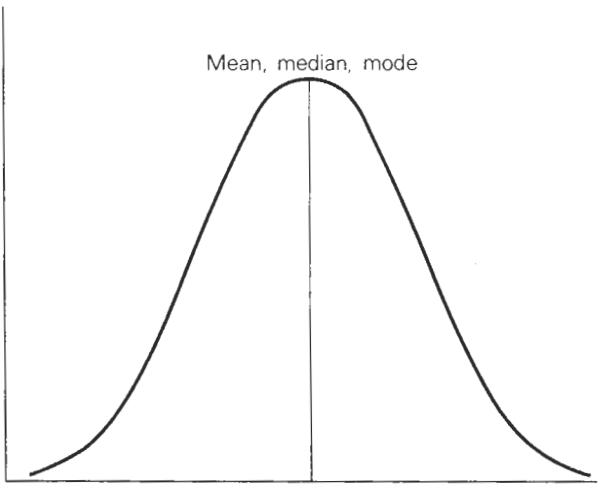

Figure 4.1 The normal distribution curve

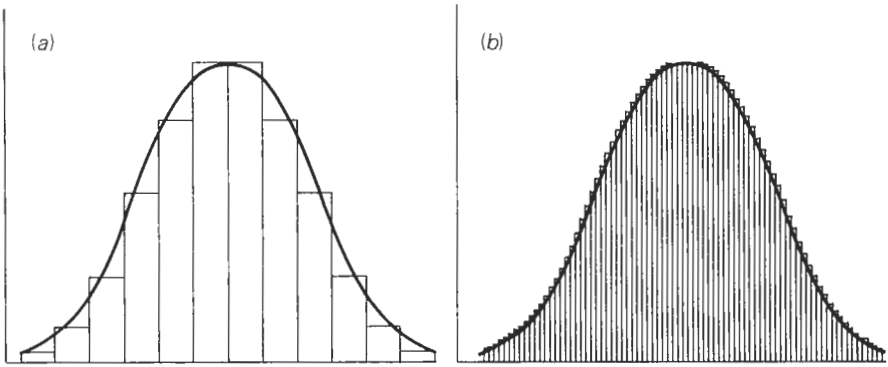

Figure 4.2 **Smoothing of the curve as interval width decreases: (a) wide intervals; (b) narrow intervals**

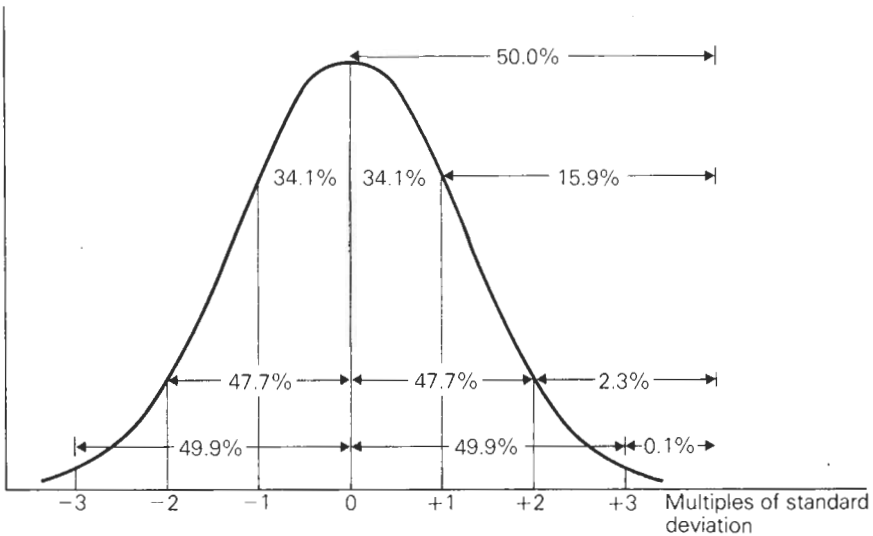

Figure 4.3 **Areas under the normal distribution curve**

drawn one standard deviation above the mean cuts off a tail containing 0.1587 of the total area under the curve; a line drawn at twice the value of the standard deviation cuts off 0.022 8 of the area; and a line drawn three standard deviations above the mean produces a tail with just 0.001 3 of the area. Since the curve is symmetrical, the same is true for lines drawn below the mean. These areas are shown as percentages of the total area in figure 4.3.

In order to use this property of the normal curve, since the area proportions are expressed in terms of standard deviation multiples away from the mean, we shall have to recast particular values of a variable in these terms. To do this, we calculate a 'standardised normal variable', otherwise known as a standard score or *z-score*:

$$z = \frac{x - \bar{x}}{\sigma}$$

or, for a sample,

$$z = \frac{x - \bar{x}}{s}$$

where

x is any particular value of the variable
$\bar{x}$ is the mean
σ or s is the standard deviation (of a population or sample respectively).

The z-score is thus simply the deviation of any given value from the mean, expressed as a multiple of the standard deviation.

Let us now consider how this z-score can be used. Suppose that we have measured the times taken by a very large number of people to utter a particular sentence, and have shown these times to be normally distributed (we shall see later how we can test for normality) with a mean of 3.45 sec and a standard deviation of 0.84 sec. Armed with this information, we can answer various questions. What proportion of the (potentially infinite) population of utterance times would be expected to fall below 3 sec? What proportion would lie between 3 and 4 sec? What is the time below which only 1 per cent of the times would be expected to fall? Let us try to answer these questions in order.

Using the formula for z above, we can calculate a z-score for a value of 3 sec:

$$z = \frac{3 - 3.45}{0.84} = -0.54.$$

That is, a value of 3 sec lies 0.54 standard deviations below the mean, the fact that it is *below* the mean being reflected in the

negative value of z. Looking up the table of areas under the normal curve for various values of z (table A2 of appendix 1), we find that a value of 0.54 cuts off a tail containing 0.294 6 of the total area. That is, the proportion of values expected to fall below 3 sec is 0.294 6, or 29.5 per cent.

Now to our second question: what proportion of the times would be expected to lie between 3 and 4 sec? If we can determine the proportion lying above 4 sec, then, since we already know what proportion lies below 3 sec, we can arrive at our answer by subtracting both of these from 100 per cent. The z-score for 4 sec is

$$z = \frac{4 - 3.45}{0.84} = 0.66.$$

From the table of areas, a value of 0.66 for z cuts off a tail containing 0.254 6 of the area under the curve. The proportion lying below 3 sec was 0.294 6, so the proportion falling between 3 and 4 sec is

$$100 - 29.46 - 25.46 = 45.1 \text{ per cent.}$$

Finally, what time value is such that only 1 per cent of the times should fall below it? To answer this question, we need to know what z-value cuts off a tail containing 0.01 of the area. The table of areas shows that a z-score of 2.33 means that 0.009 9 of the area is cut off. If x is the time value we require, then

$$-2.33 = \frac{x - 3.45}{0.84}.$$

Thus

$$x = (-2.33 \times 0.84) + 3.45 = 1.49 \text{ sec.}$$

Note carefully that we use a negative value for z because we are interested in what happens *below* the mean. Readers who are worried by this may find it useful to look at the situation in another way. A z-score of 2.33 (irrespective of sign) means that we are talking about a value that is 2.33 standard deviations away from the mean. In our case, this is equivalent to (2.33 × 0.84) or

1.96 sec. Since we are interested in values below the mean, we must subtract this from the mean value, giving a value of (3.45 − 1.96) or 1.49 sec.

We thus see that the special properties of the normal distribution allow us to make predictions about the proportions of observations lying above, below or between particular limits, defined in terms of multiples of the standard deviation away from the mean.

4.3 Testing for normality

It is often necessary to know whether or not a variable is approximately normally distributed. If it is, we can go on to perform calculations of the type just discussed, and we shall also be able to use certain statistical tests considered in later chapters.

One simple way to test roughly for normality is to draw a histogram or frequency polygon of the distribution, to see if it is at least symmetrical and unimodal. We can also test whether the mean, median and mode are close together: as we have seen, these three measures coincide for an exactly normal distribution.

A rather more sophisticated test is based on the fact, discussed earlier, that a fixed proportion of the data will fall between particular values of the variable, if the distribution is normal. We saw that about 34 per cent of the observations should lie between the mean and one standard deviation (on each side of the mean), 47.7 per cent between the mean and two standard deviations, and 49.9 per cent between the mean and three standard deviations. In practice, the last figure means that virtually all the data should lie within three standard deviations of the mean. We can check these theoretical predictions against the actual data for our distribution.

Let us carry out such an exercise for the data from our language teaching experiment, introduced in chapter 1. In chapter 3, we calculated the means and standard deviations for these data. The values were as follows:

Group A: mean score = 14.93 marks;
standard deviation = 2.49 marks
Group B: mean score = 11.77 marks;
standard deviation = 3.31 marks

We can now predict that for each group, if the data are indeed normally distributed, about 68 per cent of the scores will lie in

the range mean ± 1 standard deviation, and about 95 per cent in the range mean ± 2 standard deviations. Thus we expect:

Group A: about 68% between 12.44 and 17.42;
about 95% between 8.95 and 20.91
Group B: about 68% between 8.46 and 15.08;
about 95% between 5.09 and 18.39

Remembering that, for example, 12.44 is very near the lower limit for the interval containing numbers that are rounded to 13, and also that the maximum score on the test is 20, we can translate the above ranges into whole mark scores as follows:

Group A: about 68% between 13 and 17;
about 95% between 9 and 20
Group B: about 68% between 9 and 15;
about 95% between 5 and 18

The figures actually observed were:

Group A: 13–17: 20 out of 30 = 67%;
9–20: 30 out of 30 = 100%
Group B: 9–15: 20 out of 30 = 67%;
5–18: 30 out of 30 = 100%

The distribution is perhaps a little short on the extreme values, but this may well be because the sample is quite small: remember that only 5 per cent of the values are expected to lie outside the two standard deviation limit, and in this case such a proportion represents only one or two scores. The predicted and observed proportions within one standard deviation of the mean agree very well indeed, and support our impressions, gained from graphical representations of the distribution, that the variable is approximately normally distributed.

A more exact test for the normality of a distribution will be considered in chapter 9. For most practical purposes, however, the approximate tests discussed above will suffice.

Exercises

1 (i) Explain what is meant in statistics by the term 'normal distribution', and outline its important properties.

(ii) Suppose that the following scores have been obtained on testing a group of 50 children for proficiency in spoken German. By means of a histogram, and by calculating the proportions of observations lying within certain distances from the mean, show that the scores are approximately normally distributed.

51	66	84	57	44
22	67	54	59	48
69	43	52	65	52
44	45	52	27	55
46	50	37	15	59
61	58	35	51	63
63	62	45	35	74
30	20	48	41	28
73	33	50	49	76
25	71	55	35	50

(iii) If a further large sample were taken from the population of students represented by the above group of 50, what percentage of the sample would be expected to score (a) less than 40 marks on the test? (b) between 60 and 70 marks on the test?

2 A group of 100 people have been asked to read a particular sentence, and the maximum sound intensity has been measured for each, in decibels from an arbitrary norm. It is found that the results are normally distributed, with a mean of 23.4 dB and a standard deviation of 5.8 dB. If a further sample were taken from the same population, what proportion of the maximum intensities would be expected to fall

(i) below 10 dB?
(ii) above 30 dB?
(iii) between 20 and 25 dB?

3 The following scores are obtained by 50 subjects on a language aptitude test:

42	62	44	32	47	42	52	76	36	43
55	27	46	55	47	28	53	44	15	61
18	59	58	57	49	55	88	49	50	62
61	82	66	80	64	50	40	53	28	63
63	25	58	71	82	52	73	67	58	77

(i) Draw a histogram to show the distribution of the scores.
(ii) Calculate the mean and standard deviation of the scores.
(iii) Use the values obtained in (ii) to show that the scores are approximately normally distributed.