Seventh Edition

STATISTICS FOR ECONOMICS, Accounting and Business Studies

MICHAEL BARROW

Pearson

Statistics for Economics, Accounting and Business Studies

The darker-shaded areas represent the probabilities of two or more sixes and together their area represents 13.2% of the whole distribution. This illustrates an important principle: that probabilities can be represented by areas under an appropriate probability distribution. We shall see more of this later.

Exercise 3.1

(a) If the probability of a randomly drawn individual having blue eyes is 0.6, what is the probability that four people drawn at random all have blue eyes?

(b) What is the probability that two of the sample of four have blue eyes?

(c) For this particular example, write down the Binomial formula for the probability of r blue-eyed individuals, $r = 0 \ldots 4$. Confirm that the calculated probabilities sum to one.

Exercise 3.2

(a) Calculate the mean and variance of the number of blue-eyed individuals in the previous exercise.

(b) Draw a graph of this Binomial distribution and on it mark the mean value and the mean value $+/-$ one standard deviation.

Having introduced the concept of probability distributions using the Binomial, we now move on to the most important of all probability distributions, the Normal.

The Normal distribution

The Binomial distribution applies when there are two possible outcomes to an experiment, but not all problems fall into this category. For instance, the (random) arrival time of a train is a continuous variable and cannot be analysed using the Binomial. There are many probability distributions in statistics, developed to analyse different types of problem. Several of them are covered in this text, and the most important of them is the Normal distribution, to which we now turn. It was discovered by the German mathematician Gauss in the nineteenth century (hence it is also known as the Gaussian distribution), in the course of his work on regression (see Chapter 7).

Many random variables turn out to be Normally distributed. Men's (or women's) heights are Normally distributed. IQ (the measure of intelligence) is also Normally distributed. Another example is of a machine producing (say) bolts with a nominal length of 5 cm which will actually produce bolts of slightly varying length (these differences would probably be extremely small) due to factors such as wear in the machinery, slight variations in the pressure of the lubricant, etc. These would result in bolts whose length varies, in accordance with the Normal distribution. This sort of process is extremely common, with the result that the Normal distribution often occurs in everyday situations.

The Normal distribution tends to arise when a random variable is the result of many independent, random influences added together, none of which dominates the others. A man's height is the result of many genetic influences, plus environmental factors such as diet, etc. As a result, height is Normally distributed. If one takes the height of men and women together, the result is not a Normal distribution, however. This is because one influence dominates the others: gender. Men are, on average, taller than women. Many variables familiar in economics are not

Normal, however – incomes, for example (although the logarithm of income is approximately Normal). We shall learn techniques to deal with such circumstances in due course.

Now that we have introduced the idea of the Normal distribution, what does it look like? It is presented below in graphical and then mathematical forms. Unlike the Binomial, the Normal distribution applies to continuous random variables such as height, and a typical Normal distribution is illustrated in Figure 3.5. Since the Normal distribution is a continuous one, it can be evaluated for any values of x, not just for integers as was the case for the Binomial distribution. The figure illustrates the main features of the distribution:

- It is unimodal, having a single, central peak. If this were men's heights, it would illustrate the fact that most men are clustered around the average height, with a few very tall and a few very short people.
- It is symmetric, the left and right halves being mirror images of each other.
- It is bell-shaped.
- It extends continuously over all the values of x from minus infinity to plus infinity, although the value of $f(x)$ becomes extremely small as these values are approached (the presentation of this figure being of only finite width, this last characteristic is not faithfully reproduced). This also demonstrates that most empirical distributions (such as men's heights) can only be an approximation to the theoretical ideal, although the approximation is close and good enough for practical purposes.

Note that we have labelled the y-axis '$f(x)$' rather than 'Pr(x)' as we did for the Binomial distribution. This is because it is *areas under the curve* that represent probabilities, not the heights. With the Binomial, which is a discrete distribution, one can legitimately represent probabilities by the heights of the bars. For the Normal, although $f(x)$ does not give the probability per se, it does give an indication: you are more likely to encounter values from the middle of the distribution (where $f(x)$ is greater) than from the extremes.

In mathematical terms, the formula for the Normal distribution is (x is the random variable)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{3.5}$$

The mathematical formulation is not so formidable as it appears. μ and σ are the parameters of the distribution, like n and P for the Binomial (even though they have different meanings); π is 3.1416 and e is 2.7183. If the formula is

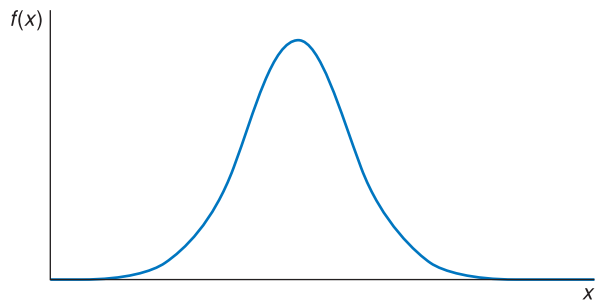

Figure 3.5
The Normal distribution

evaluated using different values of x, the values of $f(x)$ obtained will map out a Normal distribution. Fortunately, as we shall see, we do not need to use the mathematical formula in most practical problems.

Like the Binomial, the Normal is a family of distributions differing from one another only in the values of the parameters μ and σ. Several Normal distributions are drawn in Figure 3.6 for different values of the parameters.

Whatever value of μ is chosen turns out to be the centre of the distribution. Since the distribution is symmetric, μ is its mean. The effect of varying σ is to narrow (small σ) or widen (large σ) the distribution. σ turns out to be the standard deviation of the distribution. The Normal is another two-parameter family of distributions like the Binomial, and once the mean μ and the standard deviation σ

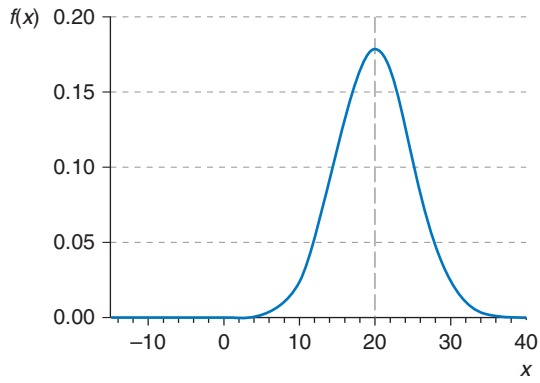

Figure 3.6(a)
The Normal distribution, $\mu = 20, \sigma = 5$

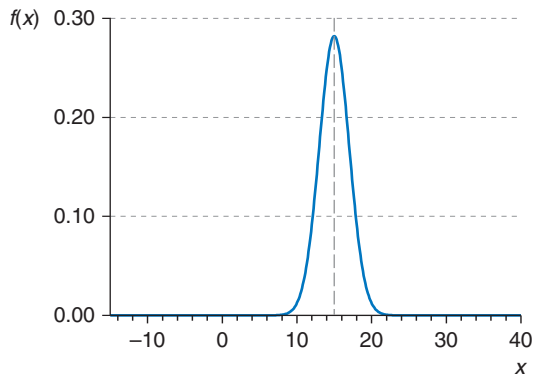

Figure 3.6(b)
The Normal distribution, $\mu = 15, \sigma = 2$

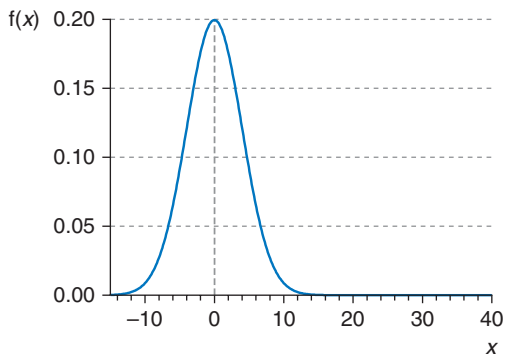

Figure 3.6(c)
The Normal distribution, $\mu = 0, \sigma = 4$

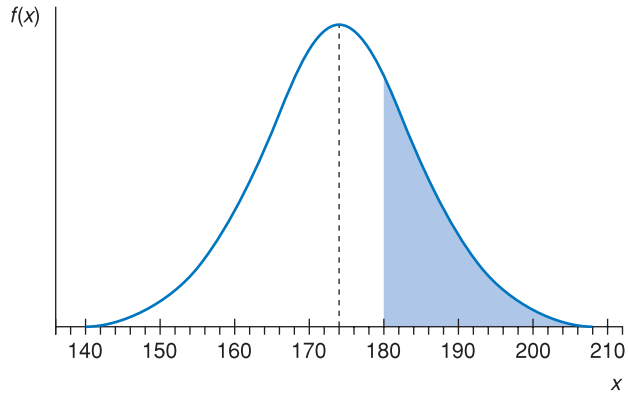

Figure 3.7
The height distribution of men

(or equivalently the variance, σ^2) are known, the whole of the distribution can be drawn. The shorthand notation for a Normal distribution is

$$x \sim N(\mu, \sigma^2) \tag{3.6}$$

meaning 'the variable x is Normally distributed with mean μ and variance σ^2'. This is similar in form to the expression for the Binomial distribution, although the meanings of the parameters are different.

Use of the Normal distribution can be illustrated using a simple example. The height of adult males is Normally distributed with mean height $\mu = 174$ cm and standard deviation $\sigma = 9.6$ cm. Let x represent the height of adult males; then

$$x \sim N(174, 92.16) \tag{3.7}$$

and this is illustrated in Figure 3.7. Note that (3.7) contains the variance rather than the standard deviation.

What is the probability that a randomly selected man is taller than 180 cm? If all men are equally likely to be selected, this is equivalent to asking what proportion of men are over 180 cm in height. This is given by the area under the Normal distribution, to the right of $x = 180$, i.e. the shaded area in Figure 3.7. The further from the mean of 174, the smaller the area in the tail of the distribution. One way to find this area would be to use equation (3.5), but this requires the use of sophisticated mathematics.

Since this is a frequently encountered problem, the answers have been set out in the tables of the **standard Normal distribution**. We can simply look up the solution. However, since there is an infinite number of Normal distributions (one for every combination μ and σ^2), it would be impossible to tabulate them all. The standard Normal distribution, which has a mean of zero and variance of one, is therefore used to represent all Normal distributions. Before the table can be consulted, therefore, the data have to be transformed so that they accord with the standard Normal distribution.

The required transformation is the z score, which was introduced in Chapter 1. This measures the distance between the value of interest (180) and the mean, measured in terms of standard deviations. Therefore, we calculate

$$z = \frac{x - \mu}{\sigma} \tag{3.8}$$

and z is a Normally distributed random variable with mean 0 and variance 1, i.e. $z \sim N(0, 1)$.

This transformation shifts the original distribution μ units to the left and then adjusts the dispersion by dividing through by σ, resulting in a mean of 0 and variance 1. z is Normally distributed because x is Normally distributed. The transformation in (3.8) retains the Normal distribution shape, despite the changes to mean and variance. If x followed some other distribution, then z would not be Normal either.

It is easy to verify the mean and variance of z using the rules for E and V operators encountered in Chapter 1:

$$\mathrm{E}(z) = \mathrm{E}\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma}(\mathrm{E}(x) - \mu) = 0 \quad (\text{since } \mathrm{E}(x) = \mu)$$

$$\mathrm{V}(z) = \mathrm{V}\left(\frac{x-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{V}(x)\frac{\sigma^2}{\sigma^2} = 1$$

Evaluating the z score from our data, we obtain

$$z = \frac{180 - 174}{9.6} = 0.63 \tag{3.9}$$

This shows that 180 is 0.63 standard deviations above the mean, 174, of the distribution. This is a measure of how far 180 is from 174 and allows us to look up the answer in tables. The task now is to find the area under the standard Normal distribution to the right of 0.63 standard deviations above the mean. This answer can be read off directly from the table of the standard Normal distribution, included as Table A2 in the Appendix. An excerpt from Table A2 is presented in Table 3.2.

The left-hand column gives the z score to one place of decimals. The appropriate row of the table to consult is the one for $z = 0.6$, which is shaded. For the second place of decimals (0.03) we consult the appropriate column, also shaded. At their intersection we find the value 0.2643, which is the desired area and therefore probability. In other words, 26.43% of the distribution lies to the right of 0.63 standard deviations above the mean. Therefore 26.43% of men are over 180 cm in height.

Use of the standard Normal table is possible because, although there is an infinite number of Normal distributions, they are all fundamentally the same, so that the area to the right of 0.63 standard deviations above the mean is the same for all of them. As long as we measure the distance in terms of standard deviations, then we can use the standard Normal table. The process of standardisation turns all Normal distributions into a standard Normal distribution with a mean of zero and a variance of one. This process is illustrated in Figure 3.8.

Table 3.2 Areas of the standard Normal distribution (excerpt from Table A2)

z	0.00	0.01	0.02	0.03	...	0.09
0.0	0.5000	0.4960	0.4920	0.4880	...	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	...	0.4247
⋮	⋮	⋮	⋮	⋮	...	⋮
0.5	0.3085	0.3050	0.3015	0.2981	...	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	...	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	...	0.2148

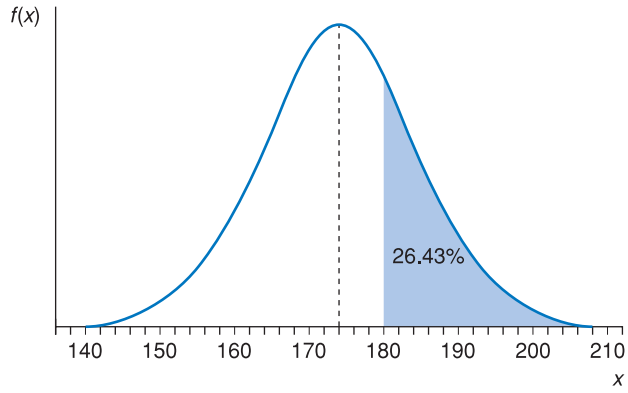

Figure 3.8(a)
The Normal distribution, $\mu = 174, \sigma = 9.6$

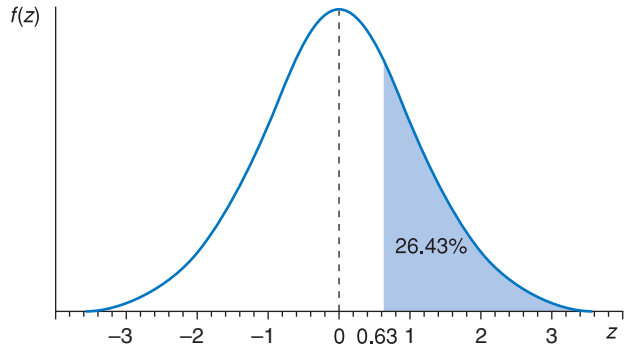

Figure 3.8(b)
The standard Normal distribution corresponding to Figure 3.8(a)

The area in the right-hand tail is the same for both distributions. It is the standard Normal distribution in Figure 3.8(b) which is tabulated in Table A2. To demonstrate how standardisation turns all Normal distributions into the standard Normal, the earlier problem is repeated but taking all measurements in inches. The answer should obviously be the same. Taking 1 inch = 2.54 cm, the figures are

$$x = 70.87 \quad \sigma = 3.78 \quad \mu = 68.50$$

What proportion of men are over 70.87 inches in height? The appropriate Normal distribution is now

$$x \sim N(68.50, 3.78^2) \tag{3.10}$$

The z score is

$$z = \frac{70.87 - 68.50}{3.78} = 0.63 \tag{3.11}$$

which is the same z score as before and therefore gives the same probability.

Worked example 3.2

Packets of cereal have a nominal weight of 750 grams, but there is some variation around this as the machines filling the packets are imperfect. Let us assume that the weights follow a Normal distribution. Suppose that the standard deviation around the mean of 750 is 5 grams. What proportion of packets weigh more than 760 grams?