The random variable X, which represents the number of successes in n trials of this experiment, is said to have a **binomial distribution**.

A consequence of the last condition is that the probability of failure will also be a constant, equal to $1-p$. This probability is usually denoted by q, which means that $q=1-p$.

In a binomial distribution, the random variable X has a probability distribution given by

$$\mathrm{P}(X=x)=\binom{n}{x}p^x q^{n-x} \qquad \text{for } x=0,1,2,\ldots,n.$$

You will see the reason for the name 'binomial' a little later. Provided you are given the values of n and p, you can evaluate all of the probabilities in the distribution table. The values of n and p are therefore the essential pieces of information about the probability distribution. In the example, n was 5 and p was $\frac{1}{3}$. You did not need to be told q, because its value is always $1-p$, so in the example the value of q was $\frac{2}{3}$.

The values of n and p are called the **parameters** of the binomial distribution. You need to know the parameters of a probability distribution to calculate the probabilities numerically.

To denote that a random variable X has a binomial distribution with parameters n and p, you write $X \sim \mathrm{B}(n,p)$. So for the probability distribution in Table 7.2 you write $X \sim \mathrm{B}(5,\frac{1}{3})$.

Example 7.1.1

Given that $X \sim \mathrm{B}(8,\frac{1}{4})$, find (a) $\mathrm{P}(X=6)$, (b) $\mathrm{P}(X \leqslant 2)$, (c) $\mathrm{P}(X>0)$.

(a) Using the binomial probability formula with $n=8$ and $p=\frac{1}{4}$ you get

$$\mathrm{P}(X=6)=\binom{8}{6}\times\left(\tfrac{1}{4}\right)^6\times\left(\tfrac{3}{4}\right)^2=28\times\left(\tfrac{1}{4}\right)^6\times\left(\tfrac{3}{4}\right)^2=0.003\,85,$$

correct to 3 significant figures.

(b) $\mathrm{P}(X \leqslant 2)=\mathrm{P}(X=0)+\mathrm{P}(X=1)+\mathrm{P}(X=2)$

$$\begin{aligned}
&=\binom{8}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^8+\binom{8}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^7+\binom{8}{0}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^6\\
&=0.1001\ldots+0.2669\ldots+0.3114\ldots\\
&=0.6785\ldots=0.679, \text{correct to 3 significant figures.}
\end{aligned}$$

(c) The easiest way to find $\mathrm{P}(X>0)$ is to use the fact that $\mathrm{P}(X>0)$ is the complement of $\mathrm{P}(X=0)$.

So $\mathrm{P}(X>0)=1-\mathrm{P}(X=0)$

$$\begin{aligned}
&=1-\binom{8}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^8 \qquad \text{(from part (b))}\\
&=1-0.1001\ldots\\
&=0.8998\ldots=0.900, \text{correct to 3 significant figures.}
\end{aligned}$$

To check that the binomial formula does represent a probability distribution you must show that $\sum_{x=0}^{n} \mathrm{P}(X=x)=1$. Consider the example involving the spinner, but use p and q instead of $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Table 7.3 shows the distribution.

x	$\mathrm{P}(X=x)$	
0	q^5	$=q^5$
1	$\binom{5}{1}\times p\times q^4$	$=5pq^4$
2	$\binom{5}{2}\times p^2\times q^3$	$=10p^2q^3$
3	$\binom{5}{3}\times p^3\times q^2$	$=10p^3q^2$
4	$\binom{5}{4}\times p^4\times q$	$=5p^4q$
5	p^5	$=p^5$

Table 7.3. Probability distribution for the number of times out of 5 that the spinner shows black.

If you sum the probabilities in the right column you get

$$\sum_{x=0}^{5} \mathrm{P}(X=x)=q^5+5pq^4+10p^2q^3+10p^3q^2+5p^4q+p^5.$$

The right side of this equation is the binomial expansion of $(q+p)^5$ (see P1 Chapter 9). You could check for yourself by multiplying out $(q+p)(q+p)(q+p)(q+p)(q+p)$, so

$$\sum_{x=0}^{5} \mathrm{P}(X=x)=(q+p)^5=1^5=1.$$

You can use a similar argument to show that

$$\sum_{x=0}^{n} \mathrm{P}(X=x)=\sum_{x=0}^{n}\left[\binom{n}{x}\times p^x\times q^{n-x}\right]=(q+p)^n=1^n=1.$$

The individual probabilities in the binomial distribution are the terms of the binomial expansion of $(q+p)^n$: these are two similar uses of the word 'binomial'.

On the next page there is a summary of the binomial distribution.

Binomial distribution

- A single trial has exactly two possible outcomes (success and failure) and these are mutually exclusive.
- A fixed number, n, of trials takes place.
- The outcome of each trial is independent of the outcome of all the other trials.
- The probability of success at each trial is constant.

The random variable X, which represents the number of successes in the n trials of this experiment, has a probability distribution given by

$$\mathrm{P}(X=x)=\binom{n}{x}p^x q^{n-x} \qquad \text{for } x=0,1,2,\ldots,n, \qquad (7.1)$$

where p is the probability of success and $q=1-p$ is the probability of failure.

When the random variable X satisfies these conditions, $X \sim \mathrm{B}(n,p)$.

Exercise 7A

In this exercise give probabilities correct to 4 decimal places.

1 The random variable X has a binomial distribution with $n=6$ and $p=0.2$. Calculate

(a) $\mathrm{P}(X=3)$, (b) $\mathrm{P}(X=4)$, (c) $\mathrm{P}(X=6)$.

2 Given that $Y \sim \mathrm{B}\left(7,\frac{2}{3}\right)$, calculate

(a) $\mathrm{P}(Y=4)$, (b) $\mathrm{P}(Y=6)$, (c) $\mathrm{P}(Y=0)$.

3 Given that $Z \sim \mathrm{B}(9,0.45)$, calculate

(a) $\mathrm{P}(Z=3)$, (b) $\mathrm{P}(Z=4 \text{ or } 5)$, (c) $\mathrm{P}(Z \geqslant 7)$.

4 Given that $D \sim \mathrm{B}(12,0.7)$, calculate

(a) $\mathrm{P}(D<4)$, (b) the smallest value of d such that $\mathrm{P}(D>d)<0.90$.

5 Given that $H \sim \mathrm{B}\left(9,\frac{1}{2}\right)$, calculate the probability that H is

(a) exactly 5, (b) 5 or 6, (c) at least 8, (d) more than 2.

6 Given that $S \sim \mathrm{B}\left(7,\frac{1}{6}\right)$, find the probability that S is

(a) exactly 3, (b) at least 4.

7* If $X \sim \mathrm{B}(n,p)$, show that $\mathrm{P}(X=r+1)=\mathrm{P}(X=r)\times\dfrac{p(n-r)}{q(r+1)}$ for $r=0,1,\ldots,n-1$.

If you have access to a spreadsheet, use this formula to construct tables for binomial probabilities.

Why is it better to use this formula than to calculate $\binom{n}{r}p^r q^{n-r}$ directly?

8* Use the formula of Question 7 to prove that the mode of a binomial distribution (that is, the value of r with the highest probability) satisfies $(n+1)p-1 \leqslant \text{mode} \leqslant (n+1)p$.

When is there equality?

7.2 Using the binomial distribution as a model

Before using the binomial distribution as a model for a situation you need to convince yourself that all the conditions are satisfied. The following example illustrates some of the problems that can occur.

Example 7.2.1

A school car park has 5 parking spaces. A student decides to do a survey to see whether this is enough. At the same time each day, she observes the number of spaces which are filled. Let X be the number of spaces filled at this time on a randomly chosen day. Is it reasonable to model the distribution of the random variable X with a binomial distribution?

She looks at each parking space to see whether it is occupied or not. This represents a single trial.

Are there exactly two outcomes for each trial (parking space), and are these mutually exclusive? In other words, is each parking space either occupied by a single car or not? The answer will usually be yes, but sometimes poorly parked vehicles will give the answer no.

Are there a fixed number of trials? The answer is yes. On each day there are 5 parking spaces available so the number of trials is 5.

Are the trials independent? This is not likely. Drivers may be less inclined to park in one of the centre spaces if it is surrounded by cars, because getting out of their own car may be more difficult.

Is the probability p of success (in this case a parking space being filled by a car) constant? Probably not, because people may be more likely to choose the space closest to the school entrance, for example.

You can see that, when you are proposing to model a practical situation with a binomial distribution, many of the assumptions may be questionable and some may not be valid at all. In this case, however, provided you are aware that the binomial model is far from perfect, you could still use it as a reasonable approximation. You might also have realised that you do not know the value of p in this example, so you would have to estimate it. To do this you would divide the total number of cars observed by the total number of available car parking spaces, which in this case is

$5 \times$ (the number of days for which the survey was carried out).

Example 7.2.2
State whether a binomial distribution could be used in each of the following problems. If the binomial distribution is an acceptable model, define the random variable clearly and state its parameters.

(a) A fair cubical dice is rolled 10 times. Find the probability of getting three 4s, four 5s and three 6s.

(b) A fair coin is spun until a head occurs. Find the probability that eight spins are necessary, including the one on which the head occurs.

(c) A jar contains 49 balls numbered 1 to 49. Six of the balls are selected at random. Find the probability that four of the six have an even score.

(a) In this case you are interested in three different outcomes: a 4, a 5 and a 6. A binomial distribution depends on having only two possible outcomes, success and failure, so it cannot be used here.

(b) The binomial distribution requires a fixed number of trials, n, and this is not the case here, since the number of trials is unknown. In fact, the number of trials is the random variable of interest here.

(c) Whether a binomial model is appropriate or not depends on whether the selection of the balls is done with replacement or without replacement. If the selection is without replacement, then the outcome of each trial will not be independent of all the other trials. If the selection is with replacement, then define the random variable X to be the number of balls with an even score out of six random selections. X will then have a binomial distribution with parameters 6 and $\frac{24}{49}$. You write this as $X \sim \mathrm{B}\left(6, \frac{24}{49}\right)$. You are assuming, of course, that the balls are thoroughly mixed before each selection and that every ball has an equal chance of being selected.

Example 7.2.3
A card is selected at random from a standard pack of 52 playing cards. The suit of the card is recorded and the card is replaced. This process is repeated to give a total of 16 selections, and on each occasion the card is replaced in the pack before another selection is made. Calculate the probability that

(a) exactly five hearts occur in the 16 selections,

(b) at least three hearts occur.

Let X be the number of hearts in 16 random selections (with replacement) of a playing card from a pack. Then X satisfies all the conditions for a binomial distribution.

- Each trial consists of selecting a card from the pack, with replacement.
- Each trial has exactly two possible outcomes, and these are mutually exclusive; getting a heart is a success and not getting a heart is a failure.

You may think that there are 52 *possible outcomes for each trial, but you are only interested in whether the card is a heart or not a heart.*

- The outcome of each trial is independent of any other trial. This is true since each card is replaced before the next one is selected. But you must ensure that each selection is random and that the cards are thoroughly shuffled before each selection.

- The probabilities of success and failure are constant. As the cards are replaced, P(selecting a heart) = P(success) = $\frac{1}{4}$ and P(not selecting a heart) = P(failure) = $\frac{3}{4}$, so this condition is fulfilled.

X therefore has a binomial distribution with parameters $n = 16$ and $p = \frac{1}{4}$. That is,

$$X \sim \mathrm{B}\left(16, \tfrac{1}{4}\right).$$

(a) Using the binomial formula,

$$\mathrm{P}(X = 5) = \binom{16}{5} \times \left(\tfrac{1}{4}\right)^5 \times \left(\tfrac{3}{4}\right)^{11} = 0.180, \text{ correct to 3 significant figures.}$$

(b) To find $\mathrm{P}(X \geqslant 3)$, use the fact that $\mathrm{P}(X \geqslant 3) = 1 - \mathrm{P}(X \leqslant 2)$.

$$\begin{aligned}
\mathrm{P}(X \geqslant 3) &= 1 - \mathrm{P}(X \leqslant 2) \\
&= 1 - \binom{16}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^{16} - \binom{16}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^{15} - \binom{16}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^{14} \\
&= 1 - 0.010\,02 - 0.053\,45 - 0.133\,63\ldots \\
&= 1 - 0.197\,11 \\
&= 0.802\,88\ldots = 0.803, \text{correct to 3 significant figures.}
\end{aligned}$$

7.3 Practical activities

1 Penalties or shots

(a) Select a group of students and ask them each to take either 8 penalties at football or 8 shots at basketball. For each student record the number of successful penalties or shots.

(b) Does the binomial distribution provide a reasonable model for these results? Is it necessary to use the same goalkeeper for all of the football penalties?

(c) Does the skill level of each person matter if the binomial distribution is to be a reasonable model? Is the basketball example more likely to be fitted by a binomial model than the football example?

Exercise 7B

1 In a certain school, 30% of the students are in the age group 16–19.

(a) Ten students are chosen at random. What is the probability that fewer than four of them are in the 16–19 age group?

(b) If the ten students were chosen by picking ten who were sitting together at lunch, explain why a binomial distribution might no longer have been suitable.

2 A factory makes large quantities of coloured sweets, and it is known that on average 20% of the sweets are coloured green. A packet contains 20 sweets. Assuming that the packet forms a random sample of the sweets made by the factory, calculate the probability that exactly seven of the sweets are green.

If you knew that, in fact, the sweets could have been green, red, orange or brown, would it have invalidated your calculation?

3 Eggs produced at a farm are packaged in boxes of six. Assume that, for any egg, the probability that it is broken when it reaches the retail outlet is 0.1, independent of all other eggs. A box is said to be bad if it contains at least two broken eggs. Calculate the probability that a randomly selected box is bad.

Ten boxes are chosen at random. Find the probability that just two of these boxes are bad.

It is known that, in fact, breakages are more likely to occur after the eggs have been packed into boxes, and while they are being transported to the retail outlet. Explain why this fact is likely to invalidate the calculation.

4 On a particular tropical island, the probability that there is a hurricane in any given month can be taken to be 0.08. Use a binomial distribution to calculate the probability that there is a hurricane in more than two months of the year. State two assumptions needed for a binomial distribution to be a good model. Why may one of the assumptions not be valid?

5 It is given that, at a stated time of day, 35% of the adults in the country are wearing jeans. At that time, a sample of twelve adults is selected. Use a binomial distribution to calculate the probability that exactly five out of these twelve are wearing jeans. Explain carefully two assumptions that must be made for your calculation to be valid. (If you say 'sample is random' you must explain what this means in the context of the question.)

6 Explain why a binomial distribution would not be a good model in the following problem. (Do not attempt any calculation.)

Thirteen cards are chosen at random from an ordinary pack. Find the probability that there are four clubs, four diamonds, three hearts and two spades.

7 Explain why the binomial distribution $\mathrm{B}(6,0.5)$ would not be a good model in each of the following situations. (Do not attempt any calculations.)

(a) It is known that 50% of the boys in a certain school are over 170 cm in height. They are arranged, for a school photograph, in order of ascending height. A group of six boys standing next to each other is selected at random. Find the probability that exactly three members of the sample are over 170 cm in height.

(b) It is known that, on average, the temperature in London reaches at least 20 °C on exactly half the days in the year. A day is picked at random from each of the months January, March, May, July, September and November. Find the probability that the temperature in London reaches 20 °C on exactly three of these six days.

8 A bag contains six red and four green counters. Four counters are selected at random, without replacement. The events A, B, C and D represent obtaining a red counter on the first, second, third and fourth selection, respectively.

Use a tree diagram to show that $\mathrm{P}(A)=\mathrm{P}(B)=\mathrm{P}(C)=\mathrm{P}(D)=0.6$.

Explain why the total number of red counters could not be well modelled by the distribution $\mathrm{B}(4,0.6)$.

The purpose of this and the preceding question is to illustrate that the properties 'the probability of a success is constant' and 'the outcomes are independent' are not the same, and you should try to distinguish carefully between them. Notice also that 'the outcomes are independent' is not the same thing as 'sampling with replacement'.

Miscellaneous exercise 7

1 The probability of a novice archer hitting a target with any shot is 0.3. Given that the archer shoots six arrows, find the probability that the target is hit at least twice. (OCR)

2 A computer is programmed to produce at random a single digit from the list $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$. The program is run twenty times. Let Y be the number of zeros that occur.

(a) State the distribution of Y and give its parameters.

(b) Calculate $\mathrm{P}(Y < 3)$.

3 A dice is biased so that the probability of throwing a 6 is 0.2. The dice is thrown eight times. Let X be the number of '6's thrown.

(a) State the distribution of X and give its parameters.

(b) Calculate $\mathrm{P}(X > 3)$.

4 Joseph and four friends each have an independent probability 0.45 of winning a prize. Find the probability that

(a) exactly two of the five friends win a prize,

(b) Joseph and only one friend win a prize. (OCR)

5 A bag contains two biased coins: coin A shows Heads with probability 0.6, and coin B shows Heads with probability 0.25. A coin is chosen at random from the bag, and tossed three times.

(a) Find the probability that the three tosses of the coin show two Heads and one Tail in any order.

(b) Find the probability that the coin chosen was coin A, given that the three tosses result in two Heads and one Tail. (OCR)

6 (a) A fair coin is tossed 4 times. Calculate the probabilities that the tosses result in $0, 1, 2$, 3 and 4 heads.

(b) A fair coin is tossed 8 times. Calculate the probability that the first 4 tosses and the last 4 tosses result in the same number of heads.

(c) Two teams each consist of 3 players. Each player in a team tosses a fair coin once and the team's score is the total number of heads thrown. Find the probability that the teams have the same score. (OCR)

7 State the conditions under which the binomial distribution may be used for the calculation of probabilities.

The probability that a girl chosen at random has a weekend birthday in 1993 is $\frac{2}{7}$. Calculate the probability that, among a group of ten girls chosen at random,

(a) none has a weekend birthday in 1993,

(b) exactly one has a weekend birthday in 1993.

Among 100 groups of ten girls, how many groups would you expect to contain more than one girl with a weekend birthday in 1993? (OCR)

8 Show that, when two fair dice are thrown, the probability of obtaining a 'double' is $\frac{1}{6}$, where a 'double' is defined as the same score on both dice. Four players play a board game which requires them to take it in turns to throw two fair dice. Each player throws the two dice once in each round. When a double is thrown the player moves forward six squares. Otherwise the player moves forward one square. Find

(a) the probability that the first double occurs on the third throw of the game,

(b) the probability that exactly one of the four players obtains a double in the first round,

(c) the probability that a double occurs exactly once in 4 of the first 5 rounds. (OCR)

9 Six hens are observed over a period of 20 days and the number of eggs laid each day is summarised in the following table.

Number of eggs	3	4	5	6
Number of days	2	2	10	6

Show that the mean number of eggs per day is 5.

It may be assumed that a hen never lays more than one egg in any day. State one other assumption that needs to be made in order to consider a binomial model, with $n = 6$, for the total number of eggs laid in a day. State the probability that a randomly chosen hen lays an egg on a given day.

Calculate the expected frequencies of 3, 4, 5 and 6 eggs. (OCR)

10 A Personal Identification Number (PIN) consists of 4 digits in order, each of which is one of the digits $0, 1, 2, \ldots, 9$. Susie has difficulty remembering her PIN. She tries to remember it and writes down what she thinks it is. The probability that the first digit is correct is 0.8 and the probability that the second digit is correct is 0.86. The probability that the first two digits are correct is 0.72. Find

(a) the probability that the second digit is correct given that the first digit is correct,

(b) the probability that the first digit is correct and the second digit is incorrect,

(c) the probability that the first digit is incorrect and the second digit is correct,

(d) the probability that the second digit is incorrect given that the first digit is incorrect.

The probability that all four digits are correct is 0.7. On 12 separate occasions Susie writes down independently what she thinks is her PIN. Find the probability that the number of occasions on which all four digits are correct is less than 10. (OCR)

8 Expectation and variance of a random variable

This chapter shows you how to calculate the mean and variance of a discrete random variable. When you have completed it, you should

- know the meaning of the notation $\mathrm{E}(X)$ and $\mathrm{Var}(X)$
- be able to calculate the mean, $\mathrm{E}(X)$, of a random variable X
- be able to calculate the variance, $\mathrm{Var}(X)$, of a random variable X
- be able to use the formulae $\mathrm{E}(X) = np$ and $\mathrm{Var}(X) = np(1-p)$ for a binomial distribution.

8.1 Expectation

A computer is programmed to produce a sequence of integers, X, from 0 to 3 inclusive, with probabilities as shown below.

x	0	1	2	3
$\mathrm{P}(X = x)$	0.4	0.3	0.2	0.1

Suppose that a sequence of 100 integers is produced by the computer. What would you expect the mean of these 100 values to be? It is not possible to answer this question exactly because you cannot tell how often each value will actually turn up in the sequence. However, it is possible to obtain an *estimate* of the mean value. You can estimate the frequency with which each integer occurs using

$$\text{frequency} \approx \text{total frequency} \times \text{probability}$$

(see Section 6.3). You might expect there to be about $100 \times 0.4 = 40$ '0's, $100 \times 0.3 = 30$ '1's, $100 \times 0.2 = 20$ '2's and $100 \times 0.1 = 10$ '3's. The sum of these integers would be

$$(0 \times 40) + (1 \times 30) + (2 \times 20) + (3 \times 10) = 100$$

so their mean would be $\frac{100}{100} = 1$.

If you look at this calculation carefully, you will see that it is independent of the number of integers in the sequence. For example, if you had a sequence of 1000 integers, then the sum of the integers would be 10 times as great, but the estimate of the mean would stay the same.

The same result can be obtained more directly by multiplying each value by its probability and summing. Using p_i as a shortened form of $\mathrm{P}(X = x_i)$, this gives

$$\sum x_i p_i = (0 \times 0.4) + (1 \times 0.3) + (2 \times 0.2) + (3 \times 0.1) = 1.$$

The value which has just been calculated is a theoretical mean. It is denoted by μ (which is read as 'mu'), the Greek letter m, standing for 'mean'. The new symbol is used in order to distinguish the mean of a probability distribution from $\bar{x}$, the mean of a data set. The mean, μ, of a probability distribution does not represent the mean of a finite sequence of numbers. It is the value to which the mean tends as the length of the sequence gets larger and larger, or, as mathematicians say, 'tends to infinity'. In practice, it is helpful to think of μ as the mean you would expect for a very very long sequence. For this reason, μ is often called the **expectation** or **expected value** of X and is denoted by $\mathrm{E}(X)$.

> The expectation of a random variable X is defined by $\mathrm{E}(X) = \mu = \sum x_i p_i$.

Example 8.1.1

Find the expected value of each of the variables X, Y and W, which have the probability distributions given below.

(a)

x	1	2	3	4	5	6	Total
$\mathrm{P}(X=x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	1

(b)

y	1	2	3	4	5	7	8	9	10	11	12	Total
$\mathrm{P}(Y=y)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	$\frac{1}{36}$	1

(c)

w	2	3	4	5	6	7	8	9	10	11	12	Total
$\mathrm{P}(W=w)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

(a) $\mathrm{E}(X) = \sum x_i p_i = \left(1\times\frac{1}{6}\right)+\left(2\times\frac{1}{6}\right)+\left(3\times\frac{1}{6}\right)+\left(4\times\frac{1}{6}\right)+\left(5\times\frac{1}{6}\right)+\left(6\times\frac{1}{6}\right)=3\frac{1}{2}$.

You may have spotted that there is a quicker way to find the mean in this example. Since the distribution is symmetrical about $3\frac{1}{2}$, the mean must equal $3\frac{1}{2}$.

(b) This distribution is not symmetrical and so the mean has to be calculated.

$$\begin{aligned}\mathrm{E}(Y) = \sum y_i p_i &= \left(1\times\tfrac{1}{6}\right)+\left(2\times\tfrac{1}{6}\right)+\left(3\times\tfrac{1}{6}\right)+\left(4\times\tfrac{1}{6}\right)+\left(5\times\tfrac{1}{6}\right)+\left(7\times\tfrac{1}{36}\right)\\ &\quad+\left(8\times\tfrac{1}{36}\right)+\left(9\times\tfrac{1}{36}\right)+\left(10\times\tfrac{1}{36}\right)+\left(11\times\tfrac{1}{36}\right)+\left(12\times\tfrac{1}{36}\right)\\ &=(1+2+3+4+5)\times\tfrac{1}{6}+(7+8+9+10+11+12)\times\tfrac{1}{36}\\ &=15\times\tfrac{1}{6}+57\times\tfrac{1}{36}=4\tfrac{1}{12}.\end{aligned}$$

(c) As in part (a), the probability distribution is symmetrical, in this case about 7, so $\mathrm{E}(W)=7$.

The variables X, Y and W were discussed in Section 6.1 in connection with the number of squares moved in a turn at three different board games. This calculation shows that you move round the board fastest in Game C and slowest in Game A.

Example 8.1.2

A random variable R has the probability distribution shown below.

r	1	2	3	4
$\mathrm{P}(R=r)$	0.1	a	0.3	b

Given that $\mathrm{E}(R)=3$, find a and b.

Since $\sum \mathrm{P}(R=r)=1$,

$$0.1+a+0.3+b=1, \quad \text{so} \quad a+b=0.6.$$

Also $\mathrm{E}(R)=3$, so $\sum r\mathrm{P}(R=r)=3$,

$$1\times 0.1+2\times a+3\times 0.3+4\times b=3, \quad \text{so} \quad 2a+4b=2.$$

Solving these two equations simultaneously gives $a=0.2$ and $b=0.4$.

8.2 The variance of a random variable

Example 8.1.1 showed that the random variables X, Y and W have different means. If you compare the probability distributions (which are illustrated in Fig. 6.6), you will see that X, Y and W also have different degrees of spread. Just as the spread in a data set can be measured by the standard deviation or variance, so it is possible to define a corresponding measure of spread for a random variable. The symbol used for the standard deviation of a random variable is σ (a small Greek s, read as 'sigma') and its square, σ^2, the variance of a random variable, is denoted by $\mathrm{Var}(X)$.

Before deriving a formula for $\mathrm{Var}(X)$, it is helpful to look at another method of arriving at the formula for $\mathrm{E}(X)$. Suppose that you had a sequence of n integers produced by the computer described in Section 8.1, and that the sequence contained f_1 '0's, f_2 '1's, f_3 '2's and f_4 '3's. The mean for these n integers is given by

$$\bar{x}=\frac{\sum x_i f_i}{n}=\frac{0\times f_1}{n}+\frac{1\times f_2}{n}+\frac{2\times f_3}{n}+\frac{3\times f_4}{n}.$$

The right side of the expression can be written slightly differently in the form

$$\bar{x}=0\times\frac{f_1}{n}+1\times\frac{f_2}{n}+2\times\frac{f_3}{n}+3\times\frac{f_4}{n}=\sum\left(x_i\times\frac{f_i}{n}\right).$$

Now consider what happens as n becomes very large: the value of $\bar{x}$ tends to μ, and the ratio $\frac{f_i}{n}$, which is the relative frequency, tends to the corresponding theoretical probability, p_i. This gives

$$\mu=\mathrm{E}(X)=\sum x_i p_i. \qquad (8.1)$$

which was the result obtained in Section 8.1.

Now consider the formula given in Equation 3.3 for the variance of a data set. Replacing $\sum f_i$ by n and rearranging gives

$$\text{variance} = \frac{\sum (x_i - \bar{x})^2 f_i}{n} = \sum (x_i - \bar{x})^2 \times \frac{f_i}{n}.$$

Again consider what happens when n becomes large. The ratio $\frac{f_i}{n}$ tends to p_i, and $\bar{x}$ tends to μ, giving

$$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 p_i. \qquad (8.2)$$

Alternatively, starting from Equation 3.4 for the variance of a data set, replacing $\sum f_i$ by n and rearranging gives

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

When n becomes large, $\frac{f_i}{n}$ tends to p_i and $\bar{x}$ tends to μ, giving

$$\sigma^2 = \text{Var}(X) = \sum x_i^2 p_i - \mu^2. \qquad (8.3)$$

The **variance** of a random variable X is defined by

$$\sigma^2 = \text{Var}(X) = \sum (x_i - \mu)^2 p_i = \sum x_i^2 p_i - \mu^2.$$

The **standard deviation** of a random variable is σ, the square root of $\text{Var}(X)$.

In practice it is usually simpler to calculate $\text{Var}(X)$ from Equation 8.3 rather than from Equation 8.2.

Example 8.2.1
Calculate the standard deviation of the random variable X in Example 8.1.1, using Equation 8.3.

First calculate $\sum x_i^2 p_i$:

$$\begin{aligned}\sum x_i^2 p_i &= \left(1^2 \times \tfrac{1}{6}\right) + \left(2^2 \times \tfrac{1}{6}\right) + \left(3^3 \times \tfrac{1}{6}\right) + \left(4^2 \times \tfrac{1}{6}\right) + \left(5^2 \times \tfrac{1}{6}\right) + \left(6^2 \times \tfrac{1}{6}\right)\\ &= \left(1^2 + 2^2 + 3^2 + 4^2 + 5^2 + 6^2\right) \times \tfrac{1}{6} = 91 \times \tfrac{1}{6} = 15\tfrac{1}{6}.\end{aligned}$$

From Example 8.1.1, $\mu = \text{E}(X) = 3\tfrac{1}{2}$.

Using Equation 8.3:

$$\sigma^2 = \text{Var}(X) = \sum x_i^2 p_i - \mu^2 = 15\tfrac{1}{6} - \left(3\tfrac{1}{2}\right)^2 = \tfrac{35}{12}.$$

Then calculate the standard deviation:

$$\sigma = \sqrt{\frac{35}{12}} = 1.71, \text{ correct to 3 significant figures.}$$

The standard deviations for the random variables Y and W in Example 8.1.1 are $\sqrt{\frac{10\,395}{1296}} = 2.83$ and $\sqrt{\frac{35}{6}} = 2.42$ respectively, both given to 3 significant figures.

You could check these values. If you look again at Fig. 6.6 you will see how the size of the standard deviation is related to the degree of spread of the distribution. Although Y and W have very similar ranges, W has a smaller standard deviation because the probability distribution rises to a peak at the centre.

Example 8.2.2

In a certain field, each mushroom which is growing gives rise to a number X of mushrooms in the following year. None of the mushrooms present in one year survives until the next year. The random variable X has the following probability distribution.

x	0	1	2
$\mathrm{P}(X = x)$	0.2	0.6	0.2

If there were two mushrooms present in one year, find the probability distribution of Y, the number of mushrooms present in the following year. Hence find the mean and variance of Y.

The possible values of Y are given below, where the first value of X is the value of X for one mushroom, and the second value of X is the value of X for the second mushroom.

		First value of X		
		0	1	2
	0	0	1	2
Second value of X	1	1	2	3
	2	2	3	4

The corresponding probabilities are given below.

		First value of X		
		0	1	2
	0	0.2×0.2	0.2×0.6	0.2×0.2
Second value of X	1	0.6×0.2	0.6×0.6	0.6×0.2
	2	0.2×0.2	0.2×0.6	0.2×0.2

Combining these two sets of results gives the probability distribution of Y, from which $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$ can be found.

y	$\mathrm{P}(Y=y)$	$y\mathrm{P}(Y=y)$	$y^2\mathrm{P}(Y=y)$
0	0.04	0	0
1	$0.12+0.12=0.24$	0.24	0.24
2	$0.04+0.36+0.04=0.44$	0.88	1.76
3	$0.12+0.12=0.24$	0.72	2.16
4	0.04	0.16	0.64
	Totals: $\sum \mathrm{P}(Y=y)=1$	$\sum y\mathrm{P}(Y=y)=2$	$\sum y^2\mathrm{P}(Y=y)=4.8$

From the last row of the table $\sum y\mathrm{P}(Y=y)=2$, and $\sum y^2\mathrm{P}(Y=y)=4.8$.

Then $\mathrm{E}(Y)=\sum y\mathrm{P}(Y=y)=2$, and

$$\mathrm{Var}(Y)=\sum y^2\mathrm{P}(Y=y)-(\mathrm{E}(Y))^2=4.8-2^2=0.8.$$

Exercise 8A

In this exercise all variables are discrete. Give numerical answers to 4 significant figures when appropriate.

1 Find the mean of the random variables X and Y which have the following probability distributions.

(a)

x	0	1	2	3	4
$\mathrm{P}(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$

(b)

y	-2	-1	0	1	2	3
$\mathrm{P}(Y=y)$	0.15	0.25	0.3	0.05	0.2	0.05

2 The random variable T has the probability distribution given in the following table.

t	1	2	3	4	5	6	7
$\mathrm{P}(T=t)$	0.1	0.2	0.1	0.2	0.1	0.2	0.1

Find $\mathrm{E}(T)$ and $\mathrm{Var}(T)$.

3 Find the exact expectation and variance of the random variable Y, which has the following probability distribution.

y	3	4	5	6	7
$\mathrm{P}(Y=y)$	$\frac{1}{18}$	$\frac{5}{18}$	$\frac{7}{18}$	$\frac{1}{18}$	$\frac{4}{18}$

4 The six faces of a fair cubical dice are numbered 1, 2, 2, 3, 3 and 3. When the dice is thrown once, the score is the number appearing on the top face. This is denoted by X.

(a) Find the mean and standard deviation of X.

(b) The dice is thrown twice and Y denotes the sum of the scores obtained. Find the probability distribution of Y. Hence find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

5 A construction company can bid for one of two possible projects and the finance director has been asked to advise on which to choose. She estimates that project *A* will yield a profit of \$150 000 with probability 0.5, a profit of \$250 000 with probability 0.2 and a loss of \$100 000 with probability 0.3. Project *B* will yield a profit of \$100 000 with probability 0.6, a profit of \$200 000 with probability 0.3 and a loss of \$50 000 with probability 0.1. Determine which project the finance director should support.

6 Some of the eggs at a market are sold in boxes of six. The number, X, of broken eggs in a box has the probability distribution given in the following table.

x	0	1	2	3	4	5	6
$\mathrm{P}(X = x)$	0.80	0.14	0.03	0.02	0.01	0	0

(a) Find the expectation and variance of X.

(b) Find the expectation and variance of the number of unbroken eggs in a box.

(c) Comment on the relationship between your answers to part (a) and part (b).

7 Find $\mathrm{E}(H)$ and $\mathrm{Var}(H)$ for the H defined in Exercise 6A Question 4.

8 The random variable X has the probability distribution given in the following table.

x	1	2	3	4	5
$\mathrm{P}(X = x)$	a	0.3	0.2	0.1	0.2

Find the values of a, μ and σ for the distribution.

9 The random variable Y has the probability distribution given in the following table.

y	2	3	4	5	6	7
$\mathrm{P}(Y = y)$	0.05	0.25	a	b	0.1	0.3

Given that $\mathrm{E}(Y) = 4.9$, show that $a = b$, and find the standard deviation of Y.

10 A game is played by throwing a fair dice until either a 6 is obtained or four throws have been made. Let X denote the number of throws made. Find

(a) the probability distribution of X, (b) the standard deviation of X.

The number of 6s obtained in the game is denoted by Y. Find $\mathrm{E}(Y)$.

If the player throws a 6 in the course of the game, then the player wins 100 points. If a 6 is not thrown, then 150 points are lost. Find the expectation of the number of points received by a player after one game.

11 The dice of Question 4 is thrown and then an unbiased coin is thrown the number of times indicated by the score on the dice. Let H denote the number of heads obtained.

(a) Show that $\mathrm{P}(H=2)=\frac{13}{48}$.

(b) Tabulate the probability distribution of H.

(c) Show that $\mathrm{E}(H)=\frac{1}{2}\mathrm{E}(X)$, where X denotes the score on the dice.

(d) Calculate $\mathrm{Var}(H)$.

8.3 The expectation and variance of a binomial distribution

Suppose that you want to find the mean and variance of the random variable X, where $X \sim \mathrm{B}(3,\frac{1}{4})$. One way would be to write out the probability distribution and calculate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ using Equations 8.1 and 8.3.

Example 8.3.1
Calculate $\mathrm{E}(X)$ and $\mathrm{Var}(X)$ for $X \sim \mathrm{B}(3,\frac{1}{4})$ from the probability distribution.

The probability distribution is found using the binomial probability formula

$$\mathrm{P}(X=x)=\binom{n}{x}p^x(1-p)^{n-x} \text{ (see Equation 7.1).}$$

x	$\mathrm{P}(X=x)$	$x\mathrm{P}(X=x)$	$x^2\mathrm{P}(X=x)$
0	$\binom{3}{0}\left(\frac{1}{4}\right)^0\left(\frac{3}{4}\right)^3=\frac{27}{64}$	0	0
1	$\binom{3}{1}\left(\frac{1}{4}\right)^1\left(\frac{3}{4}\right)^2=\frac{27}{64}$	$\frac{27}{64}$	$\frac{27}{64}$
2	$\binom{3}{2}\left(\frac{1}{4}\right)^2\left(\frac{3}{4}\right)^1=\frac{9}{64}$	$\frac{18}{64}$	$\frac{36}{64}$
3	$\binom{3}{3}\left(\frac{1}{4}\right)^3\left(\frac{3}{4}\right)^0=\frac{1}{64}$	$\frac{3}{64}$	$\frac{9}{64}$
	Totals: $\sum p_i=1$	$\sum x_ip_i=\frac{48}{64}$	$\sum x_i^2p_i=\frac{72}{64}$

Remember that p_i is the same as $\mathrm{P}(X=x_i)$.

$$\mu=\mathrm{E}(X)=\sum x_ip_i=\frac{48}{64}=\frac{3}{4}.$$

$$\sigma^2=\mathrm{Var}(X)=\sum x_i^2p_i-\mu^2=\frac{72}{64}-\left(\frac{3}{4}\right)^2=\frac{36}{64}=\frac{9}{16}.$$

There is, however, a quicker method for doing these calculations. If you consider the general case $X \sim \mathrm{B}(n,p)$, then the mean would be given by $\sum x_ip_i=\sum x\binom{n}{x}p^x(1-p)^{n-x}$.

Each term in this sum depends on the parameters n and p and so it would be reasonable to assume that E(X) also depends on n and p. Although this sum looks very complicated, it can be shown that it simplifies to np.

The working is not shown here, but intuitively you might expect this result, since in n trials

$$\text{number of successes} \approx \text{number of trials} \times \text{probability of success at a single trial}$$
$$= n \times p = np.$$

Now $\sum x_i^2 p_i$ also depends on n and p and so, therefore, does $\text{Var}(X)$. It can be shown that $\text{Var}(X) = np(1-p) = npq$, where $q = 1-p$.

For a random variable with a binomial distribution, $X \sim \text{B}(n,p)$,

$$\text{E}(X) = np, \qquad (8.4)$$
$$\text{Var}(X) = np(1-p) = npq, \text{ where } q = 1-p. \qquad (8.5)$$

Example 8.3.2
Calculate $\text{E}(X)$ and $\text{Var}(X)$ for $X \sim \text{B}(3,\frac{1}{4})$ using Equations 8.4 and 8.5.

Using Equations 8.4 and 8.5,

$$\text{E}(X) = \mu = np = 3 \times \tfrac{1}{4} = \tfrac{3}{4},$$
$$\text{Var}(X) = \sigma^2 = np(1-p) = 3 \times \tfrac{1}{4} \times \tfrac{3}{4} = \tfrac{9}{16}.$$

You can see that the second method is much quicker than the first. The formulae for calculating the mean and variance of a binomial distribution are particularly useful when n is large, as in the following example.

Example 8.3.3
Nails are sold in packets of 100. Occasionally a nail is faulty. The number of faulty nails in a randomly chosen packet is denoted by X. Assuming that faulty nails occur independently and at random, calculate the mean and standard deviation of X, given that the probability of any nail being faulty is 0.04.

Since faulty nails occur independently and at random and with a fixed probability, the distribution of X can be modelled by the binomial distribution with $n = 100$ and $p = 0.04$. Therefore

$$\text{E}(X) = \mu = np = 100 \times 0.04 = 4,$$
$$\text{Var}(X) = \sigma^2 = np(1-p) = 100 \times 0.04 \times 0.96 = 3.84.$$

Therefore $\sigma = \sqrt{3.84} = 1.96$, correct to 3 significant figures.

To calculate μ and σ using Equations 8.1 and 8.3 you have to write out a probability distribution with 101 *terms!*

The following examples give further illustrations of the use of Equations 8.4 and 8.5.

Example 8.3.4

(a) Given that $X \sim \mathrm{B}(10,0.3)$, find $\mathrm{E}(X)$ and $\mathrm{Var}(X)$.

(b) For $X \sim \mathrm{B}(10,0.3)$, calculate $\mathrm{P}(\mu-\sigma < X < \mu+\sigma)$.

(a) Since $n=10$ and $p=0.3$,

$$\mathrm{E}(X)=\mu=np=10\times 0.3=3,$$

$$\mathrm{Var}(X)=\sigma^2=np(1-p)=10\times 0.3\times 0.7=2.1.$$

(b)
$$\begin{aligned}\mathrm{P}(\mu-\sigma < X < \mu+\sigma) &= \mathrm{P}\left(\mu-\sqrt{2.1} < X < \mu+\sqrt{2.1}\right)\\ &= \mathrm{P}(3-1.44\ldots < X < 3+1.44\ldots)\\ &= \mathrm{P}(1.55\ldots < X < 4.44\ldots)\\ &= \mathrm{P}(X=2)+\mathrm{P}(X=3)+\mathrm{P}(X=4)\\ &= \binom{10}{2}0.3^2 0.7^8+\binom{10}{3}0.3^3 0.7^7+\binom{10}{4}0.3^4 0.7^6\\ &= 0.2334\ldots+0.2668\ldots+0.2001\ldots=0.7004\ldots.\end{aligned}$$

Thus $\mathrm{P}(\mu-\sigma < X < \mu+\sigma)=0.700$, correct to 3 significant figures.

Example 8.3.5

Given that $Y \sim \mathrm{B}(n,p)$, and $\mathrm{E}(Y)=24$ and $\mathrm{Var}(Y)=8$, find the values of n and p.

Using Equations 8.4 and 8.5,

$$\mathrm{E}(Y)=np=24 \quad\text{and}\quad \mathrm{Var}(Y)=np(1-p)=8.$$

Substituting the value of np from the first equation into the second equation gives:

$$24(1-p)=8, \quad\text{so}\quad 1-p=\tfrac{1}{3} \quad\text{and}\quad p=\tfrac{2}{3}.$$

Using $np=24$,

$$n=\frac{24}{p}=\frac{24}{\frac{2}{3}}=24\times\tfrac{3}{2}=36.$$

Exercise 8B

1 Given that $X \sim \mathrm{B}(20,0.14)$, calculate

(a) $\mathrm{E}(X)$ and $\mathrm{Var}(X)$, (b) $\mathrm{P}(X \leqslant \mathrm{E}(X))$.

2 A batch of capsules of a certain drug contains 2% of damaged capsules. A bottle contains 42 of these capsules. Calculate the mean and standard deviation of the number of damaged capsules in such a bottle, assuming that each capsule was randomly selected for inclusion in the bottle.

3 In a certain examination 35% of all candidates pass. Calculate the expectation and variance of the number of passes in a group of 30 randomly chosen candidates who take the examination.

4 The random variable X has a binomial distribution with mean 3 and variance 2.25. Find $\mathrm{P}(X=3)$.

5 The random variables X and Y are such that $X \sim \mathrm{B}(n,p)$ and $Y \sim \mathrm{B}(m,p)$. Given that $\mathrm{E}(X)=3$, $\mathrm{Var}(X)=2.4$ and $\mathrm{E}(Y)=2$, find $\mathrm{Var}(Y)$.

6 For the random variable Y, for which $Y \sim \mathrm{B}(16,0.8)$, calculate $\mathrm{P}(Y>\mu+\sigma)$.

Miscellaneous exercise 8

1 The number of times a certain factory machine breaks down each working week has been recorded over a long period. From these data, the following probability distribution for the number, X, of weekly breakdowns was produced.

x	0	1	2	3	4	5	6
$\mathrm{P}(X=x)$	0.04	0.24	0.28	0.16	0.16	0.08	0.04

(a) Find the mean and standard deviation of X.

(b) What would be the expected total number of breakdowns that will occur over the next 48 working weeks?

2 Some of the eggs sold in a store are packed in boxes of 10. For any egg, the probability that it is cracked is 0.05, independently of all other eggs. A shelf contains 80 of these boxes. Calculate the expected value of the number of boxes on the shelf which do not contain a cracked egg.

3 The random variable X is such that $X \sim \mathrm{B}(5,p)$. Given that $\mathrm{P}(X=0)=0.010\,24$, find the values of $\mathrm{E}(X)$, $\mathrm{Var}(X)$ and $\mathrm{P}(X=\mathrm{E}(X))$.

4 The independent random variables X and Y have the following probability distributions.

x	0	1	2	3
$\mathrm{P}(X=x)$	0.3	0.2	0.4	0.1

y	3	4	5
$\mathrm{P}(Y=y)$	0.5	0.2	0.3

Find $\mathrm{E}(X)$, $\mathrm{Var}(X)$, $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

The sum of one random observation of X and one random observation of Y is denoted by Z.

(a) Obtain the probability distribution of Z.

(b) Show that $\mathrm{E}(Z)=\mathrm{E}(X)+\mathrm{E}(Y)$ and $\mathrm{Var}(Z)=\mathrm{Var}(X)+\mathrm{Var}(Y)$.

5 A possible criterion for an outlier of a set of data is if it lies outside the interval from $\mu-2\sigma$ to $\mu+2\sigma$. For a set of observations of a random variable X, where $X \sim \mathrm{B}(20,0.4)$, determine whether the following values constitute outliers according to this criterion.

(a) 2 (b) 5 (c) 10 (d) 15

Find $\mathrm{P}(X<\mu-2\sigma)$ and $\mathrm{P}(X>\mu+2\sigma)$.

6 An absent-minded mathematician is attempting to log on to a computer, which is done by typing the correct password. Unfortunately he can't remember his password. If he types the wrong password he tries again. The computer allows a maximum of four attempts altogether. For each attempt the probability of success is 0.4, independently of all other attempts.

(a) Calculate the probability that he logs on successfully.

(b) The total number of attempts he makes, successful or not, is denoted by X (so that the possible values of X are $1, 2, 3$ or 4). Tabulate the probability distribution of X.

(c) Calculate the expectation and variance of X. (OCR)

7 A committee of six men and four women appoints two of its members to represent it. Assuming that each member is equally likely to be appointed, obtain the probability distribution of the number of women appointed. Find the expected number of women appointed.

8 The discrete random variable X takes the values $1, 2, 3, 4$ and 5 only, with the probabilities shown in the table.

x	1	2	3	4	5
$\mathrm{P}(X=x)$	a	0.3	0.1	0.2	b

(a) Given that $\mathrm{E}(X) = 2.34$. show that $a = 0.34$, and find the value of b.

(b) Find $\mathrm{Var}(X)$. (OCR)

9* The number of eggs, X, laid by the female tawny owl (*Strix aluco*) has the probability distribution given in the following table.

x	2	3	4
$\mathrm{P}(X=x)$	0.1	0.2	0.7

For any egg, the probability that it is hatched is 0.8, independently of all other eggs. Let Y denote the number of hatched eggs in a randomly chosen nest.

(a) Obtain the probability distribution of Y.

(b) Find $\mathrm{E}(Y)$ and $\mathrm{Var}(Y)$.

9 The normal distribution

This chapter investigates a very commonly occurring distribution, called the normal distribution. When you have completed it, you should

- understand the use of the normal distribution to model a continuous random variable
- be able to use the normal distribution function tables accurately
- be able to solve problems involving the normal distribution
- be able to find a relationship between x, μ and σ given the value of $\mathrm{P}(X > x)$ or its equivalent
- recall conditions under which the normal distribution can be used as an approximation to the binomial distribution
- be able to solve problems using the normal approximation, with a continuity correction.

9.1 Modelling continuous variables

In Section 6.1, you met the idea of a discrete random variable and its probability distribution. An example is given in Table 9.1, which shows the probability distribution of the outcome of a single throw of a dice. The table lists each possible value of the variable together with its associated probability.

x	1	2	3	4	5	6
$\mathrm{P}(X = x)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Table 9.1. Probability distribution of the outcome of a single throw of a dice.

Is it possible to specify the distribution of a continuous random variable in the same way? Consider the lengths, in millimetres, of 50 leaves that have fallen from a coffee tree.

60	31	72	57	99	46	68	47	54	57
42	48	39	40	67	89	70	68	42	54
52	50	85	56	50	53	57	83	79	63
63	72	57	53	90	52	58	47	34	102
70	60	94	43	85	67	78	66	57	44

Although at first sight it appears that the length of a leaf takes discrete values, this is only because the length has been measured to a given degree of accuracy. For example, when the length is given as 63 mm, it means that the length, l, of the leaf lies in the interval $62.5 \leqslant l < 63.5$. This suggests that a probability model for the length, L, of a leaf should give the probability that L lies in a certain interval rather than that L takes a particular value. Thus the approach to modelling continuous variables must be different from that used for discrete variables.

The way in which the length of the leaves is distributed can be illustrated by a histogram. First, the data are assembled into a grouped frequency table, as in Table 9.2.

Length (mm)	Frequency	Relative frequency	Class boundaries	Class width	Relative frequency density
30–39	3	0.06	29.5–39.5	10	0.006
40–49	9	0.18	39.5–49.5	10	0.018
50–59	15	0.30	49.5–59.5	10	0.030
60–69	9	0.18	59.5–69.5	10	0.018
70–79	6	0.12	69.5–79.5	10	0.012
80–89	4	0.08	79.5–89.5	10	0.008
90–99	3	0.06	89.5–99.5	10	0.006
100–109	1	0.02	99.5–109.5	10	0.002

Table 9.2. Grouped frequency table for the coffee tree leaves.

The third column of this table gives the relative frequencies: these are found by dividing each frequency by the total frequency, in this case 50. The relative frequency gives the experimental probability that the length of a leaf lies in a given interval. The fourth and fifth columns give the class boundaries and class widths respectively. The last column gives the **relative frequency density**: this is found by dividing the relative frequency by the class width. The data are illustrated by the histogram in Fig. 9.3.

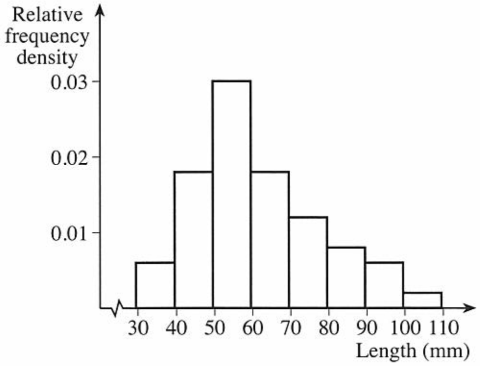


Fig. 9.3. Histogram of the data in Table 9.2.

A histogram is usually plotted with frequency density rather than *relative* frequency density on the vertical axis. The reason for using relative frequency density here is that area then represents *relative* frequency and hence experimental probability. It follows that the total area of the histogram must be equal to 1. The histogram can be used to give other experimental probabilities. For example, the shaded area in Fig. 9.4 gives the probability that the length, L, lies in the interval $54.5 \leqslant L < 71.5$. This area is equal to $5 \times 0.03 + 10 \times 0.018 + 2 \times 0.012 = 0.354$.

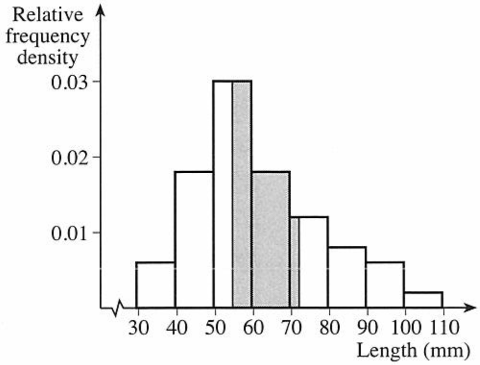


Fig. 9.4. Histogram of the data in Table 9.2.

If you want the probability that $L = 51.2$, say, the answer is zero. Although it is theoretically possible for L to equal 51.2 exactly, that is 51.200 000 ..., the probability is actually zero. This means that $\mathrm{P}(51 < L < 53) = \mathrm{P}(51 \leqslant L < 53) = \mathrm{P}(51 < L \leqslant 53) = \mathrm{P}(51 \leqslant L \leqslant 53)$. This is characteristic of continuous distributions.

The probabilities calculated from the histogram in Fig. 9.4 could be used to model the length of a coffee tree leaf. However, the model is crude; first, because of the limited amount of data and, secondly, because of the small number of classes into which the leaves are assembled and the resulting 'steps' in the histogram. Collecting more data and reducing the class width can improve the model. Table 9.5 gives the frequency table for 100 leaves, with the size of the class widths halved. This table is illustrated by the histogram in Fig. 9.6. The area of this histogram is again 1 but the steps in the histogram are smaller.

Length (mm)	Frequency	Relative frequency	Class boundaries	Class width	Relative frequency density
30–34	2	0.02	29.5–34.5	5	0.004
35–39	4	0.04	34.5–39.5	5	0.008
40–44	7	0.07	39.5–44.5	5	0.014
45–49	10	0.10	44.5–49.5	5	0.020
50–54	14	0.14	49.5–54.5	5	0.028
55–59	15	0.15	54.5–59.5	5	0.030
60–64	13	0.13	59.5–64.5	5	0.026
65–69	9	0.09	64.5–69.5	5	0.018
70–74	8	0.08	69.5–74.5	5	0.016
75–79	7	0.07	74.5–79.5	5	0.014
80–84	4	0.04	79.5–84.5	5	0.008
85–89	3	0.03	84.5–89.5	5	0.006
90–94	2	0.02	89.5–94.5	5	0.004
95–99	1	0.01	94.5–99.5	5	0.002
100–104	1	0.01	99.5–104.5	5	0.002

Table 9.5. Grouped frequency table for the lengths of 100 coffee tree leaves.

The model could be further refined by repeating the process of collecting more data and reducing the class width. If this process were to be continued indefinitely, then the outline of the histogram would become a smooth curve instead of a series of steps. The sort of curve which you might expect is shown in Fig. 9.7. Since the distribution of the leaf length is approximately symmetrical, a symmetrical curve would seem appropriate. The axis of symmetry of the curve will be positioned at the mean length. The curve gives a model for the length of a coffee tree leaf. The probability that the length of a leaf lies between the values a and b is given by the area under the curve between a and b, as shown in Fig. 9.8.

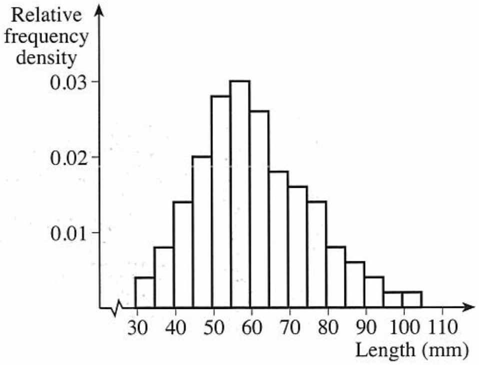


Fig. 9.6. Histogram of data in Table 9.5

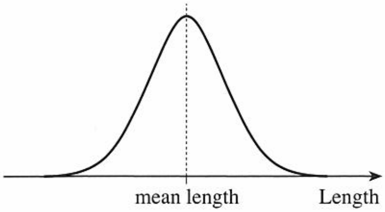


Fig. 9.7. Distribution of coffee tree leaf length.

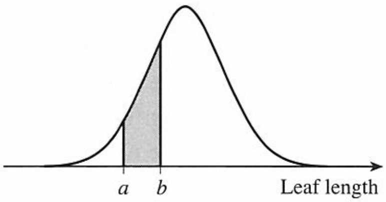


Fig. 9.8. Shaded area gives the probability that leaf length lies between a and b.

9.2 The normal distribution

The histograms of many other variables have features in common with Fig. 9.6.

(a) The distribution has a modal class somewhere in the middle of the range of values.
(b) The distribution is approximately symmetrical.
(c) The frequency density tails off fairly rapidly as the values of the variable move further away from the modal class.

This is just the sort of histogram which you would expect for the dimensions and masses of physical objects. Most of the values are close to the 'average' with just a few very large and a few very small values.

If the histogram of a variable shows these properties, then a bell-shaped curve, like that in Fig. 9.7, provides a suitable model. This model is called the **normal distribution**.

*The normal distribution is sometimes called the **Gaussian distribution** after Carl Friedrich Gauss (1777–1855), who introduced it in connection with the theory of errors. Although many continuous variables are normally distributed, many are not. Since the latter are in no way 'abnormal', the name 'Gaussian' is sometimes preferred.*

Although the curves for different variables that are normally distributed will have a similar bell shape, they will have different locations and spread. For example, the mean length of orange tree leaves will be different from that of coffee tree leaves and so will the dispersion. In order to specify the distribution completely, you need to give the mean, μ, and the variance, σ^2. The notation $X \sim \mathrm{N}(\mu,\sigma^2)$ is used to denote a continuous variable which is normally distributed with mean, μ, and variance, σ^2.

Fig. 9.9 illustrates how the appearance of the normal distribution varies for different values of the parameters, μ and σ^2. Look at the top diagram. The curve for X_1 is centred on the mean 13, while that for X_2 is centred on the mean 20. The curves have the same spread because the variance is 9 for both variables. Now look at the bottom diagram. Both curves are centred on the mean 15, but the spread for X_2 is greater than that for X_1 because it has the greater variance. The area under all the curves is the same and equal to 1. This is why the curve for X_2 is flatter than that for X_1 in the bottom diagram.

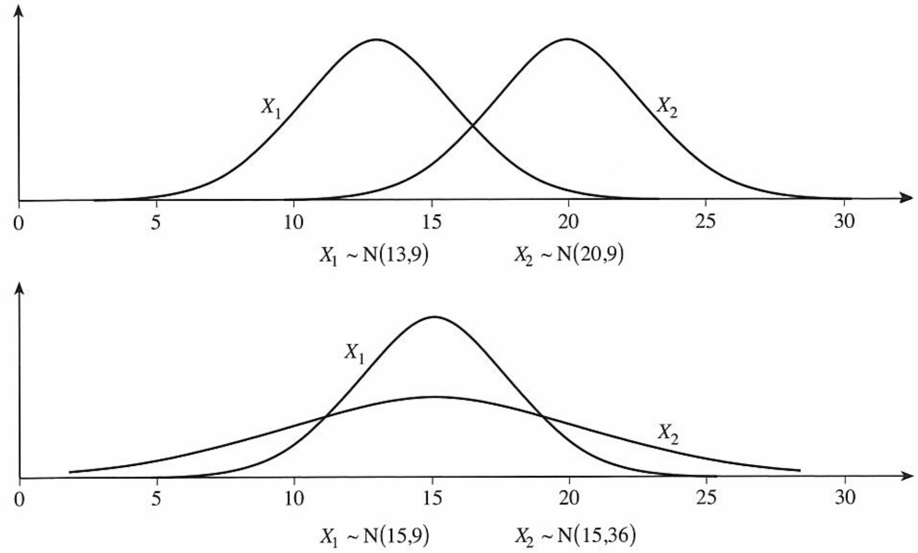


Fig. 9.9. Normal distributions with different values of μ and σ^2.

For any variable which is normally distributed, about $\frac{2}{3}$ of the values lie within 1 standard deviation of the mean, about 95% of values lie within 2 standard deviations of the mean and nearly all the values lie within 3 standard deviations of the mean. These properties are illustrated in Fig. 9.10. You can check that the diagrams in Fig. 9.9 also show these properties.

The curve of the distribution $X \sim \mathrm{N}(\mu,\sigma^2)$ can be described mathematically by the function

$$\mathrm{f}(x) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for all real values of x.

You will see that this curve extends from $-\infty$ to $+\infty$, which implies that x can take negative values. The length of a leaf, for example, can never be negative. However, the curve falls away so sharply that the probability that the normal model would predict a negative value for the length of a leaf is negligible.

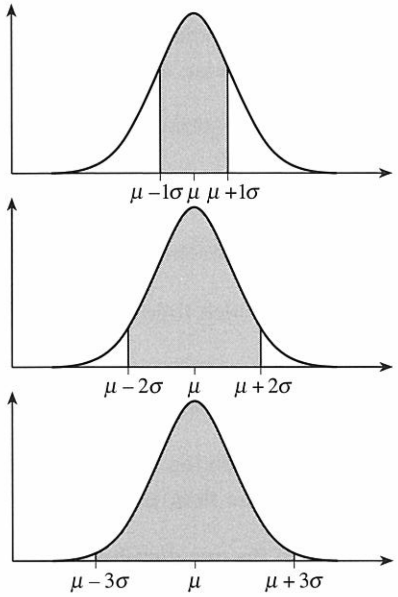


Fig. 9.10. Properties of the normal distribution: top, approximately $\frac{2}{3}$ of values lie within $\pm 1\sigma$ of the mean; middle, approximately 95% of values lie within $\pm 2\sigma$ of the mean; bottom, nearly all the values lie within $\pm 3\sigma$ of the mean.

9.3 The standard normal distribution

Probabilities for a variable which is normally distributed can be found from tables. At first sight this might appear impossible because a different table would be needed for each distribution, depending on the values of its parameters. However, it turns out that if the variable is **standardised** then one table is sufficient for all normal distributions. The standardised value, Z, is calculated from the value of the variable X by

$$Z = \frac{X - \mu}{\sigma}. \qquad (9.1)$$

You can see that Z measures the deviation of X from the mean, μ, in units of the standard deviation, σ. The random variable Z has a distribution which is $\mathrm{N}(0,1)$. This is called the **standard normal distribution**. It is illustrated in Fig. 9.11. Like all other normal distributions, about $\frac{2}{3}$ of the values lie within 1 standard deviation of the mean, that is between –1 and 1, about 95% of values lie within 2 standard deviations of the mean, that is between –2 and 2, and nearly all the values lie within 3 standard deviations of the mean, that is between –3 and 3.

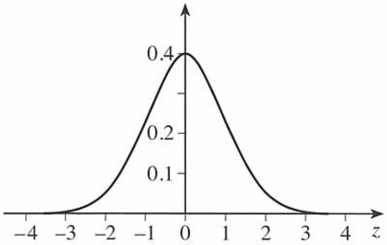


Fig. 9.11. Graph of the standard normal distribution.

$$\boxed{\text{If } X \sim \mathrm{N}(\mu,\sigma^2) \text{ and } Z = \frac{X-\mu}{\sigma}, \text{ then } Z \sim \mathrm{N}(0,1).}$$

The discussion which follows explains why $Z \sim \mathrm{N}(0,1)$.

Suppose that the random variable X has a normal distribution with parameters μ and σ^2. Then the bell-shaped curve of the normal distribution will be centred on μ.

Let $Y = X - \mu$. Then the distribution of Y will have a typical bell shape but it will be centred on 0, rather than μ, because all the values have been reduced by μ.

The spreads of the two distributions are identical, so $Y \sim \mathrm{N}(0,\sigma^2)$. Fig. 9.12 shows the relation between the distributions of X and Y.

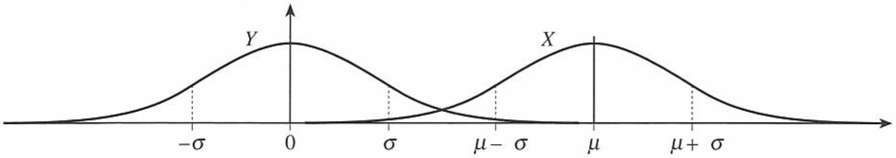


Fig. 9.12. Normal ditributions for $X \sim \mathrm{N}(\mu,\sigma^2)$ and $Y \sim \mathrm{N}(0,\sigma^2)$, where $Y = X - \mu$.

Now let $Z = \dfrac{Y}{\sigma}$. This has the effect of altering the spread of the distribution since when $Y = \pm\sigma$, then $Z = \pm 1$. Also when $Y = \pm 2\sigma$, then $Z = \pm 2$, and when $Y = \pm 3\sigma$, then $Z = \pm 3$. So, as Fig. 9.13 shows, Z has a standard normal distribution.

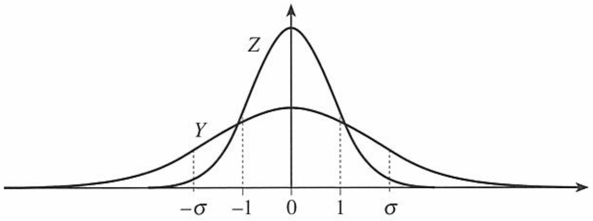


Fig. 9.13. Normal ditributions for $Y \sim \mathrm{N}(0,\sigma^2)$ and $Z \sim \mathrm{N}(0,1)$.

On page 172 there is a table of areas under the standard normal distribution. The table gives the value of $\Phi(z)$, the **normal distribution function**, where $\Phi(z) = \mathrm{P}(Z \leqslant z)$. The shaded region in Fig. 9.14 shows the area tabulated.

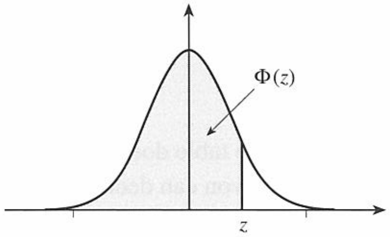


Fig. 9.14. The function $\Phi(z)$.

The symbol Φ, pronounced 'fi', is the Greek letter 'F'.

If your calculator has a routine which gives you the area under the standard normal distribution, then you should use this rather than the table.

To see what the numbers in the normal distribution function table mean, consider the value $\Phi(2)$. From the table on page 172, $\Phi(2) = 0.9772$. This means that for a random variable Z with a $\mathrm{N}(0,1)$ distribution, $\mathrm{P}(Z \leqslant 2) = 0.9772$. Fig. 9.15 illustrates this situation.

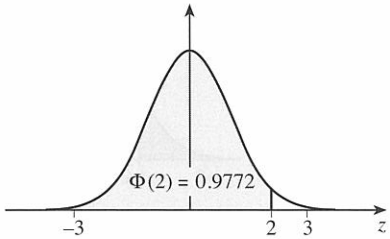


Fig. 9.15. $\Phi(2) = 0.9772$.

The table enables you to find $\mathrm{P}(Z \leqslant z)$ for values of z, given correct to 3 decimal places, from $z = 0$ up to $z = 3$. Table 9.16 is taken from the table on page 172.

z	0	1	2	3	4	5	6	7	8	9	1	2	3	4	5	6	7	8	9
0.6	0.7257	0.7291	**0.7324**	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549	3	7	10	**13**	16	19	23	26	29

Table 9.16. Extract from the normal distribution function table on page 172.

Suppose that you wish to use the table to find $\mathrm{P}(Z \leqslant 0.624)$, where $Z \sim \mathrm{N}(0,1)$. The first two digits 0.6 of 0.624 indicate that you must look in the row labelled 0.6 in the z-column. The digit in the second decimal place is 2, so now look at the entry in the first column marked 2, giving the value 0.7324. The digit in the third decimal place is 4, so you must add 13 ten-thousandths (from the '4' column on the right) to 0.7324, giving 0.7337. Thus

$$\mathrm{P}(Z \leqslant 0.624) = 0.7337.$$

To find the area between two values, $z = a$ and $z = b$, use the fact that the area between $z = a$ and $z = b$ can be written as

(area between $z = a$ and $z = b$) = (area up to $z = b$) − (area up to $z = a$).

In symbols,

$$\mathrm{P}(a \leqslant Z \leqslant b) = \mathrm{P}(Z \leqslant b) - \mathrm{P}(Z \leqslant a) = \Phi(b) - \Phi(a).$$

For example,

$$\mathrm{P}(1.20 \leqslant Z \leqslant 2.34) = \Phi(2.34) - \Phi(1.20) = 0.9904 - 0.8849 = 0.1055.$$

It is sensible to round all answers obtained from the table to 3 decimal places since the entries are given correct to 4 decimal places. So write

$$\begin{aligned}\mathrm{P}(1.20 \leqslant Z \leqslant 2.34) &= \Phi(2.34) - \Phi(1.20) = 0.9904 - 0.8849 \\ &= 0.1055 = 0.106, \text{ correct to 3 decimal places.}\end{aligned}$$

Notice that the table does not give the values of $\Phi(z)$ for negative values of z. The reason is that you can deduce the value of $\Phi(z)$ for negative z from the symmetry of the distribution.

The diagrams in Fig. 9.17 show how to calculate $\mathrm{P}(Z \leqslant -1.2)$.

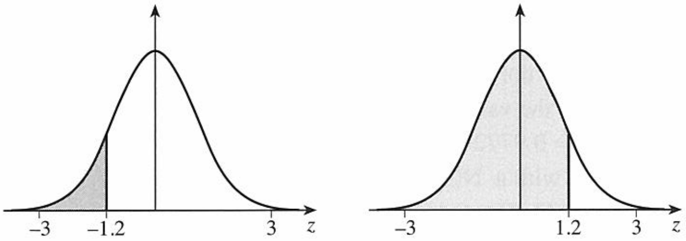


Fig. 9.17. Diagram showing how to calculate $\Phi(z)$ when z is negative.

Fig. 9.17 shows that the area shaded in the left diagram is equal to the area unshaded under the graph in the right diagram. Since the total area, shaded and unshaded, under the graph in each diagram is equal to 1,

$$\begin{aligned}\Phi(-1.2) &= 1 - \Phi(1.2) = 1 - 0.8849 \\ &= 0.1151 = 0.115, \text{ correct to 3 decimal places.}\end{aligned}$$

This is an example of the identity

$$\Phi(-z) \equiv 1 - \Phi(z),$$

which applies for all values of z.

This identity means that it is not necessary to tabulate values of $\Phi(z)$ for negative values of z.

If you are using a calculator, you may not need to use the formula $\Phi(-z) \equiv 1 - \Phi(z)$ *because the calculator gives output for negative values of* z*. But whichever method you use, table or calculator, make sure that you show your working clearly.*

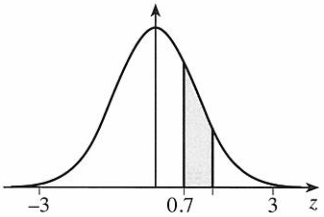
It will help to draw a sketch of the region whose area you are finding. Sketches will be drawn in this section but omitted in following sections. However, you are advised always to make a sketch.

Example 9.3.1

The random variable Z is such that $Z \sim \mathrm{N}(0,1)$. Find the following probabilities.

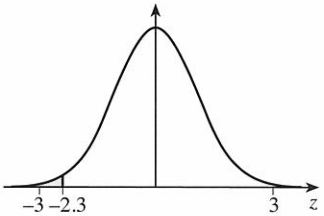
(a) $\mathrm{P}(0.7 \leqslant Z < 1.4)$ (b) $\mathrm{P}(Z \leqslant -2.3)$ (c) $\mathrm{P}(Z > 0.732)$ (d) $\mathrm{P}(-1.4 \leqslant Z \leqslant 1)$

(a) $$\begin{aligned}\mathrm{P}(0.7 \leqslant Z < 1.4) &= \Phi(1.4) - \Phi(0.7)\\ &= 0.9192 - 0.7580\\ &= 0.1612\\ &= 0.161, \text{ correct to 3 decimal places.}\end{aligned}$$

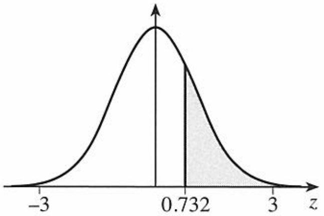


(b) Using the identity $\Phi(-z) \equiv 1 - \Phi(z)$,

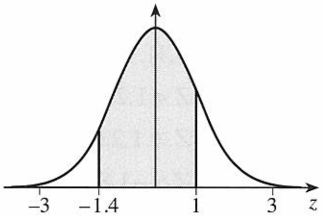
$$\begin{aligned}\mathrm{P}(Z \leqslant -2.3) &= 1 - \mathrm{P}(Z \leqslant 2.3)\\ &= 1 - \Phi(2.3)\\ &= 1 - 0.9893\\ &= 0.0107\\ &= 0.011, \text{ correct to 3 decimal places.}\end{aligned}$$



(c) $$\begin{aligned}\mathrm{P}(Z > 0.732) &= 1 - \mathrm{P}(Z \leqslant 0.732)\\ &= 1 - \Phi(0.732)\\ &= 1 - (0.7673 + 0.0006)\\ &= 1 - 0.7679 = 0.2321\\ &= 0.232, \text{ correct to 3 decimal places.}\end{aligned}$$



(d) $$\begin{aligned}\mathrm{P}(-1.4 \leqslant Z \leqslant 1) &= \mathrm{P}(Z \leqslant 1) - \mathrm{P}(Z \leqslant -1.4)\\ &= \mathrm{P}(Z \leqslant 1) - \left(1 - \mathrm{P}(Z \leqslant 1.4)\right)\\ &= \mathrm{P}(Z \leqslant 1) - 1 + \mathrm{P}(Z \leqslant 1.4)\\ &= 0.8413 - 1 + 0.9192\\ &= 0.7605\\ &= 0.761, \text{ correct to 3 decimal places.}\end{aligned}$$



A rough sketch of the graph of the $\mathrm{N}(0,1)$ *distribution with the ends of the sketch marked as –3 and 3 can give you some indication of whether your answer is approximately correct.*

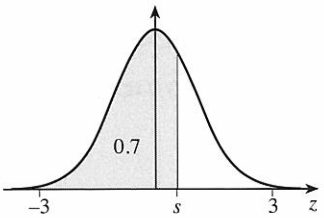
It is also sometimes necessary to use the table 'in reverse'. For example, if you know the probability of $\mathrm{P}(Z \geqslant k)$, you may need to find the corresponding value of k. As the function Φ is one–one (see P1 Section 11.6) you can always use the table in this way.

Example 9.3.2

The random variable Z is such that $Z \sim \mathrm{N}(0,1)$. Use the normal distribution function table to find

(a) the value of s such that $\mathrm{P}(Z \leqslant s) = 0.7$,

(b) the value of t such that $\mathrm{P}(Z > t) = 0.8$.

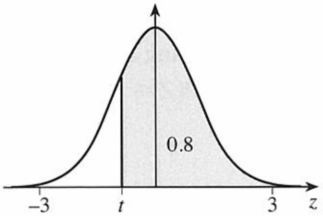


(a) In this case you need to use the table in reverse. You know that $\Phi(s) = 0.7$.

From the table,

$$\Phi(0.524) = 0.6999 \quad \text{and} \quad \Phi(0.525) = 0.7002,$$

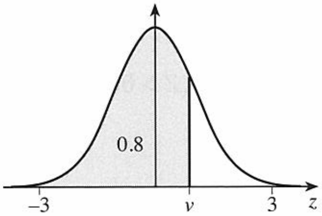
so $s = 0.524$, correct to 3 decimal places.



(b) From the diagram (*right*), it is clear that the value of t such that $\mathrm{P}(Z > t) = 0.8$ is negative.

Since the value of t is negative, it cannot be found directly from the normal distribution function table.

However, from the diagram (*below*), $t = -v$ where $\mathrm{P}(Z \leqslant v) = 0.8$, using symmetry.



You know that $\Phi(v) = 0.8$.

From the table, $\Phi(0.842) = 0.8000$, so $v = 0.842$, and hence $t = -0.842$, correct to 3 decimal places.

Exercise 9A

In this exercise use either your calculator or the table on page 172.

1 $Z \sim \mathrm{N}(0,1)$. Find the following probabilities.

(a) $\mathrm{P}(Z < 1.23)$ (b) $\mathrm{P}(Z \leqslant 2.468)$ (c) $\mathrm{P}(Z < 0.157)$

(d) $\mathrm{P}(Z \geqslant 1.236)$ (e) $\mathrm{P}(Z > 2.378)$ (f) $\mathrm{P}(Z \geqslant 0.588)$

(g) $\mathrm{P}(Z > -1.83)$ (h) $\mathrm{P}(Z \geqslant -2.057)$ (i) $\mathrm{P}(Z > -0.067)$

(j) $\mathrm{P}(Z \leqslant -1.83)$ (k) $\mathrm{P}(Z < -2.755)$ (l) $\mathrm{P}(Z \leqslant -0.206)$

(m) $\mathrm{P}(Z < 1.645)$ (n) $\mathrm{P}(Z \geqslant 1.645)$ (o) $\mathrm{P}(Z > -1.645)$

(p) $\mathrm{P}(Z \leqslant -1.645)$

2 The random variable Z is distributed such that $Z \sim \mathrm{N}(0,1)$. Find these probabilities.

(a) $\mathrm{P}(1.15 < Z < 1.35)$ (b) $\mathrm{P}(1.111 \leqslant Z \leqslant 2.222)$

(c) $\mathrm{P}(0.387 < Z < 2.418)$ (d) $\mathrm{P}(0 \leqslant Z < 1.55)$

(e) $\mathrm{P}(-1.815 < Z < 2.333)$ (f) $\mathrm{P}(-0.847 < Z \leqslant 2.034)$

3 The random variable Z is distributed such that $Z \sim \mathrm{N}(0,1)$. Find these probabilities.

(a) $\mathrm{P}(-2.505 < Z < 1.089)$

(b) $\mathrm{P}(-0.55 \leqslant Z \leqslant 0)$

(c) $\mathrm{P}(-2.82 < Z < -1.82)$

(d) $\mathrm{P}(-1.749 \leqslant Z \leqslant -0.999)$

(e) $\mathrm{P}(-2.568 < Z < -0.123)$

(f) $\mathrm{P}(-1.96 \leqslant Z < 1.96)$

(g) $\mathrm{P}(-2.326 < Z < 2.326)$

(h) $\mathrm{P}(|Z| \leqslant 1.3)$

(i) $\mathrm{P}(|Z| > 2.4)$

4 The random variable $Z \sim \mathrm{N}(0,1)$. In each part, find the value of s, t, u or v.

(a) $\mathrm{P}(Z < s) = 0.6700$

(b) $\mathrm{P}(Z < t) = 0.8780$

(c) $\mathrm{P}(Z < u) = 0.9842$

(d) $\mathrm{P}(Z < v) = 0.8455$

(e) $\mathrm{P}(Z > s) = 0.4052$

(f) $\mathrm{P}(Z > t) = 0.1194$

(g) $\mathrm{P}(Z > u) = 0.0071$

(h) $\mathrm{P}(Z > v) = 0.2241$

(i) $\mathrm{P}(Z > s) = 0.9977$

(j) $\mathrm{P}(Z > t) = 0.9747$

(k) $\mathrm{P}(Z > u) = 0.8496$

(l) $\mathrm{P}(Z > v) = 0.5$

(m) $\mathrm{P}(Z < s) = 0.0031$

(n) $\mathrm{P}(Z < t) = 0.0142$

(o) $\mathrm{P}(Z < u) = 0.0468$

(p) $\mathrm{P}(Z < v) = 0.4778$

(q) $\mathrm{P}(-s < Z < s) = 0.90$

(r) $\mathrm{P}(-t < Z < t) = 0.80$

(s) $\mathrm{P}(-u < Z < u) = 0.99$

(t) $\mathrm{P}(|Z| < v) = 0.50$

9.4 Standardising a normal distribution

The standardisation equation (9.1)

$$Z = \frac{X - \mu}{\sigma}$$

allows you to change a statement about a $\mathrm{N}(\mu, \sigma^2)$ distribution into an equivalent statement about a $\mathrm{N}(0,1)$ distribution.

To see how standardisation works, consider finding the probability $\mathrm{P}(X \leqslant 230)$, where $X \sim \mathrm{N}(205, 20^2)$

Using the standardisation equation, $Z = \frac{1}{20}(X - 205)$, you know that $Z \sim \mathrm{N}(0,1)$.

Then

$$\begin{aligned}
\mathrm{P}(X \leqslant 230) &= \mathrm{P}\left(Z \leqslant \tfrac{1}{20}(230 - 205)\right) \\
&= \mathrm{P}(Z \leqslant 1.25) \\
&= \Phi(1.25) \\
&= 0.8944 \\
&= 0.894, \text{ correct to 3 decimal places.}
\end{aligned}$$

Example 9.4.1

Given that $X \sim \mathrm{N}(4,25)$, find the following probabilities.

(a) $\mathrm{P}(X < 4.5)$ (b) $\mathrm{P}(5 \leqslant X \leqslant 6)$ (c) $\mathrm{P}(2 \leqslant X \leqslant 7)$ (d) $\mathrm{P}(X > 1)$

Let $Z = \frac{1}{5}(X-4)$. Then $Z \sim \mathrm{N}(0,1)$.

(a) $\mathrm{P}(X < 4.5) = \mathrm{P}\big(Z < \frac{1}{5}(4.5-4)\big) = \mathrm{P}(Z < 0.1)$
$= \Phi(0.1) = 0.5398$
$= 0.540$, correct to 3 decimal places.

(b) $\mathrm{P}(5 \leqslant X \leqslant 6) = \mathrm{P}\big(\frac{1}{5}(5-4) \leqslant Z \leqslant \frac{1}{5}(6-4)\big) = \mathrm{P}(0.2 \leqslant Z \leqslant 0.4)$
$= \mathrm{P}(Z \leqslant 0.4) - \mathrm{P}(Z \leqslant 0.2)$
$= \Phi(0.4) - \Phi(0.2)$
$= 0.6554 - 0.5793 = 0.0761$
$= 0.076$, correct to 3 decimal places.

(c) $\mathrm{P}(2 \leqslant X \leqslant 7) = \mathrm{P}\big(\frac{1}{5}(2-4) \leqslant Z \leqslant \frac{1}{5}(7-4)\big) = \mathrm{P}(-0.4 \leqslant Z \leqslant 0.6)$
$= \mathrm{P}(Z \leqslant 0.6) - \mathrm{P}(Z \leqslant -0.4)$
$= \Phi(0.6) - \Phi(-0.4)$
$= \Phi(0.6) - (1 - \Phi(0.4))$
$= 0.7257 - (1 - 0.6554) = 0.3811$
$= 0.381$, correct to 3 decimal places.

(d) $\mathrm{P}(X > 1) = \mathrm{P}\big(Z > \frac{1}{5}(1-4)\big) = \mathrm{P}(Z > -0.6)$
$= \mathrm{P}(Z < 0.6)$ (by symmetry)
$= \Phi(0.6) = 0.7257$
$= 0.726$, correct to 3 decimal places.

Example 9.4.2

Given that $X \sim \mathrm{N}(6,4)$, find, correct to 3 significant figures, the values of s and t such that

(a) $\mathrm{P}(X \leqslant s) = 0.6500$, (b) $\mathrm{P}(X > t) = 0.8200$.

Let $Z = \frac{1}{2}(X-6)$. Then $Z \sim \mathrm{N}(0,1)$.

(a) The statement $\mathrm{P}(X \leqslant s) = 0.6500$ is equivalent to $\mathrm{P}\big(Z \leqslant \frac{1}{2}(s-6)\big) = 0.6500$.

Therefore $\Phi\big(\frac{1}{2}(s-6)\big) = 0.6500$.

From the table,

$$\Phi(0.385) = 0.6498 \quad \text{and} \quad \Phi(0.386) = 0.6502.$$

Therefore, by interpolation, $\Phi(0.3855) = 0.6500$, so $\frac{1}{2}(s-6) = 0.3855$ giving

$s = 6 + 2 \times 0.3855 = 6.771$, which is 6.77, correct to 3 significant figures.

(b) The statement $P(X > t) = 0.8200$ is equivalent to $P\left(Z > \frac{1}{2}(t-6)\right) = 0.8200$, The problem is now similar to Example 9.3.2(b).

The value of $\frac{1}{2}(t-6)$ which you are looking for is negative. Let $v = -\frac{1}{2}(t-6)$. Then by symmetry, $P(Z \leqslant v) = 0.8200$. Therefore $v = \Phi^{-1}(0.8200)$, and from the table $v = 0.9155$.

This means that $-\frac{1}{2}(t-6) = 0.9155$. Rearranging, $t = 6 + 2 \times (-0.9155) = 4.169$, which is 4.17, correct to 3 significant figures.

Exercise 9B

You are strongly advised to draw rough sketches for these questions.

1 Given that $X \sim N(20,16)$, find the following probabilities.

(a) $P(X \leqslant 26)$ (b) $P(X > 30)$ (c) $P(X \geqslant 17)$ (d) $P(X < 13)$

2 Given that $X \sim N(24,9)$, find the following probabilities.

(a) $P(X \leqslant 29)$ (b) $P(X > 31)$ (c) $P(X \geqslant 22)$ (d) $P(X < 16)$

3 Given that $X \sim N(50,16)$, find the following probabilities.

(a) $P(54 \leqslant X \leqslant 58)$ (b) $P(40 < X \leqslant 44)$ (c) $P(47 < X < 57)$

(d) $P(39 \leqslant X < 53)$ (e) $P(44 \leqslant X \leqslant 56)$

4 The random variable X can take negative and positive values. X is distributed normally with mean 3 and variance 4. Find the probability that X has a negative value.

5 The random variable X has a normal distribution. The mean is μ (where $\mu > 0$) and the variance is $\frac{1}{4}\mu^2$.

(a) Find $P(X > 1.5\mu)$. (b) Find the probability that X is negative.

6 Given that $X \sim N(44,25)$, find s, t, u and v correct to 2 decimal places when

(a) $P(X \leqslant s) = 0.9808$, (b) $P(X \geqslant t) = 0.7704$,

(c) $P(X \geqslant u) = 0.0495$, (d) $P(X \leqslant v) = 0.3336$.

7 Given that $X \sim N(15,4)$, find s, t, u, v and w correct to 2 decimal places when

(a) $P(X \leqslant s) = 0.9141$, (b) $P(X \geqslant t) = 0.5746$, (c) $P(X \geqslant u) = 0.1041$,

(d) $P(X \leqslant v) = 0.3924$, (e)* $P(|X - 15| < w) = 0.9$.

8 Given that $X \sim N(35.4,12.5)$, find the values of s, t, u and v correct to 1 decimal place when

(a) $P(X < s) = 0.96$, (b) $P(X > t) = 0.9391$,

(c) $P(X > u) = 0.2924$, (d) $P(X < v) = 0.1479$.

9 X has a normal distribution with mean 32 and variance σ^2. Given that the probability that X is less than 33.14 is 0.6406, find σ^2. Give your answer correct to 2 decimal places.

10 X has a normal distribution, and $\text{P}(X > 73.05) = 0.0289$. Given that the variance of the distribution is 18, find the mean.

11 X is distributed normally, $\text{P}(X \geqslant 59.1) = 0.0218$ and $\text{P}(X \geqslant 29.2) = 0.9345$. Find the mean and standard deviation of the distribution, correct to 3 significant figures.

12 $X \sim \text{N}(\mu, \sigma^2)$, $\text{P}(X \geqslant 9.81) = 0.1587$ and $\text{P}(X \leqslant 8.82) = 0.0116$. Find μ and σ, correct to 3 significant figures.

9.5 Modelling with the normal distribution

The normal distribution is often used as a model for practical situations. In the following examples, you need to translate the given information into the language of the normal distribution before you can solve the problem.

Example 9.5.1
Look back to the data on lengths given in Table 9.2, and to the associated histogram in Fig. 9.3. Assuming that the distribution is normal, how many of the 50 leaves would you expect to be in the interval $59.5 \leqslant l \leqslant 69.5$?

You can check that the mean and the standard deviation of the original data on page 133 are 61.4 and 16.8, correct to 1 decimal place.

Using a random variable, L with a $\text{N}(61.4, 16.8^2)$ distribution you can calculate the expected frequency for each class.

Given that $L \sim \text{N}(61.4, 16.8^2)$, let $Z = \dfrac{L - 61.4}{16.8}$. Then $Z \sim \text{N}(0,1)$.

$$\begin{aligned}\text{P}(59.5 \leqslant L \leqslant 69.5) &= \text{P}\left(\frac{59.5 - 61.4}{16.8} \leqslant Z \leqslant \frac{69.5 - 61.4}{16.8}\right)\\ &= \text{P}(-0.113\ldots \leqslant Z \leqslant 0.482\ldots)\\ &= \Phi(0.482\ldots) - \Phi(-0.113\ldots)\\ &= \Phi(0.482\ldots) - (1 - \Phi(0.113\ldots)) \qquad \text{(using symmetry)}\\ &= 0.6851 - (1 - 0.5450) = 0.2301.\end{aligned}$$

This means that the expected frequency for the class $59.5 \leqslant l \leqslant 69.5$ is $50 \times 0.2301 = 11.5$, correct to 1 decimal place.

Therefore in a group of 50 leaves you would expect about 11 or 12 leaves to have lengths in the class $59.5 \leqslant l \leqslant 69.5$.

The observed frequency was actually 9. Does this mean that the $\text{N}(61.4, 16.8^2)$ distribution is a poor model for these data? To answer this question sensibly, you really need to calculate the expected frequencies for all eight classes.

Using the method shown above find the expected frequencies for the remaining seven classes and then review the results to consider whether the $\text{N}(61.4, 16.8^2)$ *distribution is a suitable model for these data.*

Example 9.5.2

Two friends Sarah and Hannah often go to the Post Office together. They travel on Sarah's scooter. Sarah always drives Hannah to the Post Office and drops her off there. Sarah then drives around until she is ready to pick Hannah up some time later. Their experience has been that the time Hannah takes in the Post Office can be approximated by a normal distribution with mean 6 minutes and standard deviation 1.3 minutes. How many minutes after having dropped Hannah off should Sarah return if she wants to be at least 95% certain that Hannah will not keep her waiting?

Let T be the time Hannah takes in the Post Office on a randomly chosen trip. Then $T \sim \mathrm{N}(6, 1.3^2)$.

Let t be the number of minutes after dropping Hannah off when Sarah returns; you then need to find t such that $\mathrm{P}(T \leqslant t) \geqslant 0.95$.

After standardising, this expression becomes

$$\mathrm{P}\left(Z \leqslant \frac{t-6}{1.3}\right) \geqslant 0.95 \quad \text{or} \quad \Phi\left(\frac{t-6}{1.3}\right) \geqslant 0.95.$$

Therefore $\frac{t-6}{1.3} \geqslant \Phi^{-1}(0.95) = 1.645$, which, on rearranging, gives $t \geqslant 8.1385\ldots$.

Sarah should not return for at least 8.14 minutes, correct to 3 significant figures, if she wants to be at least 95% sure that Hannah will not keep her waiting.

Example 9.5.3

A biologist has been collecting data on the heights of a particular species of cactus (*Notocactus rutilans*). He has observed that 34.2% of the cacti are below 12 cm in height and 18.4% of the cacti are above 16 cm in height. He assumes that the heights are normally distributed. Find the mean and standard deviation of the distribution.

Let the mean and standard deviation of the distribution be μ and σ respectively. Then, if H is the height of a randomly chosen cactus of this species,

$$H \sim \mathrm{N}(\mu, \sigma^2).$$

The biologist's observations can now be written

$$\mathrm{P}(H < 12) = 0.342 \quad \text{and} \quad \mathrm{P}(H > 16) = 0.184.$$

After standardising using $Z = \frac{H-\mu}{\sigma}$, these equations become

$$\mathrm{P}\left(Z < \frac{12-\mu}{\sigma}\right) = 0.342 \quad \text{and} \quad \mathrm{P}\left(Z > \frac{16-\mu}{\sigma}\right) = 0.184.$$

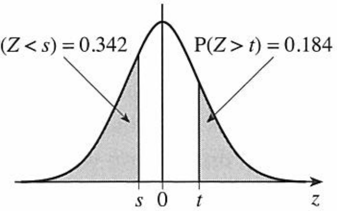
Writing $\frac{12-\mu}{\sigma} = s$ and $\frac{16-\mu}{\sigma} = t$, the two equations become

$$P(Z < s) = 0.342 \text{ and } P(Z > t) = 0.184$$

or, in terms of the normal distribution function,

$$\Phi(s) = 0.342 \quad \text{and} \quad 1 - \Phi(t) = 0.184.$$

This information is summarised in the diagram.



Using the table, after writing $s = -v$, gives $\Phi(v) = 0.658$, $v = 0.407$ and $s = -0.407$.

Since $1 - \Phi(t) = 0.184$, $\Phi(t) = 0.816$ giving $t = 0.900$.

Therefore $s = \dfrac{12 - \mu}{\sigma} = -0.407$ and $t = \dfrac{16 - \mu}{\sigma} = 0.900$.

These give the two simultaneous equations

$$12 - \mu = -0.407\sigma,$$
$$16 - \mu = 0.900\sigma.$$

Solving these equations gives $\mu = 13.2$ and $\sigma = 3.06$, correct to 3 significant figures.

9.6 Practical activities

1 Leaves Collect about 50 fallen leaves from a bush and measure the length of each leaf. You may need to consider what 'length' means, but any consistently applied definition should be satisfactory. Summarise the data in a grouped frequency table similar to Table 9.2 and calculate estimates of the mean length and the standard deviation. Use a normal distribution with the estimated mean and standard deviation to calculate expected frequencies for each class and compare these expected frequencies with the observed frequencies. Was the normal distribution a good model? You could repeat this experiment with a larger number of leaves, say 100. Does this have any effect upon your conclusions?

2 Other situations You can investigate a number of other situations to see whether or not they show normal characteristics: weight, height, foot length, finger length etc. for people of the same age and gender; masses of pebbles in sample of gravel, lengths of nails of the 'same' size; heights jumped in the Practical activities for Chapter 1; the error (plus or minus) in bisecting a line 30 cm long by eye; masses of coins of various ages (measured very accurately on a scientific balance); lengths of songs in minutes taken from a CD.

Exercise 9C

1 The time spent waiting for a prescription to be prepared at a chemist's shop is normally distributed with mean 15 minutes and standard deviation 2.8 minutes. Find the probability that the waiting time is

(a) more than 20 minutes, (b) less than 8 minutes,

(c) between 10 minutes and 18 minutes.

2 The heights of a group of sixteen-year-old girls are normally distributed with mean 161.2 cm and standard deviation 4.7 cm. Find the probability that one of these girls will have height

(a) more than 165 cm, (b) less than 150 cm,

(c) between 165 cm and 170 cm, (d) between 150 cm and 163 cm.

In a sample of 500 girls of this age estimate how many will have heights in each of the above four ranges.

3 The lengths of replacement car wiper blades are normally distributed with mean of 25 cm and standard deviation 0.2 cm. For a batch of 200 wiper blades estimate how many would be expected to be

(a) 25.3 cm or more in length, (b) between 24.89 cm and 25.11 cm in length,

(c) between 24.89 cm and 25.25 cm in length.

4 The time taken by a garage to replace worn-out brake pads follows a normal distribution with mean 90 minutes and standard deviation 5.8 minutes.

(a) Find the probability that the garage takes longer than 105 minutes.

(b) Find the probability that the garage takes less than 85 minutes.

(c) The garage claims to complete the replacements in 'a to b minutes'. If this claim is to be correct for 90% of the repairs, find a and b correct to 2 significant figures, based on a symmetrical interval centred on the mean.

5 The fluorescent light tubes made by the company Well-lit have lifetimes which are normally distributed with mean 2010 hours and standard deviation 20 hours. The company decides to promote its sales of the tubes by guaranteeing a minimum life of the tubes, replacing free of charge any tubes that fail to meet this minimum life. If the company wishes to have to replace free only 3% of the tubes sold, find the guaranteed minimum it must set.

6 The lengths of sweetpea flower stems are normally distributed with mean 18.2 cm and standard deviation 2.3 cm.

(a) Find the probability that the length of a flower stem is between 16 cm and 20 cm.

(b) 12% of the flower stems are longer than h cm. 20% of the flower stems are shorter than k cm. Find h and k.

(c) Stem lengths less than 14 cm are unacceptable at a florist's shop. In a batch of 500 sweetpeas estimate how many would be unacceptable.

7 The T-Q company makes a soft drink sold in '330 ml' cans. The actual volume of drink in the cans is distributed normally with standard deviation 2.5 ml.

To ensure that at least 99% of the cans contain more than 330 ml, find the volume that the company should supply in the cans on average.

8 The packets in which sugar is sold are labelled '1 kg packets'. In fact the mass of sugar in a packet is distributed normally with mean mass 1.08 kg.

Sampling of the packets of sugar shows that just 2.5% are 'underweight' (that is, contain less than the stated mass of 1 kg).

Find the standard deviation of the distribution.

9 The life of the Powerhouse battery has a normal distribution with mean 210 hours. It is found that 4% of these batteries operate for more than 222 hours.

Find the variance of the distribution, correct to 2 significant figures.

10 In a statistics examination, 15% of the candidates scored more than 63 marks and 10% of the candidates scored less than 32 marks. Assuming that the marks were distributed normally find the mean mark and the standard deviation.

9.7 The normal distribution as an approximation to the binomial distribution

If you were asked to estimate the probability that a school of 1000 students contains more than 150 left-handed students, how would you try to solve such a problem? Perhaps one sensible approach would be to take a reasonably large sample, say of size 50, and count the number of left-handed students in the sample. From this information you could estimate the probability that a randomly chosen student is left-handed.

For example, if your sample contains 8 left-handed people you would estimate the probability that a randomly chosen person is left-handed as $\frac{8}{50}$, or 0.16. If you define L as the number of left-handed people in a random sample of 1000 people, you could then use the distribution $\mathrm{B}(1000, 0.16)$ as a model for the distribution of L. You only need to find $\mathrm{P}(L > 150)$. This is given by

$$\begin{aligned}\mathrm{P}(L>150) &= 1-\mathrm{P}(L\leqslant 150) = 1-\mathrm{P}(L=0)-\mathrm{P}(L=1)-\mathrm{P}(L=2)-\ldots-\mathrm{P}(L=150)\\ &= 1-\binom{1000}{0}0.16^0\,0.84^{1000}-\binom{1000}{1}0.16^1\,0.84^{999}\\ &\qquad -\binom{1000}{2}0.16^2\,0.84^{998}-\ldots-\binom{1000}{150}0.16^{150}0.84^{850}.\end{aligned}$$

To calculate this is horrendous; there are 151 separate calculations to be carried out.

Fortunately there is an approximate method which uses the normal distribution and involves far less work. This section shows how to carry out such approximations.

Fig. 9.18 shows bar charts of the binomial distribution for different values of n and p.

The diagrams in the top row of Fig. 9.18 show $p = 0.1$, with three values of n, 12, 20 and 60. When $n = 12$ and $p = 0.1$ the bar chart is positively skewed. However, as n gets larger, the bar chart becomes more symmetrical and bell-shaped in appearance.

The diagrams in the bottom row of Fig. 9.18 show that when $p = \frac{1}{2}$ the shape of the bar chart resembles the bell shape that you associate with a normal distribution. This is true even when n is quite small as the diagram for $n = 12$ and $p = \frac{1}{2}$ shows.