6 In a mathematics competition, students try to find the correct answer from five options in a multiple choice exam of 25 questions. Alex decides his best strategy is to guess all the answers.

a State an appropriate model for the random variable A = the number of questions Alex gets correct.

Find the probability that the number of questions that Alex gets correct is:

b at most five **c** at least seven

d no more than three.

e Write down $E(A)$ and interpret this value.

f Find the probability that Alex scores more than expected.

g In the test, a correct answer is awarded 4 points. An incorrect answer incurs a penalty of 1 point. If Alex guesses all questions, find the expected value of his total points for the examination.

h Four students in total decide to guess all their answers. Find the probability that at least two of the four students will get seven or more questions correct.

7 Calcair buys a new passenger jet with 538 seats. For the first flight of the new jet all 538 tickets are sold. Assume that the probability that an individual passenger turns up to the airport in time to take their seat on the jet is 0.91.

a Write down the distribution of the random variable T = the number of passengers that arrive on time to take their seats, stating any assumptions you make.

b Find $P(T = 538)$ and interpret your answer.

c Find $P(T \geq 510)$ and interpret your answer.

d Calcair knows that it is highly likely that there will be some empty seats on any flight unless it sells more tickets than seats. Find the smallest possible number of tickets sold so that $P(T \geq 510)$ is at least 0.1.

e Determine the number of tickets Calcair should sell so that the expected number of passengers turning up on time is as close to 538 as possible.

f For this number of tickets sold, find $P(T = 538)$ and $P(T > 538)$. Interpret your answers.

8 You are given $X \sim B(n, p)$. Analyse the variance of X as a function of p to find the value of p that gives the most dispersion (spread) of the probability distribution.

Developing inquiry skills

Look back at the opening scenarios.

Can you solve one of the questions in the opening scenario with the binomial distribution? If so, what assumptions would you have to make?

7.7 Modelling measurements that are distributed randomly

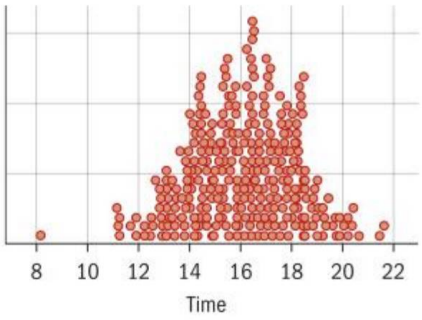
In this section you will model an example of a **continuous** random variable.

For example, consider the height Y metres of an adult human chosen at random. Recall from Section 7.5 that Y is a measurement and would therefore be an example of **continuous** data.

We use the following terminology for random variables that are found by measuring:

Terminology	Explanation
Y is a **continuous random variable.**	Continuous: Y can be found by measuring and is therefore a real number.
	Random: Y is the result of a random process.
	Variable: Y can take any value in a domain that is a subset of R.
	(Y has domain $0.67\text{ m} \leq Y \leq 2.72\text{ m}$ according to the Guinness book of world records.)

Figure 1



For example, 300 batteries are tested in a quality control exercise. The lifetime of each battery is measured to the nearest second.

The lifetime of a battery chosen at random, L, is a continuous random variable.

Figure 1 represents the 300 data points.

Figure 2

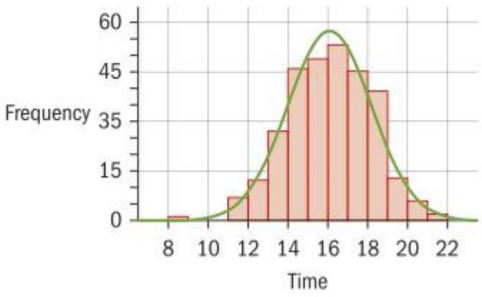


Figure 2 represents the same data set.

The frequency histogram is broadly symmetric and can be modelled by the "bell-shaped" curve shown.

Batteries lasting relatively long or short periods of time are rare.

If a set of continuous data can be modelled with this shape, we say that the data is **distributed normally** or that the data **follows a normal distribution.**

In this section you will make only **subjective** judgments of how well data follow the normal distribution. In the next chapter, you will learn techniques with which to make more objective judgments.

In the following investigation, you will explore whether some data sets are distributed normally or not.

International-mindedness

The normal curve is also known as the Gaussian curve, and is named after the German mathematician, Carl Friedrich Gauss (1777–1855), who used it to analyse astronomical data. This is seen on the old 10 Deutsche Mark notes.

Investigation 20

For this investigation, if a data set has a symmetric bell shape, consider that it follows a normal distribution. Click the icon to view the data set.

You are going to analyse W, the length of the wing from the carpal joint to the wing tip, of 4000 blackbirds. This data is in column E of the spreadsheet.

1 **Factual** Is W a discrete random variable?

Representing and exploring the data: procedure

E	F	G	H	I	J	K	L
Wing	Weight	Day	Month	Year	Time		
133	95	21	12	2006	14		114
134	106	25	11	2012	9		144
135	125	29	1	1994	9		112
135	113	5	2	1994	10		114
135	111	12	2	1994	8		116
134	105	15	2	1994	8		118
136	111	11	2	2004	8		120
127	103	23	2	2004	13		122
127	102	26	2	2004	9		124
126	104	25	2	2004	11		126
125	97	26	2	2004	9		128
135	102	27	2	2004	10		130
135	126	27	2	2004	15		132
135	121	2	3	2004	12		134
125	88	27	5	2004	14		136
123	90	8	6	2004	17		138
135	101	30	6	2004	13		140
129	100	7	9	2004	8		142
129	98	4	3	2005	16		144
129	97	6	3	2005	14		146
132	107	28	12	2005	10		

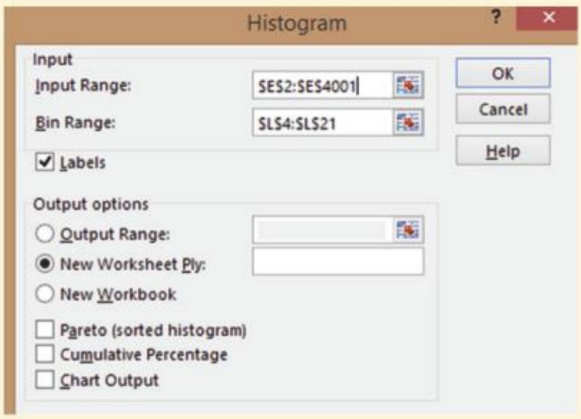
In cell L2 (green) type "$=\min(E2{:}E4001)$".

In cell L3 (blue) type "$=\max(E2{:}4001)$".

These give the maximum and minimum values of W in this sample of 4000.

Type 112, 114 and 116 as shown and drag down to 146 in cell L21.

These will give your class intervals for a histogram. These are referred to as bins by the software.



Now click on DATA on the top of your screen and Data Analysis on the far right.
Choose Histogram from the list.
Complete the dialogue box as shown.
The software will put the data into class intervals. The output will begin:

114	1	This means thatthe frequency of W in the interval $(112,\ 114]$ is 1, the frequency in $(118,\ 120]$ is 28 etc.
116	3	
118	5	
120	28	
...	...	

Use the output to create a histogram of the 4000 values of W.

2 **Factual** Does the histogram for W show a symmetric bell-shaped curve?

3 **Factual** Can it be modelled by a normal distribution?

Repeat the procedure for E, life expectancy at birth. The data for 224 countries can be found here

https://www.cia.gov/library/publications/the-world-factbook/rankorder/2102rank.html

4 **Factual** Is T a continuous random variable?

5 **Factual** Does the histogram for T show a symmetric bell-shaped curve?

6 **Factual** Can it be modelled by a normal distribution?

Repeat the procedure using other data sets that interest you.

7 **Factual** Which of these measurements do you feel could be modelled by a normal distribution?

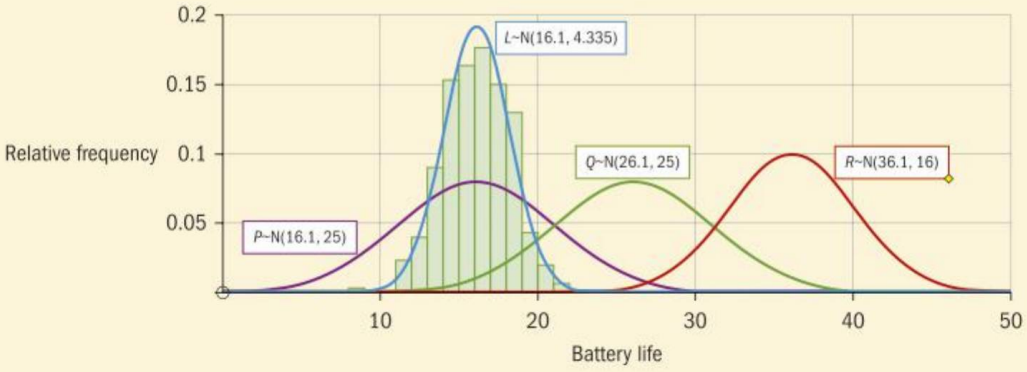
- babies' birth weight
- ages of mother of new baby
- number of hairs on head of a new baby
- number of toes on a new baby
- annual salary of adults aged between 20 and 25 years
- life expectancy in Sweden
- the number of births on each day of January 1978 in the USA
- ages of people in Haiti
- journey time of a delivery van
- heights of sunflowers
- annual salary of professional footballers
- IQ scores of 2501 undergraduate students.

8 **Conceptual** In which contexts would you expect the normal distribution to be an appropriate model?

Now that you have experienced the bell-shaped curve used to model normally distributed data in different contexts, you can learn how it is applied to data with different measures of central tendency and spread.

Investigation 21

Notation: The data in Figure 1 is distributed normally with parameters 16.1 and 2.08^2.
You write this as $L \sim \mathrm{N}(16.1,\ 2.08^2)$. The same data set is represented in the diagram below along with the curves that model three other data sets of 300 batteries sampled in the quality control exercise: $P \sim \mathrm{N}(16.1,\ 5^2)$, $Q \sim \mathrm{N}(26.1,\ 5^2)$ and $R \sim \mathrm{N}(36.1,\ 4^2)$.



Reflect on how the parameters of these normal distributions affect the location and shape of the curves.
Relate the answers to the following four questions to the parameters of the distribution.

1 **Factual** Which of the data sets has the greatest mean?

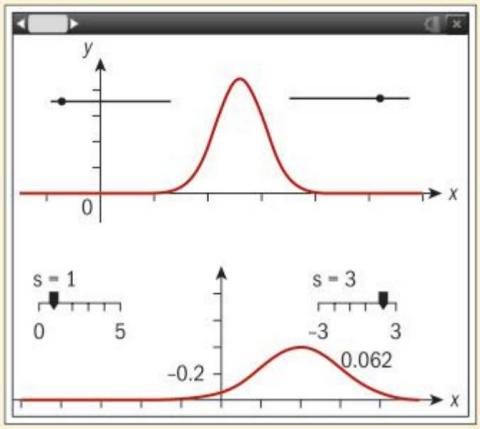
2 **Factual** Which of the data sets has the lowest mean?

Continued on next page

3 **Factual** Which of the data sets has the widest spread?

4 **Factual** Which of the data sets has the narrowest spread?

5 **Conceptual** How can you predict the axis of symmetry from $X \sim N(\mu, \sigma^2)$?



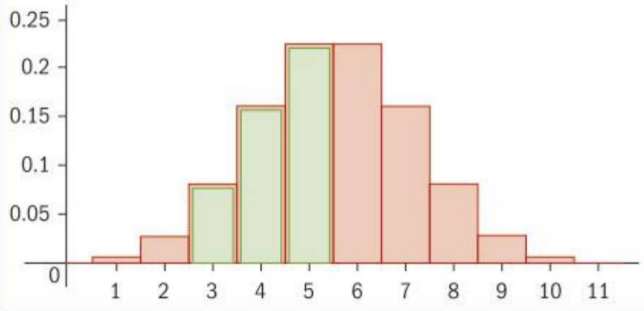
You can explore further either by using geometry software,

or by using a GDC.

In either case, alter the sliders and observe the effect that changing the parameters has on the shape of the curve.

6 **Conceptual** What do the parameters of $X \sim N(\mu, \sigma^2)$ model?

You have learned and applied **discrete** probability distributions to contexts in Sections 7.5 and 7.6. Some facts relevant to this section are summarized here:

This bar chart represents $X \sim \mathbf{B(11, 0.5)}$	Facts
Figure 3 	• The shape of the bar chart is symmetric about the mean value of 5.5. • $P(X = x)$ decreases the more the value of x differs from the mean. • Events such as $P(3 \leq X \leq 5)$ can be found from $P(X = 3) + P(X = 4) + P(X = 5)$ ie adding the heights of the bars shown. • The total of all the heights of the bars is 1, the total probability.

Notice that the probability $P(3 \leq X \leq 5)$ is related to an area in the bar chart.

Similarly, with our representation of the continuous battery data, probabilities can be found, as shown in the following investigation.

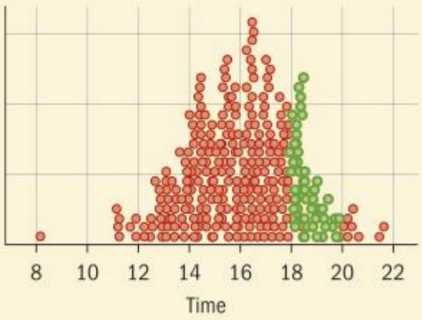
International-mindedness

French mathematicians Abraham De Moivre and Pierre Laplace were involved in the early work of the growth of the normal curve.

De Moivre developed the normal curve as an approximation of the binomial theorem in 1733 and Laplace used the normal curve to describe the distribution of errors on 1783 and in 1810 to prove the central limit theorem.

Investigation 22

Figure 4



Recall that the lifetime of a battery chosen at random is L.

1 Write down an estimate of the mean battery lifetime.

In total, there are 300 data points, of which 60 are black.

2 Write down estimates for $P(18 \le L \le 20)$, $P(12 \le L \le 14)$ and $P(14 \le L \le 18)$.

3 How does the shape of the dot plot help you make these estimates?

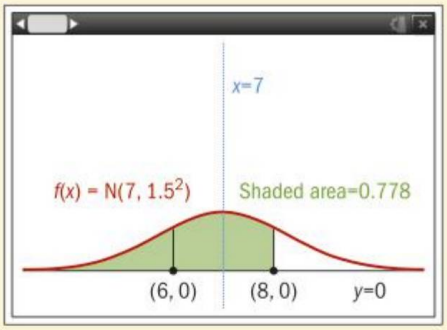
4 Is counting data points accurate and efficient in general?

Notice that Figures 3 and 4 use areas to find the probabilities. Probabilities from the normal distribution are **always** found by using technology or symmetry to find an **area** under the curve.

The total area under the normal distribution curve is 1.

For example,

Given $X \sim N(7,\ 1.5^2)$, you can find the probability $P(X \le 8) = 0.778$ as shown.



It is useful to bear in mind the symmetry of the curve in order to find related probabilities.

5 In the diagram above, why does the fact that $8 > 7$ guarantee that $P(X \le 8) > 0.5$?

6 Use the graph to find $P(X \ge 8)$, $P(X \le 6)$, $P(X \ge 6)$ and $P(6 \le X \le 8)$.

Check your answers with technology.

7 Let $X \sim N(10,\ 1.7^2)$. Use technology to find the following probabilities:

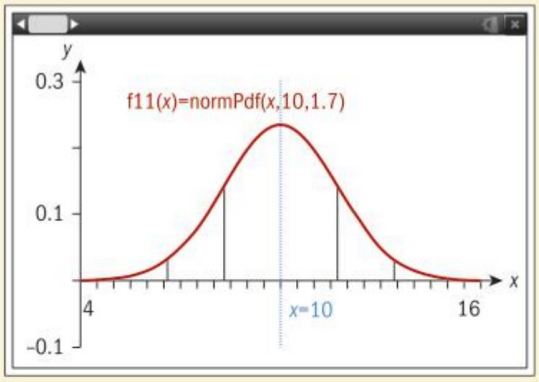
$P(10 - 1.7 \le X \le 10 + 1.7)$, $P(10 - 2 \times 1.7 \le X \le 10 + 2 \times 1.7)$, $P(10 - 3 \times 1.7 \le X \le 10 + 3 \times 1.7)$.

Copy this diagram three times, complete the labelling of the regions and shade in the areas that represent your answers.

You are given that in $X \sim N(\mu, \sigma^2)$, μ is the mean and σ is the standard deviation.

8 Repeat **7** with your own values of μ and σ. What do you notice? Write your findings in a table and compare your results with those of others in your class.

If $X \sim N(\mu, \sigma^2)$, then:		
$P(\mu - \sigma \le X \le \mu + \sigma) =$		Approximately ___% of the data lies within _ standard deviations of the mean
$P(\mu - 2\sigma \le X \le \mu + 2\sigma) =$		Approximately ___% of the data lies within _ standard deviations of the mean
$P(\mu - 3\sigma \le X \le \mu + 3\sigma) =$		Approximately ___% of the data lies within ___ standard deviations of the mean



9 How do you know when you can find normal distribution probabilities without technology?

10 How do you know when you must find normal distribution probabilities with technology?

You can now apply your knowledge and understanding of the normal distribution to find probabilities and to solve problems.

Example 26

T is the waiting time in seconds for a customer care representative of TechCo to respond to a customer's first question in an online chat session. You are given that $T \sim N(19.1, 3^2)$. Find:

a $P(T < 17)$

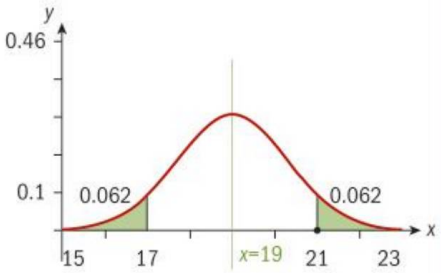
b $P(T \leq 21)$

c $P(T \geq 20.3)$

d the expected number of times the waiting time is less than 21 seconds in a sample of 107 chats collected by TechCo management.

a $P(T < 17) = 0.0620$

b The answer to **a** can be represented on a sketch:



Noticing that 17 and 19 are both 2 units from the mean enables you to use the symmetry of the curve.

Hence $P(T > 21) = P(T < 17)$

So $P(T \leq 21) = 1 - P(T < 17)$
$= 0.938$

Apply the symmetry of the curve and complementary events. Note that $P(T \leq 21) = P(T < 21)$. In fact, $<$ and $\leq$ and $>$ and $\geq$ are interchangeable in events with continuous random variables.

c $P(T \geq 20.3) \approx \frac{1 - 0.68}{2} = 0.16$

This follows from applying the fact that 20.3 is one standard deviation from the mean.

d $107 \times 0.938 = 100.36$

Approximately 100 chats are predicted to have a waiting time of less than 21 seconds for the first question to be responded to.

Apply and interpret the formula for the expected number of occurrences.

TOK

Do you think that mathematics is a useful way to measure risks?

To what extent do emotion and faith play a part in taking risks?

Example 27

The lengths of trout in a fish farm are normally distributed with a mean of 39 cm and a standard deviation of 6.1 cm.

a Find the probability that a trout caught in the fish farm is less than 35 cm long.

b Cliff catches five trout in an afternoon. Find the probability that at least two of the trout are more than 35 cm long. State any assumptions you make.

c Find the probability that a trout caught is longer than 42 cm given that it is longer than 40 cm.

d Determine if the events $L > 42$ and $L > 40$ are independent.

a Let L represent the length of a randomly selected trout. Then $L \sim N(39, 6.1^2)$.	Write down the random variable, the distribution and the event to clarify your thoughts and to demonstrate your knowledge and understanding.
$P(L < 35)$ can be found on your GDC:	Use technology to find the probability and give the answer to three significant figures.
$P(L < 35) = 0.256$	
b Let C represent how many of the five fish are more than 35 cm long.	Five fish caught can be represented as five trials.
Then $C \sim B(5, 0.744)$, assuming that the length of each fish caught is independent of the others.	Write down the distribution, using
$P(C \geq 2)$ can be found on your GDC:	$1 - 0.256 = 0.744$ as the probability of success.
$P(C \geq 2) = 0.983$	
c $P(L > 42 \mid L > 40) = \dfrac{P(L > 42)}{P(L > 40)}$	Apply the formula for conditional probability.
$P(L > 42 \mid L > 40) = 0.716$	$P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
Since $P(L > 42 \mid L > 40) = 0.716$ and $P(L > 42) = 0.311$, the events are not independent.	Apply the definition of independent events. If $P(A \mid B) = A$ then A and B are independent.

Exercise 7M

1 The length of a local bus journey from Anrai's home to school is a normally distributed random variable T_L with mean 31 minutes and standard deviation 5 minutes. Sketch the following events on three separate diagrams:

a $P(T_L \geq 31)$ **b** $P(29 \leq T_L < 32)$

c $P(T_L < 36)$.

2 The length of an express bus journey from Anrai's home to school is a normally distributed random variable T_E with mean 25 minutes and standard deviation 3 minutes. Sketch the following events on three separate diagrams:

a $P(T_E \geq 31)$ **b** $P(29 \leq T_E < 32)$

c $P(T_E < 36)$.

3 A large sample of bottles of shampoo are inspected. The contents of the bottles S is distributed normally with mean 249 ml and standard deviation 3 ml.

a Sketch a diagram to represent this information.

b Hence estimate the probability that a randomly selected bottle of shampoo will contain less than 246 ml.

c Verify your answer using technology.

d The bottle is labelled "Contents 250 ml". Predict the number of bottles in a sample of 200 that will contain at least the amount claimed.

4 Data is collected from a large number of rush hour commuter trains.

The time T for all the passengers to board a train is distributed normally with mean 186 seconds and standard deviation 14 seconds.

a Sketch a diagram to represent this information.

b Hence estimate the probability that for a randomly selected commuter train it will take at least 214 seconds for all passengers to board.

c Verify your answer using technology.

d Hence predict the number of rush hour trains in a sample of 176 that will take longer than 200 seconds to be fully boarded.

5 $T \sim N(17.1, 3.1^2)$. Estimate these probabilities without technology:

a $P(T < 17.1)$ **b** $P(T < 14)$

c $P(T > 20.2)$ **d** $P(14 \leq T < 23.3)$

e $P(T < 7.8)$ **f** $P(T < 23.3 | T > 20.2)$.

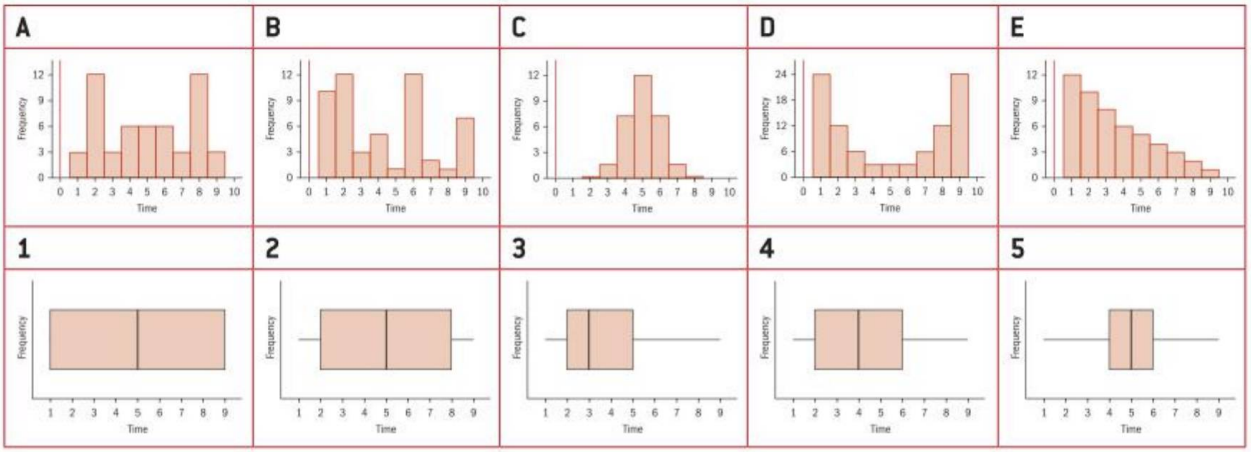
6 $Q \sim N(4.03, 0.7^2)$. Find these probabilities with technology:

a $P(Q < 4)$ **b** $P(Q < 3.4)$

c $P(Q > 5)$ **d** $P(3.5 \leq Q < 4.5)$

e $P(Q < 4.9 | Q > 2.9)$.

7 **a** Match these five histograms with the correct box-and-whisker diagram.



b Identify the one histogram of a data set that follows the normal distribution.

c State, with a reason, which statement is true:

p: A data set whose histogram is symmetric can be represented by a symmetric box-and-whisker diagram.

q: A data set that follows the normal distribution must have a symmetric box-and-whisker diagram.

r: A data set with a symmetric box-and-whisker diagram must be normally distributed.

In this section you have learned how to find probabilities of the form $P(a \leq X \leq b)$ where X is a normally distributed random variable and the values of a and of b are **known**.

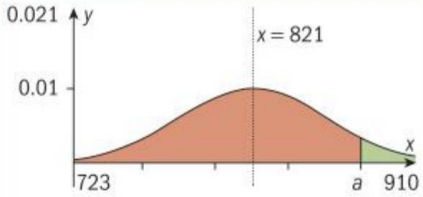
It is also possible to find the values of a and of b if, instead, $P(a \leq X \leq b)$ is known. This reverse process is illustrated in the next example.

International-mindedness

Belgian scientist Lambert Quetelet applied the normal distribution to human characteristics (l'homme moyen) in the 19th century. He noted that characteristics such as height, weight and strength were normally distributed.

Example 28

The weights W of cauliflowers purchased by a supermarket from their suppliers are distributed normally with mean 821 g and standard deviation 40 g. The heaviest 8% of cauliflowers are classified as oversized and re-packaged. Find the range of weights of cauliflowers classified as oversized. Express your answer correct to the nearest gram.



A sketch helps you orientate your reasoning in the correct place and is a good problem-solving strategy.

$P(W < a) = 1 - 0.08 = 0.92$

You can see already that the lower limit for classification as oversized, labelled a, must be greater than 821.

Use complementary events to find $P(W < a)$.

Cauliflowers weighing at least 877 g will be classified as oversize.

This gives you the input necessary for the inverse normal function on your GDC.

Interpret the result to state the range.

Exercise 7N

1 The weight of a bag of rice is distributed normally with mean 998 g and standard deviation 10 g. It is known that 20% of the packs weigh less than r g.

Find the value of r.

2 The weight of a pack of three bananas is distributed normally with mean 372 g and standard deviation 13 g. It is known that 17% of the packs weigh more than t g.

Find the value of t.

3 The weight of bags W measured in grams at an airport is distributed normally with mean 22 129 g and standard deviation 300 g.

a Identify the correct statement(s) about the upper quartile of the weights of the bags Q_3:

s : Q_3 can be found by solving $P(W > Q_3) = 0.75$.

t : Q_3 can be found by solving $P(W > Q_3) = 0.25$.

u : Q_3 can be found by solving $P(W < Q_3) = 0.75$.

b Hence calculate the interquartile range of W.

4 The speeds of cars passing a point on a highway are analysed by the police force. It is found that the speeds follow a normal distribution with mean 115.7 km/h and standard deviation 10 km/h.

- **a** Find the probability that a car chosen at random will be travelling between 110 km/h and 120 km/h.
- **b** A sample of eight cars is taken. Find the expected number of the sample that are travelling between 110 km/h and 120 km/h.
- **c** Find the probability that in the sample of eight, more than five cars are travelling between 110 km/h and 120 km/h. State the assumptions you must make.

5 An electronics company produces batteries with a lifespan that is normally distributed with a mean of 182 days and a standard deviation of 10 days.

- **a** Find the probability that a randomly selected battery lasts longer than 190 days.
- **b** In a sample of seven batteries chosen for a quality control inspection, find the probability that no more than three of them last longer than 190 days.
- **c** If a battery is guaranteed to last up to 165 days, find the probability that the battery will cease to function before the guarantee runs out.
- **d** Hence predict the number of batteries in a batch of 10000 that would not last the duration of the guarantee.

6 The distance travelled to and from work each day by employees in a central business district is modelled by a normal distribution with mean 16 km and standard deviation 5 km.

- **a** Find the probability that a randomly chosen employee travels between 13 km and 15.3 km each day.
- **b** 13% of employees travel more than x km each day to and from work. Find the value of x.
- **c** Records show that when snow falls, 91% of employees who live further than 14 km from the central business district will fail to get to work. Predict how many of the 23109 employees will fail to get to work on a snow day.

7 A nurse has a daily schedule of home visits to make. He has two possible routes suggested to him by an app on his phone for the journey to his first patient, Nur. Assume that the journey times are normally distributed in each case.

Route A has a mean of 42 minutes and a standard deviation of 8 minutes.

Route B has a mean of 50 minutes and a standard deviation of 3 minutes.

- **a** Identify the advantages and disadvantages of each route.
- **b** The nurse starts his journey at 8.15am and must be at Nur's house by 9.00am. State the route he should take.
- **c** If on five consecutive days, the nurse leaves home at 8.15am and takes route A, find the probability that he arrives at Nur's house:
  - **i** by 9.00am on all five days
  - **ii** by 9.00am on at least three of the five days
  - **iii** by 9.00am on exactly three consecutive days.

8 Catarina finds a set of ages X measured in years that follow a normal distribution with mean 70 years and variance 25 years. She represents the data with a box-and-whisker diagram.

- **a** Calculate the upper quartile of X.
- **b** Hence determine whether the length of the box represents more than, less than or equal to 10 years.

Developing your toolkit

Now do the Modelling and investigation activity on page 376.