15 Statistics

6.1 Concepts of population, sample, random sample and frequency distribution

6.2 Presentation of data: frequency tables and diagrams, box and whisker plots

Grouped data: mid-interval values, interval width, upper and lower interval boundaries, frequency histograms

6.3 Mean, median, mode; quartiles, percentiles

Range; interquartile range; variance; standard deviation

6.4 Cumulative frequency; cumulative frequency graphs; use to find median, quartiles, percentiles

15.1 Populations and samples

A collection of objects is often called a **population** in statistical terms. Any subset of objects is called a **sample** from the population. In real life such samples are very common and are used to make estimates about the whole population. For example, in the run-up to an election, samples will be drawn from the voting population to make predictions about what the final election outcome might be.

In statistics it is commonplace to lack certain information about a population. Taking a sample enables **estimation** to take place. For example the mean of a population may not be known. By taking a sample and measuring characteristics of the sample, an estimation can be made about the characteristics of the population.

A **random sample** from a finite population is a selection of objects from the population such that all possible selections of the objects are equally likely to be chosen.

15.2 Diagrammatic representation of data

There are three common types of data: qualitative, discrete and continuous. **Qualitative** data consists of descriptions using names, for example red, white, blue. **Discrete** data consists of exact numerical values, for example 2, 3, 4. **Continuous** data consists of numerical values in cases where it is not possible to make a list of all the outcomes, for example, measurements such as weight, time and length.

For continuous data, you can use ranges of values such as 0–4, 5–8, 9–12, etc.

Lists of numbers can be difficult to understand and interpret straight away. Sometimes it is not easy to just look at a set of data and make some immediate conclusions. Therefore it is often useful to draw tables and diagrams to show what is happening. Some ways of representing data are tally charts, stem and leaf diagrams, frequency tables, bar charts, pictograms, line graphs and pie charts. Most of these should be familiar to you already, and they are summarised here.

Tally charts and frequency tables

Data recorded in a list can sometimes be hard to work with. For example, look at the number of cars sold per week by a garage over a 20-week period:

10, 12, 11, 10, 14, 8, 9, 10, 12, 15,
14, 11, 12, 13, 12, 12, 10, 9, 13, 12

It is not easy to see which observation is the most common from the raw data. A **tally chart** provides a simple summary.

	Tally
8	|
9	||
10	||||
11	||
12	|||| |
13	||
14	||
15	|

The diagonal line/tally used against the observation 12 is used to group each set of five observations. This is often referred to as a five-bar gate.

You can use the tally chart to draw up a **frequency table.**

No. of cars sold in a week	Tally	Frequency
8	|	1
9	||	2
10	||||	4
11	||	2
12	|||| |	6
13	||	2
14	||	2
15	|	1
	Total	20

Frequency tables are the most commonly used method of storing large amounts of data.

Note that the total of the frequencies is 20, the total number of observations.

Stem and leaf diagrams

Tally charts are not always suitable if the range of possible values is very large. In these cases a **stem and leaf diagram** summarises the data better. The stem represents the most significant figure and the leaves are the less significant figures. For example, the ages of people staying in a hotel on one particular night are:

10, 13, 12, 20, 9, 21, 14, 32, 21, 6,
10, 56, 41, 4, 51, 12, 33, 8, 31, 23

This can be summarised as:

If you order the data you get:

stem (tens)	leaves (units)
0	9 6 4 8
1	0 3 2 4 0 2
2	0 1 1 3
3	2 3 1
4	1
5	6 1

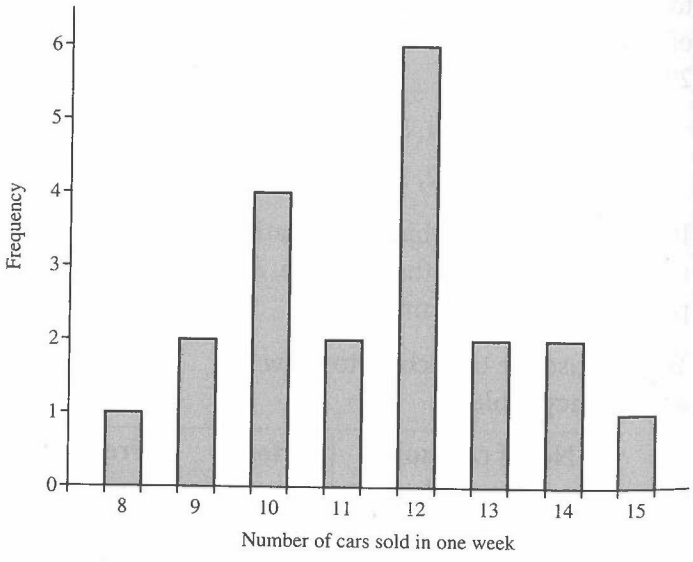
$\Rightarrow$

stem (tens)	leaves (units)
0	4 6 8 9
1	0 0 2 2 3 4
2	0 1 1 3
3	1 2 3
4	1
5	1 6

The stem is the tens and the leaves are the units.
0 | 9 means 9
3 | 2 means 32

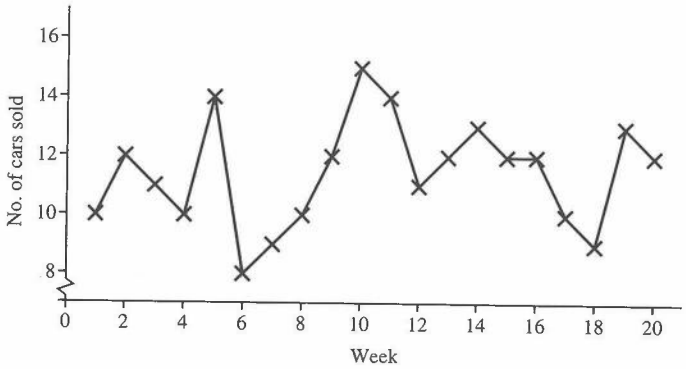
Bar charts

You can represent the number of cars sold per week by a garage over a 20-week period, as recorded in the tally chart on page 385, in a bar chart. The lengths of the bars are proportional to the number of observations of each outcome.



Line graphs

In a line graph, the data points are plotted on axes and joined by straight lines. The car sales data are shown here as a line graph.



The line graph shows how sales vary over time.

Pictograms

Pictograms are bar charts in which the bars are replaced by symbols (pictures) representing the subject of the data. They are less precise than bar charts.

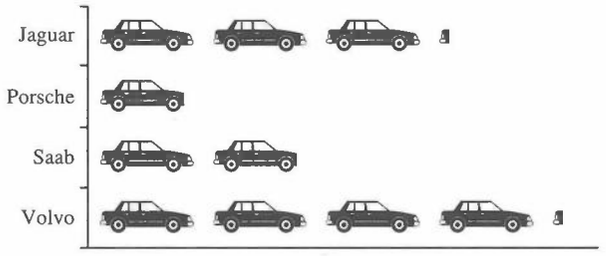
Example 1

The table shows January sales figures for four manufacturers of cars.

Manufacturer	Jaguar	Porsche	Saab	Volvo
Number of cars	1550	466	925	2050

Represent the data using a suitable pictogram.

Using a picture of a car to represent 500 car sales, the pictogram is as shown.



Pie charts

In a pie chart, the areas of the portions of the pie are in proportion to the quantities being represented.

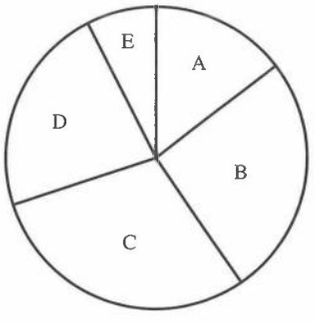
Example 2

The table shows the number of 240 students who achieved grades A to E in their mathematics examinations in a particular year.

Grade	A	B	C	D	E
Number of students	35	62	71	54	18

Represent these data in a pie chart.

First convert the frequencies to angles. These are the angles that represent each grade as a proportion of the pie chart.



Grade	**Number of students**	**Angle**
A	35	$\frac{35}{240} \times 360° = 52.5°$
B	62	$\frac{62}{240} \times 360° = 93°$
C	71	$\frac{71}{240} \times 360° = 106.5°$
D	54	$\frac{54}{240} \times 360° = 81°$
E	18	$\frac{18}{240} \times 360° = 27°$
Total	240	360°

Exercise 15A

1 Here are the points scored by the final 30 competitors in a talent contest:

74, 75, 77, 82, 80, 77, 79, 79, 81, 82, 82, 82, 79, 80, 78, 81, 86, 79, 80, 81, 81, 83, 83, 84, 81, 84, 80, 80, 77, 81

a) By first drawing a tally chart, obtain a frequency table.

b) Represent these scores on a bar chart.

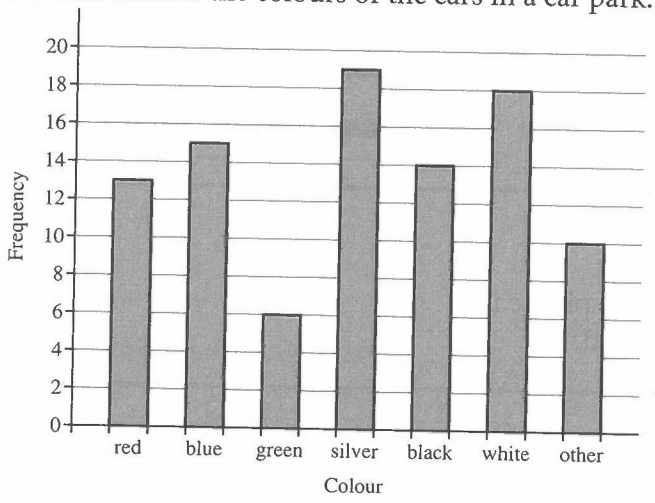
2 Here are the numbers of puppies in 40 different litters.

8, 6, 7, 2, 10, 5, 7, 3, 3, 4, 8, 8, 4, 5, 7, 5, 6, 5, 9, 5, 6, 7, 7, 5, 7, 4, 7, 8, 8, 8, 5, 12, 4, 9, 7, 10, 5, 12, 2, 6

a) By first drawing a tally chart, obtain a frequency table.

b) Represent these scores on a bar chart.

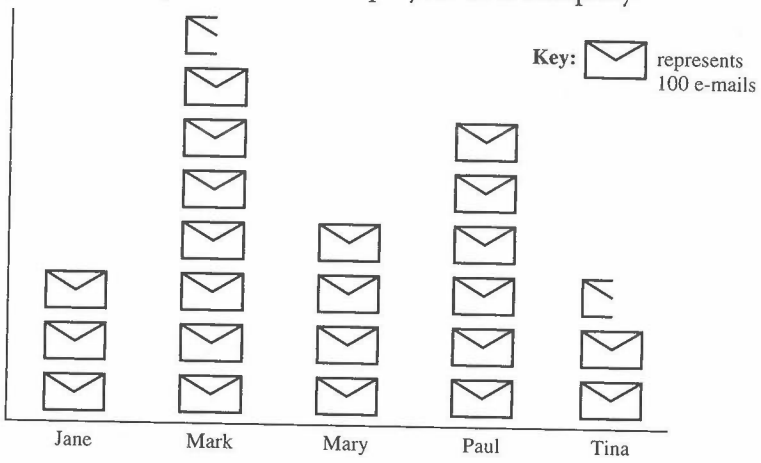
3 The bar chart shows the colours of the cars in a car park.



a) How many cars were in the car park?

b) What percentage of the cars were silver?

4 The pictogram shows the number of e-mails received in a given week by each of five employees of a company.



a) How many e-mails did Mark receive?

b) How many e-mails were received in total by the five employees?

5 Here are the marks of a group of students in a Chemistry exam.

45, 56, 67, 73, 82, 91, 67, 85, 94, 88, 65, 76, 48, 62, 84,
57, 74, 88, 73, 93, 82, 45, 58, 72, 81, 60, 91, 73, 77, 81

Construct a stem and leaf diagram to summarise these marks.

6 The points scored by a rugby team in all its matches in one season are given here.

23, 32, 15, 27, 0, 29, 34, 46, 17, 8,
33, 63, 24, 12, 48, 37, 26, 0, 36, 41

Construct a stem and leaf diagram to summarise these results.

7 The table shows the number of tubers under 50 potato plants.

Number of tubers	0	1	2	3	4	5	6	7
Frequency	6	2	5	8	14	9	4	2

Construct a line graph to summarise these results.

8 The frequency table shows the age at which a group of mothers gave birth to their first child.

Age	16–20	21–25	26–30	31–35	36–40	41–45
Frequency	7	12	9	7	3	2

Using mid-interval values, construct a line graph to summarise these results.

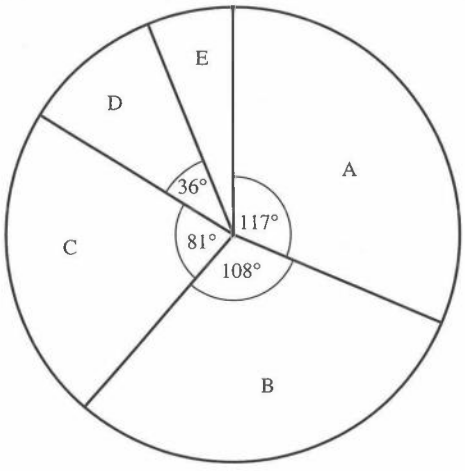
The mid-interval value for the class 16–20 is 18.

9 The shoe sizes of the children in a class are:

4, 5, 7, 6, 4, 6, 7, 7, 8, 5, 4, 7, 6, 4, 7, 8, 6, 5, 7, 4

Construct a pie chart to represent these data.

10 In a drawing competition all the exhibits are awarded a grade from A to E. The results are summarised in the pie chart.



a) Given that 13 drawings were awarded an A, how many were awarded a C?

b) How many drawings were in the competition altogether?

15.3 Mode, median, mean

There are three common ways of measuring the typical, or **average**, value of a set of data. These averages are called the mean, median and mode.

The mode

The **mode** of a set of data is the single value that occurs most often. If two outcomes occur with the greatest frequency then there is no unique mode. The data are described as being **bimodal**. If there are three or more such outcomes then the data are described as being multimodal.

Example 1

A fair die is thrown ten times and the following results obtained:

5, 4, 5, 1, 3, 5, 6, 2, 1, 4

What is the modal score?

.....

The score 5 occurs the most (three times) and therefore the modal score is 5.

In the case of continuous data, the modal class is the class which has the greatest frequency.

Example 2

The results of a survey of the age of 110 cars passing a particular point are given in the table.

Age (years)	1–2	2–3	3–4	4–5	5–6
Frequency	15	27	36	21	11

What is the modal class?

.....

The greatest frequency is 36, corresponding to 3–4 years.
The modal class is 3–4 years.

When the data are grouped into classes (see page 392), you can't tell exactly what the mode is, only which class it is in.

The median

The **median** is the middle value when all the observations or outcomes are arranged in order of magnitude.

Example 3

Find the median of these sets of data.

a) In an office block the amount spent on lunch by a cross-section of office workers on a particular Friday was recorded as €4, €14, €2, €6, €6, €4, €24, €10, €12, to the nearest euro.

b) The ages of university students in a tutorial group were recorded as 20, 23, 18, 19, 28, 26, 22, 18.

.....

a) Arrange in order of magnitude:

2, 4, 4, 6, 6, 10, 12, 14, 24

The middle value is 6 and therefore the median is €6.

b) Arrange in order of magnitude:

18, 18, 19, 20, 22, 23, 26, 28

In this case no one number is in the middle as there is an even number of observations. The observations 20 and 22 are in the middle. The median is taken to be

$\frac{1}{2}(20 + 22) = 21$

The mean

The **mean** is the sum of all the observed values divided by the total number of observations. It is written as $\bar{x}$ and calculated by the formula:

$$\bar{x} = \frac{1}{n}\sum x$$

This is also called the arithmetic mean.

where x_i represents the observed values/outcomes.

Example 4

Calculate the mean of the set of data 3.7, 4.2, 4.0, 2.5 (cm).

The mean is given by

$$\bar{x} = \frac{3.7 + 4.2 + 4 + 2.5}{4} = \frac{14.4}{4} = 3.6$$

The mean of the data is 3.6 cm.

The mean of a frequency distribution

Very often, data are presented in the form of a **frequency distribution**. For example, the lengths of stems of a group of 35 plants were recorded, to the nearest 5 cm, and the data recorded in a table.

Length (x cm)	10	15	20	25
Frequency (f)	6	12	13	4

The table shows that there were 6 plants whose stems were 10 cm long, to the nearest 5 cm. There were 12 plants whose stems were 15 cm long, to the nearest 5 cm, and so on.

Notice that the total number of observations/outcomes is the total of the frequencies:

$n = 6 + 12 + 13 + 4 = 35$

To find the sum of the observed values, draw the table vertically and add another column:

Length (x cm)	Frequency (f)	Total for each group ($x \times f$)
10	6	$10 \times 6 = 60$
15	12	$15 \times 12 = 180$
20	13	$20 \times 13 = 260$
25	4	$25 \times 4 = 100$
	$\Sigma f_i = 35$	$\Sigma f_i x_i = 600$

Therefore

$$\bar{x} = \frac{60 + 180 + 260 + 100}{35} = \frac{600}{35} = 17.1 \text{ (3 s.f.)}$$

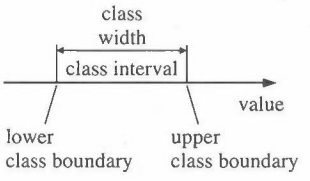
There is an important distinction between the values x representing observations of an entire population and the observations x representing a sample. For example, the observations may be the set of measurements relating to the entire workforce in a factory – the whole population. Alternatively, the observations may relate to only 10 employees and will be used as a sample from which conclusions can be drawn about the entire workforce.

For a frequency distribution, the mean is given by

$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

Grouped frequency tables

When you have a lot of numerical data, it is helpful to group it into classes or **class intervals**. Each class interval lies between an **upper class boundary** and a **lower class boundary**.



For example, consider the following class intervals representing lengths in cm:

0–10 10–20 20–30 30–40

For the class interval 30–40, 30 is the lower class boundary, 40 is the upper class boundary, and the class width is 10.

Example 5

The heights of flowers in a flower bed were recorded and the grouped frequency table summarises the results.

Length (cm)	0–10	10–20	20–30	30–40
Number of flowers	28	34	17	6

Find an estimate of the mean.

Redraw the table vertically and add additional columns for further calculations. You do not know the actual height of every flower, only that 28 of them have heights in the range 0–10 cm. So to calculate an approximate value for the mean you must use the mid-point value of each interval.

Length (cm)	Mid-point (x)	No. of flowers (f)	$x \times f$
0–10	5	28	140
10–20	15	34	510
20–30	25	17	425
30–40	35	6	210
		$n = 85$	$\Sigma f_i x_i = 1285$

The mean is given by

$$\bar{x} = \frac{\Sigma f_i x_i}{n} = \frac{140 + 510 + 425 + 210}{85} = \frac{1285}{85} = 15.1$$

The mean length is 15.1 cm.

For grouped data, the formula for the mean is

$$\bar{x} = \frac{\Sigma fx}{n}$$

Exercise 15B

1 A statistician throws a fair die 20 times and records the following results.

4, 2, 1, 6, 4, 5, 3, 4, 6, 2, 3, 2, 2, 4, 6, 3, 5, 1, 2, 3

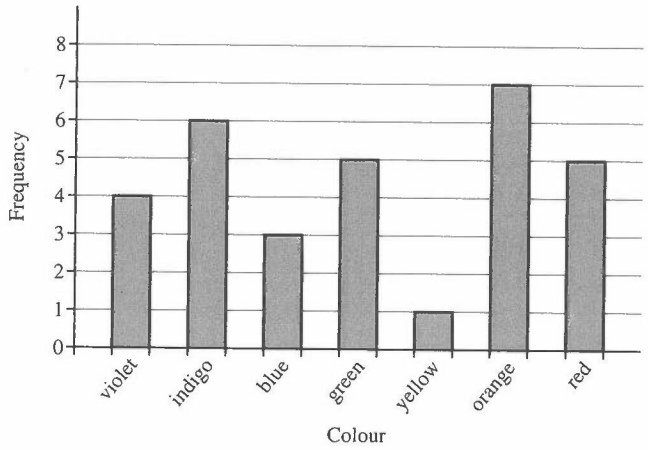
What is the modal score?

2 Each of the children in a class was asked to name the day of the week on which they were born. The results are given in the table.

Day	Mon	Tue	Wed	Thu	Fri	Sat	Sun
Frequency	2	3	2	5	4	1	6

What is the modal day?

3 Each of the passengers on a bus was asked to name a favourite colour of the rainbow. The results are given in the bar chart.



What is the modal colour?

4 Fifteen boxes of matches were purchased, and the number of matches in each box was recorded. The results are:

52, 48, 49, 51, 51, 47, 50, 55, 48, 52, 52, 46, 53, 47, 46

What is the median number of matches?

5 The masses, in kg, of the 11 players in a football team are:

67, 74, 59, 82, 76, 58, 72, 66, 85, 67, 63

What is the median mass?

6 The masses, in kg, of the eight oarsmen in an VIII are:

76, 82, 83, 78, 92, 85, 98, 75

What is the median mass?

7 An athlete records his time over 200 m over five successive races. The times, in seconds, are:

24.2, 25.1, 22.7, 23.5, 24.0

What is his mean time?

8 The four children in a family have ages 5 years, 6 years, 11 years and 14 years. Calculate their mean age.

9 The heights, in cm, of the seven policemen on duty in a police station are:

169, 183, 171, 178, 184, 172, 189

What is the mean height?

10 Four apples have a mean mass of 124 g. Another apple is added to the original four, and the mean mass of the five apples is now 132 g. What is the mass of the additional apple?

11 The table shows the salaries of 80 secretaries in a company.

Salary (€)	8000–10 000	10 000–12 000	12 000–15 000	15 000–20 000
Frequency	45	18	12	5

Calculate an estimate of the mean salary.

12 The table shows the masses of 50 chickens on a supermarket shelf.

Mass (kg)	2.2–2.4	2.4–2.6	2.6–3.2	3.2–3.8	3.8–4.6
Frequency	12	14	11	9	4

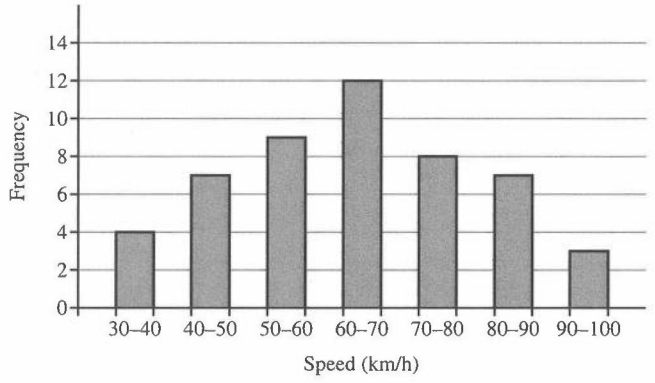
Calculate an estimate of the mean mass.

13 The table shows the IQs of 70 students.

IQ	80–100	100–120	120–130	130–140	140–160
Frequency	6	22	25	14	3

Calculate an estimate of the mean IQ.

14 The bar chart shows the speeds of cars on a stretch of road.



a) How many cars were recorded?

b) Calculate an estimate of the mean speed.

15.4 Measures of dispersion

Range

Look at these two sets of data:

2, 3, 4, 5, 6 [1]

−3, 2, 3, 5, 13 [2]

They both have a mean of 4. However, you can see that data set [2] is more spread out than data set [1]. The mean doesn't tell you this.

To represent the data more accurately, you need the mean plus a measure of the spread or dispersion of the data. One simple measure of dispersion is the **range**.

The range of a set of data is the highest value minus the lowest value.

In this case the range of set [1] is $6 - 2 = 4$.
The range of set [2] is $13 - (-3) = 16$.

The range is easy to calculate, but there are other measures of spread that are more useful.

Mean deviation squared from the mean

A common measure of spread is the mean of the sum of the squares of the deviations from the mean:

$$\frac{\Sigma(x - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}$$

$x - \bar{x}$ is the deviation, or difference, of the value x from the mean value, $\bar{x}$.

The more variation in the data, the larger the value of $\frac{\Sigma(x - \bar{x})^2}{n}$.

For data set [1] in the previous example:

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= (2 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2 \\ &\quad + (6 - 4)^2 \\ &= 10\end{aligned}$$

$$\therefore \frac{\Sigma(x - \bar{x})^2}{n} = \frac{10}{5} = 2$$

For data set [2]:

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= (-3 - 4)^2 + (2 - 4)^2 + (3 - 4)^2 + (5 - 4)^2 \\ &\quad + (13 - 4)^2 \\ &= 136\end{aligned}$$

$$\therefore \frac{\Sigma(x - \bar{x})^2}{n} = \frac{136}{5} = 27.2$$

The value 27.2 compared with the value 2 clearly shows a greater variation in data set [2].

Variance and standard deviation

The mean of the deviations from the mean squared is called the **variance**, usually written as σ^2.

σ is the Greek letter sigma.

The variance is calculated as:

$$\sigma^2 = \frac{1}{n}\Sigma(x - \bar{x})^2$$

Sometimes it is easier to calculate a variance using the alternative formula:

$$\sigma^2 = \frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2$$

This is often more useful when performing calculations without the use of the GDC.

The **standard deviation**, σ, is defined as the square root of the variance:

$$\sigma = \sqrt{\frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2} = \sqrt{\frac{1}{n}\Sigma(x - \bar{x})^2}$$

Example 1

The heights of eight plants are measured 3 months after they are fed with a new plant food.

Calculate the mean and standard deviation of the heights:

30 cm, 17 cm, 32 cm, 25 cm, 31 cm, 28 cm, 35 cm, 26 cm

Putting the data into the GDC and calculating the mean gives 28 cm.

The GDC will give you the standard deviation 5.15, from which you can calculate the variance as 26.5.

Using the formula would give the mean as:

$$\frac{\Sigma x}{n} = \frac{30 + 17 + 32 + \ldots + 26}{8} = 28$$

The variance is:

$$\sigma^2 = \frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2$$

$$= \frac{1}{8}(30^2 + 17^2 + 32^2 + \ldots + 26^2) - 28^2$$

$$= 26.5$$

Therefore the standard deviation is $\sigma = \sqrt{26.5} = 5.15$

The mean height is 28 cm and the standard deviation is 5.15 cm.

Ensure that your GDC is in statistical mode.

Example 2

The heights, in metres, of 10 children of a particular age in a school were recorded as

1.0, 1.2, 1.3, 1.1, 1.2, 1.4, 1.1, 0.9, 1.3, 1.2

Calculate a) the mean b) the standard deviation.

Inputting the data into the GDC gives the following results:

a) the mean is 1.17 m

b) $\sigma = 0.142$ m

Standard deviation for frequency distributions

In this case the x values are the mid-points of the class intervals. The formula for variance becomes

$$\sigma^2 = \frac{1}{n}\Sigma fx^2 - \left(\frac{\Sigma fx}{n}\right)^2$$

where f represents the frequency of the x observation, and $n = \Sigma f$.

Example 3

A restaurant serves a variety of wines of different prices. In a week chosen at random the numbers of bottles of wine sold were recorded by price and the results are shown in the table.

€x	$0 \leqslant x < 5$	$5 \leqslant x < 10$	$10 \leqslant x < 15$	$15 \leqslant x < 20$	$20 \leqslant x < 25$
Number of bottles of wine sold	27	15	8	3	1

Determine the mean and the standard deviation.

Extending the table to include the mid-values gives:

€x	Mid-value	No. of bottles of wine sold
0–5	2.5	27
5–10	7.5	15
10–15	12.5	8
15–20	17.5	3
20–25	22.5	1

The mid-point of the class 10–15 is $\frac{10 + 15}{2} = 12.5$

The mean is given by

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{(27 \times 2.5) + (15 \times 7.5) + \ldots + (1 \times 22.5)}{54} = 6.57$$

The variance is given by

$$\sigma^2 = \frac{1}{n}\Sigma fx^2 - \left(\frac{\Sigma fx}{n}\right)^2$$

$$= \frac{1}{54}[(27 \times 2.5^2) + (15 \times 7.5^2) + \ldots + (1 \times 22.5^2)] - (6.57)^2$$

$$= 25.1$$

Therefore the standard deviation is $\sigma = \sqrt{25.1} = 5.01$

The mean price is €6.57 and the standard deviation is €5.01.

Exercise 15C

1 Use your GDC to calculate the mean and standard deviation of each of these sets of numbers.

a) 3, 5, 7, 8, 9, 10

b) 5, 5, 7, 8, 9, 10, 12

c) 0, 0, 1, 3, 5, 6, 8, 9

d) $-2, -1, 3, 4, 6$

2 Use your GDC to estimate the mean and standard deviation for each of these frequency distributions.

a)

x	0–10	10–20	20–30	30–40
Frequency	4	6	7	3

b)

x	6–10	11–15	16–20	21–25	26–30	31–35
Frequency	2	3	8	6	5	1

c)

x	20–22	22–24	24–26	26–28	28–30
Frequency	8	14	14	10	4

3 The table shows the prices of the 25 cars which are for sale in a garage.

Price ($s)	1000–2000	2000–4000	4000–6000	6000–10 000	10 000–15 000
Frequency	6	7	5	4	3

a) Use your GDC to calculate an estimate of the mean and the standard deviation of the prices of the cars in the garage.

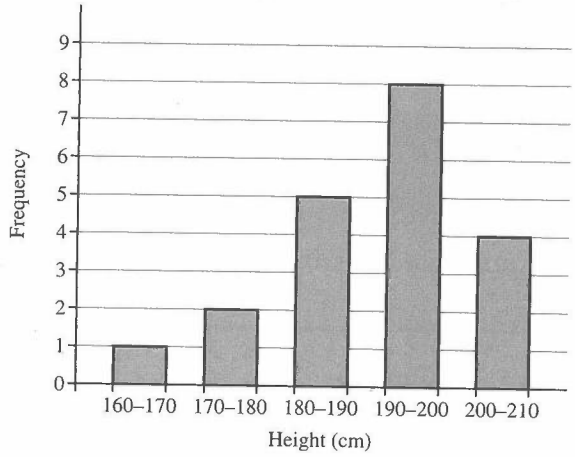
A second garage has 30 cars for sale at the following prices.

Price ($s)	1000–2000	2000–4000	4000–6000	6000–10 000	10 000–15 000
Frequency	8	13	6	2	1

b) Use your GDC to calculate an estimate of the mean and the standard deviation of the prices of the cars in this garage.

c) Comment on the difference in the cars for sale in the two garages.

4 The bar chart shows the heights of players in a basketball squad.



a) How many players are in the squad?

b) Calculate an estimate of the mean height.

c) Calculate an estimate of the standard deviation.

5 a) Calculate the mean and standard deviation of these numbers:

1, 2, 3, 4, 5

b) Deduce the mean and standard deviation of each data set:

i) 11, 12, 13, 14, 15

ii) 45, 46, 47, 48, 49

iii) 10, 20, 30, 40, 50

iv) 43, 46, 49, 52, 55

15.5 Cumulative frequency

The **cumulative frequency** is the total frequency up to a particular value or class boundary. The following example illustrates how to construct a cumulative table and then draw a **cumulative frequency curve**.

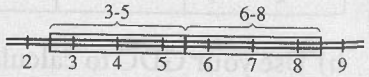
Example 1

The heights to the nearest centimetre of a type of plant were recorded 6 months after planting. The frequency distribution is shown in the table.

Height (cm)	Frequency
3–5	1
6–8	3
9–11	6
12–14	10
15–17	12
18–20	4

Show these results on a cumulative frequency curve.

The interval 6–8 ranges from 5.5 to 8.5.

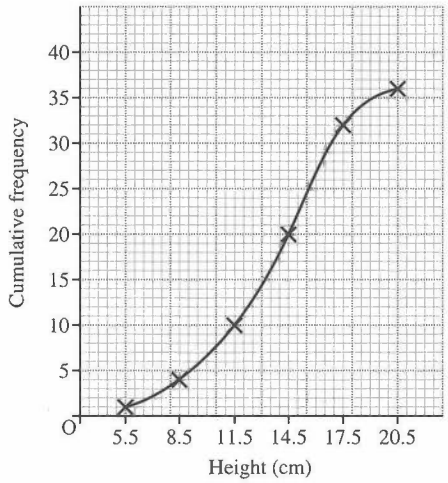


Construct the cumulative frequency table, including the upper class boundary.

Height (cm)	Upper class boundary	Frequency	Cumulative frequency
3–5	5.5	1	1
6–8	8.5	3	4
9–11	11.5	6	10
12–14	14.5	10	20
15–17	17.5	12	32
18–20	20.5	4	36

1 + 3 = 4
4 + 6 = 10
10 + 10 = 20
and so on.

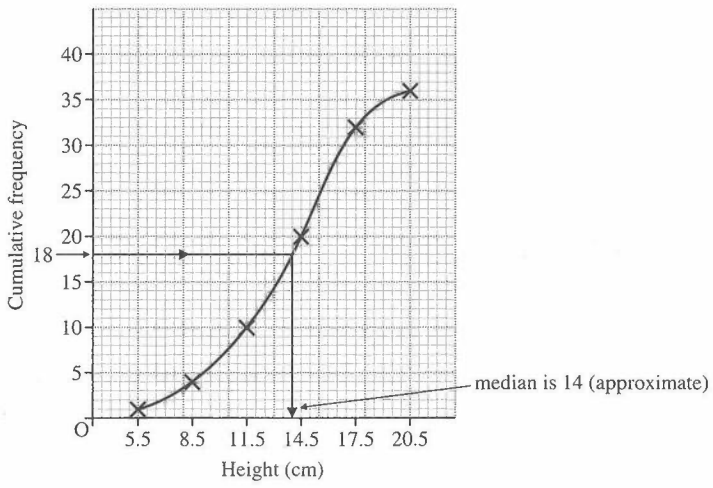
Plot the cumulative frequencies against the corresponding upper class boundaries to give the cumulative frequency curve shown.



Join the points with a curve.

Median, quartiles and interquartile range

Cumulative frequency provides a convenient way of estimating the median. In Example 1, the middle plant is roughly the 18th, since $36 \div 2 = 18$. From the curve, read down from 18 on the vertical axis to give:

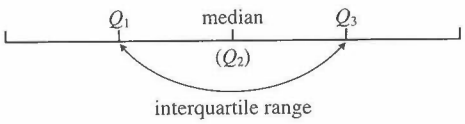


This tells you that there are 18 plants with height less than 14 cm.

The range of values includes any extreme measurements included in the data. It is useful to discard any such extreme measurements. The most common way of doing this is to exclude the top and bottom quarters of the distribution.

The lower point, chosen so that one quarter of the values are less than or equal to it, is called the **first quartile** and denoted as Q_1.

The first quartile Q_1 is also called the **lower quartile**.

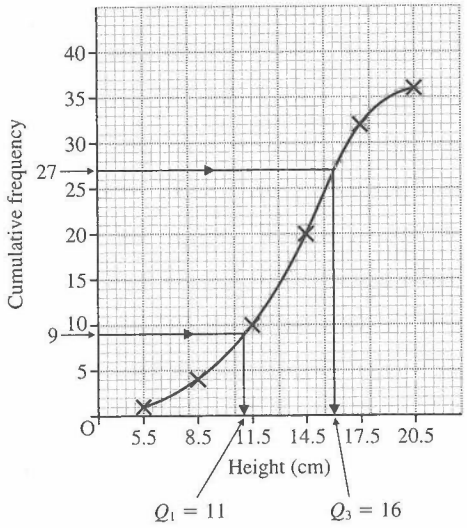


The upper point, with one quarter of the values greater than it, is the **third quartile** and is denoted as Q_3.
(Q_2 is the median.)

The third quartile Q_3 is also called the **upper quartile**.

The difference between the two ($Q_3 - Q_1$) is called the **interquartile range.**

Look again at the cumulative frequency polygon for the heights of plants.



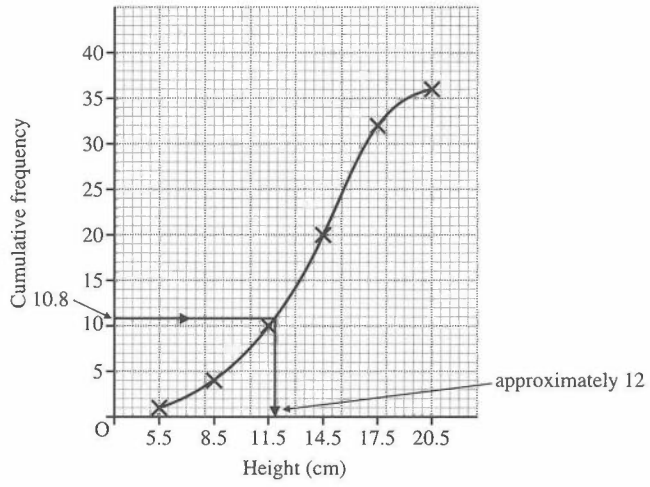
The quartiles are approximately the heights of the 9th ($36 \div 4$) plant and the 27th ($3 \times 36 \div 4$) plant. From the cumulative curve you see that these values are 11 and 16.

Therefore the interquartile range is $16 - 11 = 5$.

Percentiles

The distribution may be split into a greater number of parts.
A very large sample may be split into 100 parts called **percentiles**.
If it is divided into ten parts they are called **deciles**.

In Example 1, the 3rd decile is found by calculating $\frac{3}{10} \times 36 = 10.8$, and projecting down on the cumulative frequency curve as shown.



The 3rd decile is approximately 12.

The 3rd decile is also called the 30th percentile.

Instead of forming a curve, you can join the points with straight lines to make a **cumulative frequency polygon**.

Example 2

In a cricket match the 40 completed innings gave this distribution of scores.

Score	0–9	10–19	20–29	30–39	40–49
Frequency	8	10	6	5	6

Score	50–59	60–69	70–79	80–89	90–99
Frequency	2	0	2	0	1

a) Show these results on a cumulative frequency polygon.

b) Use your polygon to estimate the median score.

c) From your polygon find the upper and lower quartiles, and hence estimate the interquartile range.

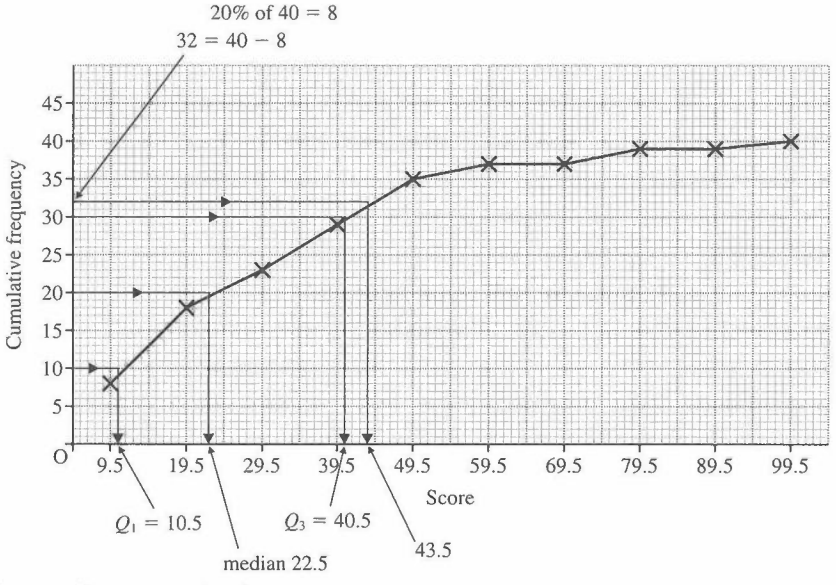
d) The cricket club decide to award prizes to the top 20% of players in the match. What score should be used as a minimum to award prizes?

.....

a) The cumulative frequencies are shown together with the upper class boundaries.

Score	Upper boundary	Frequency	Cumulative frequency
0–9	9.5	8	8
10–19	19.5	10	18
20–29	29.5	6	24
30–39	39.5	5	29
40–49	49.5	6	35
50–59	59.5	2	37
60–69	69.5	0	37
70–79	79.5	2	39
80–89	89.5	0	39
90–99	99.5	1	40

The cumulative frequency polygon is shown.



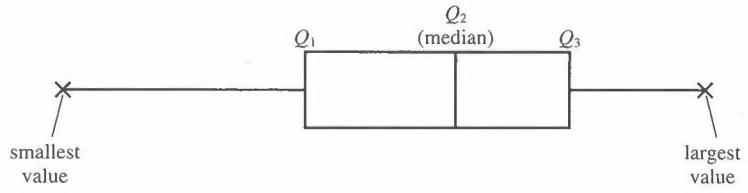
b) The median score is about 22.5.

c) $Q_3 - Q_1 = 40.5 - 10.5 = 30$

d) Score of 43.5 should be used as a minimum.

Box and whisker plots

A box and whisker plot is a useful way of representing data. It shows at a glance the median and quartiles and gives an idea of the spread of the data.

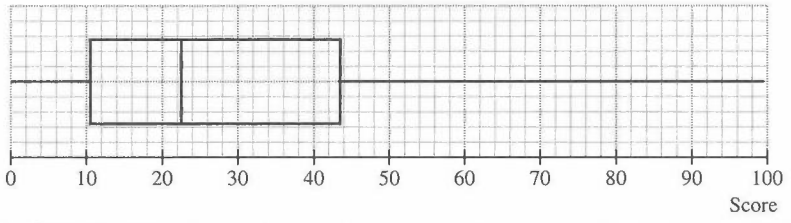


The box represents the central 50% of the data ($Q_3 - Q_1$), and the 'whiskers' show the smallest and largest values.

Example 3

Illustrate the data from Example 2 using a box and whisker plot.

$Q_1 = 10.5$, $Q_2 = 22.5$, $Q_3 = 43.5$, lowest value = 0, highest value = 99.5



Note that the box and whisker plot clearly illustrates where the central 50% of the data lies relative to the smallest and largest values.

Exercise 15D

1 The masses of 800 eggs were recorded and the results are summarised in the table.

Mass (grams)	50–55	55–60	60–65	65–70	70–75	75–80
Frequency	88	152	212	234	78	36

a) Show these results on a cumulative frequency curve.

b) Use your curve to estimate the median mass of an egg.

c) From your curve read off the upper quartile and lower quartile, and hence estimate the interquartile range.

d) Represent this information on a box and whisker plot.

2 The heights of 600 trees in a garden centre were recorded and the results are summarised in the table.

Height (cm)	140–150	150–160	160–170	170–180	180–190	190–200	200–210
Frequency	26	38	56	85	125	212	58

a) Show these results on a cumulative frequency curve.

b) Use your curve to estimate the median height of a tree.

c) Use your curve to read off the upper quartile and lower quartile, and hence estimate the interquartile range.

d) Represent this information on a box and whisker plot.

3 The marks of 120 students in a Physics exam were recorded, and the results are summarised in the table.

Mark	0–20	21–40	41–60	61–80	81–100
Frequency	6	11	24	62	17

a) Show these results on a cumulative frequency curve.

b) Given that the pass mark is 46, estimate the percentage of students who pass.

c) If the chief examiner decides that only 76% of the pupils should pass, to what should he raise the pass mark?

4 The lifetimes of 400 electric lightbulbs were recorded, and the results are summarised in the table.

Lifetime (h)	700–800	800–900	900–1000	1000–1100	1100–1200
Frequency	62	116	98	76	48

a) Show these results on a cumulative frequency polygon.

b) Use your polygon to estimate the median lifetime of a lightbulb.

c) Use your polygon to read off the upper quartile and lower quartile, and hence estimate the interquartile range.

5 As part of a survey, the times of arrival of 160 trains were recorded, and the number of minutes that each train was late is recorded in the table.

Minutes late	0	0–5	5–10	10–20	20–30	30–60
Frequency	24	34	32	28	20	22

a) Show these results on a cumulative frequency polygon.

b) Use your polygon to estimate the 2nd decile, or 20th percentile.

c) Use your polygon to estimate the 7th decile, or 70th percentile.

6 A survey of the ages of 240 cars in a car park gives these results.

Age (years)	0–2	2–4	4–8	8–12	12–16
Frequency	114	36	48	28	14

a) Show these results on a cumulative frequency polygon.

b) Use your polygon to estimate the 42nd percentile.

c) Use your polygon to estimate the 84th percentile.

7 Two hundred students each sat a Maths exam and an English exam. Their marks are summarised in the table.

Mark	0–20	21–40	41–60	61–80	81–100
Frequency (Maths)	14	30	76	56	24
Frequency (English)	6	8	18	122	46

a) Show these results on two cumulative frequency curves, drawn on the same set of axes.

b) Use your curves to read off the median mark and the interquartile range for each of the two exams.

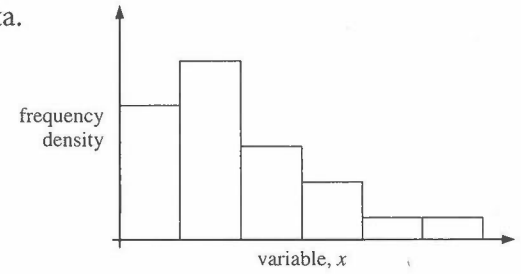
c) Comment on the differences between the marks achieved in each of the two exams.

15.6 Histograms

Histograms are used to illustrate grouped or continuous data. Although histograms may look the same as bar charts, there are some important differences.

In the case of a bar chart the variable axis represents discrete data and is therefore simply divided into spaces.

A histogram represents continuous data and therefore the variable axis is a continuous number line.



In a histogram, rectangles representing frequency may have different widths. The area of each rectangle is proportional to the class frequency.

It is useful to work out the **frequency density** for each class. This is defined as

$$\text{frequency density} = \frac{f}{w}$$

where f = frequency and w = class width.

The Mathematics Standard Level syllabus only deals with histograms with equal class intervals.

When all the class widths are equal, a histogram is easy to construct.

Example 1

100 students take a mathematics test which has a maximum possible score of 40. The results are shown in the table.

Mark	1–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40
Number of students	3	10	22	28	20	10	2	5

Show the results on a histogram.

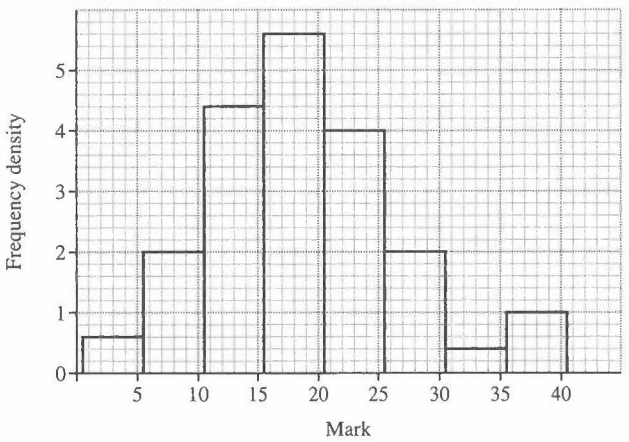
Work out the frequency densities:

Mark	**Class width (w)**	**No. of students (f)**	**Frequency density ($f \div w$)**
1–5	5	3	0.6
6–10	5	10	2
11–15	5	22	4.4
16–20	5	28	5.6
21–25	5	20	4
26–30	5	10	2
31–35	5	2	0.4
36–40	5	5	1

Note: The class widths are calculated as $5.5 - 0.5 = 5$, $10.5 - 5.5 = 5$ etc.

The heights of the sections of the histogram are in proportion to the frequency densities. In this case it makes sense to use 0.5–5.5, 5.5–10.5, etc. on the horizontal axis due to fractional marks and rounding.

Constructing the histogram gives:



The next example illustrates how to construct a histogram with unequal class widths.

Example 2

A survey of family salaries in a holiday resort revealed the following results.

Salary (€)	Number of families
20 000–30 000	2
30 000–40 000	6
40 000–60 000	18
60 000–80 000	15
80 000–100 000	3
100 000–140 000	4

Note: Histograms with unequal class width will not be examined. This example is for interest only.

Show these results on a histogram.

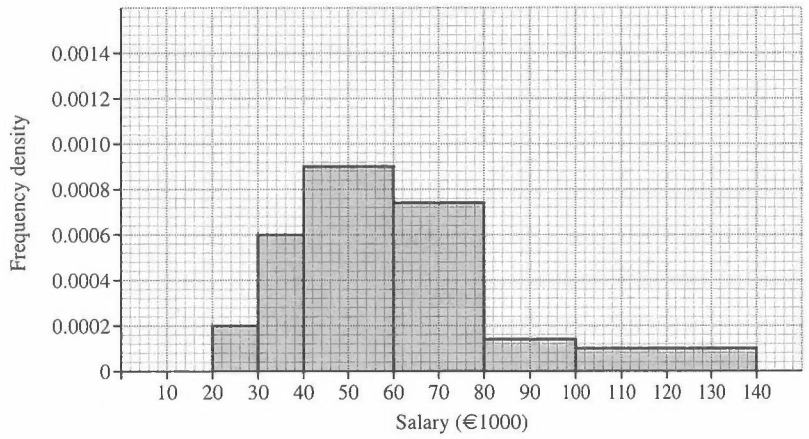
.....

Notice in this case that the class intervals are unequal. The first two are €10 000, the third, fourth and fifth are €20 000 and the sixth is €40 000.

Calculating the frequency densities gives:

Salary (€)	Class width (w)	No. of families (f)	Frequency density ($f \div w$)
20 000–30 000	10 000	2	0.0002
30 000–40 000	10 000	6	0.0006
40 000–60 000	20 000	18	0.0009
60 000–80 000	20 000	15	0.00075
80 000–100 000	20 000	3	0.00015
100 000–140 000	40 000	4	0.0001

Constructing the histogram gives:



Notice that it is the area, not the height, of each bar that represents the frequency.

Exercise 15E

1 A teacher records the Intelligence Quotients of the 24 children in her class. The results are:

125, 132, 117, 98, 151, 146, 133, 106, 114, 128, 126, 141,
137, 122, 152, 108, 137, 121, 142, 102, 148, 133, 136, 95

a) Construct a frequency table of these results with classes 90–99, 100–109, ...

b) Show these results on a histogram.

2 A biologist records the length of 20 worms. The results in mm are:

152, 183, 125, 88, 194, 164, 129, 182, 173, 137, 91,
102, 148, 162, 156, 183, 146, 117, 166, 172

a) Construct a frequency table of these results with classes 80–99, 100–119, ...

b) Show these results on a histogram.

3 Sixteen athletes record their times to run 400 m. The results, in seconds, are:

62, 58, 72, 66, 63, 78, 67, 56, 78, 83, 69, 75, 71, 68, 61, 73

a) Construct a frequency table of these results with classes 55–59, 60–64, ...

b) Show these results on a histogram.

4 A farmer records the masses of 30 piglets at birth. The results in kg are:

3.12, 4.23, 3.67, 4.16, 4.45, 3.56, 3.92, 4.16, 5.02, 3.86, 4.39, 5.13, 4.58, 3.74, 4.12, 3.77, 4.32, 4.73, 3.36, 4.55, 5.23, 4.34, 4.61, 3.98, 4.32, 3.80, 3.54, 4.52, 4.68, 3.74

a) Construct a frequency table of these results with classes 3.00–3.49, 3.50–3.99, ...

b) Show these results on a histogram.

5 The marks obtained by 100 students in a Science exam are given in the table.

Marks	0–19	20–39	40–59	60–79	80–99
Frequency	8	18	26	42	6

a) Show these results on a histogram.

b) Calculate an estimate of the mean mark.

6 The heights of the 34 players in a rugby match are given in the table.

Height (cm)	160–169	170–179	180–189	190–199
Frequency	6	13	11	4

a) Show these results on a histogram.

b) Calculate an estimate of the mean height.

7 On a section of road with a 50 km h^{-1} speed limit, 40 cars are caught on a speed camera. Their speeds are given in the table.

Speed (km h^{-1})	51–60	61–70	71–80	81–90	91–100	101–110
Frequency	13	10	8	5	3	1

a) Show these results on a histogram.

b) Calculate an estimate of the mean amount by which the cars were exceeding the speed limit.

15.7 Random variables

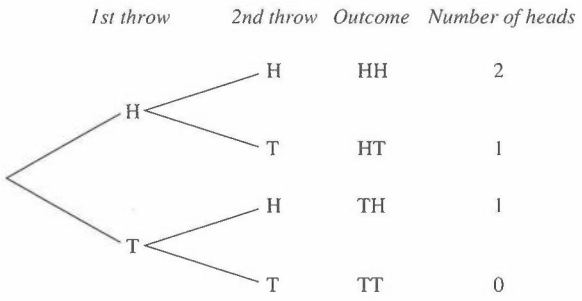
A **random variable**, denoted by X, is a measurable quantity which can take any value or range of values. Its value is the result of a random observation or experiment. Actual measured values are represented by x.

A **discrete random variable** means that a list of its possible numerical values could be made. That is, the data only has certain discrete values.

Probability distributions

Suppose you throw a fair coin twice.

You can draw a probability tree diagram to show this.



If you define the random variable X as 'the number of heads obtained', you can summarise the information in a table:

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

$P(X = x)$ is the probability that the number of heads is x.

Example 1

For each of these examples, draw a table of possible values of x, together with the associated probability $P(X = x)$.

a) A box contains 2 red marbles and 6 green marbles. Two marbles are chosen at random with replacement, and X is the number of green marbles obtained.

b) A fair die has faces labelled 1, 1, 1, 2, 3, 3 and X is the score when the die is thrown.

c) Two fair dice are thrown and X is the difference between the higher score and the lower score.

a) The table shows the possible ways of choosing green marbles together with the associated probabilities.

Marble 1	Marble 2	Number of green marbles	Probability
R	R	0	$\frac{2}{8} \times \frac{2}{8} = \frac{1}{16}$
R	G	1	$\frac{2}{8} \times \frac{6}{8} = \frac{3}{16}$
G	R	1	$\frac{6}{8} \times \frac{2}{8} = \frac{3}{16}$
G	G	2	$\frac{6}{8} \times \frac{6}{8} = \frac{9}{16}$

As the marbles are replaced, each event is independent. So you multiply the probabilities (see page 375).

x	0	1	2
$P(X = x)$	$\frac{1}{16}$	$\frac{3}{16} + \frac{3}{16} = \frac{6}{16}$	$\frac{9}{16}$

Notice that $\frac{1}{16} + \frac{6}{16} + \frac{9}{16} = 1$, which shows that all possible combinations have been considered for the random variable X.

b) The probability distribution for X is:

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

c) In this case the possible outcomes can be best summarised in a table as shown.

		2nd die					
		1	**2**	**3**	**4**	**5**	**6**
1st die	**1**	0	1	2	3	4	5
	2	1	0	1	2	3	4
	3	2	1	0	1	2	3
	4	3	2	1	0	1	2
	5	4	3	2	1	0	1
	6	5	4	3	2	1	0

There are 36 possible outcomes. The probability distribution is summarised in this table.

x	0	1	2	3	4	5
$P(X = x)$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{8}{36}$	$\frac{6}{36}$	$\frac{4}{36}$	$\frac{2}{36}$

Expectation

The mean value of the random variable X is called the **expected value** of X and is written $E(X)$.

The expected value of X is

$$E(X) = \Sigma x P(X = x)$$

Example 2

A fair coin is thrown twice and X is the number of heads obtained. Find $E(X)$.

The table below gives the possible outcomes of throwing a fair coin twice, together with the associated probabilities.

1st throw	2nd throw	Number of heads (x)	Probability
T	T	0	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
T	H	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
H	T	1	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$
H	H	2	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

This is summarised as:

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

Therefore,

$$E(X) = \Sigma x \mathrm{P}(X = x)$$
$$= \left(0 \times \tfrac{1}{4}\right) + \left(1 \times \tfrac{1}{2}\right) + \left(2 \times \tfrac{1}{4}\right)$$
$$= 1$$

Therefore the expected number of heads is 1.

Example 3

The random variable X can only take the values 1, 2 and 3. Given that the value 3 is twice as likely as each of the values 1 and 2, and values 1 and 2 are equally likely,

a) draw a table of possible values of x together with $\mathrm{P}(X = x)$

b) determine the expectation of X.

a) Let $\mathrm{P}(X = 1) = \mathrm{P}(X = 2) = p$, then $\mathrm{P}(X = 3) = 2p$. The table of possible values and probabilities is shown on the right.

x	1	2	3
$\mathrm{P}(X = x)$	p	p	$2p$

Since X can only take the values 1, 2 and 3,

$$p + p + 2p = 1$$
$$\therefore p = \tfrac{1}{4}$$

The completed table is shown on the right.

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

b) $E(X) = \left(1 \times \tfrac{1}{4}\right) + \left(2 \times \tfrac{1}{4}\right) + \left(3 \times \tfrac{1}{2}\right)$

$$= \tfrac{9}{4}$$

Example 4

The random variable X has a probability distribution given by

$$\mathrm{P}(X = x) = \frac{x}{k}, \quad x = 1, 2, 3, 4$$

a) Find the value of the constant k.

b) Calculate $E(X)$.

a)

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

You know that

$$\frac{1}{k}+\frac{2}{k}+\frac{3}{k}+\frac{4}{k}=1$$

$$\frac{10}{k}=1$$

$$\therefore k=10$$

b) $E(X)=\left(1\times\frac{1}{10}\right)+\left(2\times\frac{2}{10}\right)+\left(3\times\frac{3}{10}\right)+\left(4\times\frac{4}{10}\right)$

$=\frac{1}{10}+\frac{4}{10}+\frac{9}{10}+\frac{16}{10}$

$=3$

Exercise 15F

In questions **1** to **7**, for each random variable X, draw a table of possible values, x, together with the matching probabilities, $P(X=x)$.

1 A fair coin is thrown three times, and X is the number of heads obtained.

2 A bag contains four red discs and five blue discs. Two discs are taken out at random without replacement, and X is the number of red discs obtained.

3 Three cards are selected at random without replacement, from a pack of 52, and X is the number of spades.

4 Two fair dice are rolled and X is the sum of the scores on the two faces.

5 In a box are four cards numbered 5, 10, 15 and 20. Two cards are taken out at random without replacement, and X is the total of the numbers on the two cards.

6 A fair die is thrown and X is the square of the number scored.

7 A fair coin is thrown three times and X is the cube of the number of heads showing.

8 Find the value of p in each of these probability distributions. Calculate also the value of $E(X)$.

a)

x	1	2	3
$P(X=x)$	0.2	0.3	p

b)

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

c)

x	0	1	2	3	4
$P(X = x)$	0.15	p	p	0.22	0.31

d)

x	-2	0	1	3
$P(X = x)$	$\frac{2}{9}$	$\frac{1}{9}$	p	$\frac{2}{9}$

e)

x	-4	-1	2	5	8
$P(X = x)$	0.12	0.24	p	$2p$	0.07

f)

x	5	7	9	11	13
$P(X = x)$	p	p	p	0.12	0.04

9 X has probability distribution $P(X = x) = kx$, for $x = 1, 2, 3, 4, 5$.

a) Find the value of the constant k.

b) Calculate $E(X)$.

10 X has probability distribution $P(X = x) = kx^2$, for $x = 1, 2, 3, 4, 5, 6$.

a) Find the value of the constant k.

b) Calculate $E(X)$.

11 X has probability distribution $P(X = x) = \frac{k}{x}$, for $x = 1, 2, 3, 4$.

a) Find the value of the constant k.

b) Calculate $E(X)$.

12 A probability distribution is given by $P(X = x) = A(1 + x)(2 + x)$, for $x = 0, 1, 2, 3$.

a) Find the value of the constant A.

b) Calculate $E(X)$.

13 A probability distribution is given by $P(X = x) = C(4 - x)(5 - x)$, for $x = 0, 1, 2, 3, 4, 5, 6$.

a) Find the value of the constant C.

b) Calculate $E(X)$.

14

x	1	2	3	4	5
$P(X = x)$	0.1	0.2	a	b	0.2

Given that $E(X) = 3.1$, calculate the values of a and b.

15.8 The binomial distribution

A **binomial distribution** is a discrete distribution defined by two parameters, 'the number of trials', n, and the 'probability of a success', p.

You write $X \sim B(n, p)$, to indicate that the discrete random variable X is binomially distributed.

Here is an illustration of the binomial distribution.

Annie is trying her skills at a shooting range. In a round she has three shots and the probability of a success on any shot is $\frac{1}{4}$. Annie is interested in the number of 'hits' or successes in any round and lists the possibilities in a table.

Possibilities	Number of successes, i.e. hits	Probability
HHH	3	$\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{1}{64}$
HHM	2	$\frac{1}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{3}{64}$
HMH	2	$\frac{1}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{3}{64}$
MHH	2	$\frac{3}{4} \times \frac{1}{4} \times \frac{1}{4} = \frac{3}{64}$
HMM	1	$\frac{1}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{9}{64}$
MHM	1	$\frac{3}{4} \times \frac{1}{4} \times \frac{3}{4} = \frac{9}{64}$
MMH	1	$\frac{3}{4} \times \frac{3}{4} \times \frac{1}{4} = \frac{9}{64}$
MMM	0	$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \frac{27}{64}$

H = hit
M = miss

Let X be the number of hits, i.e. the number of successes.
X has the distribution shown:

x	0	1	2	3
$P(X = x)$	$\frac{27}{64}$	$3 \times \frac{9}{64}$	$3 \times \frac{3}{64}$	$\frac{1}{64}$

In this example, note that the outcomes of the 'experiment' only comprise success (hit) and failure (miss). Also note that we have assumed that the probability of a success was the same for each trial and that each trial was independent of the others.

The distribution of X is a binomial distribution.

The binomial distribution function is given by

$$P(X = r) = {}^nC_r p^r (1 - p)^{n-r}$$

where p is the probability of success and ${}^nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$

You met ${}^nC_r$ in Chapter 6 when you studied binomial expansions. You can use your GDC to work it out.

It is common to write $q = 1 - p$, which gives

$$P(X = r) = {}^nC_r\, p^r q^{n-r}$$

Notice that ${}^nC_r$ is the coefficient of $p^r q^{n-r}$ in the binomial expansion of $(p + q)^n$ (see page 201).

Example 1

A fair coin is thrown 10 times. Find the probability that 6 heads will occur.

Let success be obtaining a head, then $p = \frac{1}{2}$ and $n = 10$.

Let the random variable X be the number of heads obtained, so $X \sim B(10, \frac{1}{2})$. Therefore

$$P(X = r) = {}^{10}C_r\left(\tfrac{1}{2}\right)^r\left(\tfrac{1}{2}\right)^{10-r}$$

Therefore

$$P(X = 6) = {}^{10}C_6\left(\tfrac{1}{2}\right)^6\left(\tfrac{1}{2}\right)^4 = 0.205$$

Use your GDC to work out ${}^{10}C_6$.

Example 2

The probability that a particular page in a mathematics book contains a misprint is 0.2. Find the probability that of 12 pages in the book

a) four of them contain a misprint

b) fewer than two of them contain a misprint.

Let success be a 'misprint on a page', then $p = 0.2$ and $n = 12$.

Let the random variable X be the number of pages containing a misprint, so $X \sim B(12, 0.2)$.

$$P(X = r) = {}^{12}C_r(0.2)^r(0.8)^{12-r}$$

a) $P(X = 4) = {}^{12}C_4(0.2)^4(0.8)^8 = 0.133$

b) $P(X < 2) = P(X = 0) + P(X = 1)$.

$$P(X = 0) = {}^{12}C_0(0.2)^0(0.8)^{12} = 0.0687$$

$$P(X = 1) = {}^{12}C_1(0.2)(0.8)^{11} = 0.2062$$

Therefore $P(X < 2) = 0.0687 + 0.2062 = 0.275$

Returning to Annie practising her shooting, you know that her probability of successfully hitting a target is $\frac{1}{4}$.

If the probability of Annie hitting a target is p then in n attempts she would expect to hit the target $n \times p$ times. This leads to the generalisation for the mean of the binomial distribution $B(n, p)$.

This means that out of every 4 attempts Annie would expect to hit the target on 1 occasion. In 8 attempts she would expect to hit the target twice and so on.

For a binomial distribution $X \sim B(n, p)$, the mean value of X is given by np. This is also the expected value of X.

Example 3

An unbiased die is thrown 24 times. Find the expected number of threes thrown.

Let success be obtaining a three, then $p = \frac{1}{6}$ and $n = 24$.

Let the random variable X be the number of threes obtained, so $X \sim B(24, \frac{1}{6})$.

The expected number of threes is given by the mean,

$$np = 24 \times \tfrac{1}{6} = 4$$

The expected number of threes is 4.

Example 4

a) A bag contains 2 red counters and 3 green counters. Two counters are drawn at random from the bag without replacement. Find the probability that exactly 2 green counters are drawn.

b) Find the probability that out of 15 attempts of drawing two counters (replaced after each two drawn) more than 12 of them result in exactly 2 green counters being drawn.

c) What is the expected number of attempts in which exactly 2 green counters will be drawn?

a) $\mathrm{P}(\text{exactly 2 green}) = \frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$

b) Let a success be choosing 2 green counters, then $p = \frac{3}{10}$.

Let the random variable X be the number of attempts in which 2 green counters are drawn. Then $X \sim \mathrm{B}(15, \frac{3}{10})$ and

$$\mathrm{P}(X = r) = {}^{15}C_r\left(\tfrac{3}{10}\right)^r\left(\tfrac{7}{10}\right)^{15-r}$$

You require $\mathrm{P}(X > 12) = \mathrm{P}(X = 13) + \mathrm{P}(X = 14) + \mathrm{P}(X = 15) = 8.72 \times 10^{-6}$.

c) The expected number of attempts in which exactly 2 green counters will be drawn is given by $15 \times \frac{3}{10} = 4.5$.

Exercise 15G

1 A fair six-sided die has the faces numbered from 1 to 6. The die is thrown four times. Calculate the probability of exactly two fives.

2 A coin is biased in such a way that the probability it lands heads is $\frac{1}{3}$. The coin is thrown six times. Calculate the probability of exactly four heads.

3 One in every five cars on the road is red. Calculate the probability that exactly one out of the next four cars to pass my house is red.

4 There are 20 children in a class. Calculate the probability that exactly four of them were born on a Sunday.

5 In the game of roulette a ball rolls into one of 37 slots. All the slots are equally likely. Eighteen of the slots are red, eighteen are white, and one is green. The ball is rolled ten times. Calculate the probability of exactly six reds.

6 Three out of every four adult males have dark eyes. Calculate the probability that exactly eight of the adult males in a soccer team of eleven have dark eyes.

7 40% of the books in a library are paperbacks. I select twelve books at random. Calculate the probability that exactly five are paperbacks.

8 15% of cars on the road have faulty brakes. As part of a survey, 25 cars are randomly stopped and tested. Calculate the probability that exactly four have faulty brakes.

9 30% of the population have had a particular vaccination. Sixteen people are selected at random. Calculate the probability that exactly three have had the vaccination.

10 When a darts player attempts to hit the centre he has a 0.7 probability of success. He makes 50 attempts at the centre. Calculate the probability that he is successful with exactly 40 of these attempts.

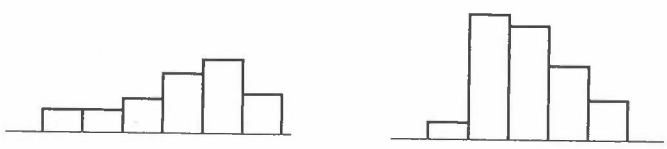
11 A machine produces ornamental plates. It is known that 10% of the plates are cracked. A random sample of 12 plates is selected.

a) Calculate the probability that
   i) none is cracked
   ii) exactly one is cracked
   iii) exactly two are cracked
   iv) there are fewer than three cracked plates.

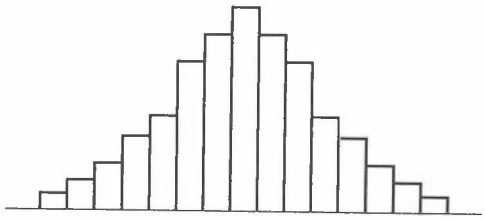
b) Calculate the expected number of cracked plates in the sample.

15.9 The normal distribution

In Exercise 15E you were asked to plot some histograms. These histograms represented a small sample from a large population, for example the length of 20 worms or the masses of 30 piglets. Rough sketches of two such histograms are shown below.



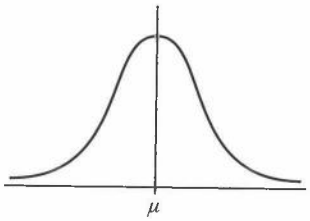
If, instead of restricting yourself to small samples of size 20 or 30, you take samples of size 1000 or 10 000, the histograms would look much more like this:



Notice that the distribution is symmetrical, and bell-shaped, with fewer observations as you move away from the central value. These are properties of the **Normal Distribution**, which is the most important distribution in statistics.

A **normal distribution** is a continuous distribution defined by two parameters, the mean μ and the variance σ^2.

Because of the symmetrical shape of the normal curve, the mean is equal to the mode and the median.

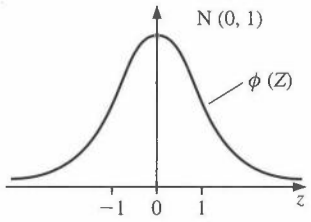


A normal distribution has a bell-shaped curve.

The normal distribution is written as $X \sim N(\mu, \sigma^2)$, to indicate that the continuous random variable X is normally distributed with mean μ and variance σ^2.

The standard normal distribution

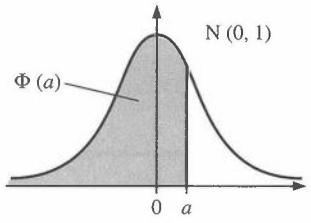
The **standard normal distribution** has mean 0 and variance 1. In this case the random variable is called Z.



The standard normal distribution

Note that the area under the curve is 1.

The probability density function for Z is usually denoted by ϕ. The distribution function is denoted by Φ.



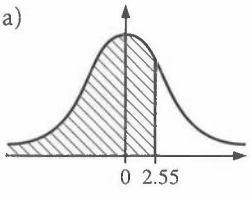
You can use tables of the normal distribution function together with a GDC to calculate areas under the N(0, 1) curve.

Tables of the normal distribution are provided in the formula book, and are duplicated here on page 468.

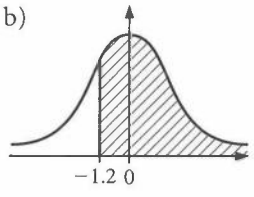
Example 1

Use your GDC to find each of the shaded areas under the N(0, 1) curve shown.

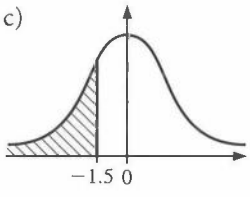
a)



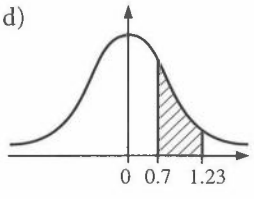
b)



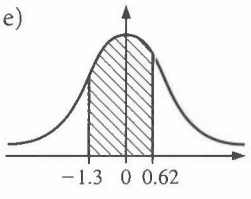
c)



d)



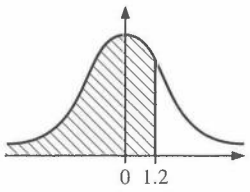
e)

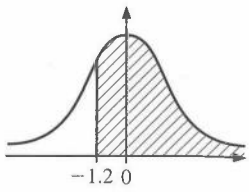


a) $\Phi(2.55) = 0.99461$

b) Because the normal curve is symmetrical, the two areas shown are equal:

Therefore the required area is $\Phi(1.2) = 0.8849$.

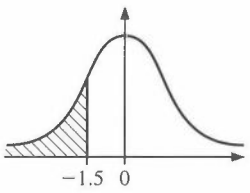


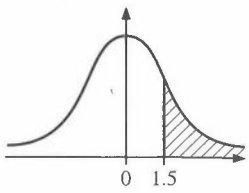


c) Due to symmetry the two areas shown are equal:

Therefore

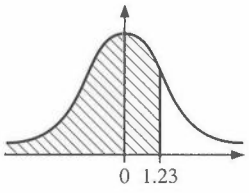
$$\Phi(-1.5) = 1 - \Phi(1.5)$$
$$= 1 - 0.9332$$
$$= 0.0668$$

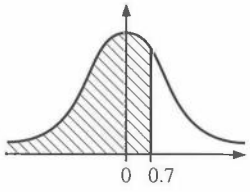




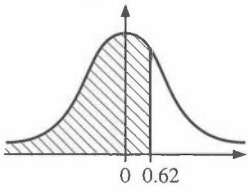
d) The required area is the difference between the areas shown.

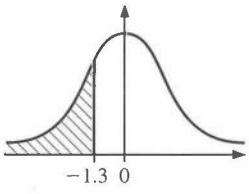
$$\Phi(1.23) - \Phi(0.7) = 0.8907 - 0.7580$$
$$= 0.1327$$





e) The required area is the difference between the areas shown.



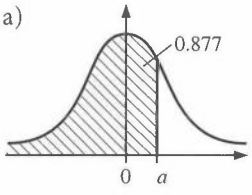


$$\Phi(0.62) - \Phi(-1.3) = \Phi(0.62) - (1 - \Phi(1.3))$$
$$= 0.7324 - 0.0968$$
$$= 0.6356$$

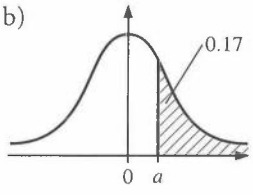
Example 2

Use your GDC to find each of the values a.

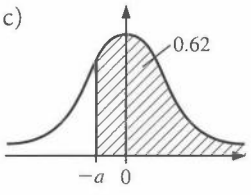
a)



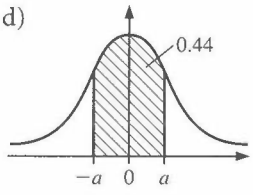
b)



c)



d)



a) $\Phi(a) = 0.877$

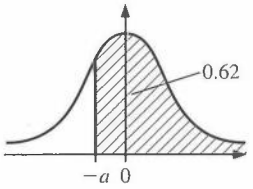
$\therefore a = 1.16$

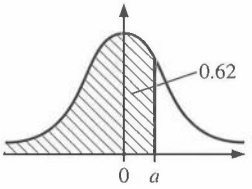
b) $1 - \Phi(a) = 0.17$

$\Phi(a) = 1 - 0.17 = 0.83$

$\therefore a = 0.95$

c)



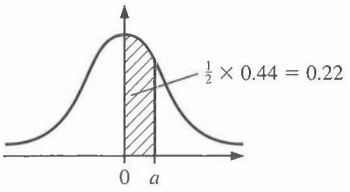


By symmetry, the two shaded areas are equal.

$\Phi(a) = 0.62$

$\therefore a = 0.31$

d) By symmetry



Therefore

$\Phi(a) = 0.5 + 0.22 = 0.72$

$\therefore a = 0.58$

The area under the normal curve represents a **probability**.

If $Z \sim N(0, 1)$ then $\Phi(a)$ represents the probability $P(z < a)$.

Example 3

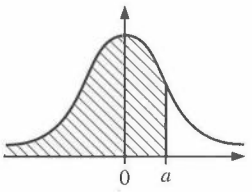
Given that $Z \sim N(0, 1)$, find each of these probabilities:

a) $P(z < 2.68)$

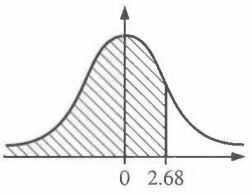
b) $P(z > -1.41)$

c) $P(0.2 < z < 1.71)$

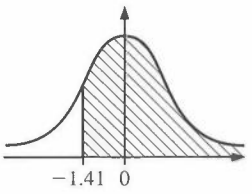
d) $P(|z| < 1.1)$



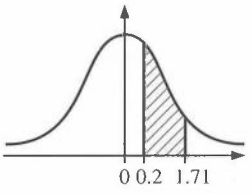
a) $P(z < 2.68) = \Phi(2.68)$
$= 0.9963$



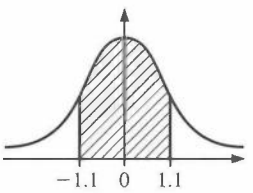
b) $P(z > -1.41) = \Phi(1.41)$
$= 0.9207$



c) $P(0.2 < z < 1.71) = \Phi(1.71) - \Phi(0.2)$
$= 0.9564 - 0.5793$
$= 0.3771$

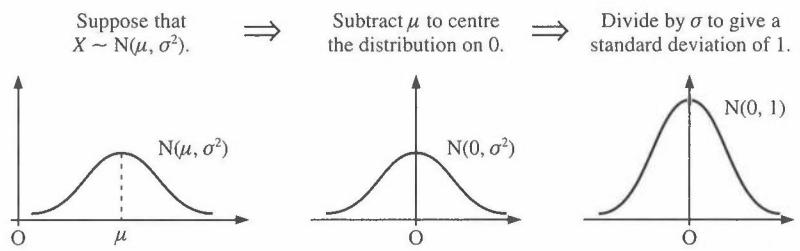


d) $P(|z| < 1.1) = P(-1.1 < z < 1.1)$
$= \Phi(1.1) - \Phi(-1.1)$
$= \Phi(1.1) - (1 - \Phi(1.1))$
$= 0.8643 - 0.1357$
$= 0.7286$



Probabilities for other normal distributions

You now know what to do with the standard normal distribution – but few normal distributions will actually have a mean of zero and a standard deviation of 1. However it is not difficult to transform any set of normal data into the standard normal.



Suppose the random variable $X \sim N(\mu, \sigma^2)$. In such cases the transformation

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

is used to give a normal distribution with a mean of 0 and a standard deviation of 1. You write $Z \sim N(0, 1)$.

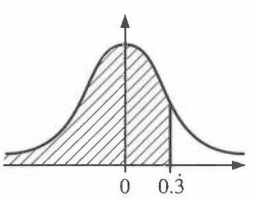
The transformation combines a translation $(-\mu)$ and a change of scale (dividing by σ).

Example 4

Given that the random variable $X \sim N(2, 3^2)$, determine $P(X < 3)$.

3 in standard units is $Z = \frac{3 - 2}{3} = \frac{1}{3}$.

$P(X < 3) = P\left(Z < \frac{1}{3}\right)$
$= \Phi(0.\dot{3})$
$= 0.6306$

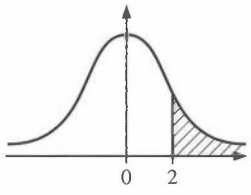


Example 5

Given that the random variable $X \sim N(6, 2^2)$, determine $P(X > 10)$.

10 in standard units is $Z = \dfrac{10 - 6}{2}$.

$$\begin{aligned} P(X > 10) &= P(Z > 2) \\ &= 1 - \Phi(2) \\ &= 1 - 0.97725 \\ &= 0.0228 \end{aligned}$$



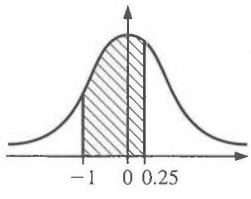
Example 6

Given that the random variable $X \sim N(20, 4^2)$, determine $P(16 < X < 21)$.

16 in standard units is $Z = \dfrac{16 - 20}{4} = -1$.

21 in standard units is $Z = \dfrac{21 - 20}{4} = 0.25$.

$$\begin{aligned} P(16 < X < 21) &= P(-1 < Z < 0.25) \\ &= \Phi(0.25) - \Phi(-1) \\ &= 0.5987 - 0.1587 \\ &= 0.44 \end{aligned}$$



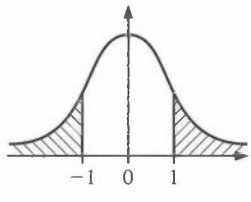
Example 7

Given that $X \sim N(-1, 3^2)$, determine $P(|X| > 2)$.

-2 in standard units is $Z = \dfrac{^-2 - ^-1}{3} = ^-1$.

2 in standard units is $Z = \dfrac{2 - ^-1}{3} = 1$.

$$\begin{aligned} P(|X| > 2) &= P(Z > 1) + P(Z < -1) \\ &= 0.1587 + 0.1587 \\ &= 0.3174 \end{aligned}$$



Example 8

Given that $X \sim N(10, 3^2)$, determine x where $P(X > x) = 0.08$.

$$P(X > x) = 0.08$$

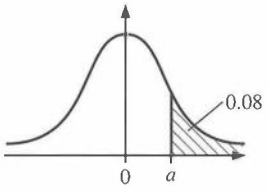
$$P\left(Z > \frac{x - 10}{3}\right) = 0.08$$

Now

$$1 - \Phi(a) = 0.08$$

$$\Phi(a) = 1 - 0.08 = 0.92$$

$$\therefore\ a = 1.4053$$



x in standard units is $Z = \dfrac{x - 10}{3}$.

$$\therefore\ \frac{x - 10}{3} = 1.4053$$

$$x = 14.2$$

Example 9

The lifetime of a particular type of battery is normally distributed with a mean life of 40 hours and a standard deviation of 0.9 hours. Find the probability that a randomly selected battery lasts longer than 42 hours.

Let L be the lifetime of a battery, then $L \sim N(40, 0.9^2)$.

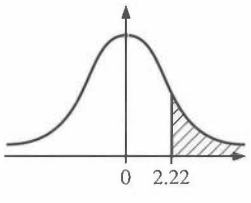
42 in standard units is $Z = \dfrac{42 - 40}{0.9}$.

$$P(L > 42) = P(Z > 2.22)$$

$$= 1 - \Phi(2.22)$$

$$= 1 - 0.9868$$

$$= 0.0132$$



Many real-life quantities are normally distributed.

Example 10

Cartons of juice are such that the volumes of the contents are normally distributed with mean 950 ml and standard deviation 10 ml.

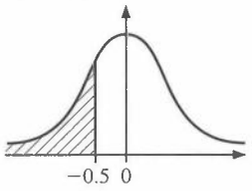
a) Find the probability that a randomly selected carton contains less than 945 ml.

b) Given that 6% of the cartons are rejected for containing too much juice, find the maximum volume, to the nearest ml, that a carton must contain if it is to be accepted.

Let V be the volume of the contents, then $V \sim N(950, 10^2)$.

a) 945 in standard units is $Z = \dfrac{945 - 950}{10} = -0.5$

$$\begin{aligned} P(V < 945) &= P(Z < -0.5) \\ &= 1 - \Phi(0.5) \\ &= 1 - 0.6915 \\ &= 0.3085 \end{aligned}$$



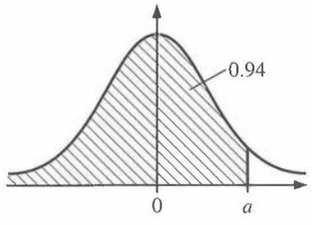
b) Let m be the maximum volume that a carton must contain to be accepted. M in standard units is $\dfrac{M - 950}{10}$.

$$P(V > m) = 0.06$$

$$P\left(Z > \frac{m - 950}{10}\right) = 0.06$$

Now $1 - \Phi(a) = 0.06$

$$\begin{aligned} \Phi(a) &= 1 - 0.06 \\ &= 0.94 \\ \therefore a &= 1.555 \end{aligned}$$



Therefore

$$\frac{m - 950}{10} = 1.555$$

$$m = 965.55$$

The maximum volume is 966 ml, to the nearest ml.

Example 11

A machine produces components whose lengths are normally distributed with a mean of 20 cm. Given that 8% of the components produced by the machine have a length greater than 20.5 cm, find the standard deviation.

Let L be the length of components, then $L \sim N(20, \sigma^2)$.

You know that

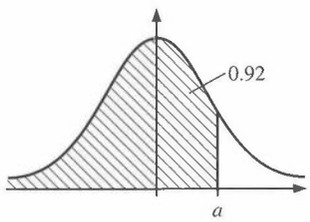
$$P(L > 20.5) = 0.08$$

$$\therefore P\left(Z > \frac{20.5 - 20}{\sigma}\right) = 0.08$$

Now

$$1 - \Phi(a) = 0.08$$

$$\begin{aligned} \Phi(a) &= 0.92 \\ \therefore a &= 1.4053 \end{aligned}$$



20.5 in standard units is $\dfrac{20.5 - 20}{\sigma}$.

$$\frac{20.5 - 20}{\sigma} = 1.4053$$

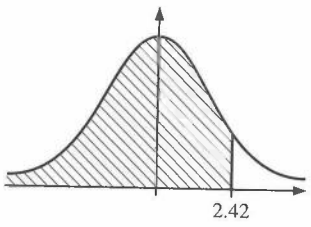
$$\therefore \sigma = 0.356$$

Exercise 15H

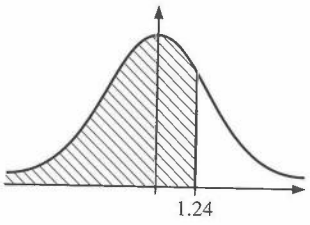
In questions **1** to **3** the random variable Z has a standard normal distribution with mean zero and variance 1.

1 Use your GDC to find each of these shaded areas.

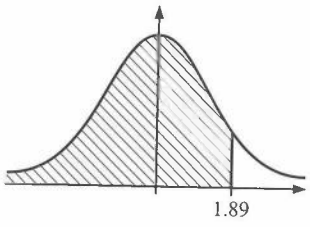
a)



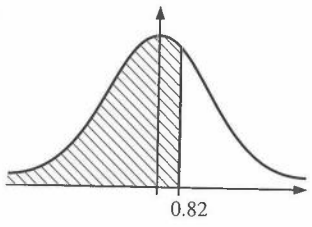
b)



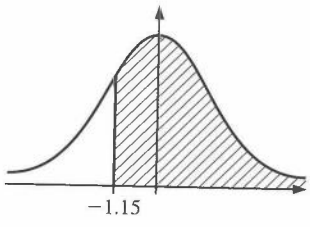
c)



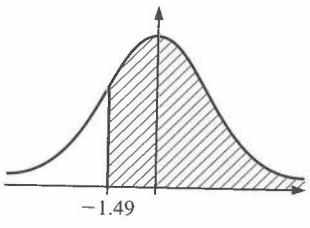
d)



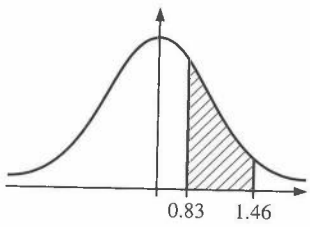
e)



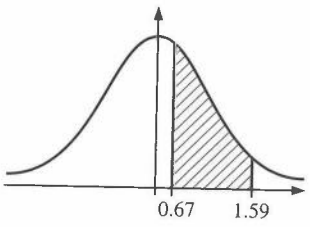
f)



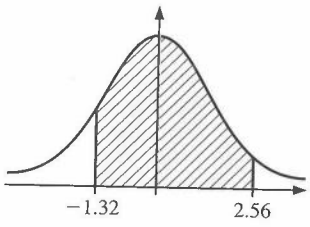
g)



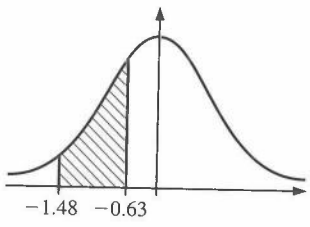
h)



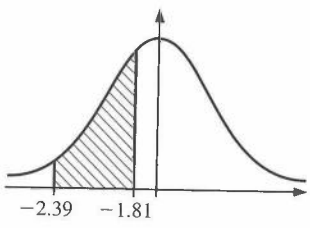
i)



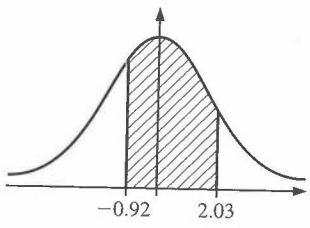
j)



k)

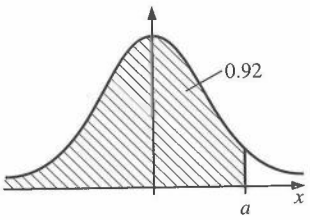


l)

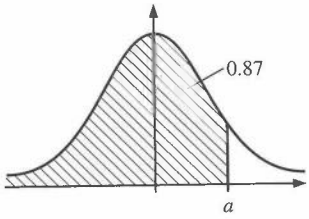


2 Use your GDC to find each of these values of a.

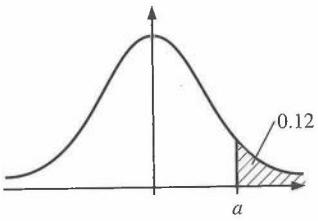
a)



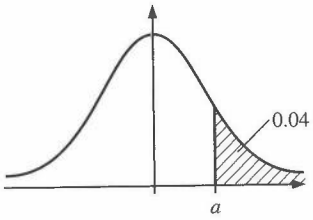
b)



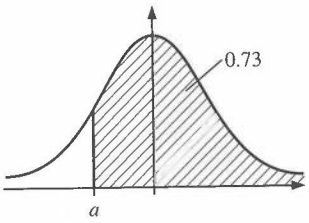
c)



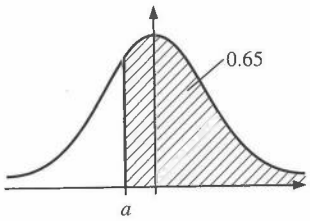
d)



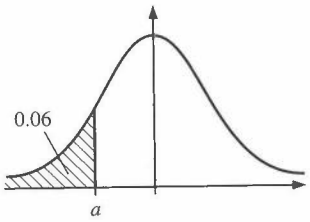
e)



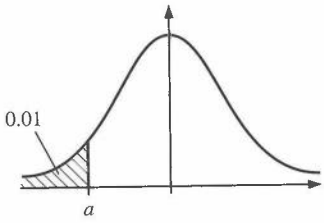
f)



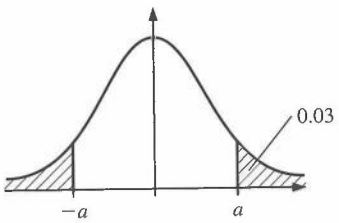
g)



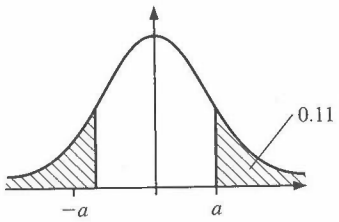
h)



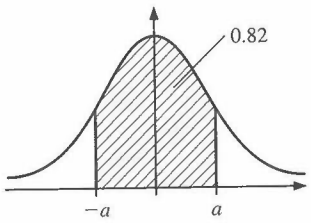
i)



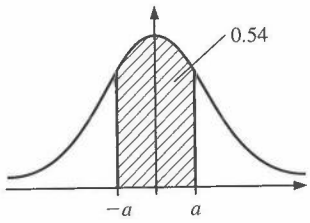
j)



k)



l)



3 Find each of these probabilities, for $Z \sim N(0, 1)$.

a) $P(Z < 1.3)$
b) $P(Z < 2.18)$
c) $P(Z > 1.53)$
d) $P(Z > 1.9)$
e) $P(Z < -0.8)$
f) $P(Z > -1.64)$
g) $P(Z > -2.27)$
h) $P(0.2 < Z < 1.4)$
i) $P(1.13 < Z < 1.72)$
j) $P(-0.3 < Z < 0.4)$
k) $P(-2.02 < Z < 1.29)$
l) $P(|Z| < 1.23)$

4 Given that $X \sim N(14, 4^2)$, calculate these probabilities.

a) $P(X > 19)$ b) $P(X < 15)$

c) $P(X < 12)$ d) $P(X > 7)$

5 Given that $X \sim N(22, 5^2)$, calculate these probabilities.

a) $P(X < 27)$ b) $P(X < 10)$

c) $P(X > 15)$ d) $P(X > 31)$

6 Given that $X \sim N(42, 10^2)$, calculate these probabilities.

a) $P(43 < X < 56)$ b) $P(46 < X < 57)$

c) $P(40 < X < 50)$ d) $P(35 < X < 45)$

7 Given that $X \sim N(37, 8^2)$, calculate these probabilities.

a) $P(23 < X < 43)$ b) $P(23 < X < 35)$

c) $P(15 < X < 55)$ d) $P(|X - 37| < 8)$

8 Given that $X \sim N(-3, 5^2)$, calculate these probabilities.

a) $P(-5.7 < x < -3.1)$ b) $P(-1.1 < X < 2.6)$

c) $P(|X| < 2.2)$ d) $P(|X| > 5)$

9 The mass of an adult male is normally distributed with a mean of 79.4 kg and a standard deviation of 20 kg. Find the probability that a randomly selected adult male has a mass of at least 65 kg.

10 The recorded speeds of cars on a section of motorway are normally distributed with a mean of 134.7 km h^{-1} and a standard deviation of 18 km h^{-1}. Calculate the percentage of cars which are travelling at less than 120 km h^{-1}.

11 The mid-day temperature in Gaiole in July is normally distributed with a mean of 35.8 °C and a standard deviation of 4 °C. Calculate the probability that it will be between 30 °C and 40 °C at mid-day in Gaiole on 12th July next year.

12 The length of earthworms is known to be normally distributed with a mean of 138 mm and a standard deviation of 40 mm. Calculate the proportion of earthworms which are between 100 mm and 200 mm in length.

13 Chicken eggs have a mean mass of 63.2 g and a standard deviation of 10 g. Eggs under 50 g are classified as *small*. Those between 50 g and 70 g are classified as *medium* and those above 70 g are classified as *large*.

a) What percentage of chicken eggs are classified as *small*?

b) What percentage of chicken eggs are classified as *medium*?

c) What percentage of chicken eggs are classified as *large*?

14 Conifers in a nursery have a mean height of 153 cm and a standard deviation of 50 cm. Those under 120 cm are classified as *short*. Those between 120 cm and 180 cm are classified as *medium* and those above 180 cm are classified as *tall*.

a) What percentage of conifers are classified as *short*?

b) What percentage of conifers are classified as *medium*?

c) What percentage of conifers are classified as *tall*?

15 The heights of candidates for the post of air hostess are normally distributed with a mean of 172 m and a standard deviation of 9 cm. 8% of candidates are rejected for being too tall. Calculate the critical height for an air hostess.

16 The speeds of lorries through a village are normally distributed with a mean of 65 km h^{-1} and a standard deviation of 12 km h^{-1}. Given that 10.56% of lorries are breaking the speed limit, what is that speed limit?

17 The masses of jars of home-made sweets are normally distributed with a mean of 245 g and a standard deviation of 23 g. Those jars with a mass below L grams are rejected. Given that 5% of jars are rejected, calculate the value of L.

18 The average daily temperature during the winter months is normally distributed with a mean of 12 °C and a standard deviation of 8 °C. The government decide to make *cold weather payments* when this average falls below a certain value, T °C. Given the government make cold weather payments on 1.5% of days in the winter months, calculate the value of T.

19 The time for an athlete to run 100 m is normally distributed with a mean of 12.4 seconds, and unknown standard deviation, σ seconds. Given the athlete records a time of more than 13.1 s in 4% of runs, calculate the value of σ.

20 The volume of beer in a glass is normally distributed with mean 523 cm^3, and unknown standard deviation, σ cm^3. Given that 20% of glasses have less than 500 cm^3 of beer, calculate the value of σ.

21 6% of the babies to be born in a hospital have a mass of over 3200 g, and 14% have a mass of less than 2600 g. Given the masses are normally distributed, calculate the mean and the standard deviation.

Summary

You should know how to ...

- Draw appropriate statistical diagrams, including:
  - stem and leaf diagrams
  - bar charts
  - line graphs
  - pictograms
  - cumulative frequency graphs
  - box and whisker plots
  - histograms
  - pie charts
- Calculate measures of location.
  - The mode is the value that occurs most often.
  - The median is the middle value in order of size.
  - The mean $\bar{x}$ is given by the formula:

    $\bar{x} = \frac{1}{n}\Sigma x$, or $\bar{x} = \frac{\Sigma fx}{\Sigma f}$ for a frequency distribution.
- Calculate measures of dispersion.
  - The variance σ^2 is given by the formula:

    $\sigma^2 = \frac{1}{n}\Sigma(x - \bar{x})^2$, or $\sigma^2 = \frac{1}{n}\Sigma x^2 - \left(\frac{\Sigma x}{n}\right)^2$
  - The standard deviation σ is the square root of the variance.
  - For a frequency distribution:

    $\sigma^2 = \frac{1}{n}\Sigma fx^2 - \left(\frac{\Sigma fx}{n}\right)^2$
- Estimate quartiles and percentiles for large sets of data.
- Calculate with discrete random variables.
  - The expected value of X is

    $E(X) = \Sigma x\, P(X = x)$
- Use the binomial distribution
  - $P(X = r) = {}^nC_r p^r (1 - p)^{n-r}$, where $X \sim B(n, p)$.
  - The mean value of X is np.
- Use the normal distribution
  - The standard normal distribution is $Z \sim N(0, 1)$, where the mean is 0 and the variance is 1.
  - For a normal distribution $X \sim N(\mu, \sigma^2)$, $Z = \frac{X - \mu}{\sigma}$.

Revision exercise 15

1 Given the following frequency distribution, find
a) the median
b) the mean.

Number (x)	1	2	3	4	5	6
Frequency (f)	5	9	16	18	20	7

© *IBO[2001]*

2 Three positive integers a, b and c, where $a < b < c$, are such that their median is 11, their mean is 9 and their range is 10.
Find the value of a.

© *IBO[2002]*

3 From January to September the mean number of car accidents per month was 630. From October to December the mean was 810 accidents per month.

What was the mean number of car accidents per month for the whole year?

© *IBO[2002]*

4 A supermarket records the amount of money d spent by customers in their store during a busy period. The results are as follows:

Money in \$ (d)	0–20	20–40	40–60	60–80	80–100	100–120	120–140
Number of customers (n)	24	16	22	40	18	10	4

a) Find an estimate for the mean amount of money spent by the customers, giving your answer to the nearest dollar (\$).

b) Copy and complete the following cumulative frequency table and use it to draw a cumulative frequency graph.
Use a scale of 2 cm to represent \$20 on the horizontal axis, and 2 cm to represent 20 customers on the vertical axis.

Money in \$ (d)	<20	<40	<60	<80	<100	<120	<140
Number of customers (n)	24	40					

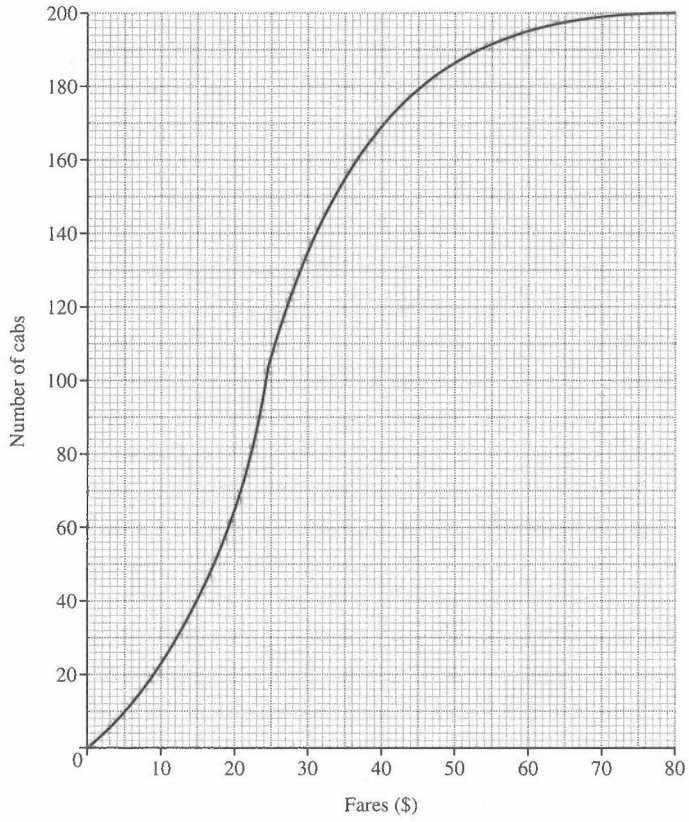
c) The time t (minutes) spent by customers in the store may be represented by the equation

$$t = 2d^{\frac{2}{3}} + 3$$

i) Use this equation and your answer to part a) to estimate the mean time in minutes spent by customers in the store.
ii) Use the equation and the cumulative frequency graph to estimate the number of customers who spent more than 37 minutes in the store.

© *IBO[2000]*

5 A taxi company has 200 taxi cabs. The cumulative frequency curve shows the fares in dollars ($) taken by the cabs on a particular morning.



a) Use the curve to estimate

i) the median fare

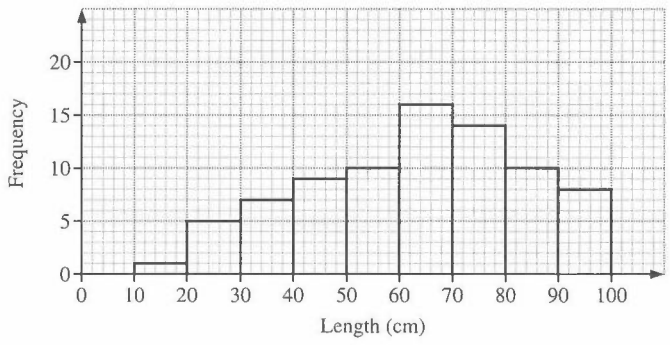
ii) the number of cabs in which the fare taken is $35 or less.

The company charges 55 cents per kilometre for distance travelled. There are no other charges. Use the curve to answer the following.

b) On that morning, 40% of the cabs travel less than a km. Find the value of a.

c) What percentage of the cabs travel more than 90 km on that morning?

6 The diagram represents the lengths, in cm, of 80 plants grown in a laboratory.



a) How many plants have lengths in cm between

i) 50 and 60

ii) 70 and 90?

b) Calculate estimates for the mean and the standard deviation of the lengths of the plants.

c) Explain what feature of the diagram suggests that the median is different from the mean.

d) The following is an extract from the cumulative frequency table.

Length in cm less than	Cumulative frequency
...	...
50	22
60	32
70	48
80	62
...	...

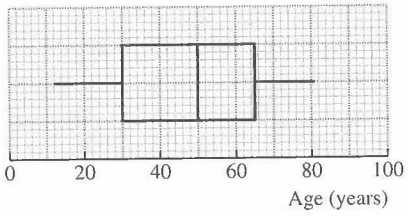
Use the information in the table to estimate the median.
Give your answer to **two** significant figures.

© *IBO [Spec.]*

7 The ages of people living in a certain street were recorded in years. The minimum age was 1 year and the maximum age was 91 years. The median age was found to be 42 years, and the quartiles were 25 and 60 years.

a) Draw a box and whisker plot to illustrate this information.

For a second street the data were presented in the following box and whisker plot.



b) State the range and interquartile range for this street.

c) Write two statements comparing the distribution of ages in the two streets.

8 A factory makes computer components.
Over a long period it is found that 7% of components are faulty.
A sample of 30 components is tested.
Find the probability that

a) exactly two components are faulty

b) none are faulty

c) two or fewer are faulty

d) more than two are faulty.

9 An airline knows from past experience that the probability of a person booking a seat and then not turning up is 0.04.
A small plane has 50 seats and 55 bookings are made.

a) A binomial distribution is used to model this situation.
What assumption must be made?
Comment on how reasonable this assumption is.

b) What is the expected number of no-shows?

c) What is the probability that more people turn up than there are seats available?

10 The masses of a group of students are normally distributed with mean 65 kg and standard deviation 10 kg.

a) Find the probability that a student, chosen at random, has a mass which is
i) less than 60 kg
ii) more than 70 kg
iii) between 60 and 70 kg.

b) Three students are chosen at random. Find the probability that only one of them has a mass of more than 70 kg.

11 A daily rail service operates to Euphoria City. The trains arrive T minutes after 08:00, where T is a random variable with a normal distribution. The mean value of T is 27, and there is a 5% probability that a train arrives after 08:38.

a) Show that the standard deviation of T is approximately 6.7.

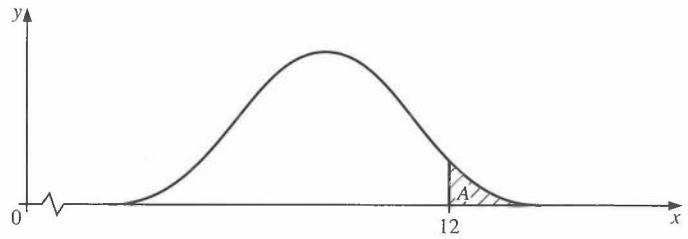
b) Find the probability that a train arrives before 08:20.

The trains are scheduled to arrive at 08:23.

c) Show that approximately 20% of the arrivals are within 2 minutes of the scheduled time.

© *IBO[1998]*

12 The graph shows a normal curve for the random variable X, with mean μ and standard deviation σ.



It is known that $P(X \geqslant 12) = 0.1$.

a) The shaded region A is the region under the curve where $x \geqslant 12$.
Write down the area of the shaded region A.

It is also known that $P(X \leqslant 8) = 0.1$.

b) Find the value of μ, explaining your method in full.

c) Show that $\sigma = 1.56$ to an accuracy of three significant figures.

d) Find $P(X \leqslant 11)$.

© *IBO[1999]*