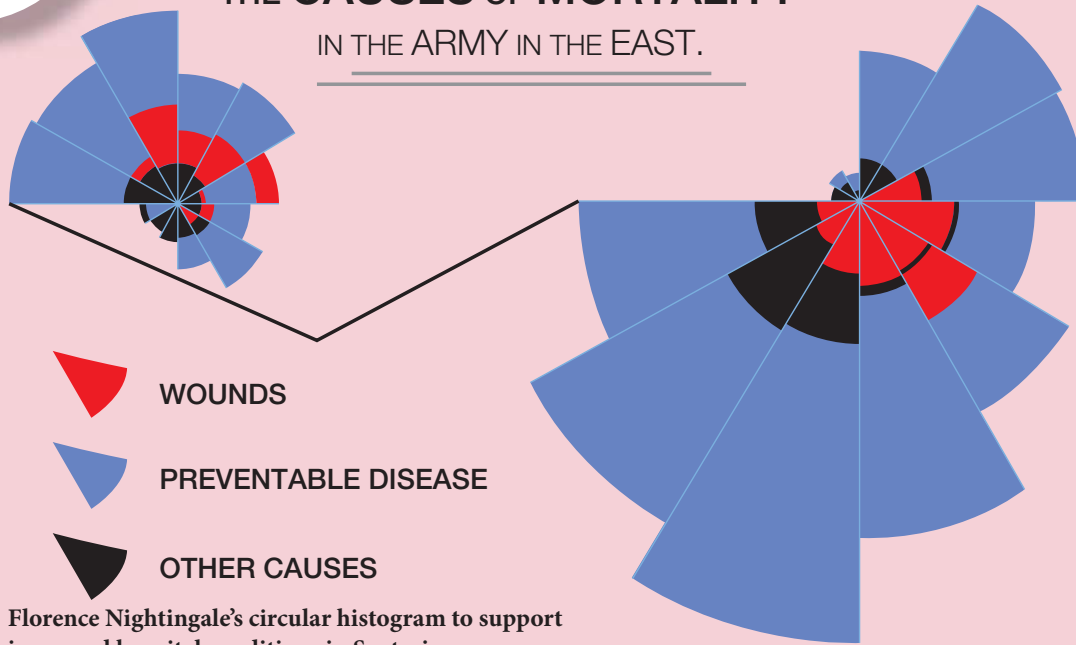


2

Descriptive statistics

THE CAUSES OF MORTALITY

IN THE ARMY IN THE EAST.



Florence Nightingale's circular histogram to support improved hospital conditions in Scutari.

In 1854, Russia was at war with Turkey. England and France went to Turkey's aid and sent soldiers to fight in the Crimea. Conditions were dreadful, and soldiers of all the armies were dying of hunger and disease as well as from wounds sustained in battle.

Florence Nightingale went to Scutari, one of the worst hospitals, and was appalled by the conditions that she found there. However, before she could make improvements, she faced the difficult task of convincing the authorities in London that improvements were necessary.

While she was growing up, Florence loved mathematics and had special tutors to help her study. She was only eight when she drew her first statistics diagram. While at Scutari, she kept careful records of all the patients, documenting the reasons for their death or recovery. She realised that pages of figures and tables are difficult to interpret and to explain. So she drew one of the first circular histograms, or 'rose diagrams', to illustrate the causes of soldiers' death during the war.

The diagram has twelve sectors, one for each month of the year. The largest area, the outer layer coloured blue, shows the deaths from infectious but preventable diseases. The red areas show the deaths from wounds, and the black areas are for 'all other causes'.

The impact was immediate: pages of figures were condensed into a single, clear graphic image. It is fair to imagine that Florence Nightingale's diagram was an important part of her campaign to improve the conditions in military hospitals, and helped her to achieve the transformation for which she is remembered.

Prior learning topics

It will be easier to study this topic if you:

- know how to collect data
- are able to use tally marks to create a frequency table for data
- can recognise and interpret bar charts, pie charts and pictograms
- can draw your own bar charts, pie charts and pictograms.

Chapter 5 Classification and display of data



The world is flooded with data. There is more information available now than there has ever been before, and much of it is available to anyone who wishes to access it. This abundance of easily accessible information is fantastic but can also create problems. A mass of data can be difficult to make sense of until it has been organised into a form that has meaning.

Statistics is the branch of mathematics that organises data and presents it in a format that is understandable and useful. It is concerned with the ways in which to analyse data, and how to show the significance of that data after it has been analysed.

When presented with a large quantity of data, one of the first decisions you need to make is how to classify the data. If you do a seaside survey for your Biology coursework, you will have pages filled with information: lengths and masses, colours and numbers, areas and names of species. How many of these pieces of information can you put together? Can you compare the area of seaweed with its mass, or the colour of a shell with its diameter?

In this chapter, you will study some of the diagrams that you can use to make data easier to understand.

5.1 Classifying data

The first thing you need to do when faced with a collection of data is to classify it. The type of data that you are working with will determine the kind of diagram you will use to present it and the calculations that you can do with it.

Qualitative data is data that cannot be counted — for example, the colour of a leaf.

In this chapter you will learn:

- about classifying data as discrete or continuous
- how to deal with simple discrete data and to draw up frequency tables
- how to treat grouped discrete or continuous data and how to draw up frequency tables
- the definitions of mid-interval values and upper and lower boundaries of such data
- how to use a GDC to draw frequency histograms
- how to produce cumulative frequency tables for grouped discrete data and for grouped continuous data
- how to find medians and quartiles from cumulative frequency curves
- how to find five-figure summaries and draw box and whisker diagrams.



Statistics is an important research tool in many fields, including biology, psychology and economics.

Quantitative data is data that comes from counting or measuring. The mass of a shell, the length of a stream, and the number of birds in a flock are all examples of quantitative data.

Quantitative data can be classified as either discrete or continuous.

Discrete data can take only certain distinct values, such as whole numbers, or fall into distinct categories. For example, the number of birds in a flock is discrete data: you can count 7 or 8 birds, but not $7\frac{1}{2}$ birds. The units of measurement for discrete data cannot be split into smaller parts. Be careful — for example, shoe size in the UK is measured in whole and half sizes so it is possible to have a shoe size of 6 or $6\frac{1}{2}$. Discrete data might not always be a whole number. UK shoe size is restricted to a whole number or the half sizes between those whole numbers and no other division is possible, so it is discrete data.

Continuous data can take any values within a certain range of real numbers. This type of data usually comes from measuring. For example, the length of a bird's wing could be 10 cm, 10.1 cm, 10.12 cm, 10.12467 cm or any other value from 10 cm to 50 cm depending on the species of bird and the degree of accuracy to which the measurement was taken. You can choose how accurately you measure the data. For example, the length of a bird's wing can be measured in centimetres or millimetres; the mass of a pebble can be measured in grams or kilograms.

Populations, samples and bias when collecting data

It is important that any data is collected with a clear purpose. A large quantity of information is very difficult to analyse unless you have decided what question(s) you are seeking to answer, and why you are asking it. You also need to decide on the population that you are studying, and how you are going to take samples from that population.

The **population** is the particular group of objects or people that are being studied. For instance, if you are studying the distribution of a particular kind of shellfish, the population would consist of all the shellfish of that type in the region you are looking at. You could also study different **variables** relating to that species, such as their mass or the depth of water in which they are found.

Populations can be very large, such as all the people in your town, all the plants in a park, or all the cars sold by a local dealer, so it might not be practical to collect data for the whole population. Once you have decided on the population and the question(s) that you want to answer, you will need to think about how you are going to choose samples from this population that are of a practical size to collect and to analyse.

exam tip

You will not be set any questions on sampling in your examinations, but learning about sampling can be useful for projects.

A **sample** is part of the population being studied. It should be chosen with care because if you choose badly the data that you obtain could be of little or no use.

A **representative sample** is a small proportion of the population that is supposed to be representative of the whole population being studied.

A **random sample** is a sample chosen in such a way that all members of the population have an equal chance of being selected.

A **biased sample** is a sample chosen by a method where some members of the population are more likely to be selected than others; this will result in distorted or unfair data.

Suppose that you are doing a survey at your school that asks ‘should we replace the swimming pool with a basketball court?’. Then:

- the **population** consists of all the students in your school
- to collect a **representative sample**, you would need to select students in order to make sure that you get a cross-section of opinions; so make sure you have students from each year, an equal mixture of sporty and non-sporty students, a variety of ability levels, both sexes, mixture of interests, etc.
- to collect a **random sample**, use a method of selection where every student has an equal chance of being chosen (name out of a hat, date of birth, etc.)
- you would be collecting a **biased sample** if you survey only your friends or if you survey only the basketball team.

So, for any statistical collection of data that you plan, you need to ask the following questions:

- Have I decided on a clear research question?
- Have I defined the population?
- Have I chosen a method of taking samples that will not be biased?



Data needs to be looked at critically. Any data should be scrutinised for the reason that it was collected and for **bias**. Try asking yourself ‘Who paid for the data to be collected and analysed?’.

Worked example 5.1

Q. Javed wants to study the statistics of the Athletics World Championships, with a particular interest in the high jump competition. Copy and complete the table to give an example of possible parameters of his survey.

Type of data	Example
Qualitative data	
Quantitative data	
Discrete data	
Continuous data	
Population	
Variables	
Representative sample	
Random sample	
Biased sample	



There are other possibilities, these are just some options.

continued . . .

A.

Type of data	Example
Qualitative data	The names of the countries taking part.
Quantitative data	The number of attempts each athlete makes.
Discrete data	The number of athletes in each team.
Continuous data	The height of the 'personal best' jump for each athlete.
Population	All the athletes competing in the high jump competition.
Variables	Each athlete's age, number of years on the team.
Representative sample	Javed goes through the team lists, and chooses athletes that represent a cross-section of each team.
Random sample	Each athlete is assigned a number, and Javed selects his sample by using a random number generator.
Biased sample	Javed picks his favourite athletes.

Exercise 5.1

1. State whether each of the following examples of quantitative data is discrete or continuous:
 - (a) the time it takes to run 100 metres
 - (b) the number of runs scored in a cricket match
 - (c) the distance jumped by Philip in the Olympic triple jump
 - (d) the number of words in the first sentence on each page of a book
 - (e) the number of items found in the pencil cases of students in a class
 - (f) the length of the footprints of children in a school
 - (g) the mass of each potato in a sack
 - (h) the number of electrons in the outer shell of different atoms
 - (i) the number of leaves on each branch of a tree
 - (j) the arm spans of basketball players in a team.

5.2 Simple discrete data

Simple discrete data consists of a list of individual values. The data can be organised by putting it into a **frequency table**. Frequency tables list each variable (measurement) and its corresponding frequency, and make it easier to see the shape, or distribution, of the data.

Frequency tables can sometimes include a tally chart but it is not essential.

Worked example 5.2

Q. Dee asks all the children at her younger brother's school how many siblings they have. Put her results into a frequency table and include a tally chart. What does her data say about the number of siblings?

A.

Number of siblings	Tally	Frequency
0		12
1		21
2		14
3		9
4		4
Total number of children questioned		60

The table shows that 21 children have only one brother or sister, while 4 children have 4 brothers or sisters each. The most common number of siblings from this sample is 1.

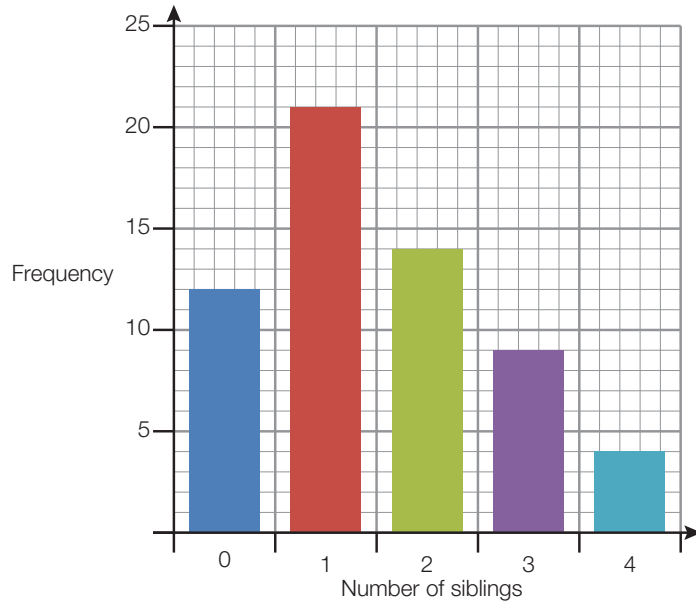
The best diagrams for displaying discrete data are pie charts and bar charts.

Bar charts for discrete data have **gaps** between adjacent bars. They are usually drawn with vertical bars where the discrete data (or categories) are plotted along the horizontal axis and frequency is plotted on the vertical axis. You can also have horizontal bars (where the discrete variable is plotted on the vertical axis and the frequency on the horizontal axis) but this is less common and more likely to be seen in newspapers or magazines than in mathematics books.

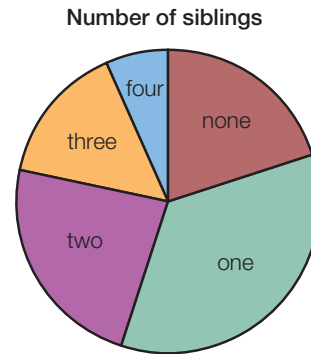
Pie charts display the data as parts of a whole, where the circle represents the total frequency of your data and the sectors represent different measurements. You plot data on a pie chart by calculating the frequency of a given measurement as a fraction of the whole and converting it into degrees (so that it is a fraction of 360°).

Let's see how Dee could display her data.

Dee displays her data in a bar chart:



Dee also used a pie chart to display her data like this:



Exercise 5.2

In each question you are given a set of discrete data. Draw up a frequency table, bar chart and pie chart to summarise the data. Make a statement about your data.

1. The homework marks (out of 10) of 25 students:

9, 8, 7, 7, 6, 7, 6, 7, 10, 5, 10, 8, 6, 9, 7, 5, 9, 8, 8, 5, 6, 10, 9, 5, 7

2. The number of rejects in each of 36 batches of laptop batteries:

1, 0, 0, 2, 5, 1, 3, 0, 0, 2, 0, 0, 0, 0, 2, 1, 4, 2,

3, 3, 3, 4, 1, 3, 5, 0, 0, 2, 4, 4, 3, 2, 3, 1, 1, 2

3. The number of home runs in 17 baseball matches:

28, 23, 24, 24, 25, 26, 26, 27, 27, 27, 27, 28, 28, 28, 31, 31, 34

5.3 Grouped discrete data

When you have a large sample of discrete data with a big difference between the smallest data value and the largest data value, you might find that the total frequency is spread thinly across the many values between the smallest and largest values. When this happens, a frequency table like the one created for Dee in Worked example 5.2 might not be very informative. It would lead to a very long frequency table that will not show clearly any shape or pattern in the data, making it difficult to analyse.

Grouping the data into 'groups' makes it easier to summarise discrete data when the data set is large and/or there is a big difference between the smallest and the largest values. Preferably, the data set should be divided into five to fifteen groups, but the number of groups to take will depend on the size of the sample and the spread of the data. It is easier to make a fair analysis of the distribution of the frequency if the size of each group (the range of data values it includes) is equal. For example, each group might cover four data values. Remember however, that although **grouped data** allows us to see general patterns more clearly, some detail will be lost.

exam tip

You can group the data using unequal sized groups where one group might cover ten values and another includes five, but the data needs to be handled slightly differently in order to make a fair analysis.

There are 50 data values ranging from 7 to 20, so it is sensible to group the data. We have chosen to use seven equal-sized groups (even though two of them will have no entries in them); other groups are possible. Each group includes three values; for example, the group 6–8 includes the marks of 6, 7 and 8.

Worked example 5.3

Q. Ahmed records the project marks of 50 IB students on the Mathematical Studies course:

7	14	16	10	20	10	19	13	8	14
15	8	13	17	11	17	11	15	20	12
14	18	6	11	12	13	6	11	12	17
20	9	13	7	15	12	11	14	11	20
11	17	15	18	8	14	13	19	17	9

Enter the data into a grouped frequency table and comment on the results.

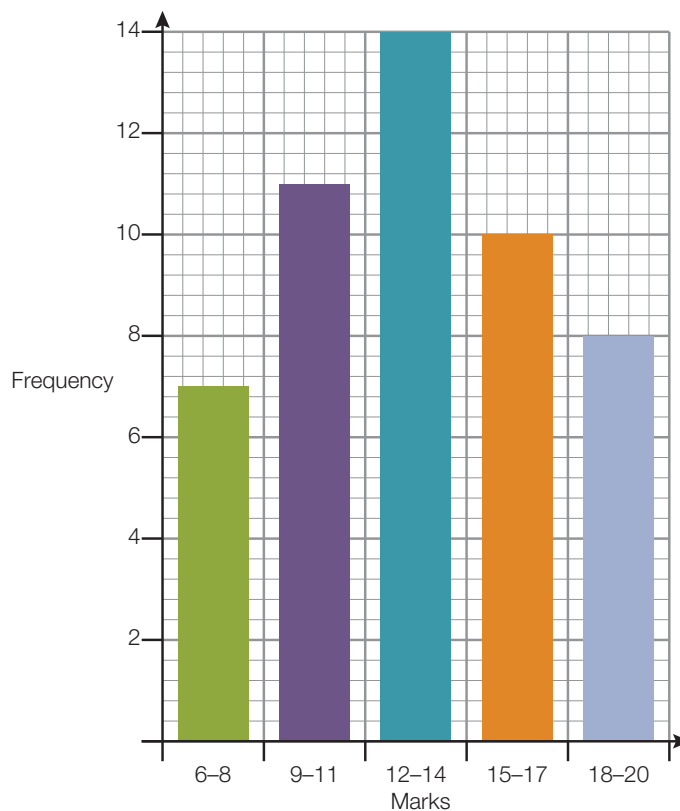
A.

Marks	Tally	Frequency
0–2		0
3–5		0
6–8		7
9–11		11
12–14		14
15–17		10
18–20		8
	Total frequency	50

continued . . .

From the frequency table, it is not possible to see how many students scored 6 marks, or 7 marks, or 8 marks; we can only see that there are seven people altogether whose marks are in the 6–8 group, so some detail has been lost because the data has been grouped. However, we can see a general pattern that suggests students tend to score between 9 and 17 marks.

A bar chart can be used to display grouped discrete data. Here, the groups are plotted along the horizontal axis; again, there should be spaces between the bars to indicate that the data is discrete. The frequency table in Worked example 5.3 can be converted to the following bar chart:



Exercise 5.3

1. The following data shows the number of students enrolled on the Mathematical Studies course in 28 accredited IB schools:

15, 24, 20, 11, 15, 12, 8, 20, 11, 21, 18, 5, 10, 5, 13, 11, 4, 10, 4, 9,

12, 12, 14, 14, 5, 15, 17, 4, 18, 13, 13, 21, 9, 11, 12, 14, 15, 21, 4, 24

Complete a frequency table with groups 0–4, 5–9, 10–14, etc.

2. The number of a certain band's CDs that were sold per month over 2.5 years are as follows:

52, 51, 71, 49, 50, 68, 52, 65, 48, 51, 67, 56, 58, 65, 68,

66, 66, 52, 67, 62, 67, 59, 74, 65, 70, 57, 64, 71, 67, 54

Complete a frequency table with groups, 45–49, 50–54, 55–59, etc.

In questions 3–6, draw a bar chart to represent the given set of data.

3. Examination marks of a group of students:

Exam mark	1–20	21–40	41–60	61–80	81–100
Number of students	3	5	6	12	8

4. Age of teachers, in completed years, at a high school:

Age of teachers	21–30	31–40	41–50	51–60
Frequency	12	23	17	8

5. Length of service, in full years, of head-teachers in a sample of county schools:

Length of service (years)	0–6	7–13	14–20	21–27	28–34
Frequency	9	18	25	12	6

6. The number of books borrowed from a school library by a group of students in an academic year:

Number of books	0–3	4–7	8–11	12–15	16–19
Number of students	3	6	15	11	7

5.4 Grouped continuous data

Continuous data can take infinitely many possible values depending on the degree of accuracy used in the measurement. Therefore, the total frequency will be spread very thinly across many data values and it is important to group the data before we can summarise it in a frequency table.

Grouping continuous data is not as straightforward as grouping discrete data; it can be quite difficult to define the smallest and largest value of each group because they can take an infinite number of values! Groups of continuous data are usually called **classes** instead, and the smallest and largest values of each class are called the class **boundaries**. The boundaries are defined by a given set of parameters and values are assigned to each class based on these parameters. There should be no gaps between the boundaries, so that no data values will be missed.

There are two main conventions for defining the class boundaries. Whichever one you use, the important point to remember is that all data

should be able to fit into a class, and that no data value should be able to fit into more than one class.

Convention 1:

lower boundary \leq data value $<$ upper boundary

In the first convention, the boundaries of a class are chosen so that all the measurements in that class are smaller than the upper boundary but greater than or equal to the lower boundary.

For example, the following frequency table gives the heights of 40 plants. The height of each plant was measured to the nearest centimetre, and then recorded. The data has been grouped in such a way that:

$0 \leq h < 5$ is the class containing all plant heights greater than or equal to 0 cm but less than 5 cm (it is often written simply as 0–5),

$5 \leq h < 10$ is the class containing all plant heights greater than or equal to 5 cm but less than 10 cm (it is often written simply as 5–10), and so on.

Height h (cm)	Frequency
$0 \leq h < 5$	5
$5 \leq h < 10$	7
$10 \leq h < 15$	13
$15 \leq h < 20$	8
$20 \leq h < 25$	5
$25 \leq h < 30$	2

Note that with this convention, the upper limit of each class is the same as the lower limit of the next class. This makes sure that all data values will fit into a class.

Convention 2:

lower boundary $- x \leq$ data value $<$ upper boundary $+ x$

In the second convention, there is a gap between the limits of consecutive classes as if you were grouping discrete data. But, each class actually begins or ends in the middle of the gap; any value that falls into a gap between two classes is rounded up or rounded down using the ' < 5 or ≥ 5 rule', to the closest class limit. x in the example above indicates the appropriate degree of accuracy to which the class boundaries are defined.

For example, the masses of players in a football competition were measured before the tournament began. The mass of each player was measured in kilograms to an accuracy of one decimal place, and the masses were grouped to the nearest kilogram; so using the ' < 5 or ≥ 5 rule', $x = 0.5$.

The following frequency table was compiled using the second convention of defining class boundaries:

61–65 is the class containing all masses greater than or equal to 60.5 kg (which will be rounded up to 61 kg) and less than 65.5 kg (which will be rounded down to 65 kg),

66–70 is the class containing all masses greater than or equal to 65.5 kg and less than 70.5 kg, and so on.

So the actual class boundaries are 60.5 ($61 - 0.5$), 65.5 ($65 + 0.5$), 70.5 ($70 + 0.5$) and so on, that is, values that lie in the middle of the gap between the upper limit of one class and the lower limit of the next.

Mass (kg)	Frequency
61–65	8
66–70	15
71–75	21
76–80	14
81–85	6
86–90	2

Be careful — sometimes data will be grouped according to a method that is neither of the two conventions described above. The boundaries of a class may be chosen so that all measurements in that class are less than or equal to the upper boundary and strictly greater than the lower boundary. For example:

Time (minutes)	Frequency
$40 < t \leq 50$	7
$50 < t \leq 60$	13
$60 < t \leq 70$	10
$70 < t \leq 80$	6
$80 < t \leq 90$	3

Always look carefully at how the classes are defined in a grouped frequency table.

As with discrete data, grouping continuous data will result in the loss of some original information. After the data is grouped, you can only tell how many items there are in each class, but not their original values.

Grouping the data does give a clearer overall picture, however, and prepares the data for any subsequent calculations.

Exercise 5.4

Each question gives the results of 30 students participating in a particular event on College Sports Day. Draw up a frequency table for each set of data, using appropriate classes.

- 100 m race times in seconds:

17.72	16.95	17.07	18.17	22.59	19.54	17.83	25.19	21.68	19.27
23.4	13.85	19.07	21.26	21.3	20.71	18.6	13.84	14.7	16.55
16.65	22.92	20.48	30.58	24.81	17.64	21.04	18.04	13.89	12.02

2. Shot put distances in metres:

10.55	7.6	8.91	9.3	8.6	6.31	11.25	11.62	11.41	9.56
6.73	9.3	7.96	7.58	13.67	5.47	7.2	8.2	6.62	5.81
12.35	6	10	6.17	11.17	7.97	14.04	6.82	9.16	9.19

3. Long jump distances in metres:

1.95	5.36	2.76	3.91	4.13	4.81	4.2	4.66	4.87	4.17
2.49	2.79	3.91	5.65	5.94	4.96	3.9	4.44	4.9	4.37
5.3	6.05	3.65	6.1	4.52	4.45	6.49	3.93	3.76	6.34

4. 400 m race times in seconds:

78.03	68.21	79.06	61.16	66.51	72.22	73.13	64.03	72.12	64.49
75.13	68.96	68.54	73.44	68.14	72.49	67.52	73.1	74.77	67.13
81.38	74.1	69.23	75.67	70.45	81.38	72.44	67.12	76.77	72.18

5. Javelin distance in metres:

26.29	32.67	41.36	39.93	39.55	38.08	32.5	37.24	29.28	27.45
32.55	37.38	33.92	35.72	44.97	33.98	42.08	29.42	35.92	41.02
40.23	36.82	33.09	32.25	48.79	32.46	47.24	26.61	37.82	30.48

5.5 Mid-interval values and upper and lower boundaries

You have already seen that grouped data is useful to identify general patterns but that the detail of the data is lost. What if you wanted to calculate some statistics about the data but you are only given the grouped frequency table? How can you do this without the detailed data? The mid-interval value allows you to **estimate** statistical values of the data.

In grouped data, the **mid-interval value** is the value that is half-way between the upper and lower boundaries of the group or class. It is the 'average' value of that particular group. When using the mid-interval value to make conclusions about our data, we assign the frequency for that group to this value as it is easier to deal with a 'representative' value for that group than it would be to deal with all the values of that group. This is why it only allows you to *estimate* statistical data, because not all points in that group will really be at the mid-interval value.

To calculate the mid-interval value for discrete data:

- The lower boundary of a group is the lowest value in that group.
- The upper boundary of a group is the highest value in that group.
- To find the mid-interval value of a group, add the upper and lower boundaries and divide the sum by two.



You will learn about measures of central tendencies (averages) in Chapter 6.

For example, for the group 12–14 in a set of discrete data:

- The lowest value is 12, so the lower boundary is 12.
- The highest value is 14, so the upper boundary is 14.
- $(12 + 14) \div 2 = 13$, so the mid-interval value is 13.

To calculate the mid-interval value for continuous data, the definition of the mid-interval value depends on how the data has been rounded and which convention has been used to define the classes.

Look again at the example of plant heights from the previous section. The classes are:

$0 \leq h < 5$, $5 \leq h < 10$, $10 \leq h < 15$, $15 \leq h < 20$, etc. and were defined using convention 1.

- The lower class boundaries are 0, 5, 10, ...
- The upper class boundaries are 5, 10, 15, ...
- Use the same sum as for discrete data to find the mid-interval value [(lowest class boundary + highest class boundary) \div 2]:
 $(0 + 5) \div 2 = 2.5$ is the mid-interval value for the first class;
 $(5 + 10) \div 2 = 7.5$ is the mid-interval value for the second class, and so on.

In summary, the full calculations for the plant height data are:

Height (cm)	Class boundaries	Frequency	Class width	Mid-interval value
$0 \leq h < 5$	0–5	5	$5 - 0 = 5$	$(0 + 5) \div 2 = 2.5$
$5 \leq h < 10$	5–10	7	5	7.5
$10 \leq h < 15$	10–15	13	5	12.5
$15 \leq h < 20$	15–20	8	5	17.5
$20 \leq h < 25$	20–25	5	5	22.5
$25 \leq h < 30$	25–30	2	5	27.5

Look again at the masses of the footballers from the previous section. The classes are:

61–65, 66–70, 71–75, etc. were defined using convention 2.

So in this case the lowest boundary is not, for example, 61 but is instead 60.5 because any value between 60.5 and 61 would have been rounded up to 61. Similarly, the highest boundary is not, for example, 65 but is instead 65.5 because any value between 65 and 65.5 would have been rounded down to 65.

- The lower class boundaries are 60.5, 65.5, 70.5, ...
- The upper class boundaries are 65.5, 70.5, 75.5, ...
- Use the same sum as before: $(60.5 + 65.5) \div 2 = 63$ is the mid-interval value for the first class; $(65.5 + 70.5) \div 2 = 68$ is the mid-interval value for the second class, and so on.



The upper and lower boundaries are harder to define for continuous data than they are for discrete data, and often different statisticians will come up with different answers! Think about why this situation has arisen.

The full calculations for the footballers' mass data are:

Mass (kg)	Class boundaries	Frequency	Class width	Mid-interval value
61–65	$60.5 \leq w < 65.5$	8	$65.5 - 60.5 = 5$	$(60.5 + 65.5) \div 2 = 63$
66–70	$65.5 \leq w < 70.5$	15	5	68
71–75	$70.5 \leq w < 75.5$	21	5	73
76–80	$75.5 \leq w < 80.5$	14	5	78
81–85	$80.5 \leq w < 85.5$	6	5	83
86–90	$85.5 \leq w < 90.5$	2	5	88

Remember the following points:

- If you have collected your own data and then organised it into classes, make sure that you have described the upper and lower boundaries of each class accurately.
- If you are using someone else's data, look at it carefully and make sure that you understand how they have defined the upper and lower class boundaries.

Exercise 5.5

1. The following table shows the times (in seconds) taken to run the 110-hurdles race in a decathlon competition.

Copy and complete the table by filling in the class boundaries, class width and mid-interval values

Time (s)	Class boundaries	Frequency	Class width	Mid-interval value
$18 \leq t < 20$	18–20	3	2	$(18 + 20) \div 2 = 9$
$20 \leq t < 22$		4		
$22 \leq t < 24$		6		
$24 \leq t < 26$		10		
$26 \leq t < 28$		3		
$28 \leq t < 30$		2		

2. Copy and complete tables like the one below for the sets of discrete data given in questions 3–6 of Exercise 5.3.

	Class boundaries	Frequency	Class width	Mid-interval value

3. Use your answers in Exercise 5.4 to complete tables like the one above for the sets of continuous data given in each question of the exercise.

5.6 Frequency histograms

To understand grouped continuous data and get a good idea of its shape, the most useful diagram is often a frequency **histogram**.

A frequency histogram is not the same as a bar chart, although they may look similar. In a histogram, the frequency of a data class is represented by the **area** of the bar, whereas in a bar chart it is represented by the height of the bar.

In a histogram, the width of a bar is defined by the lower and upper boundaries of that class of data. Since we are looking at continuous data, the upper boundary of each class will be the same as the lower boundary of the next class, so there will be **no gaps** between adjacent bars.

In this course we will consider only histograms which have bars of **equal width**, meaning that all the data classes have equal size. Note that if every bar has the same width, then the area of each bar is directly proportional to the height of that bar, and so the frequency of that data class will be represented by the height of the bar, just as it is in bar charts for discrete data. Bear in mind, though, that in histograms with bars of unequal width, the frequency of a data class is **not** directly represented by the bar height.

Also be aware that some histograms may look as if they have gaps between the bars – but that would be because some of the data classes are empty.

You can draw histograms by hand or by using your GDC or a computer graphing package. See section ‘5.2 Drawing a histogram’ on page 664 of the GDC chapter for a reminder of how to use your GDC if you need to.

When drawing a histogram by hand, you would plot the class intervals along the horizontal axis and the frequency for each data class along the vertical axis to get the bars. A GDC constructs a histogram using the **mid-interval** value and frequency of each class to draw the bars.

Let us plot histograms for the plant heights and footballers’ mass data given in section 5.4.

Plant heights:



TEXAS

L1	L2	L3	2
2.5	5	-----	
7.5	7		
12.5	13		
17.5			
22.5			
27.5			

L2(7) =			



CASIO

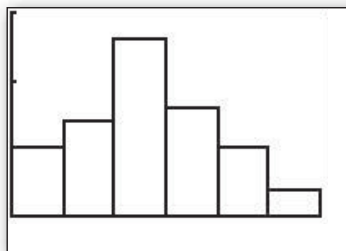
SUB	List 1	List 2	List 3	List 4
1	2.5	5		
2	7.5	7		
3	12.5	13		
4	17.5	86		
				5
GRAPH CALC TEST INTR DIST				

exam tip

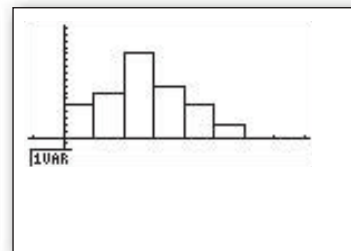
Any histograms that you encounter in examinations will have class intervals of equal width.



TEXAS



CASIO



Footballers' masses:

TEXAS

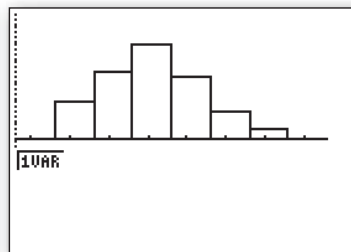
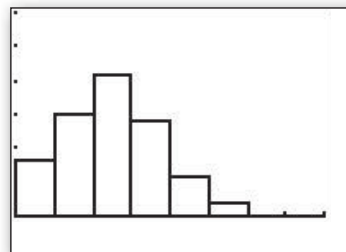
L1	L2	L3	2
63	8		
68	15		
73	21		
78	14		
83	6		
88	2		

L2(?) =			

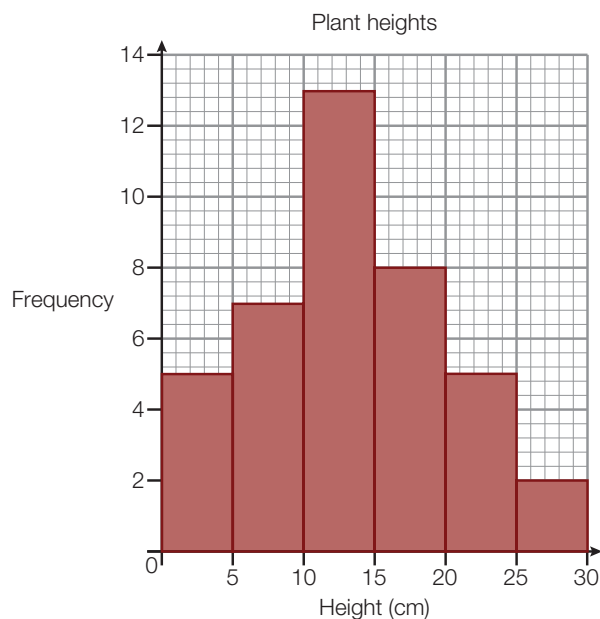
CASIO

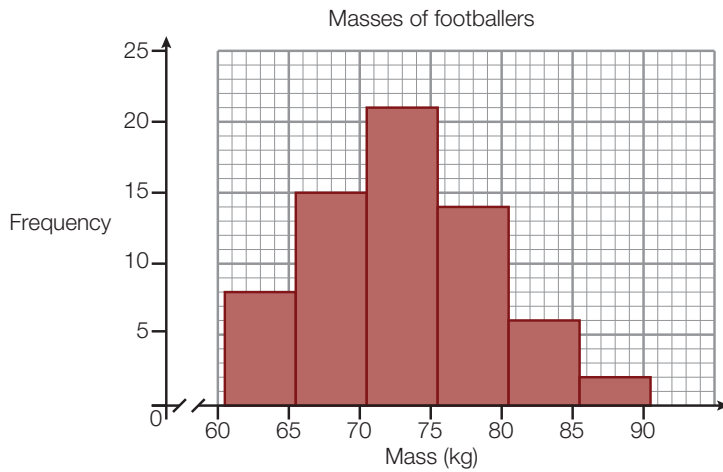
SUB	List 1	List 2	List 3	List 4
2	68	15		
3	73	21		
4	78	14		
5	83	6		

GPH1 GPH2 GPH3 SEL SET				



If you draw the histogram by hand or with a computer statistics package, you will be able to add more detail to your graph, such as labelling the axes and giving it a title:





Worked example 5.4

Q. Every day at noon, Fingal records the rainfall at his home near Sligo. The rainfall is measured in millimetres. These are his results for July 2012:

87	48	108	69	78	89	18	23	5	13	25	41	32
76	132	136	49	95	105	48	10	15	21	38	49	62
61	42	35	18	14								

Use the data to:

- (a) complete a frequency table (b) draw a histogram.

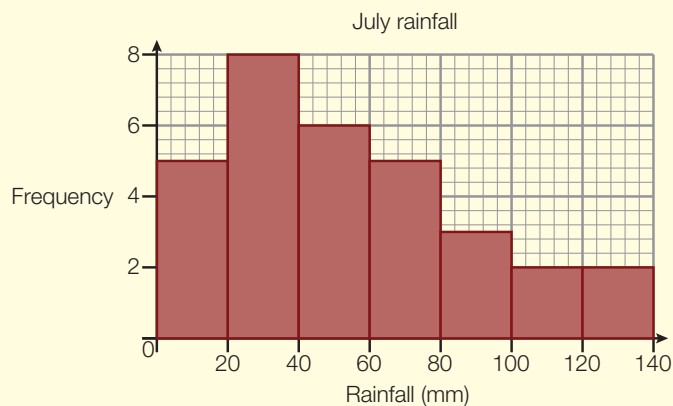
A.

Rainfall	Tally	Frequency
$0 \leq r < 20$		5
$20 \leq r < 40$		8
$40 \leq r < 60$		6
$60 \leq r < 80$		5
$80 \leq r < 100$		3
$100 \leq r < 120$		2
$120 \leq r < 140$		2
Total		31

Here, we have grouped the data using classes of width 20. This gives seven classes, which is a reasonable number.

Use a tally chart to find the frequency for each class.

Plot the histogram by hand.



Exercise 5.6

Each question presents a set of data. Copy and complete the given frequency table and hence draw the corresponding frequency histogram.

1. The number of days spent on revision for a final examination by a group of students:

10, 34, 37, 33, 12, 15, 34, 34, 34, 13, 28, 23, 34, 11, 13,

36, 23, 38, 31, 16, 16, 36, 26, 28, 12, 19, 19, 33, 19, 12

Number of days	Tally	Frequency
10–14		
15–19		
20–24		
25–29		
30–34		
35–39		

2. The thickness, in mm, of books on a library shelf:

11, 19, 49, 36, 63, 38, 56, 23, 26, 10, 38, 51, 86, 43, 83, 80, 63, 71, 28, 18,

80, 22, 70, 12, 68, 57, 21, 59, 68, 55, 47, 50, 62, 68, 46, 80, 21, 19, 14, 29

Thickness (mm)	Tally	Frequency
0–14		
15–29		
30–44		
45–59		
60–74		
75–89		

3. The time, to the nearest minute, taken to complete a given homework assignment:

43, 79, 85, 70, 71, 60, 78, 79, 77, 82, 80, 91, 57, 67, 75,

54, 64, 86, 69, 84, 72, 89, 79, 85, 63, 60, 74, 89, 62, 88,

80, 69, 63, 92, 93, 70, 80, 59, 81, 89, 42, 93, 56, 65, 82

Time (minutes)	Tally	Frequency
$40 < t \leq 50$		
$50 < t \leq 60$		
$60 < t \leq 70$		
$70 < t \leq 80$		
$80 < t \leq 90$		
$90 < t \leq 100$		

Note: the classes here are defined differently from either of the two conventions described in section 5.4.

4. The masses to the nearest gram of a random selection of apples before they are bagged on a farm:

Mass (g)	Frequency
151–190	2
191–230	12
231–270	17
271–310	6
311–350	3

Mass (g)	Class boundaries	Class width	Mid-interval value	Frequency
151–190	150.5–190.5	$190.5 - 150.5 = 40$	$(150.5 + 190.5) \div 2 = 170.5$	2
191–230				
231–270				
271–310				
311–350				

5. The volume of fuel (to the nearest litre) bought by 50 different drivers at a filling station in a two-hour period:

Volume (litres)	Frequency
20–39	5
40–59	12
60–79	20
80–99	9
100–119	4

Volume (litres)	Class boundaries	Class width	Mid-interval value	Frequency
20–39				5
40–59				12
60–79				20
80–99				9
100–119				4

6. A country's monthly iron ore production in millions of tonnes (for this question select your own classes):

8.3, 7.5, 8.6, 8.7, 8.9, 8.0, 7.5, 7.2, 7.6, 7.7, 8.1, 8.2, 8.2, 8.5,

7.9, 8.3, 8.1, 8.2, 8.3, 8.2, 8.3, 7.7, 8.3, 8.7, 8.5, 7.9, 8.1

Iron ore production (millions of tonnes)	Tally	Frequency

7. The masses of 30 tennis balls were recorded at the beginning of a tournament. The masses of the balls, in grams, are listed as follows:

56.1, 56.7, 55.3, 57.2, 56.2, 58.4, 58.1, 55.8, 59.3, 55.4,
58.0, 57.7, 58.3, 57.3, 57.0, 57.9, 56.9, 59.4, 58.1, 57.7,
57.1, 55.9, 56.7, 58.5, 56.6, 57.1, 57.8, 56.4, 57.8, 58.5

Mass, w (grams)	Class boundaries	Class width	Mid-interval value	Frequency
$55.0 < w \leq 55.5$				
$55.5 < w \leq 56.0$				
$56.0 < w \leq 56.5$				
$56.5 < w \leq 57.0$				
$57.0 < w \leq 57.5$				
$57.5 < w \leq 58.0$				
$58.0 < w \leq 58.5$				
$58.5 < w \leq 59.0$				
$59.0 < w \leq 59.5$				

Note: the classes here are defined differently from either of the two conventions described in section 5.4.

5.7 Cumulative frequency

Cumulative frequency is the total frequency up to a certain data value. For grouped data, you can calculate the cumulative frequency at the upper boundary of each data class by adding the frequency of that class to the cumulative frequency up to the class before. In other words, you keep track of the ‘running total’ of the frequency at each class boundary, showing how the frequency accumulates. The cumulative frequency allows you to make general statements about your data, for example ‘60% of children in my school have at least one pet’.

Cumulative frequency tables

To make a cumulative frequency table, you add an extra column to a frequency table. This column is used to keep track of the running total.

Worked example 5.5

Q. Add cumulative frequencies to the table of IB project marks that Ahmed compiled in Worked example 5.3 and make some observations about the data.

In the 'cumulative frequency' column you add the frequency of the class to the cumulative frequency of the class before it.

A.

Marks	Frequency	Cumulative frequency
0–2	0	0
3–5	0	0
6–8	7	7
9–11	11	18
12–14	14	32
15–17	10	42
18–20	8	50
Total frequency	50	

$$(0 + 0)$$

$$(0 + 7)$$

$$(7 + 11)$$

$$(18 + 14)$$

$$(32 + 10)$$

$$(42 + 8)$$

The value at the end of the cumulative frequency column is the same as the total frequency.

From the cumulative frequency you can see that:

18 students gained 11 marks or fewer.

32 students gained 14 marks or fewer.

42 students gained 17 marks or fewer.

Exercise 5.7

- The following table shows the hammer throw distances recorded in a competition. Complete the cumulative frequency table

Distance d (metres)	Frequency	Cumulative frequency
$55 < d \leq 60$	5	
$60 < d \leq 65$	7	
$65 < d \leq 70$	5	
$70 < d \leq 75$	8	
$75 < d \leq 80$	8	
$80 < d \leq 85$	9	
$85 < d \leq 90$	3	

2. The following table shows the mean homework marks of students in a school term. Complete the cumulative frequency table

Mark (%)	Frequency	Cumulative frequency
20–29	1	
30–39	2	
40–49	5	
50–59	10	
60–69	12	
70–79	4	
80–89	3	

3. Complete the following cumulative frequency table.

Test mark (%)	Frequency	Cumulative frequency
31–40	2	
41–50	6	
51–60	9	
61–70	15	
71–80	10	
81–90	7	
91–100	3	

Cumulative frequency curves (ogives)

Once you have created a cumulative frequency table, you can use it to draw a cumulative frequency curve or graph. This graph has a distinctive ‘S-shape’, and is sometimes called an **ogive**. You can use the curve to answer questions about the data up to a certain data value and you can use it make estimates about the data such as what frequency a certain data value might occur at, or what the frequency is likely be of a particular data value. This can useful if you want to use existing data to make predictions about data in the future.

You can also use the cumulative curve to estimate the ‘average’ data value. This is the data value at which 50% of the total frequency lies and the one you expect to occur most often. This is called the **median**. You can also use the cumulative curve tell you at what data point 25% of the frequency lies (the **lower quartile**), and where 75% of it lies (**upper quartile**). These three values can help you to spot patterns in your data.



You will learn more about the median and other measures of central tendency (averages) in Chapter 6, and more about quartiles in Chapter 7.

When you draw a cumulative frequency curve:

1. Mark the values of the **upper** class boundaries along the horizontal axis.
2. Plot the cumulative frequency values along the vertical axis.
3. Label the axes clearly.
4. Give the graph a title.

When using the graph to estimate the median or the values of the quartiles:

1. Draw a dashed line straight across from the vertical axis until you meet the cumulative curve.
 - (a) For the median, the value on the vertical axis will be at ' $50\% \times \text{total frequency}$ '.
 - (b) For the lower quartile, the value will be ' $25\% \times \text{total frequency}$ '.
 - (c) For the upper quartile, the value will be ' $75\% \times \text{total frequency}$ '.
2. From the point where you meet the curve, draw a straight vertical line (dashed) down to the horizontal axis and read off the value.

It is important that you include these dashed lines in your working to demonstrate how you obtained each value.

Worked example 5.6

- Q. The total lung capacity (TLC) of 120 female members of a sports club was recorded. The data was grouped into classes of width 0.25 litres and the cumulative frequency was calculated. Draw a cumulative frequency curve of the data and make observations about the data.

Capacity (litres)	Frequency	Cumulative frequency
3.00–3.25	0	0
3.25–3.50	4	4
3.50–3.75	11	15
3.75–4.00	23	38
4.00–4.25	32	70
4.25–4.50	21	91
4.50–4.75	13	104
4.75–5.00	8	112
5.00–5.25	5	117
5.25–5.50	3	120

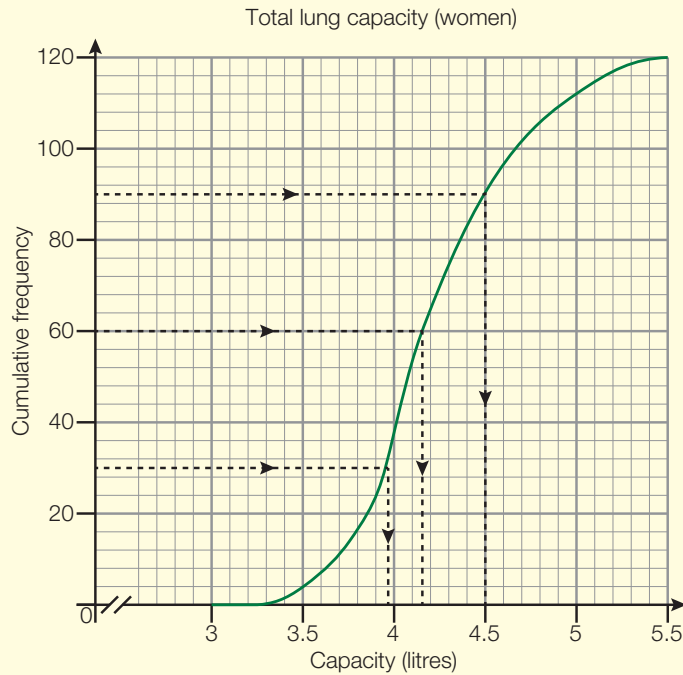
continued . . .

The lung capacity, in litres, is plotted on the horizontal axis and the cumulative frequency on the vertical axis.

The first point plotted is (3.25, 0), the second is (3.50, 4), the third is (3.75, 15), and so on, up to the final point (5.50, 120). Note that the x -coordinates of the plotted points are the upper class boundaries.

Draw a horizontal line from 60 on the vertical axis to the curve, and then vertically down to the horizontal axis. The dashed line reaches the horizontal axis somewhere between 4.1 and 4.2, so an estimate for the median would be 4.15 litres.

To obtain these estimates, you would draw a vertical line from the relevant value on the horizontal axis (3.6 litres or 4.8 litres) **up** to the curve and then horizontally to the left until you reach the vertical axis.



- A. From the graph it is possible to read off estimates for a number of statistical data:
- The TLC achieved by 50% of the women (60 women) is about 4.15 litres. This is an estimate of the **median**.
 - The TLC achieved by 75% of the women (90 women) is about 4.5 litres. This is the **upper quartile**.
 - The TLC achieved by 25% of the women (30 women) is about 3.95 litres. This is called the **lower quartile**.

From the cumulative frequency curve it is also possible to estimate that, for example:

- About 8 women have a TLC of less than 3.6 litres.
- About 105 women have a TLC of less than 4.8 litres.
- 15 ($= 120 - 105$) women have a TLC greater than 4.8 litres.

exam tip

If you draw your cumulative frequency graph by hand, the values that you estimate from it may differ from those given by a software package or GDC. If you estimated values for the median and quartiles by drawing lines correctly on your graph, the examiner will accept those values because you have shown how you obtained them.

Worked example 5.7

- Q. A company has 180 employees, and the company record summarises their salaries as follows:

Salary s (\$,000)	Frequency	Cumulative frequency
$10 \leq s < 15$	8	8
$15 \leq s < 20$	13	21
$20 \leq s < 25$	19	
$25 \leq s < 30$	26	
$30 \leq s < 35$	32	
$35 \leq s < 40$	35	
$40 \leq s < 45$	20	
$45 \leq s < 50$	13	
$50 \leq s < 55$	10	
$55 \leq s < 60$	4	

- (a) Complete the cumulative frequency column.
 (b) Draw a cumulative frequency curve to represent the information.
 (c) Use the graph to find an estimate for the median salary.
 (d) Use the graph to calculate the percentage of the employees with:
 (i) a salary below \$22,000
 (ii) a salary above \$58,000.
 (e) Use your graph to estimate the upper quartile and the lower quartile.

Calculate the cumulative frequency in each row by adding the frequency in that row to the cumulative frequency of the previous row.

- A. (a)

Salary s (\$,000)	Frequency	Cumulative frequency
$10 \leq s < 15$	8	8
$15 \leq s < 20$	13	$8 + 13 = 21$
$20 \leq s < 25$	19	$21 + 19 = 40$
$25 \leq s < 30$	26	$40 + 26 = 66$
$30 \leq s < 35$	32	$66 + 32 = 98$
$35 \leq s < 40$	35	$98 + 35 = 133$
$40 \leq s < 45$	20	$133 + 20 = 153$
$45 \leq s < 50$	13	$153 + 13 = 166$
$50 \leq s < 55$	10	$166 + 10 = 176$
$55 \leq s < 60$	4	$176 + 4 = 180$

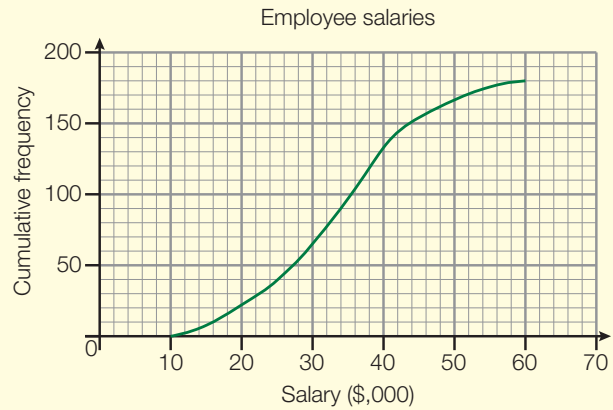
continued . . .

Use the upper class boundaries and corresponding cumulative frequencies to plot the curve.

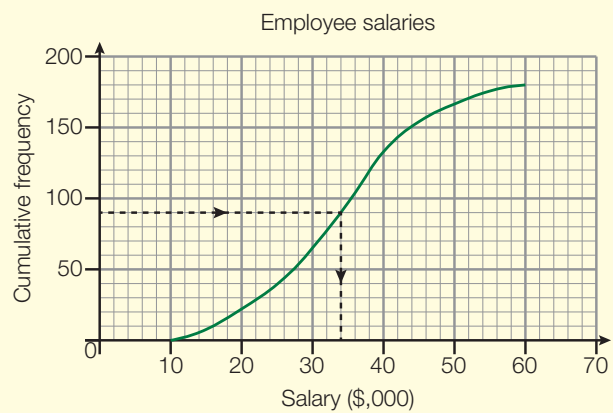
$50\% \times 180 = 90$. On the cumulative frequency graph, draw a line horizontally across from 90 on the vertical axis to the curve, and then vertically down until it meets the horizontal axis.

Draw a line vertically up from \$22,000 on the horizontal axis to the curve, and then horizontally to the left to the vertical axis. This tells you that up to 28 employees earn less than \$22,000. The question asks what 'percentage' of employees, so you need to work out what percentage of 180 employees 28 is.

(b)

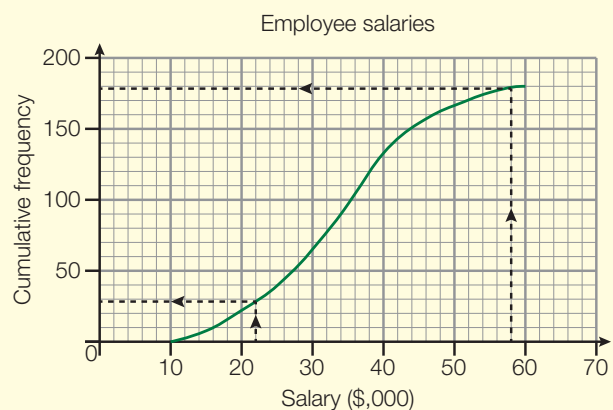


(c)



The median salary is approximately \$34,000.

(d)



(i) Approximately 28 employees earn less than \$22,000:

$$\frac{28}{180} \times 100\% = 15.6\%$$

So, about 16% (to the nearest per cent) of employees earn less than \$22,000.



continued . . .

Draw a line vertically up from 58 on the horizontal axis to the curve, and then horizontally to the left until it hits the vertical axis.

$$75\% \times 180 = 135;$$

$$25\% \times 180 = 45.$$

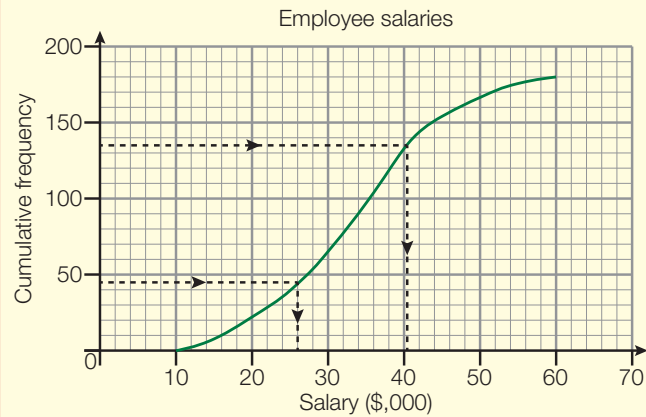
Draw lines horizontally across from 135 and 45 on the vertical axis to the curve, and then vertically down until they meet the horizontal axis.

(ii) Approximately 179 employees earn less than \$58,000, so around $180 - 179 = 1$ employee earns more than \$58,000:

$$\frac{1}{180} \times 100\% = 0.556\%$$

1% of employees (to the nearest percent) earn more than \$58,000.

(e)



The upper quartile of salaries is about \$40,500. The lower quartile of salaries is about \$26,000.

This means 75% of employees earn up to \$40,500 and about 25% of employees earn \$26,000 or less.

Exercise 5.8

- The following table shows the times achieved by Olympic gold medallists in the men's 100-metres final from 1896 to 2008.

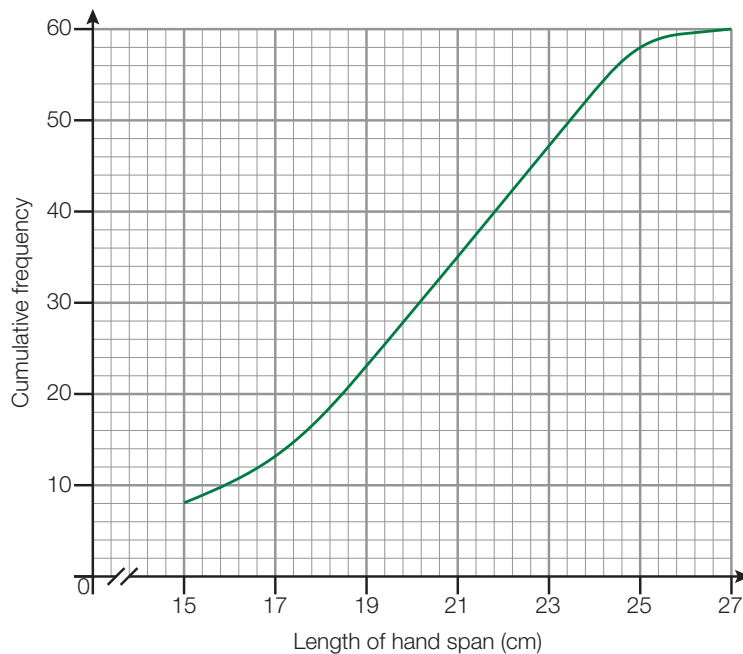
Time (seconds)	Frequency	Cumulative frequency
$9.60 \leq t < 10.00$	20	
$10.00 \leq t < 10.40$	30	
$10.40 \leq t < 10.80$	12	
$10.80 \leq t < 11.20$	14	
$11.20 \leq t < 11.60$	2	
$11.60 \leq t < 12.00$	0	
$12.00 \leq t < 12.40$	1	
$12.40 \leq t < 12.80$	2	

- Complete the cumulative frequency table.
- Draw a cumulative frequency curve to represent the information.
- Find an estimate of the median time.

2. The following table shows the distances achieved by medallists in the women's long jump in the Olympic Games from 1948 to 2008.

Distance, d (metres)	Frequency	Cumulative frequency
$5.50 \leq d < 5.75$	3	
$5.75 \leq d < 6.00$	1	
$6.00 \leq d < 6.25$	5	
$6.25 \leq d < 6.50$	4	
$6.50 \leq d < 6.75$	7	
$6.75 \leq d < 7.00$	11	
$7.00 \leq d < 7.25$	15	
$7.25 \leq d < 7.50$	1	

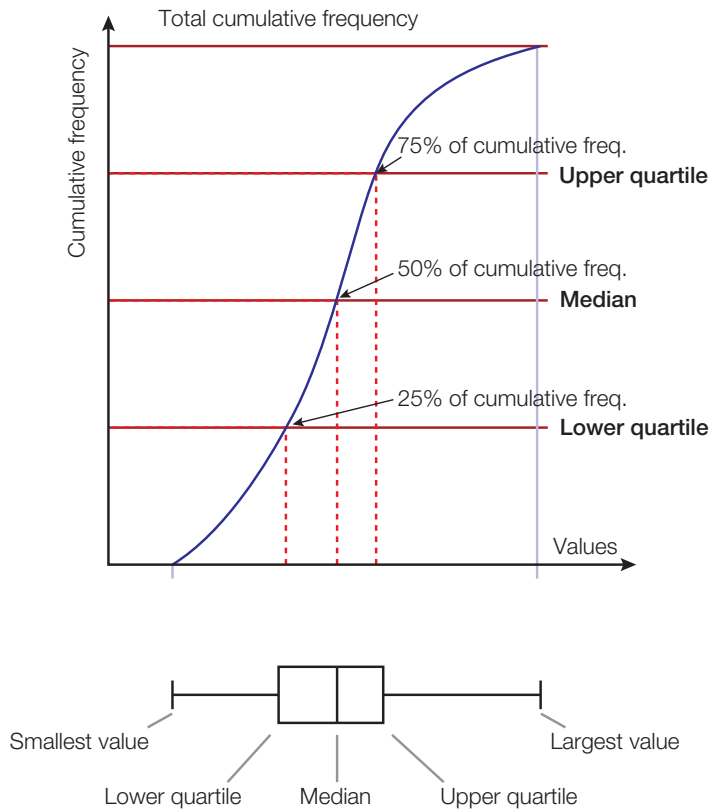
- (a) Complete the cumulative frequency table.
- (b) Draw a cumulative frequency curve to represent the information.
- (c) Find an estimate for the median distance achieved.
- (d) Find estimates for the lower quartile and the upper quartile.
3. The hand spans of 60 female students are represented on the cumulative frequency diagram below.



- (a) Find an estimate of the median length of hand span.
- (b) Estimate the number of students with a hand span less than 22.5 cm.
- (c) Calculate the percentage of students with a hand span greater than 24 cm.

5.8 Box and whisker diagrams

The picture below shows how cumulative frequency curves are related to another sort of statistical diagram called a **box and whisker diagram** or box plot.



A box and whisker diagram summarises five important values from a set of data, and gives you a simple picture of the data. It is also useful when you have two or more sets of data and wish to make comparisons between them.

The information displayed by a box and whisker diagram is sometimes called a **five-figure summary**. The five figures are:

- the smallest value (minimum) of the data
- the lower quartile (Q_1)
- the median (Q_2)
- the upper quartile (Q_3)
- the largest value (maximum) of the data.

Revisiting the example on total lung capacity in the previous section (Worked example 5.6), the five figures are:

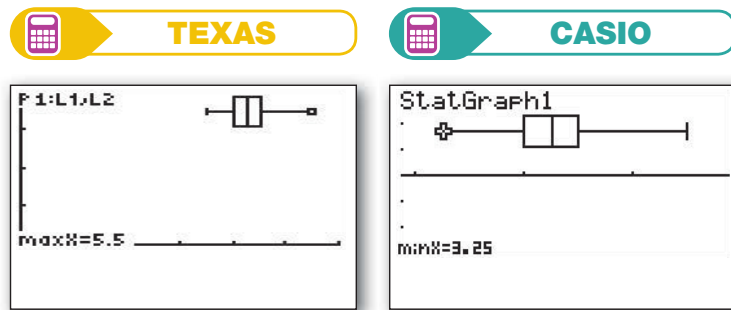
- minimum 3.25 litres
- lower quartile 3.95 litres

- median 4.15 litres
- upper quartile 4.5 litres
- maximum 5.5 litres



You can use your GDC to draw box and whisker diagrams. See section '5.3 Drawing a box and whisker diagram' on page 665 of the GDC chapter if you need a reminder of how to do this.

The box and whisker diagram for the lung capacity data looks like this:



If you are drawing the box and whisker diagram on paper (from scratch or by copying it from your calculator), remember to put in a heading and, more importantly, a scale.

Note that:

- half, or 50%, of the data lies between Q_1 and Q_3
- a quarter, or 25%, of the data lies between the smallest value and Q_1
- three quarters (75%) of the data lies between the smallest value and Q_3 (note that this also means that a quarter of the data lies between Q_3 and the largest value).

Exercise 5.9

1. For each of the following sets of data, draw a box and whisker diagram by hand.

- | | |
|---------------------------|------|
| (a) Minimum mark (%): | 35 |
| Lower quartile: | 48 |
| Median: | 56.5 |
| Upper quartile: | 67 |
| Maximum mark: | 82 |
| (b) Minimum distance (m): | 10 |
| Lower quartile: | 22 |
| Median: | 48 |
| Upper quartile: | 65 |
| Maximum distance: | 86 |

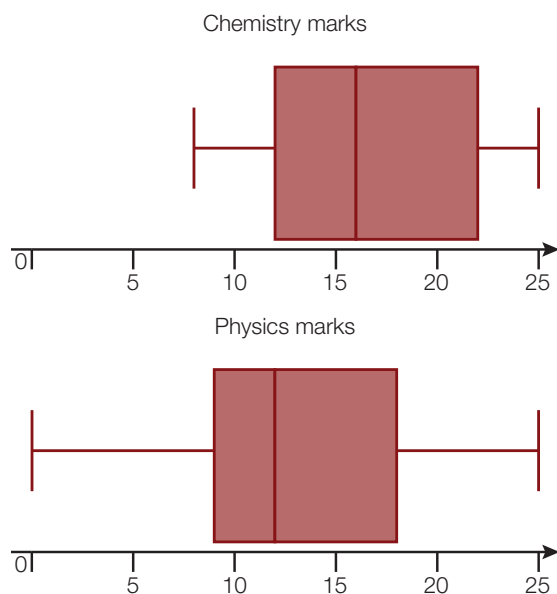
(c) Minimum time (s):	12.0
Lower quartile:	17.0
Median:	18.8
Upper quartile:	21.3
Maximum time:	25.2
(d) Minimum height (m):	61.2
Lower quartile:	68.2
Median:	71.8
Upper quartile:	73.5
Maximum height:	81.4

2. Use your GDC to draw box and whisker diagrams for the following sets of data. State the five-figure summary in each case.

- (a) 59, 35, 64, 43, 83, 46, 51, 71, 54, 61, 89, 77, 47, 71, 74, 84, 76, 54, 51, 86, 61, 65, 63
- (b) 49.4, 48.8, 42.6, 49.1, 45.6, 45.3, 50.6, 35.5, 45.6, 48.5, 52.1, 32.9, 56.8
- (c) 33.72, 39.87, 48.51, 23.05, 41.93, 36.76, 43.9, 40.74, 28.07, 49.1, 54.73, 53.16
- (d) 141.13, 84.6, 188.44, 172.45, 175.82, 152.03, 155.83, 166.07, 159.94, 163.01, 150.08, 117.09, 133.81, 152.64, 171.24, 111.98, 142.78, 119.22

Using box and whisker diagrams

The box and whisker diagrams below are being used to compare the marks of two study groups. The upper diagram gives the five-figure summary for the group studying IB Chemistry HL. The lower diagram presents the five-figure summary for a group of students studying IB Physics HL.



exam tip

The treatment of outliers will not be part of the examinations. However, you may come across outliers when you are doing coursework for other subjects, or when you are working on your project for this course.

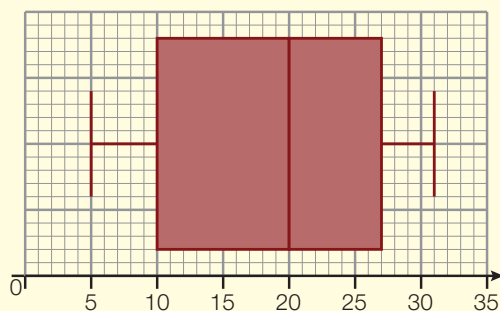
These diagrams show that:

- The median mark for Chemistry is higher than that for Physics.
- The distance between Q_1 and Q_3 is smaller for Physics than for Chemistry.
- The lowest mark in Physics is lower than the lowest mark in Chemistry. (In fact, the lowest Physics mark is so low that it has skewed the diagram; an unusually low or high value in the data is called an **outlier**, a piece of data that is so different from the rest that it can cause a distortion in calculations done with the data.)
- The top marks are the same for each study group.

Worked example 5.8

Q. As part of her project, Cécile counts the number of words in the sentences of a newspaper. The box and whisker diagram for her data is shown below. Use it to write down:

- (a) the maximum and minimum number of words in a sentence
- (b) the median number of words in the newspaper sentences
- (c) the upper quartile (Q_3) and the lower quartile (Q_1).



Look at the right and left ends of the whiskers.

A. (a) The maximum number of words in a sentence is 31; the minimum number is 5.

Look at the vertical line inside the box.

(b) The median number of words is 20.

Look at the right and left ends of the box.

(c) The upper quartile is 27 and the lower quartile is 10.

Worked example 5.9

- Q. Every day at noon, Fingal records the temperature at his home near Sligo. The temperatures are expressed in degrees Celsius ($^{\circ}\text{C}$) to the nearest degree. These are his results for July 2012.

21	13	17	16	20	12	15	22	14	20	22	14	16
12	13	12	18	21	15	15	13	12	21	22	16	15
22	21	21	18	14								

Use the data to:

- make a frequency table
- draw a box and whisker diagram on your GDC
- find the values of the lower quartile, median and upper quartile using your GDC.

Make a tally chart. To avoid losing any detail, we do not group the data.

A.

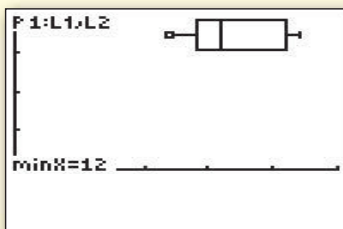
Temperature ($^{\circ}\text{C}$)	Tally	Frequency
12		4
13		3
14		3
15		4
16		3
17		1
18		2
19		0
20		2
21		5
22		4
Total		31

Draw a box and whisker diagram using your GDC. See section '5.3 Drawing a box and whisker diagram' on page 665 of the GDC chapter if you need a reminder of how to do this.

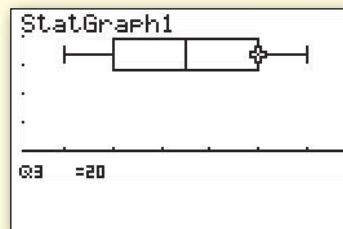
(b)



TEXAS



CASIO

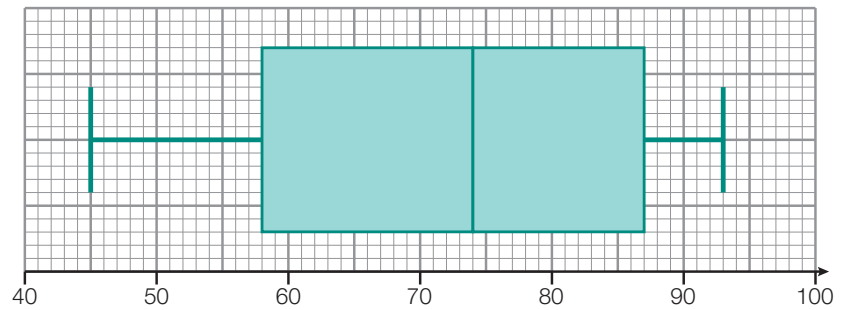


Read off the median and quartiles and write down the answers appropriately.

(c) From GDC: $Q_1 = 14$, $Med = 17$, $Q_3 = 20$

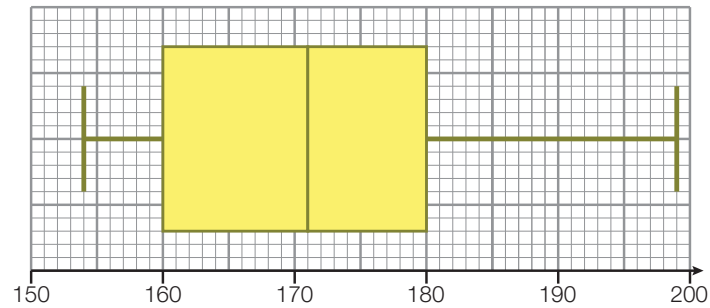
Exercise 5.10

1. The examination results of students in a class are summarised in the following box and whisker diagram.



Using the information from the diagram, find:

- the minimum and maximum scores
 - the median
 - the lower quartile
 - the upper quartile.
2. The box and whisker diagram below shows the heights of members of a sports club.



From the diagram, write down:

- the median
 - the highest and lowest values
 - the lower quartile and the upper quartile.
3. The following list shows the total number of tries scored by each team in a rugby league.

16	17	21	16	15	14	23	15	23	15	14	21	17
15	22	20	21	17	18	18	15	20	22	22	14	19
14	14	18	22	23	19	18	20	19	21	21	17	23
19	20	18	14	21	18	16	19	18	16	16		

(a) Copy and complete the given frequency table.

Number of tries	Tally	Frequency
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
Total		

(b) Draw a box and whisker diagram.

(c) Give the values of the lower quartile, median and upper quartile found with your GDC.

Summary

You should know:

- the classification of data as qualitative or quantitative, and the classification of quantitative data as discrete or continuous
- how to draw up frequency tables and present simple discrete data
- how to deal with grouped discrete or continuous data and draw up frequency tables
- how to define upper and lower class boundaries and mid-interval values
- how to use a GDC to draw frequency histograms
- how to produce cumulative frequency tables and curves for grouped discrete or continuous data
- how to find medians and (upper and lower) quartiles from cumulative frequency curves
- how to draw box and whisker diagrams using your GDC and find five-figure summaries.

Mixed examination practice

Exam-style questions

1. The following table shows the number of items in customers' shopping baskets at a check-out.

Number of items	1–3	4–6	7–9	10–12	13–15	16 or more
Number of shoppers	2	10	12	10	8	5

Draw a bar chart to represent the data.

2. The heights of 36 students, to the nearest centimetre, are given in the list below:

172, 162, 175, 172, 173, 175, 175, 162, 168, 166, 163, 179,

182, 171, 175, 186, 169, 165, 168, 172, 171, 164, 165, 170,

169, 168, 172, 170, 177, 175, 169, 166, 187, 162, 175, 161

- (a) Complete the following frequency table.

Height h (cm)	Tally	Frequency
$160 < h \leq 165$		
$165 < h \leq 170$		
$170 < h \leq 175$		
$175 < h \leq 180$		
$180 < h \leq 185$		
$185 < h \leq 190$		

- (b) Draw a frequency histogram to represent the data.

3. The following data represents the distances, in metres, recorded in a triple jump competition:

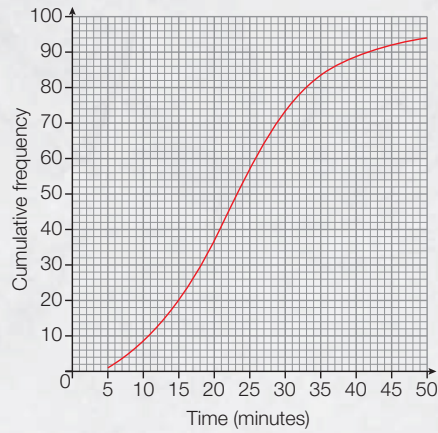
8.7	8.57	7.53	7.1	7.99	10.89	6.15	9.23	8.97	11.05
7.56	7.09	10.35	9.38	6.79	11.23	9.93	11.69	9.19	9.08
10.71	7.75	10.46	11.45	10.46	6.38	10.25	8.92	6.25	12.34

- (a) Complete the frequency table.

Distance d (metres)	Frequency
$5.00 \leq d < 6.50$	
$6.50 \leq d < 8.00$	
$8.00 \leq d < 9.50$	
$9.50 \leq d < 11.00$	
$11.00 \leq d < 12.50$	

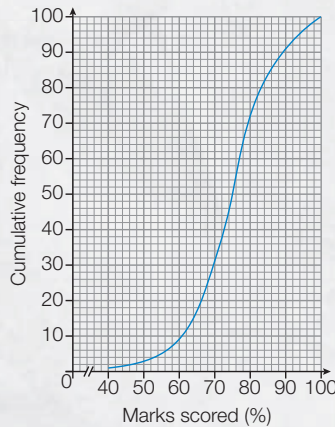
- (b) Create a cumulative frequency table and hence use your table to draw a cumulative frequency curve.
- (c) Use the appropriate summary data from your answer in part (b) to draw a box and whisker diagram.

4. The cumulative frequency graph shown below represents the time taken to travel to school by a group of 94 students.



From the graph, answer the following questions.

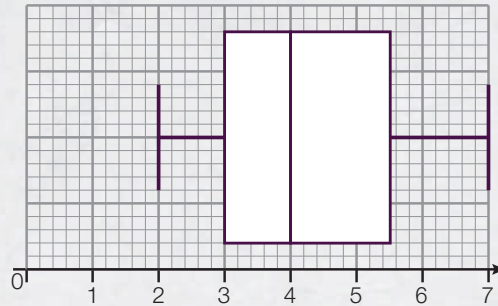
- Write down the median time.
 - Write down the lower quartile and the upper quartile for the times taken to travel to school.
 - Estimate the number of students who take longer than 38 minutes to travel to school.
 - Given that the minimum and maximum times are 5 minutes and 50 minutes, respectively, draw and label a box and whisker diagram for the data.
5. The scores of all the students who sat a mock examination are summarised in the cumulative frequency curve below.



- From the graph, find:
 - the number of students sitting the examination
 - the median examination mark
 - the upper and lower quartile marks.
- How many students scored between 61% and 75% inclusive?

- (c) Given that the minimum score was 40% and the maximum was 100%, draw and label a box and whisker diagram for the data.
- (d) Any student scoring more than 84% was awarded a Level 7. Estimate the number of students who were awarded a Level 7.

6. The test results of 25 students in Group A1 are displayed in the box and whisker diagram below.

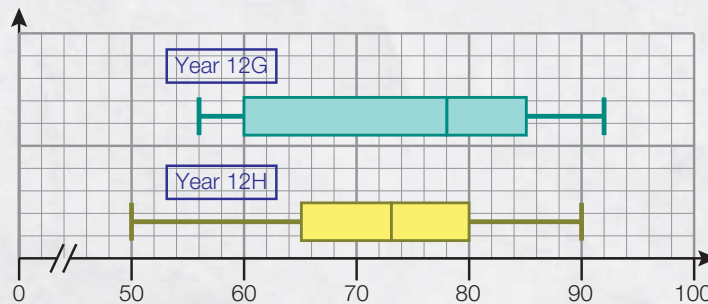


- (a) Find the range of the results (that is, the difference between the maximum and minimum marks).
- (b) Find the lower quartile and the upper quartile mark.
- (c) What is the median mark?

The test results of students in Group A2 are as follows:

1, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 7

- (d) Represent these results in a box and whisker diagram.
- (e) Compare the results of the two groups by citing two differences in their performance.
7. The performance of two Year 12 groups in the same test is illustrated on the two box and whisker diagrams below.



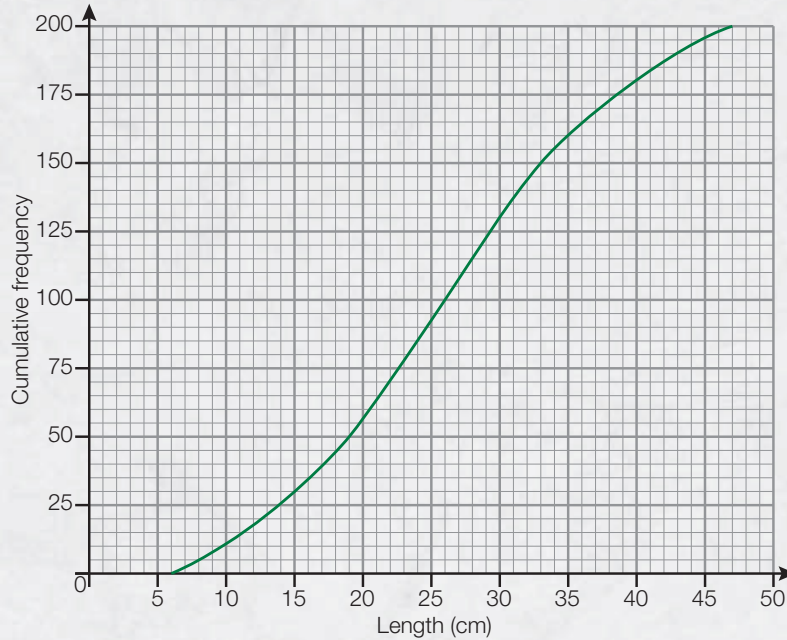
(a) Complete the following table of summary statistics.

	Year 12G	Year 12H
Median		
Lower quartile		
Upper quartile		

(b) Use the results from part (a) to compare the performance of the two groups in the test.

Past paper questions

1. A random sample of 200 females measured the length of their hair in cm. The results are displayed in the cumulative frequency curve below.

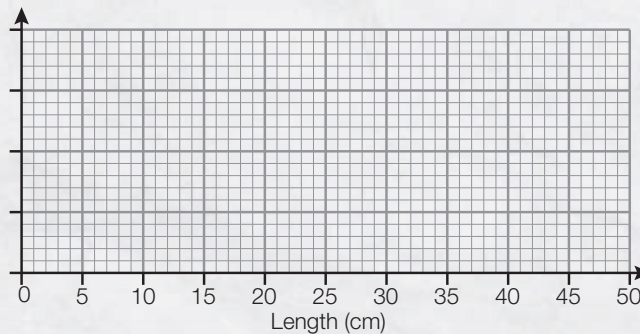


- (a) Write down the median length of hair in the sample. [1 mark]
- (b) Find the interquartile range for the length of hair in the sample. [2 marks]



The interquartile range is the difference between Q_1 and Q_3 . You will learn more about this in Chapter 7.

- (c) Given that the shortest length was 6 cm and the longest 47 cm, draw and label a box and whisker plot for the data on the grid provided below.



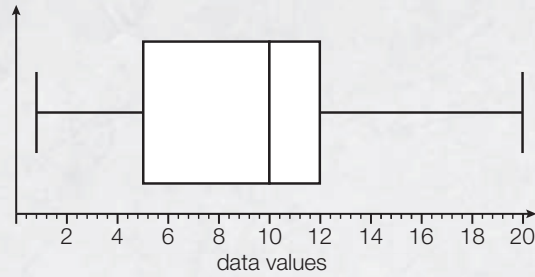
[3 marks]

[May 2008, Paper 1, Question 13] (© IB Organization 2008)

2. (a) State which of the following sets of data are discrete.
- Speeds of cars travelling along a road.
 - Numbers of members in families.

- (iii) Maximum daily temperatures.
- (iv) Heights of people in a class measured to the nearest cm.
- (v) Daily intake of protein by members of a sporting team.

The boxplot below shows the statistics for a set of data.



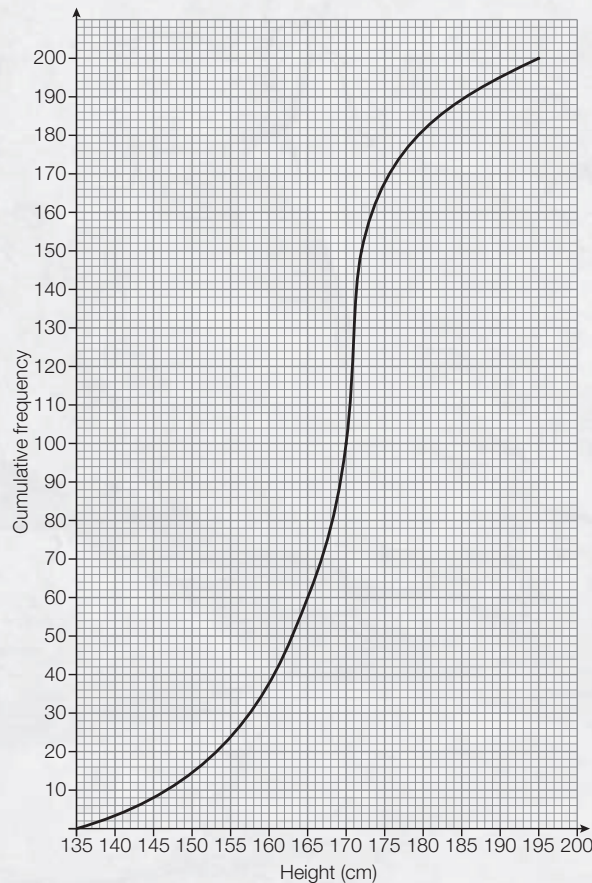
(b) For this data set write down the value of

- (i) the median
- (ii) the upper quartile
- (iii) the minimum value present

[6 marks]

[May 2007, Paper 1, Question 2(a)(b)] (© IB Organization 2007)

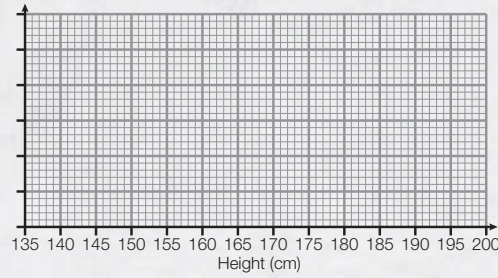
3. A cumulative frequency graph is given below which shows the height of students in a school.



- (a) Write down the median height of the students. [1 mark]
- (b) Write down the 25th percentile. [1 mark]
- (c) Write down the 75th percentile. [1 mark]

The height of the tallest student is 195 cm and the height of the shortest student is 136 cm.

- (d) Draw a box and whisker plot on the grid below to represent the heights of the students in the school.



[3 marks]

[May 2009, Paper 1, Question 5] (© IB Organization 2009)

exam tip

The 25th and 75th percentiles are the same as the lower and upper quartile.

Chapter 6 Measures of central tendency

In this chapter you will learn:

- several ways of measuring central tendency (mean, median and mode)
- how to find the median for simple discrete data
- how to find the mean and mode for simple discrete data, for grouped discrete data and for continuous data
- how to calculate an estimate for the mean from grouped data
- how to recognise the modal class.



"Add the numbers, divide by how many numbers you've added and there you have it—the average amount of minutes you sleep in class each day."



Can you describe yourself as 'average'? Do you know anyone whom you think is average? Is this concept helpful to organisations such as insurance companies, public health bodies or governments?

The word 'average' is often used to summarise a whole set of data, or even an entire population, with one single number.

In statistics, the everyday word 'average' is one example of a **measure of central tendency**.

The three most common ways of describing the average of a population or sample of data are described below:

- The **median** is the central (middle) value of a data set whose values have been arranged in order of size.
- The **mean** is the sum of all the data values divided by the total number of data values in the set.
- The **mode** is the data value that occurs most frequently.

To see the difference between these three measures, let's look at how they are calculated for a given set of data.

Suppose that Rita records the length of time that she spends on every assignment that her Maths teacher sets over the course of one term. She would like to know the 'average' length of time that she takes to complete her homework. She measures her times to the nearest minute.

In one term the times are 43, 51, 53, 52, 75, 95, 36, 43, 37, 67, 87, 58 and 56 minutes.

First, she puts the times in order from least to greatest time; then she picks out the middle value. This gives her the **median**. This is 53 minutes.

36, 37, 43, 43, 51, 52, **53**, 56, 58, 67, 75, 87, 95

There are 13 data values. The central value is '53' because there are the same number of values above it as there are below it (six above and six below).

Next, she adds up all the times and divides by the total number of assignments she was set (13). This gives the **mean**:

$$\begin{aligned} &(43 + 51 + 53 + 52 + 75 + 95 + 36 + 43 + 37 + 67 + 87 + 58 + 56) \div 13 \\ &= 753 \div 13 \\ &= 57.9 \text{ minutes.} \end{aligned}$$

She also looks for the value that occurs most frequently, this is the **mode**. This is 43 minutes, as this time is listed twice and all other times are only recorded once.

43, 51, 53, 52, 75, 95, 36, **43**, 37, 67, 87, 58, 56

Rita's results are:

Mean	57.9 minutes
Median	53 minutes
Mode	43 minutes

You can see that each measure of central tendency gives a slightly different answer. Which one should Rita use?

Rita's mean is the highest measure of the three at 57.9 minutes. If you look at the list of data, you can see that it is influenced by two values, 87 and 95, that are considerably larger than any of the other values.



Adolph Quetelet, (1796–1874) was born in Ghent (in present-day Belgium). He was a gifted student and was interested in poetry, art and languages as well as mathematics and astronomy. In 1823 he went to Paris to study astronomy and the theory of probability, returning to Belgium to give his first lectures on probability in 1824. From this time on, he worked hard to find a mathematical way of describing 'the average man', but could not find one measure that was suitable for all sets of data and situations. His work was fundamental to the later development of statistics, and to understanding the way in which statistical results can be used.



Which figure do you think Rita will quote when she is talking to her friends? Which one do you think she will tell her Maths teacher? Is this difference important?

hint

As Rita's homework times were measured, the data was continuous. However, she rounded the times to the nearest minute and treated the data as discrete. This practice is quite common when data is being collected and analysed. It is a good idea to inspect any set of data and think about whether the figures have been rounded in this way, and whether the rounding might have affected the results.

In finding the median, the two larger values are balanced by two lower values when ordering the data.

Rita's mode is quite a lot lower than either the median or the mean and, looking at all the data points, does not seem like a very good representation of the average time.

Rita can use any of the three values, which are all correct, but as they are all different you can see why she can't simply just say the 'average'. She needs to decide which measure to use in a given situation, and be able to explain why she has chosen that particular measure.

There are different methods of working with the mean, median and mode that depend on the type of data being analysed.

6.1 Finding the median for simple data

The median is the middle value when the data is presented in an ordered list. So, if the data is presented as a simple list of numbers, then to find the median:

1. Put the numbers in order of size, usually from smallest to largest.
2. Identify what position the middle value is in as follows:
 - (a) If the total frequency is an **odd number**, use the formula $\frac{1}{2}(n+1)$, where n is the (odd) number of data values. For example:
1, 5, 8, 9, 9, 9, 9, 10, 11, 14, 15

There are 11 data values so $n = 11$; substitute this into the formula: $11 + 1 = 12$, and $12 \div 2 = 6$, so the sixth value is the middle value and the value at this position is the median.

$$n = 11, \text{ so } \frac{1}{2}(11+1) = 6$$

Value at position 6 is 9.
The median is 9.

- (b) If there is an **even number** of data values, then there will be **two** middle values and the median is calculated by working out the mean of these two values. The two middle values are those that are on either side of the value obtained by using the formula $\frac{1}{2}(n+1)$. For example:

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

Find the mean of the two middle values:
 $(7 + 8) \div 2 = 7.5$.
The median is 7.5.

There are 10 data values so $n = 10$; substitute this into the formula: $10 + 1 = 11$, and $11 \div 2 = 5.5$. The positions either side of 5.5 are the 5th position and the 6th position; the two middle values are 7 and 8.



You can also use your GDC to find the median of a list of numbers. See section '6.1 (a) Finding the mean, median, quartiles and standard deviation for a simple list of data (single variable, no frequency)', on page 666 of the GDC chapter if you need a reminder of how. In the list of statistics provided by your GDC, the median is labelled 'Med' or 'Q₂'.

**TEXAS****CASIO**

L1	L2	L3	1
9	-----	-----	
10			
2			
4			
2			
10			
6			
L1(1)=9			

SUB	List 1	List 2	List 3	List 4
1	9			
2	10			
3	7			
4	4			
9				
GRAPH CALC TEST INTR DIST				

1-Var Stats	
n	=10
minX	=2
Q1	=4
Med	=7.5
Q3	=10
maxX	=12

1-Variable	
n	=10
minX	=2
Q1	=4
Med	=7.5
Q3	=10
maxX	=12

Exercise 6.1

- The following data shows the number of misprints in different chapters of a draft copy of a new textbook:

2, 3, 0, 1, 5, 7, 1, 3, 0, 4, 5, 5, 0, 8, 1, 2, 6, 6

Find the median number of misprints per chapter of the book.

- The list below shows the number of revision lessons attended by each student of a Mathematics class before a mock examination:

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

Find the median number of revision lessons attended by a student.

- The prices of 'super value tyres' stocked by a garage are shown below:

£33, £29, £35, £40, £45, £46, £47, £47, £53, £54, £64, £66, £50


Find the median price of the tyres stocked in the garage.

6.2 Finding the mean for discrete and continuous data

When you calculate the median, you are not interested in the actual values of *all* the data given; you are only interested in the central value(s).

A measure that *does* consider all the values in the data set is the **mean**.

The formula for finding the mean is as follows:

 The mean \bar{x} of a set of data x_1, x_2, \dots, x_n is $\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$, where $n = \sum_{i=1}^k f_i$

exam tip

The Σ (sigma) symbol means 'add up all the individual pieces'. This symbol is a useful mathematical shorthand for formulae that involve sums. Although you are not expected to use the symbol yourself, you need to know how to interpret it when it appears in formulae in your Formula booklet.

x_i is the value at the general position i .

f_i is the frequency of the value at position i .

$\sum_{i=1}^k f_i x_i$ tells you that you need to multiply each data value (x_i) by its frequency (f_i), and then add all the products together.

$\sum_{i=1}^k f_i$ represents the sum of all the frequencies, which is the same as the total number of data values.

Simple data

Simple data is considered to be a list of just one variable where each variable only has a frequency of 1. So, for example, 4, 5, 9, 10 is a simple list of data. To find the mean of a simple list of values, add up all the values and divide by the total frequency.

For example, take the numbers 9, 10, 7, 4, 2, 10, 6, 12, 3, 8.

There are 10 of them. So the mean is:

$$(9 + 10 + 7 + 4 + 2 + 10 + 6 + 12 + 3 + 8) \div 10 = \frac{71}{10} = 7.1$$

Even if the data is discrete, it is not unusual for the mean to have a value that is not a whole number, or not the same as any of the values in the original list. For example, the number of children in every family is discrete data, but according to United Nations figures, the mean number of children in a family in 2011 was 2.5. In this context, this value represents that most families in the United Nations have between 2 and 3 children, not that some people can have half a child!

You can use your GDC to calculate the mean of a simple list of data. See section '6.1 (a) Finding the mean, median, quartiles and standard deviation for a simple list of data (single variable, no frequency)' on page 666 of the GDC chapter if you need a reminder of how. In the list of statistical data on your GDC, the mean is labelled as \bar{x} .



 **TEXAS**

L1	L2	L3	1
9	-----	-----	
10			
7			
4			
2			
10			
6			
12			
3			
8			
L1(10)=9			

 **CASIO**

sub	List 1	List 2	List 3	List 4
1	9			
2	10			
3	7			
4	4			
GRAPH CALC TEST DATA DIST				



TEXAS

```

1-Var Stats
x̄=7.1
Σx=71
Σx²=603
Sx=3.314949304
sx=3.144837039
n=10

```



CASIO

```

1-Variable
x̄=7.1
Σx=71
Σx²=603
sx=3.14483703
sx=3.3149493
n=10

```

Discrete data organised in a frequency table

Suppose that the sizes of shoes sold in a ladies' shoe shop in San Gimignano are collected in a frequency table. To find the mean size of shoes sold, you could write out all the values (repeating each one according to its frequency) and then add them up as we did with the list of simple data in the previous section. But it is easier and quicker to use the formula:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

Worked example 6.1

Q. A new shoe design is coming to the shop in San Gimignano for the new season. The manager wants to make sure he can meet demand but cannot afford to over-order stock. He has been keeping a record of how many women buy shoes of each size and he wants to know which size is the most popular. Calculate the mean shoe size of the women who shop in his store.

A.

Shoe size, x_i	Frequency, f_i	Size \times frequency $x_i \times f_i$
34	7	$34 \times 7 = 238$
35	11	$35 \times 11 = 385$
36	15	$36 \times 15 = 540$
37	19	$37 \times 19 = 703$
38	23	$38 \times 23 = 874$
39	25	$39 \times 25 = 975$
40	18	$40 \times 18 = 720$

The calculation $\sum_{i=1}^k f_i x_i$ is

demonstrated in the third column of the table.

Each shoe size is multiplied by its frequency.

continued...

Shoe size, x_i	Frequency, f_i	Size \times frequency $x_i \times f_i$
41	13	$41 \times 13 = 533$
42	8	$42 \times 8 = 336$
43	3	$43 \times 3 = 129$
Total	142	5433

The product of $x_i \times f_i$ for each shoe size is added together to give the value of $\sum_{i=1}^k f_i x_i$

$n = \sum_{i=1}^k f_i$ is shown at the bottom of the second column

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n} = \frac{5433}{142} = 38.3$$

The mean shoe size sold is 38.3.

Substitute the values of $\sum_{i=1}^k f_i x_i$ and n into the formula for calculating the mean.

Using your GDC, you would just need to enter the shoe size as one list of data and the frequency as a second list of data. Your GDC can then calculate the mean (and other statistics) for you. See section '6.1 (b) Finding the mean, median, quartiles and standard deviation for grouped data (single variable with frequency)' on page 667 of the GDC chapter for a reminder of how to do this.



TEXAS



CASIO

L1	L2	L3	2
34	7		
35	11		
36	15		
37	19		
38	23		
39	25		
40	18		

L2(1)=7

Sub	List 1	List 2	List 3	List 4
1	34	7		
2	35	11		
3	36	15		
4	37	19		

GRAPH CALC TEST DATA DIST

1-Var Stats
 $\bar{x} = 38.26056338$
 $\Sigma x = 5433$
 $\Sigma x^2 = 208567$
 $s_x = 2.223917626$
 $\sigma_x = 2.216073095$
 $\downarrow n = 142$

1-Variable
 $\bar{x} = 38.2605633$
 $\Sigma x = 5433$
 $\Sigma x^2 = 208567$
 $\sigma_x = 2.21607309$
 $s_x = 2.22391762$
 $n = 142$

Grouped discrete data

To calculate the mean of grouped data, you use the same formula as before:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

But, the mean calculated from a table of grouped data will always be an **estimate** of the true mean.

If the data has been grouped, you need a single value that can represent all the data within a given group. A sensible value to choose as the group representative would be the one half-way between the class boundaries: the **mid-interval value**.

Look again at Ahmed's table data from Worked example 5.3. Suppose we want to find an estimate of the mean mark. The first thing we need to do is identify the mid-interval value for each group. This value will be substituted into the formula for the mean as the ' x_i ' value.

Marks	Frequency, f_i	Mid-interval value, x_i	$x_i \times f_i$
0–2	0	1	$0 \times 1 = 0$
3–5	0	4	$0 \times 4 = 0$
6–8	7	7	$7 \times 7 = 49$
9–11	11	10	$11 \times 10 = 110$
12–14	14	13	$14 \times 13 = 182$
15–17	10	16	$10 \times 16 = 160$
18–20	8	19	$8 \times 19 = 152$
Total	50		653

Using the formula and substituting in known values:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n} = 653 \div 50 = 13.1$$

Therefore an estimate of the mean mark is 13.1.

You can use your GDC to calculate the mean of grouped data in much the same way as you would for a simple list of data with frequency: simply enter the mid-interval in one list and the frequency in another list. See '6.1 (b) Finding the mean, median, quartiles and standard deviation for grouped data (single variable with frequency)' on page 667 of the GDC chapter if you need a reminder.



Recall from Chapter 5 that once data is grouped it loses some detail.



Return to section 5.5 if you need a reminder of how to calculate the mid-interval value of a group when using discrete data.



exam tip

In examinations, if a question is dealing with grouped data, it will always say 'Find an estimate of the mean'. Do not worry about this wording — the examiner is just acknowledging that the data has been grouped, so that not every one of the original values will be used.



TEXAS

CASIO

```
L1 | L2 | L3 | 2
7 | 7 | | 
10 | 11 | | 
13 | 14 | | 
16 | 10 | | 
19 | 8 | | 
---|---| | 
L2(6) =
```

```
List 1 | List 2 | List 3 | List 4
SUB
1 | 7 | 7 | 
2 | 10 | 11 | 
3 | 13 | 14 | 
4 | 16 | 10 | 
7
GRAPH | CALC | TEST | DATA | DIST |
```

```
1-Var Stats
x̄=13.06
Σx=653
Σx²=9257
Sx=3.856666637
σx=3.817905185
↓n=50
```

```
1-Variable
x̄ =13.06
Σx =653
Σx² =9257
σx =3.81790518
sx =3.856666637
n =50 ↓
```

RR Refer back to section 5.5 if you need a reminder of how to calculate the mid-interval value of a class when using grouped continuous data.

Grouped continuous data

As with grouped discrete data, the mean of grouped continuous data will also be an estimate. The calculation is very similar to the method for grouped discrete data, but you need to take care to use the correct class boundaries when calculating the mid-interval value.

For example, let's revisit the data on the mass of footballers from section 5.4. To find the mean mass, you first need to identify the class boundaries, then the mid-interval values, and then do the appropriate calculations for the formula:

$$\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n}$$

Mass (kg)	Class boundaries	Mid-interval value, x_i	Frequency, f_i	$x_i \times f_i$
61–65	60.5–65.5	$(60.5 + 65.5) \div 2 = 63$	8	$63 \times 8 = 504$
66–70	65.5–70.5	68	15	$68 \times 15 = 1020$
71–75	70.5–75.5	73	21	$73 \times 21 = 1533$
76–80	75.5–80.5	78	14	$78 \times 14 = 1092$
81–85	80.5–85.5	83	6	$83 \times 6 = 498$
86–90	85.5–90.5	88	2	$88 \times 2 = 176$
Total			66	4823

So $\bar{x} = \frac{\sum_{i=1}^k f_i x_i}{n} = 4823 \div 66 = 73.1$. The mean mass is 73.1 kg.

You can use your GDC in the same way as you did for discrete grouped data.

TEXAS

L1	L2	L3	1
63	8		
68	15		
73	21		
78	14		
83	6		
88	2		

L1(?)=

1-Var Stats

$\bar{x}=73.07575758$
 $\Sigma x=4823$
 $\Sigma x^2=355019$
 $Sx=6.293612405$
 $\sigma x=6.24575154$
 $n=66$

CASIO

List 1	List 2	List 3	List 4
63	8		
68	15		
73	21		
78	14		

GRAPH CALC TEST DATA DIST

1-Variable

$\bar{x}=73.07575758$
 $\Sigma x=4823$
 $\Sigma x^2=355019$
 $\sigma x=6.24575154$
 $Sx=6.2936124$
 $n=66$

RR We learned about histograms in Chapter 5. See section '5.2 Drawing a histogram' on page 664 of the GDC chapter for a reminder of how to use your GDC if you need to.

Enter the mid-interval values in one list and the frequencies in another.

Worked example 6.2

Q. Fifty people were asked to say when they think a time period of one minute had elapsed. The times they estimated were recorded to the nearest second. The results are listed below.

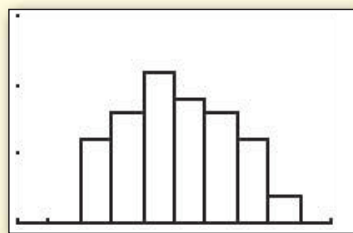
Time (s)	50–52	53–55	56–58	59–61	62–64	65–67	68–70
Frequency	6	8	11	9	8	6	2

- (a) Use the table to draw a histogram on your GDC.
- (b) Estimate the mean value of the results.
- (c) Use the histogram and mean value to comment on your results.

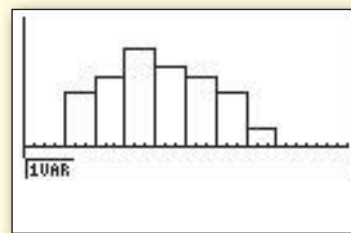
A. (a)

Time (s)	50–52	53–55	56–58	59–61	62–64	65–67	68–70
Frequency	6	8	11	9	8	6	2
Mid-interval x_i	51	54	57	60	63	66	69

TEXAS



CASIO



continued ...

(b)



TEXAS

```

1-Var Stats
x̄=58.86
Σx=2943
Σx²=174483
Sx=5.066939663
σx=5.016014354
↓n=50

```



CASIO

```

1-Variable
=58.86
Σx=2943
Σx²=174483
σx=5.01601435
sx=5.06693966
n=50
↓

```

Use your GDC to get an estimate of \bar{x} . Write down the answer appropriately.

exam tip

Data is sometimes described as **skewed** to the left or to the right. This is not a term that will appear in the examinations, but you may meet it when reading other books.

The mean is estimated to be 58.9 seconds.

(c) The histogram is skewed towards the left, and the mean is less than 60 seconds. This indicates that most of the people taking part in the experiment underestimated the length of a minute.

Exercise 6.2

1. The price of a 4th generation iPod Touch 8GB in ten different shops is shown below.

£164, £166.21, £167.00, £169.99, £167,

£170.00, £172.00, £174.00, £174.99, £179.99

Calculate the mean price for this model of iPod.

2. In 2010, the countries with GDP per capita in the top ten of the world had the following GDP figures:

\$82,600	\$69,900	\$62,100	\$57,000	\$54,600
\$51,600	\$49,600	\$48,900	\$141,100	\$179,000

Calculate the mean per capita GDP of these countries.

3. The sizes in KB of emails in a person's inbox are listed below.

10, 10, 15, 2, 27, 3, 323, 38, 4, 4, 4, 4, 439, 6, 6, 6, 7, 8, 926

Calculate the mean size of these emails.

4. (a) The following table shows the distribution of ages of 22 horses from the same race-course.

Age (years)	2	3	4	5	6	7
Frequency	3	4	8	3	2	2

Calculate the mean age of the horses.

(b) The following table shows the carrying mass of the same horses.

Carrying mass (lbs)	110	112	113	114	115	116
Frequency	1	2	2	1	1	2
Carrying mass (lbs)	118	120	121	122	123	126
Frequency	3	5	1	1	2	1

Calculate the mean carrying mass of these horses.

5. The table below shows the number of goals scored by top goal-scorers in the English Premier League for the 2010 football season.

Number of goals	10	11	12	13	15	17	18	20
Frequency	8	1	4	6	1	1	1	2

(a) Find the total number of goals scored by these players.

(b) Calculate the mean number of goals scored.

The list of top 40 goal-scorers in the same season contains the following additional information:

Number of goals	7	8	9
Frequency	4	5	7

(c) Calculate the overall mean number of goals scored by the players from the list of top 40 goal-scorers.

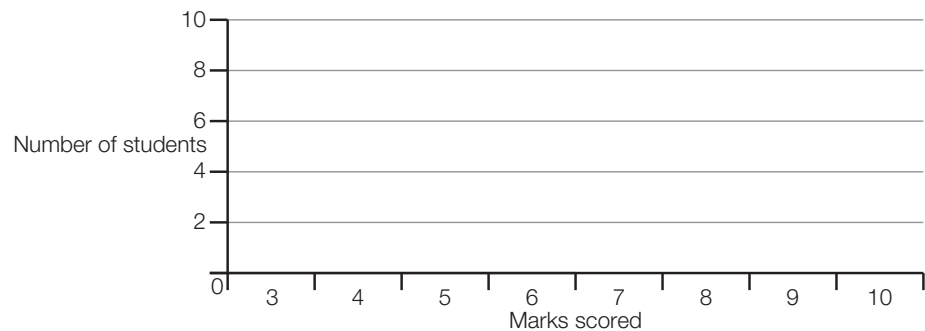
6. The table below shows the price list of cars stocked by a car dealer.

	4-Door Sedan	2-Door Coupe	Hatchback
1	\$9,990	\$11,990	\$9,985
2	\$11,965	\$12,490	\$10,990
3	\$11,965	\$14,990	\$12,115
4	\$11,965	\$15,605	\$12,115
5	\$11,965	\$15,605	\$12,605
6	\$12,295	\$16,995	\$13,155
7	\$12,295	\$17,200	\$13,300
8	\$12,445	\$18,275	\$13,895
9	\$13,200	\$18,575	
10	\$13,200	\$18,999	
11	\$13,359		
12	\$13,365		

- (a) Using the prices from the table, calculate the mean and median prices for the three types of car.

	Type of car		
Average	4-Door Sedan	2-Door Coupe	Hatchback
Mean price(\$)			
Median price (\$)			

- (b) Which of the two measures do you think is better for comparing the prices of the types of cars? State your reason.
- (c) Explain why your answers from part (a) may not be a fair way of comparing the average prices of the cars. Suggest a fairer method.
7. The bar chart below shows the marks scored by students in a French spelling test.



- (a) How many students took the test?
- (b) What was the mean mark scored by the students?
- (c) Which mark represents the median score?
8. The consumption of softwood throughout the UK for the period 2000–2009 is shown in the table below. The figures are given in thousands of green tonnes.

(Source: <http://www.forestry.gov.uk/forestry/HCOU-4UBEJZ>)

Mass, m (thousands of green tonnes)	Frequency
$500 < w \leq 750$	16
$750 < w \leq 1000$	4
$1000 < w \leq 1250$	0
$1250 < w \leq 1500$	5
$1500 < w \leq 1750$	5
$1750 < w \leq 2000$	1
$2000 < w \leq 2250$	3
$2250 < w \leq 2500$	4
$2500 < w \leq 2750$	2

- (a) What is the estimated total consumption of softwood by the UK in the given period?
- (b) Calculate an estimate of the mean mass of softwood consumed in the given period.
9. The table below shows the average price per litre of fuel in 194 countries in 2010. Fuel prices refer to the pump prices of the most widely sold grade of gasoline. Prices have been converted from the local currency to US dollars.

(Source: <http://data.worldbank.org>)

Fuel price per litre, p (USD)	Frequency
$0.00 < p \leq 0.40$	12
$0.40 < p \leq 0.80$	16
$0.80 < p \leq 1.20$	70
$1.20 < p \leq 1.60$	52
$1.60 < p \leq 2.00$	39
$2.00 < p \leq 2.40$	3
$2.40 < p \leq 2.80$	2

Work out an estimate of the mean price per litre of fuel in 2010 across these countries.

10. The table from Exercise 5.8, question 1, showing the times of Olympic gold medallists in the men's 100-metres final from 1896 to 2008 has been reproduced below.

Time (seconds)	Frequency
$9.60 \leq t < 10.00$	20
$10.00 \leq t < 10.40$	30
$10.40 \leq t < 10.80$	12
$10.80 \leq t < 11.20$	14
$11.20 \leq t < 11.60$	2
$11.60 \leq t < 12.00$	0
$12.00 \leq t < 12.40$	1
$12.40 \leq t < 12.80$	2

Calculate an estimate of the mean winning time.

6.3 Identifying the mode or modal class

The mode is the value that occurs most frequently in a set of data. If the data is grouped, then the class with the largest frequency is called the **modal** class.

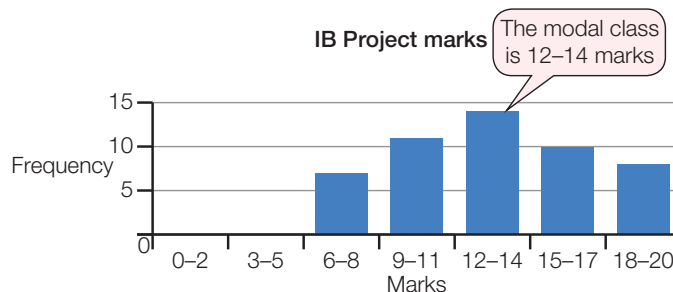
If you have made a frequency table for the data, the mode or modal class will correspond to the highest frequency in that table. For example, look at the data on the number of siblings from Worked example 5.2:

Number of siblings	Frequency
0	12
1	21
2	14
3	9
4	4
Total number of children questioned	60

The largest frequency is 21, so the mode is 1 sibling.

The mode is 1.

If you have drawn a bar chart or histogram to display the data, the mode or modal class is the data value or data class with the highest bar.



6.4 Comparing the median, mean and mode

At the beginning of this chapter, Rita calculated the mean, median and mode for the time spent on her Maths assignments, and obtained three different answers.

This outcome occurs quite often, especially for sets of values that are very spread out, or where the data is sparse. It is important to be aware of how and why these 'averages' can be so different, and be able to choose the most appropriate value to summarise a particular data set.

Suppose that Rita asks her classmates to count the number of music downloads they make in one month. She collects the following data:

19, 23, 15, 16, 17, 13, 21, 12, 18, 20, 12, 14, 10, 22, 12

She finds that:

Mean	16.3
Median	16
Mode	12



Governments, aid agencies, corporations, in fact most organisations, gather data about their clients or customers to use in assessing what has been achieved and in planning for the future. It is important that statistics should always be used with an understanding of where and how the data has been collected, and how the statistical measures have been calculated. It is too easy to think that figures look reasonable, when in fact they have been calculated to mislead.

She and her classmates discuss the results. Ari says that they should not use the mode; it is low in comparison with the other figures and does not give a good representation of the other data. Niamh is worried about the median; it seems high in relation to the mode, and she thinks it does not give any indication that there are a lot of low figures in the data set. The class as a whole decides that the best measure for this data set is the mean, because it has used all the figures and taken into account both high and low numbers. This is the result of their discussion for this particular data set; a different group of people could come to a different conclusion and decide to choose the median as the best measure. The important point is that you think about and are able to explain any differences in the values of the three measures.

Rita's data on music downloads was discrete. We now look at an example where the data is continuous. Similar problems can arise if the three measures of central tendency give different values.

Suppose that a forestry company wanted to know the average growth of a group of trees. They measured the heights of those trees in successive years. The frequency table below gives the heights of the trees in 2012.

Height (m)	Frequency
5–9	12
10–14	18
15–19	18
20–24	8
25–29	3
Total	59

To help calculate the mean, we add extra columns for the mid-interval values x_i and the products $f_i x_i$.

Height (m)	Frequency (f_i)	Mid-interval value (x_i)	$f_i \times x_i$
5–9	12	7	84
10–14	18	12	216
15–19	18	17	306
20–24	8	22	176
25–29	3	27	81
Total	59		863

The mean is $863 \div 59 = 14.6$ m.

The mode can be found by looking at the first table: there are actually two modal classes, 10–14 m and 15–19 m, which have the same frequency. Data sets with two modes are described as **bimodal**.

In this case, neither the mean nor the mode is a good measure of central tendency:

- The mean is influenced too much by the heights of the three tall trees.
- Since there are two modes, this measure is not very useful.

hint

When someone **gives** you an average, it is important to know whether it is the median, mean or mode. When you **calculate** an average, make it clear which measure you have chosen and why you think it gives the most representative value.



When presenting statistics, a responsible organisation will make it clear how the data was collected, which statistical measure is being used to summarise it, and why. They will also use the correct terms. For example, a development agency might collect many small donations from individuals as well as some large ones from corporations. Using the median to represent the 'average' donation will allow them to balance these small contributions against the large ones; if they use the mean, it may make the large donations seem too important.

Sometimes it is possible to calculate unknown values within your data set if you know the mean, median or mode.

There are six numbers, so the median is the mean of the 3rd and 4th ones, which are 16 and p . Solve the equation for p .

Now we only have one unknown left, q . We can use the information about the mean to find it.

Worked example 6.3

Q. The following set of ordered numbers has mean and median both equal to 16.5.

$$12, 15, 16, p, 18, q$$

Find the values of p and q .

A.
$$\frac{16+p}{2} = 16.5$$

$$16+p=33$$

$$p=17$$

$$\text{Mean} = (12 + 15 + 16 + 17 + 18 + q) \div 6 = 16.5$$

$$(12 + 15 + 16 + 17 + 18 + q) = 16.5 \times 6$$

$$78 + q = 99$$

$$q = 21$$

Worked example 6.4

Q. A clinic recorded the blood pressure of its patients as follows:

Blood pressure (mmHg)	70–79	80–89	90–99	100–109	110–119	120–129
Frequency	8	11	15	19	22	18

Blood pressure (mmHg)	130–139	140–149	150–159	160–169	170–179
Frequency	17	15	12	9	6

- Write down the total number of patients.
- Calculate an estimate of the mean blood pressure.
- Write down the modal class.
- An 'ideal' blood pressure is considered to lie between 90 and 120 mmHg. What percentage of patients had an 'ideal' blood pressure?

Add all the frequencies to get the total number.

A. (a) $8 + 11 + 15 + 19 + 22 + 18 + 17 + 15 + 12 + 9 + 6 = 152$

To estimate the mean, we first need to identify the class boundaries and the mid-interval values.

- (b) The boundaries of the first class are 69.5 and 79.5, so the mid-interval value is $(69.5 + 79.5) \div 2 = 74.5$. The second class has boundaries 79.5 and 89.5, with mid-interval value 84.5, and so on.

continued ...

Use your GDC to calculate the mean. See section '6.1 (b) Finding the mean, median, quartiles and standard deviation for grouped data ...' on page 667 of the GDC chapter if you need a reminder of how.

Write down the answer appropriately.

The modal class is the class with the highest frequency.

The number of people with blood pressure between 90 and 120 mmHg is the number of people in the classes 90–99, 100–109 and 110–119.



TEXAS

L1	L2	L3	1
74.5	8		
84.5	11		
94.5	19		
104.5	19		
114.5	22		
124.5	18		
134.5	17		
L1(G)=124.5			

1-Var Stats	
\bar{x}	=121.8684211
Σx	=18524
Σx^2	=2365238
s_x	=26.71252012
σ_x	=26.62450499
n	=152



CASIO

SUB	List 1	List 2	List 3	List 4
1	74.5	8		
2	84.5	11		
3	94.5	19		
4	104.5	19		
1VAR 2VAR REG SET				

1-Variable	
\bar{x}	=121.868421
Σx	=18524
Σx^2	=2.3652E+06
σ_x	=26.6245049
s_x	=26.7125201
n	=152

The mean is 122 mmHg (3 s.f.)

(c) The modal class is 110–119 mmHg.

(d) $15 + 19 + 22 = 56$

$$\text{Percentage} = \frac{56}{152} \times 100 = 36.8\%$$

Exercise 6.3

1. For each of the following sets of data, find:

- (i) the mean (ii) the median (iii) the mode

and then compare the three averages.

- (a) 5, 7, 3, 2, 1, 2, 8
 (b) 70, 57, 57, 61, 64, 70, 69
 (c) 29, 28, 27, 38, 29, 35, 29
 (d) 110, 140, 160, 110, 105, 109, 120, 107, 111, 107
 (e) 33, 44, 37, 48, 50, 42, 47, 42, 44, 40, 36

2. The mean of the following twelve numbers is 30.

30, 27, 28, 33, 27, x , 27, 32, 33, 31, 29, 31

- (a) Determine the value of x .

- (b) Find the median.
- (c) What is the modal value?
- (d) Which of the two measures, the median or the mode, summarises the data better?
3. The mean of the following twelve numbers is 18.
11, 20, 19, x , 18, 10, 19, 20, 13, 21, 11, 19
- (a) Determine the value of x .
- (b) Find the median.
- (c) Find the mode.
- (d) Which of the two measures, the median or mode, summarises the data better?
4. For each of the following sets of data you are given the mean value. Determine the value of:
- (i) x (ii) the median (iii) the mode.
- (a) $2x$, 13, 18, 19, $2x$, x , 17, 12; mean = 13
- (b) 12, 16, $2x$, 18, 23, 18, 26, 24, x , $2x$; mean = 19
- (c) 7, 11, 7, x , 9, 14, 16, 11, $2x$, $4x$; mean = 11

5. For a certain class, the number of student absences from school during the summer term is shown on the table below.

Absences (days)	0	1	2	3	4	5	6	7
Frequency	8	6	3	3	2	1	1	1

- (a) State the total number of students in the class.
- (b) Calculate the total number of absences.
- (c) Find the mean number of absences.
- (d) State the modal number of absences.
- (e) Determine the median number of absences.
- (f) Discuss which of the averages is more likely to be used when:
- (i) comparing attendance among different groups
- (ii) dealing with targeted individual attendance.
6. The scores of the top 26 golfers for rounds 1 and 2 in a major tournament are given below.

Round 1:

Score	66	68	69	70	71	72	73	74	75
Number of golfers	1	3	2	4	5	2	5	2	2

Round 2:

Score	68	69	70	71	72	73	74	75	78
Number of golfers	1	8	1	5	4	4	1	1	1

(a) Copy and complete the following table of summary statistics for the first two rounds of the competition.

	Round 1	Round 2	Rounds 1 and 2 combined
Mean			
Mode			
Median			

(b) Compare the performance of the golfers in the two rounds.
 (c) Which of the three measures better describes the overall performance of the golfers?

7. For each of the following sets of grouped data:

- (i) Work out an estimate of the mean.
 (ii) State the modal group.

(a) The time taken by a group of 15-year-old students to complete a homework task:

Time (minutes)	Number of students
27–33	3
34–40	6
41–47	13
48–54	10
55–61	8
62–68	7
69–75	3

(b) The circulation of a group of daily newspapers:

Circulation, c , (thousands)	Frequency
$0 \leq c < 150$	50
$150 \leq c < 300$	35
$300 \leq c < 450$	7
$450 \leq c < 600$	4
$600 \leq c < 750$	1
$750 \leq c < 900$	0
$900 \leq c < 1050$	1

Summary

You should know:

- that the three measures of central tendency are the mean, median and mode
- how to calculate the mean, median and mode for simple discrete data
- how to calculate an estimate of the mean, and to identify the modal class, for grouped discrete or continuous data
- that the mean, median and mode for the same set of data might be different but are all 'correct'; you need to choose which measure to use and explain why.

Mixed examination practice

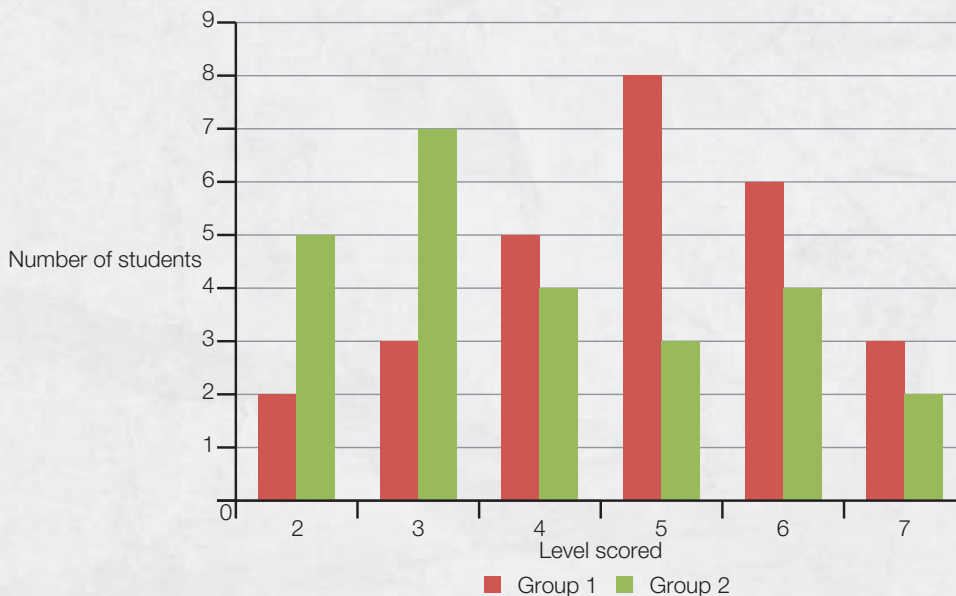
Exam-style questions

1. The scorecard of the first innings of a county cricket club is shown below.

Batsman	Number of runs
1	36
2	33
3	22
4	51
5	
6	30
7	30
8	27
9	17
10	7
11	1

Mean number of runs 28

- (a) What was the total number of runs scored by the team?
- (b) Work out the number of runs scored by the fifth batsman.
- (c) Determine the median number of runs.
2. The bar chart shows the performance of two groups in their Mathematical Studies mock examination.



- (a) How many students took the examination in each group?
- (b) What was the mean grade for each of the groups?

- (c) Compare the performance of the two groups of students.
- (d) Find the combined mean for the two groups.
3. The following table shows the distances achieved by medal winners in the women's long jump in the Olympic Games from 1948 to 2008. (It is a copy of the data from Exercise 5.8 question 2.)

Distance, d (metres)	Frequency
$5.50 \leq d < 5.75$	3
$5.75 \leq d < 6.00$	1
$6.00 \leq d < 6.25$	5
$6.25 \leq d < 6.50$	4
$6.50 \leq d < 6.75$	7
$6.75 \leq d < 7.00$	11
$7.00 \leq d < 7.25$	15
$7.25 \leq d < 7.50$	1

Work out an estimate of the mean distance jumped by these athletes.

4. The table below summarises the populations of 48 African countries in 2005.

Population, p (in millions)	Number of countries
$0 < p \leq 10$	27
$10 < p \leq 20$	11
$20 < p \leq 30$	3
$30 < p \leq 40$	3
$40 < p \leq 50$	1
$50 < p \leq 60$	1
$60 < p \leq 70$	1
$70 < p \leq 80$	1

- (a) What is the estimated total population of the 48 countries?
- (b) Calculate an estimate of the mean population of these countries.
- One of the countries omitted from the list has a population of 138 million people.
- (c) Use your result from part (a) to calculate the overall mean for the 49 countries.
5. Each of the following sets of data has been listed in numerical order. The mean and the median have been given. Determine the values of x and y in each case.
- (a) 37, 38, 39, x , 39, 43, 45, y (mean is 41, median is 39)
- (b) 11, x , 14, 14, y , 19, 19, 20 (mean is 16, median is 16)
- (c) 42, 42, 43, x , y , 48, 50, 53, 55, 56 (mean is 48, median is 47)

6. The number of merit awards received by a group of students in one month is shown below.

Number of merits	0	1	2	3	4	5	6
Frequency	2	15	17	10	9	5	3

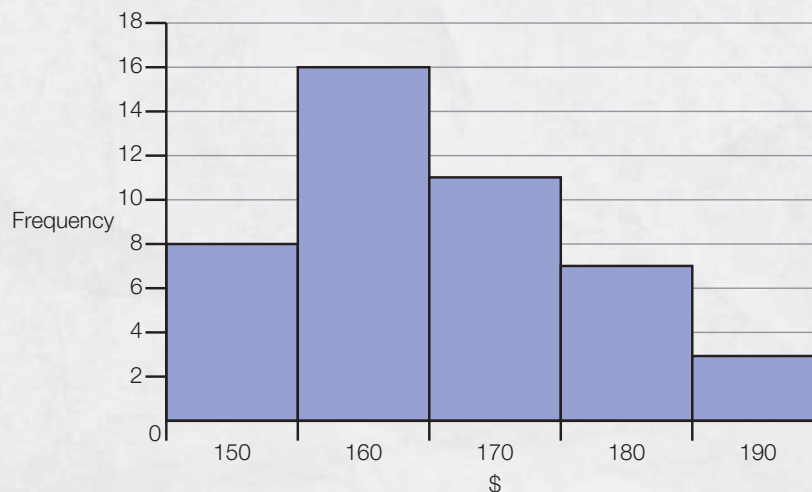
- Find the total number of students in the group.
 - Calculate the total number of merits.
 - Find the mean number of merits.
 - State the modal number of merits.
 - Determine the median number of merits.
 - Compare the three averages, indicating which of them is a better summary measure of the number of merits.
7. The table below shows the number of sixes scored by 30 cricketers in their professional careers.

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

- Find the mean number of sixes.
- State the modal number of sixes.
- Determine the median number of sixes.

Past paper questions

1. The histogram below shows the amount of money spent on food each week by 45 families. The amounts have been rounded to the nearest 10 dollars.



- Calculate the mean amount spent on food by the 45 families.
- Find the **largest possible amount** spent on food by a single family in the **modal** group.

(c) State which of the following amounts could **not** be the total spent by all families in the modal group:

- (i) \$2430 (ii) \$2495 (iii) \$2500 (iv) \$2520 (v) \$2600

[6 marks]

[May 2006, Paper 1, Question 5] (© IB Organization 2006)

2. The temperatures in °C, at midday in Geneva, were measured for eight days and the results are recorded below.

7, 4, 5, 4, 8, T , 14, 4

The mean temperature was found to be 7°C.

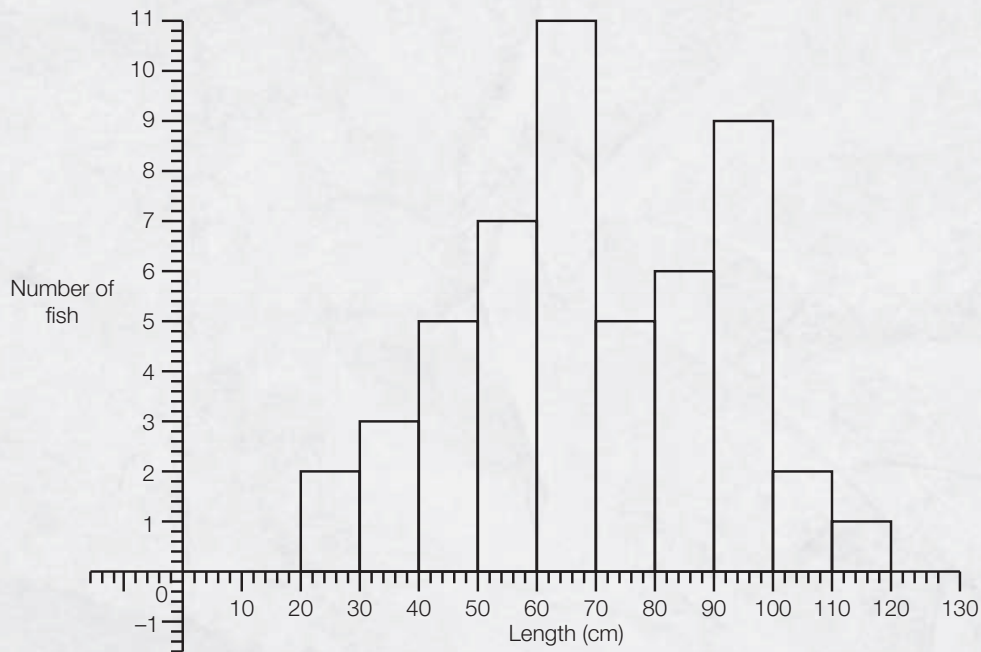
(a) Find the value of T . [3 marks]

(b) Write down the mode. [1 mark]

(c) Find the median. [2 marks]

[Nov 2009, Paper 1, Question 1] (© IB Organization 2009)

3. The figure below shows the lengths in centimetres of fish found in the net of a small trawler.



(a) Find the total number of fish in the net. [2 marks]

- (b) Find
- (i) the modal length interval
 - (ii) the interval containing the median length
 - (iii) an estimate of the mean length.

[5 marks]

[May 2007, Paper 2, Question 1(a),(b)] (© IB Organization 2007)

Chapter 7 Measures of dispersion

In this chapter you will learn:

- the different ways of measuring the dispersion of a set of data
- how to calculate the range, interquartile range and standard deviation.

In Chapter 6 we looked at different ways in which an ‘average’ can be calculated, and discussed how to decide whether the mean, median or mode gives the best description of a particular set of data. Although the ‘average’ is a useful single number to summarise the data, it is not the only measure that is important when we are analysing and interpreting data. For example, if you are told that the mean temperature in Mexico City in January is 13°C, this could suggest that the temperature stays around 13°C all January; but it could also suggest that the temperature ranges between 5°C and 21°C, with a ‘centre’ at around 13°C.

This chapter looks at various methods of calculating the **dispersion** of a set of data; this is an estimate of how ‘spread out’ (i.e. dispersed) the data is. This will give you an indication of how well the mean, median or mode represents the data overall.

Jaime and her father enjoy playing golf together. They also enjoy arguing about who is the better golfer.

Here are their scores over the past eight games:

Jaime	90	92	92	90	87	91	86	85
Father	90	83	95	93	86	88	96	82

You can see that Jaime had the better score three times, her father had the better score four times, and there was one tied game. Who is the better golfer?

Let’s look at the statistics:

	Mean	Median	Mode	Range (highest score – lowest score)
Father	89	89	None	96 – 82 = 14
Jaime	89	90	90 and 92	92 – 85 = 7

The average scores reveal very little: the mean scores are the same, and the median scores are nearly the same. The last column of the table, ‘Range’, gives more useful information: Jaime has a small range of scores, which means that her game is quite consistent; her father’s range is much wider — he has good days and bad days, so his performance is more variable. Does this extra information help you to decide who is the better golfer?

A measure of central tendency (median, mean or mode) gives you one piece of information about the data set, but it is also important to know how spread out the data is. If the spread of data is small, the mean or median gives a more accurate measure than if the data is widely dispersed.



7.1 Range and interquartile range

The simplest way of measuring the spread of a set of data is to subtract the lowest value from the highest value. This is called the **range**. The range can be calculated for any set of data, though it may not always provide useful information: when one or two data values are unusually high or low (outliers), the range will give an unrealistically large value for the spread.

Another number measures the spread about (above and below) the **median** and is obtained by subtracting the lower quartile from the upper quartile. It is called the **interquartile range** (IQR); it measures the spread of the central 50% of data values. The interquartile range is often used in preference to the range because it is not affected by outliers, but it still doesn't take into account all of the data. The formula for interquartile range is given in your Formula booklet as:

$a = \pi r^2$

$$\text{IQR} = Q_3 - Q_1$$

You know from Chapter 5 that:

- half (50%) of the data lies between Q_1 and Q_3 ; we now know this is called the interquartile range
- a quarter (25%) of the data lies between the smallest value and Q_1
- three quarters (75%) of the data lies between the smallest value and Q_3 (note that this also means that a quarter of the data lies between Q_3 and the largest value).

For both the range and the interquartile range:

- A smaller measure of spread tells you that the data values lie close to the 'average value'.
- A larger measure of spread tells you that the data is more widely dispersed from the 'average value'.

We now look at how to find the interquartile range for different types of data sets.

Simple data

For **simple data** with an **even number** of values:

- Put the numbers in order.
- Divide the set into two halves.
- Find the median for each **half**; these will be the quartiles Q_1 and Q_3 .
- Subtract the lower quartile from the upper quartile to get the interquartile range: $\text{IQR} = Q_3 - Q_1$.



In Chapter 5 you learned how to use a cumulative frequency curve to find the values of the upper and lower quartiles.



You learned how to find the median in Chapter 6.

For example, take the numbers 12, 8, 9, 14, 15, 10, 16, 18, 19, 13, 9, 11.

Find the lower quartile by taking the median of the lower half; as there is an even number of values, this is obtained by finding the mean of the third and fourth values.

Put them in order and divide into two halves, then find the median of each half: 8, 9, 9, 10, 11, 12 | 13, 14, 15, 16, 18, 19

Lower half: 8, 9, 9, **10**, 11, 12

$$Q_1 = (9 + 10) \div 2 = 9.5$$

Upper half: 13, 14, **15**, **16**, 18, 19

$$Q_3 = (15 + 16) \div 2 = 15.5$$

The interquartile range is $Q_3 - Q_1 = 15.5 - 9.5 = 6$.

Find the upper quartile by taking the median of the upper half in the same way as you did with the lower half.

For **simple data** with an **odd number** of values:

- Put the numbers in order.
- Find the median of the **whole** data set; the median separates the data set into two halves.
- Look at each half, **not including the median of the whole set**, and find the median for the **lower** half and then for the **upper** half; the medians for the two halves will be the quartiles Q_1 and Q_3 .
- Subtract the lower quartile from the upper quartile to get the interquartile range: $IQR = Q_3 - Q_1$.

For example, take the numbers 9, 10, 7, 4, 2, 10, 6, 12, 3, 8, 5.

Put them in order and find the median, then divide the list into two halves and find the median of each half:

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

Median is the central value of the ordered data, which is 7.

Lower half (7 not included): 2, 3, **4**, 5, 6.

$$Q_1 = 4$$

Find the lower quartile by taking the median of the lower half; as there is an odd number of values, the median is the central value.

Upper half (7 not included): 8, 9, **10**, 10, 12.

$$Q_3 = 10$$

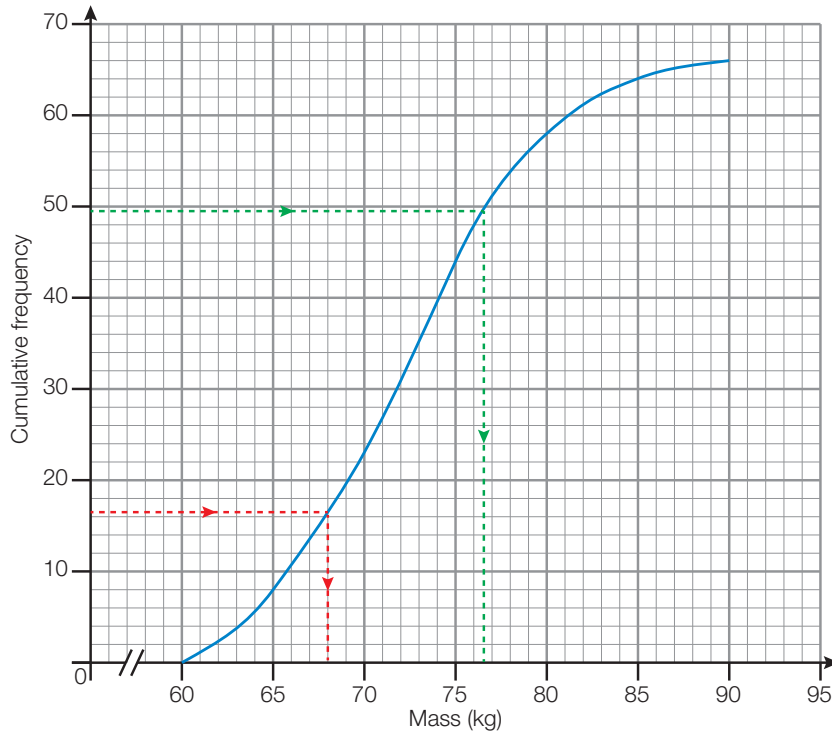
Find the upper quartile by taking the median of the upper half; the median is the central value.

The interquartile range is $Q_3 - Q_1 = 10 - 4 = 6$.

Grouped data

For grouped data organised in a grouped frequency table, it is best to use a **cumulative frequency graph** to find the quartiles and the interquartile range. You learned how to do this in Chapter 5.

Here is a quick reminder: let us revisit the data on the mass of footballers from section 5.4. We use the frequency table to draw a cumulative frequency curve.



There are 66 footballers in total, so:

- Lower quartile represents a quarter of the total frequency, so this is a frequency of $\frac{1}{4} \times 66 = 16.5$ (or you could do $25\% \times 66$).
- Upper quartile represents three quarters of the total frequency, so this is a frequency of $\frac{3}{4} \times 66 = 49.5$ (or you could do $75\% \times 66$).

Draw horizontal lines from 16.5 and 49.5 on the vertical axis until they meet the curve; then from each intersection draw a vertical line down to the horizontal axis and read off the value. This gives:

- $Q_3 = 76.5$
- $Q_1 = 68$

You can then use the values of Q_1 and Q_3 to calculate the interquartile range:

$$\text{IQR} = Q_3 - Q_1 = 76.5 - 68 = 8.5$$

hint

Remember to plot the **upper boundary** of each class against the cumulative frequency.

Mass (kg)	Frequency	Cumulative frequency
61–65	8	8
66–70	15	23
71–75	21	44
76–80	14	58
81–85	6	64
86–90	2	66

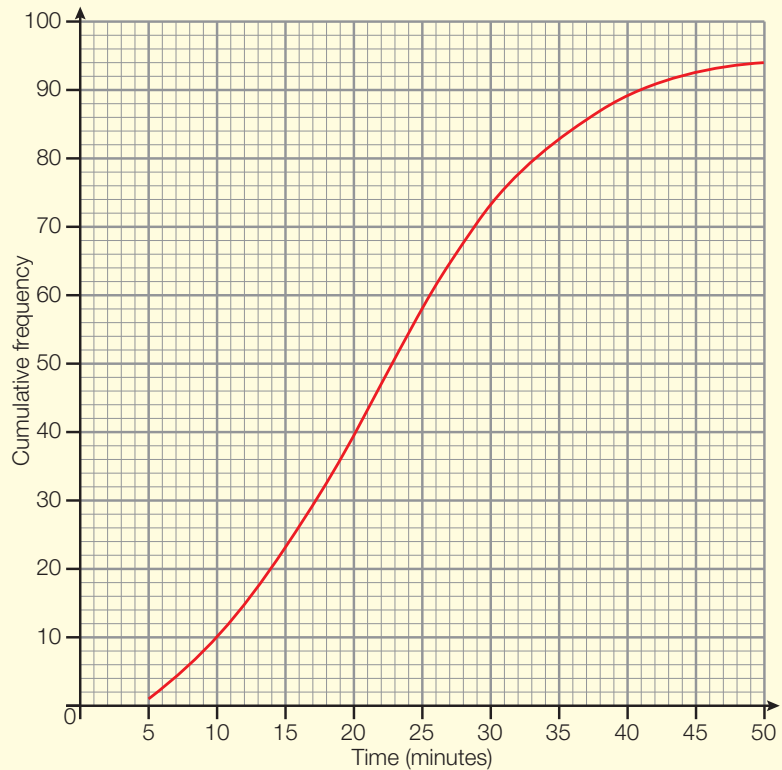


If the points used to plot a cumulative frequency graph are joined with straight lines rather than with curves, will your answers change? Would it matter? Some statisticians prefer using straight lines to join plotted points, to show that they are making no assumptions about the spread of data between successive points.

Note that with grouped data, your GDC does **not** provide a reliable way to calculate the upper and lower quartiles. It is much better to use a cumulative frequency graph.

Worked example 7.1

- Q. The cumulative frequency graph shown below represents the times taken to travel to school by a group of 94 students.



From the graph:

- Write down the median time.
- Calculate the interquartile range for the times taken to travel to school.
- Estimate the number of students who take longer than 38 minutes to travel to school.
- Given that the minimum and maximum times are 5 minutes and 50 minutes respectively, draw a box and whisker diagram for the data.

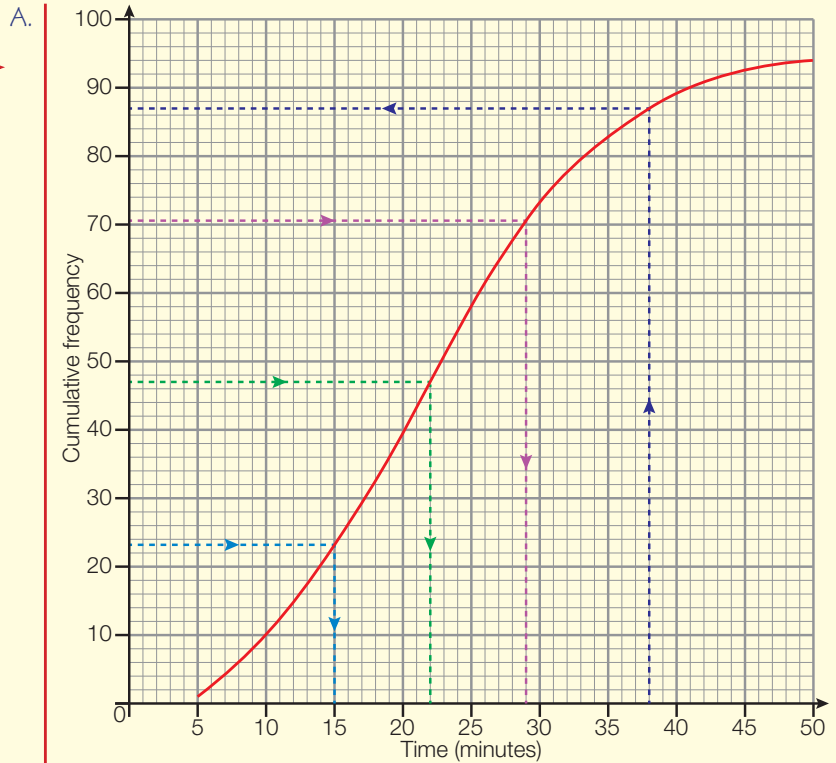


**You learned
about box
and whisker
diagrams in
Chapter 5.**



continued...

We draw lines on the cumulative frequency graph to find the answers.



Draw a line horizontally across from 47 on the vertical axis to the curve, and then vertically down until it meets the horizontal axis.

(a) $50\% \times 94 = 47$
Median = 22 minutes.

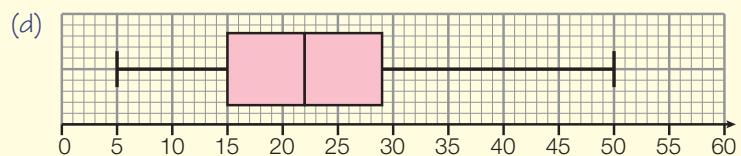
Draw lines horizontally across from 23.5 and 70.5 on the vertical axis to the curve, and then vertically down until they meet the horizontal axis.

(b) $25\% \times 94 = 23.5$, $75\% \times 94 = 70.5$
From the graph, $Q_1 = 15$, $Q_3 = 29$
so IQR = $29 - 15 = 14$ minutes.

Draw a line vertically up from 38 on the horizontal axis to the curve, and then horizontally to the left until it hits the vertical axis.

(c) 87 students take less than 38 minutes to travel to school,
so $94 - 87 = 7$ take longer than 38 minutes.

Use the median, upper and lower quartiles, and maximum and minimum values to construct the box and whisker diagram.



Exercise 7.1

1. For each of the following sets of data, determine:

(i) the median (ii) the range (iii) the interquartile range.

(a) 6.77, 6.67, 6.72, 6.66, 6.60, 7.06, 7.04, 7.01, 6.96, 6.81, 6.80

(b) 226, 222, 224, 222, 220, 235, 235, 234, 232, 227, 227, 247, 241

(c) 65, 86, 64, 50, 77, 96, 72, 66, 72, 65, 77, 69, 75, 73

(d) 14.3, 22.3, 14.1, 19.2, 11.7, 30.9, 21.7, 13.6, 17.2, 20.1, 18.6, 25.1

(e) 580, 550, 300, 350, 300, 344, 500, 263, 330, 230, 330, 196, 200, 608

(f) 97.9, 96.9, 99.4, 98.1, 97.7, 97.1, 95.9, 96.7, 98.5, 96.6, 97.1

2. For each of the following sets of data, calculate:

(i) the median (ii) the range (iii) the interquartile range.

(a) The number of goals scored over a football season per match:

Number of goals	0	1	2	3	4	5
Number of matches	10	11	9	5	3	2

(b) The height of students in a class of 14- and 15-year-olds:

Height (cm)	154	157	165	171	175	176	181
Frequency	5	9	12	6	7	8	3

(c) The number of marks scored by a class in a maths test:

Marks	60	61	62	63	64	65	66
Frequency	3	5	6	13	7	2	1

(d) The number of emails a class of students receive in a day:

Number of emails	0	1	2	3	4	5	6
Number of students	1	9	11	13	8	4	2

(e) The mass of passengers' luggage on a flight from Paris to New York:

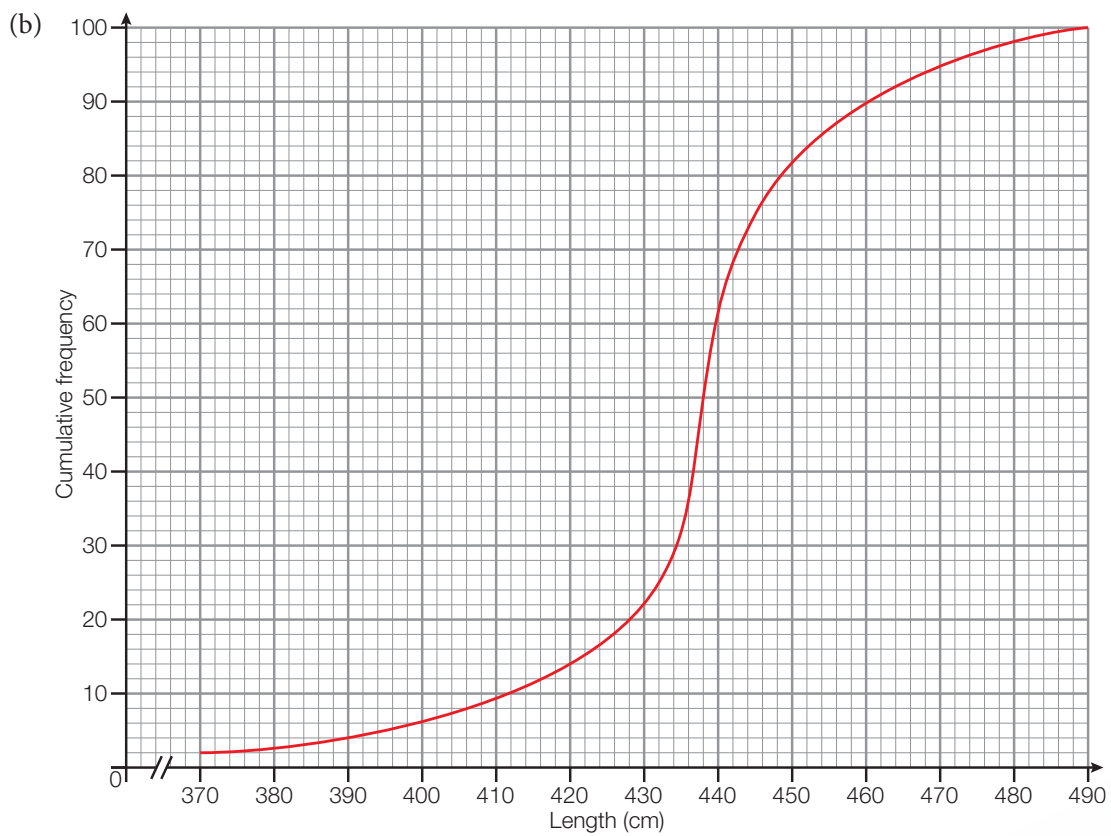
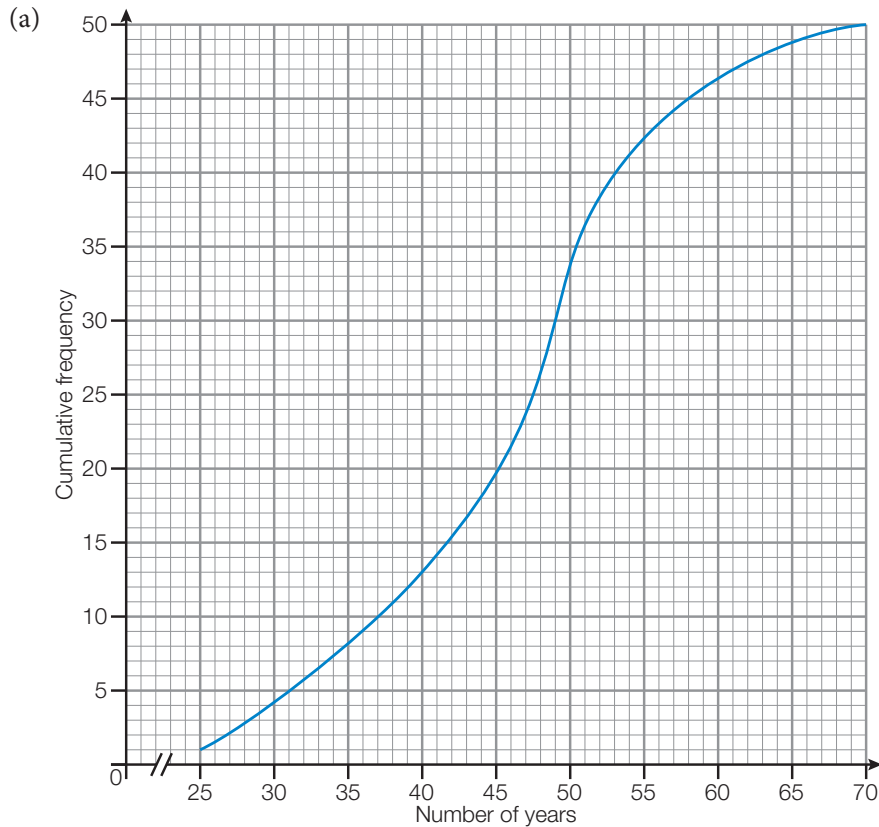
Mass of luggage (kg)	25	26	27	28	29	30	31	32
Frequency	3	7	9	11	10	28	2	1

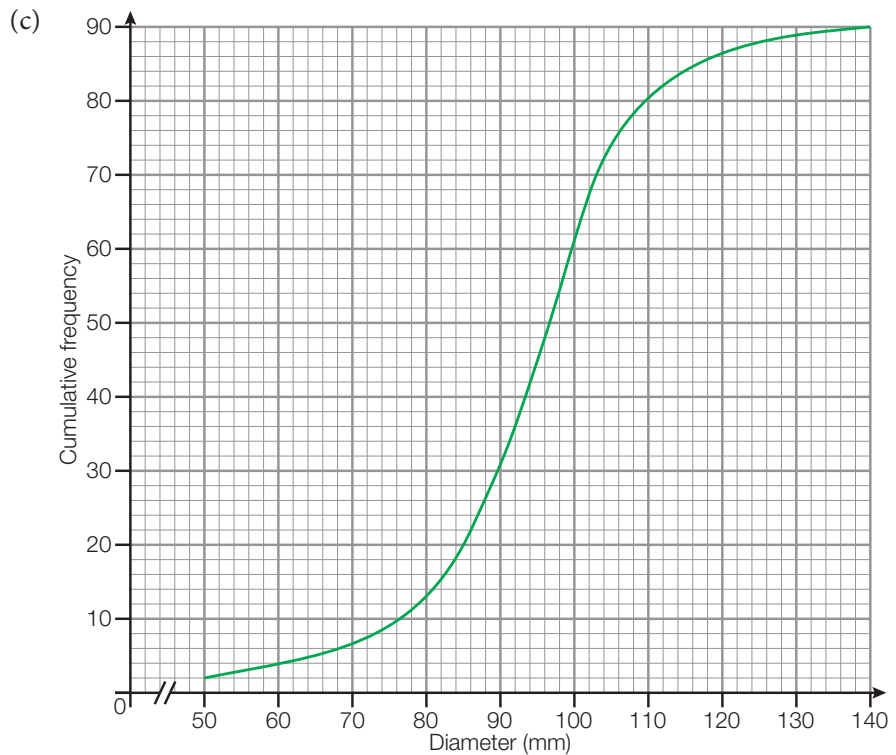
(f) The price of 1997 three-door Alfa Romeos in Rome:

Price (\$)	1349	1499	1599	1849	2399	2899	3349	3399	5799
Frequency	15	18	20	22	25	12	8	7	3

3. For each of the following cumulative frequency curves, estimate:

(i) the median (ii) the lower and upper quartiles (iii) the interquartile range.





Francis Galton
(1911–1922)

developed the concept of standard deviation, and pioneered its use in statistics. He was passionately interested in measuring all aspects of daily life. The data he collected on the length of a man's forearm in relation to his height led him to formulate the concept of correlation.

exam tip

In examinations, questions are set so that you can always use your GDC to find standard deviations. As part of your project, you may wish to explore the standard deviation using the method described in Learning links 7A, but this is outside the syllabus.

7.2 Standard deviation

The **standard deviation** is another measure of spread about the 'average'. It tells you the spread of the data about the **mean** value. The standard deviation measures the 'average' distance of all the data points from the mean. It takes into account **all** of the data in the set. The smaller the standard deviation, the closer the data values are to the mean and the more representative the mean is of the data set.

We will look at how to calculate the standard deviation for various types of data. There is a formula that can be used to calculate the standard deviation, but it is quicker to use a statistics program on your GDC.

Simple data

The Formula booklet does not contain a formula for the standard deviation, and in the examinations you are expected to use your GDC. However, to understand what the standard deviation means, you might find it helpful to calculate a few without using a statistical program on your calculator. See Learning links 7A if you would like to see how to calculate the standard deviation using more traditional methods.

Learning links

7A Calculating the standard deviation step by step

The standard deviation is based on finding the distance between each data value and the mean, and then averaging those distances.

To do this:

1. Calculate the mean value of the data.
2. Find the distance of each data value from the mean.
3. Square these distances.
4. Calculate the mean of all the squared distances; that is, add them up and divide by the total number of values. The result is called the **variance**.
5. Take the square root of your result in the previous step. This is the standard deviation.

The steps are easier to follow if you put the calculations in a table. For example, taking the six values 1, 2, 4, 11, 12 and 15, we first find their mean:

$$\bar{x} = (1 + 2 + 4 + 11 + 12 + 15) \div 6 = 7.5$$

Then we calculate the distances of each value from the mean, and their squares, in a table:

x	$(x - \bar{x})$	$(x - \bar{x})^2$
1	$1 - 7.5 = -6.5$	42.25
2	$2 - 7.5 = -5.5$	30.25
4	$4 - 7.5 = -3.5$	12.25
11	$11 - 7.5 = 3.5$	12.25
12	$12 - 7.5 = 4.5$	20.25
15	$15 - 7.5 = 7.5$	56.25
Total		173.5

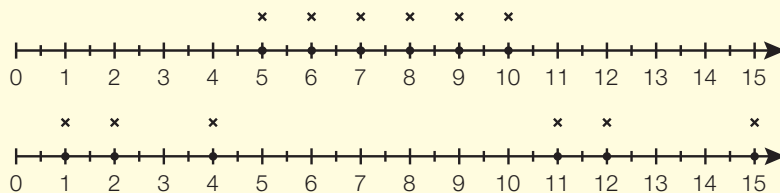
The variance is $173.5 \div 6 = 28.916\dots$

So the standard deviation is $\sqrt{28.916\dots} = 5.38$ (3 s.f.)

If you have a large data set, this would be a long calculation. You can save time by using a spreadsheet or other software package.

Worked example 7.2

Q. The following diagrams show two sets of data.



For both data sets the mean is 7.5. Calculate the standard deviation for both data sets and determine which mean is a more accurate representation of the data set.

Use your GDC to calculate the standard deviation. For a reminder of how to do this, see '6.1 (a) Finding the mean, median, quartiles and standard deviation for a simple list of data...' on page 666 of the GDC chapter.

A.

TEXAS

CASIO

```
1-Var Stats
x̄=7.5
Σx=45
Σx²=355
Sx=1.870828693
σx=1.707825128
↓n=6
```

```
1-Variable
x̄=7.5
Σx=45
Σx²=355
σx=1.70782512
sx=1.87082869
n=6 ↓
```

The standard deviation for the first set of data is $\sigma_x = 1.71$ (3 s.f.)

Write down the answer appropriately.

Use your GDC to calculate the standard deviation of the second set of data.

TEXAS

CASIO

```
1-Var Stats
x̄=7.5
Σx=45
Σx²=511
Sx=5.89067059
σx=5.377421935
↓n=6
```

```
1-Variable
x̄=7.5
Σx=45
Σx²=511
σx=5.37742193
sx=5.89067059
n=6 ↓
```

The standard deviation of the second set of data is $\sigma_x = 5.38$ (3 s.f.)

Write down the answer appropriately.

GDC

Your GDC actually gives two standard deviation values: s_x and σ_x . In this course we will only be using the value σ_x , so be careful that you read off the correct result from your GDC.

The standard deviation of the first set of data is much lower than the standard deviation of the second set of data. This suggests that the data values in the first set are more closely distributed around the mean; this can also be seen in the diagrams. The mean of 7.5 is a more accurate representation of the average value in the first set of data than it is in the second. Since 5.38 is more than three times as big as 1.71, the spread of data about the mean is more than three times greater in the second set than in the first set.

Discrete data organised in a frequency table

For ungrouped data in a table, you can calculate the standard deviation with your GDC by entering the data values in List 1 and the frequencies in List 2. See '6.1 (b) Finding the mean, median, quartiles and standard deviation for grouped data...' on page 667 of the GDC chapter if you need a reminder of how.

For example, recall Dee's table of the number of siblings of her brother's schoolmates from Worked example 5.2.

Number of siblings	Frequency
0	12
1	21
2	14
3	9
4	4
Total number of children questioned	60



The standard deviation is used in many statistical applications. It plays a key role in understanding the normal distribution (see Chapter 11), a distribution that is widely used in fields as diverse as finance, psychology, quality control and population studies.



TEXAS



CASIO

L1	L2	L3	Z
0	12	---	
1	21	---	
2	14	---	
3	9	---	
4	4	---	
L2(6) =			

SUB	List 1	List 2	List 3	List 4
3	2	14		
4	3	9		
5	4	4		
6				
GRAPH CALC TEST DISTR DIST				

1-Var Stats	
\bar{x}	=1.533333333
Σx	=92
Σx^2	=222
s_x	=1.171217918
σ_x	=1.161416759
n	=60

1-Variable	
\bar{x}	=1.533333333
Σx	=92
Σx^2	=222
σ_x	=1.161416759
s_x	=1.17121791
n	=60

From GDC, $\sigma_x = 1.16$.

Grouped discrete or continuous data

For grouped data in a frequency table, enter the **mid-interval values** of the data classes in List 1 and the frequencies in List 2. Again, see section 6.1 (b) on page 667 of the GDC chapter if you need to.

For example, let's revisit the table showing heights of trees from section 6.4.

Height (m)	Mid-interval value	Frequency
5–9	7	12
10–14	12	18
15–19	17	18
20–24	22	8
25–29	27	3
Total		59



TEXAS

CASIO

L1	L2	L3	2
7	12	-----	
12	18		
17	18		
22	8		
27	8		
-----	-----		
L2(G) =			

SUB	List 1	List 2	List 3	List 4
1	7	12		
2	12	18		
3	17	18		
4	22	8		
				12
[1VAR] [2VAR] [REG]				[SET]

```

1-Var Stats
x̄=14.62711864
Σx=863
Σx²=14441
Sx=5.598331972
σx=5.550685727
↓n=59
    
```

```

1-Variable
x̄=14.6271186
Σx=863
Σx²=14441
σx=5.55068572
sx=5.59833197
n=59
↓
    
```

exam tip

To estimate the mean or standard deviation for a grouped data set, remember to always use the mid-interval value.

From GDC, $\sigma_x = 5.55$ (3 s.f.).

Exercise 7.2

- The table below shows data from Meteogroup UK for 19 July 2011 across 11 towns and cities.

Town/city	Sunshine (hours)	Rainfall (inches)	Temperature (°C)	
			Minimum	Maximum
Edinburgh	1.0	0.12	11	19
Glasgow	3.6	0.01	13	19
Hull	2.8	0.51	13	20
Ipswich	6.9	0.03	11	21
Leeds	6.0	0.00	12	21
Lincoln	1.9	0.24	11	20
London	1.6	0.55	11	19
Manchester	0.4	0.24	12	16
Southampton	3.3	0.28	11	23
St Andrews	3.5	0.63	11	20
Stornoway	0.0	0.02	11	13

Source: Meteogroup UK.

Calculate the mean and standard deviation across all 11 locations for:

- hours of sunshine
- amount of rainfall in inches
- minimum temperature
- maximum temperature.

2. A five-day temperature forecast across nine cities is shown in the table below.

Temperature (°C)	14	15	16	17	18	19	20	21	22
Frequency	2	1	5	7	13	10	4	2	1

Using the information from the table above, calculate:

- (a) the mean temperature for the five-day period
 (b) the corresponding standard deviation for the data.
3. The price and number of bookings for various cruises in a three-month period are represented in the table below.

Price (£)	Number of bookings
1700	12
1850	18
2150	24
2530	26
2880	35
3100	15
3500	12
4300	6
5600	2

- (a) Calculate the mean cost of a cruise.
 (b) Calculate the standard deviation of the cost of these cruises.
4. The maximum spectator numbers for two different football leagues are shown below. Calculate the mean and standard deviation of attendances across the stadia for each league.

(a)

Attendance n (thousands)	Number of stadia
$10 \leq n < 26$	8
$26 \leq n < 42$	12
$42 \leq n < 58$	15
$58 \leq n < 74$	9
$74 \leq n < 90$	5

(b)

Attendance n (thousands)	Number of stadia
$10 \leq n < 26$	11
$26 \leq n < 42$	17
$42 \leq n < 58$	5
$58 \leq n < 74$	2
$74 \leq n < 90$	4
$90 \leq n < 106$	1

5. The performance of a group of students in a History project is illustrated in the table below. Calculate the mean and standard deviation of the students' scores.

Score (out of 70)	Number of students
21–25	3
26–30	7
31–35	15
36–40	20
41–45	7
46–50	4
51–55	4
56–60	1
61–65	4

7.3 Using the interquartile range and standard deviation to make comparisons

Once you have collected data and calculated the mean, median, interquartile range and/or the standard deviation, what do you do with this information? Statistical values such as these are useful for making general statements about your data, which in turn could be used to draw conclusions or predict future data.

Two IB students, Abi and Ben, are analysing the statistical information that they have gathered on a Biology field trip. They are studying shellfish and want to find out if limpets are larger when they are closer to the sea. They measure the diameter of groups of limpets at different points on a beach. Before they compare the different groups, they want some reassurance that the samples they've collected for each area are representative of the limpets in that area. For one group of 80 limpets, they collected the following results.



A group of limpets.

Diameter d (cm)	Frequency
$0 < d \leq 1$	7
$1 < d \leq 2$	12
$2 < d \leq 3$	15
$3 < d \leq 4$	19
$4 < d \leq 5$	17
$5 < d \leq 6$	10
Total	80

Abi drew a cumulative frequency curve and calculated the median and interquartile range of the data:

- Median = 3.3 cm
- Interquartile range = 2.3 cm

Ben calculated the mean and standard deviation:

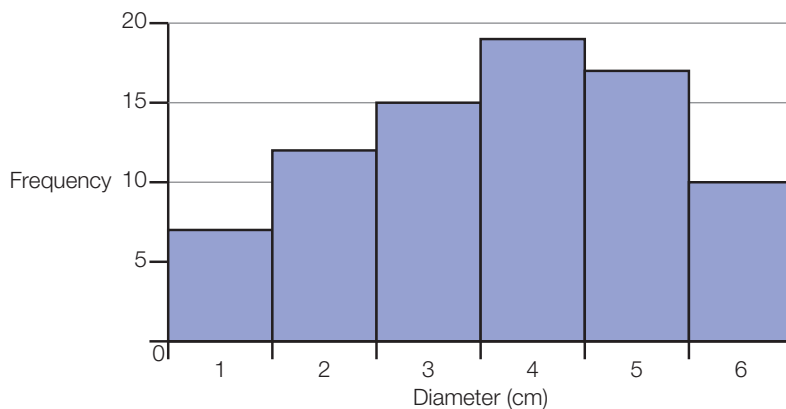
- Mean = 3.21 cm
- Standard deviation = 1.48 cm

Ben and Abi looked at the frequency table to find the modal class:

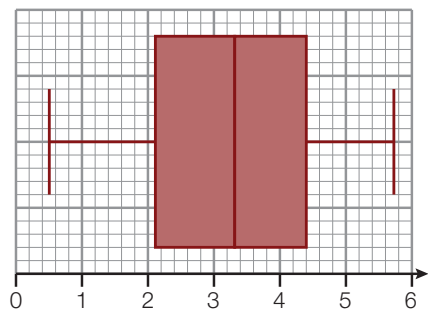
- Modal class is $3 < d \leq 4$

From this information we can see that the mean and median are close together, and both of these values lie within the modal class. This suggests that the limpets within the sample are all similar sized (suggesting they fairly represent the size of limpets in that area); but be careful – these values on their own are not proof that their sample is good. They need further support.

By drawing a histogram, Abi and Ben could see that the data is not symmetrical but is very slightly leaning towards higher values. But we can also see that it is not distorted completely to one direction either (as it would be if most of the data were concentrated at one end of the range of values), nor is it distorted by the data being very spread out.



The box and whisker diagram also seems to suggest that the data leans slightly towards the higher values:



In their research, scientists often compare populations from around the world to understand more about the topic they are studying. Linking up with another IB college to share data would enable students to take advantage of working this way too.

exam tip

When you use your GDC to draw a histogram or box and whisker plot, the scale will not be shown on the axes, so you will need to define this in what you write down.

The interquartile range tells us that the central 50% of data (the 'box' in the box and whisker diagram) spans about 2.3 cm.

The standard deviation is small, which means that the data is quite closely clustered around the mean.

From the above combined analyses of this data set, you could say that it is highly likely that Abi and Ben have collected a good sample on their field trip. The sizes are not biased in one direction and the data is close to the measure of central tendency, which suggests the mean or median shell size taken from this sample is a good representation of the shell size of all limpets in this area.

Worked example 7.3



- Q. An ornithologist weighed 12 adult male sparrows (*Passer domesticus*) in spring and again in autumn. The masses were recorded in grams.



Spring mass (g)	36.6	33.7	25.2	31.4	28.0	27.8	38.0	26.4	26.3	34.1	30.6	26.0
Autumn mass (g)	35.3	38.7	34.5	29.8	31.9	33.6	32.6	29.6	32.2	30.7	33.5	38.9

Calculate the mean and standard deviation of the spring masses using your GDC. See '6.1 (a) Finding the mean, median, quartiles and standard deviation for a simple list of data...' on page 666 of the GDC chapter if you need a reminder.

- A. (a) Calculate the mean and standard deviation of the masses in spring and the masses in autumn.
 (b) Comment on your results.

(a)  **TEXAS**  **CASIO**

```
1-Var Stats
x̄=30.34166667
Σx=364.1
Σx²=11260.91
Sx=4.405669557
σx=4.218107461
↓n=12
```

```
1-Variable
x =30.34166666
Σx =364.1
Σx² =11260.91
σx =4.21810746
sx =4.40566955
n =12
```

$$\bar{x} = 30.3 \text{ g}, \sigma_x = 4.22 \text{ g}$$



continued . . .

Calculate the mean and standard deviation of the autumn masses using your GDC.



TEXAS



CASIO

```
1-Var Stats
x̄=33.44166667
Σx=401.3
Σx²=13522.35
Sx=3.048235953
σx=2.918463732
↓
n=12
```

```
1-Variable
x̄=33.44166666
Σx=401.3
Σx²=13522.35
σx=2.91846373
sx=3.04823595
n=12 ↓
```

Compare the means.
Compare the standard deviations.

$$\bar{x} = 33.4 \text{ g}, \sigma_x = 2.92 \text{ g}$$

- (b) In autumn the mean mass of the birds is greater than it is in spring. In spring the standard deviation is larger, suggesting that the data is more spread out. So the autumn masses are both heavier and more consistent.

This data supports the hypothesis that in autumn birds are heavier because they have gained mass over the summer when food is abundant. In spring, birds have just gone through a period of limited food sources during the winter and so are less able to gain mass. However, some birds may have retained more mass through the winter because of access to bird feeders or habitats with plenty of grain and seeds, which could account for the greater spread of masses in the spring.

Exercise 7.3

1. The following data shows the number of goals scored by each of the ten teams in the ANZ Netball Championship during the 2010 and 2011 seasons.

2010 season	621	646	599	533	662	664	684	677	647	758
2011 season	571	594	679	651	696	717	644	682	681	704

- (a) Calculate:
- the mean number of goals scored per team in each season
 - the standard deviation of the number of goals scored per team in each season.
- (b) Use your answers from part (a) to compare the goal-scoring performance of the teams in the two seasons.

2. Mrs Chan has two IB Mathematical Studies groups. The marks scored by her students in the last class test are shown below

Group MS-A	25	68	93	30	42	22	33	35	35	67	77	50	97	98	95
Group MS-B	53	70	52	27	57	52	77	48	87	50	40	62	58	78	50

- (a) Calculate:
- the median mark for each group
 - the quartiles and hence the interquartile range of the test marks for each group.
- (b) Use your answers from above to compare the performance of the two groups in the test.
3. The following are the waiting times in minutes for blood tests at two separate laboratories.

Lab A	52	43	44	42	51	51	57	47	52	50	41	44	42	47	52	54	45	49
	50	42	43	40	46	49	54	44	51	51	43	47	46	59	58	58	50	51
Lab B	52	39	40	40	59	55	56	44	49	50	37	38	40	55	58	57	45	49
	51	39	38	39	52	55	57	44	53	53	42	40	43	51	55	62	47	51

- (a) Create a grouped frequency table for each set of data.
(*Hint: you can use the groups 37–39, 40–42, 43–45, etc.*)
- (b) Construct separate cumulative frequency tables for the two sets of data.
- (c) Draw cumulative frequency curves for the two sets of data.
- (d) From your curves, estimate the median and the interquartile range of the waiting times at each laboratory.
- (e) Use your results from part (d) to compare the waiting times for blood tests in the two laboratories.
4. The following table shows ten countries ranked by population in 2011.

Rank	Country	Population (millions)
1	China	1337
2	India	1189
3	United States	311
4	Indonesia	246
5	Brazil	203
6	Pakistan	
7	Nigeria	166
8	Bangladesh	159
9	Russia	139
10	Japan	127
	Mean	406

- (a) Given that the mean population of the ten countries was 406 million, find an estimate of the population of Pakistan in 2011.
- (b) Determine the median population of the ten countries.
- (c) Calculate the standard deviation of the populations.
5. The world mid-year population for 2011 of people under 90 years old by age and sex is summarised in the table below.

Age	Male (millions)	Female (millions)
0–14	949	885
15–29	901	859
30–44	745	725
45–59	539	549
60–74	275	303
75–89	83	119

- (a) Work out estimates for the mean and standard deviation of the male population and of the female population.
- (b) Calculate the combined mean and standard deviation.
- (c) Use your answers from parts (a) and (b) to fill in the following table:

	Males	Females	Combined
Mean			
Standard deviation			

- (d) Comment on the differences and similarities between the male and female population distributions.

Summary

You should know:

- that the dispersion of a data set estimates how spread out the data set is, and therefore indicates how good a representation of the data the measure of central tendency is
- that there are three ways of measuring dispersion: range, interquartile range and standard deviation
- how to calculate the range, interquartile range and standard deviation
- how to make sensible comments about your data based on these values and the additional support of histograms, cumulative frequency curves and/or box and whisker diagrams.

Mixed examination practice

Exam-style questions

1. The mean of the 12 numbers listed below is 7.

5, x , 10, 6, 9, 10, y , 3, 8, 9, 1, 8

- Write a simplified equation connecting x and y .
 - Given that the mode of the numbers is 8 and $x < y$, find the values of x and y .
 - Hence find the median of the numbers.
 - Determine the lower and upper quartiles.
 - Find the interquartile range.
2. The table below shows the tuition fees for 65 schools in the USA.

Fees F (US dollars)	Number of schools
$5,000 \leq F < 10,000$	2
$10,000 \leq F < 15,000$	4
$15,000 \leq F < 20,000$	8
$20,000 \leq F < 25,000$	8
$25,000 \leq F < 30,000$	25
$30,000 \leq F < 35,000$	18

- In which interval does the median lie?
 - What is the modal group?
 - Work out an estimate of the mean value of the tuition fees.
 - Construct a cumulative frequency graph for the data.
 - Use your graph to estimate the following values:
 - the median
 - the interquartile range.
 - Given that the minimum and maximum tuition fees are \$5000 and \$35,000 respectively, draw a box and whisker diagram to represent the given data.
3. The following table shows the time spent in the staff room by teachers during one lunch break.

Time (minutes)	Frequency	Cumulative frequency
0–4	1	1
5–9	4	5
10–14	7	12
15–19	6	18
20–24	4	x

Time (minutes)	Frequency	Cumulative frequency
25–29	10	32
30–34	y	34
35–39	9	43
40–44	5	48

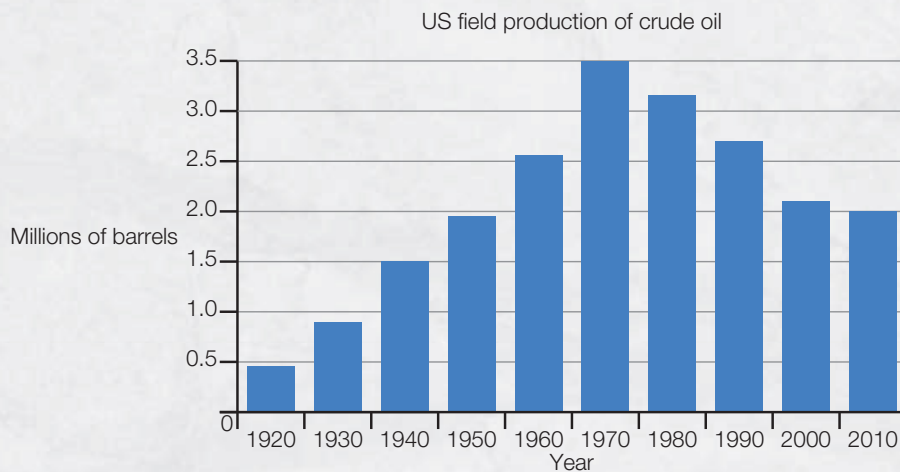
Determine the values of x and y .

4. The following table shows the number of text messages received by a group of students in a 24-hour period.

Number of text messages	Frequency	Cumulative frequency
3	1	1
4	3	4
5	c	8
6	6	14
7	9	23
8	10	33
9	d	e
10	6	47
11	f	50

- (a) How many students were there in the group?
 (b) Work out the values of c , d , e and f .
5. Data on the US field production of crude oil (in millions of barrels) between 1920 and 2010 is displayed in the diagram below.

(Source: U.S. Energy Information (2012); <http://www.eia.gov/>)

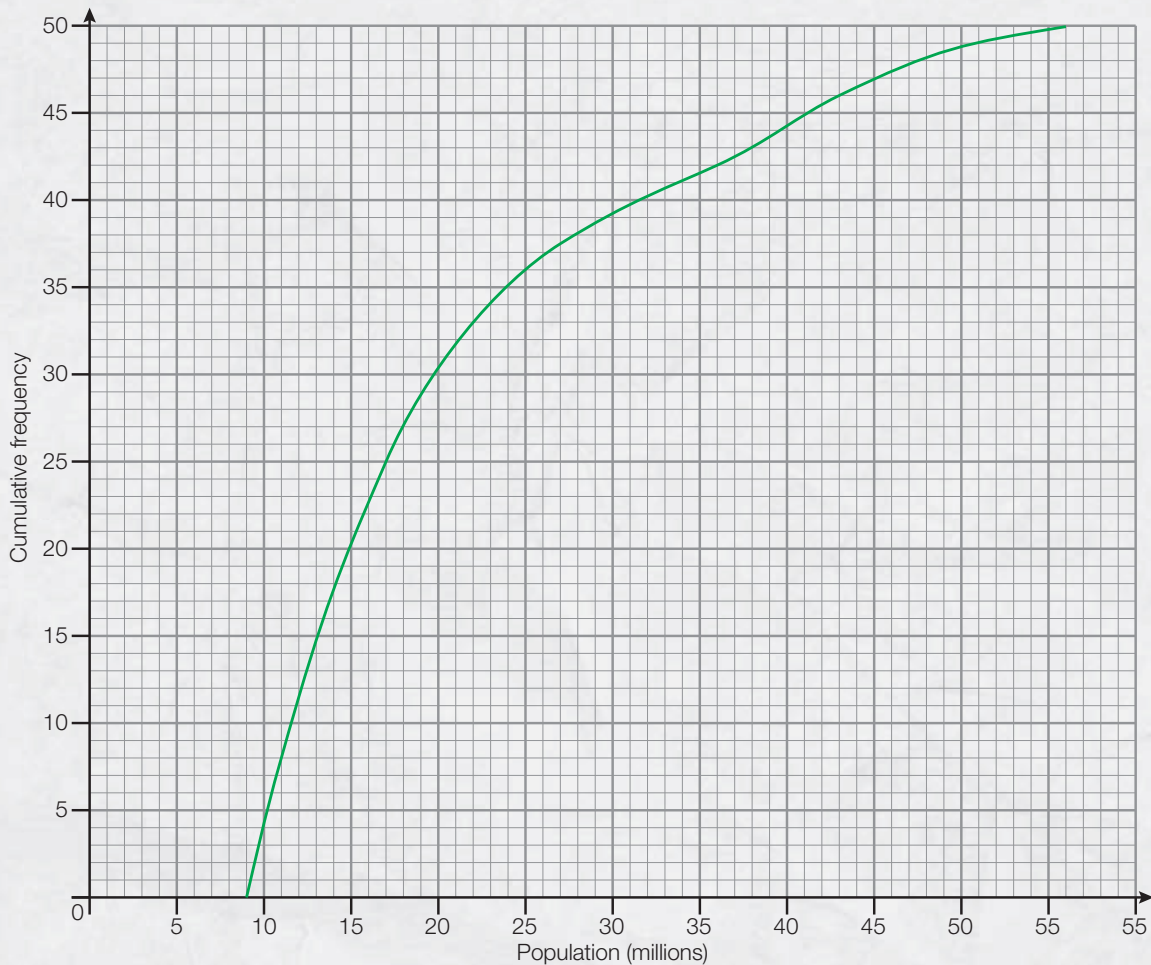


- (a) Find an estimate of the total volume of crude oil produced in the given years.
 (b) Work out the mean volume of crude oil produced per year.

6. An Under-16s Youth Football League has 12 teams in Division One and 13 teams in Division Two. At the end of the 2010 season, the mean number of points earned by the teams was 32.5 for Division One and 36 for Division Two.

After a disciplinary hearing, six of the Division One matches were cancelled at the end of the season. As a result, all 18 points earned by the teams in these matches (3 for each match) were withdrawn.

- (a) Work out the reduced mean number of points for teams in Division One after the disciplinary hearing.
- (b) What is the combined mean number of points for both divisions after the disciplinary hearing?
7. The graph below is the cumulative frequency graph of the populations of 50 countries with between 8 million and 56 million people.



- (a) Use the graph to estimate:
- the median population
 - the number of countries with a population less than 43 million
 - the percentage of countries with a population greater than 36 million.
- (b) Draw a box and whisker diagram to represent the population distribution of these countries.

8. 90 students were asked how long it took them to get ready for school in the morning. Their responses are summarised in the table below.

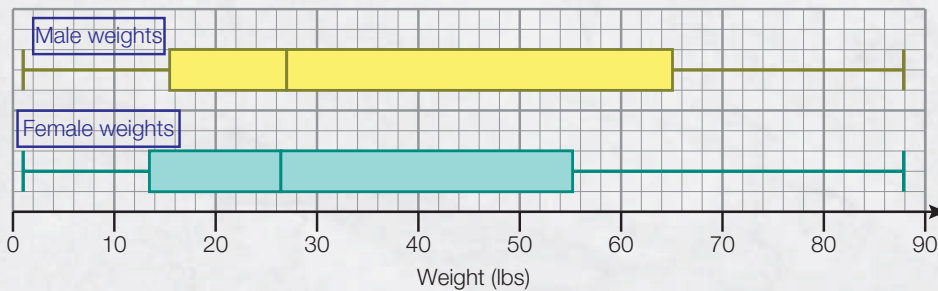
Time t (minutes)	$10 \leq t < 15$	$15 \leq t < 20$	$20 \leq t < 25$	$25 \leq t < 30$	$30 \leq t < 35$	$35 \leq t < 40$
Frequency	6	14	16	34	18	2

- In which interval does the median lie?
 - What is the modal group?
 - Calculate estimates for the mean and standard deviation of the times.
 - Draw a frequency histogram to represent the data.
 - Use your answers from above to comment on the data.
9. The table below summarises the weights of 22 male and 22 female foxes.

Weight w (lbs)	Frequency	
	Male	Female
$0 \leq w < 15$	6	7
$15 \leq w < 30$	7	6
$30 \leq w < 45$	1	1
$45 \leq w < 60$	1	4
$60 \leq w < 75$	5	3
$75 \leq w < 90$	2	1

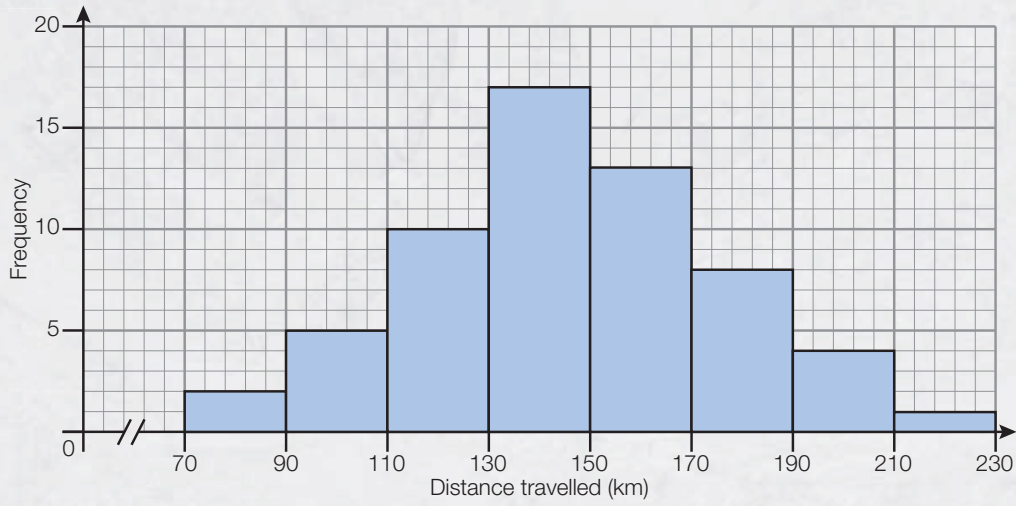
- Draw two separate frequency histograms to represent the information.
- Write down the modal groups for the weights of the foxes.
- Calculate estimates of the mean and standard deviation of the weights.

The box and whisker diagrams for the data are shown below.



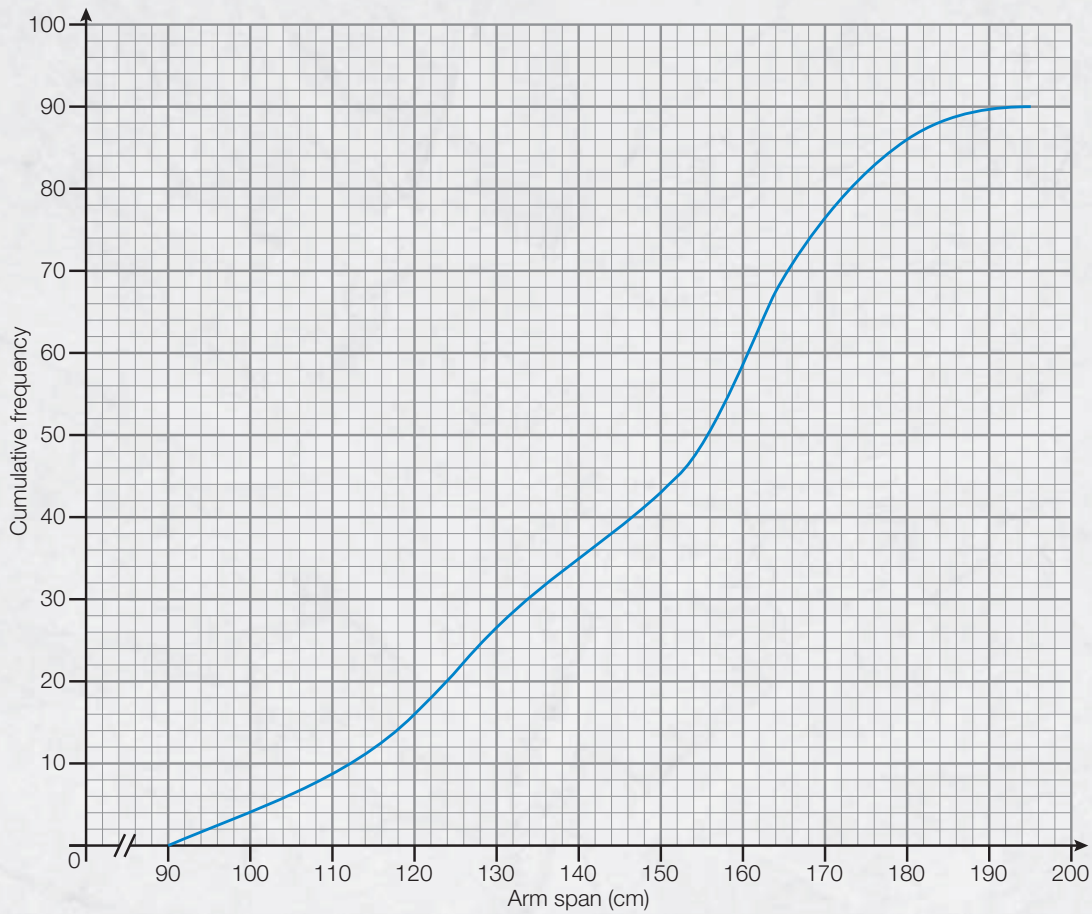
- Use the information from the diagrams to comment on the differences/similarities between the weights of the male and female foxes.

10. The following histogram shows the distances travelled by salesmen of an advertising company during a randomly chosen week.



- Determine the total number of journeys made by the salesmen during the week.
- Write down the interval containing:
 - the modal distance
 - the median distance
- Work out estimates of:
 - the mean distance travelled
 - the standard deviation of the distances
- Use your answers from above to comment on the journeys made by the salesmen.

11. The cumulative frequency diagram below shows the lengths of the arm spans of members of a sports club.



- (a) From the graph, find:
- the median length
 - the interquartile range of the lengths.
- (b) Draw a box and whisker diagram to represent the data.

The graph was drawn from the following table:

Length L (cm)	Frequency
$90 \leq L < 105$	6
$105 \leq L < 120$	u
$120 \leq L < 135$	13
$135 \leq L < 150$	v
$150 \leq L < 165$	24
$165 \leq L < 180$	w
$180 \leq L < 195$	4

- (c) Using information from the graph, find the values of u , v and w .
- (d) Calculate an estimate of the mean length of the arm spans.