2 Descriptive statistics

CHAPTER OBJECTIVES:

4.1-4.3 Discrete and continuous data: frequency tables; mid-interval values; upper and lower boundaries. Frequency histograms

4.4 Cumulative frequency tables; cumulative frequency curves; median and quartiles. Box and whisker diagrams

4.5 Measures of central tendency: mean; median; mode; estimate of a mean; modal class

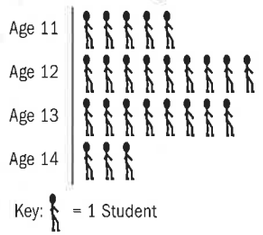
4.6 Measures of dispersion: range, interquartile range, standard deviation

Before you start

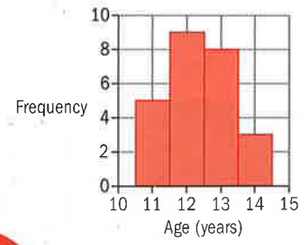
You should know how to:

1 Collect and represent data using

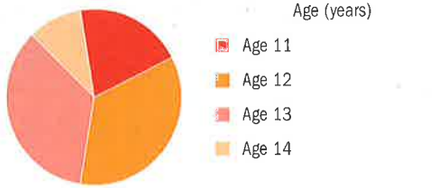
a a pictogram



b a bar chart



c a pie chart



2 Set up axes on graphs using given scales.

Skills check

1 Maerwen wants to find out information about the numbers of men, women, boys and girls using a library. Design a suitable data-collection sheet to collect the information.

2 These data show the number of different colored sweets in a packet.

Color	blue	green	red	orange	yellow
Frequency	5	7	8	4	6

a Draw a pictogram to represent these data.

b Draw a bar chart to represent these data.

c Draw a pie chart to represent these data.

3 On graph paper, draw a set of axes such that 1 cm represents 2 units on the x-axis and 1 cm represents 10 units on the y-axis.

Every country needs basic information on its population so that it can plan and develop the services it needs. For example, to plan a road network you need to know the size of the population so you can estimate the amount of traffic in an area.

To collect information on a population, governments often organize a census. A census is a survey of the **whole population** of a country.

The information collected includes data on age, gender, health, housing, employment and transport. The data are then analyzed and presented in tables, charts and spreadsheets. All data should be processed so that information on individuals is protected. The United Nations recommends that population censuses should be taken at least every 10 years.

In what other areas of society is mathematics used in a practical way?

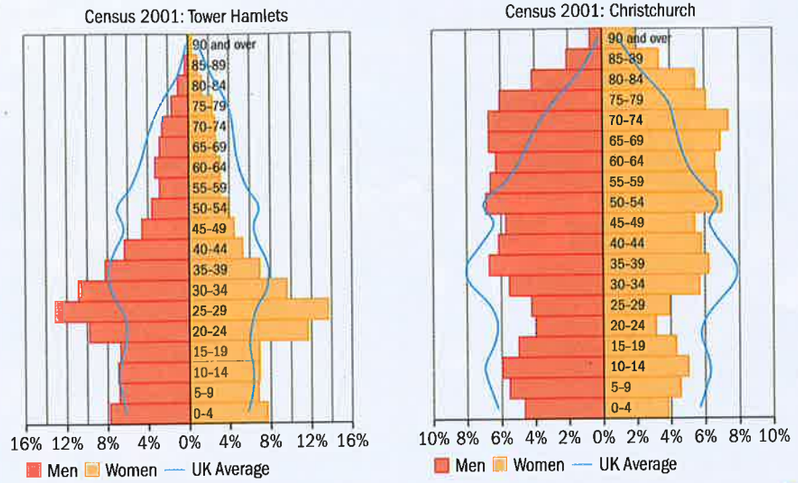
What are the benefits of sharing and analyzing data from different countries?

When was the last census in your country? Is the census information published in the public domain? How has technology changed the way census data is collected and presented?

Investigation – population distribution

In the United Kingdom, there is a census every 10 years.

These population pyramids are based on information collected in the 2001 census. They show the distribution of age ranges in Tower Hamlets, London, and Christchurch, Dorset.



Compare the population pyramids for Tower Hamlets and Christchurch.

Simply based on these data, make a number of conjectures about these two areas.

Fully research the areas to test your conjectures. How accurate were you?

All information from the 2001 census can be found at www.ons.gov.uk by searching for '2001 census data'.

In this chapter you will organize data in frequency tables, graph data in a variety of diagrams, and analyze data using a range of measures.

How is education data used to investigate the link between the level of education and patterns of creating families and fertility?

2.1 Classification of data

There are two main types of data: **qualitative** and **quantitative**. Qualitative data are data that are not given numerically, for example, favorite color. Quantitative data are numerical.

Quantitative data can be further classified as **discrete** or **continuous**.

→ **Discrete data** are either data that can be **counted** or data that can only take **specific values**.

Examples of data that can be counted include the number of sweets in a packet, the number of people who prefer tea to coffee, and the number of pairs of shoes that a person owns.

Examples of data that can only take specific values include shoe size, hat size and dress size.

Is the number of grains of salt in a salt cellar discrete?

→ **Continuous data** can be **measured**. They can take any value within a range.

Examples of continuous data include weight, height and time.

Continuous data can be expressed to a required number of significant figures. The greater the accuracy required, the more significant figures the data will have.

⚠ The number of shoes and shoe size are examples of discrete data.

The weighing scale was invented at a time when countries began trading materials and a standard measurement was required to ensure fair trading.

Time is a continuous measure because it can take any numerical value in a particular range. For example, the time taken for world-class sprinters to run 100 m can be recorded as any fraction of a second.

Population and sample

When conducting a statistical investigation, the whole of a group from which we may collect data is known as the **population**. It is not always possible, or even necessary, to access data for a whole population.

⚠ A weighing scale gives us continuous data.

You can make conclusions about a population by collecting data from a sample. It is usually cheaper and quicker to collect data from a sample.

A **sample** is a small group chosen from the population.

A **random sample** is one where each element has the same chance of being included.

A **biased sample** is one that is not random.

It is important that a sample is random and not biased – it must be **representative** of the elements being investigated. To ensure that the different members of the population have an equal probability of being selected you could choose people by picking names out of a hat. Or you could assign a number to each member of the population and then choose numbers at random using the random number function on a GDC.

Can the wording of a survey question and the way the data are presented introduce bias?

Sampling will not be examined. However, if you use sampling when writing your Mathematical Studies project, you will need to discuss how you picked your sample and convince the moderator that it is indeed a random sample.

Are exit polls a good way of predicting the results of an election?

Example 1

Kiki wants to find out if there is any connection between eating breakfast and grades among students in her school. However, there are too many students in the school to ask everyone. She needs to pick a sample.
How can she make sure that the sample she picks is a random sample?

Answer

Kiki can use her GDC to generate random numbers and use the students who have those numbers on the school register.

Does each student have the same chance of being included in her sample? If yes, it is a random sample.

In market research, a sample of the population is interviewed in order to collect data about customers. Many research methods have been developed since companies began to carry out formal market research in the 1920s.

Example 2

Ayako is conducting a survey to find out how much money women who live in London spend on fashion in a month. She only interviews women coming out of Harrods (a very exclusive store). Is this a random sample?

Answer

No, because the sample will not come from the total population of women in London and some of the women she interviews may not even belong to the population.

Is Ayako only asking 'women who live in London'?
Do all women who live in London shop at Harrods?

Exercise 2A

1 State whether these data are discrete or continuous.

a The number of sweets in a packet
b The heights of students in Grade 8
c The dress sizes of a girls' pipe band
d The number of red cars in a parking lot
e The weights of kittens
f The marks obtained by Grade 7 in a science test
g The times taken for students to write their World Literature paper
h The weights of apples in a 5 kg bag
i The number of cm of rain each day during the month of April
j The number of heads when a coin is tossed 60 times
k The times taken for athletes to run a marathon
l The number of visitors to the Blue Mosque each day.

2 State whether the following samples are random or biased.

a When researching if people eat breakfast, only interview the people in the canteen.
b When researching spending habits, interview every third person you meet.
c When researching spending habits on cars, Josh interviews men exiting a garage.
d When comparing GDP to child mortality, Eizo chooses the countries from a numbered list, by generating random numbers on his GDC.
e When researching the sleeping habits of children, Adham distributes a questionnaire to the students in his school.

GDP, gross domestic product, is the total value of goods produced and services provided in a country in a year.

2.2 Simple discrete data

When there is a large amount of data, it is easier to interpret if the data are organized in a **frequency table** or displayed as a graph.

Example 3

The numbers of sweets in 24 packets are shown below.

22 23 22 22 23 21 22 22 20 22 24 21
22 21 22 23 22 22 24 20 22 23 22 22

Organize this information in a frequency table.

Answer

Number of sweets	Tally	Frequency
20	||	2
21	|||	3
22	𝍸 𝍸 |||	13
23	||||	4
24	||	2
	TOTAL	**24**

Draw a chart with three columns.

Write the possible data values in the 'Number of sweets' column.

Use tally marks to record each value in the 'Tally' column.

For each row, count up the tally marks and write the total in the 'Frequency' column.

Add up the values in the 'Frequency' column to work out the total frequency.

Now you can see how many packets have each number of sweets.

Exercise 2B

1 The numbers of goals scored by Ajax football team during their last 25 games were:

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

Organize this information in a frequency table.

2 The numbers of heads obtained when twelve coins were tossed 50 times are recorded below.

8 3 5 7 1 9 2 10 5 12 7 6 6 8 12 4 10 2 6 6 8 4 5 11 3
4 6 8 6 7 5 3 11 2 10 5 6 7 5 8 9 2 10 11 0 12 3 6 6 5

Organize this information in a frequency table.

3 The ages of the girls in a hockey club are:

10 11 12 10 9 11 15 13 12 16 11 13 14 12 10 10 11 9 9 10
10 12 15 16 12 11 13 10 15 13 12 11 15 16 11 12 10 9 10 11

Organize this information in a frequency table.

4 It is stated that there are 90 crisps in a box.
Viktoras checked 30 boxes and the numbers of crisps in them are recorded below.

90 90 91 90 89 89 90 90 92 90 90 88 89 90 90
91 90 89 90 88 89 90 91 90 92 88 89 90 90 90

Organize this information in a frequency table.

5 Sean threw a dice 50 times. The numbers that appeared are shown below.

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

Organize this information in a frequency table.

EXAM-STYLE QUESTION

6 The numbers of games played in matches at a badminton tournament are recorded below.

8 8 10 11 9 7 8 7 11 12 7 8 10 10 11 9 9 8 11 7 9 8

The raw data have been organized in the frequency table.

Games	Frequency
7	4
8	m
9	4
10	n
11	4
12	1

Write down the values of m and n.

2.3 Grouped discrete or continuous data

When there are a lot of data values spread over a wide range it is useful to **group** the data. Depending on the number of data values, there should be between 5 and 15 groups, or classes, of equal width.

The classes must cover the range of the values and they must not overlap – each data value must belong to only one class.

You can organize both discrete and continuous data in **grouped frequency tables**.

Example 4

Loni made 30 telephone calls one week. The times of the calls, in minutes, were recorded.

3.1 12.2 9.6 8.1 2.2 1.2 15.0 4.8 21.2 13.6
17.3 22.3 1.5 4.6 31.2 26.7 7.8 18.2 35.4 1.6
2.9 5.5 12.8 28.3 16.9 1.3 5.6 7.8 2.3 6.9

Organize this information in a grouped frequency table.

The frequency table gives a much clearer picture of the data.

Answer

Time (t)	Frequency
$0 \le t < 5$	10
$5 \le t < 10$	7
$10 \le t < 15$	3
$15 \le t < 20$	4
$20 \le t < 25$	2
$25 \le t < 30$	2
$30 \le t < 35$	1
$35 \le t < 40$	1

First decide on the size and the number of classes:

Smallest number = 1.2 so classes start at 0.

Largest number = 35.4 so classes finish at 40.

Using a class width of 5, there will be (40 ÷ 5 =) 8 classes in total.

Exercise 2C

1 Organize each of these sets of data in a grouped frequency table.

a 2 5 12 21 7 9 25 31 17 19 22 23 15 24 5
34 45 32 13 43 7 11 32 6 18 40 23 32 22 8

b 10 24 31 29 42 19 55 65 46 72 35 48 68 56 92
12 33 77 56 45 82 76 56 34 12 78 89 45 59 32
26 97 67 54 34 18 77 59 34 27 13 19 63 65 22

c 1 3 8 12 4 2 6 3 9 10 11 9 7 5 14 2 3 16
9 5 13 14 4 8 17 3 15 19 5 3 9 10 11 14 15

Upper and lower boundaries

To find the **upper** and **lower boundaries** of a class, calculate the mean of the upper value from one class and the lower value from the following class.

Example 5

This table shows the heights of flowers in a garden.
Write down
a the upper boundary of the first class
b the lower boundary of the third class.

Height (x cm)	Frequency
$0 \le x < 10$	5
$10 \le x < 20$	12
$20 \le x < 30$	21
$30 \le x < 40$	15
$40 \le x < 50$	6

Answers

a $\frac{10+10}{2} = 10$

Upper value of the first class is 10. Lower value of the second class is 10. The upper boundary of the first class is the mean of these two values.

b $\frac{20+20}{2} = 20$

Upper value of the second class is 20. Lower value of the third class is 20. The lower boundary of the third class is the mean of these two values.

Example 6

The table shows the numbers of pairs of shoes of each size sold in a shop one day.
Write down
a the upper boundary of the first class and the last class
b the lower boundary of the first class and the fourth class.

Shoe size	Frequency
15–19	3
20–24	9
25–29	12
30–34	22
35–39	45
40–44	31

These are European shoe sizes. What are the equivalent shoe sizes in your country?

How could the shoe shop manager use this data?

Answers

a Upper boundary of the first class: $\frac{19+20}{2} = 19.5$

Upper boundary of the last class: $\frac{44+45}{2} = 44.5$

Upper value of the first class is 19. Lower value of the second class is 20. The upper boundary of the first class is the mean of these two numbers. Similarly for last class.

b Lower boundary of the first class: $\frac{14+15}{2} = 14.5$

Lower boundary of the fourth class: $\frac{29+30}{2} = 29.5$

Upper value of the previous class would be 14. Lower value of the first class is 15. The lower boundary of the first class is the mean of these two numbers. Similarly for fourth class.

Exercise 2D

1 Copy these tables and fill in the missing lower and upper boundary values.

a

Class	Lower boundary	Upper boundary
9–12		12.5
13–16		
17–20	16.5	
21–24		

b

Time (t seconds)	Lower boundary	Upper boundary
$2.0 \le t < 2.2$		
$2.2 \le t < 2.4$		
$2.4 \le t < 2.6$		

Frequency histograms

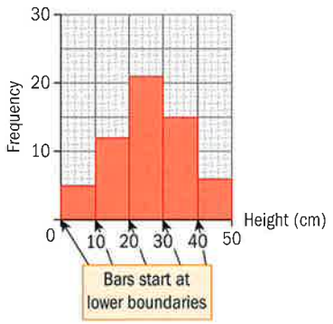
A **frequency histogram** is a useful way to represent data visually.

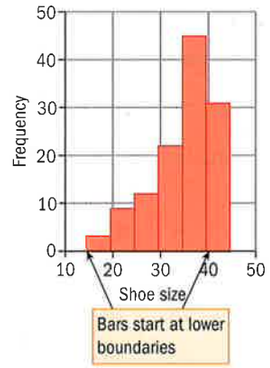
→ To draw a frequency histogram, find the lower and upper boundaries of the classes and draw the bars between these boundaries. There should be no spaces between the bars.

In the Mathematical Studies course you will only deal with frequency histograms that have equal class intervals.

The class boundaries are plotted on the x-axis and the frequency values on the y-axis.

Here are the frequency histograms for Examples 5 and 6.





English statistician Karl Pearson (1857–1936) first used the term 'histogram' in 1895.

Exercise 2E

1 The costs, in euros, of 80 dinners are shown in the table.
Draw a histogram to display this information.

Cost of dinner in euros (c)	Frequency
$10 \le c < 15$	2
$15 \le c < 20$	8
$20 \le c < 25$	11
$25 \le c < 30$	25
$30 \le c < 35$	14
$35 \le c < 40$	11
$40 \le c < 45$	6
$45 \le c < 50$	3

2 The table shows the age distribution of teachers at Genius Academy.

a Write down the lower and upper boundaries of each class.

b Draw a histogram to represent the information.

Age (x)	Frequency
$20 \le x < 30$	4
$30 \le x < 40$	8
$40 \le x < 50$	10
$50 \le x < 60$	9
$60 \le x < 70$	3

3 The masses of 150 melons are recorded in the table.

a Write down the lower and upper boundaries of the third class.

b Draw a histogram to represent the information.

Mass ($x$kg)	Frequency
$0.4 \le x < 0.6$	21
$0.6 \le x < 0.8$	36
$0.8 \le x < 1.0$	34
$1.0 \le x < 1.2$	29
$1.2 \le x < 1.4$	18
$1.4 \le x < 1.6$	12

4 The lengths of 100 worms (to the nearest cm) are given in the table.

a Write down the lower and upper boundaries of each class.

b Draw a histogram to represent the information.

Length (cm)	4	5	6	7	8	9	10
Frequency	18	20	26	15	8	6	7

5 50 people were asked how often they traveled by train each month. The results were:

8	7	10	5	23	4	16	9	62	28
14	53	29	11	34	33	68	75	12	79
22	54	67	55	13	32	41	58	36	2
26	80	65	38	52	71	2	16	36	40
18	24	52	64	76	16	6	18	28	40

a Organize this information in a grouped frequency table.

b Draw a histogram to represent the information graphically.

6 Yuri decided to count the number of weeds in one square metre of grass. He chose 80 plots of one square metre. The results for each square metre are:

22	24	21	12	8	14	34	62	54	6	28	42	35	22	14	18	9	24	12	18
31	47	17	9	35	24	41	52	38	19	5	23	31	65	32	46	15	13	74	22
9	13	22	55	47	52	14	13	21	19	52	33	71	12	22	17	58	42	31	16
2	15	31	73	45	31	12	8	4	33	42	57	61	48	43	27	14	5	14	26

a Organize this information in a grouped frequency table.

b Draw a histogram to represent the information graphically.

7 Simi recorded the numbers of vans per five minutes that drove down her street over a period of eight hours. Her results were:

Number of vans (x)	Frequency
$1 \le x \le 5$	12
$6 \le x \le 10$	23
$11 \le x \le 15$	31
$16 \le x \le 20$	13
$21 \le x \le 25$	9
$26 \le x \le 30$	5
$31 \le x \le 35$	2
$36 \le x \le 40$	1

a Write down the lower and upper boundaries of the fourth class.

b Draw a histogram to represent the information.

EXAM-STYLE QUESTION

8 The numbers of visitors per hour to the Taj Mahal are recorded in the table.

Time (t)	Number of visitors
$09{:}00 \le t < 10{:}00$	324
$10{:}00 \le t < 11{:}00$	356
$11{:}00 \le t < 12{:}00$	388
$12{:}00 \le t < 13{:}00$	435
$13{:}00 \le t < 14{:}00$	498
$14{:}00 \le t < 15{:}00$	563
$15{:}00 \le t < 16{:}00$	436
$16{:}00 \le t < 17{:}00$	250
$17{:}00 \le t < 18{:}00$	232

Draw a histogram to represent this information.

2.4 Measures of central tendency

Data can be summarized by using a measure of central tendency such as the mode, median or mean.

→ The **mode** of a data set is the value that occurs most frequently.

The **median** of a data set is the value that lies in the middle when the data are arranged in size order.

The **mean** of a data set is the sum of all the values divided by the number of values.

When there are two 'middle' values, the median is the midpoint between the two middle values. To find the midpoint, add the two middle values and divide by two.

How do you know which measure of central tendency is the best to use?

Can you mislead people by quoting statistics? For example, the numbers 1, 1, 100 have mode = 1, median = 1 and mean = 34.

You have to be aware that there may be outliers (isolated points outside the normal range of values) that skew the statistics.

What are the ethical implications of using statistics to mislead people?

Example 7

Here is a set of data: 5 4 8 4 4 7 8 9 11 1 5
Find the mode, median and mean.

Answer

5 (4) 8 (4) (4) 7 8 9 11 1 5
Mode = 4

The value '4' occurs three times.

1 4 4 4 5 (5) 7 8 8 9 11
Median = 5

First arrange the data in size order. There are 11 entries, so the median is the $\frac{11+1}{2}$ *= 6th value.*

Mean = $\frac{1+4+4+4+5+5+7+8+8+9+11}{11}$

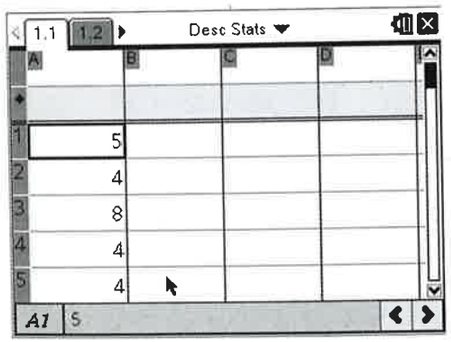
$= \frac{66}{11}$

Mean = 6

The mean is the $\frac{\text{sum of all the values}}{\text{number of values}}$

GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

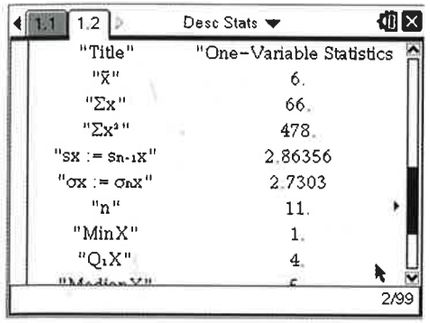
You can also use a GDC to calculate the median and mean.
Enter the data values:



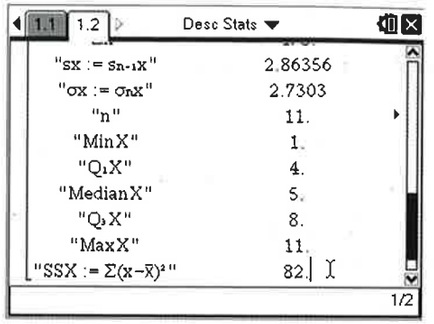
For help with entering data values, see Chapter 12, Section 2.1.

The GDC screen is too small to display all of the values in the list. Scroll down to see the remaining values.

The value of the mean is given by $\bar{x}$ (pronounced 'x-bar'):



The value of the median is shown as 'MedianX':



Exercise 2F

1 Calculate the mode, median and mean for each data set.

a 7 3 8 9 1 10 1

b 3 4 8 2 5 6 11 13 3 5 6 5

2 Calculate the values of **a**, **b**, **c**, **d** and **e** in this table.

Data	Median	Mode	Mean
Height (m): 1.52, 1.74, 1.83, 1.52, 1.67, 1.91	**a**	**b**	1.70
Age (years): 21, 34, 17, 22, 56, 38	28	none	**c**
Weight (kg): 54.7, 48.6, 63.2, 55.1, 77.9, 48.6	**d**	48.6	**e**

The German psychologist Gustav Fechner (1801–1887) popularized the use of the median, although French mathematician and astronomer Pierre-Simon Laplace (1749–1827) had used it previously.

3 The weights of eight pumpkins are 26.3 kg, 12.6 kg, 33.5 kg, 8.9 kg, 18.7 kg, 22.6 kg, 31.8 kg and 45.3 kg.

a Find the median weight.

b Calculate the mean weight.

EXAM-STYLE QUESTIONS

4 For these data the mode is 5, the median is 6 and the mean is 6.5.

1 1 2 3 s 5 5 7 8 9 10 t 12 12

Given that $s < t$, find the values of s and t.

5 Jin obtained marks of 76, 54 and 65 in his Physics, Biology and History examinations respectively.

a Calculate his mean mark for the three examinations.

b Find the mark that Jin must achieve in Mathematics so that the mean mark for the four examinations is exactly 68.

EXAM-STYLE QUESTION

6 Zoe and Shun compared their test scores. Zoe had a mean of 81 after taking five tests and Shun had a mean of 78 after taking three tests. Each of them took one more test and ended up with the same mean score of 80.

a Find the grade that Zoe gained on her sixth test.

b Find the grade that Shun gained on his fourth test.

Mean, median and mode from a frequency table

➔ For data in a frequency table, the **mode** is the entry that has the largest frequency.

The **median** is the middle entry as the entries in the table are already in order. For n pieces of data, the median is the $\frac{n+1}{2}$th value.

The next example shows how to calculate the mean from a frequency table.

Example 8

Calculate the mode, median and mean of these data.

Number of sweets	Frequency
20	2
21	3
22	13
23	4
24	2
TOTAL	24

Sometimes a question asks for the 'modal value'. This means 'the mode'.

Answer

Mode = 22

Median = 22

'22' has the highest frequency (13).

Median is the $\frac{24+1}{2} = 12.5$*th entry, so it is between the 12th and 13th entry. Both the 12th and 13th entries are 22, so the median = 22.*

To calculate the mean: Label the first column x_i*. Label the second column* f_i*. Add a third column and label it* $f_i x_i$*. Work out* $f_i \times x_i$ *for each row:*

$2 \times 20 = 40$

$3 \times 21 = 63$

$13 \times 22 = 286$

$4 \times 23 = 92$

$2 \times 24 = 48$

Work out the total of the f_i *column and the total of the* $f_i x_i$ *column.*

Number of sweets, x_i	Frequency, f_i	$f_i x_i$
20	2	40
21	3	63
22	13	286
23	4	92
24	2	48
TOTAL	24	529

Mean $= \frac{529}{24} = 22.0$ (to 3 sf)

Mean $= \frac{\text{total of } f_i x_i}{\text{total of } f_i}$

➔ The **mean** from a frequency table is:

$$\text{mean} = \frac{\text{total of } f_i \times x_i}{\text{total frequency}}$$

where f_i is the frequency of each data value x_i and $i = 1, \ldots, k$, where k is the number of data values.

The IB formula for the mean is:

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

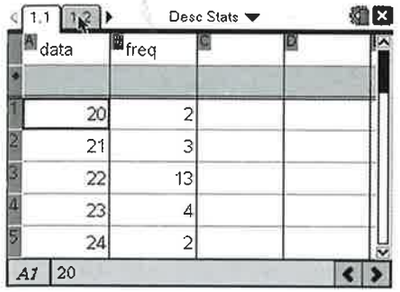
$n = \sum_{i=1}^{k} f_i$

The Σ notation simply means 'sum'.

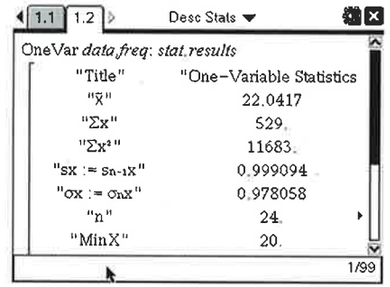
This formula is given in the Formula booklet.

You can also use your GDC to calculate the mean and median from a frequency table.

Enter the data values:



The value of the mean is given by $\bar{x}$:



The value of the median is shown as 'MedianX':

Exercise 2G

1 A dice is thrown 29 times and the score noted. The results are shown in the table.

Score	Frequency
1	4
2	7
3	3
4	8
5	5
6	2

a Write down the modal score.

b Write down the median score.

c Calculate the mean score.

EXAM-STYLE QUESTION

2 The table shows the frequency of the number of visits to the doctor per year for a group of children.

Number of visits	0	1	2	3	4	5
Frequency	4	3	8	5	4	1

a How many children are in the group?

b Write down the modal number of visits.

c Calculate the mean number of visits.

EXAM-STYLE QUESTIONS

3 A bag contains six balls numbered 1 to 6. A ball is drawn at random and its number noted. The ball is then returned to the bag.
The numbers for the first 30 draws are:

Number	1	2	3	4	5	6
Frequency	4	5	3	n	6	5

a Write down the value of n.
b Calculate the mean number.
c Write down the modal number.

4 The table gives the frequency of grades achieved by students in an IB school.

Grade	Frequency
1	1
2	6
3	19
4	34
5	32
6	18
7	10

a Calculate the mean grade.
b What percentage of students achieved a grade 4 or 5?
c Write down the modal grade.

Mean, median and mode for grouped data

For grouped data, you can find the modal class and an **estimate of the mean**.

➔ For grouped data, the **modal class** is the group or class interval that has the largest frequency.

The next example shows how to calculate an estimate of the mean.

Example 9

The times, in seconds, taken to complete 200 bouts of sumo wrestling are shown in the table.

Time (t seconds)	Frequency
$0 \le t < 20$	37
$20 \le t < 40$	62
$40 \le t < 60$	46
$60 \le t < 80$	25
$80 \le t < 100$	11
$100 \le t < 120$	9
$120 \le t < 140$	6
$140 \le t < 160$	4
TOTAL	200

You do not know the exact data values for each group. Use the midpoint of each class interval as an estimate of the values in each group. You may also find the midpoint referred to as the 'mid-interval value'.

To find the midpoint of a class, find the mean of the class limits.
midpoint = $\frac{\text{lower boundary} + \text{upper boundary}}{2}$

Calculate **a** the modal class and **b** an estimate of the mean.

▶ Continued on next page

Answer

a Modal class = $20 \le t < 40$

This class interval has the largest frequency (62).

b

Time (t seconds)	Frequency, f_i	Midpoint, x_i	$f_i x_i$
$0 \le t < 20$	37	10	370
$20 \le t < 40$	62	30	1860
$40 \le t < 60$	46	50	2300
$60 \le t < 80$	25	70	1750
$80 \le t < 100$	11	90	990
$100 \le t < 120$	9	110	990
$120 \le t < 140$	6	130	780
$140 \le t < 160$	4	150	600
TOTAL	200		9640

Mean = $\frac{9640}{200}$ = 48.2 (to 3 sf)

To work out an estimate of the mean, you must first work out the ***midpoint*** *of each class interval. Add a third column and label it 'Midpoint, x_i'. Work out each midpoint:*

Midpoint of $0 \le t < 20$: $\frac{0+20}{2} = 10$

Midpoint of $20 \le t < 40$: $\frac{20+40}{2} = 30$

Midpoint of $40 \le t < 60$: $\frac{40+60}{2} = 50$

Next add a fourth column and label it '$f_i x_i$'.
Then work out $f_i \times x_i$ for each row:
$9 \times 110 = 990$
$6 \times 130 = 780$

Work out the total of the f_i column and the total of the $f_i x_i$ column.

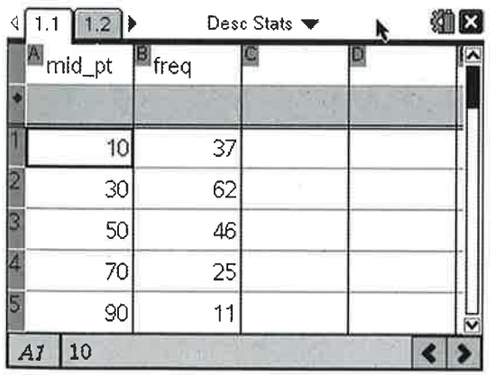
Mean = $\frac{\text{total of } f_i x_i}{\text{total } f_i}$

➔ To calculate an estimate of the **mean** from a grouped frequency table, use $\frac{\text{total of } f_i \times x_i}{\text{total frequency}}$ where f_i is the frequency and x_i is the corresponding midpoint of each class.

Why does this give an estimate of the mean and not an exact value?

You can also use a GDC to calculate an estimate of the mean from a grouped frequency table.

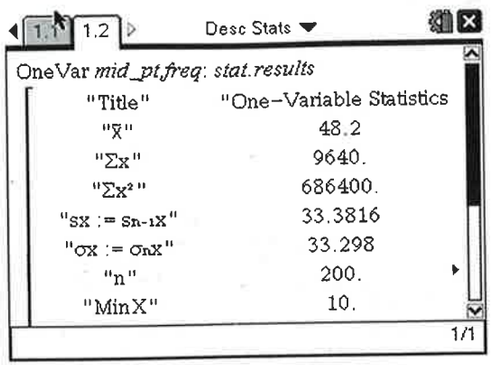
Enter the data values:



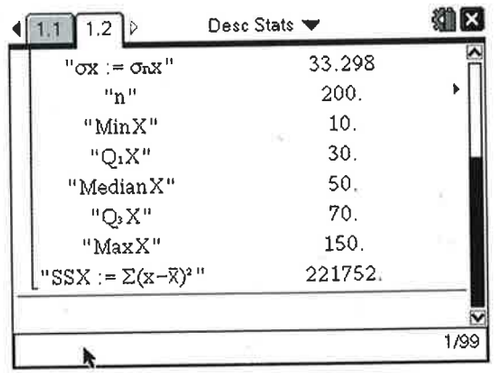
For help with entering data values, see Chapter 12, Section 2.2.

GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

The value of the mean is given by $\bar{x}$:



You were not asked for the median but the GDC works it out as part of the calculation screen (this too is only an estimate as we do not have all the individual values):



Exercise 2H

EXAM-STYLE QUESTIONS

1 The table shows the times taken for 25 cheetahs to cover a distance of 50 km.

a Write down the modal class.

b Calculate an estimate of the mean time taken.

Time taken (t minutes)	Frequency
$20 \le t < 22$	2
$22 \le t < 24$	5
$24 \le t < 26$	8
$26 \le t < 28$	4
$28 \le t < 30$	3
$30 \le t < 32$	2
$32 \le t < 34$	1

2 The speeds of vehicles passing under a bridge on a road are recorded in the table.

a Write down the modal class.

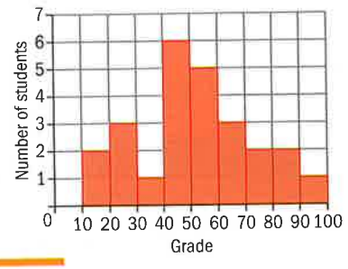
b Calculate an estimate of the mean speed of the vehicles.

Speed (s km h^{-1})	Frequency
$60 \le s < 70$	8
$70 \le s < 80$	15
$80 \le s < 90$	12
$90 \le s < 100$	10
$100 \le s < 110$	8
$110 \le s < 120$	3
$120 \le s < 130$	4

3 The results of a Geography test for 25 students are given in the diagram.

a Write down the modal class.

b Calculate an estimate of the mean grade.



2.5 Cumulative frequency curves

→ The **cumulative frequency** is the sum of all of the frequencies up to and including the new value. To draw a **cumulative frequency curve** you need to construct a cumulative frequency table, with the upper boundary of each class interval in one column and the corresponding cumulative frequency in another. Then plot the upper class boundary on the x-axis and the cumulative frequency on the y-axis.

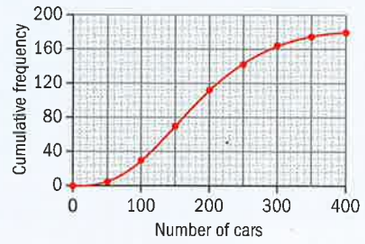
Example 10

A supermarket is open 24 hours a day and has a free car park. The number of parked cars each hour is monitored over a period of several days. The results are shown in the table.
Organize this information in a cumulative frequency table.
Draw a graph of the cumulative frequency.

Number of parked cars per hour	Frequency
0–49	6
50–99	23
100–149	41
150–199	42
200–249	30
250–299	24
300–349	9
350–399	5

Answer

Number of parked cars per hour	Frequency	Upper boundary	Cumulative frequency
0–49	6	49.5	6
50–99	23	99.5	29
100–149	41	149.5	70
150–199	42	199.5	112
200–249	30	249.5	142
250–299	24	299.5	166
300–349	9	349 5	175
350–399	5	399.5	180



Add a third column and label it 'Upper boundary'.
Work out the upper boundary of each class:

$$\text{Upper boundary} = \frac{49+50}{2} = 49.5$$

$$\text{Upper boundary} = \frac{99+100}{2} = 99.5$$

$$\text{Upper boundary} = \frac{149+150}{2} = 149.5$$

Now add a fourth column and label it 'Cumulative frequency'.
Work out the cumulative frequency for each row:

$6 + 23 = 29$
$29 + 41 = 70$
$166 + 9 = 175$
$175 + 5 = 180$

The final cumulative frequency value should equal the total frequency value. Cumulative frequency is always plotted on the **vertical** axis.

To draw the graph of the cumulative frequency, plot the value of the upper boundary against the cumulative frequency value.

Interpreting cumulative frequency graphs

We can use the cumulative frequency curve to find estimates of the **percentiles** and **quartiles**.

Percentiles separate large ordered sets of data into hundredths.
Quartiles separate large ordered sets of data into quarters.

When the data are arranged in order, the lower quartile is the 25th percentile, the median is the 50th percentile (middle value) and the upper quartile is the 75th percentile.

→ To find the **lower quartile**, Q_1, read the value on the curve corresponding to $\frac{n+1}{4}$ on the cumulative frequency axis, where n is the total frequency.
To find the median, read the value on the curve corresponding to $\frac{n+1}{2}$ on the cumulative frequency axis.
To find the **upper quartile**, Q_3, read the value on the curve corresponding to $\frac{3(n+1)}{4}$ on the cumulative frequency axis.
To find the **percentiles**, p%, read the value on the curve corresponding to $\frac{p(n+1)}{100}$ on the cumulative frequency axis.
To find the **interquartile range** subtract the lower quartile from the upper quartile: $IQR = Q_3 - Q_1$.

The cumulative frequency curve is sometimes called an ogive.

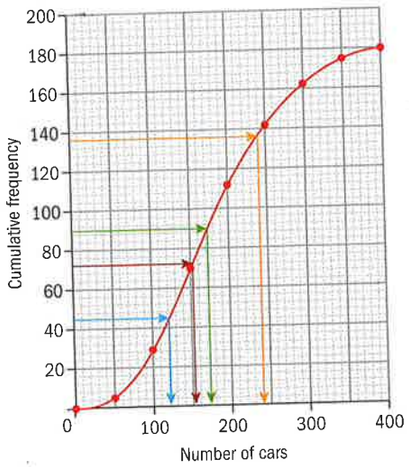
Per cent means out of 100.
$\frac{1}{4} = 25\%$
$\frac{1}{2} = 50\%$
$\frac{3}{4} = 75\%$

There are no universally agreed formulae for the quartiles. For large n and grouped data: n rather than $n + 1$ may be used.

The IQR shows the spread of the middle 50% of the data

For any set of data:
- 25% or one-quarter of the values are between the smallest value and the lower quartile
- 25% are between the lower quartile and the median
- 25% are between the median and the upper quartile
- 25% are between the upper quartile and the largest value
- 50% of the data lie between the lower and upper quartiles.

In this cumulative frequency diagram (from the data in Example 10), $n = 180$.



Lower quartile ≈ 120 (blue)
This is the value corresponding to $\frac{180+1}{4} = 45.25$.

Median ≈ 173 (green)
This is the value corresponding to $\frac{180+1}{2} = 90.5$.

Upper quartile ≈ 238 (orange)
This is the value corresponding to $\frac{3(180+1)}{4} = 135.75$.

40th percentile ≈ 153 (brown)
This is the value corresponding to $\frac{40(180+1)}{100} = 72.4$.

The interquartile range ≈ 238 – 120 = 118

Example 11

50 contestants play the game of Oware. In total they have to play 49 games to arrive at a champion. The average times for the 49 games are given in the table.

Time (t minutes)	Frequency
$3 \le t < 4$	4
$4 \le t < 5$	12
$5 \le t < 6$	18
$6 \le t < 7$	9
$7 \le t < 8$	3
$8 \le t < 9$	2
$9 \le t < 10$	1

The game of Oware is played all over the world and there is even an Oware Society.
Why must 50 contestants play 49 games to arrive at a champion? Can you prove this?

a Construct a cumulative frequency table for these data.
b Draw a cumulative frequency graph for these data.
c Use your graph to estimate
i the lower quartile **ii** the median
iii the upper quartile **iv** the interquartile range
v the 30th percentile.

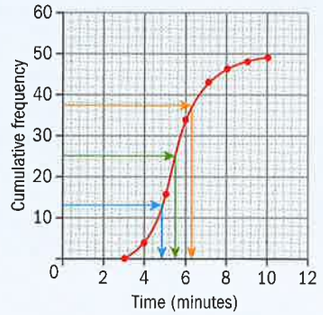
Answers

a

Time (t minutes)	Frequency	Upper boundary	Cumulative frequency
$3 \le t < 4$	4	4	4
$4 \le t < 5$	12	5	16
$5 \le t < 6$	18	6	34
$6 \le t < 7$	9	7	43
$7 \le t < 8$	3	8	46
$8 \le t < 9$	2	9	48
$9 \le t < 10$	1	10	49

Check:
Total frequency: 4 + 12 + 18 + 9 + 3 + 2 + 1 = 49
Final cumulative frequency value = 49

b



Plot each cumulative frequency at the upper boundary.

▶ Continued on next page

c i $n = 49$

$\frac{n+1}{4} = \frac{50}{4} = 12.5$

Lower quartile ≈ 4.7 minutes

25% of games last 4.7 minutes or less.

Read across from 12.5 on the vertical axis, then down to the horizontal axis.

ii $\frac{n+1}{2} = \frac{49+1}{2} = 25$

Median ≈ 5.5 minutes

50% of games last 5.5 minutes or less.

This is the value on the horizontal axis corresponding to 25 on the vertical axis.

iii $\frac{3(n+1)}{4} = \frac{3(49+1)}{4} = 37.5$

Upper quartile ≈ 6.4 minutes

75% of games last 6.4 minutes or less.

This is the value on the horizontal axis corresponding to 37.5 on the vertical axis.

iv Interquartile range = 6.4 − 4.7 = 1.7 minutes

The 'middle' 50% of games last between 4.7 and 6.4 minutes.

v $\frac{30(n+1)}{100} = \frac{30(49+1)}{100} = 15$

30th percentile ≈ 4.9 minutes

30% of games last 4.9 minutes or less.

This is the value on the horizontal axis corresponding to 15 on the vertical axis.

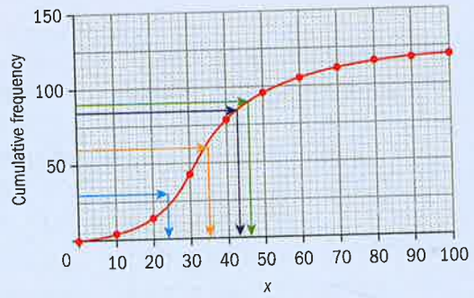
Example 12

From this cumulative frequency graph find

i the median

ii the interquartile range

iii the 70th percentile.



Answers

i $n = 120$

$\frac{n+1}{2} = \frac{121}{2} = 60.5$

Median ≈ 35

$n = 120$ from graph

Median is the value on the horizontal axis corresponding to 60.5 on the vertical axis.

ii Lower quartile $\frac{120+1}{4} = 30.25$th value

Lower quartile = 26

Upper quartile $\frac{3(120+1)}{4} = 90.75$th value

Upper quartile = 46

Interquartile range ≈ 46 − 26 = 20

Interquartile range = upper quartile − lower quartile

Upper quartile is the value corresponding to 90.75 on the vertical axis.

Lower quartile is the value corresponding to 30.25 on the vertical axis.

iii $\frac{70(120+1)}{100} = 84.7$th value

70th percentile ≈ 43

This is the value corresponding to 84.7 on the vertical axis.

Exercise 2I

EXAM-STYLE QUESTIONS

1 A dice is tossed 50 times. The number shown is recorded each time and the results are given in the table.

a Write down the value of N.

b Find the values of a, b and c.

Number	Frequency	Cumulative frequency
1	6	6
2	a	14
3	10	24
4	b	c
5	5	43
6	7	50
	N	

2 The table shows the percentages scored by candidates in a test.

Marks (%)	0–9	10–19	20–29	30–39	40–49	50–59	60–69	70–79	80–89	90–100
Frequency	1	5	7	11	19	43	36	15	2	1

Here is the cumulative frequency table for the marks.

a Calculate the values of s and of t.

b Draw a cumulative frequency graph for these data.

c Use your graph to estimate

i the median mark

ii the lower quartile

iii the pass mark, if 40% of the candidates passed.

Marks (%)	Cumulative frequency
< 9.5	1
< 19.5	6
< 29.5	s
< 39.5	24
< 49.5	43
< 59.5	86
< 69.5	t
< 79.5	137
< 89.5	139
≤ 100	140

3 A safari park is open to visitors every day of the year. The numbers of cars that pass through the park each day for a whole year were recorded and are shown in the table.

a Draw a cumulative frequency graph to represent this information.

b Find the median and the interquartile range.

c On what percentage of days were there more than 800 cars in the park?

Number of cars (n)	Frequency
$0 < n \leq 150$	25
$150 < n \leq 300$	36
$300 < n \leq 450$	68
$450 < n \leq 600$	102
$600 < n \leq 750$	64
$750 < n \leq 900$	41
$900 < n \leq 1050$	19
$1050 < n \leq 1200$	10

4 Sofia studied an article in the *Helsingborgs Dagblad*. She recorded the numbers of words in each sentence in the frequency table.

a Draw a cumulative frequency graph to represent this information.

b Work out the lower quartile, the median and the upper quartile of the data.

Number of words	Frequency
1–4	4
5–8	19
9–12	38
13–16	23
17–20	8
21–24	4
25–28	2
29–32	1
33–36	1

EXAM-STYLE QUESTIONS

5 A salmon farmer records the lengths of 100 salmon, measured to the nearest cm. The results are given in the table.

a Construct a cumulative frequency table for the data in the table.

b Draw a cumulative frequency curve.

c Use the cumulative frequency curve to find

i the median length of salmon

ii the interquartile range of salmon length.

Length of salmon (x cm)	Number of salmon
$25 < x \leq 28$	3
$28 < x \leq 31$	4
$31 < x \leq 34$	11
$34 < x \leq 37$	23
$37 < x \leq 40$	28
$40 < x \leq 43$	15
$43 < x \leq 46$	12
$46 < x \leq 49$	4
TOTAL	100

6 The table shows the times taken by 100 students to complete a puzzle.

Time (t minutes)	11–15	16–20	21–25	26–30	31–35	36–40
Number of students	6	13	27	31	15	8

a Construct a cumulative frequency table.

b Draw a cumulative frequency graph.

c Use your graph to estimate

i the median time

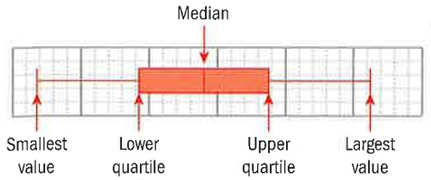
ii the interquartile range of the time

iii the time within which 75% of the students completed the puzzle.

2.6 Box and whisker graphs

Another useful way to represent data is a **box and whisker graph** (or box and whisker plot).

A box and whisker graph looks something like this.



➔ To draw a box and whisker graph, five pieces of information are needed: calculate the lower quartile, median and upper quartile for the data. Find the smallest and largest values.

Draw the box and whisker graph to scale on graph paper.

Note:

An **outlier** is a value that is much smaller or much larger than the other values.

Normally we consider an outlier to be a point with a value:

- less than 'the lower quartile − 1.5 × the interquartile range' or
- greater than 'the upper quartile + 1.5 × the interquartile range'.

Outliers will not be examined but they may be useful for projects.

Example 13

A yacht club hosts an annual race. The numbers of people in each yacht are recorded in the table.

a Find the median number of people in a yacht.

b Find the upper and lower quartiles.

c Draw a box and whisker graph to represent the information.

Number of people	Frequency
4	1
5	8
6	16
7	25
8	28
9	16
10	5
TOTAL	99

▶ Continued on next page

Answers

a $n = 99$, so the median is the number of people in the $\frac{99+1}{2} = \frac{100}{2} = 50$th yacht.

Number of people	Frequency	Cumulative frequency
4	1	1
5	8	9
6	16	25
7	25	50
8	28	78
9	16	94
10	5	99

The median number of people is 7.

The 50th yacht is in the group highlighted red.

b The lower quartile is the number of people in the $\frac{99+1}{4} = 25$th yacht.

The lower quartile is 6.

The 25th yacht is in the group highlighted green.

The upper quartile is the number of people in the $\frac{3(99+1)}{4} = 75$th yacht.

The upper quartile is 8.

The 75th yacht is in the group highlighted blue.

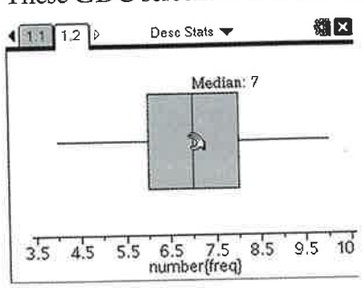
c

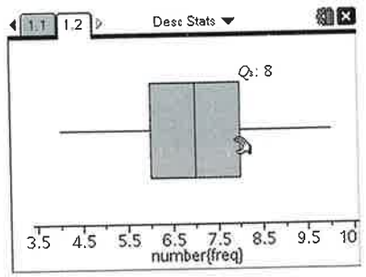
Need five pieces of information to draw a box and whisker graph:
Smallest number of people = 4
*Lower quartile = 6 (from part **b**)*
*Median = 7 (from part **a**)*
*Upper quartile = 8 (from part **b**)*
Largest number = 10

You can also find all the data for the box and whisker graph using your GDC.

Enter the 'Number of people' and 'Frequency' into lists named 'Number' and 'Freq' in a Lists & Spreadsheets page. Add a Data & Statistics page and press MENU 2: Plot Properties | 5: Add X Variable with Frequency and select the two lists. To read the values use the touchpad to move the arrow over them.

These GDC screenshots show the median and the upper quartile (Q_3).





GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

Example 14

The weights, in kilograms, of 25 koala bears are:
4.3, 7.2, 5.6, 4.8, 10.7, 9.7, 5.6, 7.8, 8.2, 11.4, 7.9, 12.6, 13.1, 5.7, 9.9, 11.3, 13.4, 8.8, 7.5, 5.8, 9.2, 10.3, 12.1, 6.5, 8.6
Draw a box and whisker graph to represent the information.

Answer

First arrange the data in ascending order:
4.3, 4.8, 5.6, 5.6, 5.7, 5.8, 6.5, 7.2, 7.5, 7.8, 7.9, 8.2, 8.6, 8.8, 9.2, 9.7, 9.9, 10.3, 10.7, 11.3, 11.4, 12.1, 12.6, 13.1, 13.4
$n = 25$

Lowest value = 4.3

Lower quartile: $\frac{25+1}{4} = 6.5$,
so between 6th and 7th value
6th value = 5.8,
7th value = 6.5,
6.5th value = $\frac{5.8+6.5}{2} = 6.15$

To find the 6.5th value, calculate the mean of the 6th and 7th values.

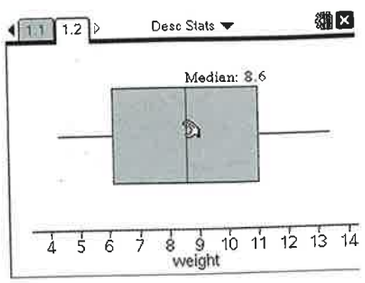
Median = 8.6
$\left(\text{the } \frac{25+1}{2} = 13\text{th value}\right)$

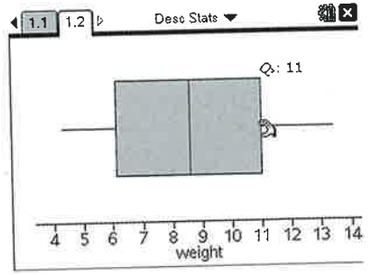
Upper quartile: $\frac{3\times 26}{4} = 19.5$,
so between 19th and 20th value
19th value = 10.7,
20th value = 11.3,
19.5th value = $\frac{10.7+11.3}{2} = 11$

Largest value = 13.4

Need five pieces of information to plot a box and whisker graph.

Using a GDC:
Enter the data into a list. You do not need to put it in order.
These GDC screenshots show the median and the upper quartile (Q_3).





GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

You cannot use a GDC to draw box and whisker graphs for grouped frequency tables.

Exercise 2J

1 The numbers of sweets in 45 bags are:

34 33 35 33 32 33 34 34 32 35 33 32 36 31 33 34
33 34 33 32 35 31 33 32 32 34 33 36 33 30 33 32
34 35 32 33 33 32 33 31 34 33 32 33 34

a Construct a frequency table to represent the information.
b Find the median, the lower quartile and the upper quartile.
c Draw a box and whisker graph to represent this information.
Use a GDC to check your answer.

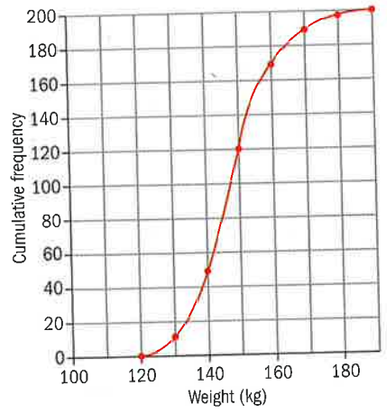
2 An experiment was performed 60 times. The scores from the experiment were recorded in the table.

Score	Frequency
1	6
2	12
3	13
4	15
5	8
6	6

a Find the median, the lower quartile and the upper quartile.
b Draw a box and whisker graph to represent this information.
Use a GDC to check your answer.

EXAM-STYLE QUESTION

3 The cumulative frequency graph shows the weights, in kg, of 200 sumo wrestlers.



a Write down
i the median
ii the lower quartile
iii the upper quartile.

The lightest wrestler weighs 125 kg and the heaviest weighs 188 kg.
b Draw a box and whisker graph to represent the information.

EXAM-STYLE QUESTIONS

4 The heights, in cm, of 180 students are given in the cumulative frequency table.

Height (x cm)	Cumulative frequency
$x \leq 145$	0
$x \leq 150$	26
$x \leq 155$	81
$x \leq 160$	119
$x \leq 165$	142
$x \leq 170$	154
$x \leq 175$	167
$x \leq 180$	174
$x \leq 185$	180

a Draw a cumulative frequency diagram to represent this information.
b Write down
i the median
ii the lower quartile and the upper quartile.
c The smallest student is 146 cm and the tallest is 183 cm. Represent this information on a box and whisker graph.

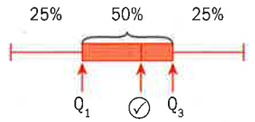
5 The table shows the heights, in cm, of 50 kangaroos.

Height (x cm)	Frequency
$200 \leq x < 210$	4
$210 \leq x < 220$	6
$220 \leq x < 230$	11
$230 \leq x < 240$	22
$240 \leq x < 250$	5
$250 \leq x < 260$	2

a Construct a cumulative frequency table and use it to draw the cumulative frequency curve.
b Write down the median.
c Find the lower quartile and the upper quartile.
The smallest kangaroo is 205 cm and the tallest is 258 cm.
d Draw a box and whisker graph to represent the information.

Interpreting box and whisker graphs

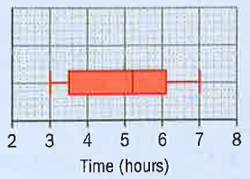
For any set of data:



- 25% or one-quarter of the values are between the smallest value and the lower quartile
- 25% are between the lower quartile and the median
- 25% are between the median and the upper quartile
- 25% are between the upper quartile and the largest value
- 50% of the data lie between the lower and upper quartiles.

Example 15

The box and whisker graph shows the times, in hours, that it takes to build an igloo.



a Write down the median time.
b Find the interquartile range.
c Write down the percentage of people who took less than 5.2 hours to build an igloo.
d x% of the people took more than 6.1 hours to build an igloo. Write down the value of x.

Continued on next page

Answers

a The median time is 5.2 hours.

b The interquartile range = 6.1 − 3.5
= 2.6 hours.

c 50% of the people took less than 5.2 hours to build an igloo.

d 25% of the people took more than 6.1 hours to build an igloo.

From the graph, upper quartile = 6.1, lower quartile = 3.5
*5.2 hours = median (from part **a**)*
50% of data are at or below this value
Upper quartile = 6.1
75% of data are at or below this value

Exercise 2K

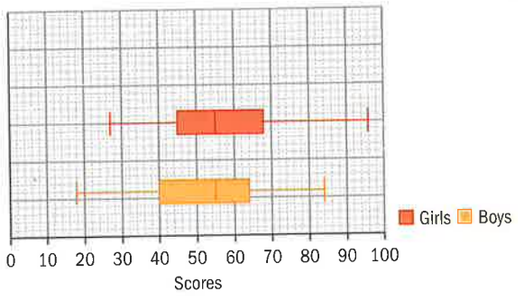
1 The box and whisker graphs represent the scores on a Psychology test for 40 boys and 40 girls.

a Find the median score for the boys and the girls.

b Write down the interquartile range for the boys' scores and the girls' scores.

c Write down the percentage of boys that scored more than 55.

d Write down the percentage of girls that scored more than 68.



2 The box and whisker graph represents the number of faults made by horses in a jumping competition.
Write down

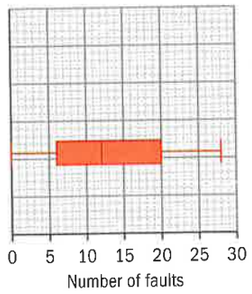
a the lowest number of faults

b the median

c the interquartile range

d the largest number of faults

e the percentage of horses that had fewer than six faults.



EXAM-STYLE QUESTION

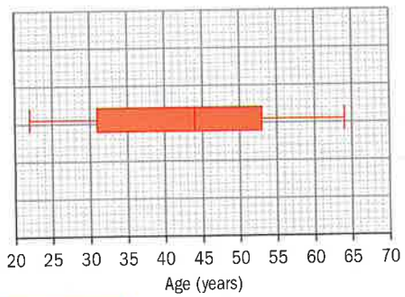
3 The box and whisker graph represents the ages of the teachers at Myschool High.

a Write down the age of the youngest teacher.

b Write down the median age.

c If 25% of the teachers are older than x, write down the value of x.

d Find the interquartile range of the ages.



Extension material on CD: *Worksheet 2 - Standard deviation, standardization and outliers*

2.7 Measures of dispersion

Measures of dispersion measure how spread out a set of data is. The simplest measure of dispersion is the **range**.

→ The **range** is found by subtracting the smallest value from the largest value.

Example 16

The numbers of piglets in the litters of 10 pigs are:
10 12 12 13 15 16 9 10 14 11
Find the range.

Answer
Range = 16 − 9 = 7

Identify the largest value (16) and the smallest value (9).

The **interquartile range** is found by subtracting the lower quartile, Q_1, from the upper quartile, Q_3: $IQR = Q_3 - Q_1$.

To work out the lower and upper quartiles the values must be arranged in size order

Example 17

Find the interquartile range of this data set.
4 5 6 6 7 8 10 10 11 14 15

Answer

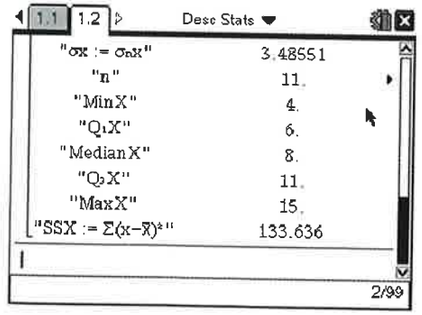
Q_1 is the $\frac{11+1}{4}$ = 3rd number,
so $Q_1 = 6$.

There are 11 numbers so $n = 11$.

Q_3 is the $\frac{3(11+1)}{4}$ = 9th number,
so $Q_3 = 11$.

$IQR = 11 - 6 = 5$

Using a GDC:
Enter the data into a list. Then use One Variable Statistics.
Scroll down to find the quartiles.
The value of Q_1 is given as 'Q_1X' and Q_3 as 'Q_3X'.



GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

You can use a GDC for drawing graphs for frequency tables but not for grouped frequency tables.

For finding the interquartile range from a cumulative frequency graph see page 62. For finding the interquartile range from a box and whisker graph see page 71.

Exercise 2L

1 For each set of data calculate

i the range **ii** the interquartile range.

a 6 3 8 5 2 9 11 21 15 8

b 5 3 6 8 9 12 10 9 8 13 16 12 9 11 8

c

Price of main course in euros	Frequency
18	6
19	4
20	5
21	8
22	3
23	2
24	5
25	4

Standard deviation

The **standard deviation** is a measure of dispersion that gives an idea of how the data values are related to the mean.

When is the standard deviation of a set of data small? Can the standard deviation equal zero?

Why do we take the square root to find the standard deviation?

Why is the standard deviation sometimes called the root-mean-square deviation?

You are expected to use a GDC to calculate standard deviations.

Example 18

Find the mean and standard deviation of this data set.
4 5 6 8 12 13 2 5 6 9 10 9 8 3 5

Answer

Mean = 7

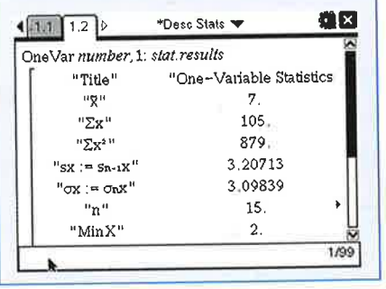
Standard deviation = 3.10 (to 3 sf)

Using a GDC:

Enter the data.

Mean is indicated by $\bar{x}$.

Standard deviation is indicated by σ_x.



GDC help on CD: *Alternative demonstrations for the TI-84 Plus and Casio FX-9860GII GDCs are on the CD.*

Example 19

50 students were asked the total number of points that they received on their IB Diploma. The results are shown in the table.

Score on IB Diploma	Boys	Girls
31	0	3
32	2	4
33	6	3
34	11	5
35	4	3
36	1	2
37	0	1
38	1	2
39	0	2

Use your GDC to calculate the mean and standard deviation for the boys and girls separately and comment on your answer.

Answer

Boys' mean = 34

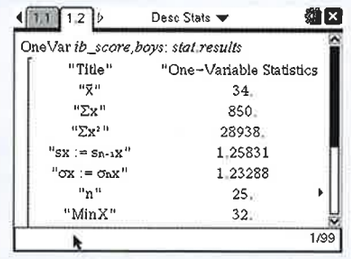
Boys' standard deviation = 1.23 (to 3 sf)

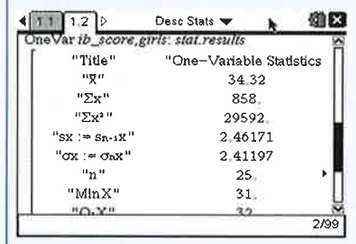
Girls' mean = 34.3 (to 3 sf)

Girls' standard deviation = 2.41 (to 3 sf)

Both the boys and the girls have a mean of about 34 points. The standard deviation for the boys is small, which implies that most boys achieved close to 34 points. However, the standard deviation for the girls is larger which implies that some girls will have much less than 34 points and some will have much more.

Using a GDC:





To make a comment, compare the mean to the corresponding standard deviation.

Is standard deviation a mathematical discovery or an invention?

It is often impossible to find the mean and standard deviation for a whole population. This could be due to time restrictions, financial constraints or other reasons.

If we have, say, a random sample of 12 babies' heights from the UK, then the standard deviation of those 12 babies' heights is given as 'σ_x' on a GDC. This is the one we use for Mathematical Studies.

If we wanted to estimate the standard deviation of all the babies' heights in the UK, based on our random sample, then we would use 's_x' on the GDC.

The IB notation for standard deviation is s_n. When you use your GDC, choose σ_x

Exercise 2M

1 For each set of data calculate the standard deviation.

a 5 3 6 8 9 12 10 9 8 13 16 12 9 11 8

b

Price of main course in euros	Frequency
18	6
19	4
20	5
21	8
22	3
23	2
24	5
25	4

2 Calculate the mean and standard deviation for these data.

6 3 8 5 2 9 11 21 15 8

3 An experiment was performed 50 times. The scores from the experiment were recorded in the table.

a Write down the range.

b Find the interquartile range.

c Find the mean and standard deviation.

Score	Frequency
1	4
2	12
3	11
4	15
5	6
6	2

4 A boat club hosts an annual race. The numbers of people in each boat are recorded in the table.

a Write down the range.

b Find the interquartile range.

c Find the mean and standard deviation.

Number of people	Frequency
4	2
5	7
6	25
7	15
8	30
9	16
10	5

5 The numbers of telephone calls to a call center were monitored every hour for a month. The data collected are shown in the table.

Use your GDC to find

a the mean number of calls per hour

b the standard deviation

c the range

d the interquartile range.

Number of calls per hour	Frequency
60	18
62	45
64	40
66	55
68	31
70	32
72	15
74	13
76	14
78	16

EXAM-STYLE QUESTIONS

6 The mean of these numbers is 33.

16 41 24 x 62 18 25

a Find the value of x.

b Calculate the standard deviation.

c Find the range.

d Find the interquartile range.

7 80 plants were measured and their heights (correct to the nearest cm) recorded in the table.

a Write down the value of m.

b Find the mean height.

c Find the standard deviation of the heights.

d Find the interquartile range of the heights.

Height (cm)	Frequency
10	7
11	m
12	21
13	22
14	11
15	7
16	3

8 The 60 IBDP students at Golden Globe Academy complete a questionnaire about the number of pairs of shoes that they own. The results are shown in the table.

a Find the range and interquartile range.

b Find the mean and standard deviation.

Pairs of shoes	Frequency
5	6
6	8
7	15
8	10
9	5
10	12
11	1
12	3

EXAM-STYLE QUESTIONS

9 The times taken for 50 students to complete a crossword puzzle are shown in the table.

Time (m minutes)	Frequency
$15 \le m < 20$	3
$20 \le m < 25$	7
$25 \le m < 30$	10
$30 \le m < 35$	11
$35 \le m < 40$	12
$40 \le m < 45$	5
$45 \le m < 50$	2

Use the midpoint of each class to estimate the mean and the standard deviation of grouped data.

Find an approximation for the mean and standard deviation.

10 The percentage marks obtained for an ITGS (Information Technology for a Global Society) test by the 25 boys and 25 girls at Bright High are shown in the table.

a Calculate an estimated value for the mean and standard deviation for the girls and the boys separately.

b Comment on your findings.

Girls' frequency	Percentage mark	Boys' frequency
0	$0 \le x < 10$	2
0	$10 \le x < 20$	1
0	$20 \le x < 30$	1
3	$30 \le x < 40$	1
5	$40 \le x < 50$	5
7	$50 \le x < 60$	9
8	$60 \le x < 70$	2
2	$70 \le x < 80$	0
0	$80 \le x < 90$	2
0	$90 \le x < 100$	2

Review exercise

Paper 1 style questions

EXAM-STYLE QUESTIONS

1 The mean of the twelve numbers listed is 6.
3 4 a 8 3 5 9 5 8 6 7 5

a Find the value of a.

b Find the median of these numbers.

2 The mean of the ten numbers listed is 5.
4 3 a 6 8 4 6 6 7 5

a Find the value of a.

b Find the median of these numbers.

EXAM-STYLE QUESTIONS

3 For the set of numbers
3 4 1 7 6 2 9 11 13 6 8 10 6

a calculate the mean

b find the mode

c find the median.

4 The lengths of nine snakes, in metres, are:
6.5 4.6 7.2 5.0 2.4 3.9 12.9 10.3 6.1

a i Find the mean length of the snakes.

ii Find the standard deviation of the length of the snakes.

b Find the median length of the snakes.

5 A survey was conducted of the number of bathrooms in 150 randomly chosen houses. The results are shown in the table.

Number of bathrooms	1	2	3	4	5	6
Number of houses	79	31	22	10	5	13

a State whether the data are discrete or continuous.

b Write down the mean number of bathrooms per house.

c Write down the standard deviation of the number of bathrooms per house.

6 The table shows the age distribution of members of a chess club.

Age (years)	Number of members
$20 \le x < 30$	15
$30 \le x < 40$	23
$40 \le x < 50$	34
$50 \le x < 60$	42
$60 \le x < 70$	13

a Calculate an estimate of the mean age.

b Draw a histogram to represent these data.

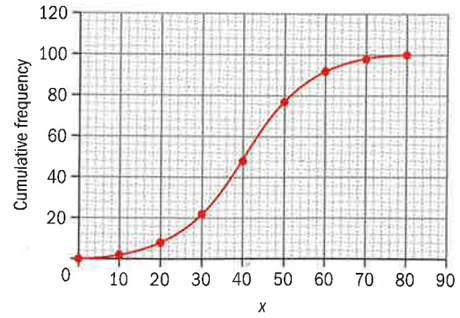
7 Using the cumulative frequency graph, write down the value of

a the median

b the lower quartile

c the upper quartile

d the interquartile range.



EXAM-STYLE QUESTION

8 The numbers of horses counted in 35 fields are represented in the table.
Draw a box and whisker graph to represent this information.

Number of horses	Frequency
8	4
10	9
12	7
15	12
21	3

Paper 2 style questions

EXAM-STYLE QUESTIONS

1 Nineteen students carried out an experiment to measure gravitational acceleration in cm s^{-2}.
The results are given to the nearest whole number.

96 97 101 99 100 98 99 94 96 100
97 98 101 98 99 96 96 100 97

a Use these results to find an estimate for
i the mean value for the acceleration
ii the modal value for the acceleration.
b **i** Construct a frequency table for the results.
ii Use the table to find the median value and the interquartile range.

2 A gardener wanted to estimate the number of weeds on the sports field.
He selected at random 100 sample spots, each of area 100 cm^2, and counted the number of weeds in each spot.

The table shows the results of his survey.

Number of weeds	Frequency
0–4	18
5–9	25
10–14	32
15–19	14
20–24	7
25–29	4

a **i** Construct a cumulative frequency table and use it to draw the cumulative frequency curve.
ii Write down the median number of weeds.
iii Find the percentage of spots that have more than 19 weeds.
b **i** Estimate the mean number of weeds per spot.
ii Estimate the standard deviation of the number of weeds per spot.

The area of the field is 8000 m^2.
iii Estimate the total number of weeds on the field.

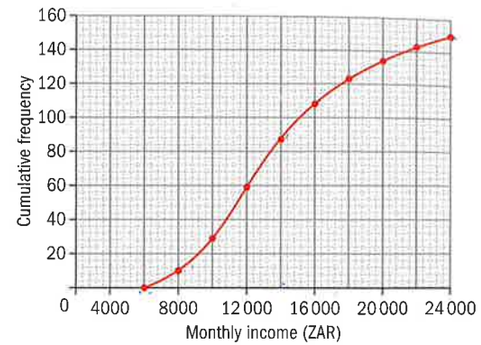
3 The marks for a test are given in the frequency table.
a Complete a cumulative frequency table and use it to draw the cumulative frequency curve.
b Find the median mark.
c Find the interquartile range.

60% of the candidates passed the examination.
d Find the pass mark.
e Given that the lowest mark was 9 and the highest was 98, draw a box and whisker graph to represent the information.

Mark, x	Frequency
$0 \le x < 10$	3
$10 \le x < 20$	14
$20 \le x < 30$	21
$30 \le x < 40$	35
$40 \le x < 50$	42
$50 \le x < 60$	55
$60 \le x < 70$	43
$70 \le x < 80$	32
$80 \le x < 90$	15
$90 \le x < 100$	10

EXAM-STYLE QUESTIONS

4 The cumulative frequency graph shows the monthly incomes, in South African Rand, ZAR, of 150 people.



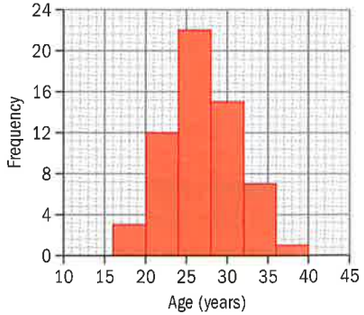
a Write down the median and find the interquartile range.
b Given that the lowest monthly income is 6000 ZAR and the highest is 23 500 ZAR, draw a box and whisker graph to represent this information.
c Draw a frequency table for the monthly incomes.
d Use your GDC to find an estimate of the mean and standard deviation of the monthly incomes.

5 The weights of 200 female athletes are recorded in the table.

Weight (w kg)	Frequency
$45 \le w < 50$	4
$50 \le w < 55$	16
$55 \le w < 60$	45
$60 \le w < 65$	58
$65 \le w < 70$	43
$70 \le w < 75$	28
$75 \le w < 80$	6

a Write down the modal group.
b Calculate an estimate of the mean and the standard deviation.
c Construct a cumulative frequency table and use it to draw the cumulative frequency graph.
d Write down the median, the lower quartile and the upper quartile.
e The lowest weight is 47 kg and the heaviest is 76 kg. Use this information to draw a box and whisker graph.

6 A group of 60 women were asked at what age they had their first child. The information is shown in the histogram.



a Calculate an approximation for the mean and standard deviation.
b Write down the modal class.
c Construct a cumulative frequency table for the data and draw the cumulative frequency curve.
d Use your graph to find the median and interquartile range.
e Given that the youngest age was 16 and the oldest was 39, draw a box and whisker graph to represent the information.

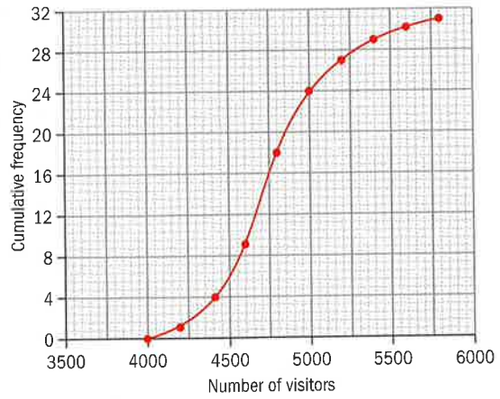
7 The average times, to the nearest second, that 100 participants waited for an elevator are shown in the table.

Time (t seconds)	Frequency
$0 \le t < 10$	5
$10 \le t < 20$	19
$20 \le t < 30$	18
$30 \le t < 40$	22
$40 \le t < 50$	16
$50 \le t < 60$	12
$60 \le t < 70$	8

a Write down the modal class.
b Calculate an estimate of the mean time and the standard deviation.
c Construct a cumulative frequency table and use it to draw the cumulative frequency graph.
d Write down the median and interquartile range.

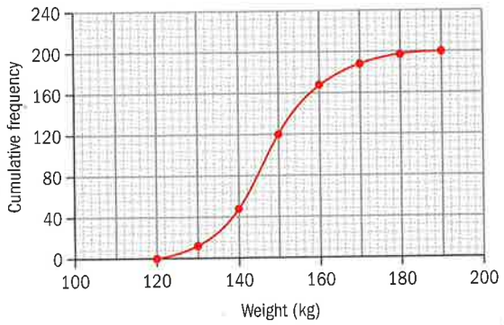
EXAM-STYLE QUESTIONS

8 The cumulative frequency graph shows the daily number of visitors to the Mausoleum on Tiananmen Square in the month of January.



- **a** Write down the median, the lower quartile and the upper quartile.
- **b** Given that the least number of visitors was 4000 and the most was 5700, draw a box and whisker graph to represent the information.
- **c** Construct a frequency table for this information.
- **d** Write down the modal class.
- **e** Calculate an estimate of the mean and the standard deviation.

9 The cumulative frequency graph shows the weights, in kg, of 200 professional wrestlers.



- **a** Construct a grouped frequency table for this information.
- **b** Write down the modal class.
- **c** Calculate an estimate of the mean weight.

CHAPTER 2 SUMMARY

Classification of data

- **Discrete data** are either data that can be counted or data that can only take specific values.
- **Continuous data** can be measured. They can take any value within a range.

Grouped discrete or continuous data

- To draw a **frequency histogram**, find the lower and upper boundaries of the classes and draw the bar between these boundaries. There should be no spaces between the bars.

Measures of central tendency

- The **mode** of a data set is the value that occurs most frequently.
- The **median** of a data set is the value that lies in the middle when the data are arranged in size order.
- The **mean** of a data set is the sum of all the values divided by the number of values.
- For data in a frequency table, the **mode** is the entry that has the largest frequency.

Continued on next page

- The **median** is the middle entry as the entries in the table are already in order. For n pieces of data, the median is the $\frac{n+1}{2}$th value.
- The **mean** from a frequency table is:

$$\text{mean} = \frac{\text{total of } f_i \times x_i}{\text{total frequency}}$$

where f_i is the frequency of each data value x_i and $i = 1, \ldots, k$, where k is the number of data values.

- For grouped data, the **modal class** is the group or class interval that has the largest frequency.
- To calculate the **mean** from a grouped frequency table, an estimate of the mean is

$$\frac{\text{total of } f_i \times x_i}{\text{total frequency}}$$

where f_i is the frequency and x_i is the corresponding midpoint of each class.

Cumulative frequency curves

- The **cumulative frequency** is the sum of all of the frequencies up to and including the new value. To draw a **cumulative frequency curve** you need to construct a cumulative frequency table, with the upper boundary of each class interval in one column and the corresponding cumulative frequency in another. Then plot the upper class boundary on the x-axis and the cumulative frequency on the y-axis.
- To find the **lower quartile**, Q_1, read the value on the curve corresponding to $\frac{n+1}{4}$ on the cumulative frequency axis, where n is the total frequency.
- To find the median, read the value on the curve corresponding to $\frac{n+1}{2}$ on the cumulative frequency axis.
- To find the **upper quartile**, Q_3, read the value on the curve corresponding to $\frac{3(n+1)}{4}$ on the cumulative frequency axis.
- To find the **percentiles**, p%, read the value on the curve corresponding to $\frac{p(n+1)}{100}$ on the cumulative frequency axis.
- To find the **interquartile range** subtract the lower quartile from the upper quartile: $\text{IQR} = Q_3 - Q_1$.

Box and whisker graphs

- To draw a box and whisker graph, five pieces of information are needed: calculate the lower quartile, median and upper quartile for the data. Find the smallest and largest values.

Measures of dispersion

- The **range** is found by subtracting the smallest value from the largest value.
- The **interquartile range** is found by subtracting the lower quartile, Q_1, from the upper quartile, Q_3: $\text{IQR} = Q_3 - Q_1$.
- The standard deviation is often referred to as the 'root-mean-square deviation' because we find the **deviation** of each entry from the mean, then we **square** these values and find the **mean** of the squared values, and, finally, we take the square **root** of this answer.

Statistically speaking

Descriptive statistics describe the basic features of a data set.

Descriptive statistics reduce lists of data into a simple summary such as a single average (a number) or a visual form such as a graph or diagram.

Morals and statistics

Case study 1

A company has 3 employees and a boss.

The employees earn 2500 euros a month and the boss earns 25 000 euros a month.

A report in the local newspaper states that the average salary in the company is 8125 euros a month.

- Which average has the newspaper used?
- Does this average give a fair representation of the average salary?
- Which would be the most appropriate average to use? Why?

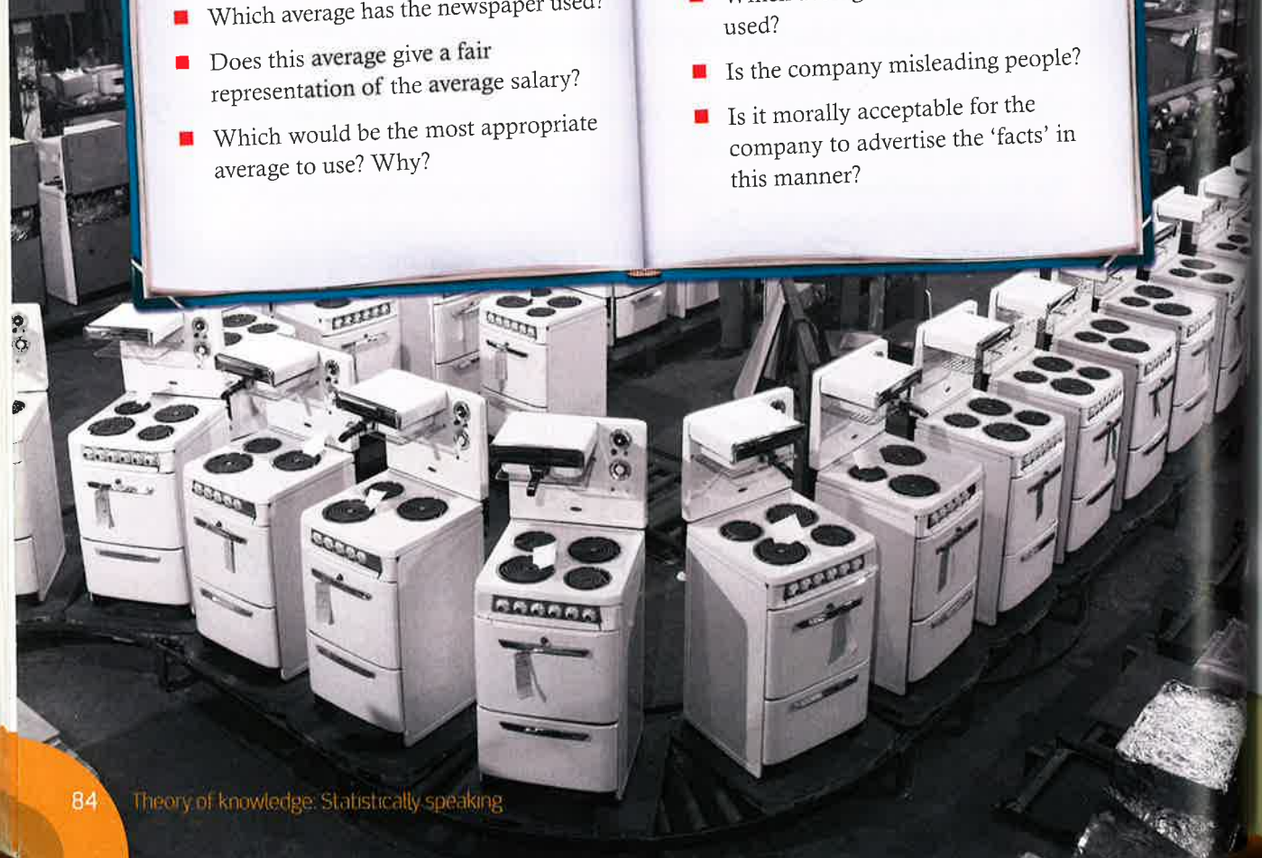
Case study 2

Ten appliances were tested and the number of faults each one had recorded below.

0 0 0 0 0 0 15 19 25 31

The company advertises that the average number of faults is 0.

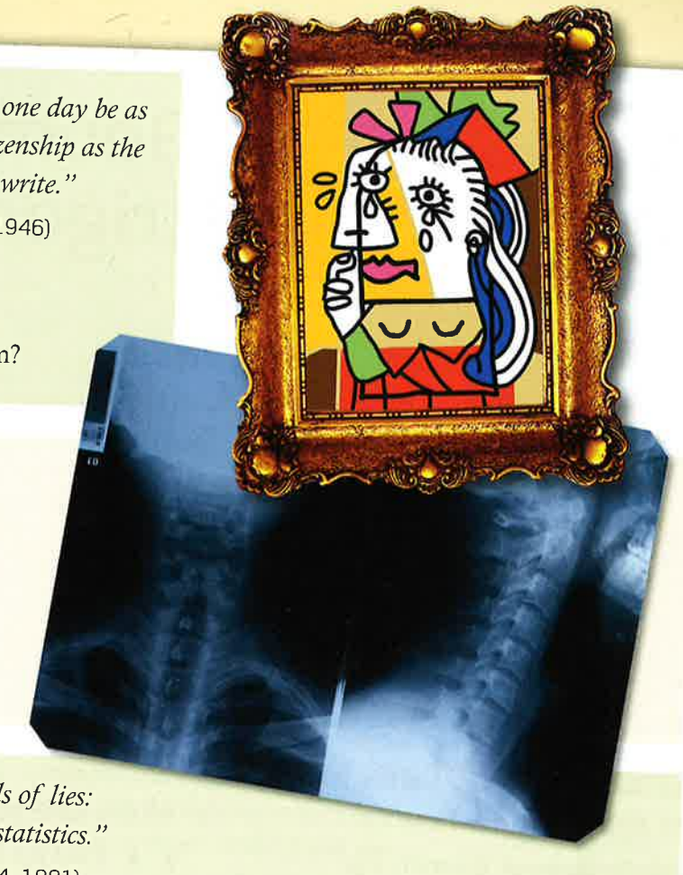
- Which average has the company used?
- Is the company misleading people?
- Is it morally acceptable for the company to advertise the 'facts' in this manner?



"Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."

H. G. Wells (1866–1946)

- What do you think H. G. Wells meant?
- Do you agree with him?



- How accurate are these visual representations:
  - X-rays
  - Snapshots
  - Paintings?

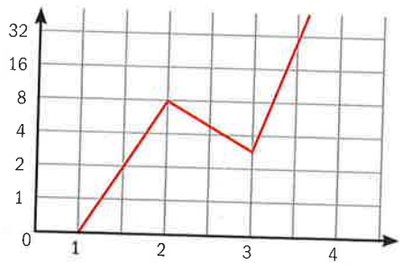
"There are three kinds of lies: lies, damned lies, and statistics."

Benjamin Disraeli (1084–1881)
Popularized by
Mark Twain (1835–1910)

- Do statistics 'lie'?
- Are all statistics 'accurate'?

Misleading graphs

- What is wrong with this graph?



- What is wrong with this 3D histogram?