

Statistics 1 for Edexcel contents

1 Organising data 6

- A Organising data by grouping 6
stem and leaf diagram, discrete and continuous variables, histogram
- B Organising data by ordering 11
median, range, quartiles, interquartile range, box plot, outliers
- C Linear interpolation 16
for median and quartiles from grouped data
- D Large data sets: percentiles 18
cumulative frequency graph

2 Summarising data 22

- A Measures of average 22
mode, median, mean, coding to calculate mean
- B Measures of spread 25
variance, standard deviation
- C Scaling and coding 29
effect on standard deviation
- D Working with frequency distributions 34
mean and standard deviation from grouped data, coding
- E Skewness 37
- F Choosing and using measures of average and spread 38
Mixed questions 40

3 Probability 43

- A Modelling 43
probability experiment, event, equally likely outcomes, favourable outcomes, sample space
- B Outcomes and events 45
set notation, Venn diagram, complementary and mutually exclusive events, sum law
- C Conditional probability 49
- D Independent events 51
conditional probability, product law
- E Tree diagrams 53
conditional probability 'in reverse'
Mixed questions 57

4 Linear regression 59

- A The least squares regression line 59
explanatory and response variables, unreliability outside the observed range
- B Explanatory variables 64
identifying response and explanatory variable, interpretation of intercept and gradient of regression line
- C Coding 66

5 Correlation 70

- A Measuring correlation 70
positive and negative correlation, product moment correlation coefficient
- B Scaling and coding 74
- C Interpreting correlation 76
spurious correlation, effect of outliers, non-linear relationships
- D Correlation and regression 78
required strength of correlation for given sample size

6 Discrete random variables 82

- A Probability functions 82
probability distribution, discrete uniform distribution
- B Formulae for probability functions 84
finding values from formula, conditions for a function to be a probability function, modelling
- C Cumulative distribution function 87
- D Mean, variance and standard deviation 90
expected value
- E Functions of a discrete random variable 94
expected value, variance
Mixed questions 97

7 Normal distribution 101

A Proportions 101

between a given number of standard deviations
above or below the mean, shape, symmetry

B The standard normal probability distribution 104

continuous random variable, area to represent
probability, standard normal distribution, use of
table

C Finding probabilities and proportions 108

scaling of data to standard normal variable

D Working backwards 111

percentage points table

Mixed questions 115

8 Modelling 117

review of modelling ideas

Excel functions 119

Tables 120

The normal distribution function 120

Percentage points of the normal distribution 121

Answers 122

Index 141

1 Organising data

In this chapter you will learn how to

- represent data using stem-and-leaf diagrams, histograms and box plots
- find and use the median and quartiles of a set of data
- use interpolation to estimate values

A Organising data by grouping (answers p 122)

People collect data for a purpose – to help answer questions about the real world. They have to decide what data they need to collect and how to process it so that it provides answers to their questions.

For example, the people living in a village want to know whether putting up a sign saying ‘Please drive carefully through our village’ will make any difference to drivers’ speeds. They employ an investigator to measure the speeds of cars as they pass through the village, before and after a sign is put up.

The investigator has to be careful about the choice of the times when she will record the speeds. For example, if she records speeds on a bright day before the sign goes up and on a dark day after, then she will not be able to tell whether any difference in speed is caused by the sign or by the weather. So she tries as far as possible to carry out both sets of measurements under the same conditions.

She decides to record speeds of cars during the same one-hour period on two similar days. Here are the results, in miles per hour (m.p.h.).

Before	After
32 25 54 61 24 30 48 58 31 40 71 62	51 22 18 30 35 64 56 32 43 50 62 73
42 20 45 36 19 26 68 56 39 45 47 63	44 11 23 45 47 39 25 65 53 51 67 32
25 48 62 55 38 56 42 55 52	30 40 28 32 47 29

This is the **raw data** – the data as collected. It is not easy to compare speeds from these lists.

A simple method of organising data is to use a **stem-and-leaf diagram**.

This can be drawn for a single set of data or, as in this case, back-to-back.

In the ‘after’ data, $|6|2$ means 62. In each row the units digits are written in numerical order.

In the ‘before’ data, $2|5|$ means 25. In each row the units digits are written in numerical order, starting from the right.

Before	After
	9 1 1 8
6 5 5 4 0	2 2 3 5 8 9
9 8 6 2 1 0	3 0 0 2 2 2 5 9
8 8 7 5 5 2 2 0	4 0 3 4 5 7 7
8 6 6 5 5 4 2	5 0 1 1 3 6
8 3 2 2 1	6 2 4 5 7
	7 3

- A1** Do you think speeds are, on the whole, greater or less after the sign is put up? What feature of the diagram leads you to your answer?

Key	
2 5	means 25
6 2	means 62

When you make a stem-and-leaf diagram, you do it in two stages. First you put each 'leaf' (usually the units digit) into the appropriate row. Then you redo the diagram with the leaves in numerical order.

A2 Here are the marks (out of 100) of a group of students in two maths papers, paper 1 and paper 2.

Paper 1

42 56 33 40 54 62 26 33 39 45
48 65 50 21 37 42 53 44 60 56
47 24 39 68 35

Paper 2

41 34 37 65 54 42 40 23 52 55
63 67 59 30 25 55 45 71 37 41
64 57 48 65 54

- (a) Make a back-to-back stem-and-leaf diagram for the two sets of marks.
- (b) Which paper was harder? How does the diagram show this?

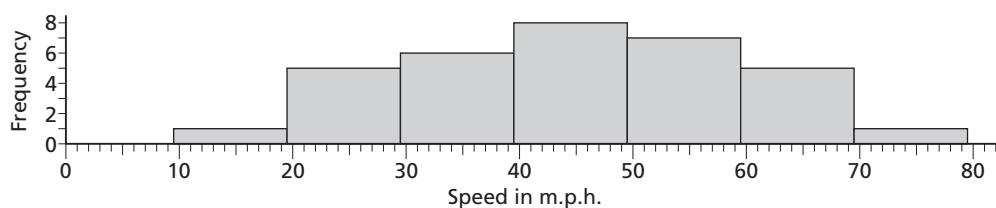
Numerical data is of two types, **discrete** and **continuous**. In discrete data the possible values are separated by gaps, for example 0, 1, 2, 3, ... or shoe sizes 5, $5\frac{1}{2}$, 6, $6\frac{1}{2}$, ... Continuous data comes from measurement; for example, the speed of a car could be 42.34657... m.p.h. In practice, measurements can only be recorded to a certain degree of accuracy and the data will appear to be discrete, as in the car speeds data on the opposite page. A speed recorded as 37 m.p.h. could be anything between 36.5 and 37.5.

Grouping is a common way of organising data. In the stem-and-leaf diagram for the car speeds, the data is organised into the groups 10–19, 20–29, ...

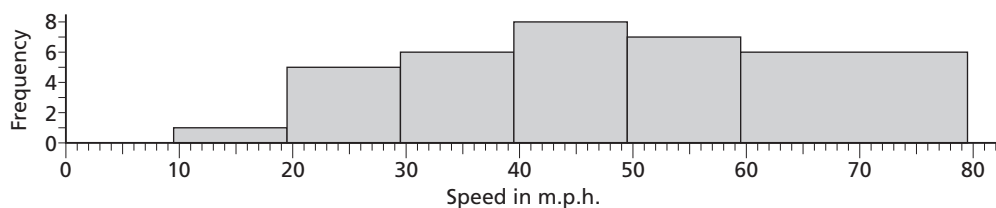
Because the speeds were rounded, the 10–19 group contains all speeds in the interval 9.5–19.5, and so on. The table below shows the frequency for each interval in the 'before' data set. There is now no gap between each interval and the next.

Speed (m.p.h.)	9.5–19.5	19.5–29.5	29.5–39.5	39.5–49.5	49.5–59.5	59.5–69.5	69.5–79.5
Frequency	1	5	6	8	7	5	1

The data can be shown in a frequency bar chart, using the intervals in the table.



Suppose that it is suggested that the last two groups should be combined into a single group 59.5–79.5 with a total frequency of 6. The chart would look like this.



A3 What is wrong with this chart?

The second chart on the previous page is misleading because the eye is drawn to the area rather than the height of the bars.

A chart in which area, not height, shows frequency is called a **histogram**.

The vertical scale on a histogram shows **frequency density** = $\frac{\text{frequency}}{\text{width of interval}}$.

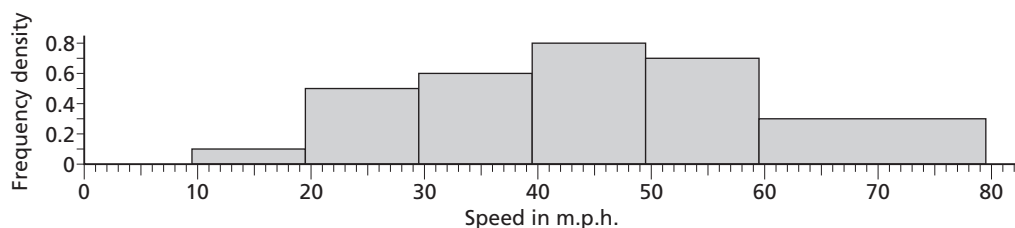
Here is the table for the 'before' car speeds data with the last two groups combined, showing the frequency density for each interval.

Speed (m.p.h.)	9.5–19.5	19.5–29.5	29.5–39.5	39.5–49.5	49.5–59.5	59.5–79.5
Frequency	1	5	6	8	7	6
Frequency density	0.1	0.5	0.6	0.8	0.7	0.3

width of interval = 10
frequency density = $\frac{1}{10} = 0.1$

width of interval = 20
frequency density = $\frac{6}{20} = 0.3$

The histogram is shown here.



The frequency for each interval is found by multiplying the width of the interval by the frequency density.

For example, the frequency for the interval 59.5–79.5 = width of interval \times frequency density
= $20 \times 0.3 = 6$

K In a histogram, area represents frequency.

The vertical scale shows frequency density = $\frac{\text{frequency}}{\text{width of interval}}$

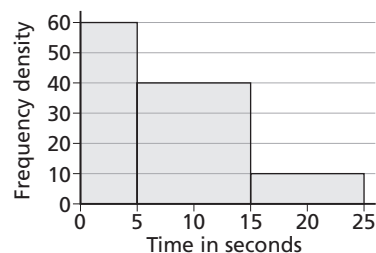
The frequency for an interval = width of interval \times frequency density

A histogram shows the distribution of the data. The histogram above shows that the speeds are concentrated towards the middle of the range, with fewer at each end.

Example 1

This histogram shows the distribution of the lengths of phone calls made from a telesales office one evening.

- (a) How many calls were made that lasted
- (i) up to 5 seconds (ii) up to 15 seconds
- (b) Estimate the number that lasted more than 20 seconds.



Solution

- (a) Area from 0 to 5 = $5 \times 60 = 300$ calls
(b) Area from 0 to 15 = $300 + 10 \times 40 = 700$ calls
(c) Assume that the calls in the 15–25 interval are evenly spread.
Area from 20 to 25 = $5 \times 10 = 50$ calls

Example 2

This is a grouped frequency table for the weights, to the nearest 10 g, of parcels sent from an office one day.

Draw a histogram to show the data.

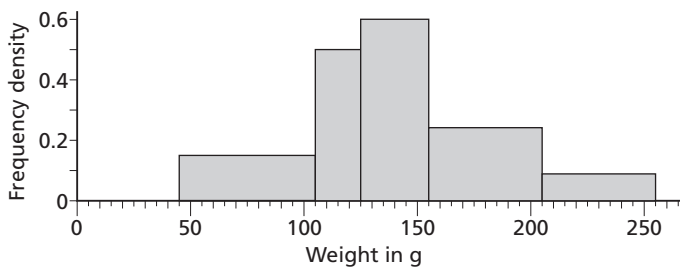
Weight (g)	Frequency
50–100	9
110–120	10
130–150	18
160–200	12
210–250	4

Solution

The weights have been rounded to the nearest 10 g. So first replace each grouping by a continuous interval. Then, using the lengths of these intervals, calculate each frequency density.

The histogram is shown below.

Weight (g)	Frequency	Frequency density
45–105	9	$\frac{9}{60} = 0.15$
105–125	10	$\frac{10}{20} = 0.5$
125–155	18	$\frac{18}{30} = 0.6$
155–205	12	$\frac{12}{50} = 0.24$
205–255	4	$\frac{4}{50} = 0.08$



Exercise A (answers p 122)

- 1 A class of students is given a test. The students' marks are shown below.

66 43 46 54 74 37 66 59 40
65 55 57 32 45 59 52 67 83
72 43 84 54 48 78 63

- (a) Draw a stem-and-leaf diagram for these marks.
(b) What percentage of the students scored fewer than 50 marks in the test?
(c) When a similar test was given to last year's students, the top 20% of students scored 76 marks or more.
What is the corresponding mark for the top 20% this year?

- 2 As part of a selection process, 103 employees of a large company were given a task. The times taken to complete the task were recorded to the nearest minute and the following grouped frequency table was made.

Time (minutes)	Frequency
5–9	18
10–14	28
15–24	24
25–39	24
40–49	9

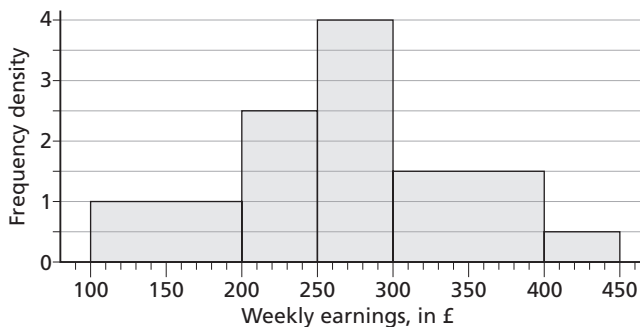
Draw a histogram to show this data.

- 3 A geography student carries out a survey of house plot sizes in an area of the country. She records the areas of the plots to the nearest 0.1 hectare and produces this grouped frequency table of her results.

Area (hectares)	Frequency
1.0–1.9	15
2.0–2.9	22
3.0–4.9	25
5.0–7.9	16
8.0–9.9	12

Draw a histogram to show this data.

- 4 This histogram gives information about the weekly earnings of the employees of a company.



- (a) How many employees earn between £250 and £300 a week?
 (b) How many employees earn less than £250 a week?
 (c) Estimate the number of employees who earn between £175 and £275 a week.
 (d) What percentage of the employees earn £400 or more a week?

B Organising data by ordering (answers p 122)

A simple way of organising a set of data is to write it in order, lowest to highest. Here is the 'before' set of car speeds data from section A, written in order.

19 20 24 25 25 26 30 31 32 36 38 39 40 42 42 45 (45) 47 48 48 52 54 55 55 56 56 58 61 62 62 63 68 71

The middle value in this list (ringed) is called the **median** speed. It can be used as an 'average' of the data.

There is an even number of values in the 'after' set of speeds, so there is no middle value. However, there is a middle pair of values, and the median is taken to be halfway between these two.

11 18 22 23 25 28 29 30 30 32 32 32 35 39 (40|43) 44 45 47 47 50 51 51 53 56 62 64 65 67 73
Median = 41.5

The fact that the first median is greater than the second suggests that on the whole the road sign has caused a reduction in speed.

Another feature of interest is the 'spread' of the data. The simplest way to measure spread is by the **range**, defined as the difference between the highest and lowest values.

The range of the 'before' data is $71 - 19 = 52$. The range of the 'after' data is $73 - 11 = 62$. This suggests that the 'after' speeds are much more spread out than the 'before' speeds.

However, data sets often contain individual extreme values at the lower or upper ends that can distort the overall picture. For this reason the **quartiles** are defined.

- K** The lower, or **first quartile**, Q_1 , has $\frac{1}{4}$ of the data values less than or equal to it.
The upper, or **third quartile**, Q_3 , has $\frac{3}{4}$ of the data values less than or equal to it.
The median itself is the **second quartile**, Q_2 .

When the number of data values is a multiple of 4, for example 12, then Q_1 and Q_3 are found like this.

Find $\frac{1}{4}$ of $12 = 3$. Q_1 is halfway between the 3rd and 4th values.

Find $\frac{3}{4}$ of $12 = 9$. Q_3 is halfway between the 9th and 10th values.

1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th	12th
-----	-----	-----	-----	-----	-----	-----	-----	-----	------	------	------

When the number of data values is not a multiple of 4, find $\frac{1}{4}$ and $\frac{3}{4}$ of it and round up to the next positions. For example, with 33 data values

$\frac{1}{4}$ of $33 = 8\frac{1}{4}$, so Q_1 is the 9th value. $\frac{3}{4}$ of $33 = 24\frac{3}{4}$, so Q_3 is the 25th value.

B1 Find the quartiles of each set of car speeds above.

- K** The difference $Q_3 - Q_1$, is called the **interquartile range**. It tells us the spread of the 'middle half' of the data. As this excludes extreme values, it is a better measure of spread than the simple range.

For the data sets above, the interquartile ranges are

before: $56 - 32 = 24$ after: $51 - 30 = 21$

On the basis of the interquartile range, the 'before' data set is more spread out.

B2 Find the median, lower quartile and upper quartile for each of these ordered data sets.

- (a) 27 33 37 41 43 44 45 47 48 48 49 52 52 54 58 62 69 73 75 79
- (b) 35 39 42 47 52 53 54 57 59 60 62 64 66 72 78 83 86
- (c) 16 18 20 21 22 25 27 30 33 33 34 37 41 42 44 46 47 48 49 52 55 55 58
- (d) 24 25 29 30 31 32 34 36 38 39 39 42 43 44 47 49 52 54

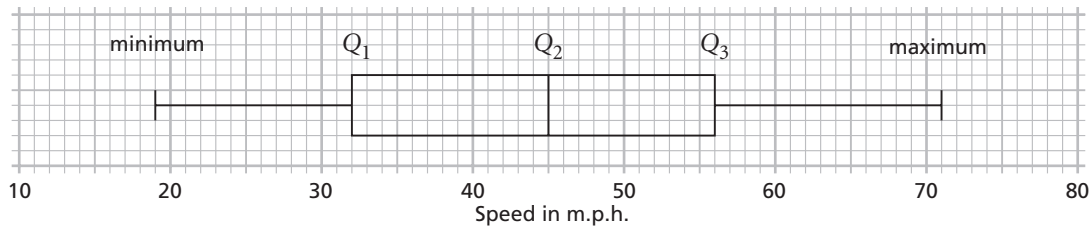
B3 This stem-and-leaf diagram gives the marks of some students in an exam. Find

- (a) Q_2
- (b) Q_1
- (c) Q_3
- (d) the interquartile range

2	2 4 5	(3)
3	0 1 3 5 7	(5)
4	1 4 4 6 6 8 9	(7)
5	2 3 4 6 6 8 8 9	(8)
6	1 2 4 7	(4)
7	0 3	(2)

2 | 4 means 24

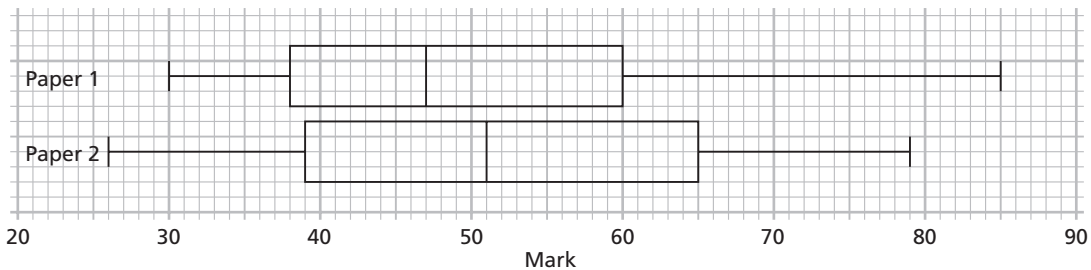
The median and quartiles can be shown in a **box plot**. The box plot for the 'before' car speeds data is shown below. The ends of the box show Q_1 and Q_3 , the line inside the box shows the median Q_2 , and the ends of the two 'whiskers' show the lowest and highest values in the data.



B4 How is the interquartile range shown in the box plot?

B5 Draw a box plot for the 'after' car speeds data.

D B6 The two box plots below show the distributions of marks in two exams.



Write a couple of sentences comparing the two sets of marks.

Outliers

An **outlier** in a data set is an exceptionally high or low value. The word ‘exceptionally’ is vague, so rules have been developed to identify outliers.

One common rule is ‘Tukey’s rule’, named after the American statistician John Tukey (1915–2000). This rule identifies an outlier as any value which is more than $(1.5 \times \text{interquartile range})$ beyond the lower or upper quartile.

For example, if $Q_1 = 45$ and $Q_3 = 67$, the interquartile range is $67 - 45 = 22$. An outlier would be any value less than $45 - 1.5 \times 22$, which is 12, or any value greater than $67 + 1.5 \times 22$, which is 100.

These values 12 and 100, beyond which values become outliers, are sometimes called lower and upper ‘fences’.

(There are other versions of the rule in which the multiplier 1.5 is varied.)

When a data set includes one or more outliers, the box plot is modified.

The outliers are shown as isolated dots or crosses and the whiskers are drawn up to the lowest and highest data values excluding the outliers.

Example 3

A student measures the pulse rate, in beats per minute, of 50 students at his school and lists them in rank order as follows.

34 52 61 62 62 63 64 64 65 65 66 68 68 70 72
73 75 77 77 78 78 79 79 80 80 80 81 81 83 83
84 84 84 85 86 87 89 89 90 90 91 91 93 94 95
99 101 108 122 123

The median of this data is 80 and the lower and upper quartiles are 68 and 89.

Identify any outliers using Tukey’s rule and use this information to draw a box plot showing the outliers.

Solution

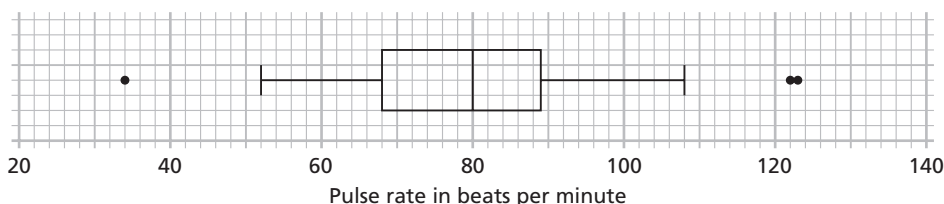
The interquartile range is $89 - 68 = 21$.

Outliers at the lower end are values less than $68 - 1.5 \times 21 = 36.5$.

Outliers at the upper end are values greater than $89 + 1.5 \times 21 = 120.5$.

So 34 at the lower end and 122 and 123 at the upper end are outliers.

Leaving out the outliers, the minimum and maximum values are 52 and 108. This is the box plot for this data.



If, after an outlier has been excluded, there is not enough information to identify the maximum or minimum value of the remaining data, then the whisker is made to end at the 'fence', as in the following example.

Example 4

The prices of flats in a district range from a minimum of £48k to a maximum of £152k. The values of Q_1 , Q_2 and Q_3 are £82k, £92k and £106k.

An outlier is defined as any value below $Q_1 - 1.5 \times (Q_3 - Q_1)$ or any value above $Q_3 + 1.5 \times (Q_3 - Q_1)$.

- (a) Determine whether the minimum and maximum prices are outliers.
- (b) Draw a box plot for the data.

Solution

(a) $Q_1 - 1.5 \times (Q_3 - Q_1) = 82k - 1.5 \times (106k - 82k) = 46k$

So the minimum value, £48k, is not an outlier.

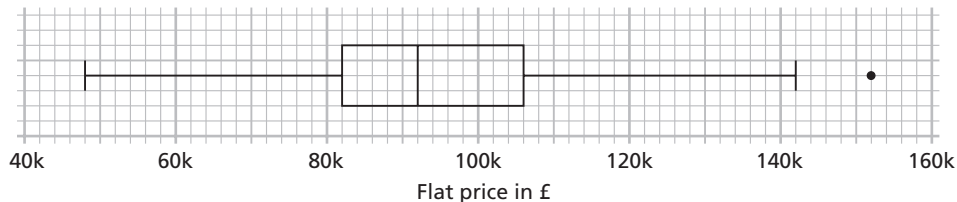
$Q_3 + 1.5 \times (Q_3 - Q_1) = 106k + 1.5 \times (106k - 82k) = 142k$

So the maximum value, £152k, is an outlier.

- (b) As the minimum value, £48k, is not an outlier, the lower whisker extends to £48k.

As the maximum value, £152k, is an outlier, the upper whisker extends to the upper fence, £142k.

Q_2 (the median) = £92k. The box plot is shown here.



Exercise B (answers p 122)

- 1 The weights in kilograms of 14 students are

48.4 47.3 50.4 55.7 52.3 48.2 50.7 44.5 49.2 56.5 52.3 51.5 48.8 55.7

Find the median weight of these students.

- 2 A citizens' aid charity opened 25 drop-in advice centres in a part of the country. The numbers of callers on the first day of opening are recorded here in order of size.

6 10 18 19 23
26 29 30 31 31
33 34 35 36 36
37 38 40 43 44
50 54 56 62 66

Find the values of the quartiles Q_1 , Q_2 and Q_3 and the interquartile range.

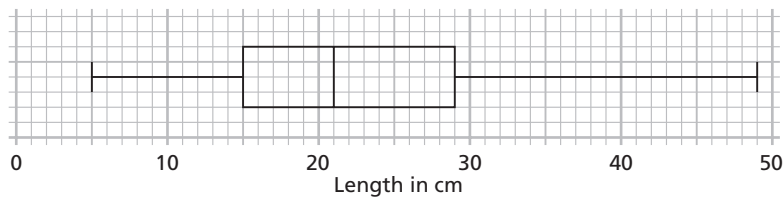
- 3 This stem-and-leaf diagram shows the number of minutes that Albion Railways trains were late arriving at their destination on one day.

0	0 0 0 1 3 4 6 7 7 9
1	2 2 4 5 5 7 8 8
2	1 3 5 6 6 7 9
3	3 4 6 9 9
4	2 3

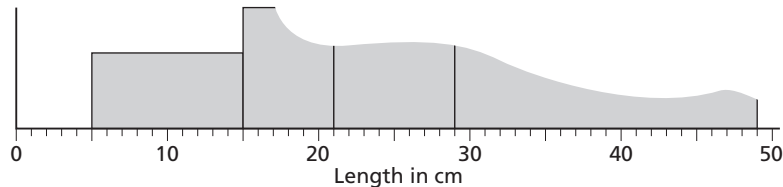
1 | 2 means 12

- (a) Find the values of Q_1 , Q_2 and Q_3 .
 (b) What is the interquartile range?

- 4 A biologist measured the unstretched lengths in cm of 60 earthworms and drew the box plot below to show the results.



A partially drawn histogram of this data is shown below.



Find the frequency density for each of the four intervals.

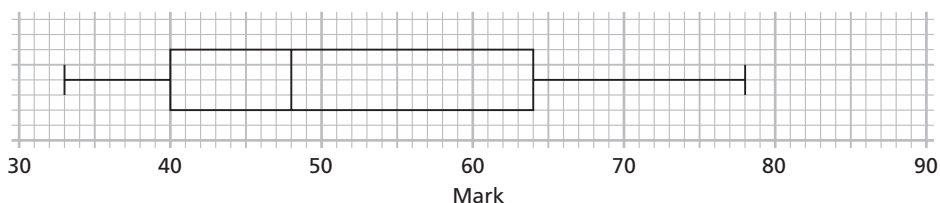
- 5 Some students enter a music exam. Their marks are as shown below.

57 33 54 43 47 52 55 40 77 58
 69 49 50 55 43 35 51 73 58 60
 80 50 37 46 67 85 44 39 47

- (a) Make a stem-and-leaf diagram to show these marks.
 (b) Find the values of Q_1 , Q_2 and Q_3 .
 (c) An outlier is defined as any value x such that $x < Q_1 - 1.5(Q_3 - Q_1)$ or $x > Q_3 + 1.5(Q_3 - Q_1)$.

Determine which of the marks, if any, are outliers.

- (d) Draw a box plot for the students' marks.
 (e) In another school a similar number of students entered for the same exam. A box plot of their marks is shown below.



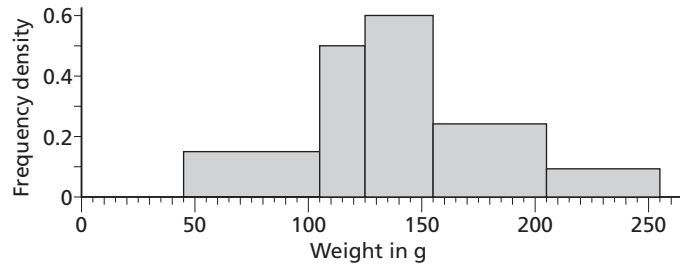
Use the two box plots to compare the performances of the two groups of students.

C Linear interpolation (answers p 123)

Linear interpolation is a method of estimating unknown values that lie between known values.

Here is the grouped frequency table and histogram for the weights of the parcels sent from an office (as in example 2 on page 9).

Weight (g)	Frequency
50–100	9
110–120	10
130–150	18
160–200	12
210–250	4



The total number of parcels is 53, so the median is the weight of the 27th parcel. But because the original data values have been lost in the grouping, it is not possible to identify the median.

However, adding the frequencies from the top of the table, we find that the 27th parcel is in the 130–150 group.

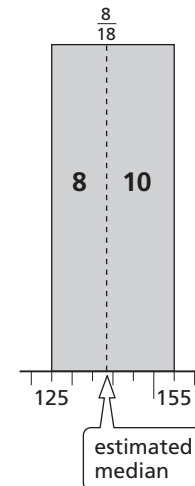
This group of 18 parcels is represented by the area of the third bar in the histogram, which is on the interval 125–155. (The interval is 125–155 rather than 130–150 because the weights of the parcels were rounded.)

The bars to the left of this bar account for 19 parcels.

So the third bar represents parcels numbers 20 up to 37.

Parcel number 27 is the 8th in this group of 18.

If we assume that the parcels are evenly spread through the group, then the 27th parcel is $\frac{8}{18}$ of the way across the bar. (See diagram.)



So we can estimate that the median weight is $\frac{8}{18}$ of the way between 125 g and 155 g, which is $125 + \frac{8}{18} \times 30 = 138$ g (to the nearest gram).

It is not necessary to draw the histogram to use this method, provided that you remember to adjust the intervals to be continuous if the data has been rounded.

C1 This is a frequency table of the lengths, to the nearest minute, of phone calls made from an office one day.

Estimate the median length of a call.

Length (min)	Frequency
0–2	8
3–5	11
6–9	16
10–15	14
16–20	9
> 20	3

Example 5

This is a frequency table for the heights of 240 female students.

Height (cm)	140–145	145–150	150–155	155–160	160–165	165–170	170–175	175–180	180–185
Frequency	3	10	21	54	72	48	25	5	2

Estimate the upper quartile of the heights.

Solution

$$\frac{3}{4} \text{ of } 240 = 180$$

The intervals in this case are already continuous.

Adding the frequencies up from the start of the table shows that the 180th student is in the interval 165–170.

The groups below this account for 160 students, leaving 20 to get to the 180th.

So the 180th student is 20th out of 48 in the interval 165–170.

Assuming equal spacing, the upper quartile is $165 + \frac{20}{48} \times 5 = 167.1$ cm (to 1 d.p.).

Exercise C (answers p 123)

- 1 A researcher in a supermarket measured the time customers spent at a checkout. Each customer was timed from when they joined the queue until when they left the checkout.

A frequency table of the times is given here.

Use linear interpolation to estimate the median time spent at the checkout.

Time (t mins)	Frequency
$0 < t \leq 2$	4
$2 < t \leq 4$	7
$4 < t \leq 6$	12
$6 < t \leq 8$	17
$8 < t \leq 10$	10
$t > 10$	5

- 2 A gardener grew onions in a plot treated with a fertiliser and in another untreated plot.

The fully grown onions were weighed. The table shows the results.

Use linear interpolation to estimate the median weight of each collection of onions and use the results to compare the two collections.

Weight (g)	With fertiliser frequency	Without fertiliser frequency
50–75	6	3
75–100	9	15
100–125	12	18
125–150	21	10
150–175	16	9
175–200	8	4

- 3 Estimate the values of Q_1 , Q_2 and Q_3 for the following data, based on weighing all the animals in a colony to the nearest kilogram.

Weight (kg)	20–24	25–29	30–34	35–39	40–44	45–49	50–54
Frequency	3	10	16	20	10	6	2

D Large data sets: percentiles (answers p 123)

The grouped frequency table below is based on the measurement of 350 sunflower plants six weeks after planting.

The table next to it is a **cumulative frequency** table. Cumulative frequencies are found by adding up the frequencies as you go down the frequency table.

Height (cm)	Frequency
2.5–6.5	10
6.5–10.5	21
10.5–14.5	114
14.5–18.5	105
18.5–22.5	54
22.5–26.5	46

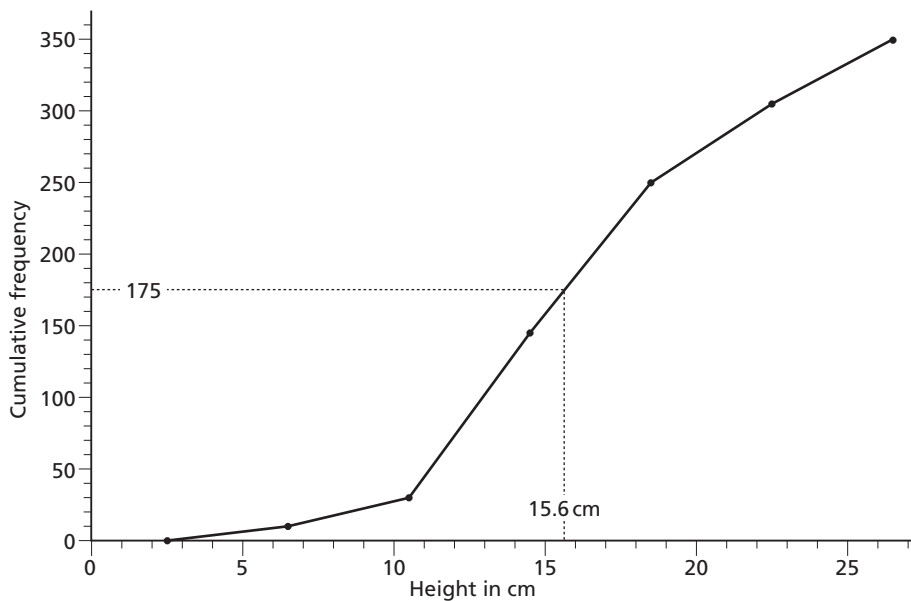
Height (cm)	Cumulative frequency
up to 6.5	10
up to 10.5	31
up to 14.5	145
up to 18.5	250
up to 22.5	304
up to 26.5	350

A **cumulative frequency graph** is drawn by plotting each cumulative frequency against the upper end of its interval (10 against 6.5, 31 against 10.5, and so on).

The shape of the graph between the plotted points is unknown. Using straight lines to join the points is equivalent to assuming that the data values are evenly spread in each interval. (This is the assumption made by the method of linear interpolation – hence its name.)

The median height of a plant is halfway up the data set, or 175 out of 350. (Strictly speaking it is halfway between the 175th and 176th. However, with a data set of this size, the difference is minimal and can be ignored.)

The line on the graph below shows that the median is estimated to be about 15.6 cm.

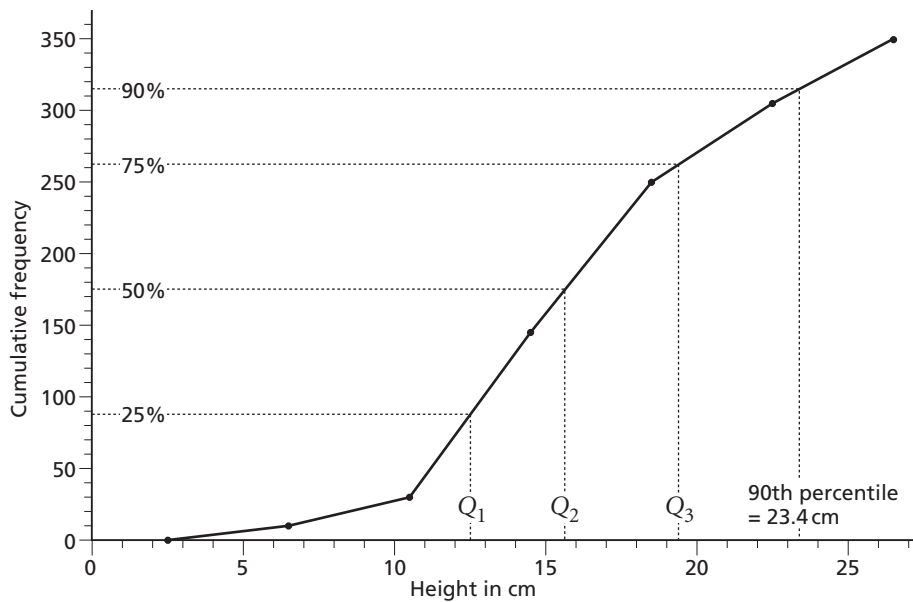


Percentiles

Just as quartiles split the data set into quarters, so **percentiles** split it into hundredths. The 90th percentile, for example, is the value that has 90% of the data set less than or equal to it.

90% of 350 = 315, so the 90th percentile can be estimated from the cumulative frequency graph as shown below.

The other lines show the 25th, 50th and 75th percentiles, which are the first, second and third quartiles Q_1 , Q_2 and Q_3 .



D1 Use the graph to estimate the interquartile range of the heights of the plants.

Unless a data set is large, there is no point in trying to calculate or estimate percentiles (except for the quartiles).

Exercise D (answers p 123)

- 1** This cumulative frequency table shows the times that 150 men spent in hospital after heart attacks. By drawing a cumulative frequency graph, or by using linear interpolation without a graph, estimate
- the median time spent
 - the 90th percentile of the times spent

Time (days)	Cumulative frequency
up to 5	40
up to 10	98
up to 15	138
up to 20	144
up to 25	148
up to 30	150

2 A tester measures the lifetimes of 200 batteries of each of two different brands, A and B.

The times, in hours, are shown in the frequency table.

Lifetime (hours)	Brand A frequency	Brand B frequency
11.0–11.5	2	8
11.5–12.0	5	14
12.0–12.5	20	43
12.5–13.0	69	50
13.0–13.5	58	48
13.5–14.0	27	30
14.0–14.5	15	5
14.5–15.0	4	2

(a) On the same axes, draw a cumulative frequency graph for each brand.

(b) Use the graphs to estimate the median lifetime for each brand.

The tester expects there to be a few exceptional batteries at both ends of the range. She decides to use the difference between the 90th and 10th percentiles as a measure of the spread.

(c) Use this measure to compare the spread of the lifetimes of the two brands.

Key points

- In a histogram, area represents frequency.
The vertical scale shows frequency density = $\frac{\text{frequency}}{\text{width of interval}}$
The frequency for an interval = width of interval \times frequency density (p 8)
- The median is the middle value (when there is an odd number of data values) or halfway between the middle pair (when there is an even number of data values). (p 11)
- The first quartile, Q_1 , has $\frac{1}{4}$ of the data values less than or equal to it.
The third quartile, Q_3 , has $\frac{3}{4}$ of the data values less than or equal to it.
The median is the second quartile, Q_2 .
The interquartile range is $Q_3 - Q_1$. (p 11)
- If n , the number of data values, is a multiple of 4, then
 Q_1 is halfway between the $\frac{n}{4}$ th and $(\frac{n}{4} + 1)$ th value, and
 Q_3 is halfway between the $\frac{3n}{4}$ th and $(\frac{3n}{4} + 1)$ th values.
If n is not a multiple of 4, find $\frac{1}{4}n$ and round up to get the position of Q_1 ;
find $\frac{3}{4}n$ and round up to get the position of Q_3 . (p 11)
- An outlier is an exceptionally low or high data value. One rule for identifying outliers is: x is an outlier if $x < Q_1 - 1.5(Q_3 - Q_1)$ or $x > Q_3 + 1.5(Q_3 - Q_1)$. (p 13)
- Linear interpolation can be used to estimate a value lying between two known values. (p 16)
- The 90th percentile is the value that has 90% of data values less than or equal to it, and similarly for other percentiles. (p 19)

Test yourself (answers p 123)

- 1** The numbers of customers for a weekday coach service over the past 9 weeks are shown in this stem-and-leaf diagram.
- | | | |
|---|---------------------------|------|
| 1 | 4 6 8 8 | (4) |
| 2 | 3 4 5 5 6 7 8 9 | (8) |
| 3 | 0 1 1 2 2 4 5 5 6 7 8 9 9 | (13) |
| 4 | 2 3 3 4 5 5 6 6 7 8 9 | (11) |
| 5 | 0 0 1 1 1 2 2 2 2 | (9) |

(a) Find the values of Q_1 , Q_2 and Q_3 .

During the previous 9 weeks the values of the quartiles were $Q_1 = 32$, $Q_2 = 40$ and $Q_3 = 46$.

1 | 4 means 14

(b) Use the quartiles to compare the two sets of data.

- 2** The areas, in hectares, of farm fields in a district range from a minimum of 45 to a maximum of 127. The values of Q_1 , Q_2 and Q_3 are 76, 90 and 103.

An outlier is defined as any value x such that $x < Q_1 - 1.0 \times (Q_3 - Q_1)$ or $x > Q_3 + 1.0 \times (Q_3 - Q_1)$.

(a) Determine whether the minimum and maximum areas are outliers.

(b) Draw a box plot for the data.

- 3** The table below shows the distribution of the weights, to the nearest 0.1 kg, of the babies born in a hospital during a 14-day period.

Weight (kg)	2.0–2.9	3.0–3.1	3.2–3.3	3.4–3.5	3.6–3.9	4.0–4.4
Frequency	3	7	10	8	4	2

(a) Draw a histogram to represent the data.

(b) Use linear interpolation to estimate the median weight.

- 4** Three swimmers Alan, Diane and Gopal record the number of lengths of the swimming pool they swim during each practice session over several weeks. The stem-and-leaf diagram below shows the results for Alan.

Lengths

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

2 | 0 means 20

(a) Find the three quartiles for Alan's results.

This table summarises the results for Diane and Gopal.

(b) Using the same scales and on the same sheet of paper, draw box plots to represent the data for Alan, Diane and Gopal.

(c) Compare and contrast the three box plots.

	Diane	Gopal
Smallest value	35	25
Lower quartile	37	34
Median	42	42
Upper quartile	53	50
Largest value	65	57